CREDIT SCORING USING SUPPORT VECTOR MACHINES WITH MAHALANOBIS DISTANCE FUNCTION

Kevin Kanyari Wachira

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Mathematical Finance and Risk Analytics of Strathmore University

Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

June 2020

Declaration

I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name: **Kevin Kanyari Wachira**

Signature:

Date: April 1, 2022

Approval

The thesis of (your name) was reviewed and approved by the following:

Dr. Rogers Ondiba

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Signature:

Date:

Abstract

Credit granting represents a significant part of most financial institutions business. The decision on who to grant or deny credit is a critical determinant of the overall profitability of these institutions. There exist many tools to help risk managers classify credit applicants as either risky or credit worthy. These tools vary both in their methodology and overall performance. With recent advancement in machine learning technology several methods have been proposed as suitable alternatives to traditional credit risk assessment models. The popularity of machine learning methods is partially due to their suitability to deal with voluminous data sets that would otherwise be impossible to handle with other traditional methods. However, they are often criticized for lacking interpretability as compared to the other methods. Significant improvements in classification accuracy means that they are still used in classification problems. While previous studies have focused on the effect of proper model framework in improving prediction accuracy the correlation between data points has rarely been studied. Preliminary review of studies relating to incorporation of Mahalanobis distance metric in classification problems suggest that a similar approach can increase prediction accuracy of SVM in credit scoring. The goal of this research is to combine SVM with Mahalanobis distance function for the purpose of credit risk assessment. A review of several studies that have combined Mahalanobis distance with SVM in classification problems show that the method leads to an overall increase in overall prediction accuracy.

Table of contents

List of abbreviations

MDA	Multiple Discriminant Analysis
WMD	Weighted Mahalanobis Distance
BP	Back propagation
SVM	Support Vector Machines
AUC	Area under Curve
PCA	Principal Component Analysis
KPCA	Kernel Principal component Analysis
ICA	Independent Component Analysis

Acknowledgement

I wish to thank the entire Strathmore Institute of mathematical sciences and my supervisor Dr. Rogers Ondiba for the support and guidance provided during the period of preparation of this research proposal. I also wish to extend my sincere gratitude to my parents and siblings for their support and words of encouragement. Most importantly, I thank God for his eternal guidance.

Dedication

This thesis is dedicated to God Almighty for giving me wisdom and good health.

Chapter 1

Introduction

1.1 Background to the study

One of the key decisions that Banks and other lending institutions have to make is which loan application to accept. This decision is key to the overall profitability of these financial institutions as loans constitute a significant proportion of their assets. Schreiner (2003) explained that the lending decision can either be made through subjective scoring or statistical scoring.Hand and Henley (1997) defined credit scoring as statistical methods that are used to classify applicants into either good or bad risk classes. Subjective scoring on the other hand relies on the input of an expert, in most cases a loan officer, to produce a qualitative judgement. Both approaches have their strengths and weakness and are often times used together. A study carried out by FSD Kenya Kenya (2008) in 2008 found that if well developed and managed credit scoring can help reduce reliance on collateral, reduce turnaround time from application to approval and also lead to risk based pricing that would mean lower rates for low-risk borrowers. In addition, Vidal and Barbon (2019) note that the use of an automated scoring solution would;

- Lower administrative costs per application
- Reduce the need for in-person applications
- Reduce the number of creditworthy applicants rejected
- Reduce the number of un-credit worthy applicants accepted
- Standardize decision making rules and minimize human error

One of the reasons why credit scoring has become popular in the recent past is due to IRB approach requirement in Basel II capital accord that required financial institutions to assign probability of default to every credit in their portfolio and charge the bank capital accordingly. Burton (2012) noted that the use of credit scoring contributed to financial inclusion by allowing credit to be granted to a wider proportion of the population in different locations and at the same time reducing losses that result from default.

Nowadays statistical models have become commonplace and there exists a variety of models used for credit scoring. Durand (1941) pioneered the use of Discriminant analysis in classification of good and bad loans. Altman (1980)) later used Multiple discriminant analysis in predicting of corporate default. The number of statistical scoring approaches has continued to grow as advances are made in the world of mathematics and computing. Support vector machines have become a popular classification tool with widespread usage in different fields including credit risk. With all these tools available, attention has shifted to improving the accuracy of the existing models as even a marginal increase in accuracy will mean an increase in the overall profitability of the lending institution.

Different statistical scoring techniques are in use to help lenders manage credit risk all over the world. In the recent past machine learning techniques have been used by researchers to improve the accuracy of this models. This has led to a consistent improvement in the performance of this models however a lot of information in the credit data goes unused in some of these models.

Support Vector Machines have become widely applied in binary classification problems and have proven to be quite useful in various fields. Credit risk is one of the areas where this machine learning technique is being used to help risk managers and financial institutions distinguish between good and risky borrowers. Data used in credit risk scoring is usually highly imbalanced and multivariate in nature. Support vector machines although powerful classifiers tend to be ineffective with imbalanced data. Our proposition is to combine SVM with Mahalanobis distance function to remedy these shortcomings. Mahalanobis distance is especially suited for multivariate and imbalanced data. By taking into consideration the correlation structure of the credit scoring data we can reduce the dimensionality of the data and improve the accuracy of SVM in distinguishing between good and risky borrowers.

1.2 Research Objectives

This research proposal has the following objectives:

- Use support vector machines with Mahalanobis distance metric for credit scoring
- Compare performance of support vector machines with Mahalanobis distance metric to svm with Euclidean distance metric in credit scoring

1.3 scope of the study

This study will mainly focus on retail credit scoring using support vector machines using data from the peer-to-peer lending firm known as lending club. The study will be focused on comparison of performance of SVM under the Mahalanobis distance metric as compared to SVM under Euclidean distance metric.

1.4 Significace of the study

- Better prediction of default will reduce losses incurred by banks and financial institutions due to non-performing loans.
- Financial inclusivity by not denying credit worthy individuals loans due to misclassification.
- The use of Mahalanobis distance metric takes advantage of any correlation in the credit data thereby reducing the number of features and increasing the computational efficiency of SVM

Chapter 2

Literature review

The two decisions that lending firms have to make are whether to grant credit to new applicants and whether to increase the credit limit of existing customers (Thomas, 2000). The practice of evaluating new applicants before granting them a loan is known as credit scoring while behavioral scoring refers to techniques used to determine what actions to take with existing customers. Thomas highlighted that most credit scoring models do not take into account rejected applications and therefore cannot tell how the applicant would have behaved. This often means that this particular group might never get the opportunity to receive credit. A response to this has been the idea of reject inference. Reichert et al. (1983) concluded that including the rejected applications added some information that was useful in predicting credit risk. Hand and Henley (1993) also noted that the only way to overcome the problem was to assume a relationship between the goods and bads that which holds for both the accepted and rejected populations.

Credit granting decision were purely judgement based before statistical scoring became widespread Goh and Lee (2019). Durand (1941) pioneered the use of discriminant analysis to distinguish between good and bad loans. This was later expanded by Altman (1968) when he used multiple discriminant analysis (MDA) to predict corporate default. Other methods used include the mathematical programming approach by Kolesar and Showers (1985) and the logit approach by Ohlson (1980). Ohlson cited certain statistical requirements that MDA imposes on the distribution of the predictors as one of its weaknesses. He also noted that the output of MDA has little intuitive interpretation and proposed the condition logit analysis as a way to avoid these problems. Goh and Lee (2019)) noted that logit turned out to be the standard scoring model due to its ability to fulfill all the requirements from the Basel II accord.

Due to the ever-growing number of models proposed for credit scoring, a number of researchers have carried out studies to document these approaches. Baesens et al. (2003) conducted a benchmark study where they used SVM and least square SVM to develop a credit scoring model and compared their performance with other popularly used models across eight different data sets. Their results showed that least square SVM and Neural networks produced better results than other models. This has contributed to increased interest in the use of SVM in credit risk modelling. Lessmann et al. (2015) conducted a large-scale benchmark of forty-one classification methods across eight credit scoring datasets where they found that sophisticated methods do not necessarily improve accuracy.

Desai et al. (1996) compared the performance of neural networks, linear discriminant analysis and logistic regression in credit scoring and found that neural networks outperformed linear discriminant analysis but performed comparably to logistic regression. Huang et al. (2007) pointed out that the problem with traditional classification method was that an assumption on the underlying probability model must be made before calculating the posterior probability upon which the classification decision is made. Huang et al. (2007) proposed the use of neural networks, genetic programming and support vector machines which are not faced by this limitation.

Support vector machine has become one of the most used classification tools used in machine learning ever since its introduction by Cortes and Vapnik (1995).Vapnik (1999) explained that support vector (SV) machine implements the following idea: It maps the input vectors x into a high-dimensional feature space Z through some nonlinear mapping, chosen a priori. (The nature of statistical learning). According to Bhavsar and Ganatra (2015) SVM constructs an optimal hyperplane as a decision surface, to divide the data points of different categories in the vector space. Kernel functions are used to extend the concept of the optimal separating hyperplane so that nonlinear data can be linearly separated.

A key influence in the overall performance of SVM is the selection of the appropriate kernel. According to Bhavsar and Ganatra (2015) each kernel function has its own characteristics. For instance, linear and polynomial kernel functions have global characteristics meaning samples far from each other can affect the value of the kernel function while the radial basis function (RBF) kernel function has local characteristic which only allows samples close to each other to influence the value of the kernel function.

Wang et al. (2007) suggested that SVM performance is limited as the data distribution information is underutilized in determining the decision hyperplane. The authors proposed a large margin learning model, max-min margin, that extended SVMs by considering class structures into decision boundary determination by utilizing the Mahalanobis distance as the distance metric. They further suggested starting with determining the data structure by performing agglomerative hierarchical clustering AHC. Wang concluded that the performance of SVM with weighted Mahalanobis distance (WMD) kernels increased as the training set grew larger and decreased as the training set shrank.

Zhang et al. (2015) explains that since the 1990s, credit risk evaluation methods have significantly improved prediction accuracy by using artificial intelligence models such as neural networks support vector machine (SVM). The authors found that SVM models were more accurate than Back propagation Neural networks in evaluating small and medium enterprises credit risk. In addition, they found that although SVM models had higher type II errors than BP neural networks they had a lower Type I error. Altman (1980) noted that the loss induced by type I error is higher than the loss induced by type II error. Type 1 error here refers to misclassification of a bad credit as good credit resulting in a bad loan. Abdou and Pointon (2011) made the same conclusion and went further and postulated that it was five times costlier to make a bad loan than to deny a non-risky borrower a loan.

Putri et al. (2021) tested four kernel models; linear kernel, polynomial, RBF, and sigmoid to determine which model predicted default best. They concluded that while all SVM models could be said to classify data well, it was the SVM model with the polynomial kernel that had the highest accuracy and AUC. Further, they noted that the model could be used to assist a Bank in deciding whether to accept or reject the application by classifying customers into either good or bad credit.

While both Logistic regression and Support vector machines are popular methods of credit scoring, they suffer from the curse of dimensionality Han et al. (2013). According to Cao et al. (2003), while all available inputs can be used as inputs of SVM, correlation among those features could lead to a deterioration of the overall performance of SVM. Cao et al. (2003) noted that several methods are available for feature selection including PCA, KPCA and Independent component Analysis (ICA). The authors compared the performance of this three methods and concluded that KPCA outperformed the other methods in feature selection. Sun et al. (2019) demonstrated the effectiveness of KPCA in extracting non linear features by combining it with SVM in a Raman spectroscopy model.

Wang et al. (2012) concluded that the use of KPCA in dimensionality reduction increased the training efficiency of SVM. Sundaram (2009) proposed reducing the support vectors in SVM by using KPCA and concluded that the method resulted in a 90% reduction of support vector with less than 1% degradation in accuracy.

Wang et al. (2007) proposed combining SVM with weighted Mahalanobis distance kernel. However, the author acknowledged that the constructed WMD kernel did not constitute a metric as the triangle inequality was not always satisfied. Jiang et al. (2018) noted that the integration of Mahalanobis distance with SVM for the purpose of credit risk assessment has not been extensively studied. The authors formulated a number of new Mahalanobis distance-based kernels and concluded that the Mahalanobis infused kernels were more robust as compared to their Euclidean counterparts.El-Banna (2015) proposed the use of Mahalanobis genetic algorithm in feature selection of imbalanced welding data. El-Banna (2015) found that both the Mahalanobis genetic algorithm and the Mahalanobis-Taguchi system were suitable for imbalanced data classification

A popularly used measure of describing the quality of a credit scoring model is the confusion matrix.Bakker et al. (2019) explained that the Average correct classification rate can be computed from the confusion matrix otherwise referred to as a classification matrix.However Bakker et al. (2019) further added that a particular weakness of the Average correct classification is that it ignores the costs of misclassification.West et al. (2005) used the estimated cost of misclassification in their study where they employed neural network ensemble strategies to predict default. These performance metrics are not restricted to finance only.El-Banna (2015) used AUC to evaluate the performance of his Mahalanobis based classification algorithm on imbalanced welding data. According to Satchel and Xia (2008) the most common metrics of discriminatory power of classifiers are the Cummulative Accuracy profile(CAP), the Accuracy ratio as well as area under ROC curve(AUC).

Chapter 3

Methodology

3.1 Objectives

This Research proposal will focus on two objectives

- Combine the SVM algorithm with Mahalanobis distance metric to predict retail loan default.
- Compare prediction accuracy of SVM model with Mahalanobis distance metric with SVM under Euclidean distance metrics

3.2 The data

The data used for this research is sourced from Kaggle https://www.kaggle.com/datasets/ethon0426/lending-club-20072020q1. Kaggle is a google owned platform for data scientists and Machine learning practitioners. The platform allows for upload of data by members and has several data sets used for credit risk analysis. The data set chosen for this research comprises of credit history information of customers from a financial institution. Machine learning algorithm work with specific data sets therefore the data will have to be prepared before it is used. The pre-processing stage will consist of several steps

- Data cleaning; this will involve checking the data for missing values and deleting them from the data set
- Converting data to the required format

3.3 Exploratory data analysis

As a preliminary step I will use various plots to visualize the data and determine if any inference can be made about the data this will include

- plot a histogram of the continuous variables in the data
- box plot of continuous variables in the data
- density plot of the numeric variables
- correlograms between variables

3.4 Dimensionality reduction using KPCA with Mahalanobis distance

Principal Component Analysis (PCA) is one the most used tools for extracting useful information from otherwise confusing datasets. PCA provides a way of reducing complex data sets to lower dimension to reveal hidden relationships. One shortcoming of PCA is that it assumes linearity and is thus restricted to expressing the data as a linear combination of its vectors. One way around this limitation is to make use of the Kernel trick. Kernelized principal component was one of the earliest applications of the kernel trick. Kernel PCA performs PCA in a mapped high dimensional inner product space instead of the input space.

KPCA

Principal component analysis is a dimensionality reduction technique that solves the eigenvalues and the corresponding eigenvectors of the matrix then, by considering the eigenvalues makes a linear combination of the eigenvectors of the matrix which extracts the main components of the data (Wang et al., 2012). Kernel principal component analysis extends this principal over a non-linear space. KPCA works by mapping the data from the input space to a feature space via a nonlinear mapping function then PCA is applied on the vector representation of the data (Wang et al., 2012)

Mahalanobis Distance Suppose we have two distinct groups C (credit worthy) and B (credit risky) with a number say (n) of relevant characteristics that define these two groups. These characteristics may include age, sex, income, purpose of loan, credit score, bank balance etc. We let X denote a random vector containing values of these characteristic for a given individual. Since we are interested in measuring the difference between the two groups we take the n-dimensional X vector having the same deviation about its mean for the two groups. We can therefore define the difference between the groups as being the difference between the mean vectors of X relative to the common within group variation(). This is mathematically defined as follows;

$$D^2 = (X - \mu)^T \Sigma^{-1} (X - \mu) \tag{3.1}$$

D^2 is the Mahalanobis distance squared

X is a vector of the observations for one individual

μ is the vector of the means of the characteristics

Σ^{-1} represents the inverse of the covariance matrix of the independent variables.

SVM model construction By combining SVM with kernelized Mahalanobis distance we can optimize the training data by trimming it to its most relevant components. This will optimize the classification performance of SVM which will in turn improve its prediction accuracy The SVM algorithm uses an optimal decision hyperplane to separate the data into different categories. Dataset are generally non-linearly separated. In order to solve the problem, SVM uses kernel trick concept on higher dimension space. Kernel is generally used for turn dataset on input space into feature space with higher dimensions (Putri et al., 2021) According to (Chatpatanasiri et al., 2008) by using the Kernel principal component analysis (KPCA), we can kernelize the Mahalanobis distance metric using the algorithm below;

Input

- 1 Training data: $\{(X_i, Y_i\}, \ldots \{(X_n, Y_n)\},$
- Test data : {X'}
- Kernel function: k(.,.)
- Mahalanobis Distance metric function: maha

proposed algorithm

- Apply KPCA $(k, \{X_i, \}, X')$ such that $\{X_i\} \rightarrow \{\phi_i\}$ and $\{X'\} \rightarrow \{\phi'\}$
- Apply maha with new inputs ϕ_i, Y_i to achieve$M^* or A^*$
- Use SVM to classify data into default and non-default

Performance evaluation To determine the performance of our model we will use **Confusion matrix**- for classification problems involving two outcomes the confusion matrix is two by two matrix that shows the true positives , false positives, true negatives and false negatives. we can use the confusion matrix to calculate;

- sensitivity of the model- this is the ratio of the true positive to sum of false Negatives and true positives
- Specificity of the model- this is the ratio of True negatives to the sum of False positives and True Negatives
- Precision- This is the ratio of true positives to sum of false positives and true positives

Receiver operating characteristic (ROC) curve- is a graphical representation of the trade off between the true positive rate and the false positive rate of the model.

AUC- Area under the ROC curve is a measure that represents the models ability to distinguish between the classes the higher the AUC the better the model.

References

Abdou, H. A. and Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent systems in accounting, finance and management*, 18(2-3):59–88.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609.

Altman, E. I. (1980). Commercial bank lending: process, credit scoring, and costs of errors in lending. *Journal of Financial and Quantitative Analysis*, 15(4):813–832.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6):627–635.

Bakker, D. K., Odundo, F., Nyakinda, J., and Paul, S. F. (2019). Performance evaluation criteria of credit scoring models for commercial lenders.

Bhavsar, H. and Ganatra, A. (2015). Support vector machine classification using mahalanobis distance function. *Int. J. Sci. Eng. Res*, 6(1):618–626.

Burton, D. (2012). Credit scoring, risk, and consumer lendingscapes in emerging markets. *Environment and Planning A*, 44(1):111–124.

Cao, L., Chua, K. S., Chong, W., Lee, H., and Gu, Q. (2003). A comparison of pca, kpca and ica for dimensionality reduction in support vector machine. *Neurocomputing*, 55(1-2):321–336.

Chatpatanasiri, R., Korsrilabutr, T., Tangchanachaianan, P., and Kijsirikul, B. (2008). On kernelizing mahalanobis distance learning algorithms. *Arxiv preprint. http://arxiv. org/abs*, 804.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Desai, V. S., Crook, J. N., and Overstreet Jr, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European journal of operational research*, 95(1):24–37.

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

El-Banna, M. (2015). A novel approach for classifying imbalance welding data: Mahalanobis genetic algorithm (mga). *The International Journal of Advanced Manufacturing Technology*, 77(1):407–425.

Goh, R. Y. and Lee, L. S. (2019). Credit scoring: a review on support vector machines and metaheuristic approaches. *Advances in Operations Research*, 2019.

Han, L., Han, L., and Zhao, H. (2013). Orthogonal support vector machine for credit scoring. *Engineering Applications of Artificial Intelligence*, 26(2):848–862.

Hand, D. J. and Henley, W. E. (1993). Can reject inference ever work? *IMA Journal of Management Mathematics*, 5(1):45–55.

Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541.

Huang, C.-L., Chen, M.-C., and Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, 33(4):847–856.

Jiang, H., Ching, W., Yiu, K., and QIU, Y. (2018). Indefinite mahalanobis kernel svm for credit risk evaluation. *Applied Soft Computing*.

Kenya, F. S. D. (2008). The potential for credit scoring for sme lending in kenya. *Nairobi: FSD Kenya*.

Kolesar, P. and Showers, J. L. (1985). A robust credit screening model using categorical data. *Management Science*, 31(2):123–133.

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, pages 109–131.

Putri, N., Fatekurohman, M., and Tirta, I. (2021). Credit risk analysis using support vector machines algorithm. In *Journal of Physics: Conference Series*, volume 1836, page 012039. IOP Publishing.

Reichert, A. K., Cho, C.-C., and Wagner, G. M. (1983). An examination of the conceptual issues involved in developing credit-scoring models. *Journal of Business & Economic Statistics*, 1(2):101–114.

Satchel, S. and Xia, W. (2008). Analytic models of the roc curve: applications to credit rating model validation. In *The analytics of risk model validation*, pages 113–133. Elsevier.

Schreiner, M. (2003). Scoring: the next breakthrough in microcredit. *Occasional paper*, 7.

Sun, H., Lv, G., Mo, J., Lv, X., Du, G., and Liu, Y. (2019). Application of kpca combined with svm in raman spectral discrimination. *Optik*, 184:214–219.

Sundaram, N. (2009). Support vector machine approximation using kernel pca. *Univ. California at Berkeley, Berkeley, CA, USA, Tech Rep. UCB/EECS-2009-94.[Online]. Available: http://www. eecs. berkeley. edu/Pubs/TechRpts/2009/EECS-2009-94. pdf.*

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. (1999). *The nature of statistical learning theory*. Springer science & business media.

Vidal, M. F. and Barbon, F. (2019). Credit scoring in financial inclusion.

Wang, D., Yeung, D. S., and Tsang, E. C. (2007). Weighted mahalanobis distance kernels for support vector machines. *IEEE Transactions on Neural Networks*, 18(5):1453–1462.

Wang, J., Zhou, Y., Du, X., and He, M. (2012). Personal credit assessment based on kpca and svm. In *2012 Fifth International Conference on Business Intelligence and Financial Engineering*, pages 25–28. IEEE.

West, D., Dellana, S., and Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers & operations research*, 32(10):2543–2559.

Zhang, L., Hu, H., and Zhang, D. (2015). A credit risk assessment model based on svm for small and medium enterprises in supply chain finance. *Financial Innovation*, 1(1):1–21.