University of Pavia, Italy

University of Bergamo, Italy

Faculty of Economics and Management

PhD Degree in Applied Economics and Management

XXXIII cycle

A novel ensemble technique and its application in credit risk management

Supervisor:

Prof. Tullio Facchinetti

Tutor:

Prof. Giorgio Rampa

Candidate:

Paritosh Navinchandra Jha

Academic Year 2020/2021

I dedicate this thesis to my wife, my little son, my elder brother, and his family, and most importantly to my parents.

Acknowledgments

Over the last three years, I had the pleasure of knowing and working with different people at the University of Pavia and the University of Bergamo, Italy, to whom I would like to express my heartfelt gratitude. First, I would like to express my gratitude to my supervisor Prof. Tullio Facchinetti and my co-supervisor (tutor) Prof. Giorgio Rampa whose support during difficult times of my Ph.D. is praiseworthy.

Prof. Gianmaria Martini and Prof. Alberto Gaggero being the management of the Ph.D. program have helped me to overcome difficult times for which I am highly thankful. I express my gratitude to them for offering help to manage my different situations during the Ph.D. program.

It has been an interesting experience in all these years to work for two of the finest financial institutions in Italy as a consultant where I had the opportunity to learn new skills and get exposure to the different facets of professional know-how. I take this opportunity to thank a number of people at Banca Mediolanum and Unicredit.

My Ph.D. Journey started with Prof. Paolo Giudici and Prof. Paola Cerchiello to whom I express my gratitude for their support in the first year of Ph.D. and on a few different occassions in the later years of Ph.D.

The development of a Ph.D. thesis started under the mentorship of Prof. Silvia Figini and Prof. Pierpaolo Uberti whose guidance, comments, and insightful suggestions helped me to gain a finer understanding of the topic. I would like to express my gratitude for their support in my Ph.D. studies.

I would like to thank all the faculty team of the Ph.D. program in Applied Economics and Management and my Ph.D. classmates with whom I spent some of the wonderful times during the Ph.D. program. In this respect, I would like to thank Prof. Francesca Maggioni and Prof. Rosella Giacometti for giving their insightful suggestions to improve my expertise in subject related matters.

Finally, I would like to express gratitude to my family members specifically my wife for keeping me motivated in the program and strengthening the confidence

in me. I am thankful to a few of the friends whom I know since childhood and supported me like a family.

Last but not the least, I would like to express my gratitude to the house owner here in Pavia in giving care to me and my family. They made us feel like home and treated us as a part of their family.

Thank you all!!.

Contents

List of Figures

List of Tables

Acronyms

BMA Bayesian Model Averaging

FMA Frequentist Model Averaging

DGM Data Generating Model

FIC Focused Information Criterion

DIC Deviance Information Criterion

BIC Bayesian Information Criterion

AIC Akaike Information Criterion

MCMC Markov Chain Monte Carlo

MMA Mallows Model Estimator

JMA Jackknife Model Averaging

RJMCMC Reversible Jump Markov Chain Monte Carlo

WBIC Widely Applicable Bayesian Information Criterion

CMA Credal Model Averaging

BACE Bayesian Averaging of Classical Estimates

BMA-EM Bayesian model averaging using expectation maximization

ARMS Adaptive Regression by Mixing with Model Screening

ROC Receiving Operating Characteristic

MOO Multiobjective Optimization

MCDA Multicriteria Decision Analysis

GA Genetic Algorithm

VEGA Vector Evaluated Genetic Algorithm

MOGA Multiobjective Genetic Algorithm

WBGA Weight Based Genetic Algorithm

NPGA Niched Pareto Genetic Algorithm

RWGA Random Weighted Genetic Algorithm

PESA Pareto Envelope-based Selection Algorithm

PAES Pareto-Archived Evolution Strategy

NSGA Nondominated Sorting Genetic Algorithm

SPEA Strength Pareto Evolutionary Algorithm

RDGA Rank-Density Based Genetic Algorithm

DMOEA Dynamic Multi-objective Evolutionary Algorithm

NSGA-II Fast Nondominated Sorting Genetic Algorithm

SPEA-II Improved Strength Pareto Evolutionary Algorithm

EMO Evolutionary Multiobjective Optimization

MOEA Multiobjective Evolutionary Algorithm

MCDM Multicriteria Decision Making

MSE Mean Squared Error

RMSE Root Mean Squared Error

SMC Sequential Monte Carlo

CV Cross Validation

MML Minimum Message Length

MDL Minimum Description Length

GIC Generalized Information Criterion

MLE Maximum Likelihood Estimation

MSQE Mean Squared Quantization Error

WAIC Widely Applicable Information Criterion

KL Kullback Leibler

MCDA Multicriteria Decision Analysis

ECB European Central Bank

EBA European Banking Authority

Fintech Technology Enabled Financial Services

Regtech Technology Enabled Regulatory Services

Insurtech Technology Enabled Insurance Services

Suptech Technology Enabled Supervisory Services

BFGS Broyden—Fletcher —Goldfarb —Shanno

L-BFGS-B Limited Memory Broyden—Fletcher—Goldfarb—Shanno

CG Conjugate Gradient

SANN Simulated Annealing Algorithm

SMOTE Synthetic Minority Oversampling Technique

ROSE Random Oversampling Examples

H Hmeasure

AUC Area Under Curve

AUCH Area Under Convex Hull

MER Minimum error rate

MWL Minimum Cost Weighted Error Rate

KS Kolmogorov Statistics

GC Gini Coefficient

Sens.Spec95 Sensitivity at 95 percent Specificity

Spec.Sens95 Specificity at 95 percent Sensitivity

CTREE Conditional Inference Trees

RPART Recursive Partitioning and Regression Trees

GLM Generalized Linear Models

RF Random Forest

BAGG Bootstrap Aggregating

BOOST Boosting

GAM Generalized Additive Model

KNN K Nearest Neighbor

NB Naive Bayes

BART Bayesian Additive Regression Trees

WTM Weighted Model

MAP Maximum a Posteriori

WMCOR Weighted method using correlation

EWM Equally Weighted Method

OWM Optimal Weighted Method

SWM Squared Weighted Method

NWM Negative Weighted Method

MOP Minimum Occurence Prediction

MEP Mean Occurence Prediction

10PO 10 Percent Omission

SS Sensitivity Equal to Specificity

MSS Maximum Sensitivity and Specificity

MK Maximum Kappa

MPC Maximum Proportion Correct

MRPD Minimum ROC Plot Distance

WSD Weighted Sum of Deviations

DCP Disciplined Convex Programming

Executive Summary

The use and adoption of machine learning models in assessing the risk associated with a product and financial instruments at financial institutions are gaining importance since traditional models in practice do not assess the risk in an efficient way.

For any commercial bank around the world, managing credit risk is an important task for enhancing business profits, therefore a larger emphasis is given to mitigate any kind of losses. Although the regulatory regime advocates the use of a simpler statistical model in measuring creditworthiness, often such a model does not capture all the abstract reality. In our opinion, an ensemble approach of different classifiers including both parametric and non-parametric is a viable solution for credit scoring and managing risks in a complex environment.

The structure of this thesis can be broadly categorized into different points mainly as the following,

- Literature review
- Proposed methodological framework of model averaging and multi-objective optimization techniques.
- Results achieved from the methodological framework to improve predictive accuracy.
- Comparison of proposed model with existing machine learning model.
- Credit risk management as one of the possible area of application.

We have extensively covered a large number of studies mostly from the field of economics, statistics, and machine learning and in this respect, our proposed model is unique and novel that brings a fresh perspective to solve a wide array of problems using data as a supporting tool for the analysis.

We studied a diverse set of parametric, non-parametric, and ensemble models to compare with our proposed models, we found a different set of performances and

the evidence of results suggests to us that our proposed model generates superior or enhanced performance with compare to a few set of existing machine learning models.

The core idea in the thesis is to develop a novel model using the knowledge of model averaging and multi-objective techniques and see the performance of such models on banking data to project the idea of effective credit risk management or risk management. The evaluation of the proposed model has been done by choosing a diverse set of performance metrics to compare performances with a few of the popular machine learning models.

Our approach in proposing the idea in this thesis is useful in many ways specifically if the error of any predictive model is influenced by variance and low co-variance between models. This proposed idea helps in addressing the problem of model uncertainty by reducing variance and enhancing the performance of the model.

The primary advantage of using any model averaging technique is that we do not have to worry about finding the best or true model. The approach combines the best individual model to bring the best performance by averaging out that depends on certain criteria for estimating weights.

In addition to the idea of model averaging and ensemble model technique proposed in this thesis, the other approach to improve the performance of the model proposes a few different strategies that are Pareto-based multi-objective optimization. This approach to a certain extent offers an alternative solutions to the limitation of a single-objective optimization problem.

The novel idea proposed in this thesis on model averaging and Pareto-based multi-objective optimization is one of the useful techniques that have the capacity to enhance the predictive accuracy of any learning model and the same idea can be applied to many different problems where data analysis is the core task of interest and is not limited only to solve classification problems.

Chapter 1

Introduction

The greatest challenge to any thinker is stating the problem in a way that will allow a solution

Bertrand Russell

While writing this thesis, we aim to provide concise detail on our proposed approach to the topic of model averaging, ensemble learning, and multi-objective optimization that could be useful to applied statisticians and data scientists. Our proposed idea is novel and takes inspiration linking theory to the field of statistics, econometrics, and machine learning that broadly covers the frequentist and Bayesian methods. The novel idea proposed in this thesis could find its potential use in solving many different real-world problems that rely on data for any decision making or inferences.

All the theory that comes under the classical statistical approach relies on the parameter estimation of a single model that is assumed to be the best model among a set of competing models. The problem with the classical approach in selecting the best model is that they ignore to explain factors that lead to biased and over-fitting estimates. Due to this limitation, model averaging can be used for the crucial task of minimizing model uncertainty, minimizing error by averaging out, and most important providing alternatives to enhance the performance of the ensemble model. Keeping this motivation incorporated, our work in this thesis does not consider any model parameter that could bias or over-fit estimation of weights and is rather a few prudent ways of enhancing the performance of the ensemble model.

The developed approach in this thesis can potentially be applied to solve classification problems of different domains like the detection of stock prices manipulation, predicting the financial distress just to give few examples among others. It can

potentially be used in other fields like medicine (to detect any neuro-cognitive disorder), different emotion recognition through image analysis, identification of factors for better forecasts of GDP growth, identification of factors for business model innovation ,and so on.

Many of the studies in model averaging centers around either frequentist methods or Bayesian methods. We compute the weighted mean of the prediction or estimates in the frequentist approach from every single model and the estimation of weights depends on certain information criteria like Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Cross Validation (CV) just to name few. For detailed information on different criteria, the reader can refer to chapter 3. In many contexts, the frequentist approach sometimes is seen as a way to trade-off bias and variance.

On the other hand, the weights are either posterior probability assuming that the model is true or obtained based on some other predictive methodology and these weights are often constrained to be non-negative in the Bayesian context.

Our approach in proposing the idea of model averaging is useful in many ways specifically if the error of any predictive model is influenced by variance and low co-variance between models. This proposed idea helps in addressing the problem of model uncertainty by reducing variance and enhancing the performance of the model.

The idea of model averaging is quite historical and has been talked in many different filed of study. The approach of combining the model dates back to 200 years ago when it was exercised to combine the coefficient of regression. Laplace in the year 1818 computed and compared the properties of estimators like least squares and weighted median. More details on the work of Laplace can be found in [118].

The primary advantage of using any model averaging technique is that we do not have to worry about finding the best or true model. The approach combines the best individual model to bring the best performance by averaging out which depends on certain criteria for estimating weights.

Beyond Bayesian and frequentist approaches to model averaging, the literature in the last 20 years has grown significantly both in terms of new theoretical approach and application to many different problems. Our proposed idea of model averaging in this thesis, although useful to many relevant problems, it is more specific to ensemble modeling from a theoretical perspective and risk management from an application perspective.

Many ensemble learning algorithms can be said as a supervised algorithm as they

require training before making predictions. Any ensemble model produces better results when we have a pool of diverse models to work with for a given problem. The *law of diminishing returns* is very much prevalent in ensemble model construction as the idea of including any less or more classifier than an ideal number of component classifiers influence directly the accuracy of combined or weighted model.

Referring to the study by [121], any ensemble method is like *wisdom of crowds* that provides an alternative for superior decision making. The ensemble model in general is diverse, independent, decentralized, and aggregation of outputs from a pool of models.

When the ensemble model is used in the context of a classification problem, it is usually based on a framework composed by the following:

- A labeled dataset that is used as a training set.
- The base learner is a type of algorithm that establishes a relationship between input and response variables.
- The generator is a type of function that is responsible for generating diverse classifiers.
- The combiner is a type of function that combines the classification of diverse classifiers.

We specifically deploy our proposed ensemble model in the context of credit risk management. Many decision involved in the credit lending business makes it of the utmost importance to rely on accurate models that can provide information about a customer in the best possible way. For this reason, machine learning models are getting significant attention as a way to better understand the creditworthiness of a customer.

Model uncertainty is obvious to see in many cases, therefore it becomes difficult to choose a single best model that can be generalized to serve different purposes. Classification trees and the random forest do answer model uncertainty problems to a certain extent by providing superior performance on a few problems but is not common to a whole set of different problems. However, the empirical evidence in our thesis supports that our proposal of weighting the model significantly improves predictive performance as a way to enhance performance with the model combination approach.

Model selection in statistical science is studied as a function of different criteria to provide some answers to model uncertainty. We often look for a certain type of

quality or measure in choosing the model and generally, there are certain trade-offs amongst them. We have the possibility to deal with such kind of situations borrowing the diverse knowledge from the field of optimization. However, it is not possible for the most simple model to optimize more than one objective function. What we all attempt to do with a single objective problem is either to minimize or maximize a single function over its domain where the function is called the objective function.

Apart from model averaging and ensemble model technique proposed in this thesis, the other approach to improve the performance of the model uses a few different strategies that are Pareto-based multi-objective optimization.

The limitation of a single-objective constrained optimization problem is well-known to solve many real-world scenarios. For instance, there are examples where the use of optimization is used or considered as a promising field of study in machine learning for minimization of different types of error rates but none of the studies so far gives a handful guide in the context of a machine learning algorithm. For more details, refer to [49]. To accurately deal with such a situation, one could extend error or performance analysis of the machine learning problem from the perspective of multi-objective functions and any such function with a constrained optimization problem. Since the error and performance of any model are closely interrelated, we always need to make sure to balance the inherent error-accuracy trade-off.

Often in practice, we come across varying performance or error reduction of a model and we get puzzled to observe divergence or convergence among models. Basically, it is hard to get absolute convergence from each of these models and essentially a model is said to converge when the value of loss function moves close to minima (local or global) with a decreasing trend.

To solve any problem through modeling, we often look for a solution that is optimal in some sense but there arise instances where we might get a different possible solution and so is the need of decision space that evaluate solution obtained from different objective functions.

1.1 Objectives of the thesis

The primary aim of this thesis is to propose a few novel methods in terms of model averaging and multi-objective optimization that provides appealing solutions to improve the performance of the machine learning model. Although the proposed idea in this thesis is applicable to many different problems that have something to do with data or the use of the machine learning model, our focus of application with

respect to the developed idea is basically attempted to predict the probability of default.

Given the state of the art in credit risk management, our effort in this thesis is to introduce a new weighted model of averaging approach which is more prudent rather than the Bayesian or information-theoretic approach in reducing prediction error. The novel idea proposed here has the capacity to enhance better performance compare to the single best model (parametric, non-parametric, and ensemble model) as evident in their empirical findings in chapter 6 and offers the possibility for an effective credit lending process to decision-makers at a financial institution.

Following this effort, we studied a range of models that were parametric, non-parametric, and ensemble models. We deployed those insights in developing a novel idea of a weighted average approach to predict the probability of default. The standard approach in making predictions does not identify a single best model due to limitation in data for several plausible combinations of predictors and therefore availability of different modeling approaches offers a solution to this problem.

Our objective in this thesis is to propose a new technique of weighted model averaging that do not take into account any averaging model parameters. The weighted model is compared based on a few key performance measure such as Hmeasure (H), Area Under Curve (AUC), Area Under Convex Hull (AUCH), Minimum error rate (MER) and Minimum Cost Weighted Error Rate (MWL) that helps to examine predictive capability, discriminatory power, and stability of the results.

The primary contribution of this thesis is to enhance predictive performance irrespective of the given number of model choice. One of the biggest advantage of using any model averaging approach is to reduce variance and understand the uncertainty involved in model parameterization and structure.

In this thesis, we also propose a few different strategies that is the Pareto-based multi-objective optimization to enhance the objectives and performance of the machine learning model, since it is well-known fact that the solution methods equipped with a single objective function are not sufficient enough to deal with many real-world problems in machine learning.

To tackle such challenges, the proposed approach provides a new perspective and perhaps a better choice to use an optimization techniques to a diverse set of machine learning problems. More specifically, our methodological approach is novel in the sense that it incorporates theories of optimization and information science to have a new understanding and performance comparison tool for learning algorithms. The methodological approach developed in this thesis is useful for many real-world

problems that seek the attention of multi-objective optimization or Pareto-based multi-objective optimization as a solution method.

1.2 Organization of the document

The thesis is organized as follows:

- **Chapter 2** presents a conscientious review of the relevant literature and state of the art methods in support of our proposed approach. The strategy to review articles mainly comes from the field of model averaging, ensemble learning, and Pareto-based multi-objective optimization. Section 2.1 is dedicated to the discussion on model averaging literature review focusing on the articles of the past 20 years. Subsection 2.1.1 discuss the literature review on ensemble learning starting from the historical reference as to how the idea of ensemble started and their contextual usefulness for solving problems in Machine Learning. The discussion on the literature review of Pareto-based multi-objective optimization is dedicated in section 2.2 starting with an initial comment on the historical idea of multi-objective optimization.

- **Chapter 3** details the proposed models from theoretical perspective. Section 3.1 explores background information on model averaging from Bayesian, frequentist, and other information criteria. Section 3.2 contributes the knowledge behind the model averaging that can be possibly applied to many different problems especially in the field of machine learning and credit risk management. Section 3.3 discuss in detail the factors that influence any estimation of weights and proposes a different strategy for estimating weights. Section 3.4 presents an additional discussion in support of the model averaging approach. Section 3.5 discuss possible properties in the context of the ensemble model. Section 3.6 describes our methodological knowledge behind multi-objective optimization from the perspective of the Pareto-based approach, information science, and machine learning as an integrated approach to enhance the performance and accuracy of the model.

- **Chapter 4** discusses the dataset in section 4.1, the nature of dataset and strategies that were adopted to treat data for the analysis in section 4.2. The chapter also describes, in brief, the software packages and software environment in which the analysis was carried on the dataset in section 4.3.

- **Chapter 5** describes in brief a set of classifiers like parametric, non-parametric, and ensemble models in section 5.1. Section 5.2 discuss in brief optimization models that were studied in developing weighted model average and Pareto-based multi-objective optimization strategies. The evaluation of proposed models with respect to classifiers and optimization models were compared using a set of diverse metrics that is briefly discussed in section 5.3.

- **Chapter 6** discuss the results achieved from the proposed model using model averaging technique in section 6.1 and using different multi-objective strategies in section 6.2. All other additional results is discussed in section 6.3.

- **Chapter 7** concludes the work demonstrated in this thesis and provides possible future direction that could be helpful to enhance the filed further.

Chapter 2

State of the art

Machine learning is a new programming paradigm, a
new way of communicating your wishes to a computer.

Anonymous

This chapter describes the state of the art in the referred literature on model averaging, ensemble learning, and Pareto-based multi-objective optimization techniques to enhance the performance of machine learning models in relation to our work in this thesis.

We structure this chapter in two sections focusing on model averaging techniques in Section 2.1 . The section discusses the relationship with our proposed approach, which is rather prudent than any Bayesian or information criterion from the logical point of view. Section 2.2 establishes the literature review relationship with our proposed methodology on multi-objective optimization from an interdisciplinary point of view. The approach to select the relevant literature in this thesis has been mostly from the field of economics, statistics, and machine learning.

Models are selected and constructed typically for a variety of purposes that try to search and explain patterns in the existing data concerning to some underlying structures. In many given problems, these underlying structures often are not known ,and this leads to prediction error among the considered models. One of the alternatives to address the model uncertainty problem is to use model averaging techniques for plausible models, which is widely known in the academic community of various disciplines.

2.1 Model averaging: literature review

Over the last two decades, the focus on model averaging literature concerned around Bayesian Model Averaging (BMA), where we set prior and treat model parameters as unknown based on a Bayesian paradigm, and Frequentist Model Averaging (FMA), where the chosen weights are determined under repeated sampling and asymptotic optimality.

In a situation where it is not possible to find a single best model, model averaging proves to be a useful technique for reducing prediction error through reduced variance. We present here relevant literature that influences our work in this thesis starting from the earliest contribution by J. Armstrong [3], to most recent studies.

Domingos [44] used a set of rules to compare bagging and partitioning methods. The paper concludes that the Bayesian model averaging error rates are consistently higher concerning to other methods, and this might be due to a marked tendency to overfit on the part of BMA . However, this is a bit contradictory in the context of our work where we use BMA as one of the ensemble models for classification problem and it seems that the error rate for BMA is lower as compared to other ensemble models. It is further possible to create a separate ensemble using the posterior distribution of the BMA results.

A review of different Bayesian procedures for model averaging like the conventional approach, Bayes information criterion, intrinsic Bayes factor, and fractional Bayes factor is reported in the study of Berger et al. [11], which illustrates many different examples with proper reasoning as to why the Bayesian approach is the best method for model selection to deal with model uncertainty. Our work relates to this with the idea of disseminating prior probability between competing models with an appropriate choice of weights. This further implies that the probability distribution of our constructed ensemble model follows a mixture of each model probability distribution.

The central idea always remains the same, which is to assign weights to models based on the proportion of time a model has been used to produce results of the highest likelihood within the set of models. This kind of approach is also known as Bayesian model averaging using expectation maximization and looks more frequentist than Bayesian.

Many authors like Watanabe [126] and Gelman et al. [59] suggest to use a new information criterion called Widely Applicable Information Criterion (WAIC) that is derived in the Bayesian framework as an alternative to AIC . This kind of criterion is based on uninformative prior and is calculated based on two logic. One way is to

keep the log pointwise predicted density across posterior simulations for each of the predicted k values as $\log \prod_{i=1}^{k} p_{posterior}(f_i)$. The other way is to use a bias-correction term as $\sum_{i=1}^{k} var(\log\ p(f_i|\theta_s))$ where the use of var is a sample variance between all samples of posterior distributions of parameter θ. The approach is more like using some kind of likelihood function that observes data for any posterior distribution. There is also a penalty parameter for model complexity that is proportional to the given variance of such likelihoods across Markov Chain Monte Carlo (MCMC) samples, which ultimately helps in deriving model weights analogously.

To derive model weights in the context of the information-theoretic framework, models that are closer to data as quantified by the Kullback-Leibler divergence receive higher weight concerning to models that are far away from data. Akaike [1], and Burnham and Anderson [86] suggest several approximations of Kullback-Leibler divergence, and all such indices can be calculated for models with likelihood function and known parameters as $AIC_k = -2\ell_k + 2p_k$, where ℓ_k is the log-likelihood of any model k. One of the few alternatives proposed by [68] and [113] is to use Mallow's criterion to reduce mean squared error by penalizing model complexity to $-2\ell_k - k + 2p_k$. In terms of penalization factor for model complexity between AIC and BIC is that AIC uses the constant 2 while BIC uses $\log(n)$. A manipulated version of AIC where the fitness of the model is assessed concerning to focal predictor value called. However, such an approach can never be seen as a superior approach since there are other variants like cross-validation and model pre-selection methods that works as a weighting procedure for enhancing the performance of the model.

A different approach of model averaging, and in particular the contrast posterior analysis gave a sampling model, is explained by Fernandez [51]. Their work tries to provide an automatic or benchmark prior structure that could be used in many such cases where there is little or no subjective prior information. In relation to benchmark prior specification in linear regression context with model uncertainty, the paper compares the predictive performance of many different priors citing examples that are classically discussed in Economics.

A study of model averaging by offering equal weight as a starting reasonable point with trimmed mean if the averaging techniques resulted from five or more methods are reported in the study of Armstrong [114]. They emphasized that different weights can be used if one has good domain knowledge or information on which method should be most accurate. The technique is useful if there is uncertainty in the model selection and if one wants to avoid a larger prediction error. Getting inspiration from their work, we followed a slightly different approaches of equal weighting which are

not in any sense the trimmed mean but simply a probabilistic measure with a larger set of models that are parametric, non-parametric, and ensemble models.

We can compare BMA with non-Bayes forms of model averaging such as stacking (which is simply a way of combining multiple models as a meta-learner [127]) where model weights are not based on posterior probabilities but rather on techniques using cross-validation. Bertrand [29] studied a sequence of examples by choosing model lists and Data Generating Model (DGM) to assess the risk performance of BMA and stacking. The robustness properties suggest that non-Bayesian techniques like stacking perform better than BMA in all possible settings.

In the context of our work in this thesis, the concept of stacking is pretty simple as we try to get a unified approach of weighting the model by minimizing co-variance between the models and is different from the frequent practice of stacking that relies on minimizing leave-one-out mean squared error.

A new criterion of model averaging was introduced by Claeskens et al. [77] and is called Focused Information Criterion (FIC). The authors present the shortcomings of other criteria like AIC, BIC, and Deviance Information Criterion (DIC). They explain a perspective that focuses on a single parameter of interest rather than multiple parameters of interest, which is a better estimate for the precision of the sub-model estimator. Following this idea, our parameter of interest is to choose few prudent ways of information criterion that allows selecting weights in an optimal way to be able to enhance predictive performance.

A slightly different approach in contrast to famous approaches was suggested by Barbieri et al. [7] for model averaging under the Bayesian framework, where they select the optimal predictive model that is often the median probability model defined as a model consisting of those co-variates that have a posterior probability greater than or equal to 1/2 of being in a model.

A concept of *thick modeling* for model averaging is discussed in the study of Granger et al. [62] that is based on many specifications in contrast to the usual technique to choose the best criterion by testing and then use them as an output. Basically, their approach is a portfolio selection and forecast combination that suggests using bootstrap techniques as a sub-optimal solution.

The study proposed in [21] states that model averaging literature is being poorly reflected in understanding the foundations of AIC and their comparison with BIC . They further extend these arguments saying that AIC and BIC for model selection should not be seen from Bayes versus a frequentist perspective. The choice of using AIC or BIC is basically an intent-based model inference that produces useful results

of model averaging. Following their work, Our intent-based model inference is a different strategy of estimating weights that are based on the criterion which gives a diverse choice of constructing ensemble model and enhances the performance of the model.

The study of Nicole Augustin [4] proposed two approaches to account for model selection uncertainty based on survival data. The first approach uses BMA for the proportional hazard model where the averaging technique on a set of possible models is done using weights estimated from bootstrap resampling. The other approach is simply based on prognostic models. The paper shows that there is a lack of formal justification and requires an additional analysis that might give better explanatory power to the considered model.

A general technique of optimal model assessment was discussed in [116] using data perturbation, which ultimately helps in model selection and model combination. Using a frequentist perspective and model combination approach, the authors develop a procedure for determining a few optimal parameters, like weights, that is used in model combination to achieve better predictive accuracy by controlling the bias and reducing the variance. Following their work, our approach for determining optimal parameters like weights is based on the solution of different optimization algorithms considered and in this respect, we have approximated value as the optimal weight for constructing the ensemble model.

In many situations, we observe that a simple combination of models that do not take into account the correlation between forecast errors is often the best approach to estimate and obtain optimal weights. In [122], Timmermann further discusses the advantages of this approach in model combination under asymmetric loss, point, interval, and probability forecast. However, the very underlying solid reasoning behind the simple combination of models is not well stated. Following their work, Our approach for model combination is primarily based on co-variance and contradicts the idea that correlation is the best approach for estimating weights. In fact, the estimated weights using correlation discussed in the empirical work in chapter 6 proves that achieving a weighted model through co-variance is a superior technique against correlation.

A different method was suggested in [34], where posterior probabilities are estimated and the model parameter averaging is done using MCMC under the Bayesian framework. There is a pool of models that is updated at each iteration where posterior probabilities are obtained by averaging continuous weights for each model and using these weights the sample average parameters are obtained from each

iteration, helping in achieving posterior densities for parameter difference between models in the context of parallel sampling.

An extension to the standard approach of Bayesian model averaging is discussed in detail by Eklund et al. (2007) [48], where the weights for the formed averaged model come from the predictive likelihood instead of the standard marginal likelihood.

Such use of predictive measure often protects from the overfitting problem of in-sample and enhances the predictive performance. This is largely due to the idea that combined weights have good large and small sample properties. Following their idea, our combination of weights does not have anything to do with sample properties and in fact, the weights are constrained accordingly to suit the construction of optimization problem stated in chapter 3 .

A new strategy was suggested by Hansen, B.(2007) [68] for selecting weights that are called Mallow's criterion and are simply an estimate of the average square error from the model average fit. The paper discusses that Mallows Model Estimator (MMA) achieves the lowest possible squared error in a class of discrete model average estimators that are asymptotically optimal.

An integrated approach of all available literature on information criterion was carried out in the study of GerdaClaeskens et al. [27] for model averaging to investigate the idea of choosing the best model among candidate models to avoid any real danger of overfitting. This work is the first of its kind to synthesize research and practice from this active field to choose model selection criteria like AIC, BIC, DIC and FIC to better understand uncertainties involved in the model selection. Following their idea, our approach for understanding uncertainty does not necessarily have to follow certain information criteria but can be done through other methods like Kullback-Leibler divergence to be precise here.

The study of Garthwaite et al. [58] presents a slightly different discussion where prior weights are chosen for the task of model averaging. In this method, models that are similar are given smaller weights with respect to models that are distinct among each other. Such an approach helps in offsetting those feature of the model that are exaggerated due to correlation value for model averaging, and predictive variance of all models are investigated using the empirical Bayes method to achieve smaller variance. Our work slightly supports this idea to choose prior weights when we use equal weight as one of the strategies for constructing a weighted model while all other weighting strategy is not necessarily a process of choosing prior weights.

A new approach was introduced by Hansen [69] for model averaging called Jackknife Model Averaging (JMA). This approach selects the weights by minimizing

a cross-validation criterion. This criterion helps in obtaining appropriate weights that come from the simple application of quadratic programming and is asymptotically optimal. This method helps in achieving the lowest possible expected squared error. The authors claim the efficiency of JMA through Monte Carlo simulations, comparing their proposed method with existing averaging methods in presence of heteroskedasticity. Following their idea, our approach is slightly similar to extract appropriate weights but with a different approaches like minimizing co-variance in the set up of quadratic programming.

The study carried out by Hoogerheide et al. [79] introduces a novel approach for the model combination using Bayesian schemes that allow for parameter uncertainty, model uncertainty, and robust time-varying model weights. The method is tested against financial and macroeconomic data to compare predictive accuracy and economic gains, and this result outperforms all other combination schemes based on time-varying model weights. Our approach in following this idea is to choose a model combination process based on fixed and random weights that do not vary with time and is a very data specific approach.

A new method of model averaging called *Bayesian adaptive sampling algorithm* was proposed by Clyde [31]. It works by sampling models without replacement from the space of models. This method orders a model in a number of iterations with potential variables under consideration and is tested against both simulated and real data. The Bayesian adaptive sampling algorithm claims to outperform Markov Chain Monte Carlo methods.

An approach where predictive models are seen as a weighted combination of linear models is evaluated in the study of Geweke et al. [60] using a log predictive scoring rule. Their optimal approach for combining models is to include all models that have positive weights, while the models that are inferior according to the scoring criteria are deleted. In our approach, we do not use the log predictive score value but a simple probability score. The constrained weights in our case is a positive weight that falls in the line of weighted model combination method but is a different approach in an abstract sense.

Hastie [73] discussed a novel way of selecting and averaging the model based on Reversible Jump Markov Chain Monte Carlo (RJMCMC), highlighting the limitations of the Bayesian approach. The analysis carried by Bayarri et al. [10] addresses the use of a model averaging technique under the Bayesian framework for variable selection. Their results claim a new model selection objective prior with useful properties.

A new criterion of model averaging called Widely Applicable Bayesian Information Criterion (WBIC) was introduced in the study of Watanable [120], which is a generalized version of BIC onto singular statistical models where the average likelihood function over the posterior distribution is defined by $1/logn$, and n is the number of training samples. The advantage of WBIC is that it can be numerically computed without having to know the true distribution.

The literature review done in [59] reflects some of the well-known information criteria like Akaike, deviance, and Watanabe-Akaike from the Bayesian perspective where their intention is to estimate expected out-of-sample-prediction error using a bias-corrected adjustment of within-sample error. The primary contribution of this paper is to review all the available information criterion from the perspective of the Bayesian predictive framework and better understand them in practice through a few small examples.

The idea of Credal Model Averaging (CMA) as an extension of BMA for strengthening robustness check is available in the study of Corani et al. [35], which reveals that the model substitutes single prior over the models by a set of priors. Unlike BMA, CMA does not behave like a random guesser as it detects prior-dependent instances that, in other sense, is a weakness of BMA .

The study by Moral-Benito et al. [98] presents the notion of model averaging in the context of Economics where uncertainty in the model selection is often ignored primarily due to biased choice of selecting a model. The biased choice is due to the fact that possible space of models and the selected model is believed as if generated from data. The paper brings a comprehensive review of model averaging techniques integrating much different literatures of Economics.

The figure 2.1 presents a graph that infers information of the highly cited authors on model averaging for considerable period of time.

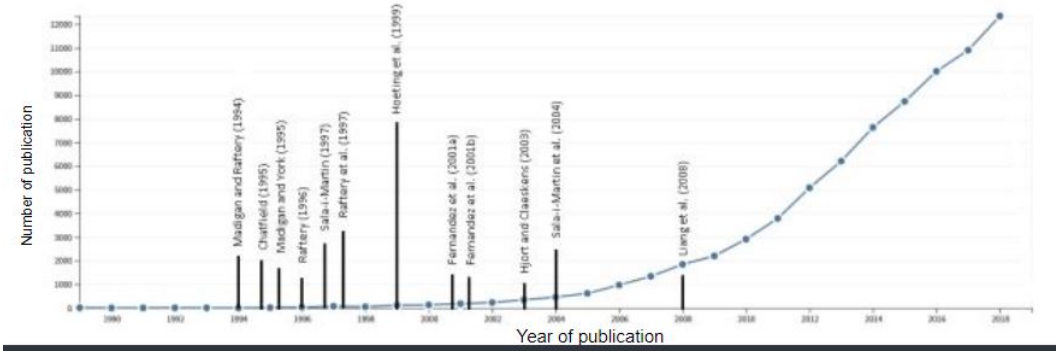

Figure 2.1: Highly cited authors in research area of model averaging. Source: [117].

The information in table 2.1 lists different model averaging technique and their computational speed in general to execute such methodology.

Table 2.1: different model averaging approach and computational speed.[45].

model averaging technique	computational speed
RJMCMC	Slow
Bayes factor	Slow
BMA-EM	Moderate
Fit-based weights	Rapid-slow
ARMS	Moderate
Bootstrapped model weights	Slow
Stacking	Slow
JMA	Slow
Minimal variance	Rapid
Cos-squared	Rapid
Model-based model combinations	Moderate
Equal weight	Rapid

The study by Lessman et al. [92] presents a review of a 10 year study as a state-of-the-art classification algorithm for credit scoring from the perspective of machine learning. They provide a unified approach of the credit scoring field integrating broken literature and updating their study based on Baesens et al. [5], which compares many novel classifications in the field of credit score modeling. The paper provides many independent assessments of scorecard methodologies and presents a viewpoint as new baseline research to be helpful for decision-makers at a financial institutions.

Very similar to the previous idea, we have considered classification models for scoring probability of default as a case-based study but our approach is actually creating a scorecard is different and unique in the sense that we have focused on constructing a novel ensemble model from a set of parametric, non-parametric and ensemble models. This is possible because of the different weighting strategies that we adopted to construct a novel ensemble model for developing a scorecard for effective credit risk management.

The information in figure 2.2 presents progressive view of articles in numbers that has been added on the topic of model averaging in the last 30 years.

A model averaging approach with optimal weights performs often poorly in many

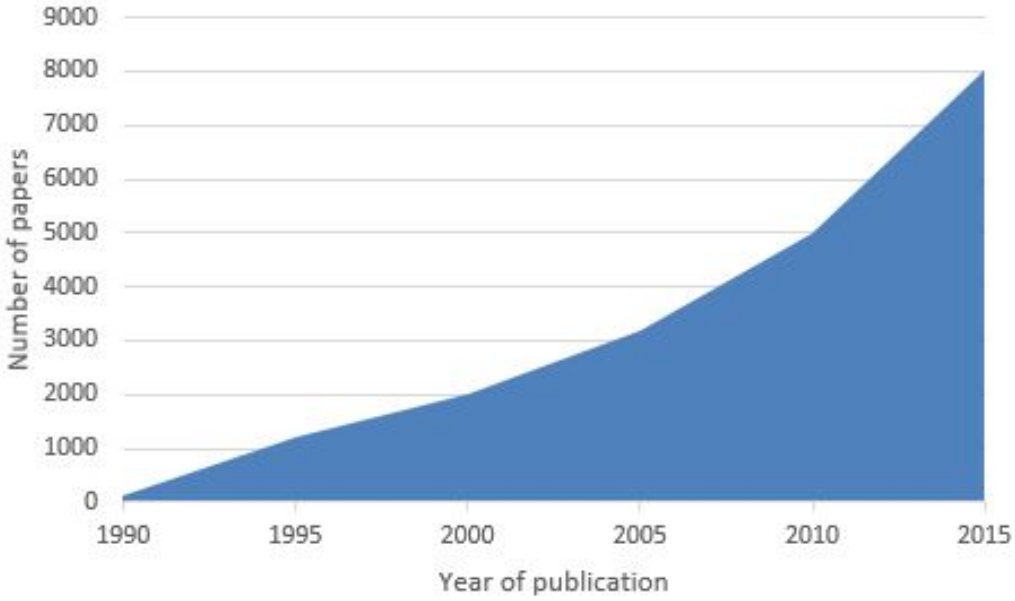

Figure 2.2: A progressive view of 30 years of model averaging articles.Source: [117] .

applications and is well described in their study by Gerda et al. [28] . The paper explains the comparison of fixed and random weights for optimal weights derivation. The approach confers the disadvantage of random weights over fixed ones which makes the model average combination a little more biased, and variance will be larger with respect to fixed weights.

In contrast to this, the story behind our approach is different and supports an alternative view of argument. For instance, the idea of choosing optimal weight based on a constructed optimization problem do provide better results and the same is true for choosing fixed and random weights.

An alternative approach to Bayesian averaging called Bayesian Averaging of Classical Estimates (BACE) to compare model reduction strategies was introduced by Blazejowski et al. [13], which helps to obtain the most probable set of determinants along with posterior parameter estimates that are based on averaged or combined model space.

Desbordes et al. [114] suggest a flexible approach of Bayesian model averaging that highlights few key insights such as not all determinants will be relevant for model uncertainty, there is a valid assumption on a homogeneous slope, cross-sectional and time-series relationship can diverge even in case of omitted variable bias.

Schomaker [41] attracts a few relevant literature reviews that explains when and

when not to use optimal model averaging technique especially in the use of standard errors and confidence intervals in dealing with model selection problem. Overall, the paper puts some benefiting arguments contradicting some popular beliefs in the forecasting literature.

2.1.1 Ensemble model: literature review

After knowing the latest developments in the field of model averaging over the last 20 years, we shift the attention of readers to model averaging in the context of ensemble learning, and more specifically with machine learning which is the primary focus of our research work. For this purpose, we describe a bit of the origin of the ensemble method from a historical perspective.

Marquis de Condorcet in the 18th century wrote an article entitled *"Essay on the application of analysis to the probability of majority decisions"*. This work introduces Condorcet's jury theorem that refers to a jury of voters to decide on a binary response variable. More specifically, if every voter has a probability p of being correct and all majority voters have L probability of being correct then generally two conditions are satisfied:

1. if $p > 0.5$ implies $L > p$, and
2. if number of voters increases to infinite then L converges to 1 for any $p > 0.5$.

However, this theorem has certain limitations, mostly concerning the assumption made, like voters are independent and there are only two possible outcomes. A correct decision is made if these two preconditions are met. Therefore it is naturally possible to combine the votes of the jury whose decision is slightly better than any random vote.

The notion of this theorem can be applied to a supervised learning problem where a strong learner behaves like an inducer for labeled training data set that can produce arbitrarily accurate classifiers, whereas a weak learner produces classifiers that are slightly better than random classifier.

Furthermore, the idea of weighting the model is conceived from history in the 19th century when Sir Francis Galton, while visiting a livestock fair, became curious by a simple weight guessing contest, and during the contest, he realized the average of all guesses from different persons is equal to the exact weight. After his visit, he shared with the scholar community the power of averaging or combining multiple simple models to achieve better predictive performance.

We know very well that any ensemble technique that combines multiple expert models keeping bagging and boosting as a base learner or commissioning model performs significantly better than any single model, and this represents the state of the art of learning approaches. Our research work mostly centers around the use of multiple classifiers where the main idea is to find a few intelligent ways of weighing the classifiers so that the combined classifier outperforms any individual classifier.

The idea of using ensemble methods to supervised learning started in the late seventies when Tukey in 1977 suggested an approach of combining two linear regression models where one model is fitted to the original data and the other model is fitted on the residuals [78] . For an extensive list of materials after the seventies that were focused on exploring the idea of ensemble methods is reflected in many studies. For instance, Dasarathy and Sheela in 1979 [37] suggested using two or more classifiers for partitioning the input space into units of smaller spaces to see the performance of classifiers in individual spaces as an alternative to combining only regression models.

The popularity of ensemble methods, however, was stronger in the nineties. For this reason, it is good to refer to the study of Hansen and Salamon (1990) [70], Schapire (1990) [112], Freund and Schapire (1996) [54] for any further additional details.

Breiman [18] introduced to the world in 2001 a famous algorithm called *random forest*, which is a combination of tree predictors and is dependent on the values of random sample independent vector which has more or less the same distribution in the forest. The algorithm is based on the random split of features that yields an error rate comparable to Adaboost but generally is robust with respect to noise. Any generalized error of the classifier in the forest depends on the strength of individual classifiers and the correlation between them. Moreover, the generalized error of the classifiers in the forest converges to almost surely to a certain limit if the tree classifier in the forest becomes large enough. Following this work, Our approach for generalizing the error in the collection of classifiers depends on co-variance between models, and correlation is used simply as one of the methods for computing alternative weights for constructing an ensemble model.

In [95], Frelicot introduced the idea of combining classifiers with two reject options that operate in a two-steps procedure and differ in terms of managing the ambiguity and distance rejection. The procedure is executed using a concept from the theory of evidence and, based on some probabilistic rules, the classes are rejected before combination. Once the combination is done a rule is established either to

classify or reject patterns due to distance or ambiguity.

The idea of building a decision tree based on the procedure of randomized sampling is suggested in 2001 by Kamath et al. [84] . By randomization, they mean using random samples of training data where a conventional tree algorithm can be run by randomizing the induction algorithm itself. The primary objective of the paper is to introduce the novel randomized tree based on the induction method that uses sampling criteria to determine the split at the node. Their suggested approach proves through experimental results that it is competitive in terms of accuracy and superior to boosting and bagging in terms of computational cost. Following their idea, we have adopted randomized sampling for training data using q-fold cross-validation for each model to achieve better accuracy with compare to other tree-based algorithm.

Seewald et al. [115] suggested a meta-classification technique to identify and correct incorrect predictions at the base level. The approach is graded predictions as a meta-level class, which is slightly different from stacking that uses predictions of the base classifier as a meta-level attribute. Their experimental results suggest superior performance when using grading and stacking as meta classification schemes with respect to voting and selection by cross-validation. Our proposed technique in this context is to construct an ensemble model that logically can be categorized as one of the meta-classification models. The combination process in our context works on the idea of estimating weights in a way that is able to enhance the performance of the model.

A new ensemble technique called *Negative Correlation Learning* (NC) was proposed by Brown et al. [20] in 2003 borrowing the concept from evolutionary computational theory where they prove this technique as a decomposition of Krogh and Vedelsby which is nothing but a simple derivative technique. There are several advantages of this method, among which is that we are able to find parameter bounds, rescaling the estimation of ensemble covariance to achieve a system that helps to have better predictive performance. This idea may or may not be directly linked to our proposed approach as one of the weighing strategies is to allow negative weights for constructing ensemble learning. The negative weights could be due to higher co-variance or correlation between models.

In [29], Bryll introduced *attribute bagging*, an alternative approach for improving and enhancing a classifier using a random subset of features. It is a wrapper method for any learning algorithm that takes an attribute subset size and randomly selects a subset of features that helps in obtaining the projections of the training set where

ensemble classifiers are built. The paper shows with further examples that the attribute classifier performs better than bagging both in terms of accuracy and stability. Relating to our work, we did not restrict to use a random subset of features for enhancing classifiers but a set of features that are ranking based features by impurity measures like Gini impurity.

Dimitriadou et al. [42] in 2003 incorporate the idea of the ensemble from regression and classification problem to the clustering algorithm as it helps to improve the quality and robustness of the results. The primary idea of aggregation or averaging relies on minimizing average dissimilarity apart from several other ideas proposed in the paper for aggregation.

In 2003, Kuncheva et al. [91] suggested that diversity is a key point while combining classifier and measuring diversity through any formal theory is often unknown. The paper describes ten statistics to measure diversity with an experimental set up to see the relationship between accuracy and diversity. Following this idea, our approach for constructing an ensemble classifier is to include diverse classifiers that is parametric, non-parametric, and ensemble models. The measure of uncertainty between models through statistical distance is a diversity measure in our case.

Melville et al. [96] presented a methodology of constructing diverse hypotheses of ensembles using an additional set of artificially trained examples. The proposed technique is general and simple in the sense that it keeps a strong learner as a base while creating a committee of ensembles. Their experimental results using decision tree induction as a base learner confirm to achieve higher accuracy compare to bagging and boosting.

The study of Tumer in 2003 [104] explores the input decimation as a method that selects feature subsets to make them able to discriminate among the classes and then later decouple the base classifier. The paper further explains that this is only possible if there is less correlation between classifiers, which in turn outperforms ensemble that uses all input features in terms of performance and accuracy.

Caruana et al. [22] in 2004 suggested a new technique of creating ensembles from libraries of thousand of models which are constructed using different learning algorithms and parameter settings. In order to maximize performance for ensemble models, the forward stepwise selection is used, which in turn also optimizes performance metrics like ROC area, accuracy, cross-entropy and mean precision.

Many ensemble techniques come with their own limitations and few of these limitations are addressed by Chawla et al. [25] in 2004 for two popular ensemble techniques that are bagging and boosting mainly dealing with massive data size. To

solve this, the authors suggest the use of voting many classifiers that are built on subsets of large data size and is considered one of the promising approach on top of bagging and boosting to learn from massive data sets. Experimental results suggest that thousands of classifiers can be set up in a distributed way to achieve faster, accurate, and scalable solutions.

Dzeroski et al. [46] presented the idea of evaluating several states of the art methods for constructing heterogeneous classifiers based on stacking methodology. The Stacking approach with a probability distribution and multi-response linear regression always performs best selecting the classifier from the ensemble if cross-validation is a criterion for selection. Moreover, the paper describes two more extensions of this methodology where one method is to extend a set of meta-level features and the other is to use multi-response model trees to learn at the meta-level.

The information in table 2.2 lists different model categories of ensemble model and their weighted approach as an ensemble scheme.

Table 2.2: Model categories and ensemble scheme.

Model	Ensemble scheme
AdaBoost	Weightining (input manipulation)
Bagging	Weightning (input manipulation)
Random forest	Weightning (ensemble hybridization)
Random subspace methods	Weightning (ensemble hybridization)
Gradient boosting machines	Weightning (output manipulation)
Error-correcting output codes	Weightning (output manipulation)
Rotation forest	Weightning (manipulated learning)
Extremely randomized trees	Weightning (partitioning)
Stacking	Meta-learning (manipulated learning)

Liu et al. [93] examined in 2004 the relationship of ensemble size with accuracy and diversity. The paper proposes the idea of compact ensembles where it is possible to keep small ensembles while maintaining accuracy and diversity with compare to full ensembles. The paper strongly favors the idea that such methodology is useful for effective learning for classification of unlabeled data.

In 2004, Rudin et al. [110] studied the convergence properties of the Adaboost algorithm by reducing the algorithm to a non-linear iterated map and considering the evolution of its weight vectors. They find the number of cycles due to the dynamical system nature of Adaboost that explicitly solves Adaboost algorithm output. Their

work is slightly related to our proposed idea that we also evaluated the possibility of convergence of optimization problems in consideration that in turn helps in achieving the optimal weights.

Polikar [106] relates decision making based on ensemble learning or averaging with many real-world examples where financial, social, medical, and other implications that really matter in order to make a decision. The primary methodology of the paper is to explain the committee of expert models with the help of many visual-art techniques that are easy to comprehend and intuitively grasp the underlying meaning of a few well-known ensemble learning algorithms. With comparison to this, Our proposed idea of the ensemble model is represented through the visual technique of the ROC curve. All the considered classifier and their performance are visualized through ROC curve that serves as one of the best technique for assessing the class label of many classification problems.

The study of Banfield et al. [6] in 2007 examines bagging and many other randomized classifiers for creating an ensemble of the classifier which were evaluated for significance test using statistical techniques. In their approach, bagging is more accurate statistically with respect to other methods but when comparing the average rank of the algorithm, it is found that boosting, random forest and randomized trees are statistically better than bagging.

Cohen et al. [33] in 2007 investigated the idea of instance space decomposition as a decision tree framework and using this framework the instance space is divided into multiple spaces and each distinct classifier is applied on that subspace. The paper actually presents a novel idea of splitting a rule where it is possible to improve the implementation of framework both in terms of accuracy and computational cost time.

In 2007, Garcia-Pedrajas et al. [57] proposed a novel idea of enhancing accuracy and making a pool of diverse individual classifiers where they construct an ensemble using non-linear projections. The primary contribution of the paper is to use projections of those instances that have been difficult for any previous classifiers instead of using random subspace. By doing this, non-linear projections are created with the help of neural networks in a consecutive way for those instances that have been misclassified and is comparable to boosting and bagging for performance increase.

The idea of P-Adaboost was proposed by [97] in 2007 as a novel use by the parallelization of Adaboost that is basically built on the dynamics of Adaboost weights and in short are an approximation of standard Adaboost which can be efficiently distributed over computing nodes. Experimental results are conducted

both on synthetic and benchmark data that supports the properties of P-Adaboost as a stochastic minimizer of Adaboost cost functional.

Tsoumakas et al. [124] is a review paper about ensemble selection that discusses the reduction of ensembles of predictive models to improve efficiency and predictive performance. The focus of the paper is to categorize those methods that are based on a greedy search of the space of all ensemble subsets. The paper highlights different directions and measures that provide a general framework of the greedy selection ensemble algorithm.

Sagi et al. [111] is a recent review article (the year 2018) on ensemble learning that provides state-of-the-art for machine learning perspective and challenges aforementioned within this field. The unique perspective of this review study is to integrate the intense study of ensemble learning in the context of deep neural networks, distributed algorithms for training ensemble models, and converting ensemble models into their simpler models.

Breiman,L. (2001) [19] is a interesting studies that discuss different cultures of statistical modeling for reaching conclusion from data. One culture advocates the underlying assumptions that the data are generated by random kinds of stochastic phenomena and other kinds of culture is the algorithmic model that treats data strategically as unknown phenomena. The paper explains the pros and cons of both cultures highlighting the need for the adoption of a diverse set of tools for solving problems of different fields using data.

2.1.2 Credit risk: literature review

Of recent, the use of machine learning models is having increased adoption in many different tasks of credit risk modelings such as identification of early warning system for predicting financial distress, corporate default, forecasting mortgage, or any other credit portfolio default risk.

Chakraborty et al.(2017) [83] advocates the use of the machine learning model to detect financial distress using balance sheet information and their studies conclude performance increase of 10 percentage points compared to logistic regression model as a preferred classical approach of financial institutions.

Khandani et al.(2010) [87] applies state-of-the-art of non-parametric machine learning models to predict the default of consumer credit risk by merging transactions and credit bureau data. Their work demonstrates that prediction of risk can be better improved using machine learning techniques in comparison to classical statistical approaches and any subsequent loss of lenders therefore can significantly be improved.

Albanesi et al.(2019) [2] applies deep learning approach as a combination of neural network and gradient boosting for high dimensional data to predict default of consumer risk. Their work shows superior performance compare to logistic regression models and is also able to adapt to the aggregate behavior of default risk easily.

The studies of Bachman et al.(2017) [43] compares the performance of machine learning models with industry-developed algorithms such as Moody's proprietary algorithm and suggests improvement of 2-3 percentage points in performance of the machine learning model. Their approach is a bit difficult to relate with the underlying firm characteristics in predicting default of credit risk although credit behavior-related variables increase the discriminatory power of the considered models.

Fantazzini and Figini (2009) [50] proposes a non-parametric approach based on random survival forests in predicting credit risk default of small-medium enterprises. The performance comparison of the proposed model with the traditional logistic regression model reveals a weak relationship of the performance between training and testing sample thereby suggesting an over-fitting problem which is mainly due to contrasting testing sample performance of logistic regression better than the proposed random survival models.

Several other studies like Kruppa et al.(2013) [90], Yuan(2015) [129], Barboza et.al (2017) [8] confirms superior performance for prediction of credit risk using machine learning compared to any other statistical approach.

2.2 Pareto-based multi-objective: literature review

The historical origin of multi-objective optimization is not known exactly. However, it is much acknowledged that the concept was borrowed from the field of Economics and is credited to Francis Y. Edgeworth (1845-1926) and Vilfredo Pareto (1848-1923) for introducing the concept of non-inferiority in the context of economics. Since then, the field of multi-objective optimization has been evolved in many diverse fields at an increasing pace. For any further detail on the historical perspective of multi-objective optimization, one can refer to [38] .

Belton et al. [125] provided in 2002 an integrated approach of multi-objective optimization, known as Multicriteria Decision Analysis (MCDA), which focuses on the development of the field in the last quarter-century from different sources. The book is concise and comprehensive to understand the underlying theories and philosophies of MCDA . The insights are drawn in the book help readers to implement any approaches in an informed manner and provides a holistic view of different

theories from broader management theory, science, and practice.

Dellnitz et al. [40] suggested in 2005 the idea of solving multi-objective optimization problems numerically that is global in nature and allows the approximation of the entire set of Pareto fronts or specifically global Pareto points. The paper describes few procedures for convergence of solutions and the results achieved are used to develop different algorithms. These algorithms are combined together for better understanding, citing real-world examples to improve the overall performance of the achieved Pareto solution. In contrast to this, our approach for approximating the Pareto front solution depends on the unordered approach and does not follow necessarily any ordered approach like a priori, posteriori, and interactive methods. The Pareto solution obtained through our proposed multi-objective problem is a set of local and global optimal points.

The study of [47] in 2005 considers many engineering and social science problems where various conflicting objectives are solved through multi-objective techniques. They provide in their explanation the mathematical reasoning to solve linear, non-linear, and combinatorial problems with multiple criteria. Their methodology provides Pareto-optimal solutions that are not within the capacity of traditional mathematical models.

Konak et al. (2006) [89] bring a tutorial perspective of multi-objective optimization using a genetic algorithm and are well captured in their studies. The use of Genetic Algorithm (GA) are a meta-heuristic technique that is particularly well suited for problems of multi-objective optimization and generally, there are a lot of approaches like utility theory, a weighted sum of methods which weigh selection problem to characterize decision-maker choices. The paper brings comprehensive understanding of many algorithm like Vector Evaluated Genetic Algorithm (VEGA), Multiobjective Genetic Algorithm (MOGA), Weight Based Genetic Algorithm (WBGA), Niched Pareto Genetic Algorithm (NPGA), Random Weighted Genetic Algorithm (RWGA), Pareto Envelope-based Selection Algorithm (PESA), Pareto-Archived Evolution Strategy (PAES), Nondominated Sorting Genetic Algorithm (NSGA), Strength Pareto Evolutionary Algorithm (SPEA), Rank-Density Based Genetic Algorithm (RDGA) and Dynamic Multi-objective Evolutionary Algorithm (DMOEA) in unique and simple style. Following their work, our proposed approach is a meta-heuristic in the sense that they are weighted and computationally evolutionary.

The table 2.3 list down the few popular multi-objective algorithm and the fitness function that is embedded in the algorithm for evolutionary computation.

Table 2.3: Popular multi-objective algorithm.

Algorithm	Fitness function
VEGA	different objective is being assessed based on subpopulation.
MOGA	ranking based on pareto approach.
WBGA	normalized objectives as a weighted average method.
NPGA	fitness function is missing and only tournament selection.
RWGA	normalized objectives as a weighted average method.
PESA	fitness function is not assigned.
PAES	In case offspring dominates, parent is replaced by pareto dominance.
NSGA	non-domination sorting way of ranking.
NSGA-II	non-domination sorting way of ranking.
SPEA	non-dominated solutions as a ranking based external archive.
SPEA-II	dominators strength are assessed.
RDGA	using solutions rank and density as objectives, problems are optimized as bi-objective problem
DMOEA	ranking based on cell method.

A review study in the year 2006 by [80] presents the strength and weakness of many multi-objective algorithms in favor of which few test problem criteria are introduced which in turn is supported by a set of definitions. Based on the motivation that many test problems are not correctly constructed, so is the poor representation of non-separable of multi-modal problems, and to meet this gap, the paper addresses a flexible toolkit for constructing the problem accompanied with empirical results that show how the proposed toolkit can be used as an optimizer in ways that traditional toolkit does not address.

In support of this study, finding an optimal solutions of multi-objective problems is difficult and this poses certain challenges in ordering the solution. The challenge in solving the multi-objective problems is not very informative and lacks mature study in many different contexts. Due to this, it can be said that such challenge poses to be one of the main weakness of the multi-objective optimization problem at least in our opinion.

Carlos et al. [32] describe evolutionary algorithms for solving multi-objective problems integrating contemporary knowledge. The paper provides an overview of many algorithms that are state-of-the-art research results available in the year 2007. They have explained many Multiobjective Evolutionary Algorithm (MOEA)

techniques using practical examples apart from explaining MOEA test functions and performance measures.

Tapia et al. [23] present a review study of multi-objective evolutionary algorithms in the field of economics and finance explaining the uniqueness, strength, and weakness of literature for five groups of applications like investment portfolio optimization, financial time series, stock ranking, risk-return analysis, and economic modeling. In contrast to this, our approach was primarily focused on using a multi-objective optimization algorithm for effective risk and portfolio management.

Juergen et al. [14] combine knowledge from the seminar series to give different perspectives on interactive and evolutionary approaches of multi-objective optimization. Their methodology focuses on continuous problems and not on discrete problems and introduces many basics of multi-objective optimization that can be extended to non-linear multi-objective optimization in the context of Evolutionary Multiobjective Optimization (EMO) and Multicriteria Decision Making (MCDM), including both interactive and non-interactive approaches.

Deb [39] provides introductory and simple explanations of multi-objective optimization that involve optimizing many objective functions simultaneously. The explanation of the approach is using an evolutionary algorithm to solve objectives that are conflicting and differ from each other and finding a set of optimal solutions for objective functions is often challenging. While solving such problem with or without constraints, a set of solutions is generated that are Pareto-optimal solutions. In this context, the basic approach is the use of population search methods known as evolutionary multi-objective optimization. The paper also highlights the journey of evolutionary multi-objective optimization through various actors as being an established field in recent times.

Ehrgott [47] published and integrated many stories of multi-objective optimization focusing on the minimization with respect to total and partial orders.

Chapter 3

Methodology

Big data is at the foundation of all of the mega-trends that are happening today, from social to mobile to the cloud to gaming.

Chris Lynch

This chapter brings our methodological contribution to the proposed idea of model averaging and multi-objective strategies. The proposed idea has the capacity to enhance the performance of the learning algorithms ,and should be viewed as one of the novel methods to construct an ensemble model to achieve superior performances. We start with a section that provides some background information of the proposed methodology and then subsequently moves to the core idea in their corresponding section.

Section 3.1 explores background information on model averaging from Bayesian, frequentist, and other information criterion. Section 3.3 discusses in detail the factors that influence any estimation of weights and proposes a different strategy for estimating weights. Section 3.4 presents an additional discussion in support of the model averaging approach. Section 3.5 describes possible properties in the context of the ensemble model.

Section 3.2 contributes the knowledge behind the model averaging that can be possibly applied to many different fields especially in the field of machine learning as a beneficial approach to the different predictive tasks. For instance, we applied the proposed methodology on a credit default prediction analysis in the scope of this thesis.

Section 3.6 describes our methodological knowledge behind multi-objective optimization from the perspective of the Pareto-based approach, information science ,and

machine learning as an integrated approach to enhance performance and accuracy of the model. At the best of our knowledge, the idea proposed in this context is novel and unique that perhaps is not been addressed so far in the literature in the way our thought process is poised in achieving greater efficiency and accuracy of the learning model.

3.1 Background Information

Majority of the study on model averaging centers around BMA that comes from the Bayesian domain where one needs to set a prior since the model parameter is unknown. The other popular approach is FMA where a prior selection of weights is important to understand the nature of estimators under repeated sampling and asymptotic optimality. There are handful of different information criterion beyond this which takes their own space and carries extended discussion in the context of model averaging.

Taking inspiration from the model averaging approach, the field of machine learning has diversified especially the importance of ensemble model where the basic goal is to combine a set of expert models with low bias and high variance to average them to achieve superior accuracy. The technique of ensemble learning is used as a cutting-edge tool in corporations and among practitioners to retrieve an efficient set of patterns from the data. Our effort in the proposed approach is an attempt to construct an ensemble model that might find its relevance and importance for many different problems that depend on data for decision making.

Beyond Bayesian, frequentist ,and information-theoretic approach, there are other approaches of model averaging whose goal is to choose weights for the model in such a way to optimize the prediction error. Any simpler way to take average weight depends on the estimate of the prediction error of every single model that is derived using CV . The CV method irrespective of how many folds are used approximates a model performance on hold out data that can be quantified in many ways with specific distribution using likelihood function as in the following Equation 3.1:

$$\ell^k{}_{cv} = \sum_{i=1}^{k} \ell(f_{[i]}|\widehat{\theta}^k_{f_{[-i]}}) \tag{3.1}$$

where l is a likelihood function, f is a density function and θ is a parameter for a given probability distribution function.

However, the approach just mentioned is prone to overfitting and to penalize any overfitting from such computation, it is important to understand how to compute

model weights and in many cases is equivalent to AIC and Kullback Leibler (KL) distance as follows in the Equation 3.2:

$$w^k{}_{cv} = \frac{e^{\ell^k{}_{cv}}}{\sum_{i \in k} e^{\ell^i{}_{cv}}} \tag{3.2}$$

We are not going into the details of all the possible CV methods with likelihood functions that are used for the model fitting task, as dedicated studies on this topic are available and one can refer to (Stone [119], Hauenstein et al. [74]) for any further details. Few of the important CV methods in the context of machine learning are jackknife model averaging and stacking to compute model weights. They are one of the finest ways to optimize model weights to reduce the error of predictive models on hold out data.Several authors like Hastie et al. [73], and Wolport [127, 128] finds that such methodology to optimize model weights is equivalent to Root Mean Squared Error (RMSE) with likelihood as evident in Equation 3.3:

$$\arg\min_{w_k} \left\{ \ell(f_{[i]} | \sum w_k \widehat{f}(X_i | \widehat{\theta}^k_{[-i]}) \right\} \tag{3.3}$$

It works with a procedure that is repeated many times to derive a vector of optimized weights that are averaged across repetitions whose sum is equal to 1 after re-scaling. Refer to [128] and [30] for many interesting case studies that also explore such topics in the Bayesian context. The stacking and jackknife model averaging conceptually is the same as they use a similar kinds of optimization function. In stacking, for each single run, the weights are optimized, while for jackknife only one optimization function is used for all leave one out cross-validation.

Very similar to our methodological approach, many studies on model averaging refer to two approaches. One is the use of a minimum variance approach (Bates and Granger [9]) that puts more weight to models with low-variance predictions. The other approach is to compute weights using the variance-covariance matrix of the model predictions in the context of multi-model generalization (Newbold and Granger [103]) as evident in Equation 3.4:

$$w_{mv} = (\mathbf{1}\sum\nolimits^{-1}\mathbf{1})^{-1}\mathbf{1}\sum^{-1} \tag{3.4}$$

Certainly, this can be said as one of the analytical solutions that assume no bias, and weights are considered as random weights since they sum up to 1. Sometimes it happens that the models among themselves are correlated, which needs adjustment. One such adjustment method is cos-squared weighting that assigns lower weight to highly correlated models and if the distribution of weights is identical then it is added

to the set to optimize weights for reducing prediction error. For any further details on such a weighting method, refer to the study by Garthwaite,P. et.al (2010) [58] .

The other adjustment method that could be relevant in this context is the idea of a super-ensemble model where different models are combined in a regression framework and can take any form like a linear model or a neural network. However, there are some drawbacks to such approaches, like the high probability of overfitting due to fitting the same data twice or multiple times. This is a bit unexplored topic both in theory and practice whose variants strongly depend on how cross-validation is done for fitting the considered models in the super-ensemble approach. Refer to Granger and Ramanathan [63] for more details.

Similarly, the other trivial weighting scheme is the use of equal weights that serve as a reference methodology to know if reducing weights actually reduce prediction error or not for a set of contributing models. It proves out to be a useful technique on many occasion with compare to other weighting methods.

A well-known information criterion that is not widely used but is worthy of attention is Hannan and Quinn criterion [67] and works on the principle of autoregressive order selection. This is to say if the data is generated by an autoregressive model of order n_0 and any selected model of order n_0 converges almost surely when the sample size goes to infinity.

Keeping it short, the method selects a model by minimizing $hq_t = -2\widehat{\ell}_{n,t} + 2c\ d_t \log\ \log\ n$ where hq refers to a small possible penalty to ensure strong consistency. Moreover, there exists a bridge criterion to provide additional support that selects the model f_t and minimizes the following Equation 3.5:

$$bc_t = -2\widehat{\ell}_{n,t} + c_n(1 + 2^{-1} + \ldots + d_t^{-1}) \tag{3.5}$$

Many other criteria like Minimum Message Length (MML), Minimum Description Length (MDL), DIC ,and Generalized Information Criterion (GIC) are motivated from other perspectives to bring consistency of model selection or model averaging approach. We explain in brief all these information criteria as follows:

- The MML is a criterion to select model weights that minimizes the following Equation 3.6:

$$-\log p(\theta) - \log p(x|\theta) + \frac{1}{2}\log|I(\theta)| + \frac{t}{2}(1 + \log t) \tag{3.6}$$

- The MML criterion also represents one of the ways to describe the best model, which leads to the best compression of given data ,and the model is selected

in this framework by minimizing the stochastic complexity as in the following Equation 3.7:

$$-\log\ p_{\theta_1}(f_1) - \sum_{t=2}^{n} \log\ p_{\theta_t}(f_t|f_1,\ldots,f_{t-1}) \tag{3.7}$$

- The DIC was proposed basically as a measure of Bayesian model complexity and this criterion can be viewed as a Bayesian counterpart of AIC . The deviance under model can be seen as $D_t(\theta) = -2\log p_t(f|\theta) + C$.
- The Maximum Likelihood Estimation (MLE) and model dimension in AIC are substituted with posterior mean while DIC substitutes effective number of parameters.
- DIC enjoys computational advantage concerning AIC over a set of complex models where the use of likelihood function could not be the right choice.
- The GIC embeds a wide class of criteria whose penalties are linear and minimizes the following Equation 3.8:

$$GIC_{\lambda_{n.t}} = \widehat{e_t} + \frac{\lambda_n \widehat{\delta_n}^2 d_t}{n} \tag{3.8}$$

we can say Mallow's criterion is a special case of GIC .

3.2 Model averaging approach

In recent years, several multi-model methods have been proposed to account for uncertainties arising from input parameters and the definition of the model structure.

In this thesis, we propose a novel methodology for model average that comes as a solution of a quadratic programming problem. The solution obtained is used as a weight to achieve a weighted model and these weighted model by construction and presumed model combination method is considered as an ensemble model to achieve higher accuracy.

Let us denote $f_1,\ldots,f_t$ as t different models. For each model under consideration, we evaluate the estimation error as $\epsilon_t = f - \widehat{f_t}$ where $\widehat{f_t}$ is the estimated value of f for a model t. Based on ϵ_t, we estimate the covariance or correlation matrix. For this purpose, the optimization problem can be solved both for co-variance and correlation matrix.

We need to understand if one of the two provides better results, keeping into account that preferring the co-variance matrix helps in enhancing the performance

of the proposed weighted model slightly better concerning to correlation matrix as evident in their empirical results in chapter 6 . In general, the optimization problem is indifferent if defined for a non-singular positive definite square matrix of the models.

We solve the optimization problem finding the vector of weights that minimizes the co-variance among models.

An average of models can improve the performance of single models when the errors of the single models are negatively correlated. Roughly speaking, an average model improves the performance concerning single models when an error of the single model is counterbalanced by a good prediction of some other model. Following this idea, the best average model is the one that minimizes the error, given the co-variance between the errors of the single models. Using the idea of our model averaging technique, we compare the constructed model with other popular models (parametric, non-parametric, and ensemble models) and the results at hand in chapter 6 suggests that our proposed model can enhance model performance.

Let us denote with $\Sigma \in \mathbb{R}^{k\times k}$ the co-variance matrix of the errors ϵ_i with $i = 1, \ldots, k$. Σ is a positive definite, symmetric, and thus a non-singular matrix. w is the $k \times 1$ vector of the weights. The average model is defined as

$$\widehat{f_w} = w_1 \times \widehat{f_1} + \ldots + w_t \times \widehat{f_k}$$

where w_t is the t-th entry of vector w. $\mathbf{1}$ is the $t \times 1$ vector of ones.

Following our considered model average technique, we formulate the optimization problem as in Equation 3.9:

$$\min_{w} \left(w^T \sum w \right) \text{ such that } w^T \mathbf{1} = 1, w^T \overline{\widehat{f}} = \overline{f} \tag{3.9}$$

where $\mathbf{1}$ is the $k \times 1$ vector of ones, and T denotes the transpose of a vector or matrix wherever applicable, $\overline{\widehat{f}}$ denotes the average value of the predicted model and $\overline{f}$ denotes the average value of the observed model.

The analytical solutions of the optimization problem in equation 3.9 produces an optimal vector of weights

$$w^* = \Sigma^{-1} \times (\mathbf{1}\overline{\widehat{f}}) \times A^{-1} \times \begin{bmatrix} 1 \\ \overline{f} \end{bmatrix}$$

where

$$A = (\mathbf{1}\overline{\widehat{f}})^T \times \Sigma^{-1} \times (\mathbf{1}\overline{\widehat{f}})$$

.

For the above-stated optimization problem, the first-order conditions are necessary and sufficient for the optimality of w^*, which is straightforward due to the assumptions made on the co-variance matrix Σ . The analytical solution assumes no bias and therefore ignores the problem that weights are random variates since weights are constrained to sum to one. Doing this, it does not necessarily ensure weights to be positive. Moreover, we do not want to use some rarely used method that adjusts for correlation or co-variance in predictions, such as assigning lower weights to highly correlated models, dividing weights if any identical model prediction is added to the set, henceforth reducing weights due to additional inclusion of the model.

The proposed approach has its advantage and disadvantage, which are listed below.
Pros of the approach:

1. Improvement of the performance compared to single original models
2. Closed-form for the solution of the optimization problem.
3. A simple interpretation of the whole theoretical structure.

Cons of the approach:

1. Interpretation of the negative weights. When the weight associated with a model is negative, intuitively we are doing the opposite compared to what the model suggests to do. It is clear that negative weights are useful in order to artificially create negative co-variances between models providing the possibility to achieve lower values of co-variance.
2. If $\widehat{f}_i$ for $i = 1, \ldots, k$ are bounded (for example in case of probability of default when modeling credit risk), the proposed approach does not guarantee that $\widehat{f_w}$ respect the bounds.

In order to overcome the potential shortcomings described above, we can think to rewrite the optimization problem as follows in Equation 3.10,

$$\begin{gathered} \min_w w^T \Sigma w \\ \text{such that} \\ w^T \mathbf{1} = 1 \\ w_i \geq 0 \text{ for } i = 1, \ldots, k \end{gathered} \tag{3.10}$$

This helps to overcome both the shortcomings of the interpretation of the negative weights and the bounded value for $\widehat{f_w}$. In this case, it is trivial to prove that $\widehat{f_w}$ is bounded between the minimum and the maximum values of the single models since the average model is a convex linear combination of the original models.

Two of the possible limitations of the new optimization problem could be the availability of no closed-form solution for the problem and that a growing number of restrictions penalizes the performance of the average model.

Moreover, we can make effort to reformulate the optimization problem 3.10 to solve them with three constraints as in the following Equation 3.11:

$$\begin{gathered} \min_{w} w^T \Sigma w \\ \text{such that} \\ w^T \mathbf{1} = 1 \\ w^T \overline{\widehat{f}} = \overline{f} \\ w_i \geq 0 \text{ for } i = 1, \ldots, k \end{gathered} \tag{3.11}$$

The analytic solution of this problem does not guarantee the estimation of positive weights nor achieving the minimum prediction error but is useful in many ways to obtain a set of weights that is capable of enhancing the performance of the model.

Many extensions are possible. One standard extension could be to allow the weights to be negative that in turn reduces the values of the co-variance matrix. The lower value assigns a lower weight to the model and compensates for the higher positive weight of the model.

The other extension could be to use a super-ensemble model also called a supra-model where each model under consideration is a covariate in a regression framework to obtain a super-ensemble model. For a deeper discussion of this approach, the reader can refer to [63] .

In our equal weighting approach, we have adopted a slightly different approach than the ones used in [82, 88, 59, 109] to put forth a new extension that could potentially serve as a good reference approach to understand if estimating weights reduces the prediction error for the given set of models. The good thing about this approach is that they do not have to depend on data provided no bias method is implied on the data.

The Majority of the model averaging discussion focuses on one simple thing, which is how to reduce the prediction error in a sense that for any given estimator $\widehat{f_k}$,

we can at least consider the Mean Squared Error (MSE) for instance among other sets of error metrics, and decompose them into components of bias and variance as presented in Equation 3.12:

$$MSE(\widehat{f_k}) = \left(bias(\widehat{f_k})\right)^2 + var(\widehat{f_k}) \tag{3.12}$$

The bias in this case refers to error components that arise from the model and more precisely can be termed as systematic model errors. The bias remains unaffected for any new addition of data points to the model. The components variance refers to the possible spread of model predictions and does fit well hypothetically for any addition of a new dataset in the model.

Using the representation of the Equation 3.12, we can investigate the error of a weighted average $\overline{f}$ of k plausible models, $\widehat{f_1}, \widehat{f_2}, \ldots, \widehat{f_k}$ as

$$\overline{f} = \sum_{i=1}^{k} w_i \widehat{f_i}$$

with $\sum_{i=1}^{k} w_i = 1$.

From Equation 3.12 we can further infer that the purpose of w_i ,in general, is to improve the prediction and reduce error with respect to the simple average having equal-weighted scheme or only considering one single model having embedded all the weight onto one model. In any case, the bias $\overline{f} - f^*$ in the model combination process do matter in reducing the variance among models and to a larger extent depends on the bias of a single model and their weights.

This contradicts those beliefs that advocates referring to an individual model is free from any bias and is never considered as a contributing model. This is precisely different for model averaging approach since reducing bias among them is primarily a major concern for many of the predictive task.

Any variance arising from k hypothetical repeated samplings in the predictive task is often composed of two terms that are a variance of each model in consideration, as modeled by Equation 3.13:

$$var\left(\widehat{f_k}\right) = \frac{1}{k-1} \sum_{i=1}^{k} \left(\overline{\widehat{f_1}} - \widehat{f_2}\right)^2 \tag{3.13}$$

To measure any co-variance between two model $f_1,\ f_2$, we can apply the following Equation 3.14:

$$cov\left(\widehat{f_1}, \widehat{f_2}\right) = \frac{1}{k-1} \sum_{i=1}^{k} \left(\overline{\widehat{f_1}} - \widehat{f_1}^{\,i}\right) \cdot \left(\overline{\widehat{f_2}} - \widehat{f_2}^{\,i}\right) \tag{3.14}$$

While taking the average of predictions $\widehat{f_1}$ and $\widehat{f_2}$, it infers that

$$var(\overline{f}) = w_1^2 var(\widehat{f_1}) + w_2^2 var(\widehat{f_2}) + 2w_1 w_2 cov(\widehat{f_1}, \widehat{f_2})$$

And using a similar method like the one above, we can average several models as in Equation 3.15:

$$var(\overline{f}) = var\left(\sum_{i=1}^{k} w_i \widehat{f_i}\right) = \sum_{i=1}^{k} f_i^2 var(\widehat{f_m}) + \sum_{i=1}^{k} \sum_{i'} w_i w_{i'} cov(\widehat{f_i}, \widehat{f_{i'}}) \tag{3.15}$$

The task to reduce the error from any of the model averaging technique with the use of several selected models is interconnected with the relationship of bias and variance. The error can increase or decrease in the average model with respect to the best model if there is a larger bias than the variance in selected models.

It is being believed with confidence that the averaging technique can reduce the error significantly with an increasing number of models if all these selected models have similar bias and variance. The averaging technique of selected models can actually make error arbitrarily small if we have a pool of unbiased models with larger variance.

The usefulness of any model averaging technique to a much larger extent depends upon the biases of any individual models and therefore prediction error is proportional to variance (increase or decrease) keeping the assumptions that co-variances are low, which in turn helps in achieving smaller mean error with respect to the variance of a single model. In any case, the bias becomes greater in comparison to the variance for any predictive model. Due to this, model averaging techniques may not be necessarily reducing variance in every situation.

The same setting can be explained for any predictive model using the correlation in place of co-variance as if the correlation increases or decreases, the co-variance plays a decisive role in reducing the overall prediction error.

When we add several models in a model averaging approach, the variance generally seems to be low as weights w become smaller, which can be approximated as proportional to $1/k$ in Equation 3.16:

$$var(\overline{f}) = \sum_{i=1}^{k} \frac{1}{k^2} var(\widehat{f_i}) + \frac{1}{k^2} \sum_{i=1}^{k} \sum_{i \neq n} cov(\widehat{f_i}, \widehat{f_n}) \approx k \frac{1}{k^2} \tag{3.16}$$

3.3 Estimation of Weights

On a broader sense, the error from any predictive model as an averaged out model predictions largely depends on the bias of model average, variance, and covariance of a model for any given set of weights. The weights in many cases are often considered as fixed and uncertain. Any estimation of weights from data brings a certain amount of uncertainty, which benefits the model averaging technique to derive optimal or sub-optimal weights.

The main challenge behind the estimation of weights is how to obtain optimal weights using the help of a good estimator and there is no closed-form solution available even for linear models. Moreover, in a purely general sense, the weights obtained through any Bayesian or information-theoretic criteria is never optimal, although they help to reduce the prediction error. However, to find more satisfying weights to enhance the predictive performance is done adopting a different techniques that is prudent for model averaging. In this way, we can say that bias-variance trade-off is also applied for the model averaging case since the estimation of model weight adds additional parameters and higher model complexity in the analysis.

Moreover, the uncertainty around obtaining optimal weights does not necessarily infer that other weighting approach is superior but offers few of the best solutions to address challenges in the estimation of weights. For instance, one possible way to enhance model accuracy is simply averaging variance and bias given that error between models is small.

It is always complicated to estimate optimally the uncertainty around the model averaging or estimation of weights for any considered approach like Bayesian, frequentist, and others. One possible way to quantify such uncertainty especially in the Bayesian method is to use posterior distribution for estimating weights. The validation of the estimated uncertainty can be summarized by credible intervals purely in Bayesian sense and 95% certainty is considered to be close to the true value in the interval. The same is true for the frequentist approach, where 95% confidence interval refers to 95% true values of the cases under repeated sampling and identical conditions.

To work with a more Bayesian and frequentist approach in model averaging techniques, we have to calculate different options for knowing predictive uncertainties of which one option would be to assume that model average predictions are unbiased and for computing variance, any bootstrapping technique is used to compute covariances of each predictive model. The other option that potentially produces better predictive performance is to assume that bias and correlation are more conservative

in model average prediction. If averaging models are independent of each other then the covariance-based on bootstrap techniques can actually compute lower variance to boost performance.

In order to estimate weights properly, we have to often fix poorly fitting models and improve good performing models to enhance the overall performance of any average or ensemble models. Although many perspectives exist on estimating weights, our approach for model averaging is more "probabilistic", in the sense that model weights are probabilities considering that f_i is any true model.

In the context of machine learning, a prudent approach could be to choose weights somehow chosen that makes the model work to enhance higher performance. The chosen weights are not considered model parameters and as such, there is no specific interpretation of the model with respect to model weights. The other benefiting approach to improve predictions is to consider a portfolio of weights that may be random, equal, negative, squared, and optimal weight without further adjusting any difference in the predictive capacity of the model.

3.4 Additional discussion

Various theory suggests that Bayes formula plays a decisive role in choosing the model among models. The theory works similar to choosing parameter values in the Bayesian framework where the posterior probability $P(f_i, \theta_i|D)$ of any model f_i with parameter vector θ_i can be stated for any given dataset observations D as

$$P(f_i, \theta_i|D) \propto L(D|f_i, \theta_i) \times p(\theta_i) \times p(f_i)$$

where $L(D|f_i, \theta_i)$ is the likelihood function of any model f_i, $p(\theta_i)$ is prior distribution of parameters with respect to any model f_i and $p(f_i)$ is any prior weight attached to the model f_i.

In many real-world examples, we are often interested in knowing simple statistics about the model such as posterior model probability. This could be high or low as a result of model prediction and model selection uncertainty. In few cases, the estimated weights are considered as the relative probability of each model in the context of marginal likelihood and are equivalent to the average of all the parameters used in the model as evident in Equation 3.17

$$P(D|f_i) \propto \int_{\theta_i} \dots \int_{\theta_k} L(D|f_i, \theta_i) p(\theta_i) d\theta_1, \dots, d\theta_k \tag{3.17}$$

Kass and Raftery [85] introduced a concept of comparing models through Bayes factor using marginal likelihoods as evident in Equation 3.18

$$\frac{P(D|f_i)}{P(D|f_j)} = \frac{\int L(D|f_i,\theta_i)p(\theta_i)d\theta_i}{\int L(D|f_j,\theta_j)p(\theta_j)d\theta_j} \tag{3.18}$$

The estimation of these ratios in practice can be challenging. To address the challenge, we rely on two numerical estimations. One option is to use direct samples from the joint posterior distribution of models and their parameters. There are many basic and advanced algorithms that exist in support of such computation but in general, they are not easy to program and therefore could be a topic of research for the future. Few of such algorithms that already exists are RJMCMC, MCMC, Sequential Monte Carlo (SMC) . one can refer to Toni et al. [123], Hartig et al. [71] and Green [64] for any additional details.

3.5 Properties: ensemble model

So far, Our discussion was focused on co-variance and other methods that plays a crucial role in the estimation of weights and construction of ensemble models. In this section, the focus of our discussion is how ensemble models behave or vary when there is correlation or no correlation between models.

Let us assume first the case of uncorrelated models in the ensemble averaging system where we refer to the properties of variance and assume models are independent. One of the possible ways to obtain optimal weight in the ensemble model system if the models are not correlated is to construct ensemble as a linear combination of the individual model.

Therefore, we consider ensemble as a linear combination of its members as $\overline{f} = \sum_j \alpha_j f_j$ and α_j are normalized to 1 to achieve the following Equation 3.19:

$$\sum_j \alpha_j^2 v(f_j) + v(f_k) + b^2 = \sum \alpha_j^2 \alpha_j^2 + (\sum_j \alpha_j b_j)^2 \tag{3.19}$$

Using this Equation 3.19, we can find the optimal coefficient as weights to the model by minimizing error and can be converted into an optimization problem as evident in the following Equation 3.20

$$\min_{w_1,\ldots,w_k} \sum_j w_j^2 \sigma_j^2 + (\sum_j w_j b_j)^2 \sum_j w_j = 1 \tag{3.20}$$

This could lead to a lesser extent the underestimation of the statistical properties of the ensemble model if the optimal weight from the coefficient is not considered.

So, we can ask ourselves if there is any way that ensures the variance of the ensemble model is lower than an individual model variance.

To understand more of this, take a case of two models having variance such that their combined variance is less than a single model variance. The combination process best works if the variance between models is not too large otherwise it is not possible to achieve ensemble model variance lower than individual models.

In this respect, We propose the following theorem and make attempt to prove this with a simplified approach. In short, we can say that

Theorem 3.5.1. *If the combined variance of two models is less than single model variance then the combination process best works only when the variance between models is minimized else it is not possible to achieve ensemble model variance lower than individual models. This is true if the models among themselves are not correlated since it makes possible to obtain the variance of ensemble model lower than single models as evident in the following inequality where v() simply denotes the variance of the model.*

$$v(\overline{f}_t) \leq v(\overline{f}_1) \leq v(\overline{f}_2) \ldots \leq v(\overline{f}_r)$$

.

Proof. Let us say that $\frac{(t-1)v(f_m^2)}{t^2} \leq \frac{(t^2-1)v(f_1^2)}{t^2}$. Therefore, we can further say that

$$(1 - \frac{1}{t^2})v(f_1^2) \geq \frac{t-1}{t^2}v(f_t^2) \geq \frac{1}{t^2}(v(f_2^2) + \ldots + v(f_t^2) \tag{3.21}$$

which in turn proves that $v(f_1^2) \geq \frac{v(f_1^2)+\ldots+v(f_t^2)}{t^2} = v(\overline{f}_t)$ □

The idea mentioned in the above theorem can be generalized to models that are correlated among each other and is possible to obtain general bounds for optimal variance using the following inequality,

$\frac{f_1}{k} \leq v(\overline{f}) \leq \frac{f_k}{k}$.

The proof sketched in the theorem 3.5.1 for uncorrelated models can equally be explored for correlated models by showing equivalent estimation for optimal variance as like in the Equation 3.22:

$$\sum_j \frac{1}{cov_j}(\sum_i u_{ij})^2 \geq \frac{1}{cov_k}\sum_j(\sum_i u_{ij})^2 = \frac{k}{cov_k} \tag{3.22}$$

Following the work of Jagannathan and Ma [81], one can solve the optimization problem as a global minimum variance problem using additional constraints on weights and is equivalent to the shrinkage estimate for co-variance matrix. This

further helps in diversifying the formulation of optimization problem for model averaging technique.

For instance, if we consider a model space of t models and p is any vector of expected accuracy or performance by the model and $\sum$ any respective co-variance matrix, we can formulate the optimization problem as follows

$$\min \frac{1}{2} w^T \sum w \text{ subject to } \mathbf{1}^T w = 1;\ w \in \phi \cap s \tag{3.23}$$

where w is the vector of weights and ϕ is the search space. If this search space belongs to $\mathbb{R}^n$ and s is the set of weights constraints, then the established optimization problem becomes a global minimum variance optimization problem.

We can elaborate the search space as $\phi = \{w \in \mathbb{R}^n : p^T w \geq p^*\}$ that helps to receive the efficient performance considering the model in search space and p^* is any desired expected performance that we expect. Since the global minimum variance optimization problem much depends on the set of weights constraints s, we can consider two different definitions of s.

The first definition is where $s = \mathbb{R}^n$ and the solution obtained through this is an unconstrained solution as $w^*(p, \sum)$. The other definition where we impose bounds is $s = (w^-, w^+)$ as $w_i^- \leq w_i \leq w_i^+$ and $\overline{w}$ is the achievable solution of the considered optimization problem.

We can further deepen the analysis on the impact of weights that may arise in the structure of optimization problem due to these bounds and it is wise to say that $\widetilde{w} = w^*(\widetilde{p}, \widetilde{\sum})$ are the approximation of observed performance of the model and co-variance matrix, where $\widetilde{p}\ and\ \widetilde{\sum}$.

We can find the solution of the global minimum variance optimization problem by taking the Lagrange function and first order conditions as follows:

$$f(w; \lambda_0) = \frac{1}{2} w^T \sum w - \lambda_0 (\mathbf{1}^T w - 1) \tag{3.24}$$

and their first order conditions can be stated as follows:

$$\begin{cases} \sum w - \lambda_0 \mathbf{1} = 0 \\ \mathbf{1}^T w - 1 = 0 \end{cases} \tag{3.25}$$

Therefore, their optimal solution can be obtained as like $w^* = \frac{1}{\mathbf{1}^T \sum \mathbf{1}} \sum^{-1} \mathbf{1}$. The solution obtained very much depends on the covariance matrix Σ and can be written as $w^* = w^*(\sum)$. However, it is always a difficult task to obtain an analytical solution of the global optimization problem if there are constraints on the weights

imposed and can be solved numerically using any convex or quadratic programming algorithm.

3.6 Multi-objective optimization approach

In this thesis, we adopted another novel idea to enhance the performance of the machine learning model that is based on a few different strategies of using the vast knowledge of multi-objective optimization.

The solutions achieved from solving the defined multi-objective problem in this thesis is a set of unordered Pareto solutions. Our analysis is interdisciplinary in nature and the insights drawn from here have the capacity to be resourceful to solve other real-world problems.

Almost every machine learning algorithm is naturally a multi-objective task depending on which cost function is adopted. There has been increasing concern in the use of machine learning methods based on Pareto multi-objective optimization methodology and the success of such methods is mostly due to the success of evolutionary algorithms and other stochastic search methods. The advantage of Pareto based multi-objective learning is many folds. It is a powerful tool with scalar cost function in addressing different topics of machine learning such as clustering, feature selection, improvement of generalization ability, knowledge extraction, and ensemble generation.

Following the idea of multi-objective learning, We can categorize all learning problems as optimization problem and often it is a task of model selection where parameter estimation depends on different criteria.

For instance, in supervised learning, the common criteria is an error function that reflects the approximation quality, whereas in clustering we try to maximize inter-cluster similarity and minimize intra-cluster similarity. For problems of reinforcement learning, the criterion is a value function that helps in predicting the reward for an agent to perform a given action in a given state.

The learning algorithm in this context can be categorized as single objective learning, scalarized multi-objective learning, and Pareto-based multi-objective learning.

Single objective learning often minimizes MSE on the training data but other error measures can be used equally. Let $f = \frac{1}{N}\sum_{i=1}^{N}(y(i) - y^d(i))^2$ where $y(i)$ and $y^d(i)$ are observed and expected output respectively, N is the number of data pairs in the training data. For instance, in the context of the clustering algorithm, we

minimize the objective function as follows

$$f = \sum_{j=1}^{k} \sum_{x \in C_j} \|x - c_j\|^2 \tag{3.26}$$

where $\|.\|$ is a chosen distance of cluster C_j between a data point x and centre c_j, and k refers to number of clusters.

The interpretation and complexity of the model are strongly interrelated to each other and in general, the lower is the complexity of the model, the easier it is to understand the model.

In this case, we have to consider often a second objective reflecting the complexity of the model which can be aggregated as a scalar objective function keeping $f = E + \lambda\Omega$ where E is a common error function and Ω is a measure for model complexity that says a number of free parameters in the model, while $\lambda > 0$ is a positive hyper-parameter defined by the user. It is clear to see through this set up that we are able to optimize two objectives using a scalar function.

Such an approach is widely used in practice such as regularizing neural networks, creating interpretable fuzzy rules, and generating negatively correlated ensemble members. However, there are two main weaknesses in the use of the scalarized objective function for multi-objective function.

Firstly, it is difficult to make an appropriate choice of hyperparameter λ, and secondly, only a single solution can be gained from which it is difficult to visualize any further additional insights into the problem.

To help such a scenario, we have to take advantage of the Pareto-based learning that may help any learning algorithm to get out of local optima thereby improving the accuracy of the learning model. The primary advantage of Pareto-based approach is that the objective function is no longer a scalar but a vector due to which a number of Pareto optimal solutions can be achieved instead of one single solution.

Let us consider m-objective minimization problem as follows:

$$\min F(x), \text{ where } F = \{f_1(x), f_2(x), \ldots, f_m(x)\} \tag{3.27}$$

Any solution X is said to dominate solution Y and is Pareto optimal if it is not dominated by any other feasible solutions. The domination can be stated as $f_j(x) \leq f_j(y) \forall j = 1, 2, \ldots, m$ and there exists $k \in \{1, 2, \ldots, m\}$ such that $f_k(x) < f_k(y)$.

In this context, we lay out four important strategies that can serve as a model enhancing procedure in enhancing the objectives of the machine learning model. To

test each of these strategy, we consider a collection of parametric, non-parametric and ensemble learning models. Let $\{1, 2, \ldots, n\}$ represents the set of these models to which we assume the allocation vector of weights $w = \{w_1, w_2, \ldots, w_n\} \in \mathbb{R}^n$.

These weights are preferential choice where w_i is any specific weight attached with any model i for $i = 1, 2, \ldots, n$ constrained as $\sum_{i=1}^{n} w_i = 1$. Let $p = \{1, 2, \ldots, n\} \in \mathbb{R}^n$ be the performance associated with each of the model and we can represent them by following function

$$f_1(w) = wp^T = \sum_{i=1}^{n} w_i p_i \tag{3.28}$$

For any given model i and j and covariance matrix Σ, we can formulate error of the model as follows

$$f_2(w) = w \sum w^T \tag{3.29}$$

Using the above equation, we can further construct any multi-objective optimization problem as a bi-objective problem with the help of the following equation

$$\underset{w \in C}{min} \begin{bmatrix} w\ p^T \\ w\ \sum w^T \end{bmatrix} \tag{3.30}$$

where $C = \{w \in \mathbb{R}^n; \sum_{i=1}^{n} w^T \mathbf{1} = 1; w^T\ \overline{\widehat{y}} = \overline{y}\ for\ i = 1, 2, \ldots, n\}$ and $\mathbf{1}$ is a unitary matrix.

Obtaining a solution for such a minimization problem is not that easy with respect to single-objective problems and therefore a scalarized approach can be used to aggregate function in a meaningful way.

One such approach is goal programming which is a special case of the multi-objective problem where we fix a goal value for each objective function and measure the deviations of the values of the objective function from their goal value over the feasible region.

The advantage of using such a methodology is that we are able to optimize a goal as a target value for each and then minimize the difference between each function and its goal rather than optimizing objective function directly.

Formally, the stated bi-objective problem can be reformulated as goal programming problem by assigning to each f_i a goal value g_i and minimizing the deviation $(f_i - g_i)^+ for i = 1, 2$ over a feasible region where $+$ refers to the positive part of the function.

In order to simplify our methodology, let us define $g_1 = p^*$, $g_2 = 0$ where $p*$ denotes the desired level of performance on the model and we do not expect our goal vector $g = (p*, 0) \in\in \mathbb{R}^2$ to lie in the objective space.

Therefore, we propose four different strategies that help to reformulate multi-objective problems into a single objective problems to enhance not only the performance of the model but also the objectives of any given model.

The strategic use of weighted sum of deviations and Chebyshev goal programming with goal vector $g = (g_1, g_2, \ldots, g_N) \in \mathbb{R}^N$ and weight vector $w = (w_1, w_2, \ldots, w_N) \in (0, 1)$ is to minimize deviation and support the constructed objective function of any given problem. These weights can be either equal or fixed and depends on how many functions are considered for multi-objective problem.

Strategy 1. The first strategy we consider is Weighted Sum of Deviations (WSD) and formally we can write them as follows

$$\min_{w \in C} \sum_{j=1}^{N} w_j \ (f_j(w) - g_j)^+ \tag{3.31}$$

where C is a set of constraints.

A scalarized or aggregated function can be written as $F = \sum_{j=1}^{N} w_j (f_j - g_j)^+$ which can further be formulated as convex combinations that can help us to generate a new curve as a weighted average of deviations for each objective from its goal.

Strategy 2. The second strategy we use here is called *Chebyshev goal programming* and there is not much significant difference with respect to previous strategy as we try to minimize only the maximum weighted deviation instead of minimizing the sum of deviations. When this is done, this helps in minimizing other deviations which are smaller. More formally, we can write them as follows

$$\min_{x \in C} \left[\max_{j} w_j \ (f_j(x) - g_j)^+ \right] \text{ for } j = 1, 2, \ldots, N \tag{3.32}$$

where C is any constraint defined in Equation 3.27.

Strategy 3. This strategy, called *joint entropy*, helps us to understand the uncertainty or divergence associated between two models. The joint entropy of n models can be formulated as follows

$$\begin{aligned} &H(x_1, \ldots, x_n) = -\textstyle\sum_{x_1 \in \chi_1} \cdots \\ &\textstyle\sum_{x_n \in \chi_n} P(x_1, \ldots, x_n) \ \log_2 \left[P(x_1, \ldots, x_n) \right] \end{aligned} \tag{3.33}$$

More formally, to understand how much each of these models diverge from each other, we can formalize them as $x = (x_1, \ldots, x_n) \in p_n, y = (y_1, \ldots, y_n) \in q_n then for i = 2, \ldots, n$, it holds as follows

$$z_i = \min \left\{ \sum_{j=1}^{i} p_j, \sum_{j=1}^{i} q_j \right\} - \sum_{j=1}^{i-1} z_j \tag{3.34}$$

for any $z = x \wedge y$ and $p_n\, q_n$ are respective marginal probability distributions. We have to keep in mind that such measure helps in understanding the diversification in a better way and is non-negative and concave.

Strategy 4. Another variant of strategy 3 is to use cross entropy rather than joint entropy for understanding diversification among models. The idea of using cross entropy is based on importance sampling. For instance, if we take a random sample $x_1, \ldots, x_n$ based on importance sampling with density g on χ and using unbiased estimator ℓ and likelihood ratio, we can evaluate to minimize the distance of cross entropy which is equivalent to solving maximization problem as follows

$$\max_{v} \int g^*(x)\ \ln\ f(x; v)\ dx \tag{3.35}$$

where $g^*(x) = \frac{I_{\{S(x) \geq \gamma\}}\ f(x;u)}{\ell}$ is the density measure and $f(.; v)$ is a family of densities.

So far, we have been asserting that it is possible to formulate the given bi-objective problem into a goal programming problem to generate an optimal solution.

But we do not know any such optimal solution obtained for the goal programming problem is also the optimal solution to the bi-objective problem. We can formalize a theorem in this context to see if it is true.

Theorem 3.6.1. *If x^* is the optimal solution for goal programming then this also serves as a unique minimizer or Pareto optimal point for the bi-objective problem.*

Proof. Using weighted sum of deviations method, we can approach to prove this theorem for the goal programming problem assuming that x^* is the unique global minimizer of

$$\min_{x \in C} \left[w_1 (x p^T - p^*)^+ + w_2 (x \sum x^T) \right] \tag{3.36}$$

Let us assume further that x^* is not a global optimal solutions or Pareto optimal solution for the bi-objective problem, $\exists$ a point $\widehat{x} \in C$ with condition either $\widehat{x} p^T < x^* p^T or \widehat{x} \sum \widehat{x} < x^* \sum x^{*^T}$. Therefore, we can say that $w_1 (\widehat{x} p^T - p^*)^+ + w_2 (\widehat{x} \sum \widehat{x}) < w_1 (x^* p^T - p^*)^+ + w_2 (x^* \sum x^{*^T})$ which implies that $\widehat{x}$ is a global minimizer of bi-objective problem and this is a contraction to what we assumed. □

Chapter 4

Dataset and Implementation

In this chapter, we dedicate our discussion to the dataset, the strategies that were adopted to treat data, and the software interface in which the task on this dataset was performed. We begin our discussions in section 4.1 on the dataset followed by other related discussion in section.

4.1 Data description

The dataset comes from one of the leading financial institutions which consist of 39970 data points as a loan application. The exact source of the data is not available with us since it was once hosted on a Italian University website as a public competition. We have no further information regarding this if the dataset once hosted on a Italian university website is still present or deleted after the competition. We received this dataset while collaborating with a University Professor and doctoral students.

Each record in the dataset reflects characteristics of loan applicant in 30 variables that are set of information on socio-demographic (table 4.3), client equipment (table 4.4), client history (table 4.5) and other characteristics related to customer behavior (table 4.6). The target variable "ClientStatus" in the dataset is primarily of three categories as follows,

- The *category 0* is labeled as a regular client which is considered to be a good loan applicant.
- The *category 1* is labeled as a client with some kind of litigation and is considered as a bad loan applicant.
- The *category 2* is labeled as a client with recovery status and is also considered as a bad applicant.

The table 4.1 shows a distribution of the target variable without variable transformation and the subsequent table 4.2 shows the distribution of transformed target variable as a binary response variable (category 1 and category 2 was merged as one label). All other covariates in the dataset that are categorical were transformed using the label encoding procedure for predictive modeling purposes.

Table 4.1: Distribution of target variable without transformation.

Client label	Number of clients
0	38442
1	1254
2	304

Table 4.2: Distribution of target variable after transformation.

Client label	Number of clients
0	38442
1	1558

Table 4.3: Socio-economic variable description.

Variable	Description	Type
AGE	loan applicant age	discrete
REGIONE	Location details	categorical
ANZ_BAN	Age of the current account (expressed in years)	discrete
RESIDENZA	Type of Residence (owner or tenant)	categorical
ANZ_RES	Seniority of residence in the current residence (expressed in years)	discrete
STA_CIVILE	Marital status (married, single, divorced . . .)	categorical
NUM_FIGLI	Number of child	discrete
SESSO	Gender	categorical
REDDITO_CLT	Applicant income	continuous
REDDITO_FAM	Family income	continuous
PROFESSIONE	profession	categorical
NAZ_NASCITA	Country of birth	categorical
ANZ_PROF	Working seniority (expressed in years)	discrete

A prior probability for the target variable shows that 96.11% of class label 0 and 3.9% of class label 1. Based on the results about the target variable, we underline that the data is composed of 39970 observations and 30 explanatory variables.

There are several ways to do feature selection or feature engineering and in this respect, we choose to sketch variable importance plots using in-built functions of boosted classification trees. The features are ranked on an importance scale of 1 to 100. We restricted to include only the top 10 ranked features in the model for

Table 4.4: Client equipment variable description.

Variable	Description	Type
CANALE_FIN	Financing channel (agency, web, telephone ...)	categorical
NUM_PRA_PP	Current Personal Loans - number of practices	discrete
esposizione_pp	Current personal loans - residual amount on the balance	continuous
durata_residua_pp	Current personal loans - residual duration to balance	continuous
NUM_PRA_CC	Total finalized loans in progress - number of practices	discrete
esposizione_CC	Total finalized loans in progress - remaining balance	continuous
durata_residua_CC	Total finalized loans in progress - residual maturity at the balance	continuous
NUM_PRA_CP	Card - Customer holding card	discrete
esposizione_CP	Card - Credit Card Display	continuous

Table 4.5: Client history variable description.

Variable	Description	Type
NUM_SAL_PP	Personal loans paid in the last 24 months - number of files	discrete
NUM_SAL_CC	Finalized loans paid in the last 24 months - number of practices	discrete

Table 4.6: Client behavior variable description.

Variable	Description	Type
num_men_rit	number of late payments from origin (in months)	discrete
score_cmp_qe	internal behavioral score	continuous
score_cmp_cb	credit bureau behavioral score	categorical
num_sal_rec	number of recovery ascents in the last 12 months	discrete
num_mes_rec	number of months to recovery in the last 12 months	discrete

getting better performance after evaluating different possibilities of feature inclusion. The features ranked are reported in figure 4.1 .

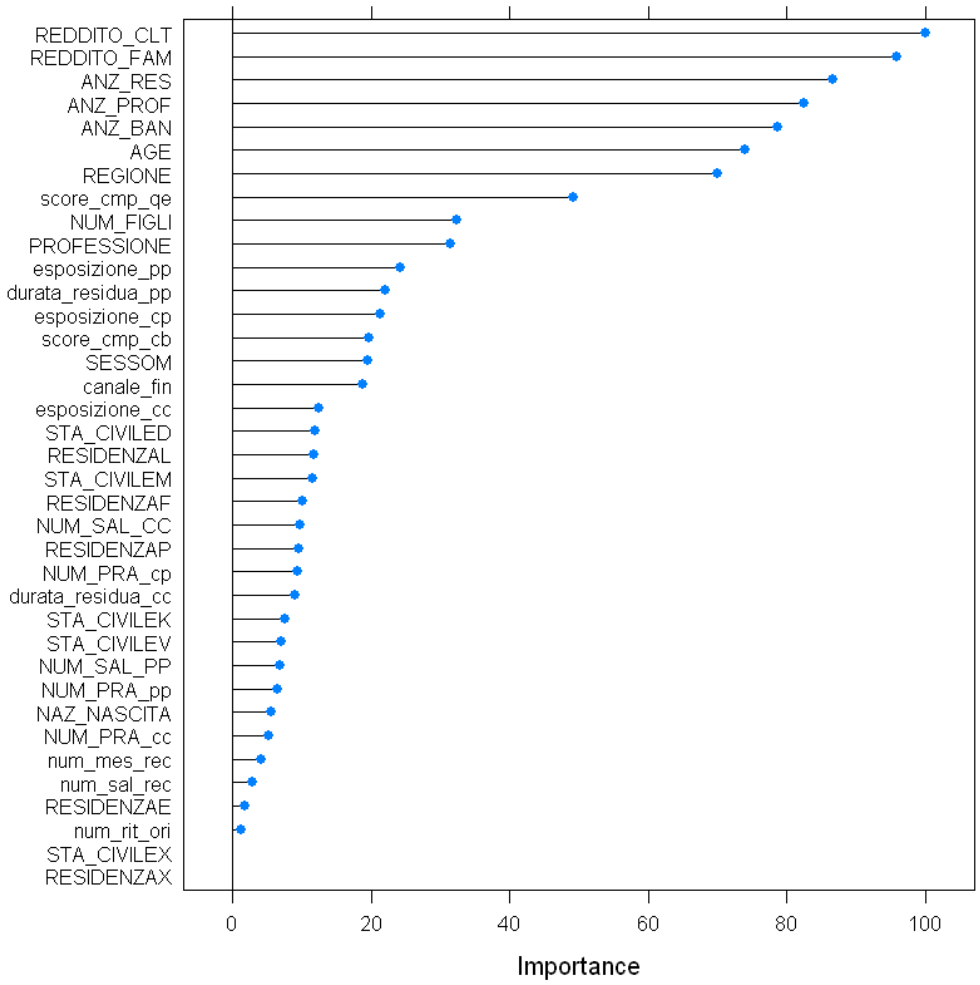

Figure 4.1: Feature importance graphical presentation.

We see the distribution of those important variables that are ranked highest on the variable importance scale and are also considered for the modeling purpose. Figure 4.2 shows the distribution of the considered variables on the diagonal. The lower inside of the diagonal shows bi-variate scatters plots with their fitted line and the upper part of the diagonal shows value of correlation and significance level indicated by stars. The stars associating to significance level takes range of p-values(0, 0.001, 0.01, 0.05, 0.1, 1).

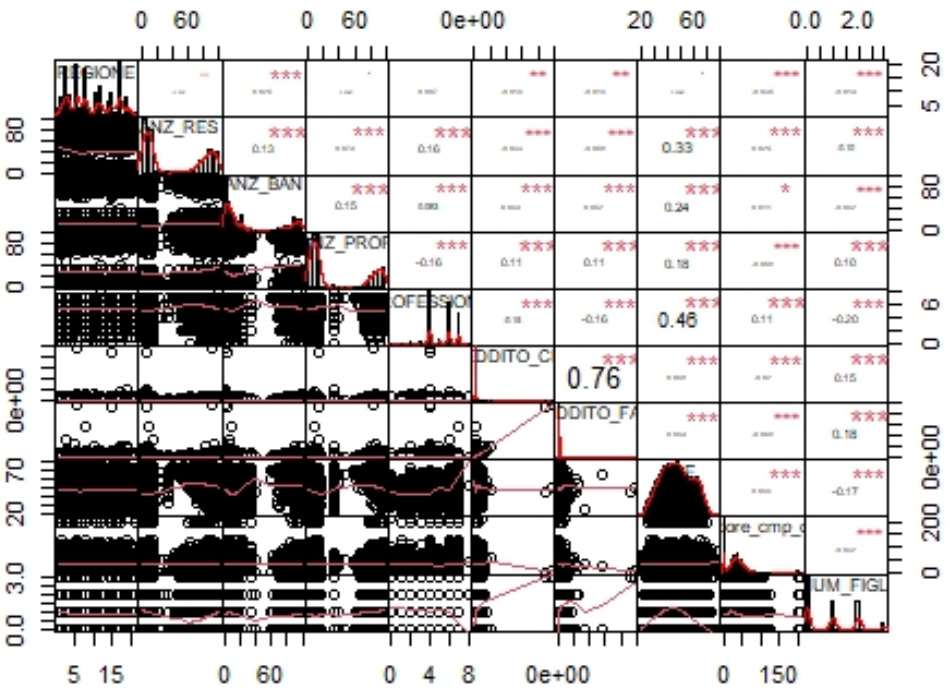

Figure 4.2: Correlation graph of the 10 highest important variable.

4.2 Data Handling

The explanatory variable in the dataset is few set of information that reflects socio-demographic characteristics, customer equipment, customer history, and other things related to customer behavior. It is obvious to see from section 4.1 the dataset has a class imbalance problem and any attempt to approach modeling tasks on this dataset would lead to over-fitting results. Therefore, it is important to treat the data with any kind of statistical technique that resolves the class imbalance problem.

For instance, Some of the well-known methods for treating the class imbalance problem is to re-sample the training set either using the under-sampling or over-sampling technique. Moreover, applying q-fold cross-validation in the right way, ensemble different re-sampled datasets, re-sample with different ratios, cluster the majority class, or design any different model are a few of the many alternatives to deal with the data imbalance problem.

In this contribution, Synthetic Minority Oversampling Technique (SMOTE) strategy is appealing to treat the imbalance problem of data. SMOTE (Chawla et.al.,2002) [24] creates synthetic observations based on available minority observations that work on the principle of k-nearest neighbors. It generates new instances that are not just copies of the available minority class, but a basic rule is to take samples of feature space for each target class and its nearest neighbors. In this way, it increases the features available to each class and makes the samples more general. The following steps explain in brief how SMOTE works,

- A total number of observations from the given dataset is set up.
- Assuming that binary class distribution is 1:1, the iterative procedure of the algorithm selects positive class randomly.
- By default, 5 is considered as the nearest neighbor of the selected positive class.
- As a next step, the algorithm generates synthetic classes of the selected class.
- To generate synthetic class, a distance metric is used between variable class and its neighbor.
- The difference obtained is multiplied by a random value from 0 to 1 which is added to the selected variable class.

Traditionally sampling methods until 1997 held a special position in many relevant studies as the goal was to create a dataset that is balanced class distribution so that

any chosen classifier can work well to distinguish between majority and minority classes. Much research over the years has proved that sampling techniques are a useful way to achieve overall accuracy from the deployed classifier.

Few of the main sampling methods that we discuss in this context are the following:

- Undersampling and Oversampling are random sampling procedures where class distribution for majority instances in undersampling is discarded at random to achieve a more balanced distribution of the considered class, while in oversampling techniques, the class instances for minority class distribution are copied and repeated until a more balanced distribution of the whole sample is achieved. However, both of these methods have serious drawbacks that bias decision making since the majority of the data are discarded and hampers the performance of classification methods. The limitations imposed in these methods are addressed by other sophisticated methods, unlike SMOTE .

- SMOTE in an appealing technique to treat the imbalance of data. Such a method creates synthetic observations based on existing minority observations that work on the principle of k-nearest neighbors. It generates new instances that are not just copies of the existing minority class, the rule is to take samples of feature space for each target class and its nearest neighbors. In this way, it increases the features available to each class and makes the samples more general. The training set due to SMOTE therefore is changed by adding up synthetically generated minority class distribution to achieve balanced samples.

- Cost-sensitive learning is another method that helps in handling classification problems dealing with an imbalanced dataset. The method handles the cost associated with misclassifying observations. Rather than creating balanced data distributions, it assigns cost matrices which help in handling misclassification cost as a way to solve the problem of working on an imbalanced dataset. For more details on such methods, refer to the study of Lopéz and Fernández,2013 [94] .

- Random Oversampling Examples (ROSE) is a different technique based on the bootstrap method that helps in the task of binary classification for handling minority classes and is capable of dealing with continuous or categorical data by assigning synthetic examples using conditional estimates of two classes.

The context of data handling in this thesis is purely based on SMOTE for generating a balanced dataset.

4.3 Software environment

The primary programming environment for the execution of tasks was carried in R software using functionalities of the following packages as a shortlist apart from others. All of these packages can be referred to the CRAN list `https://cran.r-project.org/web/packages/available_packages_by_name.html` which is a database of R software packages. Few other adjunct software were used to support the overall analysis in this thesis.

- *caret* stands for classification and regression training and is used for many tasks on predictive modeling. The package contains many in-built functions that are used for data splitting, pre-processing, feature selection and feature importance, model tunning parameter.

- *randomForest* is a package that can be used for classification and regression tasks in supervised settings. It can be also used to assess proximities among data points in unsupervised settings.

- *hmeasure* is a performance metric that is used to assess the performance of classification tasks and is capable of assessing performance across multiple scenarios. It also addresses the limitations of other performance variants like AUC and the Gini coefficient.

- *entropy* is a package that has various in-built estimator to measure similarity and difference of probability distributions of a random variable.

- *mlbench* is a package that has collections of several real-world and artificial machine learning problems to practice and learn.

- *PerformanceAnalytics* is a package that supports the performance and risk analysis through different functions.

- *DMwR* is a package that handles the various tasks of data processing and mining.

- *ISLR* is a package that has various in-built functions for statistical learning.

- *caTools* is a package that mostly used for faster calculation of AUC apart from other additional functions.
- *SDMTools* is a package that includes various functions for model comparison based on different threshold settings.
- *MASS* is a package that supports various functionalities of modern applied statistics.
- *pracma* is a package that supports the various computations of mathematical optimization, linear algebra, and other mathematical calculation.
- *stats* is a package that supports statistical calculations and random number generation.
- *e1071* is a package that supports miscellaneous functions of probability and statistics.
- *gbm* is a package considered as an extension of the Adaboost algorithm and supports various functions of regression, loss measure, and statistical distributions.
- *BAS* is a package that supports Bayesian Model averaging based on stochastic or deterministic sampling without replacement for any considered posterior distribution.
- *BMA* is a Bayesian model averaging package and variable selection used for linear models, generalized linear models and survival models.
- *mgcv* is a package that supports various functions for generalized additive modeling and generalized additive mixed modeling.
- *gam* is a package that uses a back-fitting algorithm to combine different smoothing or fitting methods which helps in fitting a generalized additive model.
- *class* is a package that supports various functions for classification tasks.
- *gmodels* is a package that supports various functions for model fitting.
- *klaR* is a package that supports various functions for classification and visualization.

- *bnlearn* is a package that supports constructing the Bayesian network, Bayesian inference, and includes various functions for Bayesian analysis.
- *BART* is a package that supports non-parametric modeling as Bayesian additive regression trees for continuous, binary, and categorical covariates.
- *bartMachine* is a package that has extended features for building Bayesian additive regression trees.
- *optim* is an optimization package that supports various minimization and maximization functions.
- *quadprog* is a package that supports to solve quadratic programming.
- *tidymodels* is a package that supports the various tasks of machine learning and is an integrated framework of the tidyverse package.
- *tidyverse* is a collection of various packages and dependencies of R that supports many functions and tasks for data science.
- *GPareto* is a collection of packages that supports various functions of multi-objective optimization.

Chapter 5

Classifiers, optimization model and performance metrics

In this chapter, we discuss a set of classifiers and optimization models that were studied, reviewed and used to support the overall analysis to produce results in this thesis. We start with a brief description of each classifiers in section 5.1 followed by discussion on optimization model in section 5.2. The performance metrics is discussed in section 5.3 for assessing and comparing the studied models with respect to proposed models in this thesis.

5.1 A set of classifiers for predictive task

Many models fit for the analysis on imbalanced dataset and parametric models is one of the few that gives better performance. However, this is not always true as the enhanced performance on imbalanced dataset is achieved often with a smaller set of models.

Our approach in this thesis is rather different as we take an heterogeneous set of models in performing the analysis on imbalanced dataset. This heterogeneous set of models include parametric, non-parametric and ensemble models.

The analysis carried on imbalanced dataset (see section 4.1) were evaluated against different set of performance metrics. These performance metrics are briefly described in section 5.3. We start the following section with brief explanation of models that were taken in consideration for our analysis on imbalanced dataset.

5.1.1 Parametric model

Parametric models are a family of distributions that can be described using a finite number of parameters. In this case, we can know which kind of model would fit the data exactly. For instance, the equation

$$f_i = \beta_0 + \beta_1 x_i + e_i$$

infers that regression will take a linear line. The term f_i simply denotes the idea of response or target variable of any arbitrary data in the real world where supervised learning could be a good fit for analysis. This kind of model is often the choice for a predictive model as it helps to estimate better statistical properties.

Generalized linear model. As per Nelder et al. [102], the generalized linear model often helps to understand binary response variables that have error distributions other than a normal distribution. More precisely, each outcome of a given target variable is considered to be distributed as per the exponential family.

For any binary response variable data, for instance, we can say that $c = 0$ indicates false classes of prediction, and $c = 1$ indicates true classes of prediction. The variable c is a general notation to denote or represent any class values of target variable from a given arbitrary dataset. So, we assume x is a column vector of P predictors whose response probability can be modeled as $\pi = Pr(c = 1|x)$ and this further can be written like logistic regression model using the link function logit as,

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta^T x \tag{5.1}$$

where α is the intercept parameter and β^T represents the coefficients of the corresponding variable.

Naive Bayes. It is simply a probabilistic classifier based on Bayes theorem [108] having strong independence assumptions between the features. The decision rule in this classifier is known to be Maximum a Posteriori (MAP) . The classifier is a function that assigns a class label as follows

$$\widehat{f} = \underset{k\in\{1,\ldots,K\}}{\text{argmax}} \; p(c_k)\prod_{i=1}^{n} \; p(x_i|c_k) \tag{5.2}$$

for any K possible outcomes or classes.

5.1.2 Non-parametric model

In this kind of model, the model structure is specified from data and it is not determined a priori. For instance, $f_i = f(x_i) + e_i$ where the function is unknown and takes the structural form from data. The statistical estimation of such a model depends on the smoothness of the chosen function.

Decision Trees. In this context, referring to Breiman et.al. [15], we have selected both Recursive Partitioning and Regression Trees (RPART) and Conditional Inference Trees (CTREE) for supporting our analysis. The criteria of univariate splits of a dependent variable based on a set of covariates are quite similar in both RPART and CTREE . However, RPART usually employs information measures (such as Gini Coefficient (GC)) for choosing the co-variate while CTREE uses a significance test to select variables. The information gain in the majority of the tree algorithms is defined as

$$H(T) = I_E(p_1, p_2, \ldots, p_J) = -\sum_{i=1}^{J} p_i \log_2 \ p_i \tag{5.3}$$

where p_1, p_2, ..., p_J are properties of class values in sample T.

Generalized Additive Model. This is a special case of the generalized linear model (Hastie, T.J, 1986 [72]) where the predictor has dependencies among each other using some kind of smooth functions. The relation of the response variable concerning predictors is captured through

$$g(E(f)) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_m(x_m) \tag{5.4}$$

where the functions f_1, f_2, ..., f_m could be any specified parametric or non-parametric form.

K Nearest Neighbor. This method is one of the standard non-parametric methods used for classification and regression tasks. The input has k closest examples scattered in any feature space and output as a class object decides whether it can be used for classification or regression. Due to this nature of the algorithm, it is also called instance-based learning or lazy learning to compute all the functions locally at the cost of retarded function. For any given data pairs $(X_1, F_1), (X_2, F_2), \ldots, (X_n, F_n)$ such that $X,\ F \in \mathbb{R}^d\ x\{1, 2\}$. These data pairs can be reordered under a given probability distributions and norm as $\|X_1 - x\| \leq \ldots \leq \|X_n - x\|$ where F is the

class label of X. This is the primary setting of k nearest neighbor. For more details on this classifier, refer to the study by Cover,T.(1968) [36] .

5.1.3 Ensemble Models

Ensemble models is a technique based on considering multiple models and combining them to enhance the predictive performance rather than relying on the single best model. Ensemble models are not based on a simple average but a weighted sum as represented in the following equation

$$\widetilde{f}(x;\alpha) = \sum_{j=1}^{p} \alpha_j f_j(x) \tag{5.5}$$

where f denotes any response variable and x denotes co-variates. There exist many types of ensembles as an averaging technique which is computationally demanding but provides better results in critical decision making task (for instance, predicting business failure). The considered ensemble models are described below.

Random Forest. This method of tree generation is sort of ensemble learning (Breiman L, 2001) [18] which is used for both classification and regression purposes based on bootstrap samples of the training data and random feature selection. After training, predictions for unseen samples $x^{'}$ can be made by averaging predictions from all the individual regression trees on $x^{'}$ or simply using a majority vote technique in case of a classification problem. The general form of bootstrap aggregation can be presented as

$$\widehat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x^{'}) \tag{5.6}$$

Bagging. It is an approach to improve the stability and accuracy of a machine learning algorithm for any classification or regression task (Breiman L, 1996) [16] . Assuming that a learning set L consists of data $\{(f_n,\ x_n),\ n = 1, \ldots, N\}$ where f is a response variable and $\varphi(x, L)$ can become a procedure for using this learning set. The repeated bootstrap samples can be developed from learning set given that the response variable is numerical $\varphi_B(x) = av_B\varphi(x, L^B)$ where av_B is the average or expected value of any bootstrapped samples. In general, they are a model averaging approach that helps to reduce variance and avoid over-fitting results.

Gradient Boosting Machine. This technique produces a predictive model in the form of an ensemble of weak prediction models (Breiman L, 1999 [17], Friedman, J. H,2001 [55]) mainly as decision trees. Unlike any general boosting methods, it develops models in each stage by generalizing them using some arbitrary differentiable loss function. The core idea behind this algorithm is that it assumes a real-valued function for response variable f and seeks an approximation $\widehat{f}(x)$ using a weighted sum of functions $h_i(x)$ for some class of weak learners as the following

$$\widehat{f}(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + \text{const} \tag{5.7}$$

where γ_i is a sort of multiplier in the form of coefficient values.

Bayesian Moving Average. This approach to predictive models is simply based on selecting and combining models (Fragoso et.al, 2018 [53]) based on their posterior probabilities. The choice of a single best model may lead to overconfident inference and henceforth riskier decision. Therefore, the Bayesian approach for predictive models is desirable in handling model uncertainty. The core of each model selection is based on posterior distribution using Bayes Theorem resulting in

$$\pi(\theta_l|F,\ K_l) = \frac{L(F|\theta_l, K_l)\pi(\theta_l|K_l)}{\int L(F|\theta_l, K_l)\pi(\theta_l|K_l)d\theta_l} \tag{5.8}$$

In the above equation, θ_1 are model specific parameters for any observed data F for any given model K and likelihood function L.

Bayesian additive regression trees. They are the "sum of trees" model where each tree is constrained by a regularization prior (HA Chipman, 2008 [26]) and is very similar to Gradient Boosting which uses Bayesian back-fitting to execute MCMC sampling from a general additive model posterior distribution. The selection of combining multiple models is very desirable as it avoids overconfident and riskier decisions arising from one particular model. Using the Bayesian Additive Regression Trees (BART) model, we can obtain posterior distribution based on Bayes Theorem resulting in

$$\pi(\theta_l|F,\ K_l) = \frac{L(F|\theta_l, K_l)\pi(\theta_l|K_l)}{\int L(F|\theta_l, K_l)\pi(\theta_l|K_l)d\theta_l} \tag{5.9}$$

as the integral in the denominator for each prior distribution represents a marginal distribution of the dataset overall parameter values specified in model K_l. Moreover, θ_1 are model specific parameters for any observed data F for any given model K and likelihood function L.

5.2 Optimization Model

The optimization problem that we have considered in the context of our research problem is not exactly convex since they do not follow Disciplined Convex Programming (DCP) rule-set and as a result to achieve an optimal solution, one has to look for the solution of the problem from the quadratic programming techniques. We studied and implemented many different optimization algorithms like Nelder-mead algorithm, BFGS algorithm, L-BFGS-B algorithm, CG algorithm, SANN algorithm.

All of these algorithms provide a common optimal solution as a globally optimal solution in each case that converges to zero with an increasing number of iterations. In the following paragraph, we discuss a short description of each of these optimization algorithms.

- The *Nelder-mead algorithm*, also known as the downhill simplex method, is a numerical method technique to find a minimum or maximum point of an objective function. It is one of the non-linear optimization technique and it is a direct search method that can converge to non-stationary points. For any given non-linear function $f : \mathbb{R}^n \to \mathbb{R}$, it deploys only function values at some points in $\mathbb{R}^n$ and does not necessarily approximate gradient values at any of these points. In simpler terms, it is a simplex method defined on a convex hull of $n+1$ points $x_0, \ldots, x_n \in \mathbb{R}^n$ that forms triangles or other geometric shapes. Any set of functions $f_j := f(x_j)$ is a non-degenerate and does not lie in the same hyperplane. The function value decreases at points after a sequence of transformations and is terminated when it is sufficiently small. Refer to [101] for further details.

- BFGS is an iterative technique or hill-climbing optimization techniques that search for stationary points of an objective function and the gradient is zero for necessary conditions of optimality. For any given or defined optimization problem, the basic goal is to minimize a scalar function $f(x)$, $x \in \mathbb{R}^n$, which produces a quasi-newton method that approximates $\nabla^2 f(x^*)$ as the sequence of iteration progress, where x^* is the new value of x obtained from iteration.

  The following steps give a short guide to do any transition from the current state of approximation to new approximation using a line search paradigm method (refer to [52] for additional details):

  - Computing a search direction $d = -H_c^{-1} \nabla f(x_c)$. The H_c is the current state of approximation.

  - Finding $x_n = x_c + \lambda d$ using a line search paradigm, where x_n is a new approximated value and λ is a scalar factor.
  - Using the current approximation x_c and new approximation x_n to update H_c and obtain H_n. The term H_n is the new state of approximation.

- L-BFGS-B is a technique that is popular for parameter estimation in machine learning. Unlike BFGS, L-BFGS-B uses the inverse Hessian matrix for finding solutions through the search space. The algorithm initially estimates an initial value x_0 and finds a better estimate of this initial value in a sequence of estimates as a derivative of the function $g_k := \nabla f(x_k)$. Thus this algorithm works like BFGS with the difference that it uses the inverse Hessian matrix for computation. The method is more adequate for working with bounded constraints as it tries to find fixed and free variables at every single step based on a simpler gradient method to achieve higher accuracy. Refer to the study by Zhu et al. (1997) [130] for any further details.

- The CG method is one of the numerical techniques for solving a system of linear equations. It is an iterative algorithm for various minimization tasks using some kind of quadratic function as $f(X) = \frac{1}{2}X^T AX - X^T b$ where $X \in \mathbb{R}^n$, where A is a positive definite symmetric matrix and X is a non-zero vector of n dimension. The second derivative for any symmetric positive definite matrix acts as a unique minimizer and solves the initial problem. We can use this algorithm to solve any optimization problem both as a direct method or iterative method. The majority of the experts view this algorithm as a direct method as it is able to produce exact solutions after a few number of iterations and the solutions obtained are highly unstable with respect to small changes. It is further possible to set up the convergence of approximate solution to an exact solution using convergence theorem. Refer to the study by Hestenus,M.(1952) [76] for additional details.

- SANN is a probabilistic technique by which we can approximate the global optimal solution of a given objective function and is a more preferable approach in many cases compared to gradient descent. SANN is a variant of the simulated annealing and relates widely to the class of stochastic global optimization methods. This algorithm is relatively slow since it uses only function values, and for achieving any acceptance probability as a value it uses Metropolis function. Such a kind of algorithm is popular in solving combinatorial optimization problems. Refer to Dimitris [12] for further details on this topic.

5.3 Performance assessment

In this section we briefly describe the performance metrics reflecting the accuracy and error of classification models. The accuracy metrics used to assess the classification models discussed in this thesis are the following:

- *H* is an alternative to the popular ROC curve performance measure for any classification or diagnostic task. It is being believed that misclassification cost is not handled properly by ROC or AUC measures, which in turn H proves to be a better metric for assessing the performance task. Furthermore, the AUC measure is classifier dependent, and generally we should choose a weight prior. This is not the case with H, as it is classifier independent and controls the cost prior to a better way. For more details on this measure, one can refer David (2009) [65] .

- *AUC* is a diagnostic tool or performance measure metrics associated with ROC and is measured at various threshold settings. AUC value signifies the degree of separability and helps to discriminate between class labels. The higher the value of AUC, the better is the performance of the classifier.

- *AUCH* is a geometrical representation ROC curve that allows us to select points on the curve under some optimality conditions of cost and class distribution. It is very much similar to Pareto-front in the case of multi-objective optimization. We can also say that the convex hull can be seen as a discretization of the scores that achieve higher AUC . It can also be inferred as a hybrid classifier that reaches any random point on the convex hull by stochastic interpolation between two neighboring classifiers. For additional details, one can refer to Provost and Fawcett, 2001 [107] .

- *MER* is a sort of decision rule that helps in minimizing the probability of error or simply error rate. The loss function considered is symmetrical or zero-one loss function which helps in minimizing error by maximizing posterior probability. For more details on this measure, one can refer H., David (2009) [65] .

- *MWL* is a technique of weighting the error for the different cost associated with misclassification. A weighting function is used either to define the cost matrix or response vector for classification. The lower is the weighting error, the better is a contribution to decision making. For more details on this measure, one can refer David (2009) [65] .

There are other metrics, like Kolmogorov Statistics (KS), GC, Sensitivity at 95 percent Specificity (Sens.Spec95), Specificity at 95 percent Sensitivity (Spec.Sens95) but is not used in the current scope of analysis as most of the performance assessment and their inference is already captured through other metrics in use. For more details on this measure, one can refer David (2009) [65] .

Chapter 6

Results

An ensemble model is a winning formula for almost
all data science and machine learning competitions.

Anonymous

This chapter discusses the results obtained from the proposed idea in chapter 3 . The solutions from the proposed approach are considered as weights to build a weighted ensemble model which is a linear combination of parametric, non-parametric, and ensemble models. The proposed weighted ensemble model WTM and its robustness were checked against all well-known parametric, non-parametric, and ensemble models. The robustness check of WTM was subsequently carried against different weighting procedure like Weighted method using correlation (WMCOR), Optimal Weighted Method (OWM), Squared Weighted Method (SWM), Negative Weighted Method (NWM) and Equally Weighted Method (EWM) and the results at hand infers that our proposed method WTM provides better performance.

6.1 Model averaging results

The main motivation to develop a novel way of ensemble model using the knowledge of model averaging is to bring an approach that seeks to minimize the error between models. We have been able to do so by measuring and minimizing the co-variance of the error between models concerning to a set of constraints as stated in the equation 3.9.

To be more precise, the difference between observed and predicted value for each model were considered as error of the model. Both co-variance and correlation were used to understand which of these two provides a better minimization of the error

between models. The results of ROC in figure 6.2 and 6.5 advocates the idea that development of ensemble model based on co-variance approach outperforms all other considered approach in this thesis.

The weighted ensemble model is then evaluated against the existing popular machine learning models (parametric, non-parametric and ensemble model as discussed in 5.1) using a set of different performance metrics. The idea to use different metrics to assess both accuracy and error of the model is to understand overall how well the proposed model is doing concerning to existing models.

What happens if we do not consider the idea of diverse weighting options to build an ensemble model and stick to just one optimal value. We explored this inquisitiveness solving the optimization algorithm discussed in section 5.2 and we obtained a unique value of 0.24 as an approximated solution for all mentioned optimization algorithm. The figure 6.1 geometrically shows the location of optimal value around its other neighboring solutions.

This unique value was considered as an optimal weight for all the considered models to develop the proposed ensemble model. How well the ensemble model performs using an optimal weight with respect to other weighting strategies is assessed using ROC in figure 6.5 .

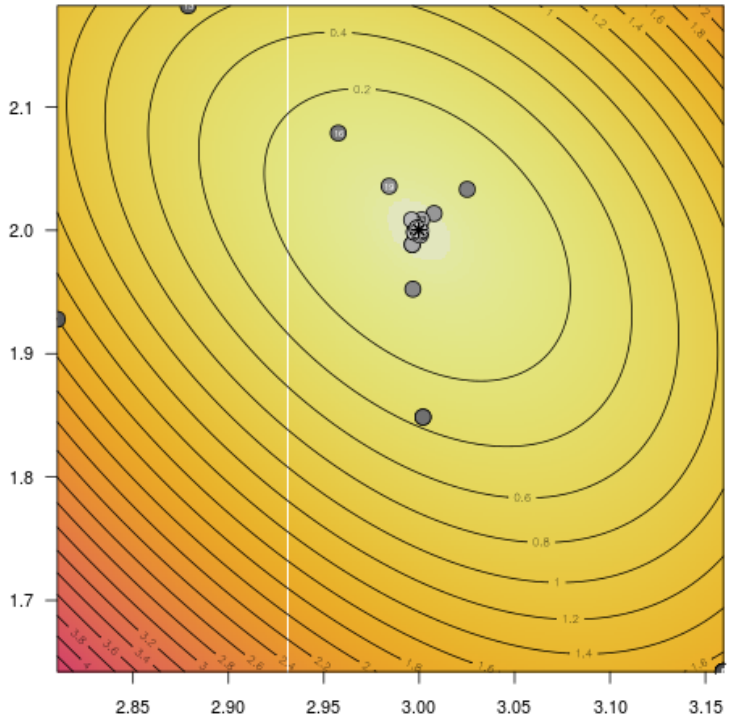

Figure 6.1: Optimal solution of the Nelder-mead, BFGS, L-BFGS-B, CG and SANN algorithm.

The set of models that were studied and compared in the analysis are *CTREE*, *RPART*, *Generalized Linear Models (GLM)*, *Random Forest (RF)*, *Bootstrap Aggregating (BAGG)*, *Boosting (BOOST)*, *BMA*, *Generalized Additive Model (GAM)*, *K Nearest Neighbor (KNN)*, *Naive Bayes (NB)*, *BART*, and proposed weighted model based on minimizing co-variance between errors of the models. The proposed weighted model in addition were evaluated against other weighting methods like optimal weight, equal weight, weight obtained from correlation, squared weight and negative weight.

The modeling approach for having predicted value from each of these models mentioned just above was kept the same and their training process was built on $k = 10$ cross-validation criteria. The performance metrics to assess accuracy and error for each of these models has been computed on out of sample data and are reported in table 6.1, table 6.2, figure 6.3, figure 6.4 .

From table 6.1, Analyzing the accuracy measure performance metrics for different models is reported as follows,

- For H metrics, it infers that BMA and GAM is the best performing model whereas NB is the worst performing model.
- For AUC and AUCH metrics, it infers that CTREE, RF, BAGG and WTM are few of the best performing model including performance overlap whereas GLM, BMA and GAM are worst performing model including performance overlap.
- For Sens.Spec95 and Spec.Sens95 metrics, RF is the best performing model and KNN is the worst performing model.

Table 6.1: Performance metrics capturing accuracy of the model.

Metrics	CTREE	RPART	GLM	RF	BAGG	BOOST	BMA	GAM	KNN	NB	BART	WTM
H	0.62	0.59	0.67	0.64	0.61	0.57	**0.78**	**0.78**	0.39	**0.38**	0.50	0.60
AUC	**0.94**	0.80	**0.67**	**0.94**	**0.92**	0.91	**0.67**	**0.67**	0.79	0.77	0.82	**0.92**
AUCH	**0.94**	0.80	**0.67**	**0.94**	**0.92**	0.91	**0.67**	**0.67**	0.79	0.77	0.82	**0.92**
Sens.Spec95	0.24	0.15	0.04	**0.63**	0.50	0.48	0.20	0.20	**0.03**	0.05	0.04	0.58
Spec.Sens95	0.10	0.07	0.05	**0.78**	0.77	0.73	0.06	0.06	**0.03**	0.06	0.04	0.73

Referring to results in table 6.2, Analyzing the error measure performance metrics for different models is reported as follows,

- For MER metrics, it infers that WTM is the best performing model whereas GLM and NB is the worst performing model.

- For MWL metrics, it infers that WTM is the best performing model whereas GLM is the worst performing model.

Table 6.2: Performance metrics reflecting error in the model

Metics	CTREE	RPART	GLM	RF	BAGG	BOOST	BMA	GAM	KNN	NB	BART	WTM
MER	0.16	0.16	**0.26**	0.13	0.14	0.15	0.15	0.25	0.16	**0.26**	0.16	**0.14**
MWL	0.18	0.19	**0.29**	0.12	0.14	0.15	0.16	0.26	0.19	0.28	0.19	**0.14**

However, the proposed model WTM in the overall accuracy and error metrics analysis is better compare to well-known parametric, non-parametric, and ensemble models.

The label mentioned in the legend of figure 6.2 is as follows,

- CTREE infers conditional tree model.
- RPART infers recursive partitioning tree model.
- GLM infers generalized linear model.
- RF infers random forest model.
- BAGG infers bagging model.
- BOOST infers gradient boosting model.
- BMA infers Bayesian moving average model.
- GAM infers generalized additive model.
- KNN infers k-nearest neighbor model.
- NB infers Naive Bayes model.
- BART infers Bayesian additive regression trees.
- WMCOR infers proposed model based on the idea of measuring the correlation of error between the models.
- EWM infers the proposed model based on the idea of equal weighting.
- WTM infers proposed model based on the idea of measuring co-variance of error between the models.

Similarly, the label mentioned in the legend of figure 6.5 is as follows,

- WTM infers proposed model based on the idea of measuring co-variance of error between the models.
- WMCOR infers the proposed model based on the idea of measuring the correlation of error between the models.
- OWM infers the idea of using optimal weight.
- SWM infers the idea of squared weight.
- NWM infers the idea of negative weight.
- EWM infers the idea of equal weighting.

An additional inference that is evident from the figure 6.2 and 6.5 is the performance overlap in a few of the performance metrics. To sort out or rank the classifiers when ROC intersects at different points is a challenging problem. It is being believed that *stochastic dominance* can be an effective tool for ranking the classifier or models based on different criteria.

However, such a solution is limited to address the issue of intersecting ROC at two different points on the curve and is not generalizable to multiple intersections of ROC at multiple points. For details on such topics, refer to the study by Figini et. al [61], Hand [66], and Muliere [99] .

Following table 6.3, Analyzing the accuracy measure performance metrics for different weighting methods is reported as follows,

- For H, AUC, and AUCH metrics, it infers that WTM is the best performing weighting method for developing ensemble model whereas OWM is the worst performing weighting method.
- For Spec.Sens95 and Sens.Spec95, it infers that WTM is the best weighting method whereas SWM and OWM is the worst performing weighting method including performance overlap.

Table 6.3: Accuracy assessment of different weighting methods.

Metrics	WTM	WMCOR	EWM	OWM	SWM	NWM
H	**0.60**	0.06	0.27	**0.01**	0.12	0.27
AUC	**0.92**	0.63	0.78	**0.55**	0.67	0.78
AUCH	**0.92**	0.63	0.78	**0.55**	0.67	0.78
Spec.Sens95	**0.58**	0.15	0.25	**0.10**	**0.10**	0.25
Sens.Spec95	**0.73**	0.11	0.33	**0.05**	**0.05**	0.33

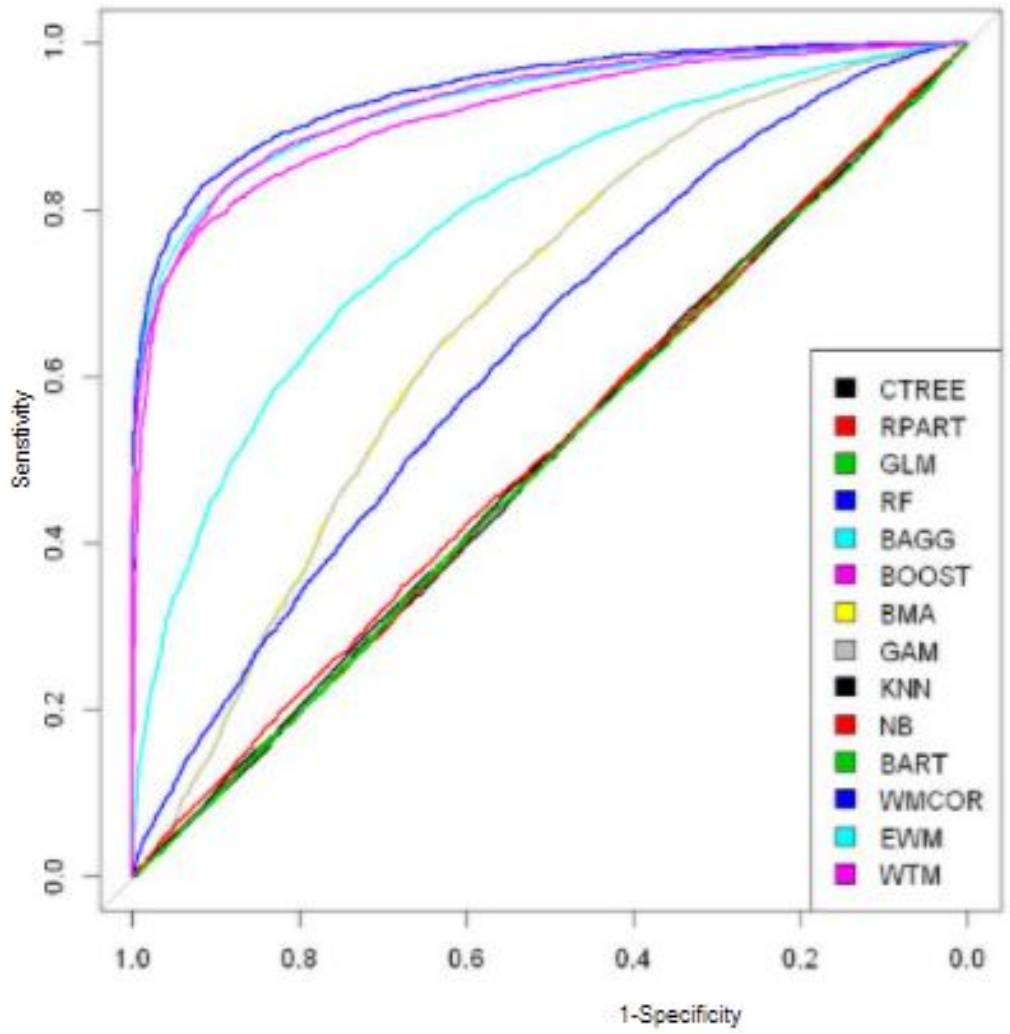

Figure 6.2: ROC curve of parametric, non-parametric, ensemble and proposed weighted model.

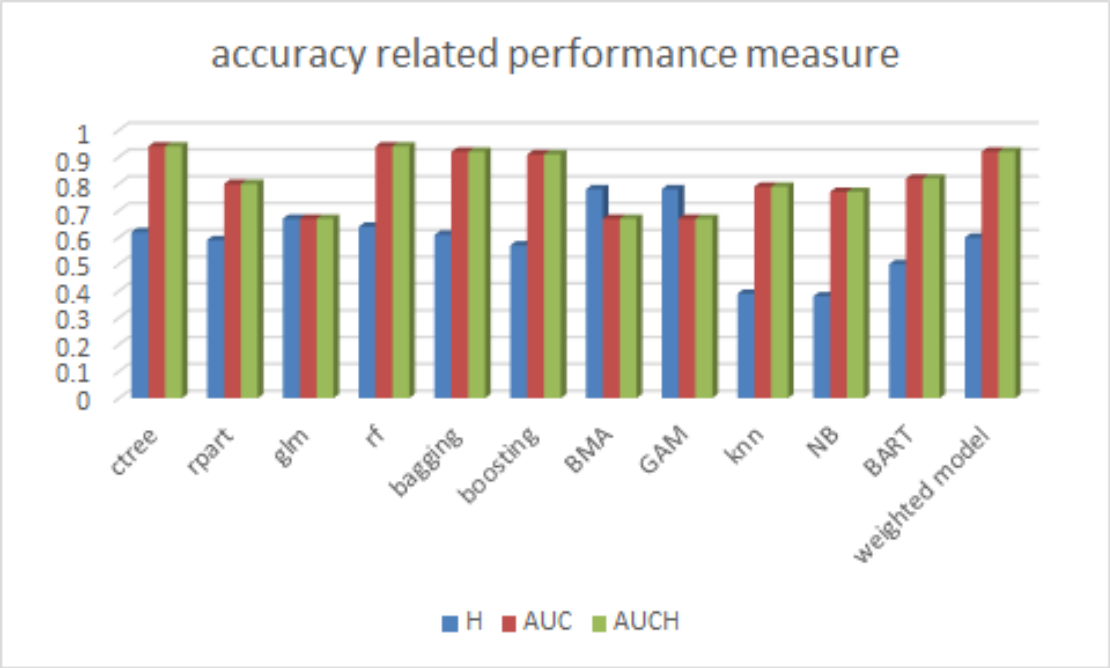

Figure 6.3: Accuracy metrics graphical representation.

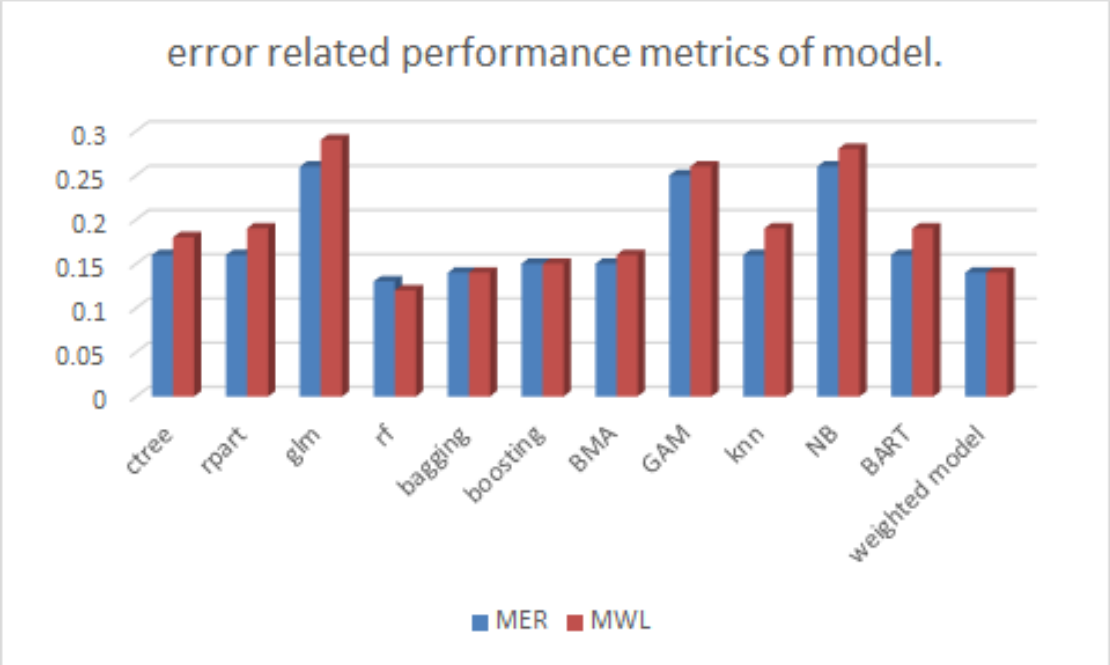

Figure 6.4: Error metrics graphical representation.

Following table 6.4, Analyzing the error measure performance metrics for different weighting methods is reported as follows,

- For MER metrics, it infers that WTM is the best performing weighting method for developing ensemble model whereas OWM is the worst performing weighting method.
- For MWL, it infers that WTM is the best weighting method whereas OWM is the worst performing weighting method.

Table 6.4: Error assessment of different weighting methods.

Metrics	WTM	WMCOR	EWM	OWM	SWM	NWM
MER	**0.14**	0.40	0.28	**0.44**	0.35	0.28
MWL	**0.14**	0.40	0.28	**0.46**	0.36	0.28

The weighted model using co-variance technique WTM is compared against other weighting methods like WMCOR based on correlation, EWM based on co-variance, OWM based on different optimization algorithm, SWM based on co-variance and NWM based on co-variance. Looking at the results in hand in table 6.3 and in figure 6.5, it infers that WTM satisfies the optimization constraints in equation 3.9 and is a better way to achieve enhanced performance from proposed ensemble model.

The idea to test the performance of the proposed model against all other weighting methods is to advocate the robustness of WTM . Moreover, the robustness check of WTM was also done looking at the predicted score distribution on a sample of the customer from the dataset mentioned in chapter 4 . Looking at the distribution of predicted default score in figure 6.6, 6.7, 6.8 and 6.9, it is clear that the score distribution obtained using WTM has a better discriminating status of defaulted or not defaulted customer in respect to other well-known methods of score distribution from parametric, non-parametric and an ensemble model.

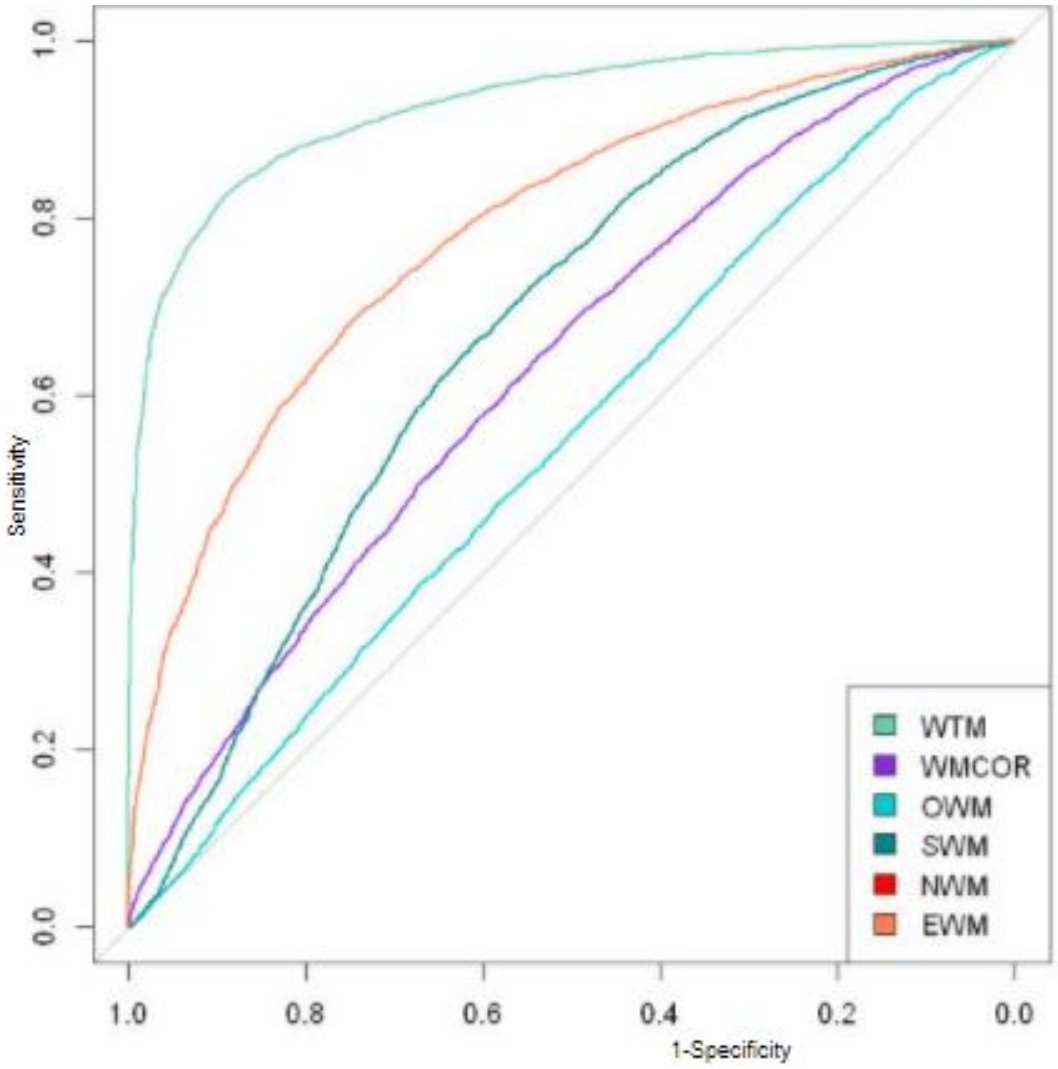

Figure 6.5: ROC curve with different weighting strategy.

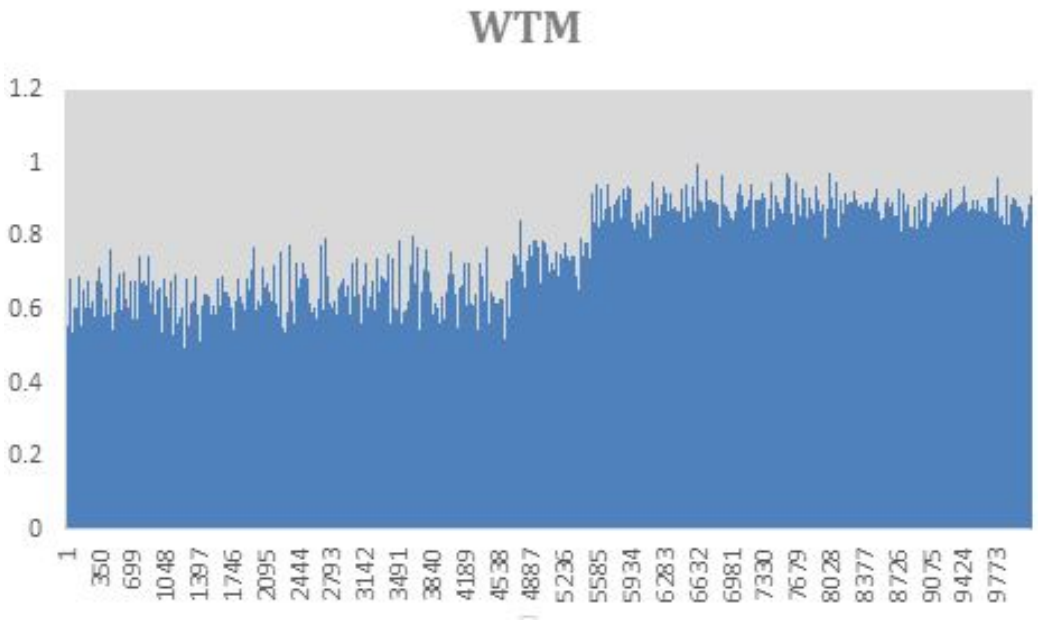

Figure 6.6: Predicted score distribution of customers using proposed model WTM.

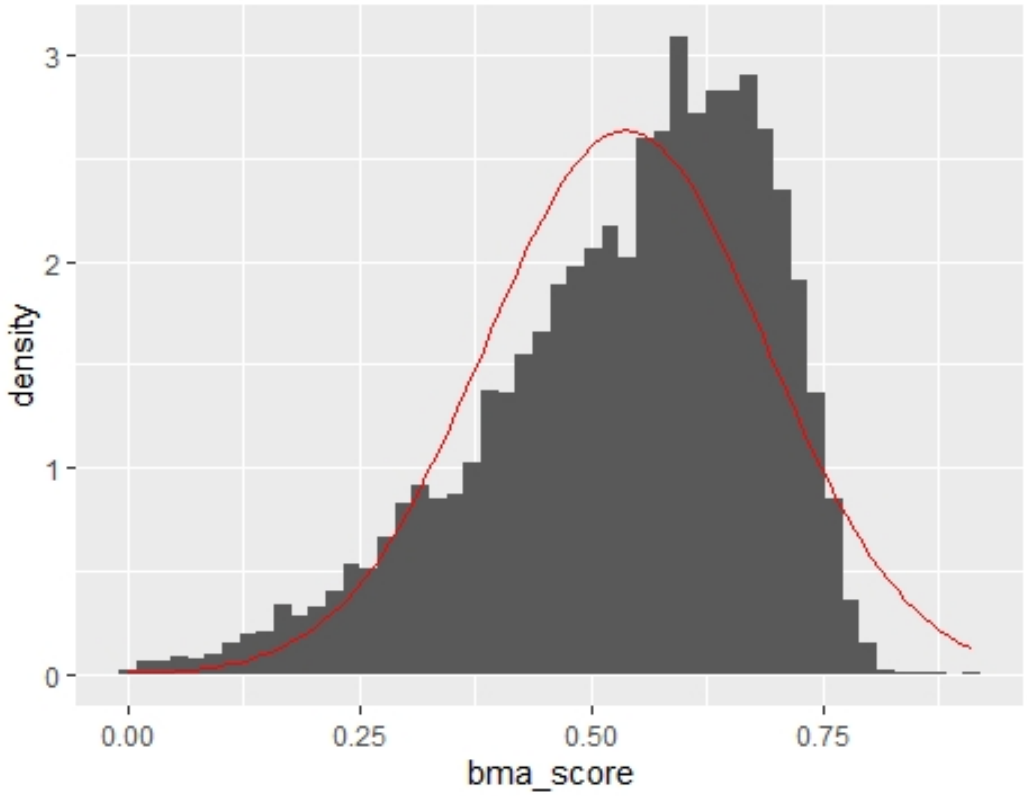

Figure 6.7: Predicted score distribution of customers using one of the well-known ensemble model.

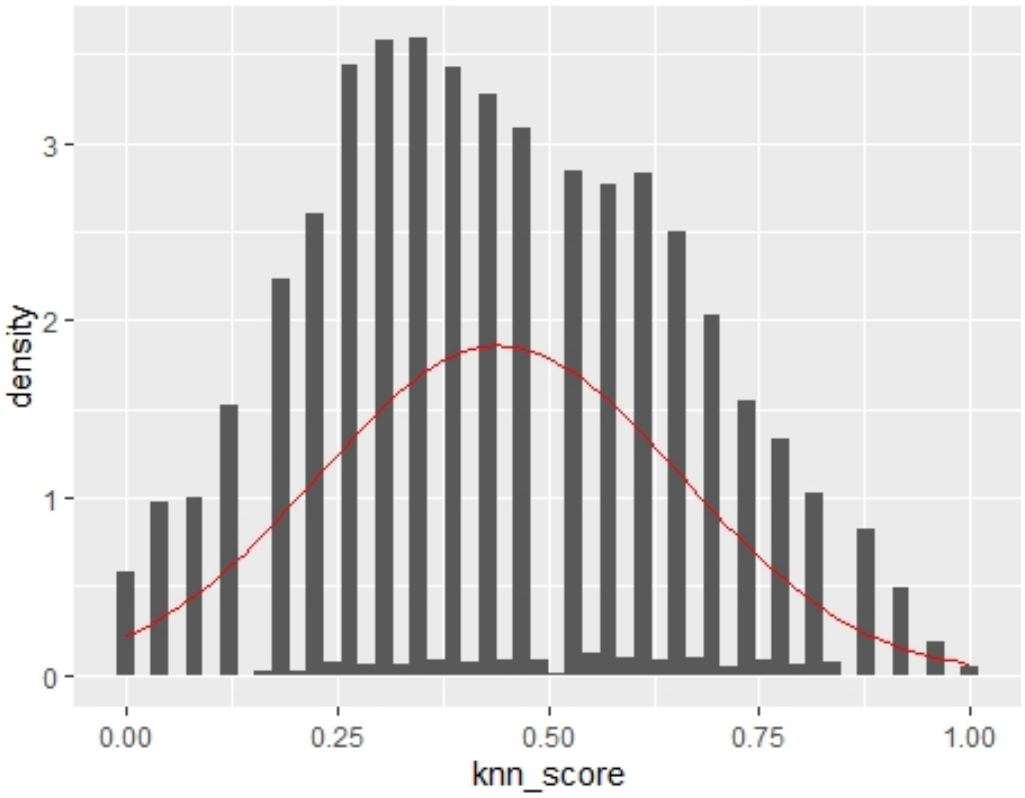

Figure 6.8: Predicted score distribution of customers using one of the well-known non-parametric model.

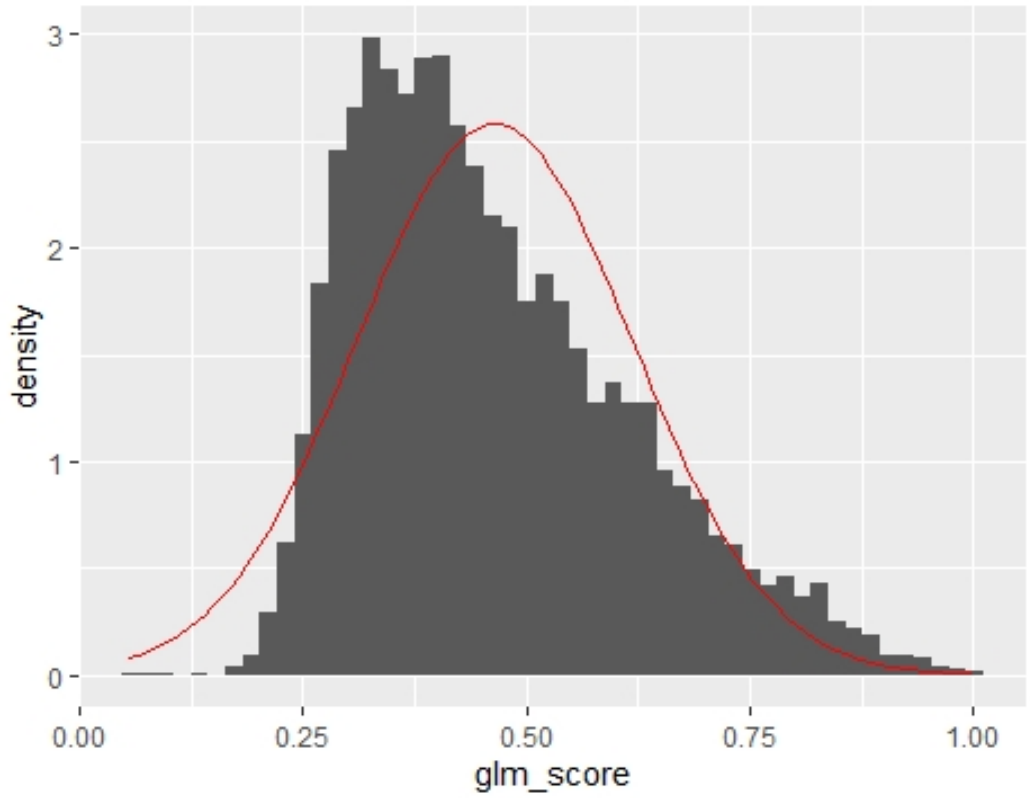

Figure 6.9: Predicted score distribution of customers using one of the well-known parametric model.

6.2 Multi-objective optimization strategies results

In this section, we discuss results that were obtained from the proposed idea of a multi-objective optimization strategy in chapter 3 . The solution obtained from the proposed idea is used as a weight to construct an ensemble model and compare performance with a well-known machine learning model. The performance evaluation is a robust indicator to see how well our proposed model is doing against other considered machine learning models.

Solving each of these strategies using a minimization framework provides a set of non-dominated solutions that are not ordered but sufficiently serves as local and global optimal values for the considered objective functions. They are Pareto efficient solutions which is used to rank the performance of machine learning models (parametric, non-parametric, and ensemble). This, in turn, helps us in mapping the relation between our objective functions which further can be changed sequentially by varying weights especially in the strategic approach of a weighted sum of deviation and Chebyshev goal programming.

The ordering analysis of the Pareto optimal solution is not considered in this scope of study in terms of *no preference method, a priori method, posterior method, hybrid method, and interactive method* since these are broadly defined topics meeting the different purpose of solving a multi-objective problem. Our approach is, to some extent, very similar to the no-preference method where we have been able to scalarize the problem taking the objectives that are normalized into a uniform dimensionless scale. For detailed insights on the ordering of Pareto solutions, refer to the study of Branke, et. al [14] .

Each of the proposed strategies is tested against parametric models (GLM, NB), non-parametric models (*decision tree*, GAM) and ensemble model average (RF, BAGG, BOOST, BMA, BART) .

A similar approach is adopted for all strategies when comparing with some key performance metrics such as H, AUC, AUCH, MER and MWL that helps to examine predictive capability, discriminatory power, and stability of the results.

Figure 6.10 presents Pareto front as an unordered point with two local minimum optimal solutions and a set of various points as a globally optimal solution with respect to strategy 1 that is based on a weighted sum of deviations. The solution achieved through this strategy is useful for the direct comparison of objectives as we know that unnecessary deviations are multiplied with weights to form a single sum for the goal or achievement function. How to set up weights or select weights is one of the active topics for research in the context of goal programming.

The idea behind each of these proposed strategies is explained in section 3.6. The solutions achieved from solving the defined multi-objective strategies in this thesis is a set of unordered Pareto solutions.

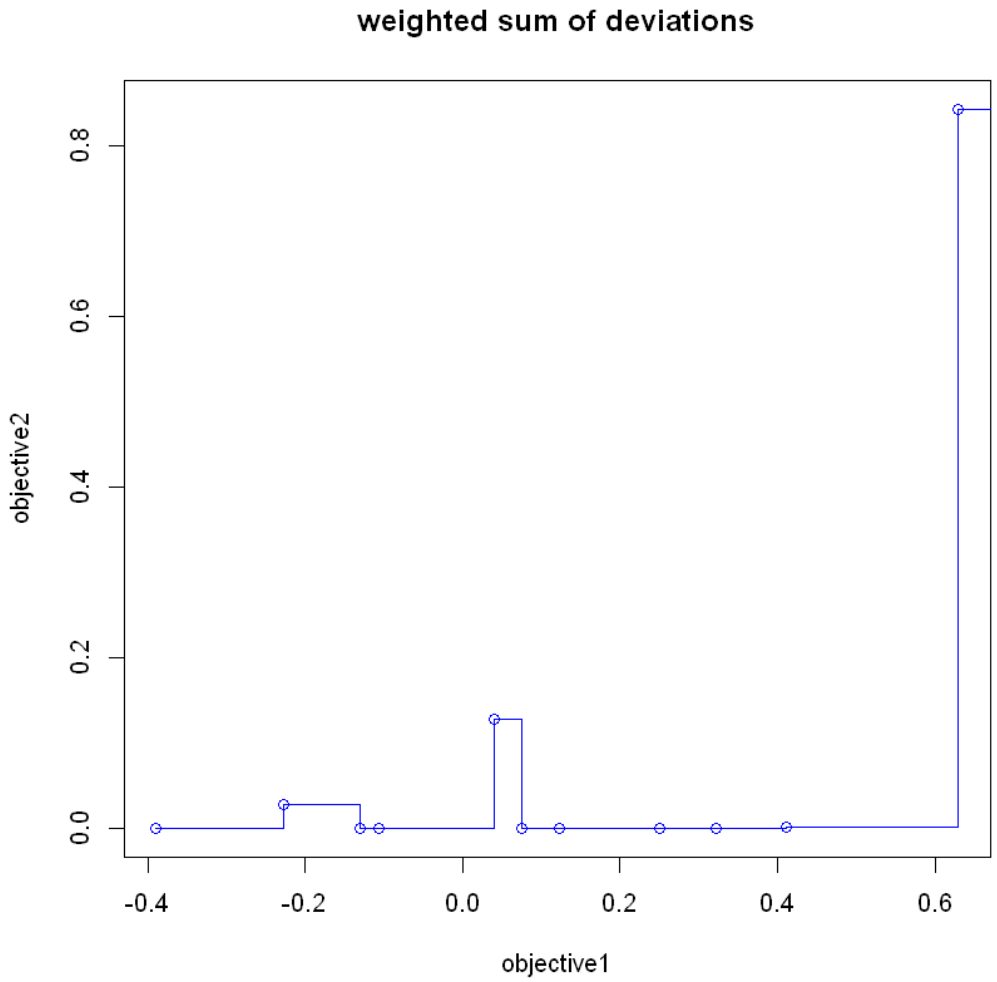

Figure 6.10: Multi-objective strategy based on weighted sum of deviations 3.6.

Figure 6.11 presents the Pareto front with respect to strategy 2 where we can see a peak at some point in their objective function value and being flattened at many other points with the intuition to minimize maximum deviation in the goal programming approach i.e to reduce maximum covariance among the chosen machine learning models. These peak points are the local minimum optimal solution and flattened points are the global minimum optimal solution.

Figure 6.12 indicates the Pareto front unordered solution of minimization problem referring to strategy 3 that is based on joint entropy and we can see multiple local

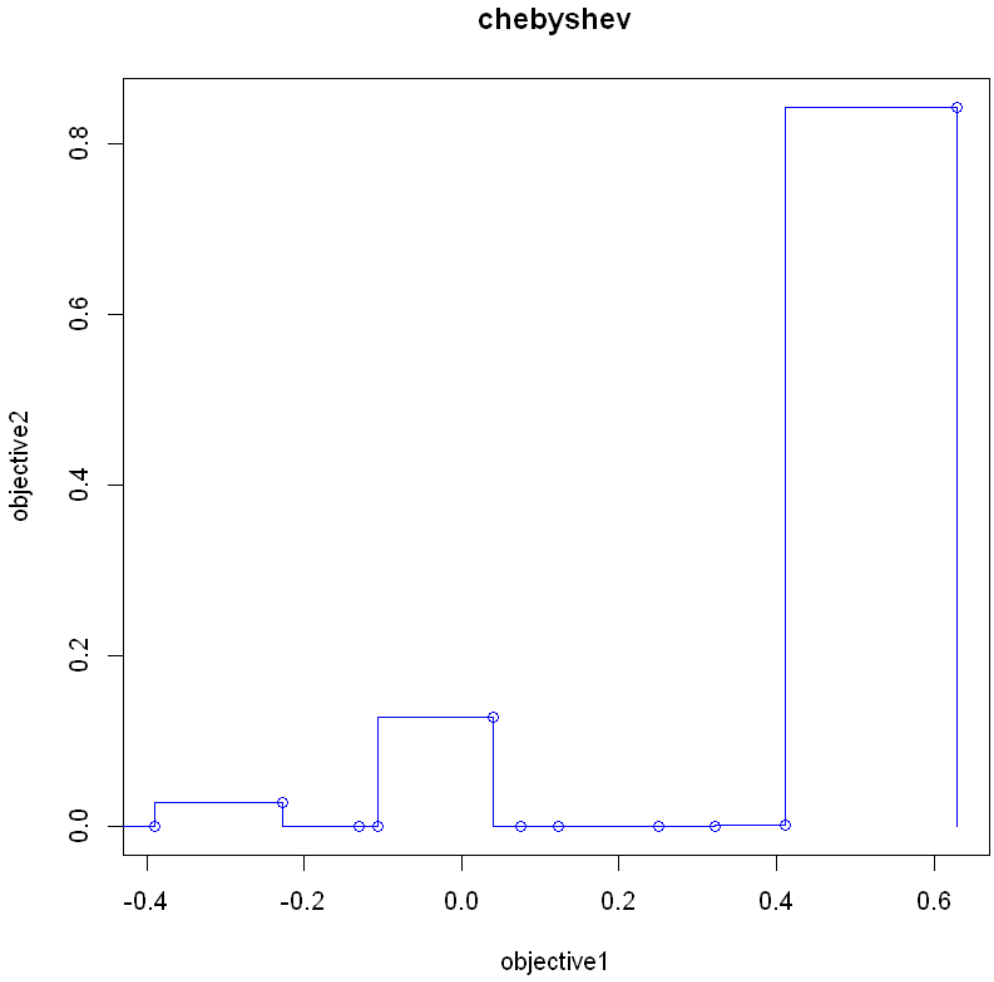

Figure 6.11: Multi-objective strategy based on chebyshev goal programming approach 3.6.

and global optimal points.

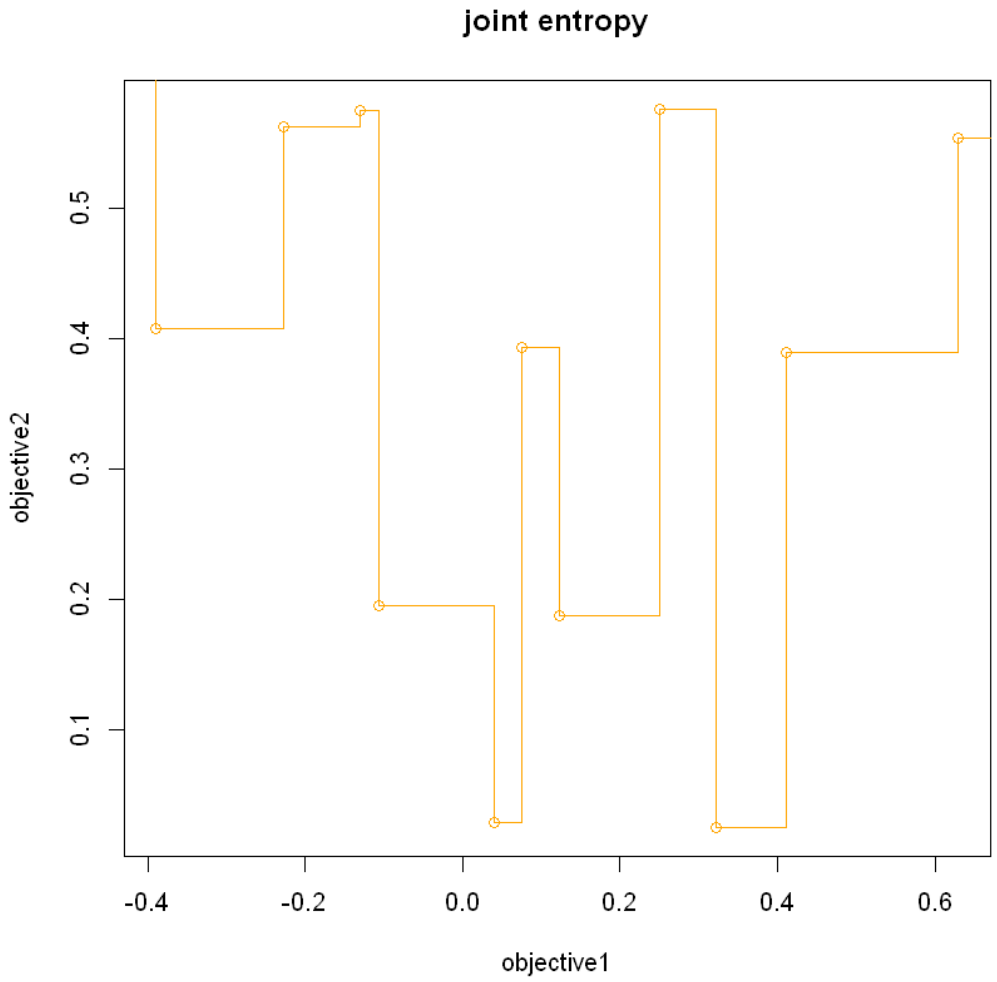

Figure 6.12: Multi-objective strategy based on joint entropy approach 3.6.

Figure 4 indicates the unordered solution of cross-entropy referring to strategy 4 that shows many local optimal solutions and one global optimal solution.

The solutions achieved through different strategies serve as an optimal weight to construct an ensemble model using the model average technique proposed in chapter 3 . They are assessed against a set of performance metrics to evaluate the possibilities to enhance the model accuracy.

Referring to the accuracy-related performance metrics in table 6.5, strategy 3 3.6 is the best performing strategy including performance overlap for AUC and AUCH metrics whereas strategy 4 3.6 is the worst performing strategy. On the other hand, for error related measures in table 6.6, strategy 3 3.6 is the obvious choice among all

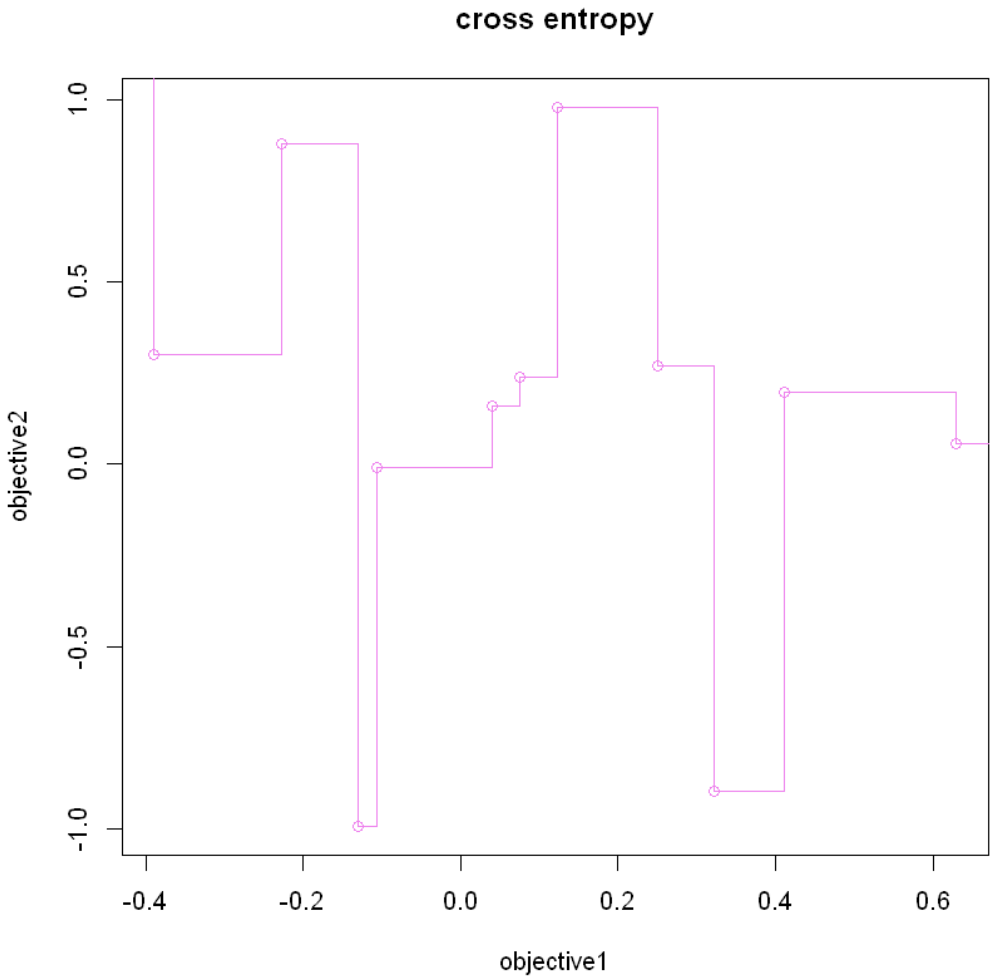

Figure 6.13: Multi-objective strategy based on cross entropy approach 3.6.

other strategies.

Table 6.5: Performance of different strategy with respect to accuracy metrics.

Strategies	H	AUC	AUCH
Strategy 1	0.60	0.92	0.92
Strategy 2	0.48	0.93	0.93
Strategy 3	**0.64**	**0.94**	**0.94**
Strategy 4	**0.06**	**0.63**	**0.63**

Table 6.6: Performance of different strategy with respect to error of the model.

Strategies	MER	MWL
Strategy 1	0.14	0.14
Strategy 2	0.12	0.12
Strategy 3	**0.11**	**0.11**
Strategy 4	**0.28**	**0.28**

6.3 Additional Results

This section bring additional insights into the results specifically from the risk analysis perspective as it helps to understand casual relationship and probabilistic inference among given set of variables in the dataset.

6.3.1 Additional results using Bayesian Network

The *Bayesian Network* is a technique that explores independence relations among data variables and distribution of data under the probability framework. The learning from Bayesian Network for modeling or predictive task is done through many different approaches but we focus mainly on the so-called constraint-based and score-based approach. To know details on such topic, refer to [105] and [75] . The constraint-based approach helps to understand the resulting network through in-dependencies of data and the score-based approach describes the data in Bayesian Network through probability distribution. The constraint-based approach uses in-built statistical tests that help the network to find a set of nodes or links that are not independent. On the other hand, the score-based approach simply works on the principle of general optimization technique where the objective is to maximize the score of node in the network. In short, the Bayesian Network is a model of in-dependencies and parametrization of the joint distribution. Each of the nodes in the network due to the score-based approach has probability distribution like multinomial distribution and conditional Gaussian distribution. These probability distributions are nothing but posterior probabilities that can be averaged out for all in degree bounded network to achieve better predictive accuracy through the Bayesian Network modeling. For more details on such topic, refer to Friedman et. al [56] .

Our approach proposed in chapter 3 on the model averaging technique can be applied in this context of the Bayesian Network using a score-based approach and posterior probability distribution. This probability distribution can be used as

a weight to construct an ensemble model to achieve enhanced predictive capacity. Figure 6.14 is a Bayesian Network that reflects the connection of the constraint-based approach and score-based approach (also known as Hill Climbing Algorithm). Both the algorithms are search algorithms for structural learning but the score-based approach is the more preferred choice since a large amount of data is required for a fully connected network using the constraint-based approach.

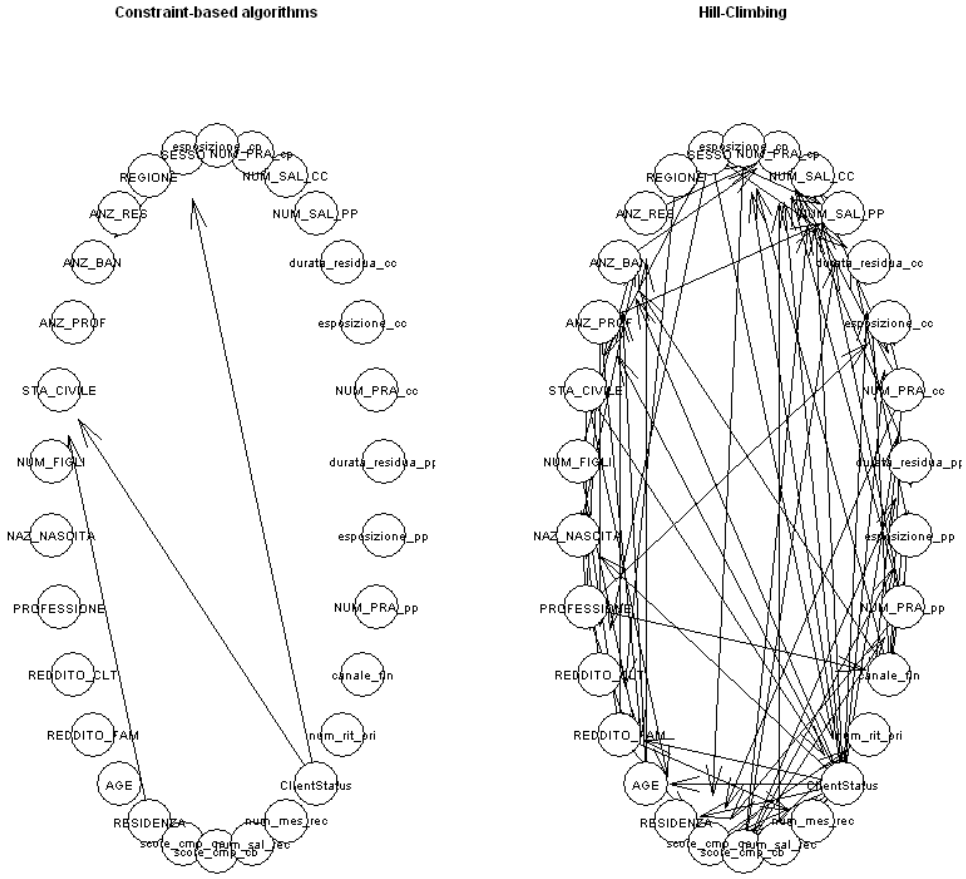

Figure 6.14: Bayesian Network using score-based and constraint-based algorithm.

The connection of each node in the network is conditionally dependent on other nodes through a joint or conditional probability distribution that gives an intuitive understanding of the causal relationship of the event and their cause.

For instance, the figure in 6.15 shows the distribution of the response variable "ClientStatus" from the dataset discussed in chapter 4 that explains the causality of

the default status of the client in the whole network considering the co-variates are randomly placed.

The graph explains the distribution of each label (denoted as a light blue bar) of the response variable with respect to a set of features that could be placed at random in the network. This is our approach as an experiment to add randomization in the network in terms of probability distribution and each number simply denotes the random positioning of nodes considered as covariates in the model. Such an approach could be useful to understand the risk dynamics of a person or any occurrence of an event from a network analysis perspective or as probabilistic inference.

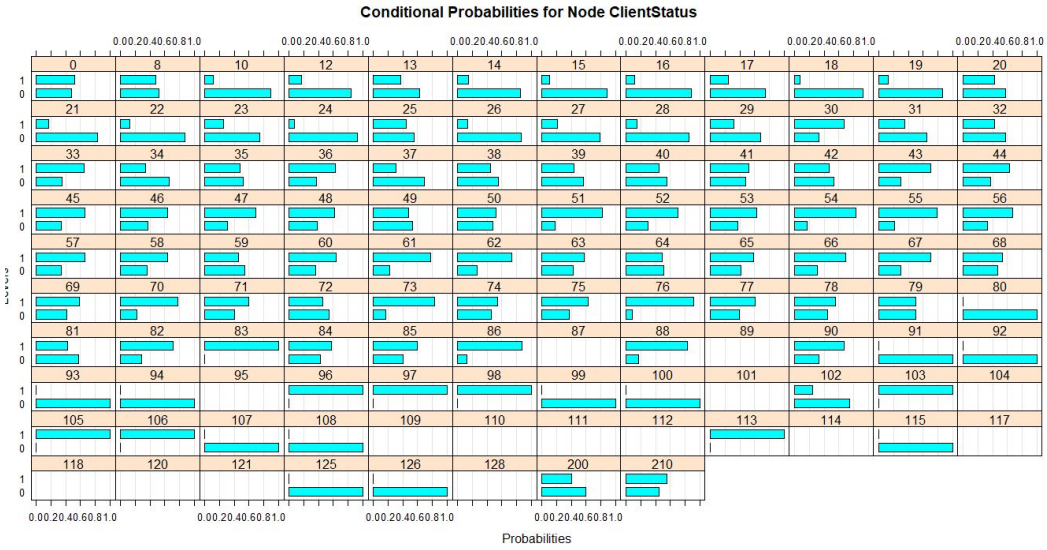

Figure 6.15: Conditional probability distribution of response variable.

6.3.2 Additional result using threshold criteria

Apart from the performance assessment metrics discussed in section 5.3, we evaluated the model based on different criteria of thresholds for comparing models as a measure of accuracy. This idea is quite helpful in understanding the ex-post analysis after the occurrence of events.

Before presenting this assessment in table 6.7, it is important to understand in brief the intuition behind threshold settings. Each of these thresholds can take values between 0 and 1. The singular value for threshold can be considered as optimal value and dual values can be considered as a range of values that might be equal for the very particular threshold selection method. The idea of using such performance metrics is based on different threshold criteria and is described in brief as follows:

- Minimum Occurence Prediction (MOP) is the value that refers to the minimum prediction for the correct predicted labels.
- Mean Occurence Prediction (MEP) is the value that refers to the mean prediction for the correct predicted labels.
- 10 Percent Omission (10PO) is the value or range that leaves 10 percent of the correct predicted labels.
- Sensitivity Equal to Specificity (SS) refers to the value at which sensitivity is equal to specificity.
- Maximum Sensitivity and Specificity (MSS) refers to the value which maximizes sensitivity and specificity.
- Maximum Kappa (MK) refers to the value that reflects the maximum kappa statistic.
- Maximum Proportion Correct (MPC) refers to the value that reflects the maximum proportion of correct and incorrect predicted labels.
- Minimum ROC Plot Distance (MRPD) refers to the threshold value where the ROC curve is a perfect fit.

They are important analysis if we want to understand the post-processing of any events that occurred, as in this case, it helps to understand the risk distribution after the default has finally occurred from a set of customers over time. This is a step to understand risk dynamics ex-post quasi distribution or analysis. Using thresholds

reported in the table 6.7, it helps to compare models and track any changes in distribution over time.

The robustness of our proposed weighted ensemble model WTM is reflected in the ex-post distribution analysis in table 6.7 where a few numbers of threshold criteria values supports that WTM is better at predicting class labels with compare to well-known parametric, non-parametric, and ensemble models.

Table 6.7: Performance assessment of the model based on different threshold value.

Threshold	CTREE	RPART	GLM	RF	BAGG	BOOST	BMA	GAM	KNN	NB	BART	WTM
MOP	**0**	**0**	0.05	**0**	**0**	0.01	0.01	0.02	**0**	**0.38**	0.05	0.05
MEP	0.46	0.46	0.46	0.23	**0.21**	0.26	0.58	0.58	0.44	0.32	0.45	**0.60**
10PO	0, 0.04	**0,0.31**	0.29	**0.01**	0,0.04	0.04	**0.42**	**0.42**	**0.17**	0.03	**0.01**	0,0.35
SS	0.46	0.44,0.6	0.43	0.54	0.53,0.56	0.50	**0.56**	**0.56**	0.42,0.43	**0.2**	0.5	0.50
MSS	**0.12**	0.32,0.35	0.74	**1**	0.97,1	0,1	0.55	0.56	0.46,0.47	0.39	0.64	0.88
MK	**0.12**	0.32,0.35	0.74	**1**	0.97,1	0,1	0.55	0.56	0.46,0.47	0.39	0.64	**1**
MPC	0,0.04	0,0.31	0.13,0.18	**0,0.01**	0.46	0.47	0,0.04	0.46	0.47	0,0.04	**0**	**0.60**
MRPD	0.39,0.40	0.61,0.65	0.43	0.97	0.97,1	0.91	0.56	0.56	0.44	0.21	**0.5**	**0.78**

The model outcome in most of the tasks is a probability distribution, and these distributions differ or diverge from each other. There are various methods to estimate the divergence of such distributions but we restrict our insights in the context of Kullback-Leibler divergence here to calculate relative entropy and cross-entropy between two probability distributions. These measure helps to understand the uncertainty among model and also can be used as a loss function for optimizing classification models.

we know that by adopting different modeling approaches, we predicted default probabilities from the given true or observed distribution. The idea carried in table 6.8 helps in understanding how similar or different is the predicted distribution from the observed distribution. The lesser is the values in the table, the better is the model in explaining the distribution closer to the observed distribution. Looking at the values in table 6.8, WTM stands to be a better model in explaining the difference of predicted and observed distribution in comparison to well-known parametric, non-parametric, and ensemble models. For more details on such topic, refer to [100] .

Table 6.8: Divergence assesssment using cross-entropy.

CTREE	RPART	GLM	RF	BAGG	BOOST	BMA	GAM	KNN	NB	BART	WTM
Cross-Entropy	**12.63**	12.90	13.22	12.86	12.75	**12.96**	**13.23**	**13.23**	13.11	12.66	12.78

Chapter 7

Conclusions

In conclusion, our proposed idea on model averaging and Pareto-based multi-objective optimization is one of the useful techniques that have the potential to enhance the performance of machine learning or any statistical model. The proposed idea can be applied to many different problems where data analysis represents the core task, and it is not only limited to classification tasks.

The conclusion drawn from the proposed idea is based on the single dataset and any generalization of this idea could be specific to the problem of interest at hand which should be checked on several further case studies.

The model averaging approach is primarily useful in reducing prediction errors but not necessarily may do so in every context. The reason for this is due to the fact few individual models among the pool of models do not contribute to the decrease of co-variance and average bias. This can be offset using a proper or diverse technique for estimating weights that in turn helps in adjusting the additional variance from weaker models.

The literature is full of different information criteria that advocate the right way of estimating weights. In our opinion, however, none of the information criteria is ideal to be applicable to every single problem. Therefore, a continuous discussion on evolving the theories and techniques of information criteria will be an important step in this direction.

The traditional approach suggests using the single best model and therefore ignores model uncertainty that may arise due to model structure and assumptions. Therefore, relying on the single best model with confidence is not a good idea as it may have adverse consequences. The committee of diverse models offers enhanced performance if it is based on model average techniques.

Model averaging studies are dominated by two approaches, that are the Bayesian

and the frequentist approach. Any different approach like the one proposed in this thesis is an attempt to offer a technique that is effective to solve the diverse problems of classification. Our proposed model averaging technique can be considered as a cutting tool that does not take parameter values for averaging. In this sense, we make the approach flexible to work on many different problems.

There is contradicting opinion if the model averaging technique is any ensemble technique unlike boosting and bagging. Such belief is mostly because model averaging is not straightforward from the computational point of view and lacks generalization abilities that can solve different problems.

However, our work in this thesis strongly supports the argument that model averaging technique outperforms bagging and boosting in many situations especially if there is model uncertainty, model bias, high variance, and if the dataset is imbalanced. Further to emphasize our proposed idea, we can say that it is similar to the ensemble technique and offers various possibilities to enhance the performance of any machine learning model.

The main idea of any ensemble technique is to weigh individual classifier and combine in a way to produce output that is better than individual classifier at predicting the task. Our proposed ensemble technique is characterized by diverse classifiers which makes any ensemble technique efficient to enhance predictive performance. The diversity of classifiers offers a serious advantage in developing an effective model averaging or ensemble technique but its inter-relationship with predictive output and errors will be an important point of investigation from a future perspective. Making an effort to keep understanding of the ensemble model simple to non-technical people would be also a wise step in this direction.

Moreover, until today, model averaging studies favor non-parametric methods for correctly estimating predictive errors, any reliable analytical method in this respect is lacking to compute frequentist confidence intervals(P-values) on averaged model predictions.

Parametric methods based on AIC and BIC may give better performance. However, this is not always true as non-parametric methods have an advantage under general considerations. Parametric methods improve predictive error if any fixed or estimated weights are used.

A major part of applied machine learning is to understand the tricks and tips around the model selection. Given a large choice of models for selection, how one model statistically differs from other models is a question of continuous investigation and testing.

The field of machine learning is evolving rapidly with its inter-connection to optimization theories and multi-objective optimization. Optimization plays a crucial role in minimizing or maximizing the different objective functions of interest that influence the performance of the learning algorithm.

The proposed idea in this thesis with respect to multi-objective optimization is evolutionary in the sense that it tries to find Pareto-optimal solutions. Researching about developing wide options of solvers would be the progressive step in the evolutionary computation of multi-objective optimization from a future point of view. Finding concrete application of multi-objective to the problems in other domains and fields would be an important development in this direction. For instance, the knowledge of multi-objective optimization to solve and understand complex systems would pave a new area of research.

One of the drawbacks of multi-objective optimization problems is that it requires a larger computational effort and often it is solved with a larger number of iterations. Reducing the computational effort and iterations would be an important development in this direction.

All the state of the art algorithms on multi-objective optimization have advantage and disadvantage. For instance, while achieving Pareto-optimal solutions, it is difficult to measure convergence and regular spacing of solutions. Explaining the computational complexity of multi-objective optimization and generalizing them as approximate solutions would be novel development from a future perspectives in this active field of research.

Incorporating the Pareto-based approach to the machine learning problem provides a new perspective to enhance the objectives of the machine learning model although this topic is discussed with their recent developments in their relevance to a limited area of research problems. What we tried in this thesis is to give a new perspective in connecting the use of the Pareto-based approach to machine learning algorithms proposing different strategies borrowing insightful knowledge from an interdisciplinary field. The approach developed here can add a new perspective to understand as to how to generate interpretable models, retrieve new insight for model selection, and model uncertainty.

In this thesis, we adopted a different strategies to assess Pareto-optimal solutions and relate with different performance metrics in order to make a comparison across a diversified pool of models. However, this comes at the cost of some advantages and disadvantages. For instance, it is difficult to guarantee and measure convergence to achieve regular spacing of solutions largely due to the dominance and diverse nature

of Pareto-based approaches.

The research activity concentrated in the area of multi-objective optimization is an active field with many challenging problems remain open in the context of uncertainty handling, computational complexity, and robustness. For instance, one such intriguing question is the influences of learning behavior or simply a property of learning curve due to the Pareto-based approach to machine learning.

Bibliography

[1] Hirotogu Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1973.

[2] Stefania Albanesi and Domonkos F. Vamossy. Predicting Consumer Default: A Deep Learning Approach. Working Papers 2019-056, Human Capital and Economic Opportunity Working Group, September 2019.

[3] James Scott Armstrong. Combining forecasts: The end of the beginning or the beginning of the end? *International journal of forecasting*, 5:585–588, 1989.

[4] Nicole Augustin, Willi Sauerbrei, and Martin Schumacher. The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modelling*, 5(2):95–118, 2005.

[5] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *The Journal of the Operational Research Society*, 54(6):627–635, 2003.

[6] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer. A comparison of decision tree ensemble creation techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):173–180, 2007.

[7] Maria Maddalena Barbieri and James O. Berger. Optimal predictive model selection. *Ann. Statist.*, 32(3):870–897, 06 2004.

[8] Flavio Barboza, Herbert Kimura, and Edward I. Altman. Machine learning models and bankruptcy prediction. *Expert Syst. Appl.*, 83:405–417, 2017.

[9] J. M. Bates and C. W. J. Granger. The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468, 1969.

[10] M. J. Bayarri, J. O. Berger, A. Forte, and G. García-Donato. Criteria for bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577, 2012.

[11] James O. Berger, Luis R. Pericchi, J. K. Ghosh, Tapas Samanta, Fulvio De Santis, J. O. Berger, and L. R. Pericchi. Objective bayesian methods for model selection: Introduction and comparison. *Lecture Notes-Monograph Series*, 38:135–207, 2001.

[12] Dimitris Bertsimas and John Tsitsiklis. Simulated annealing. *Statist. Sci.*, 8(1):10–15, 02 1993.

[13] Marcin Blazejowski and Jacek Kwiatkowski. Bayesian Averaging of Classical Estimates (BACE) for gretl. gretl working papers 6, Universita' Politecnica delle Marche (I), Dipartimento di Scienze Economiche e Sociali, August 2018.

[14] Jrgen Branke, Kalyanmoy Deb, Kaisa Miettinen, and Roman Slowinski. Multiobjective optimization, interactive and evolutionary approaches [outcome of dagstuhl seminars]. In *Multiobjective Optimization*, 2008.

[15] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone. Classification and regression trees. In *Mathematics and Statistics*, 1984.

[16] Leo Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, August 1996.

[17] Leo Breiman. Prediction games and arcing algorithms. *Neural Comput.*, 11(7):1493–1517, October 1999.

[18] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[19] Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199 – 231, 2001.

[20] Gavin Brown and Jeremy Wyatt. Negative correlation learning and the ambiguity family of ensemble methods. In *Proceedings of the 4th International Conference on Multiple Classifier Systems*, MCS-03, pages 266–275, Berlin, Heidelberg, 2003. Springer-Verlag.

[21] Kenneth P. Burnham and David R. Anderson. Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods & Research*, 33(2):261–304, 2004.

[22] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML-04, page 18, New York, NY, USA, 2004. Association for Computing Machinery.

[23] M. G. Castillo Tapia and C. A. Coello Coello. Applications of multi-objective evolutionary algorithms in economics and finance: A survey. In *2007 IEEE Congress on Evolutionary Computation*, pages 532–539, 2007.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, Jun 2002.

[25] Nitesh V. Chawla, Lawrence O. Hall, Kevin W. Bowyer, and W. Philip Kegelmeyer. Learning ensembles from bites: A scalable and accurate approach. *J. Mach. Learn. Res.*, 5:421–451, December 2004.

[26] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bart: Bayesian additive regression trees. *Ann. Appl. Stat.*, 4(1):266–298, 03 2010.

[27] Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging*. Number 9780521852258 in Cambridge Books. Cambridge University Press, March 2008.

[28] Gerda Claeskens, Jan Magnus, Andrey Vasnev, and Wendun Wang. The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762, 2016.

[29] Bertrand clarke. Comparing bayes model averaging and stacking when model approximation error cannot be ignored. *Journal of machine learning research*, 4:683–712, 2003.

[30] Bertrand Clarke. Comparing bayes model averaging and stacking when model approximation error cannot be ignored. *J. Mach. Learn. Res.*, 4(null):683–712, December 2003.

[31] Merlise A. Clyde, Joyee Ghosh, and Michael L. Littman. Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20(1):80–101, 2011.

[32] Carlos A. Coello Coello, Gary B. Lamont, and David A. Van Veldhuizen. *Evolutionary Algorithms for Solving Multi-Objective Problems (Genetic and Evolutionary Computation)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[33] Shahar Cohen, Lior Rokach, and Oded Maimon. Decision-tree instance-space decomposition with grouped gain-ratio. *Information Sciences*, 177(17):3592 – 3612, 2007.

[34] Peter Congdon. *Bayesian Statistical Modelling*. Wiley series in probability and statistics, 2 edition, 2007.

[35] Giorgio Corani and Alessio Benavoli. A bayesian approach for comparing cross-validated algorithms on multiple data sets. *Machine Learning*, 100:285–304, 2015.

[36] T. Cover. Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:50–55, 1968.

[37] B. V. Dasarathy and B. V. Sheela. A composite classifier system design: Concepts and methodology. *Proceedings of the IEEE*, 67(5):708–713, 1979.

[38] Olivier L De Weck. Multiobjective optimization: History and promise. In *Invited Keynote Paper, GL2-2, The Third China-Japan-Korea Joint Symposium on Optimization of Structural and Mechanical Systems, Kanazawa, Japan*, volume 2, page 34, 2004.

[39] Kalyanmoy Deb. Multi-objective optimisation using evolutionary algorithms: An introduction. In *Multi-objective Evolutionary Optimisation for Product Design and Manufacturing*, 2011.

[40] M. Dellnitz, O. Schütze, and T. Hestermeyer. Covering pareto sets by multilevel subdivision techniques. *J. Optim. Theory Appl.*, 124(1):113–136, January 2005.

[41] Rodolphe Desbordes, Gary Koop, and Vincent Vicard. One size does not fit all... panel data: Bayesian model averaging and data poolability. *Economic Modelling*, 75:364 – 376, 2018.

[42] Evgenia Dimitriadou, Andreas Weingessel, and Kurt Hornik. *A Cluster Ensembles Framework*, pages 528–534. IOS Press, NLD, 2003.

[43] Dr. Janet Zhao Dinesh Bacham. Machine learning: Challenges, lessons, and opportunities in credit risk modeling. 2017.

[44] Pedro M. Domingos. Bayesian averaging of classifiers and the overfitting problem. In *ICML*, 2000.

[45] Carsten F. Dormann, Justin M. Calabrese, Gurutzeta Guillera-Arroita, Eleni Matechou, Volker Bahn, Kamil Barto?, Colin M. Beale, Simone Ciuti, Jane Elith, Katharina Gerstner, Jérôme Guelat, Petr Keil, José J. Lahoz-Monfort, Laura J. Pollock, Björn Reineking, David R. Roberts, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Simon N. Wood, Rafael O. Wüest, and Florian Hartig. Model averaging in ecology: a review of bayesian, information-theoretic and tactical approaches for predictive inference. *Ecological Monographs*, 88(4):485–504, May 2018.

[46] Saso Dzeroski and Bernard Zenko. Is combining classifiers with stacking better than selecting the best one? *Mach. Learn.*, 54(3):255–273, March 2004.

[47] Matthias Ehrgott. *Multicriteria Optimization.* Springer-Verlag, Berlin, Heidelberg, 2005.

[48] Jana Eklund and Sune Karlsson. Forecast combination and model averaging using predictive measures. *Econometric Reviews*, 26(2-4):329–363, 2007.

[49] Michael T. Emmerich and André H. Deutz. A tutorial on multiobjective optimization: Fundamentals and evolutionary methods. *Natural Computing: An International Journal*, 17(3):585Ű609, September 2018.

[50] D. Fantazzini and S. Figini. Random survival forests models for sme credit risk measurement. *Methodology and Computing in Applied Probability*, 11:29–45, 2009.

[51] Carmen Fernandez, Eduardo Ley, and Mark Steel. Model uncertainty in cross-country growth regressions. Econometrics, University Library of Munich, Germany, 2001.

[52] R. Fletcher. *Practical Methods of Optimization; (2nd Ed.).* Wiley-Interscience, USA, 1987.

[53] Tiago M. Fragoso, Wesley Bertoli, and Francisco Louzada. Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, 86(1):1–28, 2018.

[54] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Proc. 13th International Conference on Machine Learning*, pages 148–146. Morgan Kaufmann, 1996.

[55] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, 29(5):1189–1232, 10 2001.

[56] N. Friedman and D. Koller. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine Learning*, 50:95–125, 2004.

[57] Nicolás García-Pedrajas, César García-Osorio, and Colin Fyfe. Nonlinear boosting projections for ensemble construction. *J. Mach. Learn. Res.*, 8:1–33, May 2007.

[58] Paul H. Garthwaite and Emmanuel Mubwandarikwa. Selection of weights for weighted model averaging. *Australian & New Zealand Journal of Statistics*, 52(4):363–382, 2010.

[59] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *STATISTICS AND COMPUTING*, 24(6):997–1016, 2014.

[60] John Geweke and Gianni Amisano. Optimal prediction pools. *Journal of Econometrics*, 164(1):130–141, 2011.

[61] Chiara Gigliarano, Silvia Figini, and Pietro Muliere. Making classifier performance comparisons when roc curves intersect. *Computational Statistics and Data Analysis*, 77(C):300–312, 2014.

[62] Clive Granger and Yongil Jeon. Thick modeling. *Economic Modelling*, 21(2):323–343, 2004.

[63] Clive W. J. Granger and Ramu Ramanathan. Improved methods of combining forecasts. *Journal of Forecasting*, 3(2):197–204, 1984.

[64] PETER J. GREEN. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 12 1995.

[65] D. Hand. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, 77:103–123, 2009.

[66] David J. Hand. Assessing the performance of classification methods. *International Statistical Review / Revue Internationale de Statistique*, 80(3):400–414, 2012.

[67] E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):190–195, 1979.

[68] Bruce Hansen. Least squares model averaging. *Econometrica*, 75(4):1175–1189, 2007.

[69] Bruce Hansen and Jeffrey Racine. Jackknife model averaging. *Journal of Econometrics*, 167(1):38–46, 2012.

[70] L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

[71] Florian Hartig, Justin M. Calabrese, Björn Reineking, Thorsten Wiegand, and Andreas Huth. Statistical inference for stochastic simulation models - theory and application. *Ecology Letters*, 14(8):816–827, 2011.

[72] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statist. Sci.*, 1(3):297–310, 08 1986.

[73] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Number 2 in 2197-568X. Springer, New York, NY, 2009.

[74] Severin Hauenstein, Simon N. Wood, and Carsten F. Dormann. Computing aic for black-box models using generalized degrees of freedom: A comparison with cross-validation. *Communications in Statistics - Simulation and Computation*, 47(5):1382–1396, 2018.

[75] David Heckerman. A tutorial on learning with bayesian networks, 2020.

[76] Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49:409–435, 1952.

[77] Nils Lid Hjort and Gerda Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899, 2003.

[78] David C. Hoaglin. John w. tukey and data analysis. *Statistical Science*, 18(3):311–318, 2003.

[79] Lennart Hoogerheide, Richard Kleijn, Francesco Ravazzolo, Herman K. van Dijk, and Marno Verbeek. Forecast accuracy and economic gains from Bayesian model averaging using time varying weight. Working Paper 2009/10, Norges Bank, June 2009.

[80] S. Huband, P. Hingston, Luigi Barone, and Lyndon While. A review of multi-objective test problems and a scalable test problem toolkit. *IEEE TRANSACTIONS OF EVOLUTIONARY COMPUTATION*, 10(5):477–506, 2006.

[81] Ravi Jagannathan and Tongshu Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4):1651–1683, 2003.

[82] Christine Johnson and Neill Bowler. On the Reliability and Calibration of Ensemble Forecasts. *Monthly Weather Review*, 137(5):1717–1720, 05 2009.

[83] A. Joseph. Working paper no . 674 machine learning at central banks chiranjit chakraborty and. 2017.

[84] C Kamath and E Cantu-Paz. Creating ensembles of decision trees through sampling. *University of North Texas Libraries, UNT Digital Library*, 7 2001.

[85] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[86] David R. Anderson Kenneth P. Burnham. *Model Selection and Multimodel Inference*. Springer-Verlag New York, 2 edition, 2002.

[87] Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking and Finance*, 34(11):2767–2787, 2010.

[88] Reto Knutti, Reinhard Furrer, Claudia Tebaldi, Jan Cermak, and Gerald A. Meehl. Challenges in Combining Projections from Multiple Climate Models. *Journal of Climate*, 23(10):2739–2758, 05 2010.

[89] Abdullah Konak, David W. Coit, and Alice E. Smith. Multi-objective optimization using genetic algorithms: A tutorial. *Reliability Engineering and System Safety*, 91(9):992 – 1007, 2006. Special Issue - Genetic Algorithms and Reliability.

[90] Jochen Kruppa, Alexandra Schwarz, Gerhard Arminger, and Andreas Ziegler. Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, 40(13):5125–5131, 2013.

[91] Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.*, 51(2):181–207, May 2003.

[92] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124 – 136, 2015.

[93] Huan Liu, Amit Mandvikar, and Jigar Mody. An empirical study of building compact ensembles. In Qing Li, Guoren Wang, and Ling Feng, editors, *Advances in Web-Age Information Management: 5th International Conference, WAIM 2004, Dalian, China, July 15-17, 2004*, volume 3129 of *Lecture Notes in Computer Science*, pages 622–627. Springer, 2004.

[94] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113 – 141, 2013.

[95] Laurent Mascarilla and Carl Frélicot. Reject strategies driven combination of pattern classifiers. *Pattern Anal. Appl.*, 5(2):234–243, 2002.

[96] Prem Melville and Raymond J. Mooney. Constructing diverse classifier ensembles using artificial training examples. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI, pages 505–510, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[97] Stefano Merler, Bruno Caprile, and Cesare Furlanello. Parallelizing adaboost by weights dynamics. *Computational Statistics and Data Analysis*, 51(5):2487 – 2498, 2007.

[98] Enrique Moral Benito. Model averaging in economics: An overview. *Journal of Economic Surveys*, 29(1):46–75, 2015.

[99] Pietro Muliere and Marco Scarsini. A note on stochastic dominance and inequality measures. *Journal of Economic Theory*, 49(2):314–323, 1989.

[100] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[101] J. Nelder and R. Mead. A simplex method for function minimization. *Comput. J.*, 7:308–313, 1965.

[102] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[103] P. Newbold and C. W. J. Granger. Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society. Series A (General)*, 137(2):131–165, 1974.

[104] Nikunj C. Oza and Kagan Tumer. Input decimation ensembles: Decorrelation through dimensionality reduction. In *Proceedings of the Second International Workshop on Multiple Classifier Systems*, MCS-01, pages 238–247, Berlin, Heidelberg, 2001. Springer-Verlag.

[105] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[106] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006.

[107] Foster Provost and Tom Fawcett. Robust classification for imprecise environments. *Mach. Learn.*, 42(3):203–231, March 2001.

[108] I. Rish. An empirical study of the naive bayes classifier. Technical report, University of Montreal, 2001.

[109] Jonathan Rougier. Ensemble Averaging and Mean Squared Error. *Journal of Climate*, 29(24):8865–8870, 11 2016.

[110] Cynthia Rudin, Ingrid Daubechies, and Robert E. Schapire. The dynamics of adaboost: Cyclic behavior and convergence of margins. *J. Mach. Learn. Res.*, 5:1557–1595, December 2004.

[111] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

[112] Robert E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5(2):197–227, July 1990.

[113] Michael Schomaker, Alan T.K. Wan, and Christian Heumann. Frequentist Model Averaging with missing observations. *Computational Statistics & Data Analysis*, 54(12):3336–3347, December 2010.

[114] James scott armstrong. Principles of forecasting. *international series in operation research and management science*, 30:XII,850, 2001.

[115] Alexander K. Seewald. Towards a theoretical framework for ensemble classification. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI-03, pages 1443–1444, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[116] Xiaotong Shen and Hsin-Cheng Huang. Optimal model assessment, selection, and combination. *Journal of the American Statistical Association*, 101(474):554–568, 2006.

[117] Mark F. J. Steel. Model Averaging and its Use in Economics. MPRA Paper 81568, University Library of Munich, Germany, September 2017.

[118] Stephen M. Stigler. Studies in the history of probability and statistics. xxxii: Laplace, fisher and the discovery of the concept of sufficiency. *Biometrika*, 60(3):439–445, 1973.

[119] M. Stone. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47, 1977.

[120] sumio watanabe. A widely applicable bayesian information criterion. *Journal of machine learning research*, 14:867–897, 2013.

[121] James Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

[122] Allan Timmermann. Forecast combinations. In G. Elliott, C. Granger, and A. Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, chapter 04, pages 135–196. Elsevier, 1 edition, 2006.

[123] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael P.H Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187–202, 2009.

[124] Vlahavas I Tsoumakas G, Partalas I. A taxonomy and short review of ensemble selection, 2008. workshop on supervised and unsupervised ensemble methods and their applications.

[125] Theo Stewart Valerie Belton. *multiple criteria decision analysis.* Springer US, 1 edition, 2002.

[126] Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, 11:3571–3594, December 2010.

[127] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241 – 259, 1992.

[128] D.M. Wolpert and M. Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, 11(7):1317 – 1329, 1998.

[129] Danny Yuan. Applications of machine learning : consumer credit risk analysis. 2015.

[130] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, December 1997.