AMERICAN UNIVERSITY OF BEIRUT

ONE-CLASS CLASSIFICATION FOR CREDIT SCORING

by
RIM FAWAZ DAYEH

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Engineering Management
to the Department of Industrial Engineering and Management
of the Faculty of Maroun Semaan Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon
September, 2020

AMERICAN UNIVERSITY OF BEIRUT

ONE-CLASS CLASSIFICATION FOR CREDIT SCORING

by
RIM FAWAZ DAYEH

Approved by:
Dr. Jimmy Azar, Assistant Professor
Department of Industrial Engineering and Management

J.C.Azar
Advisor

Dr. Jimmy Azar, Assistant Professor
Department of Industrial Engineering and Management
[Idem]

J.C.Azar
Member of Committee

Dr. Maher Nouiehed, Assistant Professor
Department of Industrial Engineering and Management
[Idem]

Maher
Member of Committee

Dr. Saif Al-Qaisi, Assistant Professor
Department of Industrial Engineering and Management
[Idem]

Member of Committee

Date of thesis defense: [September 15, 2020]

AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

Student Name:	Dayeh	Rim	Fawaz
	Last	First	Middle

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes:

□ As of the date of submission

□ One year from the date of submission of my thesis.

□ Two years from the date of submission of my thesis.

☒ Three years from the date of submission of my thesis.

Rim September 17, 2020

Signature Date

(This form is signed & dated when submitting the thesis to the University Libraries ScholarWorks)

ACKNOWLEDGEMENTS

Throughout the writing of my thesis, I have received a great deal of support and assistance.

I would first like to thank my thesis advisor, Dr. Jimmy Azar, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would also like to thank my thesis committee Doctors, Dr. Maher Nouiehed and Dr. Saif Al-Qaisi, for their valuable feedback following my thesis proposal. You provided me with the guidance that I needed to successfully complete my thesis.

In addition, I would like to thank my parents for their wise counsel and sympathetic ear. You are always there for me. Finally, I could not have completed my work without the support of my friends, Azzam Shaarani and Fadwa Dannawi, who provided stimulating discussions as well as happy distractions to rest my mind outside of my research.

ABSTRACT
OF THE THESIS OF

Rim Fawaz Dayeh for Master of Engineering Management
Major: Engineering Management

Title: One-Class Classification for Credit Scoring

Credit scoring nowadays has become a crucial tool for the bank industry since its purpose is to assess the risk associated with granting clients loans. Accurate prediction of credit risk according to historical or current financial behaviors can be challenging, partly due to imbalanced data. The number of defaulters is relatively small compared to non-defaulters. Nonetheless, these cases are very important since they can cause huge losses if not predicted correctly. In this study, a comparison between four one-class classifiers and their combinations will be tackled, and several combining strategies will be explored. The one-class classification techniques isolation forest (IForest), support vector machine (SVM), Gaussian mixture model (GMM) and Parzen classifier (Parzen) and their hybrid models were applied on three different credit scoring datasets from uci machine learning repository; the Taiwanese dataset (30000 samples), the German dataset (1000 samples) and the Australian dataset (690 samples). The Australian dataset performed the best between the three datasets, especially when combining GMM and IForest classifiers, which gave an AUC result of 0.852 (sensitivity = 60.8%, specificity = 93.2%). Hybrid models enhanced the performance of one-class classifiers especially when including the strongest predictive classifier for the specific dataset. The study highlights the importance of hybrid models and their effect on improving the classification performances and reflects some interesting findings compared to past literature.

TABLE OF CONTENTS

ILLUSTRATIONS

Figure

TABLES

Table

ABBREVIATIONS

ANN	Artificial Neural Networks
AUC	Area under the Curve
CLC	Clustering-launched classification
FP	False Positives
FN	False Negatives
Gauss	Gaussian
GMM	Gaussian Mixture Model
IForest	Isolation Forest
K-NN	K- Nearest Neighbor
LOGR	Logistic Regression
MLP	Multi-layer perceptron
NN	Neural Networks
OCC	One Class Classification
OR	Operations Research
PCC	Percentage Correctly Criterion
PD	Probability of Default
SVM	Support Vector Machine
TP	True Positives
TN	True Negatives

CHAPTER I

INTRODUCTION

The chaos threatening the global financial markets since August 2007 had originated from the wrong decisions made by the global banking industry especially in the consumer-lending sector (Baily, Litan, & Johnson, 2008). One major financial risk comes from the difficulty to assess the creditworthiness of borrowers and predict their repayment behavior. In the field of finance and banking, this issue earned an increasing attention over time and forged the way to creating new techniques to achieve credit risk assessment. It is nowadays crucial to assess the complex behavior of borrowers, which is known to be influenced by traits, trust and time (Longo, Dondio, & Barrett, 2010). Customers' behavior reflects their ability to repay their bank loans. In this context, credit scoring has generally become the new model that will help lenders in making decisions regarding granting customers loans or not, the amount granted, and the strategies to be implemented in order to assure the profitability of the borrowers to lenders. The main advantages of performing credit scoring are many, by which is evaluating credit risk, enhancing cash flows, minimizing risks and enabling credit decision almost instantaneously (Thomas et al, 2002; Abrahams and Zhang, 2008; Marques, Garcia and Sanchez, 2013). Credit scoring is considered also a helpful tool in common applications such as operations research and data mining (Baesens et al, 2009). In practice, the procedure of credit scoring goes by gathering data about the customers, analyzing it, and classifying many credit variables to transform the problem into an automated predictive problem where a new customer will be labeled into one of two classes. He or she will either be labeled as a ‘low risk’ applicant (it is the most common case, in which the customer will not default on paying their future loans) or a ‘high risk’ applicant meaning a defaulter (Kennedy, Namee, & Delany, 2013). Making the decision mainly depends on the likelihood of them repaying their future loans based on a number of observed variables or attributes related to the sample. The input to the model includes the economic conditions, and sociological and financial data about the borrower, and the prediction techniques will provide the output in terms of the customer creditworthiness (Marques, Garcia and Sanchez, 2013). Thus, data mining has proved to be an efficient tool to solve the prediction problem and give a reliable probability of default (PD) of applicants.

A. Research Question

Many data mining approaches exist in the literature to solve the default problem and accurately label the customers as defaulters or non-defaulters according to their historical or current financial behaviors and accounts transactions. However, this problem proved to be challenging with the years especially when a credit scoring model constitutes a highly uncertain problem with imbalanced data. In fact, this issue is due to the reason that the number of defaulters in a certain population is very low or negligent compared to the number of non-defaulters. Nevertheless, this number is very important to be accurately predicted, since it causes huge losses if not predicted correctly. Following this matter, many classification models were derived. One possible way to address the credit-scoring problem is the application of the one class classification (OCC) technique also known as anomaly detection (Kennedy, Namee, & Delany, 2013). It is a new technique that is gaining much attention in the data science field (Chawla et al., 2004), especially because it identifies a single class for the expected or most probable behavior and allows the distinction of unusual behavior. In this thesis, the performance of one-class classifiers will be explored. In addition, the one-class support vector machine (SVM) method and the other classification methods will be extended to give a probabilistic output and their results will be combined with each other to test whether these constructed hybrid models could improve the results obtained by the study. Therefore, the main research question of this study is:

RQ) Can combining one-class classifiers yield an improvement in the classification results for the credit-scoring problem? And what is the optimal way to combine them?

The datasets used will be treated as benchmarks to assess the methods applied in the study, and compare them to other methods used in the literature.

B. Research Objectives

The goal of this research is to show that using a hybrid model by combining the results of two or more one-class classifiers together will yield an improvement in the classification results for the credit-scoring problem. The research also focuses on finding out the optimal combination that yields the best improvement. Another goal is that the datasets used will be treated as benchmarks to assess the methods applied in the study, and compare them to other methods used in the literature.

C. Challenges

The limitations in this kind of problem always exist in form of imbalanced data, which yields to a bias in classifying a new object. In fact, imbalanced data exists when there is a class that is well represented and is called the majority of the data, while there exist fewer or rare data points that belong to another class. This makes the prediction of classes even harder and increases the classification error especially for the minority data points. This issue exists deeply in the credit-scoring problem, where the defaulters constitute the minority class. To solve this problem, many techniques were applied and will be described later in the literature. New classification methods or old ones performed in a hybrid way will prove to enhance the performance of the classifiers and reduce the classification error.

D. Document Outline

This research will be composed of five chapters and a brief overview of its content is presented below.

Literature review: which is divided into many parts. The first part will be an introduction to machine learning algorithms. Moreover, supervised machine learning will be defined and distinguished from unsupervised technique implying the fact that the study will be based on a supervised technique. The second part is informative and describes several concepts such as the concepts of one class classification technique, and the concept of two-class classification. The classification techniques will then be described and there will be an introduction on how efficient it is to combine classifiers, which will be tackled more in the specific case of the study where the performance of one-class classifiers will be combined.

The second part will be the benchmark for the study i.e. The results of the study performed will be compared to this part and conclusions will be then deducted. However, this part will particularly include comparisons of the classification techniques applied for each of the two data sets used in the study in many articles to show what techniques performed best based on changes performed on the traditional OCC techniques and by how much results improved. This part will help summarize the previous research to use it as a comparative tool in later stages.

Experiment design: describes the data sets used, and sheds light on the software and tools used for performing data analysis. Then, this chapter will include a brief description about training and testing the classifiers on the data. Chapter 3 will also tackle the evaluation metrics that will assess the results obtained from the study. Thus, the performance of the techniques will be compared with each other based on certain criteria, elaborated later in this section.

Analysis and Findings: will describe the results obtained after implementation of the classifiers mentioned in the experiment. This chapter will contain the actual observations and results obtained from running the classifiers and the new technique suggested in the study. This chapter will include evaluating the techniques based on the performance metrics. Additionally, a ranking of the results will be made.

Comparative section: is dedicated to demonstrate and evaluate the study's results. This chapter will be a tool to address the main research question. In fact, the findings achieved by the machine learning classifiers will be compared against the findings achieved in the literature in order to see what insight has the study offered.

Conclusion: will provide a conclusion of this research study and summarize the results obtained by the new classifying technique and its impact on improving results. Additionally, it will tackle possible improvements and give insights for future work.

CHAPTER II

LITERATURE

Banking industries faced a series of issues especially because of the lending sector. It is always uncertain to determine whether a customer will be paying back their loans if they have a history of defaulting or even if they will default for the first time on paying their duties. That is when the concept of credit risk assessment became a crucial need, and this process was based on collecting specific information about customers; it is mostly economic and financial information such as historical transactions and payments, or sociological information such as their marital status, age, etc. In the next step of the process the concept of credit scoring was introduced. It is an endeavor which can be formally defined as a statistical technique to predict the probability that a new borrower or an already - existing borrower will default on paying their payments or become delinquent (Mester, 1997) i.e. assess the credit worthiness of a customer. Before credit scores were created in the 1950's, the prediction process was done based on a social approach to determine the creditworthiness. In fact, in 1958, the engineer Bill Fair and the mathematician Earl Isaac finished and sold their first standardized credit scoring system (Tavin,2020).

Credit scoring's main goal is to aid the lenders quantify their risk in providing credit and helps them take critical lending decisions quickly and more objectively. The credit scoring was a common application in the research studies of OR and machine learning. These areas of research constitute a tool that helps in drawing many conclusions and shaping company's strategies especially after the analysis of the immense data provided by credit scoring domain. Since this field is an essential procedure in the lending process, many statistical techniques were previously invented and tackled to enhance the performance of application scoring.

This chapter provides a literature review of the different concepts and techniques that help answer the research problem.

Initially, the machine learning concept will be introduced and definitions of the concept of one class classification, and the concept of two-class classification will be made. Additionally, its algorithms will be described in details to clarify all the technical terminology and methods used in the study later. A new technique will be tackled in the literature, which is the trial of combining many classifiers together and see its impact on the accuracy of prediction. The last part of the literature will be handling the comparison of the classification techniques performed on three public datasets that will be used later in the study, and the benefit of applying this comparison is to summarize what was applied in past literature and set it as a benchmark to compare the results of the study.

A. Machine Learning

Machine learning is a concept in data analytics in which analytical methods were invented to automate model building. It helps decision makers learn from data, identify patterns and base strategies that will enhance companies' performances without human intervention.

In the 1950s, the world has witnessed the invention of the first data mining approaches for credit scoring (Khashei and Mirahmadi, 2015). With the years, people invented and worked on enhancing many existing ones (Lando, 2004; Thomas et al., 2002). These methods can be typically divided into two categories: the first category is the soft models. In details, soft classifiers explicitly predict the class conditional probabilities and classify the data based on estimated probabilities. The second category is the hard models. Hard classifying targets on the classification decision boundary without passing through the step of estimating the probability (Khashei and Mirahmadi, 2015).

In supervised learning, the data selected and trained is known as a labeled data. Examples of algorithms that use supervised learning are regression, decision trees, random forest, K-NN, logistic Regression (Lu, Hargreaves and Bilal, 2018). In the unsupervised learning technique, the data is unlabeled. This learning technique is mainly used for clustering purposes. Example of this approach is the K-means method. However, semi – supervised is a learning technique that uses unlabeled data to enhance the performance of supervised learning techniques when the labeled data are limited or costly (Zhu & Goldberg, 2009). The benefits of a labeled data is that it helps predict an output based on the previous experiences. That’s why, the technique that will be used in the study later on is a supervised technique, by which the collected credit scores data will be used from previous customers’ transactions and information in order to predict their behavior.

B. One-Class Classification (OCC)

The usual multi-class classification problem, aims to label a new object into one of several pre-defined categories. However, in certain problems, the new object might not belong to any of those labels. This need was the reason behind the invention of one class classification technique. OCC distinguish a target class, which is well characterized by a set of target objects trained and recognized from other objects. For its other class, called the non-target class, it has either negligent or very few objects that are not enough to be a statistically - representative sample of the non- target class. OCC are very important nowadays, and are applied widely in many applications such as detecting machine faults (Sarmiento et al., 2005), fraud detection (Juszczak et al., 2008), and identity verification (Hempstalk, 2009). The term OCC originated from Moya et al. (1993) and its synonyms that describe the same approach were mentioned in many following articles such as the word outlier detection (Ritter and Gallegos, 1997), novelty detection (Bishop, 1994), and concept learning (Japkowicz, 1999).

In 2001, Tax (Tax, 2001) introduced the concept of taxonomy and considered that the OCC algorithms can be reduced to three categories: density estimation methods, boundary decision methods and the reconstruction methods. All OCC techniques are defined by two unique components: z, which is the distance from the test object to the center of the target data, and a threshold ϴ, to which the distance is compared to. Basically, if the distance z is less than the threshold, the object belongs to the target class. In contrast, if the distance z is larger than the threshold value than the object will be labeled as an outlier.

Density based classification is a method well known of OCC where a probability distribution is fitted to the target class, examples of this method are Gaussian density estimation, Gaussian Mixture Model, Parzen density estimation. This method doesn’t require to classify the outliers. The selection of the data related to the target and the sample size are two factors that determine the success of this method. One issue is that it needs a very large sample data to be trained (Tax and Duin, 1999).

Boundary based classification aims to enclose the target data in a boundary that must be fitted optimally to the target class by being minimized. Examples of this method is the

Support vector machine method (SVM), where the target class is bounded by a hypersphere in a high dimensional feature space. Boundary decision estimation does not require an estimation of the class probabilities. The form of the boundary is obtained by computing the distances between the objects in the target class. This method results in a good performance with small sample data (Tax, 2001).

Reconstruction based classification recreates an input pattern by assuming a model of the data generation. The components of this method are assumed during the learning phase and help to classify the object to either the target or the outlier class.

C. Overview of Classification techniques

The next part of the literature will describe the one-class classification techniques that will be implemented in the study. One must note that the techniques that will be used in the study are the following:

1. Isolation Forest (IForest)

According to (Breiman 2001; Murthy, 1998), an isolation forest is a classification technique, which consists of a combination between decision trees and random forests. Usually, it is defined by several Isolation Trees where IF = {t_1… t_T} (Puggini & McLoone, 2018). For every tree t, one can compute the average number of iterations needed to assure the isolation of an object x. An anomaly score is assigned according to the depth of object x in the trees.

The purpose of the isolation forest method is as its name stems for is to explicitly isolate anomalies instead of enclosing the target instances (Liu, Ting, & Zhou, 2008). An isolation tree (iTree) is known as a tree structure that can be built effectively to isolate a single object. Isolation Forest (IForest), is about building an ensemble of isolation trees (iTrees) for a given dataset, then anomalies are identified as being those samples which have short average path lengths on the isolation trees (Liu, Ting, & Zhou, 2008). IForest performs very well on small datasets because the masking effects will be minimized in this case. An advantage of IForest is that it does not rely on distance or density measures to detect anomalies. This reduces the computational costs of distance calculations.

2. *Support Vector Machines (SVM)*

Support vector machine (SVM) was invented by V. Vapnik in 1995 and is a commonly used and powerful technique that is applied to perform machine learning. SVM is practiced in many fields such as pattern recognition, bio-informatics, and credit scoring etc. (Berry & Linoff, 1997; Rüping, 2000; Luo, Cheng and Hsieh, 2009). Given the training set, the one class SVM can be constructed by following these several steps: At first, the input feature vectors must be mapped into a feature space and this is achieved by going to a higher dimensional feature space. The mapping is subject to the type of the kernel function selected in applying SVM. Next, build a hyperplane that differentiate two or more classes (Luo, Cheng and Hsieh, 2009). Thus, the division within the feature space will be optimized. The hyperplane is usually capable to deal with a high number of features. To note that the idea behind SVM is a hard-label classification which has the purpose to reach a hyperplane that can divide the two classes in order to minimize the classification error (Luo, Cheng and Hsieh, 2009). The process of SVM is well suited for a limited training data.

The mode of operation of the SVM technique is elaborated as follows. First, the linear separable case will be explained, i.e. two classes will be separated by a linear decision boundary. Given a training set $x_i \in IR^n$ with n input objects i={1…n}, and supposing that the classifier is a linear classifier following the equation:

$$g(x)= w\ x_i +b$$

This equation looks for a hyperplane when set to 0.

$$w\ x_i +b = 0$$

The SVM classifier according to Vapnik must satisfy the following conditions:

$$\begin{cases} w\ x_i + b \geq +1\ for\ y_i = +1 \\ w\ x_i\ + b \leq -1\ for\ y_i = -1 \end{cases}$$

Which is equivalent to $y_i\ (x_i\ w + b)\ -1 \geq 0 \qquad \forall i = 1…n$

The equations of the conditions mean that if $w\ x_i + b \geq +1$, the object belongs to the class labeled by +1 and if $w\ x_i + b \leq -1$, the object belongs to the class labeled by -1. The hyperplane will be setting the boundary between both classes.

Distance to the separating hyperplane: d = $\frac{|w^T x+w0|}{||w||}$

The maximum margin in the feature space with the margin width between both hyperplanes is equal to $\frac{2}{||w||}$ (Luo, Cheng and Hsieh, 2009).

For the nonlinear SVM problems, the equations are the same but the difference is that instead of considering x_i input data, a nonlinear function $\varphi(xi)$ that is the mapped input data to a high dimensional feature space.

In primal weight space, the classifier takes the decision function as the following equation shows:

$$y(x)=\text{sign}[w\,x+ b]$$

This equation was never solved in this form. One defines the following optimization problem:

$$\text{Min I} = \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \quad \forall i = 1\ldots n$$

$$\text{Subject to } y_i\,(x_i\ w + b) \geq 1- \xi_i$$

$$\xi_i \geq 0$$

Where the positive constant C is described as the tuning (regularization parameter) in the algorithm. The slack variables ξ_i are needed in order to allow misclassification in the set of inequalities (example because of overlapping distributions).

The objective function's first part can be interpreted as trying to maximize the margin between both classes in the feature space, and the second part's job is to minimize the misclassification error (Luo, Cheng and Hsieh, 2009).

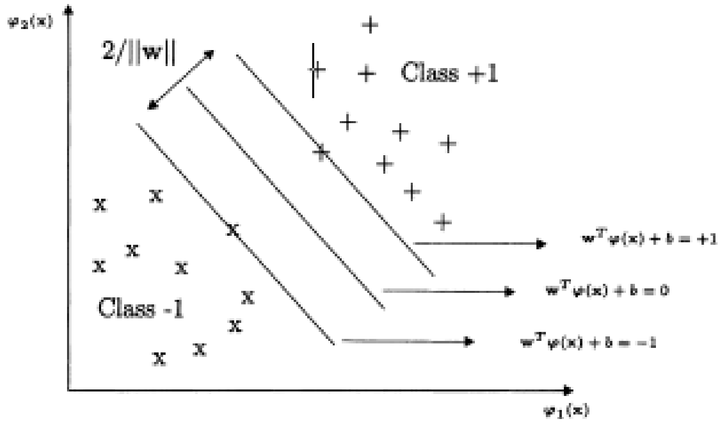


Figure 1. Illustration of SVM optimization of the margin in the feature space (Baesens et al., 2009)

Linear separation of the data is basically restricted. Going back to the nonlinear case where the previous equations are valid but with the introduction of the nonlinear function $\varphi(x_i)$ instead of $x_{i.}$ To note that this formulation is similar to LP formulation but the main difference is that it introduces a margin showed in Figure 1. to separate between the classes and permits a nonlinear decision boundary because of the use of the mapping $\varphi(.)$ (Baesens et al., 2003).

The nonlinear lagragian equation can be constructed as follows:

$$L= I - \sum_{i=1}^{N} \alpha_i \{y_i[w^T \ \varphi(x_i) + b] - 1 + \ \xi_i\} - \sum_{i=1}^{N} v_i \ \xi_i$$

Solve the minimization by deriving with respect to w, b, and ξ_i and setting the equations = 0. Thus, three equations will be obtained:

$$w = \sum_{i=1}^{N} \alpha_i y_i \varphi(x_i)$$

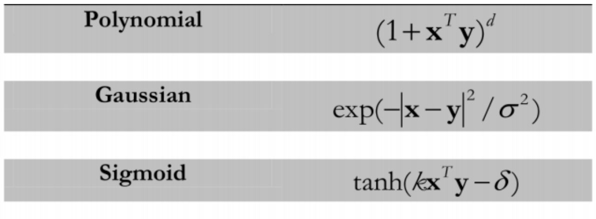
$$\sum_{i=1}^{N} \alpha_i \ y_i = 0$$

$$0 \leq \alpha_i \leq C \qquad \forall i = 1 \ldots n$$

By replacing the first equation in the primal form shown previously, the expression of y(x) becomes as follows:

$$y(x) = \text{sign}\ [\ \sum_{i=1}^{N} \alpha_i\ y_i\ k(x_i; x) + b\ ]$$

where k(x$_i$;x) = $\varphi(x_i)^T \varphi(x)$ is called the kernel function based on the kernel trick. This trick assumes that the mapping won't be calculated but will be replaced by the kernels. The table for the kernels is showed in figure 2 as follows:



Polynomial	$(1+\mathbf{x}^T\mathbf{y})^d$
Gaussian	$\exp(-\lvert\mathbf{x}-\mathbf{y}\rvert^2/\sigma^2)$
Sigmoid	$\tanh(\kappa\mathbf{x}^T\mathbf{y}-\delta)$

Figure 2. Kernel functions

At last, the Lagrange multipliers αi is obtained by maximizing the following:

$$\text{Max} - \frac{1}{2}\sum_{i,j=1}^{N} y_i\ y_j\ K\big(x_i; x_j\big)\alpha_i\alpha_j + \sum_{i=1}^{N}\alpha_i$$

$$\text{Subject to } \sum_{i=1}^{N}\alpha_i\ \ y_i = 0$$

$$0 \le \alpha_i \le C \qquad\qquad \forall i = 1\ldots n$$

The problem now simplifies to a convex quadratic programming (QP) in α_i (Baesens et al., 2003).

Two hyper parameters to be controlled while using one class SVM: the complexity of the model parameter and the threshold of the model parameter.

The key idea behind using SVM for anomaly detection, is to rely on a target class which contains the normal cases (non – defaulters) and consider that the rest are considered as anomalies (in the study, the defaulters) (Schölkopf et al., 2001, Tax and Duin, 1999). In order to generalize the SVM classifier into a novelty detection tool, Tax and Duin (Tax and Duin, 1999) suggested the idea of enclosing the normal cases in a hypersphere with center c and a minimal radius R to separate the normal instances from the outliers. The data will be mapped into a higher dimensional feature space by a kernel function.

3. Gaussian Mixture Model (GMM)

For most datasets, the assumption of convex unimodal distribution won't mostly fit the data correctly (Tax, 2001). For more flexibility, the normal distribution was extended to a Gaussian Mixture Model also known as mixture of Gaussian (Bishop, 1995; Duda and Hart, 1973) which constitutes a linear combination of k Gaussian distributions (Kennedy, Namee, & Delany, 2013). The probability density function (pdf) of the GMM is given by the following:

$$p_{GMM}(x) = \frac{1}{N}\sum_j \alpha_j \, pN(x;\, \mu_j,\, \Sigma_j)$$

where αj are the mixing coefficients (Tax, 2001). The issue with this model is that when a limited data is used, the variance turns out to be larger. The advantage is that it has a smaller bias than the single Gaussian distribution, but it needs a greater number of data to train.

In order to construct a Gaussian Mixture classifier, the trained data will be split into k clusters, and a Gaussian distribution model will represent each cluster. For every object z, the Gaussian Mixture Model will be written as:

$$f\ (z) = \sum_{i=1}^{k} \alpha_i \exp\{-(z-\mu_i)^T\ \Sigma \mathrm{i}^{-1}(z-\mu_i)\}$$

For each cluster i, α_i, μ_i, and $\sum_i$ are estimated by using the expectation maximization (EM) algorithm which facilitates maximizing the likelihood function in estimating each parameter (Fessler and Hero, 1994). Again, in order to determine whether an object belongs to the target class or not, the density value of object z will be compared to the threshold Ɵ.

4. Parzen classifier (Parzen)

Parzen density estimator (Parzen, 1962), is an extension of the Gaussian Mixture Model introduced previously. The Parzen approach is a non-parametric approach and its density estimate is usually a mixture of Gaussian kernels. In fact, each object in the target data will then represent the center of a Gaussian distribution (Kennedy, Namee, & Delany, 2013).

$$p_{p(x)} = \frac{1}{N} \sum_i pN\ (x;\ x_i\ , hI)$$

For few sample sizes, Parzen technique proves to be very sensitive to scaling the feature values of the data, since Parzen's parameter h gives an equal weight to its features. During the process of training a dataset, one must explicitly decide on the value of h. In order to check if an object is accepted to the target data or rejected, a measure of the likelihood must be made by taking the average of the probability of membership of the Gaussian distributions. The object z will be then compared to a threshold Ɵ in order to classify the data as target or non-target (Kennedy, Namee, & Delany, 2013). Given n target objects z, the probability density function of Parzen estimation is defined p(z) as follows:

$$p(z) = \frac{1}{nh} \sum_{i=1}^{n} \rho(\frac{z-zi}{h})$$

where h is a smoothing parameter, and ρ is a Gaussian kernel function:

$$\rho\ (\mathrm{z}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

h represents the width of the kernel and is computationally obtained by getting the maximum LL (Log Likelihood) (Tax, 2001); Kraaijveld & Duin, 1991).

There exist many disadvantages of the Parzen classifier. First, when big differences in density exist, the Parzen method will not be able to perform well in low density estimation areas. Additionally even though the price for training a Parzen classifier is negligible, the testing phase is costly. Like any other density techniques seen so far, this classifier also needs a large set of data to estimate the probability density function of the estimator. Analytically, this method might be a hassle since during the testing phase the distances to trained objects must also be calculated (Kennedy, Namee, & Delany, 2013). Thus, this method won't be feasible when a large amount of data in high dimensional feature space is inspected.

5. *Hybrid Model (Hybrid)*

According to Tax (Tax, 2001), multiple statistical technique have been proposed in order to solve the classification problems and reduce the classification error. Depending on the size of the data its features, what distribution or technique can be implemented for it and how well it can perform on the input data, the best classifier is typically chosen (Tax, 2001). Trying a single classifier has proven to deliver a sub optimal result in labeling the objects as targets or non-targets. The effect of opting for one method only is reflected by giving up on the results of the poorer performing classifiers however, it might deliver a valuable information (Wolpert, 1992-bookocc; Tax, 2001). Coming from the fact that a slight improvement in classification results might lead to tremendous gains to companies, the importance of enhancing the labeling process has become a need and that’s when a new method came to light: combining classifiers together which may differ in complexity and algorithm might make the difference. [Sharkey and Sharkey, 1995; Tax, 2001] has elaborated about the advantage of combining classifiers and highlighted that it doesn’t just improve the work of the classifiers but also increases their robustness. That’s what drove the researchers in literature to try to come up with specific combination rules to apply this method.

Mostly combining happens by taking the average of the posterior probabilities obtained by the classifiers. Even though this technique is not very complicated, it is proven that it delivers good results especially when the posterior probabilities are easy to compute [Hashem, 1994, Tanigushi and Tresp, 1997 – A1; Tax, 2001). One main drawback of this method is that sometimes it is hard to estimate the probability distributions for some methods for example, the K-nearest neighbor method with k =1. The posterior is not well defined in this case. A solution for that is to use the combining on the labels. For one-class classifiers the problem is different. Information is usually obtained about one class, which is the target class, and the outlier class lacks information about it. According to Tax’s thesis (Tax, 2001) in the domain of one class classification, he elaborated about combining classifiers, and the results of his experiments showed that for applying a one class classifier by itself, the best method was to apply Parzen density estimation. However, combining changed the results and it appeared that combining is done best when using a Gaussian Model (Tax, 2001).

Many researches such as Bates and Granger (Khashei and Mirahmadi, 2015) have argued that combining improves the performance of classifying. They believe that a single method is not enough to completely identify all the information needed. They highlighted that in fact other than predicting the class labels with a higher accuracy, the advantages of combining are that it minimizes the risk of applying a wrong model to a specific data, and the risk of failing in combining methods (Khashei and Mirahmadi, 2015). In recent literature, combined, also known as, hybrid classification techniques, have also been suggested and developed in order to improve accuracy results. For credit scoring application, many hybrid trials have been performed. (Lee *et al.*; Khashei and Mirahmadi, 2015) have combined back propagation neural networks (NN) with linear discriminate analysis (LDA) in credit scoring. Additionally, (Hsieh; Khashei and Mirahmadi, 2015) has proposed a new hybrid approach in creating an effective credit-scoring model, which is to integrate together some clustering algorithms and artificial neural networks (ANNs). (Luo *et al.*; Khashei and Mirahmadi, 2015) have tried to apply support vector machines technique (SVMs) with clustering-launched classification (CLC) models for scoring applications. (Li *et al.*; Khashei and Mirahmadi, 2015) came up with a linear combination between kernel functions in order to interpret credit scoring models in a better way. Other researchers such as (Chen and Li ;Khashei and Mirahmadi, 2015) have proposed a combination between SVM classifiers and LDA, decision trees, as a preparation for optimizing the feature space by eliminating duplicates, redundancy, and unrelated features. (Kim and Han ;Khashei and Mirahmadi, 2015) have proposed and implemented a hybrid technique that combines Self-Organizing Map (SOM) and case-based reasoning (CBR) in order to estimate the corporate bonds rating. (Park and Han ;Khashei and Mirahmadi, 2015)found that integrating analytic hierarchy with case-based reasoning to give weights to the features and enhance the performance of case-based reasoning (CBR) techniques especially in foreseeing failure of businesses. To avoid banking bankruptcy, (Ahn & Kim; Khashei and Mirahmadi, 2015) have combined the performance of genetic algorithm

(GA) for case-based reasoning. In order to increase the effectiveness in classification, (Akkoc ; Khashei and Mirahmadi, 2015) have suggested a three-stage hybrid neuro-fuzzy inference system (ANFIS) for applying classification for credit scoring, which is based mainly on artificial neural networks (ANNs), and fuzzy logic. (Laha; Khashei and Mirahmadi, 2015) has proposed an adaptive hybrid credit-scoring technique that is based on fuzzy rule based classifiers. Initially the rule is learned from the training set using a Self-Organizing Map (SOM) technique, and the next step would be to integrate the fuzzy K-nearest neighbor to come up with a contextual classifier that contains information from the training set. At last, (Yao; Khashei and Mirahmadi, 2015) have implemented a hybrid fuzzy support vector machine (F-SVM) for score valuation by adopting three strategies: to select the input features either (CART) classification and regression trees or the multivariate adaptive regression splines (MARS). The last strategy was to use using GA in order to optimize the model parameters.

In this paper, an adaptive hybrid classification technique of several one-class classifiers with a probabilistic output will be integrated in order to study its effect on the classification results in the domain of credit scoring. The idea behind the construction of a hybrid model in this study is to make use of the soft labels obtained from the classifiers and combining them using one of the several combining strategies; classifiers can simply be combined using the product rule, the min, the max, the median or the mean combining rules. These combining rules will provide a probabilistic output to a new tested object. Basically, after combining and compiling the classifier on the new object, the object will have a probability for being an inlier and another one for being an outlier. The higher probability will assess the label that should be assigned to the object. The hybrid rules used in the study are the mean, median and product rules for the Australian, Taiwanese and the German datasets respectively. In order to show the effectiveness and flexibility of the new proposed models, combining rules were chosen in a way to maximize the AUC results for the hybrid models with respect to the datasets used in the study. The benefit of the hybrid models in the study will be to compare their performances in the three chosen data sets with the results obtained from individual classifiers.

D. Classification techniques in the study

This study relies on using four computational models for classification. These models are the Gaussian Mixture Model (GMM), Support Vector Machine (SVM), isolation forest (IForest) and the Parzen classifier (Parzen). Most of the techniques belong to one class classification because the credit-scoring problem labels the customers as either targets or outliers. The outlier class is almost negligent compared to the target class; here lies the class imbalance in the credit-scoring problem. The focus in the study will be on the new technique applied, which is making use of the soft labels of the classifiers and being able to use their probabilistic output in order to combine them together and come up with an improved classifier.

E. Comparative benchmark for the study

In order to validate the results of a study, these results must not only be covered and discussed however, they must be benchmarked and compared to many other models results to show the accuracy and the improvement that distinguishes the technique applied from other techniques performed in other studies. Nevertheless, most of the studies compare multiple methods applied on a single data set. Comparing based on a single data set methods performance is somewhat limited by the data attributes, size, and the methods applied to it. These features are highly uncertain and subject to change from a data set to another (Zurada, Kunene, and Guan, 2014). Comparison of methods must not be based on single data sets, and a benchmark in the literature must be set in order for the researchers to validate their studies. On a micro level, every researcher must start examining the performance of several methods applied to the different characteristics of datasets.

One of the articles addressing the credit scoring classification problem is for (Kennedy, Namee, & Delany, 2013). The authors wanted to study the classification problem by comparing the performance of two class classifiers, imbalanced, oversampled and the performance of one-class classifiers on the data. The two datasets used to conduct the study were the Australian dataset and the German dataset. According to H measure and harmonic mean test results, with imbalanced data the authors were able to rank their classifiers before optimizing the threshold and at an imbalance of 99:1 (Kennedy, Namee, & Delany, 2013). The classifiers they used in their study were LOG_Norm, Gauss, k-means (10), k-NN (10), GMM, NParzen, Parzen, SVDD and AE. Among all the results presented in the article, several interesting results are worth exploring because of their relation to this study. For the OCC classifiers with a ratio of 99:1 (non-defaulters: defaulters), GMM was the 2nd best classifier for the Australian dataset with an H measure of 51.9 which is a little less than Gauss who gave the best performance in this imbalance resulting in H measure of 52.3, the highest. In contrary, GMM classifier was the 2nd worst classifier for the German dataset with an H measure of 6.5. For the German dataset, the best performance was for Parzen when it scored an H measure of 9.7. As one can notice even if Parzen performed at its best with the German dataset, it seems that all the classifiers did not deliver high predictive result in 99:1 imbalance compared to the Australian dataset results. The best H measure for the Australian dataset was 52.3 and the worst was 27.2, but for the German dataset the best H measure was 9.7 and the worst was 5.3 while using the same classifiers in both cases. This shows that the models delivered weak performances with the German dataset. After optimizing the threshold in their study and at an imbalance level 99:1, (Kennedy, Namee, & Delany, 2013) used another performance measure to assess the results of the classifiers, the harmonic mean test. The harmonic mean test showed that optimizing the threshold didn't improve the work of GMM with respect to the other classifiers improvement and GMM became ranked the 5th out of 9 classifiers for the Australian dataset while delivering a harmonic mean of 69.9, which is less than the best performing classifier the LOG_Norm that delivered a harmonic mean of 79.8. GMM performance was deteriorated to 9th for the German dataset with a harmonic mean of 55.2 (Kennedy, Namee, & Delany, 2013), which is less than the highest performing classifier Parzen having an harmonic mean of 58.8. One can still notice the fact the the harmonic mean results for the models using the German dataset are weaker than the results obtained using the Australian dataset.

Regarding Parzen classifier (Parzen), it was ranked the 5th for the Australian dataset with a low H measure of 34.4 while the best performance was for Gauss, which gave the best performance in this imbalance resulting in H measure of 52.3, the highest. For the German dataset, Parzen performed the best for the German dataset with a measure of 9.7 among the other one-class classifiers mentioned earlier (Kennedy, Namee, & Delany, 2013). The lowest performance in this imbalance was for NParzen having an H measure of 5.3. As one can notice even though Parzen was the highest performing classifier in their study, but the fact that Australian dataset H measures are very much higher than the German ones, can highlight the fact that all the classifiers of the German dataset are performing poorly on it. After optimizing the threshold, another performance metric was used to assess the performance of the classifiers on the datasets used. Optimizing the threshold led to a significant enhancement for the ranking of Parzen to the 4th out of 9 classifiers with a harmonic mean of 70.9 for the Australian dataset while the best performing classifier in this case was the LOG_Norm with a harmonic mean of 79.8. For the German dataset, Parzen ranked 1st place with a harmonic mean score of 58.8 out of 9 classifiers. One can notice that the results for the German dataset between the 9 classifiers are not very far from each other. The weakest classifier is for GMM resulting in a harmonic mean of 55.2 while Parzen the best gave a harmonic mean of 58.8. So the performance of the classifiers on the German dataset are not different from each other the first performing classifiers performance does not differ much from the worst performing classifier. Also, one must mention that the predictive power for the classifiers using the German are much lower than the Australian dataset performance, highlighting the fact that the German dataset is weaker than the Australian one using H measure or harmonic mean test as metrics to measure the performances of the classifiers using these datasets. Another study performed by (Flemotomos, 2017) shed light on the performance of Parzen classifier in credit scoring while using the German data as a base data for classifying and applying 5-fold cross validation and PCA before assessing the results. In order to assess the function of classifiers, accuracy and F-score were used. In fact, F-score can be obtained by getting the weighted average F – score, based on number of classes (in credit scoring it’s two), the size of class i and the size of the data. The author set the crucial parameter for Parzen at h=15, and obtained after optimizing his parameters a prediction of the performances of the three classifiers used in the study, SVM, k-NN, and Parzen. (Flemotomos, 2017) obtained that Parzen ranks the third with close results to other classifiers, SVM and k-NN, with an accuracy of 60.5% and F-score for the majority class(non-defaulters) of 0.7427 and 0.1505 for the minority resulting a mean F-score of 0.4466. This was while Parzen got an accuracy result of 61.5% and SVM got an accuracy of 62% with mean F-scores of 0.5029 and 0.5062 respectively. One can notice that the performances of the three classifiers don't differ much from each other using the accuracy metric or the F-score metric. These results can be considered intermediate to low performing classifiers using the German dataset. In conclusion, both articles (Kennedy, Namee, & Delany, 2013) and (Flemotomos, 2017) can give an insight for the performance of the Parzen. (Kennedy, Namee, & Delany, 2013) performed multiple evaluation technique and compared Parzen in several datasets with 8 other OCC classifiers instead of 3 other classifiers only for one dataset, so the first article allows for a generalization about Parzen more. However, the results of both articles are similar, which is the fact that Parzen classifier has an intermediate to a weak performance in classifying these datasets.

Among all the techniques presented in the article, SVM is worth exploring. (Marques, Garcia and Sanchez, 2013) made a different type of experiment then seen before. They wanted to assess the effect of oversampling, under sampling on imbalanced data, quantify the improvements and compare if LOGR is better than SVM in the study or vice versa. Their results showed that oversampling and under sampling improved the results that were assessed by AUC for low/moderate imbalance. The low/moderate imbalance results where the imbalance level was (1:4), (1:6), (1:8) gave for LOG R for the Australian dataset AUC values of 0.8, 0.79 and 0.84 respectively, which are higher than the German AUC results which were, 0.611 0.608 and 0.554 respectively. This proves that the models with the German dataset are performing weaker than with the Australian dataset. These results showed for LOGR were lower than the SVM performance especially for the Australian dataset where the values were greater than the LOG R values. For the German dataset, it was the opposite case. SVM performance on the German dataset gave AUC results of 0.51, 0.5 and 0.5 respectively, which are less than the performance of LOGR for the German dataset. As a result, for low/moderate imbalance SVM performed better than LOGR for the Australian dataset, however LOGR performed better than SVM for the case of German dataset. For a higher imbalance where the imbalance level was (1:10), (1:12), (1:14) LOG R scored for the Australian dataset AUC values of 0.660, 0.762 and 0.675 respectively, which are higher than the German AUC results which were, 0.528 0.511 and 0.535 respectively. These results showed were greater than the SVM performance especially for the Australian dataset where the values were 0.610, 0.682 and 0.568, less than the LOG R values. For the German dataset, it was the same case too. SVM performance on the German dataset gave AUC results of 0.5, 0.5 and 0.5 respectively, which are less than the performance of LOGR for the German dataset. As a result, a selection of a highly imbalanced dataset, proves that the performance of SVM deteriorates, where in this study (Marques, Garcia and Sanchez, 2013), SVM performed lower than LOGR for both German and Australian datasets. (Khashei and Mirahmadi, 2015) used the classification error metric in order to compare the performance of many classifiers together and to test the performance of their new hybrid model. The classifiers they used in their study are LDA, QDA, kNN, ANN, SVM and the new hybrid model. It turned out that their hybrid model had a classification error of 10.9%, which was the best result, and the one that has a significantly less error than what was delivered by all the other classifiers. In fact, the new hybrid model, LDA, QDA, ANN and KNN all delivered better results than SVM, the weakest performing classifier. SVM classifier had a classification error of 22.5%, the worst compared to 10.9%, the error of the hybrid model, the best performing classifier. SVM delivered the worst performance for the German data for assessing bad loans according to (Zurada, Kunene and Guan, 2014). The average correct classification showed low numbers for SVM. It seems like the German datasets does not deliver good results overall. SVM gave 47.2% accuracy for bad loans while the best classifier was neural network classifier, which delivered not a very higher result too, an accuracy of 49.7%. In contrary, for the Australian data set, (Zurada, Kunene and Guan, 2014) SVM delivered an intermediate performance for bad loans with an average correct classification accuracy rate around 80% which is higher than the German results. The best performing classifier in that case was the RBFNN, which gave an accuracy of 89%. (Flemotomos, 2017) used SVM as a traditional classifier to compare with other classifiers results. The article results proved that SVM (accuracy of 62%) delievered the best performance between the three classifiers SVM, Parzen and k-NN. Even if SVM had the best result, this does not prove that the result reflects a high predictive ability for the classifier, since 62% accuracy usually means an intermediate performance. (Luo, Cheng and Hsieh, 2009) wanted to compare the performance of clustering launched classification (CLC) and SVM on the German dataset. CLC delivered better predictive credit accuracy of 84.8% for the German data while 73.7% for the SVM classifier. As a result of all those articles, one can conclude that SVM delivers intermediate to low performances in classifying the German, Australian and Taiwan datasets.

At last, the authors (Khashei and Mirahmadi, 2015) created a hybrid model to enhance the performance of traditional classifiers and reduce classification errors. They changed the performance of artificial neural networks (ANN) multi-layer perceptron (MLP) from crisp parameters for weights and biases to fuzzy paramaters and fuzzy triangular numbers. The model is available in details in the article called “A Soft Intelligent Risk Evaluation Model for Credit Scoring Classification” (Khashei and Mirahmadi, 2015). Their hybrid model actually delivered the best results in relation to classification error, which was 10.9%, the lowest classification error between 6 classifiers. The improvement percentage from each of the other classifiers to Hybrid model were calculated and it is showed that the new model is 11.38% better than ANN, 22.14% better than LDA, 23.24% better than k-NN, 45.23% better than QDA and 51.56% better than SVM.

To note that iForest classifier had no previous record regarding the three datasets used in the thesis. So this study will include the first time where iForest is simulated on the German, Taiwanese and Australian datasets.

The following table summarizes the results of the ten articles discussed above.

Table 1. Summary table for comparative study on creditworthiness

Authors	**Title**	**Classification technique of interest applied**	**Data**	**Evaluation metrics**	**Summary results of the article**
(Kennedy, Namee, & Delany, 2013)	Using Semi-supervised Classifiers for Credit Scoring	Parzen, GMM	German Australian	H measure, Harmonic mean, Friedman’s average rank test	Before optimization: German: (low performance) Intermediate: Parzen – 1st Bad: GMM -7th Australian: Good: GMM - 2nd Intermediate: Parzen - 5th

					After optimization: German: (low performance) Good: Parzen – 1st Bad: GMM -9th Australian: Intermediate: GMM - 5th , Parzen – 4th
(Marques, Garcia and Sanchez, 2013)	On the suitability of resampling techniques for the class imbalance problem in credit scoring	SVM	German Australian	Accuracy, ROC(AUC), Friedman’s average rank test, Post hoc test	Low imbalance: German: (low performance) LOGR better than SVM Australian: SVM better than LOGR High imbalance: German: (low performance) LOGR better than SVM Australian: LOGR better than SVM
(Baesens et al., 2003)	Benchmarking state-of-the-art classification algorithms for credit scoring	SVM	German Australian	ROC(AUC), Friedman’s average rank test, PCC	German: (low performance) Intermediate: SVM Australian: Intermediate: SVM
(Khashei and Mirahmadi, 2015)	A Soft Intelligent Risk Evaluation Model for Credit Scoring Classification	SVM, hybrid	Australian	Classification error	Australian: Good: Hybrid (error = 10.9%) Bad: SVM (error = 22.5%)
(Junior et al., 2019)	An Empirical Comparison of Classification Algorithms for Imbalanced Credit Scoring Datasets	SVM	German Australian	Classification error, ROC(AUC), Friedman’s average rank test	German: (low performance) Intermediate: SVM (~51% AUC) Australian: Intermediate: SVM (~71% AUC)
(Zurada, Kunene and Guan, 2014)	The Classification Performance of Multiple Methods and Datasets: Cases from the Loan Credit Scoring Domain	SVM	German Australian	Accuracy, Classification error, Friedman’s average rank test	German: (low performance) Intermediate: SVM Australian: Intermediate: SVM
(Lu, Hargreaves and Bilal, 2018)	Prediction of Credit Card Clients Payment Status	SVM	Taiwan	Accuracy	Taiwan: Intermediate: SVM (4th place – 81.9% accuracy)

(Flemotomos, 2017)	Pattern Recognition Techniques for Credit Classification	SVM, Parzen	German	Accuracy, F-score	German: Intermediate: Parzen (60.5% accuracy), SVM (62 % accuracy)
(Liu, 2018)	Machine Learning Approaches to Predict Default of Credit Card Clients	SVM	Taiwan	Accuracy, F-score	Taiwan: Good: SVM (F-score 0.8040; higher than NN with F-score 0.4520)
(Luo, Cheng and Hsieh, 2009)	Prediction model building with clustering-launched classification and support vector machines in credit scoring	SVM	German Australian	Accuracy	German: (low performance) Intermediate: SVM (73.7% accuracy) Australian: Intermediate: SVM (80.43%)

1. Limitations

There is a need to invent and train new hybrid models. At first, one must fit a model on the training set and then assign anomaly scores to it. In fact, scores that are closer to 1 are considered as outliers. The fraction rejection, which is the fraction of training objects to reject as outliers will allow setting a threshold for this purpose. The predicted performance of the test set will also get anomaly scores assigned to it based on the output of the classifiers. These scores will be compared to the threshold and will help in obtaining the probabilistic output of the classifier.

The probabilistic outputs will help in forming the hybrid models by using many combining rules and making use of the soft labels of the classifiers. There’s the product rule, the min rule, the max rule, the median rule, and the mean rule. The product rule simply multiplies the inlier posterior probabilities and the outlier ones for every class and then the class with the maximum score will be assigned the corresponding output. The minimum rule finds the minimum score for every class between the classifiers, and assigns the class label whether inlier or outlier according to the highest posterior output, the combination has. The other combination rules follow the same concept during the formation of a hybrid model. If one chooses one of these rules for all the datasets, some hybrid models will not deliver their best results due to the combining rule used. That’s why in this study different combining rules were used in order to maximize the performance of each dataset. The hybrid models for the Australian dataset were done using the mean rule on soft labels, the hybrid models for the Taiwanese dataset were done using the median rule, and the hybrid models for the German dataset were done using the product rule.

CHAPTER III

EXPERIMENT DESIGN

A. Datasets

The data chosen for the study consists of two datasets from many financial contexts. The focus in these datasets on the social, economic and characteristics of the borrowers. One will notice that some variables in the data consist of letters and this is for confidentiality reasons. Each dataset of a certain country has a number of samples, and a set of features. All the information in the datasets will help lenders to assess the creditworthiness of the borrower and increase their predictive power of knowing if customers will default on paying their loans. It is very important to use not just one dataset, to show the effect of classifying datasets with different sizes and different attributes. The characteristics of each dataset are described in Table 3 (Zurada, Kunene and Guan, 2014). The German, Taiwanese and Australian datasets are publicly available at the UCI Machine Learning Repository with the following links:

German URL: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Taiwanese URL: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Australian URL: http://archive.ics.uci.edu/ml/datasets/statlog+(australian+credit+approval)

A description for each dataset is also available in Table 4.

Table 2. Characteristics of the datasets

Datasets	**Characteristics**			
	Size	**Features**	**Categorical**	**Numerical**
Australian	690	14	8	6
German	1000	20	12	8
Taiwan	30000	24	10	14

Table 3. Description of the datasets

Datasets	**Description**
Australian	The names of the features were not revealed and the size of the dataset is considerably small which, is 690 sample. The dataset is a combination of continuous and nominal variables and it contains missing values. It contains 14 features from which 8 are categorical and 6 are numerical.
German	The names of the attributes are shown. The dataset made of 1000 customers' data contains the economic, social, and characteristics information about the customers. It contains 8 numeric attributes, and 12 categorical attributes, and there are no missing values.
Taiwan	The names of the attributes are shown. This data differs from other datasets with its large size which is 30000 samples. The dataset also contains information about the social, family and financials of borrowers. The dataset is a combination of categorical and numerical

	variables and it contains no missing values. It is made of 24 features from which 10 are categorical and 14 are numerical.

B. Evaluation Metrics

In credit scoring applications, several evaluation metrics were used in order to assess the performance of classifiers used for labeling loan borrowers. (Baesens et al., 2003) used the percentage correctly criterion (PCC) as a criterion for measuring their classifiers performances. This performance criterion measures the proportion of correctly classified samples from the overall samples in the datasets. Then, (Baesens et al., 2003) admitted using it was wrong because it assumes the same cost for both false positives (FP) and false negatives (FN). Other techniques were opted by researchers and will be used in this study. For a one or two class classification problem, most of the metrics can be derived from the confusion matrix. It is a 2x2 matrix, quantifying four crucial predictions for every classifier; the TP, FP, FN and TN represent the true positives, false positives, false negatives and true negatives respectively (Baesens et al., 2003). When getting these one can identify the number of correctly classified objects and the number of misclassified objects yielding a general conclusion about the performance of a classifier. It is usually defined by the ratio of correct number of predictions over the whole sample data. Since the confusion matrix allows quantifying TP, FP, FN and TN for every classifier, these predictions can be used in the study to obtain the sensitivity measure which is the proportion of the positive cases that are predicted to be positive (TP/(TP + FN)). One can also get the specificity measure which is the proportion of the negative cases that are predicted to be negative

(TN/(FP + TN)). Sensitivity and specificity are influenced by the classifier's output variations especially between its extremes. These three metrics vary together as the threshold on the output varies.

Table 4. Confusion matrix table (Tax, 2001)

	Object from target class	Object from outlier class
Classified as a target object	True positive, (TP)	False positive, (FP)
Classified as an outlier object	False negative, (FN)	True negative, (TN)

Another performance metric that will be used to account for class imbalance is the ROC curve. The receiver operating characteristic (ROC) curve is a two-dimensional graphical representation of the true positives (TP) rates vs the false positives (FP) in the range [0;1] for many values of the classification threshold. Each point on the curve represents the cutoff probability. The point on the upper left corner, the closest to (1,1) is usually the optimal point on the ROC curve representing the optimal classification delivered by a classifier at a certain threshold. As a result of the ROC curve, the area under the curve (AUC) will be calculated and used in the study to quantify the performances: The area under the curve can vary between 0.5 for a worthless performance, to 1 for a perfect classifier performance.

1. *Summary of evaluation metrics in the study*

- Accuracy: accuracy measure, which shows the percentage of the correctly classified objects. (This measure is a weak measure and might not lead to accurate results if not mentioned with sensitivity and specificity measures)

$$\text{Accuracy} = \frac{TP+FP}{TP+FP+TN+FN}$$

- Sensitivity: sensitivity measure, which is the proportion of the positive cases that are predicted to be positive.

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

- Specificity: specificity measure, which is the proportion of the negative cases that are predicted to be negative.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

- ROC curve: is a two-dimensional graphical representation of the true positives (TP) rates vs the false positives (FP), for many values of the classification threshold. Each point on the curve represents the cutoff probability.
    - The area under the curve (AUC): In order to use the ROC curves as a comparison tool between different classifiers, one must calculate the area under the receiver operating characteristic curve (AUC). AUC gives an estimation of the probability that a randomly selected object of class 1 (positive object) is correctly ranked higher than a randomly selected object from class 0 (negative object).

C. Evaluation Metrics

Each dataset obtained will be explored and cleaned in order to gain insights about the variables of interest, leading to a correct classification of defaulters and non-defaulters. For each dataset, inferential and exploratory analysis will be made to extract some observations about the datasets. These analysis techniques will be performed on each dataset in later stages of the study. The next section will show the inferential and exploratory analysis applied for two of the chosen datasets which are the Taiwanese and the German datasets.

Oversampling or Undersampling techniques are usually applied before performing any classification technique when the data appears to be highly imbalanced. There is no agreement between researchers for the ratio of imbalance but assuming if the minority data (percentage of defaulters) is represented by less than 20%, the dataset will be considered imbalanced and thus the oversampling or undersampling techniques will be applied to account for the existing imbalance. Oversampling means to over sample the minority category in order to make it representative while assessing a classifier’s performance. In contrast, Undersampling techniques usually means to under sample the majority category in order to make it less represented while assessing a classifier’s performance, in order to highlight the performance of the minority class.

The datasets in the experiment were divided into three subsets: training set (60%), validation set (20%) and testing set (20%). The training set will be used to train the classifiers, the validation set will help in tuning their parameters while the test set will be used to verify their performances.

One must note that the one class classification technique is known for training the dataset with only objects belonging to the target class (non-defaulters). So that is what will be done during classification. At first, the inliers will exist in the target class (non-defaulters) which is the majority class in the datasets. Then the validation process will happen when tuning the parameters of the classifiers. Testing will happen afterwards, where a random customer will be learned by the classifier and labeled as inlier or outlier. The classifiers and their combination will then be implemented. Performance measures of each classifier and the combination will be recorded and presented in a summary table of findings in the next section. Figure illustrates the stages of the process, which the datasets in the study will go through.

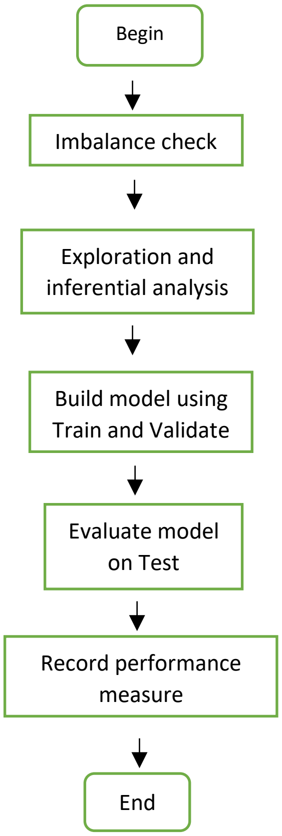

Begin
Imbalance check
Exploration and inferential analysis
Build model using Train and Validate
Evaluate model on Test
Record performance measure
End

Figure 3. Process of work

D. Combination techniques

In order to combine the classifiers in the study, one can take advantage of the soft labels. Classifiers can simply be combined using the product rule, the min, the max, the median or the mean combining rules. These combining rules will provide a probabilistic output to a new tested object. These rules are inexpensive and easy to call. One must first get the posterior probabilities of the classifiers and then apply one of the previously mentioned combining rules. After combining and compiling the new hybrid classifier on the new object, the object will have a probability for being an inlier and another one for being an outlier. The higher probability will decide the label that should be assigned to the object. The combination techniques will lead to hybrid models composed of the four classification techniques used in the study, which are the isolation forest (IForest), support vector machine (SVM), Gaussian Mixture Model (GMM) and Parzen classifier.

CHAPTER IV

ANALYSIS AND FINDINGS

The results and the findings obtained from implementing the classifiers and their combinations using R software are presented below in Table 5.

Table 5. Table showing the area under the curve (AUC), sensitivity, specificity and rank results for the German, Taiwanese and Australian datasets for the four classifiers and their hybrid models

Method	German				Taiwanese				Australian			
	AUC	Sensitivity	Specificity	Rank AUC	AUC	Sensitivity	Specificity	Rank AUC	AUC	Sensitivity	Specificity	Rank AUC
IFOREST	0.592	0.756	0.325	6	0.600	0.688	0.415	7	0.850	0.696	0.848	2
GMM	0.656	0.619	0.525	1	0.646	0.658	0.562	5	0.835	0.620	0.814	4
SVM	0.594	0.475	0.669	5	0.641	0.638	0.575	6	0.655	0.644	0.557	7
Parzen	0.587	0.594	0.575	7	0.655	0.759	0.493	3	0.735	0.658	0.695	6
Hybrid Model: 4/4	0.604	0.419	0.700	3	0.653	0.736	0.511	4	0.800	0.620	0.848	5
Top 1: Hybrid 3/4	0.614	0.225	0.925	2	0.662	0.736	0.519	1	0.836	0.620	0.898	3
Top 1: Hybrid 2/4	0.602	0.263	0.875	4	0.657	0.765	0.486	2	0.852	0.608	0.932	1

Note:

- GMM German at k = 80
- GMM Taiwanese at k = 20
- GMM Australian at k = 7
- Hybrid German using prod rule on soft labels
- Hybrid Taiwanese using median rule on soft labels
- Hybrid Australian using mean rule on soft labels
- Top 1 Hybrid 2/4 - German: SVM + GMM
- Top 1 Hybrid 2/4 - Taiwanese: Parzen + SVM

- Top 1 Hybrid 2/4 - Australian: GMM + IFOREST
- Top 1 Hybrid 3/4 - German: IFOREST + GMM + SVM
- Top 1 Hybrid 3/4 - Taiwanese: Parzen + GMM + SVM
- Top 1 Hybrid 3/4 - Australian: IFOREST + GMM + Parzen

By order, one can notice in Table 5. that the area under the curve (AUC) results for the Australian dataset were higher than the results of the Taiwanese and German datasets, followed by the Taiwanese dataset, which gave better results than the German dataset. As a first impression, one can notice that the classifiers in the study, isolation Forest (Iforest), support vector machine (SVM), Gaussian mixture model (GMM) and Parzen classifier (Parzen) have a higher predictive performance when compiled with the Australian dataset, followed by the Taiwanese dataset, and delivers intermediate to weak performances when simulated using the German dataset.

For the Australian dataset, the best predictive performance was for the hybrid models made of the IFOREST classifier and the IFOREST classifier itself. The hybrid models for the Australian dataset were obtained using the mean rule on soft labels. However, the individual classifiers delivered high results for this dataset too. Between the four classification tools, Iforest was the best individual classifier with an AUC result of 0.85 (sensitivity = 0.696, specificity = 0.848) and was the second best classifier for the Australian dataset overall. GMM and Parzen followed the Iforest classifier and ranked fourth and sixth respectively with AUC results of 0.835 (sensitivity= 0.620, specificity = 0.814) and 0.735 (sensitivity= 0.658, specificity= 0.695). The weakest performance of an individual classifier for the Australian dataset was for the SVM classifier. This classifier got 0.655 as an AUC result (sensitivity= 0.644, specificity= 0.557). The best model for the Australian dataset and overall in the study goes to a hybrid model composed of two classifiers which are the best performing individual classifier IFOREST and GMM, the second best one. This model showed a very high predictive ability with an AUC result of 0.852 (sensitivity= 0.608, specificity= 0.932). This model turned out to combine two of the best classifier performances for this dataset and delivered the highest result not only for this dataset but also for the study in general. The second best classifier for the Australian dataset is for IFOREST classifier resulting in an AUC result of 0.85 (sensitivity=0.696, specificity=0.848). The third best model for the Australian dataset is a combination of Iforest, GMM and Parzen together to deliver an AUC result of 0.836 (sensitivity= 0.620, specificity= 0.898). Followed by the hybrid models and IFOREST, which delivered the highest performances for the Australian dataset, the individual classifier GMM occupied the 4th place with an AUC of 0.835(sensitivity=0.620, specificity=0.814). GMM was followed by the hybrid model made of a combination of the cardinality of the classifiers tackled in the study who gave an AUC result of 0.800 (sensitivity = 0.620, specificity = 0.898). This performance was followed by the intermediate performance of Parzen with AUC = 0.735 (sensitivity=0.658, specificity=0.695) and the weak performance of SVM with an AUC result of 0.655 (sensitivity= 0.644, specificity= 0.557). As a conclusion, the best classifier performance for the Australian dataset goes to the combination between GMM and IFOREST with an AUC result of 0.852, and the weakest performance is for the SVM classifier with an AUC of 0.655. The weakest classifier for the Australian dataset turned out to be performing better than most of the individual and hybrid models for the Taiwanese and German datasets.

The hybrid models for the Taiwanese dataset resulted in an intermediate predictive power. The models were constructed by applying the median rule on the soft labels of the individual classifiers in the study. Furthermore, the individual classifiers resulted in intermediate performances for the Taiwanese dataset too. Between the four classification techniques, Parzen was the top individual classifier with an AUC result of 0.655 (sensitivity = 0.759, specificity = 0.493). The second high individual performance was for the GMM classifier with an AUC result of 0.646 (sensitivity= 0.658, specificity=0.562). SVM and IFOREST turned out to be the weakest individual classifiers with AUC results 0.641 (sensitivity= 0.638, specificity= 0.575) and 0.600 (sensitivity= 0.688, specificity=0.415). The best model for the Taiwanese dataset was a hybrid model composed of the strongest classifier, which is Parzen and SVM. These two resulted in an AUC of 0.662 (sensitivity= 0.736, specificity=0.519). The second best classifier is also a hybrid model composed of Parzen, SVM and GMM classifiers. These resulted in an intermediate performance with an AUC of 0.657 (sensitivity= 0.765, specificity=0.486). The third best classifier is Parzen classifier itself, which was part of the hybrid models. This shows that the strongest classifier combined with other classifiers is able to give light to a classifier performing better than this classifier alone. Parzen delivered an AUC result of 0.655 (sensitivity = 0.759, specificity = 0.493). This classifier was followed by a hybrid model made of all the classifiers tackled in the study Iforest, GMM, Parzen and SVM which gave an AUC result of 0.653 (sensitivity=0.736, specificity=0.511). This hybrid model proves that a combination of all the classifiers together regardless of their predictive performances might significantly enhance the performance of the classifiers as individuals. However, this is not always the case since for the hybrid model in the Australian dataset formed from the four classifiers did not deliver a strong result. So one can conclude from this that a combination of all the classifiers in the study gives a predictive power for the hybrid model but might not actually be the optimal way to combine the classifiers and deliver the highest result. This model was then followed by GMM classifier who gave an AUC result of 0.646 (sensitivity=0.658, specificity= 0.562). The individual classifier SVM is ranked the sixth classifier with an AUC result 0.641 (sensitivity= 0.658, specificity=0.562), followed by IFOREST which was the worst classifier for the Taiwanese dataset with an AUC of 0.600 (sensitivity = 0.688, specificity = 0.415). As a conclusion, the best classifier performance for the Taiwanese dataset was a hybrid model formed a combination between Parzen classifier and SVM with an AUC result of 0.662, and the weakest performance was for IFOREST classifier with an AUC of 0.600. The highest hybrid model in the Taiwanese dataset, with the AUC of 0.662 can outperform only the weakest classifier in the Australian dataset, which is the SVM classifier when it is applied as an individual classifier with an AUC of 0.655. The weakest classifier for the Taiwanese dataset is actually performing better than most of the classifiers of the German dataset.

The classifiers applied on the German dataset did not deliver high encouraging results and using several combining strategies did not significantly enhance the performances of the classifiers. Even if the predictive power of the classifiers in the German dataset is not that high, but they delivered the highest results following GMM classifier. In any case, the results of this datasets showed that the highest performing classifier was the GMM with an intermediate AUC result of 0.656 (sensitivity = 0.619, specificity = 0.525). This classifier is followed by a hybrid model made of the combination between GMM, IFOREST and SVM who delivered an AUC result of 0.614 (sensitivity = 0.225, specificity = 0.925). The third classifier is the combination of the four classifiers together resulting in also an intermediate AUC of 0.604 (sensitivity = 0.419, specificity = 0.700). The hybrid model made of GMM and SVM ranked fourth and delivered an AUC result of 0.602(sensitivity = 0.263, specificity

= 0.875). SVM was the fifth classifier with an AUC result of 0.594 (sensitivity = 0.475, specificity = 0.669). The two weakest classifiers for the German dataset are the Iforest classifier, and Parzen classifiers who gave AUC results of 0.592 (sensitivity = 0.756, specificity = 0.325) and 0.587 (sensitivity = 0.594, specificity = 0.575). These two results for the German dataset turned out to be the worst in the study. Only the highest performing classifier for this dataset GMM with an AUC result of 0.656 provided a good competing AUC result with the weakest classifiers for the Taiwanese and Australian datasets.

After inspecting Table 6, one can conclude that the best classifiers overall turned out to be hybrid models especially for three datasets, the Australian, the Taiwanese and the German datasets. To note the German datasets won't be considered as a reliable dataset since the classifiers performed very weakly in this dataset. Combining 2 out of 4 classifiers was the first or the second best classifier; it might be delivering a guaranteed result. The fact of combining 3 out of 4 results guarantees a high predictive performance too and makes it one of the top two strongest classifiers. This shows that combining enhanced the performance of the individual classifiers and proves that combining enhances the performance of one-class classifiers. To note also that the strongest classifier is the reason behind the high predictive performances of the hybrid models. It is explained by the fact that for the three datasets, the strongest classifier is always included in the formation of the hybrid model and is between the top three best performances for any of the three datasets. The worst classifiers in Table 6 turned out to be mostly the classifiers for the German dataset. If one looks at the overall rankings, he or she can draw the conclusion that the weakest classifier is the SVM classifier for being ranked 5th to 7th, which means delivering the worst AUC results for the Australian, Taiwanese and German datasets. GMM can be labeled as intermediate performing classifier since it is ranked between 4th and 5thth for both the Australian and Taiwanese datasets and 1st for the German dataset. IFOREST performed as the best individual classifier for the Australian dataset, but turned out to be poorly performing for the Taiwanese and German datasets. At last, Parzen classifier performed best for the Taiwanese dataset but poorly for the Australian and German datasets.

The Receiver operator curve (ROC) will help in visualizing the performances of the classifiers in the study and their hybrid models. The ROC curves for the classifiers for each dataset will be represented below. One must note that the area under the curve (AUC) results are usually deducted from these graphs, by integrating the area that lies underneath the function obtained by plotting the fraction of targets accepted versus the fraction of outliers accepted.

For the Australian dataset, the ROC curves of the classifiers are shown below:

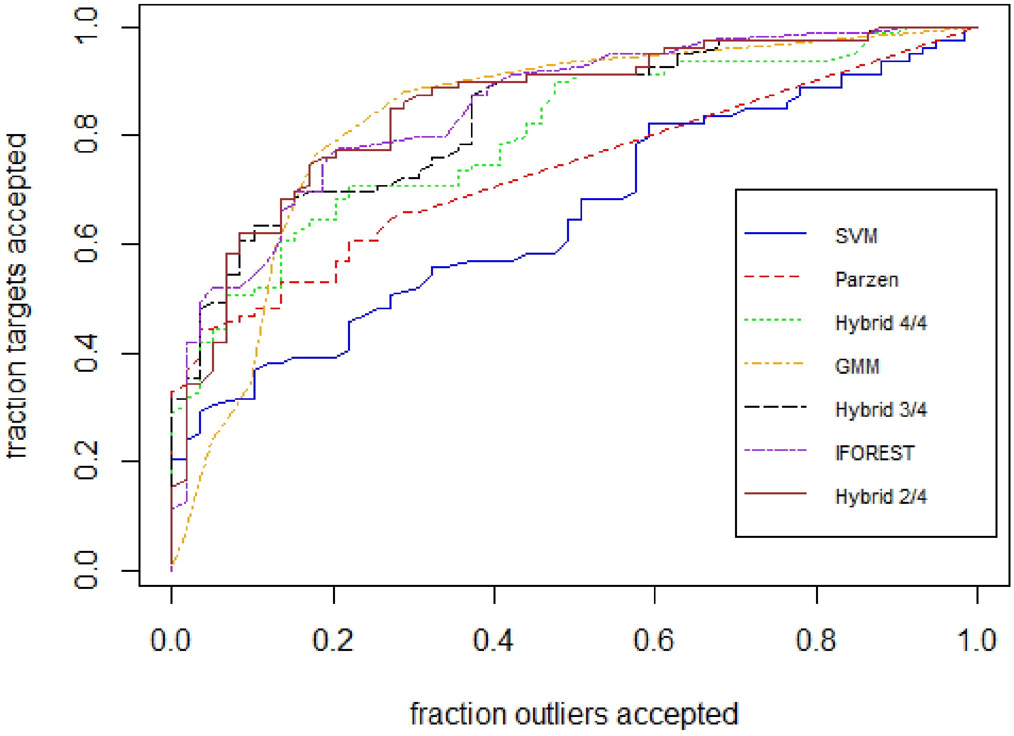


Figure 4. ROC curves – Australian dataset

As a first impression, the ROC curves for the Australian dataset visually translate the high predictive power of the classifiers in this dataset. All the curves seem to follow the left-hand and top border of the ROC space, which means that the results of the classifiers are more accurate. According to Table 5, the best performing classifier is a hybrid model made of GMM and IFOREST classifier (AUC=0.852), followed by IFOREST classifier with an AUC of (AUC=0.85). These top two results are visually represented on the ROC graph. The hybrid model made of IFOREST, GMM and Parzen (AUC=0.836), and the individual classifier GMM (AUC=0.835), then follow. The combination of all the classifiers used in the study together gave an intermediate performance resulting in an AUC (AUC=0.8). The hybrid of cardinality performed better than Parzen classifier and this is demonstrated by the AUC result of Parzen (AUC=0.735), which is less than the AUC of the hybrid. It is then clear that the worst classifier for the Australian dataset is the SVM classifier (AUC=0.655). It visually appears the farthest from the point [0;1].

For the Taiwanese dataset, the ROC curves of the classifiers are shown below:

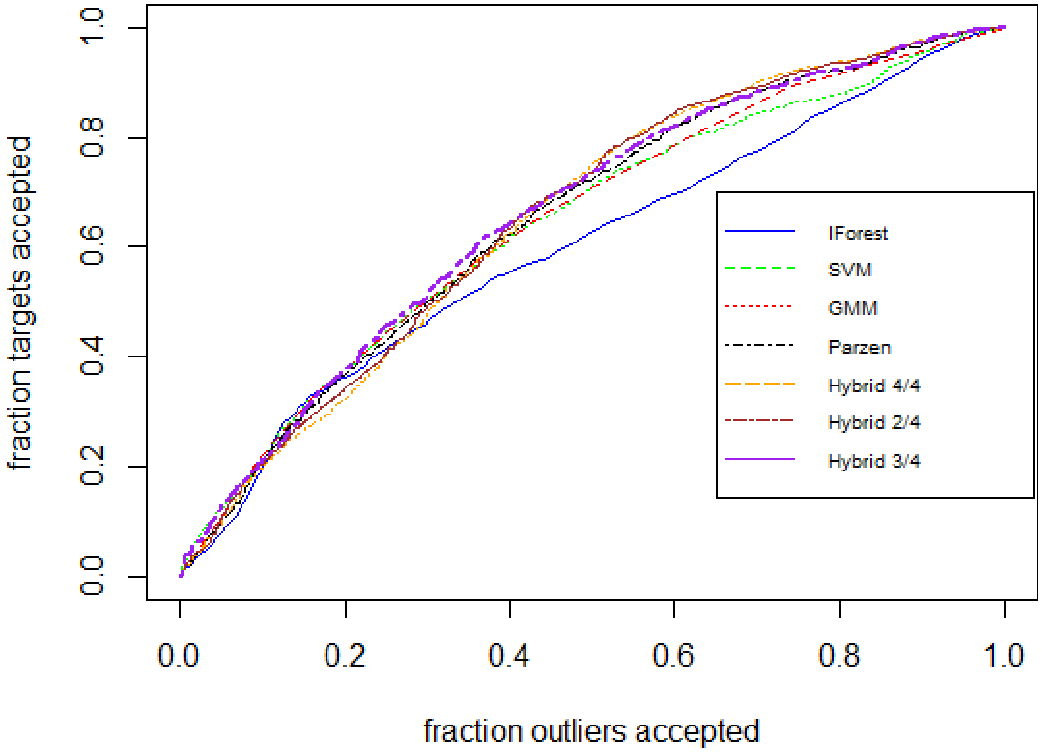


Figure 5. ROC curves – Taiwanese dataset

For the Taiwanese dataset, the ROC curves appear to be farther from the top left corner and closer to the 45-degree diagonal of the ROC space, which means that the classifiers are delivering an intermediate performance for the Taiwanese dataset. According to Table 5, the best performing classifier are the hybrid models made of Parzen, GMM and SVM, and the one made of Parzen and GMM then follow with AUC results of 0.662 and

0.657 respectively. All the other individual classifier have intermediate performances, and it is graphically clear that IFOREST is the worst classifier for the Taiwanese dataset which gave an AUC result of (AUC=0.600).

For the German dataset, the ROC curves of the classifiers are shown below:

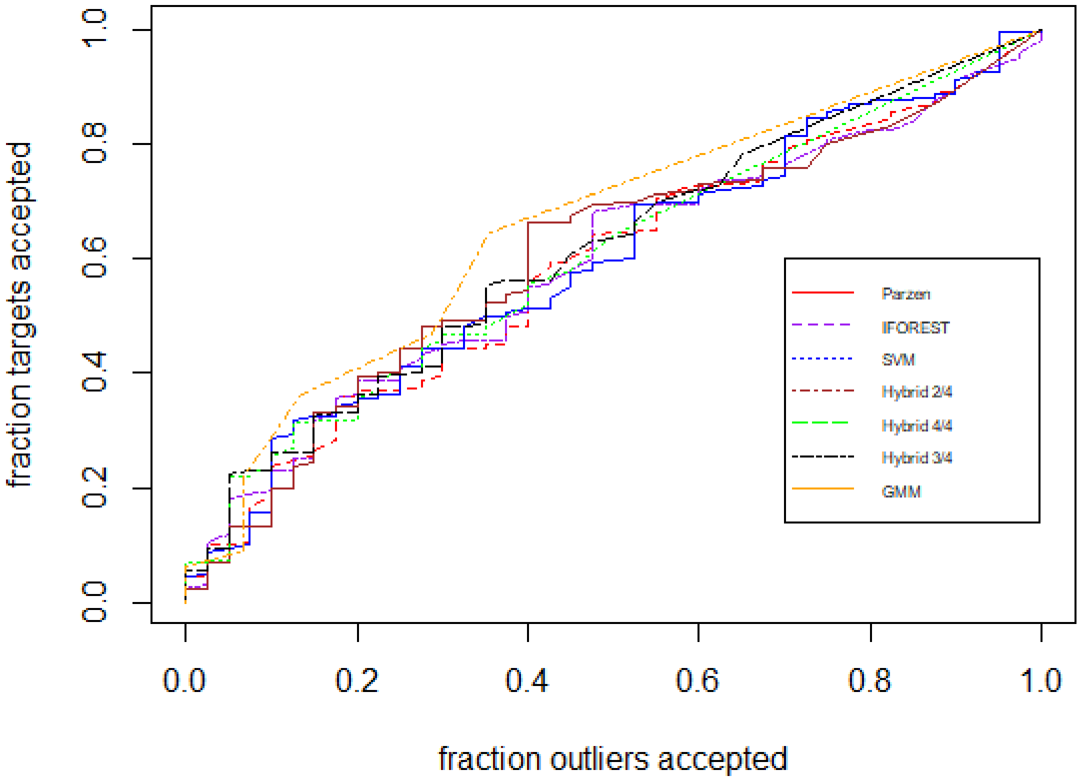


Figure 6. ROC curves – German dataset

The ROC curves for the German dataset are very close to the 45-degree diagonal of the ROC space, which proves that the classifiers for the German dataset are less accurate than their performance with other datasets. Visually, it shows that GMM topped the performance of the other classifiers with an AUC result of (AUC=0.656). The hybrid models enhanced the performances of the weak classifiers especially for Iforest, GMM and Parzen. The hybrid models formed by IFOREST, GMM and SVM resulted in (AUC=

0.614), and the hybrid made of the cardinality of the classifiers resulted in an AUC of (AUC=0.604). The hybrid made of GMM and SVM (AUC=0.602) then followed. These intermediate performances were followed by the weak performances of the individual classifiers SVM, IFOREST, Parzen who delivered AUC results close to 0.59. The worst classifier turned out to be the Parzen classifier with an AUC result of (AUC=0.587).

CHAPTER V

COMPARATIVE SECTION

In order to evaluate the study's results, this section will address one of the goals of the study and will use the findings obtained by the machine learning techniques in the results section, to compare them against the findings achieved in the literature. To note that all the past records collected are explained in details in the literature review section. The models, the performance measures and the state of the art performances for the best classifiers in each study are tackled in details in that section. This section will serve the comparison between the study's findings and the findings from previous literature.

To start with, the table of findings shows that combining yields the best performances in classification for all the three datasets. The hybrid models including the strongest individual classifier, delivered the highest predictive results between all the other combination trials. The hybrid models reflected very improved results and ranked in the top performances with AUC results varying between 0.8 and 0.852 for the Australian dataset and having approximately AUC results around 0.660 for the Taiwanese dataset. For the German dataset, the hybrid models delivered improved results but still considered as intermediate performances with AUC results varying between 0.602 and 0.614.This performance can be compared with the work of (Khashei and Mirahmadi, 2015) where they used the Australian dataset and combined multiple multi-layer perceptrons (MLPs) in a fuzzy way using fuzzy parameters. They obtained a hybrid model that delivered the best result in their study with a classification error of 10.9% (the lowest error compared to SVM 22.5%, ANN 12.3%, k-NN

14.2% and many individual traditional classifiers). This helps to draw the conclusion that even if the application of the hybrid models was minimal in the past literature but it led to deliver the highest results in the studies similar to the findings obtained from this research study.

For the Gaussian mixture model (GMM), the study's findings proved that it is one of the intermediate performing classifiers by scoring an AUC result of 0.835 (sensitivity= 0.620, specificity= 0.814) for the Australian dataset and being ranked the fourth classifier out of 7 applied in the study. This predictive performance proved to be similar to the GMM's performance in (Kennedy, Namee & Delany, 2013) work, where it ranked the 5th out of 9 classifiers and got a harmonic mean of 69.9. In contrary, for the German dataset, GMM ranked in the first place especially with an AUC result of 0.656, which is low compared to the study's results. GMM performed the worst for the German dataset in the work of (Kennedy, Namee & Delany, 2013) with a harmonic mean score of 55.2 the lowest between 9 classifiers used in the study. This can be explained by the fact that the number of classifiers used in this study are fewer than in the work of (Kennedy, Namee & Delany, 2013) and the fact that only GMM and Parzen are common between both of the studies. This cannot help in ranking the performance of GMM classifier. However, for both studies, either for AUC results or for harmonic mean results the predictive performance of GMM appeared to be intermediate in this study but very low in their work. This is due to the fact that, in both studies the German dataset does not deliver high results for any of its classifiers and does not help in predicting the performance of its classifiers. Therefore, any result deducted from the German dataset won't be very accurate to assess due to the weakness of this dataset.

The results of the Parzen classifier (Parzen), for the Australian dataset, showed that this classifier gave an AUC of 0.735 (sensitivity=0.658, specificity= 0.695) which was the worst individual classifier between the four chosen classifier's especially for the Australian and German datasets. After optimizing the thresholds in (Kennedy, Namee & Delany, 2013) work, Parzen delivered an intermediate to bad performance and ranked 5th with a harmonic mean of 70.9. For the German dataset, Parzen ranked the 7th in the study resulting in an AUC of 0.587 (sensitivity=0.594, specificity= 0.575). This result was similar to the result delivered by (Kennedy, Namee & Delany, 2013), where the parzen classifier delivered a harmonic mean result of 58.8, which is considered intermediate to low. Even though the results show that for the German dataset Parzen performed the best, but its predictive power is very low since one can notice if in this study (using AUC results) or in the (Kennedy, Namee & Delany, 2013) work (using harmonic mean results), the predicted values of the German dataset are much lower than any values predicted in both studies. As a result, this proves that the German dataset results are not very reliable.

For the support vector machine (SVM) classifier, the study showed that it delivers an intermediate to weak performance resulting in an AUC of 0.655 (sensitivity= 0.644, specificity= 0.557) for the Australian dataset. This result is lower than the AUC values of SVM for a low imbalanced dataset in (Marques, Garcia and Sanchez, 2013), which gave an AUC value of 0.71. The result of the study were higher than highly imbalanced data for the Australian dataset, which had an AUC value of 0.61. These results makes sense since the data was moderately imbalanced in the study which means that the AUC results for SVM using the Australian dataset should lie between the low and the highly imbalanced dataset results, i.e. between 0.61 and 0.71 and that's the case of this study where the AUC are 0.655.

Another work done by (Khashei and Mirahmadi, 2015) proved that SVM delivered an intermediate to low performance for both the Australian dataset. They used in their study the misclassification rate of classification models in order to evaluate the performance on the classifiers. SVM turned out to perform the worst between all the classifiers with a classification error of 22.5%. Hence, it is obvious in both cases that SVM is not the best performing classifier to use in order to predict accurate results for the Australian dataset. (Flemotomos, 2017) worked on pattern recognition techniques for credit classification and found that SVM performed with an accuracy of 66.5%, which reflects that SVM has an intermediate performance in accurate prediction. As for the German dataset, the SVM delivered an AUC result of 0.594 (sensitivity=0.475, specificity=0.669) which is higher then what was mentioned by (Junior et al., 2019) when they got the results of SVM classifier in many imbalance stages. It turned out that SVM gave an AUC result around 0.51 for moderate imbalance in the German dataset, which is the case in the study. As a result, the study gave better results for SVM than in (Junior et al., 2019) work. This work is assumed to be very low in predicting accurate classification results due to low AUC results. As for the Taiwanese dataset, SVM also reflected an intermediate performance with an AUC result of 0.641 (sensitivity=0.638, specificity= 0.575). This result was also proved by the work of (Lu, Hagreaves and Bilal, 2018) when it ranked the fourth out of 6 classifiers applied in the study with an accuracy result of 81.9%. The results for the Taiwanese dataset and for many other datasets, were hard to compare with each other since the performance metric used in the study (AUC is the most common metric to measure the performance of classification techniques) is different than some evaluation metrics used in the articles found about the classifiers for the specific datasets. That explains why the results of SVM and some other classifiers were approximated per ranking and comparison with the results of the other classifiers found in the articles in order to come up with a performance approximation.

Finally, one must note that no previous record was found for the IForest classification technique using one of the three datasets applied in this study. This might be due to the fact that IForest is a new classification technique that was created in 2008 by Fei Tony Liu, Kai Ming Ting and Zhi-Hua Zhou, and applied only in their work in 2012 and 2013. This method is completely brand new in the world of classification. However, its performance turned out to very high for the Australian dataset since it ranked as the best individual classifier for the Australian dataset with an AUC result of 0.850 (sensitivity= 0.696, specificity= 0.848) and the second best classifier for the study. IForest performed poorly for the Taiwanese and German datasets with AUC results of 0.600 (sensitivity= 0.688, specificity= 0.415) and 0.592 (sensitivity= 0.756, specificity= 0.325) respectively.

CHAPTER VI

CONCLUSION

In conclusion, the research questions were answered and this can be proved by the findings of the study. The goal of this research was to show that using a hybrid model by combining the results of two or more one-class classifiers together will yield an improvement in the classification results for the credit-scoring problem. After implementing the classifier and completing the analysis, one can deduct that the hybrid models increased the predictive ability of the one-class classification techniques by applying several combination strategies to the traditional classifiers. The models resulted in a direct improvement of the classification results especially when the combined model included the strongest predictive classifier for the dataset. Their accurate probabilistic output represent a more credible tool that will help loan lenders to assess the situation of a customer and predict accurately his/her behavior towards defaulting or not against paying their dues in order to grant or not grant them the loans in the first place. The best individual classifier for the Australian dataset turned out to be the Iforest with an AUC result of 0.85 (sensitivity=0.696, specificity= 0.848). Meanwhile for the Taiwanese dataset, the best individual classifier turned out to be the Parzen classifier with an AUC result of 0.655 (sensitivity=0.759, specificity=0.493). For the German dataset, the best classifier that delivered an intermediate performance is GMM classifier with an AUC result of 0.656 (sensitivity=0.619, specificity= 0.525). Overall, the best result in the study turned out to be a combination between Iforest and GMM for the Australian dataset which delivered an AUC of 0.852 (sensitivity=0.608, specificity=0.932).

Another goal of the study was to use the literature articles to assess the classification methods applied in the study. Records of the same classification techniques were collected from previous. The results obtained by the study were compared to the past literature. One gets to conclude that the individual and the hybrid classifiers applied in the study performed approximately in a similar way as they were recorded in the past literature. This article becomes a benchmark to be used later on by other studies in order to compare the performances of new techniques applied to these three datasets.

Another conclusion derived from the study is that the attributes for a given dataset might affect its performance. This is shown by the difference in classifiers performances for the three datasets used in the study. One can clearly notice the high predictive performances for the classifiers in the Australian dataset, the intermediate performances in the Taiwanese dataset, and the poor performances in the German dataset whether in this study and in previous literature. The features that existed in some datasets contributed in deteriorating their performances, especially for the German dataset. Features like Average number of credits (p-value = 0.1475, confidence interval [1.42, 1.36]), other debtors (between guarantors, co applicant and none, none was the one who had the highest count) and many other features turned out to be not significant versus the defaulting variable. There exist also other features which arise the need to remove them due to the fact that they seem completely irrelevant to the defaulting. These features are the number of people being liable to provide maintenance for, present residence since, telephone and foreigners. Their results also in the exploratory analysis show that these factors are not related to the defaulting problem. These 6 out of 20 features might be one of the reasons why the results for the German dataset were very low. Another reason that would explain the effect of the attributes negatively on the performance of some datasets is the fact that some models might not interact well with some features. While applying the models, the features might interact in a way that worsens the performance of the classifiers instead of enhancing it. One must note that performance of the classifier is directly related to what the classifier is trained on and tested. So this matter should always be checked when running models on specific datasets with specific features.

To note that this study focuses on the machine learning techniques applied in the study and mainly sheds light on the classifiers, and tuning the hyper parameters in order to maximize their performance and be able to get the best results out of them. However, some datasets still performed badly and didn't help in drawing any conclusion that serves the objectives of the study. A better way to enhance the performance of several datasets and make use of them is to take into consideration the business objectives of the studies. The performance of the classifier is directly related to what the classifier is trained on. That's why, there is a need to study which variables to include and start removing some of the not very interesting variables. The inferential and exploratory analysis helped in identifying the significantly statistical variables to the variable of interest whether the customers will default or not on their next payment. This study helped in coming up with this conclusion and get to the fact that some attributes need to be removed because they simply worsen the results of some models using the datasets. That's why there is a need to do feature selection and ranking of the features that are most important for the classification task. This strategy would actually help in identifying the variables to include and removing the ones that are not statistically significant, and will help overall in Constructing improved datasets yielding better prediction.

In order to improve the predictive ability of the hybrid and get the best performances out of them in future work, one can form an optimization problem after assigning weights to the individual classifiers for a specific dataset, according to their impact in the results and solve this problem. In other words, the individual classifiers can be combined linearly by weighing their output and optimizing the weights to provide the highest AUC. More generally, a trained (nonlinear) combiner can be used as a second classification stage where the classifier is trained on the probabilities resulting from the individual base classifiers.

REFERENCES

Belloti, T., & Crook, J. (2009). Support Vector Machines for Credit Scoring and Discovery of Significant Features. *Expert Systems with Applications, 35*(2), 3302-3308.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Catania, C. A., Bromberg, F., & Garino, C. G. (2012). An autonomous labeling approach to support vector machines algorithms for network traffic anomaly detection. *Expert Systems with Applications*, *39*(2), 1822-1829.

Chen, M. C., & Huang, S. H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications, 24*(4), 433-441. doi: 10.1016/s0957-4174(02)00191-4.

Durand, D. (1941). *Risk elements in consumer installment financing*. National Bureau of Economic Research, New York.

Fessler, J. A., & Hero, A. O. (1994). Space-alternating generalized expectation-maximization algorithm. *IEEE Transactions on signal processing*, *42*(10), 2664-2677.

Flemotomos, N., (2017). Pattern Recognition Techniques for Credit Classification. *Ming Hsieh Department of Electrical Engineering.*

Henley, W. E., & Hand, D. J. (1996). A k-nearest neighbour classifier for assessing consumer credit risk. *The Statistician, 45*(1), 77-95.

Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications, 33*(4), 847-856. doi: 10.1016/j.eswa.2006.07.007

Kennedy, K., Namee, B. M., & Delany, S. J. (2013). Using Semi-supervised Classifiers for Credit Scoring. *Journal of the Operational Research Society, 64*, 513-529. doi:10.1057/jors.2011.30

Khashman, A. (2009). A Neural Network Model for Credit Risk Evaluation. *International Journal of Neural Systems, 19*(4), 285-294.

Lando, D. (2009). Credit risk modeling. In *Handbook of Financial Time Series* (pp. 787-798). Springer, Berlin, Heidelberg.

Lando, D. (2004) Credit Risk Modeling. *Princeton Series in Finance, Princeton UP, USA.*

Lee, T. S., & Chen, I. F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications, 28*(4), 743-752. doi: 10.1016/j.eswa.2004.12.031

Li, S. T., Shiue, W., & Huang, M. H. (2006). The evaluation of consumer loans using support vector machines. *Expert Systems with Applications, 30*(4), 772-782. doi: 10.1016/j.eswa.2005.07.041

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 413-422). IEEE.

Luo, S. T., Cheng, B. W., & Hsieh, C. H. (2009). Prediction model building with clustering launched classification and support vector machines in credit scoring. *Expert Systems with Applications*, *36*(4), 7562-7566. doi: 10.1016/j.eswa.2008.09.028

Majid, A., Khan, A., & Mirza, A. M. (2005, December). Intelligent combination of Kernels information for improved classification. In *Fourth International Conference on Machine Learning and Applications (ICMLA'05)* (pp. 6-pp). IEEE.

Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data mining and knowledge discovery*, *2*(4), 345-389.

Puggini, L., & McLoone, S. (2018). An enhanced variable selection and Isolation Forest based methodology for anomaly detection with OES data. *Engineering Applications of Artificial Intelligence*, *67*, 126-135.

Tax, D. (2001). *One-class classification.* University of Tech.

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. Society for industrial and Applied Mathematics.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting*, *16*(2), 149-172.

Vapnik, V. (1998). Statistical learning theory Wiley. *New York*, *1*.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, *27*(11-12), 1131-1152.

Yu, L., Wang, S. Y., & Lai, K. K. (2009). An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: The case of credit scoring. *European Journal of Operational Research, 195*, 942-959. doi: 10.1016/j.ejor.2007.11.025

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, *3*(1), 1-130.

Zurada, J., & Kunene, N. (2010). Performance Assessment of Data Mining Methods for Loan Granting Decisions: A Preliminary Study. Paper presented at the *Artificial Intelligence and Soft Computing -10th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2010)*.