

2018

Using Machine Learning Techniques to Predict a Risk Score for New Members of a Chit Fund Group

Sinead Aherne
Technological University Dublin

Follow this and additional works at: <https://arrow.tudublin.ie/scschcomdis>



Part of the [Computer Sciences Commons](#)

Recommended Citation

Aherne, Sinead (2018). *Using machine learning techniques to predict a risk score for new members of a chit fund group*. Masters dissertation, DIT, 2018.

This Dissertation is brought to you for free and open access by the School of Computer Science at ARROW@TU Dublin. It has been accepted for inclusion in Dissertations by an authorized administrator of ARROW@TU Dublin. For more information, please contact arrow.admin@tudublin.ie, aisling.coyne@tudublin.ie, vera.kilshaw@tudublin.ie.

Using Machine Learning Techniques to Predict a Risk Score for New Members of a Chit Fund Group



Sinead Aherne

A dissertation submitted in partial fulfilment of the requirements of Dublin
Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

2018

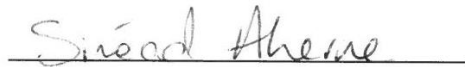
DECLARATION

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed:

A handwritten signature in cursive script, reading "Sióad Aherne", written over a horizontal line.

Date:

14 June 2018

ABSTRACT

Predicting the risk score of new and potential customers is used across the financial industry. By implementing the prediction of risk scores for their customers a chit fund company can improve the knowledge and customer understanding without relying on human knowledge. Data is collected on each customer before they have taken out credit and during the time they contribute to a chit fund. Having collected the necessary data, the company can then decide whether modelling customer risk would benefit them.

As the data is available historically, one aspect of risk score prediction will be the focus of this thesis, supervised machine learning. Supervised machine learning techniques use historic data to ‘learn a model of the relationship between a set of descriptive features and a target feature’ (Kelleher, Mac Namee, & D’Arcy, 2015). There are many supervised machine learning techniques; support vector machine (SVM), logistic regression and decision trees will be the focal point of this thesis.

The main objective of this project attempts to predict a risk score for new or potential subscribers of a chit fund company. The models generated would be suitable for use before a customer joins a chit fund group as well as while the customer is taking part in the group, measuring risk before becoming a subscriber and the behavioural risk while with the company. The objective is to extend research already carried out to predict a score from zero to one identifying the probability of default. Default, for the purpose of this project, is defined as being more than 90 days late with a payment. The data of real chit fund subscribers was used to train and test the models built for the project. A factor reduction technique was used to identify key variables, and multiple models were tested to determine which gives the best results.

The second objective of this project will look at the subscriber network. This section of the project will check for links between subscribers, and investigate a possible link between subscribers and their chance of default. Variables such as address and nominee will be the focus in this section.

The most successful supervised machine learning model was the random forest model with precision of 59% and recall of 92%. Accuracy for this model was the highest of each of the models in the experiment at 85%. However, this is not the most trustworthy evaluation measure for this project as the dataset is unbalanced.

A combination of 300 decision trees were applied in this model. Using the classification method, the class that was predicted by the majority of trees was selected as the final prediction. This achieved high accuracy of the dataset from the chit fund company, Kyepot. Social network analysis found that there was no unusual relationship between subscribers that went into default with regards to the area in which they live or their nominees.

Supervised machine learning techniques have been shown to be a useful tool in the financial industry. This project suggests that these techniques may also be useful tools for chit fund companies. This project evaluates four different techniques suggesting the random forest technique is the most useful for this chit fund company.

Key words: Chit funds, credit risk, decision tree, random forest, logistic regression, support vector machines.

ACKNOWLEDGMENTS

I would like to thank my supervisor Brian Leahy for the support, guidance and advice throughout this project.

Also, I would like to thank my family and friends for their support and encouragement, especially my parents John and Sadie, and my boyfriend Alan, for taking the time to proofread this project.

TABLE OF CONTENTS

DECLARATION	i
ABSTRACT	ii
ACKNOWLEDGMENTS	iv
TABLE OF CONTENTS	v
TABLE OF FIGURES.....	viii
TABLE OF TABLES	ix
1. INTRODUCTION.....	1
1.1 Background	1
1.2 Research Project.....	3
1.3 Research Objectives	4
1.4 Research Methodologies	5
1.5 Scope and Limitations.....	6
1.6 Document Outline	6
2. LITERATURE REVIEW AND RELATED WORK.....	8
2.1 Background	8
2.2 Chit Funds	8
2.3 Credit Risk	10
2.4 Data Exploration and Pre-processing.....	12
2.4.1 Class Imbalance Problem	13
2.4.2 Feature Reduction	14
2.5 Machine Learning Techniques.....	16
2.5.1 Support Vector Machines	17
2.5.2 Logistic Regression	19
2.5.3 Decision Trees	21
2.6 Social Network Analysis.....	23

2.7	Financial Sector	24
2.8	Conclusions.....	25
3.	DESIGN AND METHODOLOGY.....	27
3.1	Data	28
3.2	Data preprocessing.....	30
3.2.1	Dummy Variables.....	30
3.2.2	Feature Reduction.....	31
3.2.3	Sampling.....	31
3.3	Models.....	32
3.4	Evaluation	34
3.5	Social Network Analysis.....	35
3.6	Software	36
4.	IMPLEMENTATION AND RESULTS	37
4.1	Data Understanding	37
4.2	Data Pre-processing	42
4.2.1	Feature Reduction.....	43
4.2.2	Sampling	45
4.3	Dataset Description	45
4.4	Machine Learning Models	46
4.4.1	Decision Tree.....	47
4.4.2	Random Forest.....	49
4.4.3	Logistic Regression	50
4.4.4	Support Vector Machine.....	51
4.5	Social Network Analysis.....	52
4.6	Results.....	53

5. ANALYSIS, EVALUATION AND DISCUSSION	55
5.1 Strengths and Limitations	59
6. CONCLUSION	61
6.1 Research Overview	61
6.2 Problem Definition.....	62
6.3 Design/Experimentation, Evaluation & Results	63
6.4 Contributions and Impact.....	64
6.5 Future Work and Recommendations.....	65
7. BIBLIOGRAPHY	66
8. APPENDICES	73
A. Dataset Descriptions	73

TABLE OF FIGURES

Figure 2.1 Phases of the CRISP-DM Process Model for Data Mining (Wirth & Hipp, 2000)	13
Figure 2.2 W is the hyper plane separating the two classes with maximum margin (Ye, Zhang, & Law, 2009).....	17
Figure 2.3 Logistic regression formula (Hamed, Li, Xiaoming, & Xu, 2013).....	19
Figure 2.4 Decision Tree Diagram (Purdila & Pentiu, 2014).....	22
Figure 3.1 Outline of the research methodology	28
Figure 3.2 Dataset Overview	30
Figure 4.1 Categorical variables describing the subscribers	38
Figure 4.2 Numerical variables describing the subscribers	38
Figure 4.3 Details of Missing Values	40
Figure 4.4 Subscribers enrolling per year.....	40
Figure 4.5 Monthly instalment amounts paid by Subscribers	41
Figure 4.6 Average Annual Income.....	41
Figure 4.7 Cross Validation Error	48
Figure 4.8 Pruned Decision Tree	49
Figure 4.9 Graph to select the optimal mtry value	50
Figure 4.10 Default Subscribers Grouped By Address	52
Figure 4.11 All Subscribers Grouped By Address	53
Figure 5.1 Receiver Operation Characteristic (ROC).....	57
Figure 5.2 Size of Decision Trees in Random Forest Model	57
Figure 5.3 Variable Importance in the Random Forest Model	58

TABLE OF TABLES

Table 3.1 Confusion Matrix.....	34
Table 4.1 Displays the number of subscribers who have ever had a payment over 90 days late.....	42
Table 4.2 Categorical Variables	44
Table 4.3 Final Dataset Description	46
Table 4.4 Decision Tree Evaluation Results.....	47
Table 4.5 Pruned Decision Tree	48
Table 4.6 Evaluation of Tuned Random Forest Model	50
Table 4.7 Tuned Logistic Regression Results	51
Table 4.8 SVM Results.....	51
Table 4.9 Results of Supervised Machine Learning Models Built.....	53
Table 4.10 Results of Supervised Machine Learning Models Without Imputing Missing Values	54
Table A.1 Description of Subscriber Dataset	73
Table A.2 Description of Auction Dataset	74
Table A.3 Description of Group Dataset	75
Table A.4 Description of Transaction Dataset	76
Table A.5 Description of Tender Dataset	77

1. INTRODUCTION

1.1 Background

Before formal financial institutions were established in India, there was a great reliance on moneylenders to such an extent that it prompted a political commitment to introduce a structured financial industry (Tsai, 2004). The formal financial industry that was introduced in India, however, was limited to those with high incomes, and therefore those who were low risk (Tsai). This left little financial support for lower income members of the population. A recent study carried out by Mehta & Bhattacharya (2017) highlighted the fact that the support the Indian government has given to the financial sector in the country has not yet benefited the poor.

Although political moves have aimed to move towards formal finance, informal finance has not been eliminated completely. According to research carried out by Binswanger & Khandker (1995), formal finance in India expanded considerably in the 1970's to the rural communities to provide a greater availability of credit to the poor, most of whom reside in rural India, and alleviate poverty and reduce reliance on informal finance such as moneylenders. Informal finance in India, however, comes in many forms. This thesis focuses on one type; chit funds. Chit funds are a saving and lending scheme exercised in India. An all India Credit Survey carried out in 2000 to 2001 however, has shown a rise in informal finance compared to a decline in the 1990's (Jones, 2008). Both government supported financial solutions and non-governmental organisations such as chit funds have become more popular as a form of microfinance for the lower income community (Yusuf, 2014).

Banerjee, Ghosh, & Roy (2010) collected primary data from surveying two rural villages for their research. Socio-political and economic data was collected from each household surveyed. The research suggests that loans for non-production purposes are more difficult for villagers to obtain from formal financial institutions, therefore forcing them to fall back on informal finance solutions.

Chit fund organisations are currently assessing the credit risk of potential new group members manually through domain knowledge. This adds unnecessary risk to the company, as valuable employees who have developed significant knowledge and skills are liable to leave the organisation. This thesis will assess and compare the following machine learning techniques; decision trees, random forest, SVM and logistic regression. This will then lead to the identification of which technique provides the model with highest accuracy, precision, recall and specificity for the risk score prediction of a chit fund member. In turn this would alleviate the risk of employee churn, significantly damaging the domain knowledge held by the company.

The prediction of the credit risk of a potential customer has been implemented to a high standard in previous research using formal financial data. Both real world data and synthetic data have been used to build and evaluate machine learning techniques. Research by Avery, Brevoort, & Canner (2009) suggests credit scoring is widely used as a decision-making aid and has increased the availability of credit and benefits companies as they increase in efficiency.

As mentioned, credit risk scoring has been widely used in the financial industry, but risk scoring has also been used in other sectors. The health sector is one example of this. Mehran et al. (2010) used logistic regression in their research to predict the risk of major bleeding in patients with acute coronary syndromes. Another area of research using similar supervised machine learning models is that of customer churn prediction. Milošević, Živić, & Andjelković, (2017) used similar techniques such as logistic regression, random forest and decision trees also in their research to predict the customer churn.

Social network analysis will also be carried out on the chit fund data. The relationships between customers and their likelihood to go into default will be analysed. Using a bipartite network, the relationships between subscribers can be visualised. This will allow the credit risk score to be looked at by means of connections. A bipartite network contains two sets of vertices representing different features; there is no connection between vertices of the same type only between those of different type. Similar to research by Godlewski, Sanditov, &

Burger-Helmchen (2012) which focused on another bank loan network, the network would provide further insight into the subscribers of the chit fund company.

1.2 Research Project

Supervised machine learning techniques have been used in many credit scoring problems in the past with SVM techniques seen to be one of the most common in credit risk prediction. Trustorff, Konrad, & Leker (2011) suggested that SVM outperforms logistic regression. While Chen, Härdle, & Moro (2011) also found that the SVM method is best for forecasting default probabilities, they also discovered that it does not offer much understanding or significant insight into the data. Other research suggests that both logistic regression and SVM perform better than other models such as neural networks and C.45 decision tree (Perols, 2011). Based on the previous literature in this area and for reasons mentioned further on in this section, four supervised machine learning techniques will be compared when aiming to predict credit risk, the four techniques are decision tree, random forest, logistic regression and SVM.

Currently a risk score for a chit fund member is assessed manually. To automate this process, four supervised machine learning techniques have been chosen to test their predictive power in relation to this credit risk problem. Decision tree was selected as a base line model. The decision tree can identify the importance of variables or how informative the variables are as predictors and clearly shows the process in how the prediction was made (Kelleher, Mac Namee, & D'Arcy, p. 122). Random forest was then chosen as a natural follower of the decision tree. It is assumed the random forest will generate a better result than that of the decision tree technique as many decision trees will be implemented in the random forest model. As seen in research by Sharma (2011) the random forest technique out performed logistic regression when predicting a credit risk score. Logistic regression has been shown to be powerful in previous research and still has a level of transparency (Peng, Wang, Kou, & Shi, 2011). SVM is commonly used and has been a successful model in previous studies (Chen, Härdle and Moro) (Boyacioglu, Kara and Baykan).

Social network analysis will be carried out to ascertain if there are any unseen connections between the subscribers of the chit fund company who have gone into default during the period of the chit fund. Subscribers will be grouped in terms of residential area, gender and chit amount.

The research question asks *Which supervised machine learning technique; decision tree, random forest, logistic regression or support vector machine; can best predict the risk score of a chit fund member with best accuracy, precision, recall and specificity?*

1.3 Research Objectives

The main objective of this project is to identify the model that best predicts the risk of potential chit fund members as they apply to become part of a chit fund. While carrying out this research, it is expected that further insight into the factors that affect the risk score of a customer will be gained. Eliminating the need for complete manual assessment of new chit fund members would prove beneficial for both customers and the company providing the data for this project. The process would become more efficient, providing an answer for the customer much sooner.

To complete this project, firstly a review of the research previously carried out in this area will be undertaken. Data exploration will then be carried out on the data received from the Indian chit fund company. This will involve joining the data, as it currently stands, it is in five separate tables with common identifiers, cleaning the data and implementing a feature reduction technique as there are over 100 attributes in the chit fund data. Each supervised machine learning model will be trained using 66% of the dataset. The models will be tested using the remaining 34% of the dataset. This same split has been used in previous research by Yang (2007), Kruppa, Schwarz, Arminger, & Ziegler (2013) and Boyacioglu, Kara, & Baykan.

Evaluation will be performed on each model calculating the accuracy, precision, recall and specificity rates. A Receiver Operating Characteristic (ROC) curve for each model will also be evaluated. Selecting the model with highest overall predictive power as the most

valuable, a conclusion will be made, and the hypothesis will be evaluated. Finally, the limitations and further areas of research will be identified and listed.

A second stream of analysis will be carried out to identify relationships between the subscribers of the chit fund. This will aim to further explain some of the reasons the subscribers may go into default. Some relationships among subscribers may not be visible from examining machine learning models. Social network analysis will aim to uncover relationships between the subscribers who have had a late or missing payment previously.

1.4 Research Methodologies

The project aims to answer the research question listed in section 1.2.

Secondary research will be carried out using historic data from an Indian chit fund company. Quantitative methods, an experiment, will be designed and carried out and used to evaluate the hypothesis. The Cross-Industry Standard Process for Data Mining (CRISP-DM) process will be followed throughout the experiment. Machine learning algorithms will be used to build supervised machine learning models to predict a credit score for members of a chit fund group. The research carried out will be empirical research, involving testing the hypothesis with an experiment. The aim of the research is to carry out an experiment to identify the model that best predicts the risk score of chit fund members. Deductive reasoning will be used in this research.

The theory that SVM will have the highest accuracy rate has evolved from the literature review. The hypothesis is based on this theory. Each model will be created and evaluated by calculating four evaluation measures, accuracy, precision, recall and specificity. The ROC curve will also be used for the evaluation of the models. The results will be compared and the model with the highest predictive power will be selected as the ‘best’ model.

1.5 Scope and Limitations

The scope of this research is machine learning techniques predicting a credit risk score for members of a chit fund in India. One limitation of the project is the data, which is provided from one chit fund company only, and contains a subset of their customer base. This may not be representative of other chit fund companies.

As the dataset is provided by one sole chit fund company, it is possible that other chit fund companies and other informal financial companies would have a very different customer base and therefore the research carried out on the data provided by this chit fund company would not be at all representative of other chit fund companies.

1.6 Document Outline

This section outlines the thesis document.

Chapter Two discusses the literature related to predicting the risk score of customers. As supervised machine learning techniques have previously been used to predict this risk, chapter two reviews and compares previous work done in this area. It describes the use of SVM, logistic regression, decision tree and random forest in previous work, the advantages and disadvantages of each technique, and discusses what was suggested as the most valuable technique. Previous works are also compared against this project and the data available. It also looks at research carried out on the financial sector in India and the part chit funds have to play in it. This chapter also discusses the use of social network analysis research previously carried out on financial data.

Chapter Three describes how the dataset was prepared and cleaned and how the experiment was carried out and methodology used to evaluate the experiment. The software used throughout the experiment is listed and the evaluation techniques that were used are also discussed.

Chapter Four discusses the implementation and results of the experiment. The data pre-processing and exploration that was carried out along with the experiment is included. The hypothesis of the project is considered, and results of the experiment are compared to evaluate the hypothesis.

Chapter Five analyses the techniques used in the experiment and data pre-processing including the feature reduction technique. The variables used in each model are described and the results gathered from each model are evaluated and compared.

Chapter Six is the final chapter containing the conclusions of the thesis, a summary of the work undertaken and discussions of future work that may be carried out.

2. LITERATURE REVIEW AND RELATED WORK

2.1 Background

In the financial sector, understanding the potential risk of a possible new customer can be a powerful tool and can enable a company to make informed choices before funding a loan to a new customer. Many previous papers have highlighted advantages and disadvantages of different machine learning techniques which have been tested in relation to this potential risk. In work carried out by Crook, Edelman, & Thomas (2007), a review was executed on different credit scoring techniques. This paper noted that risk assessment is common for a financial institution before and during a loan. The paper concluded that the most frequent classifier used was logistic regression although many other techniques had been tested. However, in research carried out by Boyacioglu, Kara and Baykan, SVM outperformed other techniques evaluated and Twala (2010) found that decision trees and ensemble methods also provide a high accuracy measure.

2.2 Chit Funds

Chit funds are Indian saving and lending companies. They offer a saving scheme where members of a group pay equal monthly instalments to a pool. Each month the pool of money is won by one group member. They are based in India and often provide finance to the rural communities in India. Chit funds for the purpose of this study are conducted by a company called Kyepot.

In research by Satish (2001) the main use of chit funds is described as providing subscribers benefits from a monthly pool of money. They are described as being flexible with the terms of each chit fund group; the amount of the monthly instalment varies with the affordability each member of the group has. It is also mentioned that the groups tend to have a large number of subscribers which would in turn provide a large lump sum each month, but again this is flexible. It has been noted that members of a chit fund group integrate more freely into the self-help groups set up by rural bank managers, as the act of monthly saving has already been established and become normal for the members.

Santhisree & Prasad (2014) suggests that there are approximately 30,000 registered chit fund companies throughout India. Their research suggests that chit funds typically provide credit for the middle-income demographic. However, recently this has changed, and the companies now cater for the lower-income demographic. Chit funds supply credit to members of the community who are unable to secure credit from banks. This research identifies a number of problems faced by chit fund companies. One such problem is the lack of regulation of business. There is a high proportion of chit fund companies that are unregistered, which creates unhealthy competition for the registered companies. Another issue identified in the research carried out by Santhisree & Prasad was that the unregistered companies can pay-out funds to subscribers much quicker than the registered companies (immediately, compared to 15 to 30 days for registered companies). This indicates a need for a quicker decision-making process in registered chit fund companies.

Tsai (2009) suggests in their research that informal finance has existed for many years and this may have been as a result of the lack of formal finance in some parts of India, especially rural India. The paper states that although good intentions of bringing formal finance to all of India was present in the 1950's, this was not completely successful therefore leaving a gap for the informal financial industry such as chit funds to continue to operate. The formal financial institutions that did operate throughout India did not reach all rural areas. There are many types of informal finance mentioned in the research carried out by Tsai. In it, chit fund companies are described as 'Rotating savings and credit organizations (ROSCAs)—indigenously organized savings and credit groups'. Tsai concludes that informal finance does not necessarily out-do formal finance but for those out of reach of formal finance, it is an imperfect substitution.

In a study carried out by Yusuf, the use of rural finance to alleviate poverty is the focus. It is noted that the majority of the poor in India live in rural areas. It has been found that the agricultural sector in India is growing and keeping up with the increasing food demands. The access to finance for these small entrepreneurs is most efficient by way of microfinance services. Access to formal finance, as mentioned by Yusuf, is costly. The people from the rural areas of India are seeking alternative ways to invest and save. Chit funds is one of the solutions found. As mentioned by Tsai and again by Yusuf the inability of the banking

institutions to provide their service to the rural areas of India has increased the dependency on microfinance.

Research by Rao & Buteau (2018) suggests that the structure of chit fund groups is similar to that of formal banking, and that by using the service to save and access credit, the chit fund is carrying out the same purpose as formal financial institutions. Due to these similarities it was suggested that chit funds face a similar level of credit risk as modern banking. Rao & Buteau use the regression technique to predict chit fund subscribers behaviour and found that the model may assist manual decision-making. The research did not aim to create a model to eliminate the manual decision-making step in the process of accepting a new subscriber. The research highlighted the fact that the data required to build a predictive model must be of high quality, and this was not found to be the case for the dataset used in Rao & Buteau's research.

2.3 Credit Risk

Credit risk score is a powerful tool used by many financial institutions as an extra check of credit worthiness of a potential borrower and as a method of managing current customers (Avery, Brevoort, & Canner. Credit risk scoring is carried out using a statistical model that aims to predict the likelihood of an account going into default. It was also noted in research by Xanthopoulos & Nakas (2007) that credit scoring now not only covers a customer's risk of default at application, but can be monitored throughout the loan duration. Credit risk can improve the decision-making process by enabling a faster turnover of decisions and reducing the risk of manual error (Dahiya, Handa, & Singh, 2015). As mentioned in research by Santhisree & Prasad, credit risk scoring may improve the turnaround time for a decision when applying for credit. They found that registered chit fund companies had been losing out on potential customers, due to the length of time it took the company to decide if the potential subscriber was credit worthy. Once again, the importance of credit scoring is mentioned in work by Sohn & Kim, (2012). Their research consisted of examining the use of decision trees for credit scoring, and found that the decision tree technique provided a

higher accuracy than that found in previous research on the same dataset using logistic regression.

Avery, Brevoort, & Canner examined the effect of demographic data on a borrower's credit risk score and its influence on the borrower's access to credit, suggesting that the demographic data had an effect on the overall credit score given to an individual. Age was one of the demographic variables seen to affect credit score; in particular a younger subscriber was seen to be given a lower credit risk score. It was suggested in this study that credit scoring increases a person's access to credit. It also increases efficiency for the financial institution and reduces the likelihood of a staff member being biased towards a personal characteristic of the potential customer. This study found that the credit risk score was a good indication of whether a borrower would go into default during the loan term. As the credit risk score decreased the percentage of bad rate increased.

Dahiya, Handa, & Singh also used demographic data in their research when building classifiers to implement credit scoring. The research examined the use of ensemble classifiers in credit risk scoring. By using several classification techniques, it was found that the ensemble method outperformed the single use of each of the classification techniques tested. The research suggests using ensemble classifiers as the credit score model would be advantageous to a company, improving and reducing time on decision making.

Edelman (2008) noted in his study, similar to Avery, Brevoort, & Canner, that credit scoring is used worldwide. It is used both for potential new borrowers and in behavioural scoring to manage existing customers. In research by Edelman, it was suggested that credit scoring is used not only in the application for credit, but can also be used in marketing to understand the likelihood of a response to different marketing campaigns.

Carling, Jacobson, Lindé, & Roszbach (2007) suggest in their research that macroeconomic variables have a strong predictive power when building a model to predict credit risk. Using credit application data from a Swedish bank and data from the credit bureau, it was found that the risk of default was considerably higher for a loan with a short term compared to a loan over a longer term. They suggested that taking macro conditions into account in the models built helps to better predict the risk of default of firms.

Research carried out by Yap, Ong, & Husain, (2011) focuses on the use of credit scoring models to improve credit assessment techniques. Their research notes the usefulness of credit scoring in the financial industry aiding the decision-making process. The data used in this research was not as highly unbalanced as seen in other research in this area, 35% of the dataset consisted of defaulters. After comparing regression and decision tree as two techniques for credit scoring it was found that there was no clear ‘best’ model. However, each had advantages and disadvantages. As the decision tree is easily understood, it was concluded that this technique would be chosen as the ‘best’ model. Also noted in this research was the data quality issues in the dataset, which may make the results of the supervised machine learning models build unreliable. Poor data quality is a common issue found in credit scoring research.

Many different credit risk models have been evaluated in previous research. Some of the machine learning techniques that have been used are described in more detail further in this chapter. In the previous research described, the advantages of credit risk scoring in formal financial institutions have been listed. Due to the similar structure of chit funds to formal finance, as mentioned in research by Rao & Buteau, it suggests that same advantages may be found if credit risk scoring was implemented in chit fund companies.

2.4 Data Exploration and Pre-processing

To gain further understanding of the business problem, data exploration is required. The CRISP-DM methodology has been widely accepted as a methodology that can be followed in machine learning and data mining research. The CRISP-DM methodology ‘provides an overview of the life cycle of a data mining project’ (Wirth & Hipp, 2000).

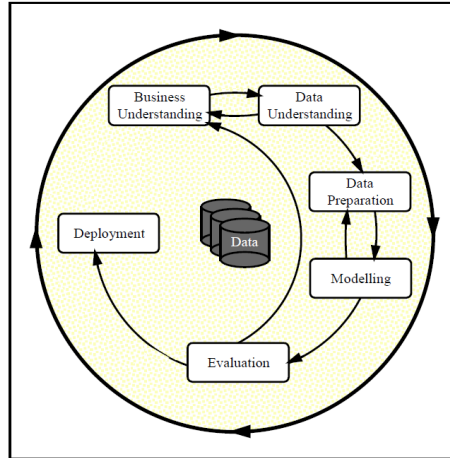


Figure 2.1 Phases of the CRISP-DM Process Model for Data Mining (Wirth & Hipp, 2000)

Wirth & Hipp go on to suggest that this model is a valuable tool in planning and documentation, following the model allowed for open communication and structured documentation of the work carried out. To enable the project to be repeated, the CRISP-DM model was deemed useful by Wirth & Hipp.

Data pre-processing, prior to fitting a machine learning model, is vital to improving the performance of the model. Pre-processing the data can improve the quality of the data, and therefore, the accuracy of the models will increase. Much of real world data collected will contain some inaccuracies or ‘bad’ data, so when relying on the quality of the data, this step proves very important.

2.4.1 Class Imbalance Problem

As seen in research by Maheshwari, Jain, & Jadon (2017) class imbalance within a dataset has become a challenge in data mining. This is a common problem within the risk score prediction area. As expected, a small number of cases in a financial institution default when compared to those who don’t, thus creating an unbalanced class.

One approach to deal with this is sampling the data. Under sampling and over sampling are two techniques analysed in the research undertaken by Maheshwari, Jain, & Jadon; under

sampling the dataset risks excluding valuable data from the dataset, while over-sampling carried the risk of overfitting the model. Another sampling method suggested by Farquad & Bose (2012) is Synthetic Minority Over-sampling Technique (SMOTE). After comparing under sampling, over sampling and SMOTE techniques, it was concluded that the SMOTE technique yielded the best results for SVM, logistic regression, multi-layer perception and random forest. SMOTE creates synthetic instances based on data provided to even up the class imbalance within the dataset, while other research suggests over-sampling is a valid and worthwhile method to significantly improve the performance of learning algorithms (Sáez, Krawczyk, & Woźniak, 2016).

Daskalaki, Kopanas, & Avouris (2006) focus their research on evaluating classifiers with an unbalanced class dataset. Their research found that an unbalanced class dataset when used with a classification algorithm may seem like the algorithm performed with a high accuracy, but at closer analysis it can be seen that all the minority class cases achieved zero accuracy. Maheshwari, Jain, & Jadon also found that accuracy is no longer the proper measure when an unbalanced dataset is in use. It is suggested that, rather than evaluating a model on accuracy alone to separately monitor true positive and true negative rates, the class distribution of the training dataset was said to influence the performance of a model (Daskalaki, Kopanas, & Avouris).

Brown & Mues (2012) carried out research to evaluate different classification algorithms when used on imbalanced datasets. Using real world datasets, ten classification techniques were compared. It was found that random forest outperformed the other techniques tested.

2.4.2 Feature Reduction

‘Feature selection is the most basic step in data processing’ (Van-Sang & Ha-Nam, 2016). It has been noted in previous research that feature selection can be a valuable step in credit risk prediction (Tsai). Feature selection helps to ‘reduce the number of descriptive features in a dataset to just the subset that is most useful’ (Kelleher, Mac Namee, & D’Arcy, p.230). Not all features within a dataset will contain any predictive power. Limiting the dataset to the most valuable and useful features can reduce the expense and computing time of an algorithm. There are many feature selection techniques, one such technique is to rank and

prune the features, which involves ranking each feature on their usefulness to predict the outcome variable and select the features with highest ranking.

Another feature selection method is Principal Component Analysis (PCA). This is an unsupervised feature selection technique. PCA reduces features, while retaining majority of value within the original dataset. This method has been shown to improve results of classification models (Doumpos & Zopounidis, 2007). Doumpos & Zopounidis carried out research on credit risk assessment with model combinations. One of the three datasets used in their experiment was unbalanced with 218 out of 1,193 firms classed as being in default. This is seen in many finance datasets. The unbalanced default class is expected. Using PCA in this experiment, it was noted that all variables with even a small predictive power should be kept in the dataset. Depending on the dataset, this result may not improve the feature selection problem in a project. PCA was also chosen as a pre-processing method in research by Boyacioglu, Kara, & Baykan. This work used machine learning techniques to predict bank financial failures. PCA was the tool used on the twenty financial ratios, and a threshold of 70% variance was set. The selected variables must explain at least 70% of the variance. Seven out of twenty factors were retained showing the feature reduction possibilities using PCA.

In a study by Huang et al. (2016), a number of data pre-processing methods such as feature selection and instance selection, with and without imputing missing values, were compared. From this analysis, it was concluded that for the datasets in the research, a classification model performs best when the data pre-processing using feature selection does not include imputing the missing values. Huang et al. granted a 10% or below missing value rate to allow a variable to be included in the experiment. However, in work carried out by Kelleher, Mac Namee, & D'Arcy (p. 73) a 60% missing value rate is noted. With real world data and the data quality issues throughout the data collection and storage process, the 60% rate seems more accessible than the significantly lower rate of 10% in this experiment. Instance selection was also tested during this experiment. Removing instances similar to under sampling would increase the risk of excluding valuable information from the models.

Van-Sang & Ha-Nam carry out both Wrapper methods and Filter methods on credit approval data from Germany and Australia. In Wrapper feature selection, each subset of features is evaluated, and the best subset of features is selected. Once the subset of features has been selected, then the classifier is tested on the test data. The Filter method selects features based on their ability to predict the target variable. The datasets used in the experiment were both relatively small in size containing just under 1,700 applications when combined. The research aimed to reduce the number of attributes used in the models, saving time and improving accuracy rates while maintaining the integrity of the data. Similarly, in research carried out by Feki, Ishak, & Feki (2012), the Wrapper method was analysed using both simulated and real-world data. Li & Sun (2011) also focus on the Wrapper technique in their research, specifically the forward feature selection method. This was carried out by ranking the data and using forward selection of features. This was found to be helpful in the research as the other option of using domain knowledge to select the best features for the model was too time consuming and expensive for the research.

Another feature selection technique that has been used in credit risk prediction is to examine the correlation between independent features in the dataset. This filter feature selection technique begins by selecting the threshold of correlation that will be allowed between features. This is mentioned in research carried out by Yu & Liu (2003), where they state that the ‘goodness’ measure must be defined. The correlation between pairs of features will be compared to the threshold to decide if a feature would be redundant in a model. This feature selection method, they found, is efficient and effective when dealing with a dataset with a large number of features.

2.5 Machine Learning Techniques

A number of machine learning techniques have previously been used in similar credit risk prediction problems.

2.5.1 Support Vector Machines

SVMs have been a common technique used in previous risk score research. The SVM technique is a supervised machine learning technique. ‘Support Vector Machines (SVMs) transform the input vectors nonlinearly into a high-dimensional feature space through a kernel function so that the data can be separated by linear models.’ (Ribeiro, Silva, Chen, Vieira, & Carvalho das Neves, 2012).

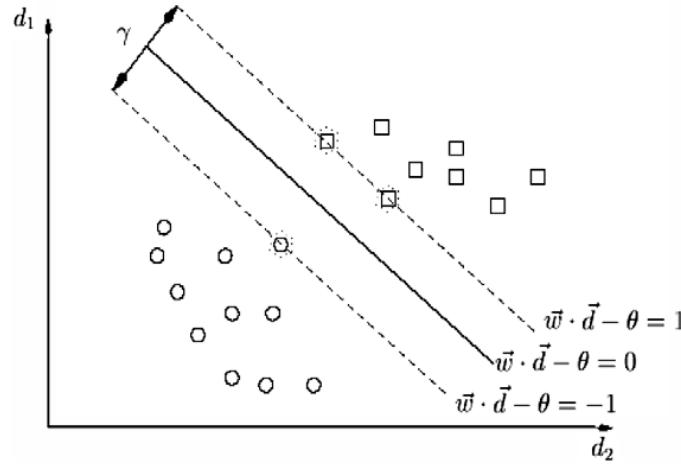


Figure 2.2 W is the hyper plane separating the two classes with maximum margin (Ye, Zhang, & Law, 2009)

It has been seen in previous research that SVMs tend to perform with the highest accuracy when compared with other models (Chen, Härdle and Moro) (Boyacioglu, Kara and Baykan).

Ribeiro, Silva, Chen, Vieira, & Carvalho das Neves focused on models which are large margin classifiers. The research focused on the SVM technique and compared four models all including an SVM technique. It was concluded that SVMs provide highly accurate results when used on financial datasets. This study could be enhanced by allowing the same dataset to be tested on other highly regarded machine learning techniques, such as logistic regression.

Chen, Härdle and Moro suggest in their research that SVMs provide high accuracy and low misclassification errors when predicting risk scores. Their research is based on real-world

data with a dataset containing 21,000 cases. Comparing SVM and logit models, the training data used was sampled from the data between 1997 and 1999, while the validation dataset was sampled from the 2000 to 2002 data. There is a risk that this training dataset may not be representative of the whole dataset. Financial datasets vary from year to year, going through highs and lows in the economic world. This may hinder the models' ability to perform accurately on more recent data.

Similarly, Boyacioglu, Kara and Baykan found that SVMs outperformed other techniques such as neural networks. However, in this research the training and validation datasets are divided using stratified sampling from the entire dataset. This creates a higher chance of the sample being representative of the entire dataset. Comparing artificial neural networks, SVMs and statistical methods, it was found that SVMs and learning vector quantization (LVQ) provided satisfying prediction results. LVQ is a classification algorithm, that can be trained as supervised or unsupervised, supervised for obtaining classifiers and unsupervised to identify clusters (Nova & Estévez, 2014).

A paper by Trustorff, Konrad and Leker considered two techniques for predicting credit risk score, least-squares SVMs and logistic regression. Using the ROC curve, it was found that SVMs surpass logistic regression when the sample is small. The ROC curve plots the false positive rates versus true positive rates and provides a visual where comparing various machine learning models is quick and easy. The dataset consisted of 31,049 instances. This is a much larger sample than that used in the research by Boyacioglu, Kara and Baykan. Trustorff, Konrad and Leker found that logistic regression models performance was affected more than the performance of the SVM.

Ye, Zhang & Law used the SVM technique to classify the sentiments of online customer reviews, while comparing this technique to Naïve Bayes algorithms. It was concluded that the SVM technique outperformed Naïve Bayes. To evaluate the machine learning models the common measures of recall, precision and accuracy were calculated from the contingency table. The recall measure is calculated by taking the number of true positive cases divided by the total number of actual true cases. The number of true positive cases relates to the number of instances in the positive class the model predicted correctly. The

number of actual true cases relates to the number of instances in the class; this information is known from the dataset. Precision is calculated by dividing the number of true positive cases by the total number of positively predicted cases. Accuracy is calculated by dividing the number of correct predictions by the total number of predictions made. The ROC curve may aid the evaluation of some machine learning algorithms also. This tool was not used in this research.

2.5.2 Logistic Regression

Logistic regression has also been found to be a favourable technique in credit risk prediction. The formula in figure 2.3 below represents logistic regression where p_i is the probability and x_i is the independent variables which predict the outcome p_i .

$$p_i = \frac{e^{x_i\beta}}{1 + e^{x_i\beta'}}$$

Figure 2.3 Logistic regression formula (Hamed, Li, Xiaoming, & Xu, 2013)

In a study on classification algorithms for financial risk prediction, it was found that linear logistic regression was ranked in the top three classifiers of the machine learning models (Peng, Wang, Kou, & Shi). Using six real-world datasets from six countries, Peng, Wang, Kou, & Shi evaluated many machine learning techniques. It was found that linear logistic regression was among the top three classifiers. It is suggested in the research that evaluating a classification model based on one or two measures is not reliable. One measure may suggest a certain model is the most useful where another measure may select another model. The process model followed in the research is similar to the CRISP-DM cycle. However, it also follows some of the Knowledge Discovery in Database process.

Zaghdoudi (2013) aimed to predict bank failures using binary logistic regression. Data regarding fourteen Tunisian banks are used for this study. From the eighteen factors provided, seven were selected for use in the model. It was found that using stepwise regression with these seven factors provided satisfactory results.

Zekić-Sušac et al. (2016) evaluated both logistic regression and neural networks when predicting company growth. There was a large number of factors in the Croatian company dataset, so factor analysis was performed to remove the redundant factors. The dataset was split into two datasets; 87.13% training and 12.87% test datasets. Compared to work by Boyacioglu, Kara, & Baykan (2009) using 66% of the dataset as training and Smeureanu, Ruxanda, & Badea (2013) using 60% of the dataset as training with the neural network technique and 80% with SMV, including over 80% of the dataset in the training set is unusual. Some further analysis with different training to test splits may have been beneficial.

In a study by Hamed, Li, Xiaoming & Xu, kernel logistic regression was used to classify videos, using this machine learning technique along with the feature reduction method of PCA to find the most valuable factors. Dealing with a large amount of data in visual data PCA was chosen as one of the reduction techniques in this research. The features were scaled to values from 0 to 1. This eliminates the risk of features with higher values dominating other features, and therefore creating bias within the model. A number of evaluation measures were used to evaluate the models. This is in line with what was suggested in research by Peng, Wang, Kou, & Shi, which allowed the models to be fairly evaluated on more than one measure. It concluded that the logistic regression technique resulted in adequate results and was easy to put in place.

Perols analysed the use of machine learning algorithms to detect fraud. This is another common problem in the financial industry and machine learning can be used in the same way as it is in credit risk prediction. Examining a number of features may help to detect fraud. Perols evaluated the performance of classification algorithms using 42 possible predictors. Ten-fold cross validation was used, ensuring data used in training was not used in testing. Ten-fold cross validation is a technique used to assess models. It was found in this research that logistic regression and SVM perform well in detecting fraud. Similar to Feki, Ishak, & Feki and Li & Sun, this research uses a wrapper method to select the features for each classifier.

Butera & Faff (2006) chose to evaluate the logistic model while developing a credit rating for private firms. As mentioned in the research, the logistic model does not require normality

of data. It was concluded that the bottom up approach to the probability of default (PD) rate was beneficial in gaining the historical PD rate. However the top down approach was valuable to obtain the credit risk assessment bases on the following year's outlook.

Kruppa, Schwarz, Armingier, & Ziegler suggest that 'Default probabilities provide more detailed information about the creditworthiness of consumers, and they are usually estimated by logistic regression'. The research carried out estimates a customers' credit risk using machine learning methods with logistic regression. The data used was not from a financial company, instead it was from a retailer offering credit on their products. Due to an increasing amount of bad debt, the company had to review their processes. No credit scoring had been previously done on the data provided by the company, and all instances in the dataset had been accepted for instalment purchase. The dependant variable of default was derived from the dataset. Some of the data was qualitative, increasing the risk that the sentiment may not have been interpreted correctly. Evaluation of the learning algorithms was carried out by creating a confusion matrix from the results. The research found that random forest outperformed a tuned logistic regression model.

2.5.3 Decision Trees

Decision trees are a common supervised machine learning technique, often used, as it provides insight into how the model made the decision. In a tree like structure the model begins with a root node and answers a series of questions to flow down the tree to a decision (leaf node). Figure 2.4 below shows a decision tree where t represents the variable tested at that node, x represents one possible feature level the variable can take, and C represents the leaf nodes and therefore the decision.

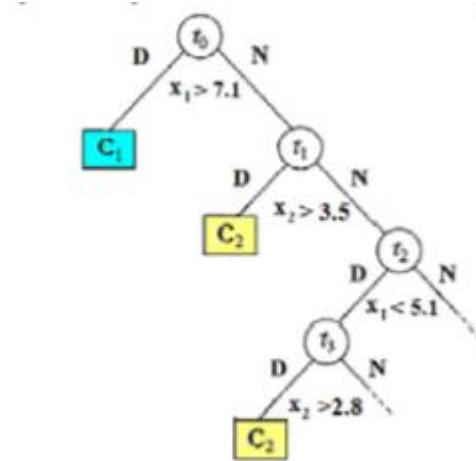


Figure 2.4 Decision Tree Diagram (Purdila & Pentiu, 2014)

Khandani, Kim and Lo (2010) used supervised machine learning to predict consumer credit risk. Using both account level transactional data, and data from the credit bureau, they aimed to reduce total default losses by 6 to 25%. It was suggested that integrating the credit bureau data with the transactional data for each customer provided more up to date details of the customers than the data that would have been sourced easily from the financial institution. As well as truncation and credit bureau data, the account balance was also integrated. A rich dataset was collected to carry out this research. Similar to Chen, Härdle and Moro, Khandani, Kim and Lo also split the training and test datasets by date. However, this approach was carried out ten times, and evaluation was then carried out on each of the ten training periods. Using generalized classification and regression trees, they found that they reached an accuracy of 85%.

Some research suggests that instance-based learning will achieve highest accuracy levels. Yang concludes that kernel learning provides better results in comparison with traditional credit scoring methods when there are many attributes and unbalanced class distributions in the data set. Yang notes that logistic regression, a traditional approach to credit scoring, needs to be built from scratch when a change is made to data or the population. With 102 factors in the dataset used for this experiment, there was no feature reduction applied. This technique will not favour the logistic regression technique.

However, in research carried out by Twala, it was suggested that the instance-based machine learning technique, k-nearest neighbour, was not achieving the highest accuracy. Instead, ensemble methods such as decision trees were found to have a high accuracy rate. Twala used four credit datasets; these datasets had much fewer features than that seen in the dataset used by Yang. Twala's datasets had at most 20 features. Considering noise in the data, it was found that the decision tree model achieved the highest accuracy. The decision tree was also considered one of the most accurate techniques when evaluating ensemble models.

Following on from the decision tree technique, it has been found in research by Sharma that the random forest technique has outperformed the logistic regression technique. Random Forest contains a group of decision trees. Combining decision trees, the random forest makes a classification prediction based on the majority vote. In a regression problem, the random forest prediction is the mean prediction of all decision trees in the model.

The random forest technique was also suggested to be the best performing model in research by Barboza, Kimura, & Altman (2017) when predicting bankruptcy. Their research used financial data and models that were built without normalising the data. It was found, similar to research by Sharma, that the random forest technique outperformed the logistic regression technique. Barboza, Kimura, & Altman used an unbalanced dataset in their research, and therefore the evaluation measures used were chosen to avoid bias in evaluation. Sensitivity and specificity were both used to evaluate the models. These measures were calculated from a confusion matrix.

The advantages of random forest listed in the research are plentiful. The model is unlikely to overfit the data, as each decision tree is built with samples of the data. The user is able to clearly see the importance and the use of the variables.

2.6 Social Network Analysis

Social network analysis evaluates and analyses the connections between certain features. It aims 'to predict the structure of relationships among social entities' (Butts, 2008). Nodes represent the social entities, and edges represent the relationships. Social network analysis is often carried out with regards to trust connections.

Godlewski, Sanditov, & Burger-Helmchen found, when analysing a bank loan network that syndicated bank loans have the ‘small world’ network properties with a short distance between them and high density in the network. The bipartite network consisted of banks that participated in syndicated loans and the events of a syndicated loan. A projection of the bipartite network was carried out to construct a network of lenders. The network was analysed as it changed over time, and it was noted that there were very few isolated nodes in each timeframe. It is suggested that small world networks improve efficiency, and betweenness centrality plays an important role for borrowing costs reductions.

Agarwal, Chomsisengphet, & Liu (2011) use social network analysis in their research on estimating the likelihood of borrowers going into default or becoming bankrupt. It was suggested that some characteristics found in a social network are related to borrower bankruptcy. It was found that a borrower is more likely to declare bankruptcy if they have moved out of the state where they were born.

2.7 Financial Sector

The machine learning techniques mentioned in this thesis have also been used in other financial sector projects. Endeavouring to predict fraud is also a common task within the financial sector. Many papers also use the support vector machine and logistic regression technique to detect fraud in the financial sector. Similar to the techniques tested in this experiment, Perols used both logistic regression and support vector machines to detect fraud and found both techniques to perform adequately.

Similarly Whiting, Hansen, McDonald, Albrecht, & Albrecht, (2012) used machine learning techniques to detect management fraud, finding ensemble methods to be powerful predictors in this area. Random forest is one example of the ensemble method examined. It was found that the random forest technique was one of the techniques which provided the highest accuracy. As seen in many datasets in this area of research, the dataset used in this research contains highly unbalanced classes. The solution found was to undersample the majority class and oversample the minority class.

Butaru et al. (2016) carried out research on the risk management of the credit card industry, aiming to use machine learning techniques to predict delinquency. Using data from six banks, analysis was also carried out on the value of each model across all six banks, the research suggests that the models are most valuable when used on the data from the bank from which they were trained, suggesting that once a supervised machine learning model was built for one financial institution, it would not be acceptable to use the same learning model in another institution.

As mentioned, research has been carried out on credit risk prediction in previous papers. However, research with regards to the risk score of a member of an informal financial group has not been completed fully. Previous research papers have examined the use of chit funds in a rural community (Tsai) (Yusuf). However, research has not been carried out for the prediction of a risk score for the use of chit fund organisations.

2.8 Conclusions

In this chapter data pre-processing methods have been defined and explained. The class imbalance problem has been evaluated, and many potential solutions have been identified (under sampling, over sampling and SMOTE). It was noted that feature reduction can be a vital step in machine learning projects. Many feature selection techniques were identified (PCA, wrapper, filter and correlation). Previous use of each of these methods was explained. It was found that for many machine learning algorithms, reducing the dimensionality of the dataset would prove beneficial to the performance of the model.

The chapter also focused on the previous machine learning models used to solve similar problems in the financial sector. Descriptions of the more popular methods were provided such as SVM, Logistic Regression and Decision Trees. The successes and failures of these techniques have been noted and discussed.

Social network analysis has also featured in this chapter as a potential method to predict customer credit score. It has been seen that this type of analysis can be beneficial in understanding the connections and similarities between customers.

Finally, the chapter finished with other financial sector problems that have used similar techniques to solve them. Fraud detection has also been seen to use SVM and Logistic Regression similar to the credit risk score papers.

3. DESIGN AND METHODOLOGY

This chapter outlines the design and methodology of the experiment that will be carried out to answer the research question of this project. The design of this experiment follows the CRISP-DM process outlined in figure 2.1. It includes an introduction to the datasets sourced for the project, as well as details of the supervised machine learning model that will be built and evaluated.

The steps below will be carried out following the CRISP-DM process:

- a. Business and data understanding will involve communication with the chit fund company providing the data. Obtaining this understanding will allow the exploration of the datasets to be carried out.
- b. Data pre-processing will involve cleaning and sampling the data and splitting the dataset into training and test datasets.
- c. Modelling, running the data through a number of supervised machine learning models (support vector machine, logistic regression, random forest and decision tree).
- d. Evaluating the models using various measures and comparing the results of each model.
- e. Using social network analysis to explore the connections between subscribers using certain variables in the dataset.
- f. Communicating the results of the thesis with the chit fund company.

The main focus of this study is to compare machine learning techniques when aiming to predict the risk score of a chit fund member.

Chit funds are informal financial institutions used by members of the population who cannot access formal finance. Each member of a chit fund group pays an agreed monthly contribution to the fund. Once the monthly subscription has been paid, it is auctioned to one of its subscribers. 70-95% of the monthly fund can be auctioned. The subscriber bidding for the lowest amount of the fund wins that amount. Each subscriber can only win the prize

once per chit fund group. The remaining balance in the pot is distributed to each subscriber. This repeats each month until all subscribers win the pot. The chit fund company charge a fixed rate each month. This information is available on the Kyepot website (<https://www.kyepot.com/how-it-works>).

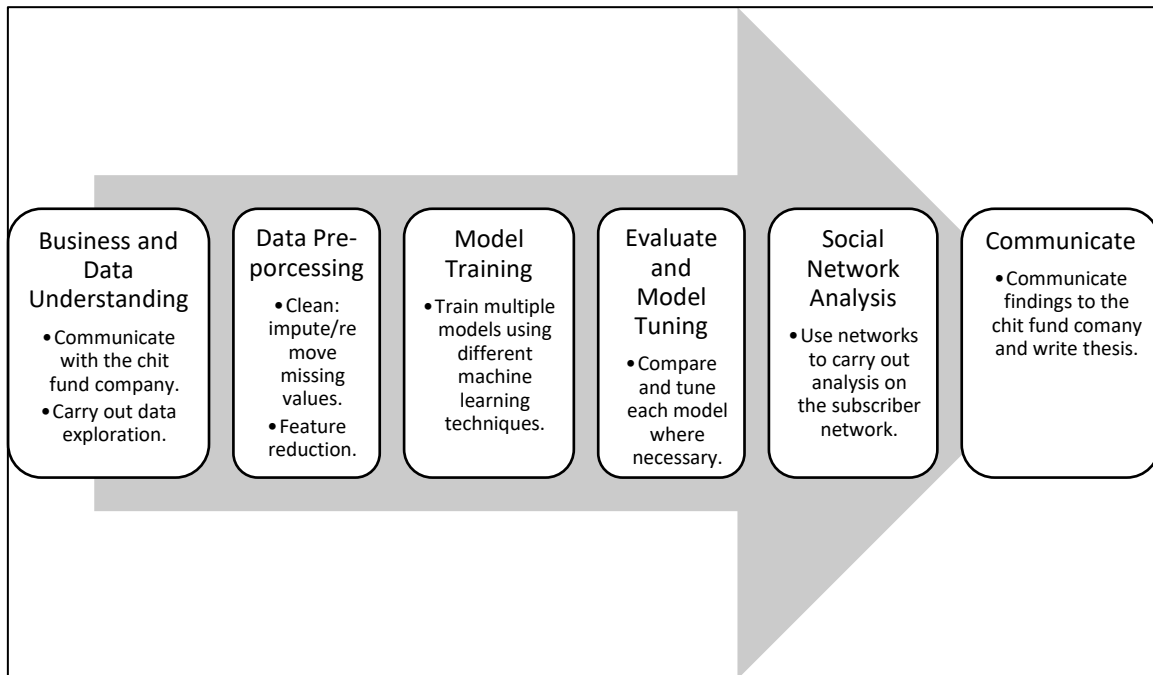


Figure 3.1 Outline of the research methodology

The thesis of this research is to demonstrate that the H0 A support vector machine achieves a higher sensitivity and specificity than that of a logistic regression model, random forest and decision tree model, when used to predict the risk score of a chit fund member.

3.1 Data

Five datasets were obtained for this study, they contained historic financial data which provided information on previous chit fund groups and the subscribers who took part in each chit fund group. ‘Historical data is a pre-requirement for any credit scoring model’ (Römer & Musshoff, 2017). The data was provided by an Indian Chit Fund company, Kyepot. The datasets provided insight into a portion of subscribers within the company, the subscribers performance of previous chit fund groups, and their lending habits. The datasets

contained information on subscribers who enrolled in a chit fund group between 1970 and 2017. The five raw datasets received were as follows; Subscriber, Auction, Transaction, Tender and Group.

The dependant variable is derived from the data provided. Auction date (the date on which the monthly instalment was due), and the payment date were used to derive a new variable detailing the number of days in default. This new variable was then used to create a 'Default' flag. Default is defined for this study as a subscriber who was greater than 90 days late with a payment. The dependant variable, 'Default', can take the values of '0' or '1', '0' meaning that the subscriber was never over 90 days late with a payment, and '1' meaning that the subscriber was over 90 days late with a payment.

The data provided included information such as employment types, income, address, age and balance of the account. It also included all details about the chit fund groups, such as the chit fund amount, who won the monthly prize, and the date payments were received. Some sensitive data was also within the datasets. Descriptions of each variable within each of the datasets is provided in Appendix A. Figure 3.2 summarizes the valuable information contained within each of the five datasets.

During the data investigation stage of the project, a number of dataset characteristics will be explored. The volume of missing values will be considered, and statistical summaries of the variables will be carried out. Exploration of the trend of subscribers over time and identification of any outliers within the data will also be explained.

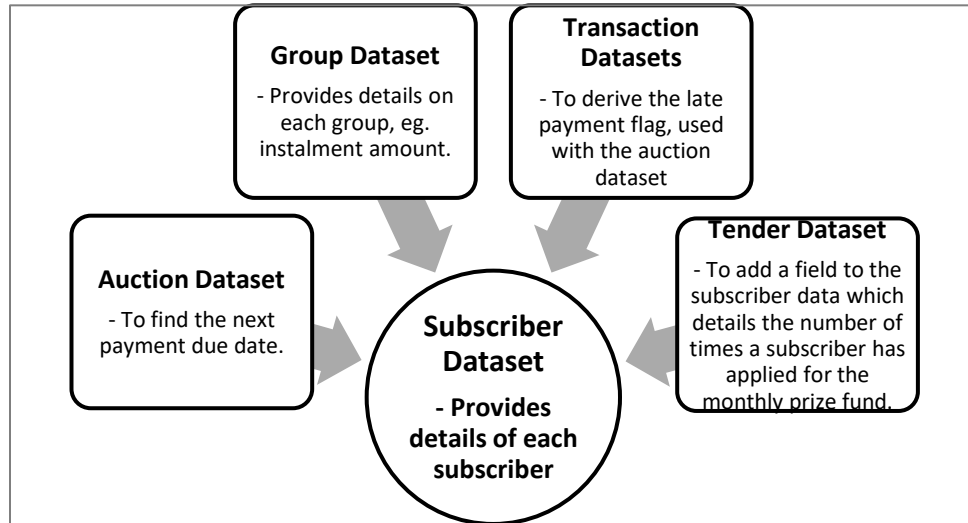


Figure 3.2 Dataset Overview

3.2 Data preprocessing

The data pre-processing stage of the project will be used to clean and prepare the data for the experiment. Any assumptions made in the data will be noted at this stage.

The five datasets will be joined using the ‘sqldf’ package in R studio to create a master dataset to be used in the models. Any issues identified at the data investigation stage will be resolved. A number of common issues previously found in datasets will be identified, such as missing values, outliers, inconsistent or unreliable entries. To resolve any missing values issue, a threshold will be selected based on previous research to remove variables where the number of values missing is above the selected threshold. Of the remaining variables containing missing values, they will then be imputed based on the values that are present within the dataset. Identifying the inconsistent or unreliable entries may not be immediately obvious within the data. However, examining the distribution or mode of the variable will help to identify this issue.

3.2.1 Dummy Variables

As there are a mix of data types in each dataset, dummy variables will be introduced to enable a correlation test to be carried out on the variables. Each categorical variable will be

converted into a series of dummy variables. This allows the correlation test to be carried out therefore reducing the number of features in the final dataset. It also allows for easy interpretation of the variables. This will be carried out in R studio.

3.2.2 Feature Reduction

Feature reduction will be applied to the dataset as there are a large number of variables available for this project. A correlation matrix will be used to remove any redundant variables. Again, a correlation threshold will be selected based on previous literature. Once two variables have a correlation above that threshold, one of the variables will be removed, as including both correlated variables will not be of any benefit to a model and will harm the performance of the model.

3.2.3 Sampling

Before an experiment can be undertaken, a decision has to be made on the class imbalance within the final dataset, as the minority class in this project is the class of paramount interest. It has been found that some machine learning techniques perform better with a balanced dataset (Kelleher, Mac Namee, & D'Arcy, p. 472). One of these techniques is the decision tree (Su, Ju, Liu, & Yu, 2015). A decision tree model trained and tested on datasets with an imbalance of 2% default and 98% non-default gave a 98% accuracy measure. This however is the same as if the decision tree had not predicted anything, but instead classed all instances to be in the same class. This real-world data problem has been seen in many research projects before.

One course of action that can be taken in the data preprocessing phase of the project is random under-sampling, another random over-sampling. As seen in work carried out by Maheshwari, Jain, & Jadon both under-sampling and over-sampling have advantages and disadvantages. Over-sampling can increase the risk of overfitting within a model and increases the workload when the dataset is large. Under-sampling in this project would leave

a small number of records in the sample dataset as the minority class consists of only 150 subscribers. There is a risk of removing important data using the under-sampling technique (Maheshwari, Jain, & Jadon).

The final dataset will then be split into training and test datasets. This split will also be decided based on the previous research carried out in this area. There is a risk that if the data is split only once, the datasets may not be representative, or an outlier may skew the results. To reduce this risk, 10-fold cross validation will be used to split the training and test datasets 10 times. Each model will then be trained on each of these 10 datasets.

3.3 Models

The objective of the research is to carry out an evaluation of machine learning techniques in order to discover which of the techniques more accurately predicts the risk of default for chit fund members, therefore mitigating the risk of loss of prediction power to the company after an employee leaves the organisation.

The first technique to be tested will be the decision tree. Decision trees used in information-based learning run a series of tests to come up with a prediction. Consisting of a root node, leaf nodes and branches, the decision tree tests descriptive features of the dataset in order to get a better prediction of the outcome variable (Kelleher, Mac Namee, & D'Arcy, p.121 - 122). Each leaf in the decision tree represents a class. Therefore, at the end of each path in a decision tree, a class is predicted. The random forest technique is then evaluated. The technique uses 'the combination of bagging, subspace sampling and decision trees' (Kelleher, Mac Namee, & D'Arcy, p.165). The random forest technique consists of multiple decision trees. The class predicted from the random forest technique is a combination of the outcome from each decision tree in the ensemble model. The instance class is predicted by all decision trees, and the most common outcome is selected as the predicted class.

Another machine learning technique that will be evaluated is logistic regression. Logistic regression has been used in many credit risk research papers previously, as it predicts risk with high accuracy. The technique can compare different variable combinations to decipher

the best combination for a given problem. The predictive power of the different combinations of variables can be evaluated (Nie, Rowe, Zhang, Tian, & Shi, 2011). This project will evaluate the technique on chit fund financial data. The target variable is binary, favouring the logistic regression technique. Also, a correlation test will be carried out to remove redundant features also favouring the logistic regression technique.

Finally, SVM will be evaluated. An SVM model is based on error-based learning. It has also become a common technique in credit risk prediction as it can handle categorical target features. 'Support Vector Machines (SVMs) technique is a classification, recognition, regression and time series technique' (Boyacioglu, Kara, & Baykan. SVMs map the input data to high-dimensional feature space, allowing linear separation to take place. An optimal hyper-plane is identified where a prediction can be made based on the hyper-plane separating both classes.

The selected machine learning models have been used in previous research predicting the risk of a customer. In research carried out by Butaru et al., three supervised machine learning models were evaluated on the success in predicting credit card delinquencies. The techniques were decision tree, random forest, and logistic regression. Chen, Härdle, & Moro used the SVM technique when building models to predict the risk of default. It was seen in chapter 2 that each technique has added value to previous research carried out on the credit risk problem.

Each machine learning model will be trained on the training set and evaluated on the previously unseen test dataset. Where possible improvements can be made, the model will be tuned using measures found in the training section of the model building. Parameters such as tree size will be introduced to the decision tree algorithm. Introducing parameters such as this will improve the performance of the model and the model can be evaluated once again.

3.4 Evaluation

To evaluate the models created during this project, a number of measures will be used. No single measure will be solely relied on to decide which model ‘best’ predicts the credit risk score as suggested in research by Maheshwari, Jain, & Jadon. In their research, it is suggested that relying on the accuracy measure as the sole evaluation toll may lead to misleading conclusions. When working with an unbalanced dataset, the accuracy measure does not take into account the number of instances in each class. It can be biased regarding class imbalance in a dataset. This is the reason not to use only one measure on its own (Maheshwari, Jain, & Jadon). Accuracy along with precision, recall and specificity will be calculated and compared for each model. The ROC curve will also be created to visually compare the models.

Figures found, when a confusion matrix of the model is created, will be used in the calculations of accuracy, precision, recall and specificity. The confusion matrix depicts the number of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) cases. The confusion matrix details the predicted class against the actual classes. TP is the number of correctly predicted positive cases. TN is the number of correctly predicted negative cases. FP is the number of incorrectly predicted positive cases and FN is the number of incorrectly predicted negative cases.

Confusion Matrix	Predicted class 0	Predicted class 1
Actual outcome of class 0	TP	FN
Actual outcome of class 1	FP	TN

Table 3.1 Confusion Matrix

If the model predicts that a subscriber is likely to remain out of default, and that subscriber has not previously gone into default, the instance is counted as a TP instance. If the model

predicts that a subscriber is likely to remain out of default, and that subscriber has previously gone into default, the instance is counted as a FP instance. If the model predicts the subscriber is likely to go into default, and that subscriber has been in default, the instance is counted as a TN. If the model predicts the subscriber is likely to go into default, and the subscriber had not previously been in default, the instance is counted as a FN instance.

The evaluation measures that will be used are calculated based on the confusion matrix. These measures are calculated as follows:

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

$$\text{Precision} = \frac{TP}{TP+FP}$$

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

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

The ROC curve plots the TP rate on the y axis and the FP rate on the x axis allowing a visual comparison of the models created. Xanthopoulos & Nakas focused their research on the benefit of using the ROC curve to evaluate credit scoring systems. It was suggested that the ROC curve model can validate credit scoring models and also allow the user to gain further insight into the performance of the model.

Accepting or rejecting the null hypothesis will be based on the evaluation measure calculated in the next chapter.

3.5 Social Network Analysis

Social network analysis will allow further information to be gained about the subscribers of the company that may not be immediately apparent from the data exploration or modelling section of the project. Social network analysis was carried out on the ‘Pcity’ variable. The ‘Pcity’ variable holds the information of the city or town where the subscriber lives. First an edge list will be extracted from the dataset. The edge list will consist of pairs, each pair

will contain a subscriber and a city. Each pair will represent an edge in the bipartite network connecting two nodes. The edge represents the relationship between subscriber and one city. The node list will then be extracted from the dataset; it will contain every subscriber and every possible city. It was hoped that analysing the bipartite network would allow the company to gain further understanding of the community of subscribers that have gone into default.

3.6 Software

Microsoft Excel was used to store the datasets, however it was not capable of storing the entire data used in this project. Therefore, R studio was the tool chosen to carry out the data storage and preparation for this project. While there are many software packages available to carry out the work for this project, R studio was chosen as there are multiple packages available to use for each stage of this study. Both the machine learning stage and social network analysis can be carried out in R studio.

Each dataset will be imported to R studio to explore the data, clean and prepare the datasets for use in the machine learning models. To visualise the social networks built from the datasets, the programme Gephi will be used. Gephi builds high quality visualisations of social networks while also generating some exploratory statistics on the network. The interactive networks produced are useful to gain further insights into the network.

4. IMPLEMENTATION AND RESULTS

This chapter outlines how the experiment was carried out, based on the steps mentioned in chapter three. It includes the data pre-processing carried out and how the machine learning models were created. The chapter also details the results of the evaluation measures calculated for each supervised machine learning model.

4.1 Data Understanding

As outlined in the previous chapter, five datasets were obtained from the chit fund company for this project. The subscriber dataset was taken as the central dataset, as the aim of the project is to predict the performance of potential new subscribers. This dataset was enriched with information from the four other data sources as visualised in figure 3.2 above. To gain further understanding of the business and the data provided, communication through Skype calls and email was used. The chit fund company also provided a data dictionary for each of the datasets which outlined a description of each variable.

The subscriber dataset contained 72 variables; the variables contained information about 12,650 of the companies' subscribers. Of these variables provided, eleven were missing 100% of the values. Four variables in this dataset were identification variables. The remaining variables included in this dataset described the subscriber in terms of their families, income, the first chit fund group they took part in, and their location. The data types were a mix between, numeric, character and date formats. Figure 4.1 and 4.2 detail the categorical and numeric variables in the subscriber dataset.

The auction dataset listed the winning bid for each of the chit fund groups. It included 16 variables and 6,213 records. Of these variables, two were identification variables. Included in the dataset was information such as the payment date, the details of the auction including the amount auctioned and commission paid to the company. The data types, similar to the subscriber dataset, were a mix of numeric, character and date formats.



Figure 4.1 Categorical variables describing the subscribers

Variable	Mean	Minimum	Maximum
Age1	32	0	82
OpenBal	94,384	0	2,300,000
Age	45	0	86
BPay	19,503	0	6,000,000
NPay	2,001	0	1,200,000
Capital	208,588	0	6,000,000
AIncome	258,902	0	25,000,000

Figure 4.2 Numerical variables describing the subscribers

The group dataset contained information on each overall group the company was running. It comprised of twenty-five variables and 535 individual chit fund groups, with eighteen of these variables being descriptive variables and the remaining being identification variables

or those missing all values. Information such as the chit amount, subscriber number, commission paid, and termination date were held here. The data was represented in numerical, character and date format.

The transaction dataset held information on each payment each subscriber made. There were eight descriptive variables of 274,548 transactions made by the subscribers to the chit fund groups. Details of each instalment paid by the subscriber, and the amount and date, were held in this dataset.

Finally, the tender dataset contained all the bids made by subscribers. With three descriptive variables, this dataset contained the number of bids made by a subscriber, and which of the bids were successful. There were 24,137 bids in this dataset.

Following the data exploration stage of this project it was noted that many variables that had been prepared would not be useful in a machine learning model. Attributes had been marked as 'ignore' by the company, as the information provided was not correct or up to date. Other attributes had a high percentage of missing values, and as mentioned previously, 11 variables had no values.

Many of the variables in each dataset had missing values. Figure 4.3 below details the variables which had one or more missing values. Some variables, for example Nominee 2, Nominee 3 or PMobile1, may not have been relevant to the particular subscriber and therefore may have been missing not because of a data error but instead, because they were not relevant.

Variable	Missing Count	Missing Percentage
Nominee1	3288	26%
Relation1	3289	26%
Age1	3301	26%
Nominee2	12447	98%
Relation2	12447	98%
Age2	12447	98%
Nominee3	12645	100%
Age3	12645	100%
Relation3	12645	100%
RepBy	12647	100%
CanCode2	12484	99%
BranchCode	2	0%
OpenBal	1	0%
PRZDate	6279	50%
GDetail	9726	77%
JVNo	10497	83%
PayDate	9591	76%
SecurityMode	9598	76%
RptAmt	12650	100%
ParentName	300	2%
Age	667	5%
PAddress	38	0%
PCity	12	0%
PDist	12609	100%
PPin	1011	8%
PPhone	8310	66%
PMobile1	4033	32%
PMobile2	11733	93%

Variable	Missing Count	Missing Percentage
CAddress	9984	79%
CCity	10870	86%
CDist	12649	100%
CPin	11056	87%
CPhone	12554	99%
CMobile1	12547	99%
CMobile2	12630	100%
LandMark	12649	100%
FirmName	921	7%
Dept	12085	96%
Desig	9724	77%
RtrmDate	12650	100%
BNature	10438	83%
PANNo	12650	100%
OSource	8691	69%
IncomeSrc	307	2%
MCustCode	2177	17%
EMailID	11358	90%
PSNo	35	7%
AggrNo	89	17%
AggrDate	96	18%
CCDate	266	50%
ComDate	6	1%
FADate	26	5%
TermDate	7	1%
PSDate	39	7%
RealDate	76691	28%

Figure 4.3 Details of Missing Values

This stage of the project allowed the data to be explored further, it was found that the number of subscribers has increased in recent years.

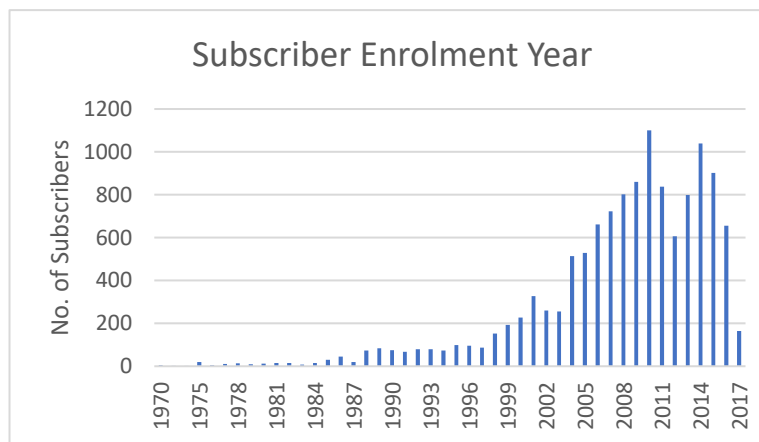


Figure 4.4 Subscribers enrolling per year

The average chit amount in a group varied with the monthly instalments ranging from less than 100 Rupees to 125,000 Rupees, with the majority at 6,000 Rupees.

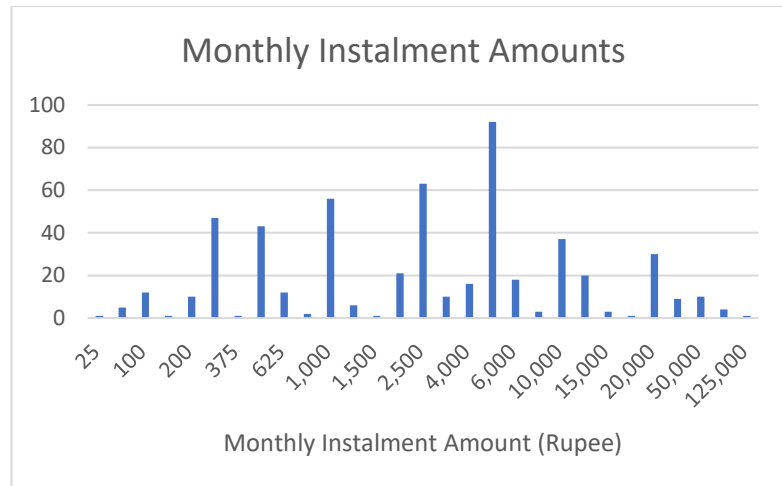


Figure 4.5 Monthly Instalment Amounts Paid by Subscribers

The average annual income of the subscriber was also seen to remain stable from 1970 to 2017.

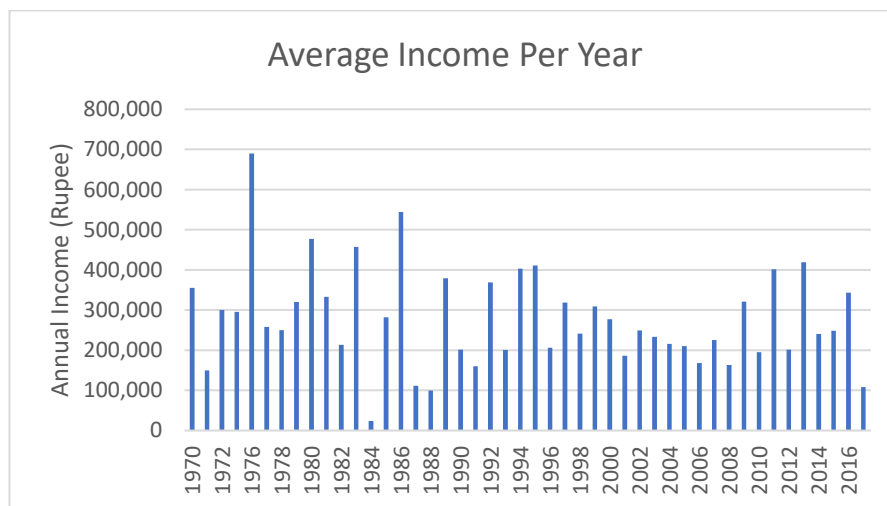


Figure 4.6 Average Annual Income

Average number of members per group is 24 and the average duration of a chit fund group is 61 months. As seen in chapter 2, Carling, Jacobson, Lindé, & Roszbach's research suggests that short term loans are at more risk of going into default compared to long term loans.

Many subscribers had taken part in more than one chit fund group, building up a strong history of payments and transactions, and therefore giving the data more dependability and trustworthiness.

4.2 Data Pre-processing

To prepare the data for the experiment, a number of data pre-processing steps were carried out. The first step was to join the five datasets to create a master dataset. The five separate datasets were imported into R studio where using the ‘sqldf’ package. They were joined using the subscriber dataset as the starting point, and then using the ‘left join’ function to tie in information from the four other datasets. The number of times a subscriber has bid for the monthly prize amount was present in the tender dataset. The group dataset was used to find the monthly instalment amount each subscriber had agreed to pay. The auction dataset provided the payment due date for each instalment, and the transaction dataset provided the payment details including the date paid.

The target variable was then derived from the data contained in the master dataset. As mentioned in the previous chapter, the payment date was taken from the auction dataset and subtracted from the payment due date to create a field containing the number of days early or late a payment was. This field was then used to create the default flag, where a value of ‘1’ meant a payment was over 90 days late and ‘0’ meant all payments were paid on time.

It has been noted that the dataset is unbalanced. As expected only a small percentage of subscribers have been late or missed a payment (1.2% of subscribers). Having only a small percentage of subscribers in default is necessary for a business to succeed. As mentioned in research carried out by West (2000) the target default variable is likely to be unbalanced in research in this area.

Default	Count of Subscribers
Yes	151
No	12,499

Table 4.1 Displays the number of subscribers who have ever had a payment over 90 days late.

To prepare the data for the social network analysis, the dataset was filtered to only contain the subscribers who had previously gone into default on a payment. It was decided that the network of default subscribers would be analyzed. The subscriber IDs and the ‘PCity’ variable were extracted from the final dataset which was used to train and test the supervised

machine learning models. Each value represented a node and each pair (subscriber id and city) represented an edge in the network. The dataset was then exported to Microsoft Excel to be visualized in Gephi.

4.2.1 Feature Reduction

The next step in the data pre-processing stage was to carry out feature reduction. This was carried out to remove the variables that would have no benefit in the models. There were a cohort of variables with only one value. As all values in a variable were equal, the variable contained no predictive power to predict the risk of default and was therefore removed from the dataset. Another cohort of variables within the final dataset were marked as ‘ignore’ by the chit fund company, the definitions of these variables were not clear, or the variables were not up to date, and therefore contained unreliable information. These were also removed from the final dataset.

The data quality of some variables was not high enough to be included in the dataset for modelling. As noted by Kelleher, Mac Namee, & D’Arcy (p. 67), missing values may be caused by a number of different factors. One reason may be that the data was not collected or inputted correctly as seen in the chit fund data. Variables with more than 60% of values missing were removed from the final dataset as ‘the amount of information stored in the feature is so low that it is probably a good idea to simply remove that feature’ (Kelleher, Mac Namee, & D’Arcy, p. 67).

The missing values for the remaining variables were imputed. One assumption made when imputing these values was that the values were missing from the dataset at random and that there was no issue with the data gathering or storing processes. Multivariate Imputation by Chained Equations (MICE) was the first R package to be considered. Another option for continuous variables was to impute the mean value of the available values for all missing entries. A third option was to use the ‘Amelia’ package in R. This also uses multiple imputation to deal with missing values. However, the variables must be normally distributed. Imputation of missing values was carried out by using the mean of the numeric

values. This was imputed for each numeric variable and the mode of the categorical variables was imputed, therefore not altering the variance in the data. Acknowledging that imputing categorical variables may not be desirable, the models were also built on the dataset before missing values were imputed and results were compared to models using the dataset which had missing values imputed.

Dummy variables were created from the categorical variables in the dataset, which allowed correlation testing to be carried out. Dummy variables were created from the twelve variables listed in table 4.2; this table also lists the number of unique values in each variable.

Variable	No. of unique values
Relation1	25
Intimation	2
IDCard	3
Status	3
PRZFlag	2
ACClosed	2
ParentType	4
PCity	96
NOJ	2
ITPayee	2
IncomeSrc	2
Default90	2

Table 4.2 Categorical Variables

A correlation matrix, Spearman's correlation coefficient was examined to identify if there were any considerable correlations between variables. Correlated variables would provide no benefit to the model and increase computing time. Based on this matrix, many features had little or no correlation to each other and were included in the dataset to be used in the models. However, a total of eight pairs of variables had a correlation of more than 0.8 or less than -0.8 and therefore eight variables were excluded from the dataset.

The names of the customer, their email addresses, other sensitive data, and identification details were also removed from the dataset in order to create a model that is more general on the descriptive variables.

One issue identified when the dummy variables were created was the misspelling or inconsistency of entries, which meant the value such as ‘Brother’ and ‘BROTHER’ in the ‘Relation1’ variable were listed twice in the dummy variables created. To solve this issue, all lowercase characters were converted to uppercase and the dummy variables were recreated.

4.2.2 Sampling

The sampling technique used was Synthetic Minority Over-sampling Technique (SMOTE) in R Studio. This sampling technique creates synthetic data for the minority class in the dataset based on the data in the dataset provided to the function, which therefore, balances the class. Using this function, inputting the default variable, the function enables the user to select the proportion of each class to be included in the final sample dataset. In the sample dataset the ratio of the class was set to 1:4. For every record where the default value was 1, there were four records where the default value was 0. This ratio was chosen as a completely balanced dataset in this area of study is not a realistic goal.

The final dataset was then split into training and test datasets. A split of 66% training and 34% test datasets was used in the models. This same split has been used in previous papers evaluating machine learning techniques while predicting credit risk. Research by Yang, Kruppa, Schwarz, Arminger, & Ziegler and Boyacioglu, Kara, & Baykan, used this 66% split for training data.

4.3 Dataset Description

There was a total of 147 variables included in the final dataset. Table 4.3 describes the variables that have remained in the dataset. Missing values have been removed or imputed, feature reduction has been carried out and the datasets five separate datasets have been joined together.

Variables	Description	Type	Dummy Variable Created
Age1	Age of the first nominee	Integer	No
BranchCode	Branch ID	Integer	No
OpenBal	Opening balance of the customer when joining the group	Numeric	No
Age	Age of the subscriber	Integer	No
BPay	Basic Pay	Numeric	No
NPay	Net Pay	Numeric	No
Capital	Income	Numeric	No
AlIncome	Annual Income	Numeric	No
Num_Bids	Number of times a subscriber has bid for the prize	Integer	No
ChitAmount	Amount the chit fund is worth	Numeric	No
CompTicketNo	The subscription number assigned to the chit company	Integer	No
CompComm	The % commission deducted by the chit company	Integer	No
Relation1	Relation with the first nominee	Integer	Yes
Intimation	Method of notice and reminders	Integer	Yes
IDCard	Identification document used	Integer	Yes
Status	Account Status	Integer	Yes
PRZFlag	Flag to indicate if this subscriber has won the prize in the specified chit group	Integer	Yes
ACClosed	Flag to indicate whether the Account of the subscriber has been closed	Integer	Yes
ParentType	Relationship with the parent.	Integer	Yes
Pcity	Permanent Address of the subscriber	Integer	Yes
ITPayee	IT Industry	Integer	Yes
IncomeSrc	Source of the subscribers income	Integer	Yes
Default90	Greater than 90 days late with a payment	Integer	Yes

Table 4.3 Final Dataset Description

4.4 Machine Learning Models

The four proposed supervised machine learning models were created to test which best predicted the risk of default of a potential chit fund customer.

4.4.1 Decision Tree

The decision tree was chosen as the first model to be created as it can help to identify the most valuable variables, and as there are a large number of descriptive variables present in the dataset. The decision tree allows for easy inspection of the variables used and paints a very clear picture of the data. Using the training dataset, a model was created using the ‘rpart’ package.

Before building the decision tree, the data was divided into training and test datasets from the final dataset which was created using the SMOTE sampling technique. The training dataset contained 66% of the data, leaving the remaining 34% for the test dataset. To ensure this split was carried out as expected, the count of instances in each dataset was checked.

The package ‘rpart’ was installed in r studio and used to create the decision tree. Rpart stands for recursive partitioning, and is a commonly used package for decision trees in R. The decision tree was trained on all training data with the method set to ‘class’. This unpruned tree provides the following results.

Evaluation Measures	Value
Accuracy	79%
Precision	74%
Recall	56%
Specificity	90%

Table 4.4 Decision Tree Evaluation Results

In order to obtain better results from this decision tree model and to avoid overfitting, the tree will be pruned. To understand what parameters work best for this model the cross-validation error was plotted as seen in Figure 4.7 below. By plotting the figures shown in Figure 4.7, it can be seen that a decision tree of size nine has the smallest relative error rate with a CP of 0.015075.

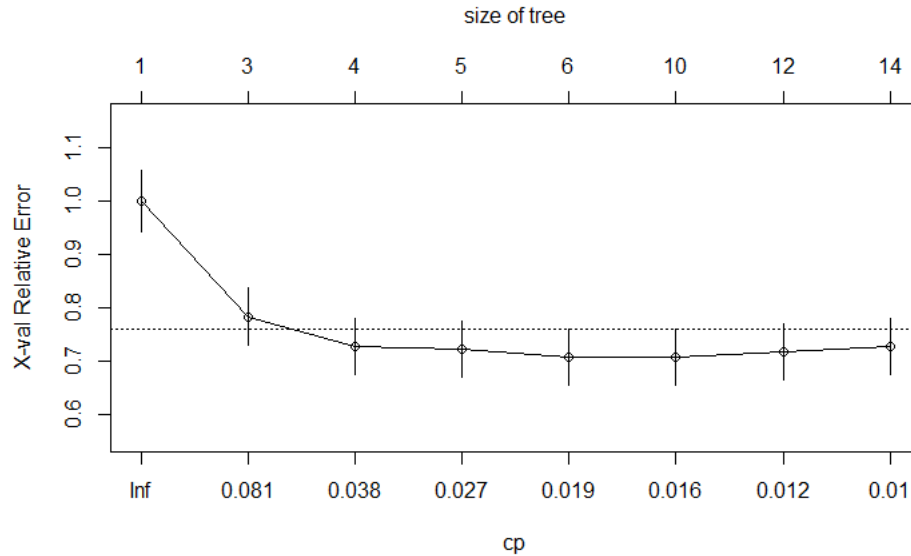


Figure 4.7 Cross Validation Error

Pruning the dataset by including the ‘rpart.control’ function with $cp = 0.015075$ provided a lower false positive rate and increased accuracy, precision and specificity as seen in Table 4.5. Recall is the only measure that remained unchanged following the pruning of the decision tree. The specificity is high at 93%. However it has been noted previously that specificity will not be as valid an evaluation measure as precision and recall, as the dataset is unbalanced. Specificity measures the correctly predicted majority class divided by the total number of instances in the majority class. This figure can be close to 100% but this measure does not show how well the minority class was predicted.

Evaluation Measures	Value
Accuracy	81%
Precision	79%
Recall	56%
Specificity	93%

Table 4.5 Pruned Decision Tree

The output in figure 4.8 clearly shows the process followed by the decision tree to predict the outcome. In each leaf node, details on the number of instances that followed that path is presented, showing the number of correctly and incorrectly identified instances. The decision tree shows the variables the algorithm found most informative to predict default.

In total nine variables were selected by the decision tree as informative. These variables are Relation1Wife, Number of Bids, Opening Balance, Chit Amount, Relation1Husband, Net Pay, Age, PCityPuducherry and Branch Code.

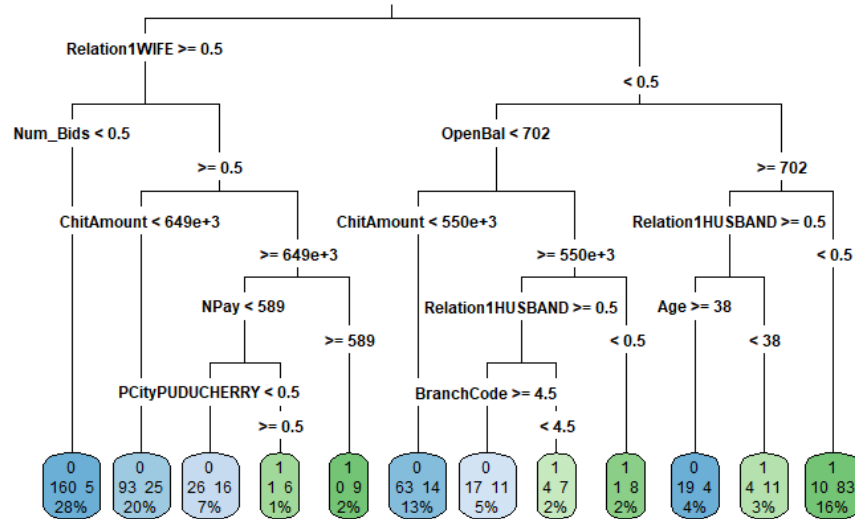


Figure 4.8 Pruned Decision Tree

4.4.2 Random Forest

Following on from the decision tree model a random forest model was built. The random forest was generated in R studio using the ‘randomForest’ package. The random forest was first generated with the default settings, 500 trees used and the Out of Bag (OOB) error rate was 17.09%.

The random forest was then tuned using the ‘tuneRF’ function. Plotting the error rate for the model, it is seen that the error rate reached its lowest point after approximately 300 trees. Therefore, the model was tuned to create up to 300 trees instead of the default of 500. Plotting the mtry, which is the number of variables tested at each split in a decision tree, against the OOB error then suggests what the optimal number mtry should be set to, within the random forest model. Figure 4.9 suggests that 12 is the optimal number of variables that should be tested for each split, as it gives the lowest OOB error.

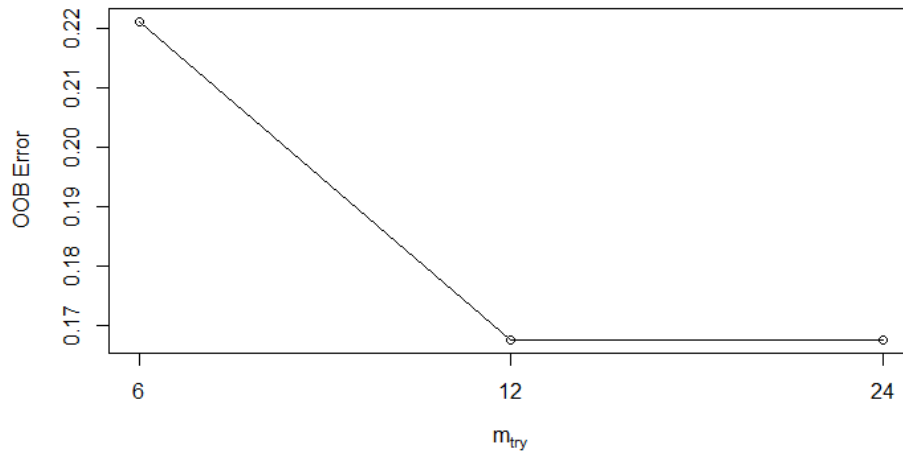


Figure 4.9 Graph to select the optimal mtry value

Inputting each of these parameters into the random forest model then improved the performance of the model. The evaluations that were focused on were precision and recall, precision of 59% was calculated with recall of 92%. A total of sixty-five variables were used in the random forest model, and results are shown in Table 4.6 below.

Evaluation Measures	Value
Accuracy	85%
Precision	59%
Recall	92%
Specificity	83%

Table 4.6 Evaluation of Tuned Random Forest Model

4.4.3 Logistic Regression

The logistic regression technique was carried out using the ‘glm’ function in R. As with the decision tree and the random forest models, the first logistic regression model was carried out with the default ‘glm’ settings. The output in R details the significance of including each variable in the logistic regression model. All variables that did not have a significance of

less than 0.1 for the model were then excluded from the model and the model was trained and tested on this new set of variables. Eight variables were significant in the logistic regression model; these were Age of the first nominee, Opening Balance, Age, Chit Amount, Relation1Wife, Relation1Husband, Relation1Father, Relation1Mother. Some similar variables were used in the decision tree and random forest models. Precision of 50% and recall of 80% were calculated.

The following results were obtained from the logistic regression model.

Evaluation Measures	Value
Accuracy	79%
Precision	50%
Recall	80%
Specificity	79%

Table 4.7 Tuned Logistic Regression Results

4.4.4 Support Vector Machine

The SVM technique was the final model to be built. The ‘svm’ function was used to build this model. The default SVM model gave the following evaluation measures, precision of only 7% and recall of 58%. Tuning was required to improve the results of the model as accuracy came in at 67% and specificity was 68% which is not expected for an unbalanced dataset.

The SVM model was tuned using the following code:

```
svm_tune_v1<-tune(svm, Default90Y~., data = train_smote,
```

```
ranges = list(epsilon = seq(0,1,0.1), cost=2^(2:9)))
```

Evaluation Measures	Value
Accuracy	67%
Precision	7%
Recall	58%
Specificity	68%

Table 4.8 SVM Results

4.5 Social Network Analysis

In order to complete the social network analysis, subscribers who had gone into default were grouped by the area in which they live. Figure 4.10 shows a bipartite network with subscribers as light green nodes and the cities they come from as dark green labelled nodes. The edges connect the subscribers to the city they are from. It can be seen that some cities are linked to many default subscribers. However, this is also in line with the proportion of overall subscribers from these cities in the chit fund company.

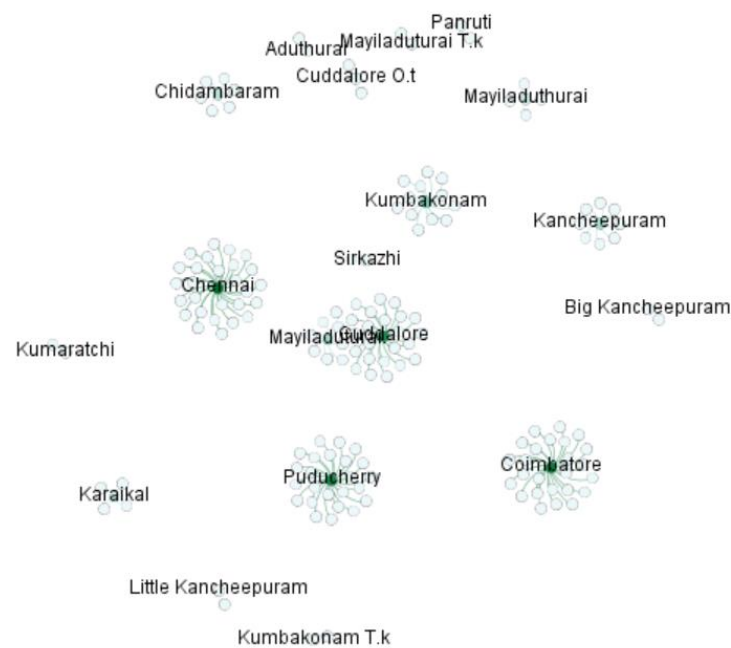


Figure 4.10 Default Subscribers Grouped By Address

This was visualised in Gephi and suggests that no unexpected relationship exists between the subscribers in default and the cities they are from. Figure 4.11 below shows the same bipartite network, however all subscribers are now included. The red nodes represent the subscribers who have gone into default and green nodes represent subscribers who have not gone into default. This network confirms that there is no unusual relationship between defaults and subscribers addresses.

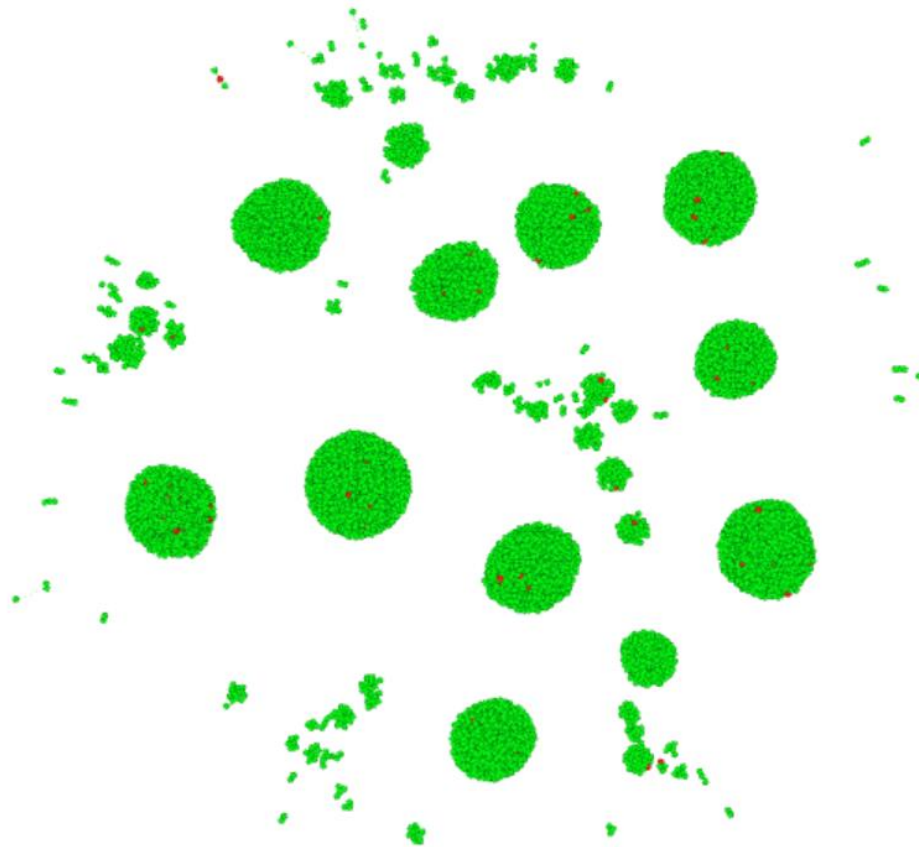


Figure 4.11 All Subscribers Grouped By Address

4.6 Results

The following results in Table 4.9 were obtained from each supervised machine learning technique used.

Evaluation Metric	Decision Tree	Random Forest	Logistic Regression	Support Vector Machines
Accuracy	81%	85%	79%	67%
Precision	79%	59%	50%	7%
Recall	56%	92%	80%	58%
Specificity	93%	83%	79%	68%

Table 4.9 Results of Supervised Machine Learning Models Built

The results in table 4.9 are from the models built and tested on the dataset which had the missing values imputed. As mentioned previously in this chapter, imputing missing values for categorical variables may not be advantageous. With this in mind the experiment was repeated. The missing values for categorical variables in the dataset have not been imputed; they have been excluded from the training and the test datasets.

Evaluation Metric	Decision Tree	Random Forest	Logistic Regression	Support Vector Machines
Accuracy	78%	83%	79%	68%
Precision	75%	57%	60%	5%
Recall	50%	86%	72%	71%
Specificity	92%	82%	82%	68%

Table 4.10 Results of Supervised Machine Learning Models Without Imputing Missing Values

It can be seen when comparing results of the model built with missing values imputed to the models built without imputing missing values, that the results after imputing missing values are better. Results in table 4.10 are based on models trained on the dataset which had all rows that contained at least one missing value removed from the final dataset. This is contrary to what was suggested in research by Huang et al. In this research, it was found that models built on the dataset without imputing missing values performed to a higher standard than the models built on the dataset where missing values were imputed.

Based on this comparison, the results in table 4.9 will be the focus of the analysis, evaluation, discussion and conclusion that follows.

5. ANALYSIS, EVALUATION AND DISCUSSION

This chapter evaluates the predictive power of the supervised machine learning models built in chapter 4. For each model, a comparison of accuracy, precision, recall and specificity will be carried out. The ROC curve will also be used to visually compare the models built. Initially the datasets were divided into 66% training set and 34% test sets. The results used in this chapter are based on the models run with the unseen test datasets.

As seen in Table 4.9 the Random Forest model has the highest accuracy and recall value of the four models created while the decision tree has the highest precision and specificity value. To understand if these models provide any value to solving the problem, the benchmark value was calculated. The benchmark value is the accuracy of a model if it classed all instances to the majority class, in this case 0. If all instances were classed as 0 in the testing dataset, the accuracy of that benchmark model would be 66%. Therefore, any model falling below that accuracy level will be disregarded.

As there is more interest in correctly predicting the default cases compared to the non-default cases, the evaluation measure of precision and recall will prove to be more important than overall accuracy. Precision is calculated by dividing the correctly identified minority class by the total cases predicted as the minority class. Recall then is calculated by dividing the correctly predicted minority class by the total number of actual cases in the minority class.

The precision of the decision tree was highest of the four models at **79%** followed by random forest at **59%**. Recall was calculated as **92%** using the random forest technique which was the highest followed by **80%** for the logistic regression technique. Overall, accuracy was highest also for the random forest model.

A ROC curve was created to further compare the models. This was created using the 'ROCR' package; it can be seen in figure 5.1. Plotting the TPR against the FPR, the Random Forest model also comes out as the most sensitive model. The area under the ROC curve (AUC) for each model was then calculated in R. For the decision tree model, the AUC was

85.78%. The Random Forest model had an AUC of **90.6%**. The Logistic Regression model has an AUC of **82.14%** and the Support vector machine has an AUC of **52.18%**.

All three supervised machine learning models, however, are above the random model which is displayed as the black line in figure 5.1. This graph shows that unlike what was seen in the literature review, random forest is suggested to be the best predictor of chit fund default. In much of the literature reviewed for this project, it was suggested that the support vector machine technique has provided the most accurate results in the area of credit risk prediction. However, with the chit fund data used for this analysis, it was not found to be the most valuable model. Trustorff, Konrad, & Leker found SVM out-performed Logistic Regression. Similarly, research by Chen, Härdle and Moro found SVM performed with the highest accuracy.

It has been seen, however, in research by Barboza, Kimura, & Altman that the random forest technique outperformed logistic regression when predicting bankruptcy. Similar to this project, the dataset used in their research was unbalanced. No normalization of the data was carried out and a correlation test was used to carry out feature reduction. It suggests similar techniques used in Barboza, Kimura, & Altman's research provided similar results in terms of supervised machine learning techniques.

This was also suggested in research by Brown & Mues, where random forest performed the best when compared to ten classification techniques. This research focused on datasets with class imbalance, similar to the dataset used in this project.

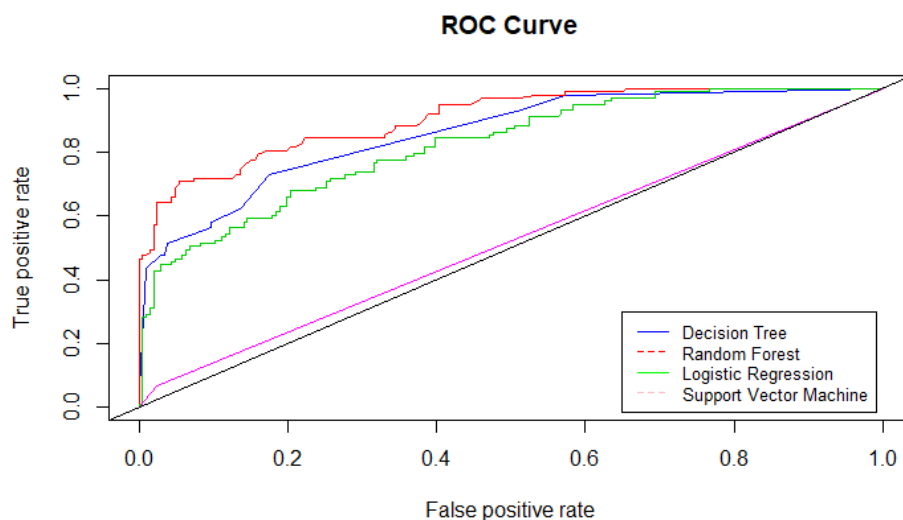


Figure 5.1 Receiver Operation Characteristic (ROC)

As expected, the random forest model out-performed the decision tree model. The random forest model was tuned to consist of 300 trees and the results were aggregated to select the most common response from each decision tree. The majority of decision trees were between a size of 50 nodes and 70 nodes, as can be seen in Figure 5.2.

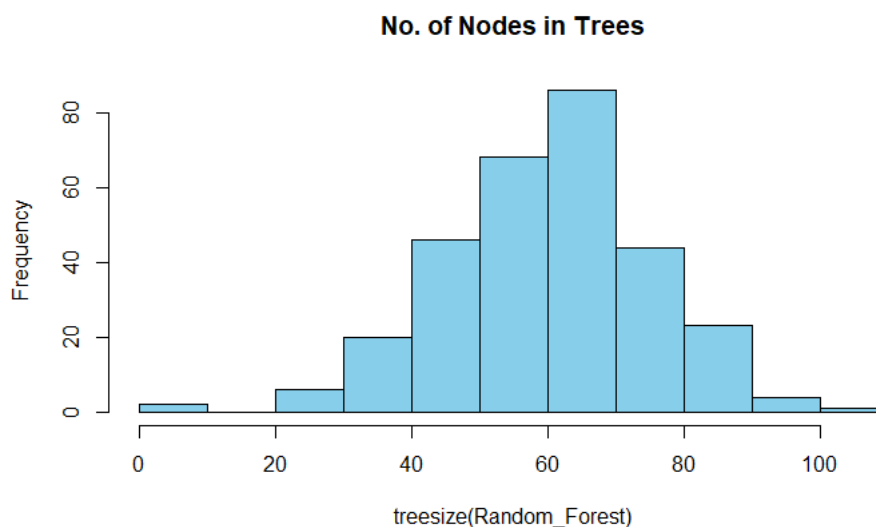


Figure 5.2 Size of Decision Trees in Random Forest Model

Comparing the variables used in each model, many variables were commonly used across all techniques suggesting that they are very informative for this project. This section looks at the importance of the variables used in three of the models in this experiment. The three models considered were decision tree, random forest and logistic regression. Support vector machine does not easily show the list of variables used and is therefore not included here. Fourteen of the variables in the training dataset were used in one, two or three of the predictive models created. Of these, four variables were used in each of the models, Relation1Wife, Opening Balance, Chit Amount and Age. A further four were used in two of the three models, Number of bids, Relation1Husband, Branch Code and Age 1. The remaining seven variables were used in one of the three models, Net pay, PCityPuducherry, CompTicketNo, Capital, Basic Pay, Relation1Father and Relation1Mother.

The Random Forest model is suggested to be the best predictive model for the dataset used in this project. The top 10 variables used are detailed below with regards to the mean decrease in accuracy if a variable is removed, and the mean decrease in Gini which measures the contribution of a variable in the decision trees. It shows that the variables listed in both graphs are the same. However, the order of importance is different for both measures.

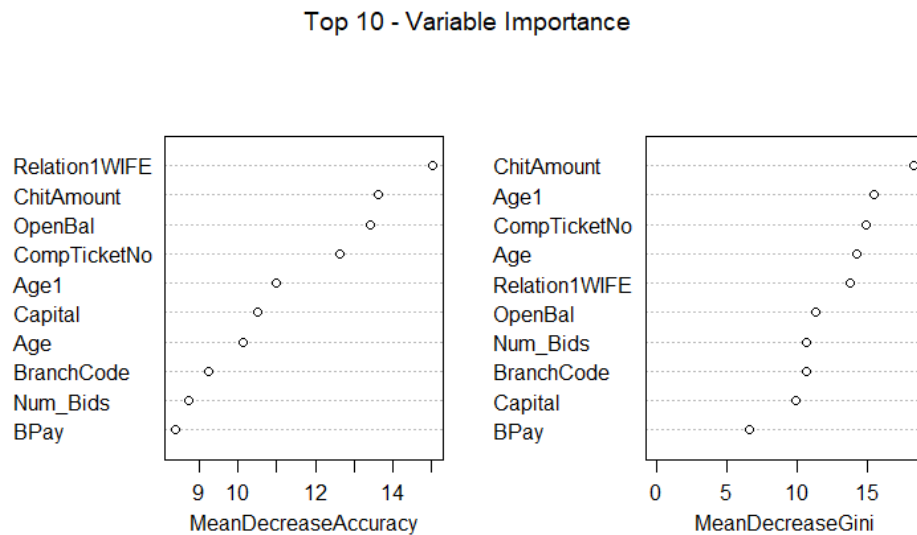


Figure 5.3 Variable Importance in the Random Forest Model

Age for both the subscriber and their nominee is seen in Figure 5.3 to be a highly predictive variable when predicting the risk of default for subscribers. This was also noted in research

by Avery, Brevoort, & Canner where it was suggested that applicants under the age of 30 were given a lower risk score than those over 30 years of age.

The hypothesis, a support vector machine achieves a higher sensitivity and specificity than that of a logistic regression model, random forest and decision tree model, when used to predict the risk score of a chit fund member, can be rejected. It can be concluded that the random forest technique outperformed the support vector machine technique.

5.1 Strengths and Limitations

The results suggest that the Random Forest model is the best predictor of chit fund default when compared to Decision Tree, Logistic Regression and SVM. This, although it was unexpected, may be beneficial. General Data Protection Regulation (GDPR) rules have recently become more restrictive in Ireland and throughout Europe. The new GDPR rules strengthen an individual's rights to request and obtain their personal data which is held and processed by a company. This includes automated processes. Random forest models may become very beneficial, as the transparency of how and why the data was processed in the model is easily accessed.

The data used in the machine learning models was from one chit fund company; this may contain company specific characteristics. A subset of the companies' customers was included in the dataset. It cannot be known if this was a biased subset of subscribers. There was a high volume of missing data present in the datasets. This meant that many variables were excluded from the dataset. Furthermore, the remaining missing values were imputed, this may have introduced bias to the dataset.

Many data pre-processing techniques were used throughout the project including handling missing values and feature reduction. There are many techniques available to handle these issues, and although multiple were considered, only one was tested for each. Performance of the models may be improved if other data pre-processing techniques were used.

The accuracy of the models has been found to be high. However as seen in the literature, the accuracy measure is not the most reliable measure to be used solely to evaluate a supervised machine learning model based on an unbalanced dataset. The dataset used in this project was highly unbalanced. Therefore, it cannot be regarded as a reliable evaluation tool in this project.

6. CONCLUSION

This chapter gives an overview of the research carried out. It summarises the results of the experiment and details the results to research previously carried out in this area. The chapter summarises the findings in relation to the research question set out at the beginning of this project; *Which supervised machine learning technique; decision tree, random forest, logistic regression or support vector machine; can best predict the risk score of a chit fund member with best accuracy, precision, recall and specificity?*

6.1 Research Overview

The goal of this research was to examine the predictive power of four supervised machine learning techniques on a chit fund dataset. Chit funds are an informal financial industry in India, and as seen in the literature review, this sector has grown in rural areas. The four machine learning techniques examined were, decision trees, random forest, logistic regression and support vector machines. These were selected based on the research reviewed for the project.

The aim of the supervised machine learning models was to predict the risk of default of a chit fund member. Many previous papers cover the credit risk score prediction area for formal financial institutions. However, the papers reviewed for this project did not cover credit risk score for informal finance such as chit funds.

The main objective was to identify the supervised machine learning model with the best accuracy when predicting credit risk. Chapter two described previous research carried out in this area, the successes of each technique and challenges faced in previous work. Chapter three detailed the method and design in which the experiment of building the machine learning models would be carried out. Chapter four outlined the experiment of building the models. Chapter five analysed the results of the four supervised machine learning models and compared their performance. A number of evaluation measures were used in this chapter, as it was mentioned in the literature review that the evaluation measure of accuracy was not the most suited to the problem as the dataset was highly unbalanced. It was

suggested that it was more appropriate to compare a number of evaluation metrics. It was found that of the four machine learning models, the random forest technique outperformed the other techniques evaluated. Therefore, the null hypothesis that; A support vector machine achieves a significantly (95% confidence) higher sensitivity and specificity than that of a logistic regression model and decision tree model, when used to predict the risk score of a chit fund member, cannot be accepted.

A second objective of research in this project looked at the social network of subscribers of the company. It was noted that of the subscribers that had gone into default, the majority were from one of seven cities in India. This, however, was proportional to the number of overall subscribers from each city. No unusual pattern was noted as it was expected that the number of subscribers from each city is proportional to the number of subscribers that went into default from each of those cities.

6.2 Problem Definition

Credit scoring has been and is currently used as a powerful tool in formal financial institutions and may also be worthwhile implementing in the chit fund company who have provided the data for this experiment. Supervised machine learning techniques have been used in many credit scoring problems in the past with SVM techniques seen to be one of the most common in credit risk prediction.

The literature review confirmed that many supervised machine learning techniques have been evaluated in the research area of credit risk prediction and many have been successfully implemented in financial institutions. This same level of research was not evident for credit risk prediction with regards to chit fund companies or other informal financial companies in India. This research attempts to understand which machine learning algorithm would best predict credit risk for chit fund companies.

Currently the chit fund employees manually assess the risk of a customer. Creating an automated way to predict the risk of a customer will illuminate the risk of employee churn which results in a loss of knowledge to the company. Building and evaluating four

supervised machine learning models, decision tree, random forest, logistic regression and SVM, allowed the most valuable machine learning model, of those tested, to be identified. This project has contributed to gaining further insight into the subscribers of this chit fund company through both machine learning and social network analysis.

To try to find the best performing machine learning technique to predict the risk score for chit fund companies the following steps were carried out.

- a. Further knowledge and business understanding was gained.
- b. Issues found in the dataset were solved, such as missing values and inconsistent entries.
- c. Supervised machine learning techniques were chosen, and models were built and analysed.
- d. Results of each model were compared.
- e. Conclusion was reached as to which model was most effective in predicting default cases for this dataset.

6.3 Design/Experimentation, Evaluation & Results

The design of this project mostly followed the CRISP-DM process. As previously mentioned, the data was provided by an Indian chit fund company. It contained information on a subset of their customer base.

Following the data exploration stage of the project, data cleaning had to be considered. Variables with missing values of above the threshold of 60% were removed, sensitive data and identifying variables were removed, and the remaining missing values were imputed. To prepare the data, feature reduction was carried out using correlation testing and the dataset was split into training and test datasets.

Four supervised machine learning models were built using the following techniques; decision tree, random forest, logistic regression and support vector machine. They were evaluated on multiple evaluation measures to account for the unbalanced dataset. The

evaluation measures used were accuracy, precision, recall and specificity. A ROC curve was also generated to evaluate the models.

One step in the CRISP-DM process is that a model will have to be evaluated and re-trained after some time passes. Understanding that the model is not future proof is important. As the chit fund business changes the model should be re-trained to enable it to account for these changes and remain informative for the company.

The random forest model provided the highest predictive power. It outperformed the other techniques when compared on the ROC curve. As mentioned precision and recall were the evaluation measures focused on, precision for the random forest model was 59% and recall was calculated as 92% for this technique. In the literature review section of the project, it was suggested that either the support vector machine technique or logistic regression would prove to be the best model for this project. However this was not found to be the case.

Social network analysis was carried out also to gain further understanding of the subscriber base of the company, and to explore any relationship between subscribers that had not been identified in the machine learning section of the project. It was found that there was no unusual pattern visible when examining subscribers who had gone into default and the cities in which they were from.

6.4 Contributions and Impact

After carrying out this experiment, the following was found:

The datasets were merged and cleaned. The issues found have been listed and would be already identified for future work on this dataset. Issues such as inconsistent values can be time consuming to identify, and understanding what issues are in the dataset may reduce time spent on data pre-processing in future work.

Supervised machine learning may be a valid way to improve the decisioning process in a chit fund company. Each technique performed better than that of a random model and may be useful to implement in a chit fund company.

Different machine learning techniques have different advantages when predicting the default risk of a subscriber. The decision tree and random forest technique can easily display the process that was used to make the prediction. The logistic regression technique, similar to the decision tree, allows the variables which were considered important to the algorithm to be identified. The SVM technique has been shown to be the most accurate technique in previous work.

6.5 Future Work and Recommendations

Some future work identified throughout the project, and that may be carried out, would be to gather data from other chit funds and informal financial companies to compare models built from datasets from different companies. It is not known how different these datasets might be as each company may have different processes in place to collect data as well as different focuses on what should be collected and analysed.

Evaluating more machine learning techniques on the dataset used in this project may be beneficial. The four techniques chosen for this project were chosen based on previous research however this is not to say that another technique may build a more useful machine learning model.

Further analysis may be carried out on the social network of the subscribers. In this project social network analysis was used as an exploratory tool to look at the subscriber network. However, it may be useful to examine the predictive power of the social network also.

7. BIBLIOGRAPHY

- Agarwal, S., Chomsisengphet, S., & Liu, C. (2011). Consumer bankruptcy and default: The role of individual social capital. *Journal of Economic Psychology*, 32(4), 632–650. <https://doi.org/10.1016/j.joep.2010.11.007>
- Avery, R. B., Brevoort, K. P., & Canner, G. B. (2009). Credit Scoring and Its Effects on the Availability and Affordability of Credit. *Journal of Consumer Affairs*, 43(3), 516–537. <https://doi.org/10.1111/j.1745-6606.2009.01151.x>
- Banerjee, T., Ghosh, C., & Roy, M. (2010). Borrowers in a Village Economy: An Analysis of Credit Contracts Across Rural Households. *Asia-Pacific Social Science Review*, 10(1). <https://doi.org/10.3860/apssr.v10i1.1582>
- Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405–417. <https://doi.org/10.1016/j.eswa.2017.04.006>
- Boyacioglu, M. A., Kara, Y., & Baykan, Ö. K. (2009). Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. *Expert Systems with Applications*, 36(2), 3355–3366. <https://doi.org/10.1016/j.eswa.2008.01.003>
- Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), 3446–3453. <https://doi.org/10.1016/j.eswa.2011.09.033>
- Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance*, 72, 218–239. <https://doi.org/10.1016/j.jbankfin.2016.07.015>
- Butera, G., & Faff, R. (2006). An integrated multi-model credit rating system for private firms. *Review of Quantitative Finance and Accounting*, 27(3), 311–340. <https://doi.org/10.1007/s11156-006-9434-7>

- Butts, C. T. (2008). Social network analysis: A methodological introduction. *Asian Journal Of Social Psychology*, 11(1), 13–41. <https://doi.org/10.1111/j.1467-839X.2007.00241.x>
- Carling, K., Jacobson, T., Lindé, J., & Roszbach, K. (2007). Corporate credit risk modeling and the macroeconomy. *Journal of Banking & Finance*, 31(3), 845–868. <https://doi.org/10.1016/j.jbankfin.2006.06.012>
- Chen, S., Härdle, W. K., & Moro, R. A. (2011). Modeling default risk with support vector machines. *Quantitative Finance*, 11(1), 135–154. <https://doi.org/10.1080/14697680903410015>
- Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3), 1447–1465. <https://doi.org/10.1016/j.ejor.2006.09.100>
- Dahiya, S., Handa, S. S., & Singh, N. P. (2015). Credit scoring using ensemble of various classifiers on reduced feature set. *Industrija*, 43(4), 163–174. <https://doi.org/10.5937/industrija43-8211>
- Daskalaki, S., Kopanas, I., & Avouris, N. (2006). EVALUATION OF CLASSIFIERS FOR AN UNEVEN CLASS DISTRIBUTION PROBLEM. *Applied Artificial Intelligence*, 20(5), 381–417. <https://doi.org/10.1080/08839510500313653>
- Doumpos, M., & Zopounidis, C. (2007). Model combination for credit risk assessment: A stacked generalization approach. *Annals of Operations Research*, 151(1), 289–306. <https://doi.org/10.1007/s10479-006-0120-x>
- Edelman, D. (2008). Credit this: how the banks decide your credit score. *Significance*, 5(2), 59–61. <https://doi.org/10.1111/j.1740-9713.2008.00287.x>
- Farquad, M. A. H., & Bose, I. (2012). Preprocessing unbalanced data using support vector machine. *Decision Support Systems*, 53(1), 226–233. <https://doi.org/10.1016/j.dss.2012.01.016>
- Feki, A., Ishak, A. B., & Feki, S. (2012). Feature selection using Bayesian and multiclass Support Vector Machines approaches: Application to bank risk prediction. *Expert Systems with Applications*, 39(3), 3087–3099. <https://doi.org/10.1016/j.eswa.2011.08.172>

- Godlewski, C. J., Sanditov, B., & Burger-Helmchen, T. (2012). Bank Lending Networks, Experience, Reputation, and Borrowing Costs: Empirical Evidence from the French Syndicated Lending Market: BANK LENDING NETWORKS AND BORROWING COSTS IN FRANCE. *Journal of Business Finance & Accounting*, 39(1–2), 113–140. <https://doi.org/10.1111/j.1468-5957.2011.02269.x>
- Hamed, A. A. M., Li, R., Xiaoming, Z., & Xu, C. (2013). Video Genre Classification Using Weighted Kernel Logistic Regression. *Advances in Multimedia*, 2013, 1–6. <https://doi.org/10.1155/2013/653687>
- Huang, M.-W., Lin, W.-C., Chen, C.-W., Ke, S.-W., Tsai, C.-F., & Eberle, W. (2016). Data preprocessing issues for incomplete medical datasets. *Expert Systems*, 33(5), 432–438. <https://doi.org/10.1111/exsy.12155>
- Jones, J. H. M. (2008). Informal finance and rural finance policy in India: historical and contemporary perspectives. *Contemporary South Asia*, 16(3), 269–285. <https://doi.org/10.1080/09584930802271315>
- Kelleher, J., Mac Namee, B., & D’Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics : algorithms, worked examples, and case studies*. Cambridge, Massachusetts : The MIT Press, 2015.
- Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11), 2767–2787. <https://doi.org/10.1016/j.jbankfin.2010.06.001>
- Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, 40(13), 5125–5131. <https://doi.org/10.1016/j.eswa.2013.03.019>
- Li, H., & Sun, J. (2011). Predicting business failure using support vector machines with straightforward wrapper: A re-sampling study. *Expert Systems with Applications*, 38(10), 12747–12756. <https://doi.org/10.1016/j.eswa.2011.04.064>

- Maheshwari, S., Jain, R. C., & Jadon, R. S. (2017). A Review on Class Imbalance Problem: Analysis and Potential Solutions. *International Journal of Computer Science Issues (IJCSI)*, 14(6), 43–51.
- Mehran, R., Pocock, S. J., Nikolsky, E., Clayton, T., Dangas, G. D., Kirtane, A. J., ... Stone, G. W. (2010). A Risk Score to Predict Bleeding in Patients With Acute Coronary Syndromes. *Journal of the American College of Cardiology*, 55(23), 2556–2566. <https://doi.org/10.1016/j.jacc.2009.09.076>
- Mehta, A., & Bhattacharya, J. (2017). Channels of financial sector development and rural-urban consumption inequality in India. *International Journal of Social Economics*, 44(12), 1973–1987. <https://doi.org/10.1108/IJSE-05-2015-0117>
- Milošević, M., Živić, N., & Andjelković, I. (2017). Early churn prediction with personalized targeting in mobile social games. *Expert Systems with Applications*, 83, 326–332. <https://doi.org/10.1016/j.eswa.2017.04.056>
- Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, 38(12), 15273–15285. <https://doi.org/10.1016/j.eswa.2011.06.028>
- Nova, D., & Estévez, P. A. (2014). A review of learning vector quantization classifiers. *Neural Computing and Applications*, 25(3–4), 511–524. <https://doi.org/10.1007/s00521-013-1535-3>
- Peng, Y., Wang, G., Kou, G., & Shi, Y. (2011). An empirical study of classification algorithm evaluation for financial risk prediction. *Applied Soft Computing*, 11(2), 2906–2915. <https://doi.org/10.1016/j.asoc.2010.11.028>
- Perols, J. (2011). Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. *AUDITING: A Journal of Practice & Theory*, 30(2), 19–50. <https://doi.org/10.2308/ajpt-50009>
- Purdila, V., & Pentiuc, S.-G. (2014). Fast Decision Tree Algorithm. *Advances in Electrical and Computer Engineering*, 14(1), 65–68. <https://doi.org/10.4316/AECE.2014.01010>

- Rao, P., & Buteau, S. (2018). Modelling credit and savings behaviour of chit fund participants. *Gates Open Research*, 2, 26. <https://doi.org/10.12688/gatesopenres.12767.1>
- Ribeiro, B., Silva, C., Chen, N., Vieira, A., & Carvalho das Neves, J. (2012). Enhanced default risk models with SVM+. *Expert Systems with Applications*, 39(11), 10140–10152. <https://doi.org/10.1016/j.eswa.2012.02.142>
- Römer, U., & Musshoff, O. (2017). Can agricultural credit scoring for microfinance institutions be implemented and improved by weather data? *Agricultural Finance Review*.
- Sáez, J. A., Krawczyk, B., & Woźniak, M. (2016). Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*, 57, 164–178. <https://doi.org/10.1016/j.patcog.2016.03.012>
- Santhisree, V. N., & Prasad, D. J. C. (2014). A Study on Problems of Chit Fund Companies and The Satisfaction Level of the Customers, 7(1), 6.
- Satish, P. (2001). Some issues in the formation of self-help groups. *Indian Journal of Agricultural Economics*, 56(3), 410. Retrieved from <http://0-search.proquest.com.ditlib.dit.ie/docview/201523036?accountid=10594>
- Sharma, D. (2011). Improving the Art, Craft and Science of Economic Credit Risk Scorecards Using Random Forests: Why Credit Scorers and Economists Should Use Random Forests. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.1861535>
- Smeureanu, I., Ruxanda, G., & Badea, L. M. (2013). Customer segmentation in private banking sector using machine learning techniques. *Journal of Business Economics and Management*, 14(5), 923–939. <https://doi.org/10.3846/16111699.2012.749807>
- Sohn, S. Y., & Kim, J. W. (2012). Decision tree-based technology credit scoring for start-up firms: Korean case. *Expert Systems with Applications*, 39(4), 4007–4012. <https://doi.org/10.1016/j.eswa.2011.09.075>
- Su, C., Ju, S., Liu, Y., & Yu, Z. (2015). Improving Random Forest and Rotation Forest for highly imbalanced datasets. *Intelligent Data Analysis*, 19(6), 1409–1432. <https://doi.org/10.3233/IDA-150789>

- Trustorff, J.-H., Konrad, P. M., & Leker, J. (2011). Credit risk prediction using support vector machines. *Review of Quantitative Finance and Accounting*, 36(4), 565–581. <https://doi.org/10.1007/s11156-010-0190-3>
- Tsai, C.-F. (2009). Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, 22(2), 120–127. <https://doi.org/10.1016/j.knosys.2008.08.002>
- Tsai, K. S. (2004). Imperfect Substitutes: The Local Political Economy of Informal Finance and Microfinance in Rural China and India. *World Development*, 32(9), 1487–1507. <https://doi.org/10.1016/j.worlddev.2004.06.001>
- Twala, B. (2010). Multiple classifier application to credit risk assessment. *Expert Systems with Applications*, 37(4), 3326–3336. <https://doi.org/10.1016/j.eswa.2009.10.018>
- Van-Sang, H., & Ha-Nam, N. (2016). Credit scoring with a feature selection approach based deep learning. In *MATEC Web of Conferences* (Vol. 54). EDP Sciences.
- Whiting, D. G., Hansen, J. V., McDonald, J. B., Albrecht, C., & Albrecht, W. S. (2012). MACHINE LEARNING METHODS FOR DETECTING PATTERNS OF MANAGEMENT FRAUD: DETECTING PATTERNS OF MANAGEMENT FRAUD. *Computational Intelligence*, 28(4), 505–527. <https://doi.org/10.1111/j.1467-8640.2012.00425.x>
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining, 11.
- Xanthopoulos, S. Z., & Nakas, C. T. (2007). A generalized ROC approach for the validation of credit rating systems and scorecards. *The Journal of Risk Finance*, 8(5), 481–488. <https://doi.org/10.1108/15265940710834762>
- Yang, Y. (2007). Adaptive credit scoring with kernel learning methods. *European Journal of Operational Research*, 183(3), 1521–1536. <https://doi.org/10.1016/j.ejor.2006.10.066>
- Yap, B. W., Ong, S. H., & Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications*, 38(10), 13274–13283. <https://doi.org/10.1016/j.eswa.2011.04.147>

- Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3), 6527–6535. <https://doi.org/10.1016/j.eswa.2008.07.035>
- Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, 8.
- Yusuf, N. (2014). Role of rural finance in reduction of poverty in the Agriculture sector: Northern India. *International Journal of Business and Economic Development (IJBED)*, 2(2).
- Zaghdoudi, T. (2013). Bank failure prediction with logistic regression. *International Journal of Economics and Financial Issues*, 3(2), 537.
- Zekić-Sušac, M., Šarlija, N., Faculty of Economics in Osijek, University of Josip Juraj Strossmayer in Osijek, Gajev trg 7, 31000 Osijek, Croatia, Has, A., Faculty of Economics in Osijek, University of Josip Juraj Strossmayer in Osijek, Gajev trg 7, 31000 Osijek, Croatia, Bilandžić, A., & Faculty of Economics in Osijek, University of Josip Juraj Strossmayer in Osijek, Gajev trg 7, 31000 Osijek, Croatia. (2016). Predicting company growth using logistic regression and neural networks. *Croatian Operational Research Review*, 7(2), 229–248. <https://doi.org/10.17535/corr.2016.0016>

8. APPENDICES

A. Dataset Descriptions

Subscriber Dataset

Provides the details of all subscribers.

Variable	Description
GroupNo	Name of the group
SubID	Subscriber ID within chit fund
SubNo	Subscriber Number
EnrlDate	The date when the subscriber is registered in the chit group
CustomerCode	Code of customer
Intimation	Method of notice and reminders
IDCard	Identification document used
BranchCode	Branch Code
Status	Subscriber Status
ChitNo	Chit Number
OpenInstNo	Opening Instalment Number
OpenBal	Opening Balance
PRZFlag	Flag to indicate if this subscriber has won the prize in the specified chit group
ACClosed	Flag to indicate whether the Account of the subscriber has been closed
PRZDate	Date when the subscriber won the prize money
GDetail	Guarantor's details taken at the time of prize distribution
SecurityMode	The form of collateral taken from the subscriber
ClosedGroup	Is the chit group closed Y/N?
CustomerName	Name of the subscriber
ParentType	Relationship with the parent
ParentName	Name of the parent or next of kin

Age	Age of the subscriber
PCity	Permanent Address of the subscriber
CCity	Current Address of the subscriber
LandMark	Nearest Landmark
NOJ	Nature of Job
FirmName	Employer address
Dept	Work Department
Desig	Designation of work
BPay	Basic Pay
NPay	Net Pay
BNature	Profession
Capital	Income
ITPayee	Industry
AIncome	Annual Income
OSource	Other income sources
IncomeSrc	Source of Income

A.1 Description of Subscriber Dataset

Auction Dataset

Provides details on when each instalment for each chit fund is due. The dataset also provides the subscriber who won the prize for the dataset that month.

Variable	Description
AuctionID	Unique identifier of the auction
GroupNo	Name of the group
AuctionDate	Date of the auction
InstNo	Instalment number of the auction
Prize	The winning prize amount to be paid to the winner
Commission	The commission paid to the chit company
Dividend	The total dividend to be distributed among all the subscribers
Discount	The winning bid discount offered

AuctionAmount	Total chit value available in the auction
UnDivBF	Carryover dividend from previous cycle
DivDistribute	Total dividend to be distributed among the subscribers of the chit group
DivSubscriber	Dividend per subscriber
UnDivCO	The remainder dividend, is carried forward as undistributed dividend to be adjusted in the next cycle
NextAucDate	Date of next auction
NextSubAmount	The next subscription amount to be collected from the subscribers of the chit group
SubNo	The subscription number who won the prize

A.2 Description of Auction Dataset

Group Dataset

Provides details of each chit fund group.

Variable	Description
GroupNo	Name of the group
PSNo	Unknown
AgrNo	Unknown
AgrDate	Unknown
CCDate	Chit certificate date
ComDate	Chit commencement date
Inst2nd	Second instalment flag
FADate	Date of first auction
ChitAmount	The total chit value
AucTimeFrom	The time of the auction
ATFAMPM	AM/PM Flag
AucTimeTo	The time of the auction
ATTAMPM	AM/PM Flag
NoofInst	Total number of instalments in the chit group

AucDay	The day of the month when auction is conducted
TermDate	Date when the group was completed
InstAmount	Monthly instalment amount
FDRDetails	Ignore- office use
CompTicketNo	The subscription number assigned to the chit company
CompComm	The % comission deducted by the chit company
GroupClosed	Not Updated
FirmCode	Office use- can ignore
PSDate	Office use- can ignore
Inst1st	First instalment flag
OpenBal	Ignore this

A.3 Description of Group Dataset

Transaction Dataset

Provides details of each payment made to the chit fund group, the date, amount and method of payment are included.

Variable	Description
ChitTransDate	Date of the transaction
SubscriberNo	Subscription number of the customer (or subscriber) within the chit group
CustomerName	Name of the subscriber
SubAmount	The amount paid as the subscription amount for the relevant instalment
RealDate	Actual date the payment was received
InstNo	The instalment number for which the payment was made
GroupID	Name of the group
ChitTransID	Transaction reference
TransID	Transaction reference of payment
TransType	The type of the transaction

RcptMode	Mode of payment
CustCode	Code of customer
Status	Subscriber Status

A.4 Description of Transaction Dataset

Tender Dataset

Provides details of each bid made by the subscribers for the monthly prize.

Variable	Description
TenderID	A unique identifier of the bid entry
AuctionID	Unique identifier of the auction
GroupNo	Name of the group
SubscriberNo	Subscription number who placed the bid
Discount	The bid discount offered
PriceFlag	Flag to indicate the prize winner

A.5 Description of Tender Dataset