TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies
Department of Software Science

Henri Ots 143892

COMPARING DIFFERENT QUANTITATIVE METHODS FOR A CREDIT SCORING MODEL BASED ON MOBILE DATA

Master's thesis

Supervisor: Innar Liiv, Ph.D
Associate Professor;
Diana Tur, MA

Tallinn 2018

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond
Tarkvarateaduse instituut

Henri Ots 143892

ERINEVATE KVANTITATIIVSETE MEETODITE VÕRDLUS KREDIIDISKOORINGU MUDELI LOOMISEKS MOBIILIANDMETE PÕHJAL

magistritöö

Juhendaja: Innar Liiv, Ph.D
Dotsent;
Diana Tur, MA

Tallinn 2018

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the materials used, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Henri Ots

04.05.2018

Abstract

The main question that concerns lenders is to whom they lend. Before 1970, credit assessment mainly relied on the work of credit officers. They looked at the customer's character and their collateral to decide on the customer's creditworthiness. For the past three decades, however, financial risk forecasting has been one of the main areas of growth in statistics and probability modelling as well. The major concern of consumer loan companies today is how could they reach unbanked customers. At the moment, more than 2 billion people in world do not have a bank account [1] and at the same time we have reached 6 billion mobile phone users worldwide [2].

One way of reaching unbanked customers is by using their mobile data to calculate their credit risk. This thesis gives an exploratory overview of the state of the art of credit scoring using mobile data.

The aim of this study is to prove that mobile data can be used to make predictions and find the best classification method for credit scoring even if the dataset is small (2,503 customers).

We use different classification algorithms to split customers into paying and non-paying ones using mobile data, and then compare the predicted results with actual results. There are three related works publicly accessible in which mobile data has been used for credit scoring, but they are based on a large dataset. Small companies are unable to use datasets as large as those used by these related papers, and so these studies are of no use for them. In this thesis we try to prove that there is value in mobile data for credit scoring even if the dataset is small.

We found that with a dataset that consists of mobile data based only on 2,503 customers, we can predict if there is credit risk. The best classification method gave us the result 0.62 AUC (area under the curve).

This thesis is written in English and is 64 pages long, including 4 chapters, 0 figures and 7 tables.

Annotatsioon

Erinevate kvantitatiivsete meetodite võrdlus krediidiskooringu mudeli loomiseks mobiiliandmete põhjal

Peamine küsimus, mis täna laenuandjaid vaevab, on kellele laenu anda. Enne 1970 aastat oli krediidiotsuse tegemine peamiselt krediidihaldurite pärusmaa. Nad vaatasid kliendi iseloomu ja tagatist, et teha otsus kliendi krediidivõimekuse kohta. Peamine probleem, millega tarbimislaenu ettevõtted täna tegelevad, on leida võimalusi jõudmaks ka sellise kliendi segmendini, kellel puudub panga ajalugu. Hetkel on maailmas rohkem kui 2 miljardit inimest, kellel ei ole oma pangakontot [1], kuid samal ajal on rohkem kui 6 miljardit inimest, kellel on mobiiltelefon [2].

Üks võimalusi jõuda panga ajaloota klientideni on kasutada selle segmendi mobiilide metaandmeid, et välja arvutada krediidirisk. Käesolev uurimustöö annabki ülevaate kõige uuematest tehnoloogiatest krediidiskooringus kasutades mobiiliandmeid.

Antud töö eesmärk on tõestada, et mobiiliandmeid kasutades saab ennustada krediidiriski isegi siis, kui tegemist on ainult 2503 inimese andmetega ning selle läbiviimiseks on parim klassifikatsiooni meetod.

Sarnasel teemal avalikke töid, kus kasutatakse krediidiskoori ennustamiseks mobiiliandmeid on hetkel kolm, kuid nad kõik baseeruvad väga suurtel andmebaasidel. Antud uurimistööga soovisin tõestada, et mobiiliandmetest on võimalik leida väärtuslikku infot krediidiotsuse tegemiseks ka väikse hulga klientide andmete kasutamisel. Uurimustöö tulemusel leidis tõestamist, et 2503 kliendi mobiili andmete põhjal on võimalik ennustada isiku krediidiriski. Parim klassifikatsiooni meetod andis tulemuse 0.62 AUC.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 64 leheküljel, 4 peatükki, 0 joonist, 7 tabelit.

List of abbreviations and terms

CRMC	The credit risk management cycle.
AUC	Area under the curve.
MDS	Management decision systems.
FI	Fair Isaac.
SMOTE	Minority over-sampling technique.
ENN	A nearing neighbour cleaning rule.
CART model	Classification and regression tree.
SVM	Support vector machine.
KKT	Karush–Kuhn–Tucker conditions.
ROC	Receiver operating characteristic.
OS	Operating system.
SDK	Software development kit.
VC-dimension	Vapnik–Chervonenkis dimension.
p-value	Probability value.

Table of contents

1 Introduction 9
2 Theoretical background of credit scoring 11
2.1 Mobile data for credit scoring 18
2.2 Related works of credit scoring based on mobile data 18
2.3 Overview of different credit scoring methods 20
2.4 Measures 26
3 Empirical research 29
3.1 Data 29
3.2 Experiment design 31
3.3 Experiments 33
3.4 Results 36
4 Conclusion 38
References 40
Appendix 1 – Logistic Regression Classifier 45
Appendix 2 – Decision Tree Classification 49
Appendix 3 – Random forest Classification 53
Appendix 3 – Support vector machine Classification 57
Appendix 3 – Artificial Neural Network 61

List of tables

Table 1. Timeline of credit history. 12
Table 2. Application scorecard example. 16
Table 3. History of credit scorecards. 21
Table 4. Showing comparative classifier performances with all variables..... 34
Table 5. Showing comparative classifier performances with variables were p>=0.05. . 34
Table 6. Showing comparative classifier performances with variables were p>=0.05. . 35
Table 7. Comparing related works. 37

1 Introduction

For the past three decades, financial risk forecasting has been one of the main areas of growth in statistics and probability modelling as well. People often think of the term ‘financial risk’ in relation with portfolio management when it comes to the pricing of options among other financial instruments. Very specific and effective models have been adopted in financial risk forecasting, which include the omnipresent Black–Scholes model [3] and the Merton bond pricing model, [4] which is seminal. However, very little information is available on the importance of credit and behavioural scoring, which are the applications of financial risk forecasting in consumer lending.

It is noteworthy that a large number of statistical methods employed in the development of credit scorecards are based on traditional techniques such as logistic regression or discriminant analysis. However, massive improvements have been made with nonlinear approaches such as the kernel support vector machine, which is currently being used in credit scoring. These new improvements have helped increase the accuracy and reliability of a high number of credit score cards [5]. Related works have not investigated mobile data for credit scoring thoroughly. There are only three publicly accessible research papers on this subject and the datasets used for them are large. Therefore, this thesis aims to study the suitability of different methods for credit scoring based on mobile data using a smaller dataset to find the best method.

There are more than 2 billion people in world who do not have a bank account [1]. This makes it difficult to perform a credit evaluation exercise for these individuals. With the rise of big data, however, various data alternatives can be used to explain the financial inclusion of these unbanked individuals. For instance, mobile data is a new data alternative that can be employed successfully. Mobile phone data has been regarded as good alternative data for credit scoring.

Customers’ mobile data is therefore regarded as an ideal and a better alternative for credit scoring. The number of mobile phone users has reached 6 billion worldwide and [2] providers have increasingly started to allow researchers and [6] commercial partners [7]

access phone logs. In this thesis, our focus is on mobile data as alternative data for credit scoring to find out if it is valuable for credit risk estimation and what method gives the best results.

The main objective of this thesis is to create a unique credit risk model based on a synthesis of earlier academic research by using mobile data. We first synthesise and theorise credit scoring based on mobile data. We then develop the best method for credit scoring based on mobile data. To solve the issue, we compare and find the statistically most effective algorithms that have been used beforehand in credit scoring with mobile data or for credit scoring based on mobile data. The data used for this empirical work is collected from a European consumer loan company.

The thesis is divided into four chapters giving an overview of the fundamentals of credit scoring, related earlier works, the main credit scoring methods, experiments with chosen credit scoring methods with mobile data, the results, and a conclusion.

2 Theoretical background of credit scoring

Credit scoring can be defined as the use of statistical models in the transformation of relevant data into numerical measures, which inform organisations assessing the credit trustworthiness of clients. Essentially, credit scoring is simply an industrialisation of trust; a logical and further development of the subjective credit ratings first provided by 19th-century credit bureaus. The subjective credit ratings were brought about by the need for objective, fast, and consistent decisions, and were enabled by technological advancements. The demerit of credit scoring lies on its dependence on data and its backward-looking nature. Although credit scores have dominated the automated high-volume low-value environments, credit ratings still imply some degree of subjective input, especially for larger loans to businesses as well as governments [8].

Credit scoring was first adopted in the 1960s as a way to determine whether individuals applying for credit would repay the debt, honour the obligation, and observe all the regulations laid down by the treasury's gatekeeper. Credit scoring in the 1960s was entirely associated with 'accept/reject' decisions brought about by the new-business application process also called application scoring. The meaning of the term has not changed in people's minds five decades after its adoption. However, the 21st century has seen some changes in the use of the term 'credit scoring'. Presently, credit scoring is being used by the majority of people in the description of any statistical models in extension and credit management, including the measurement of risk, response, revenue, and retention, whether for marketing, new-business processing, account management, collections and recoveries, or elsewhere (the credit risk management cycle, or CRMC) [8]. Credit scoring is inseparable from other elements of the decision-making process, despite its close association with risk-assessment models. Table 1 shows the timeline of credit history [8].

Table 1. Timeline of the history of credit.

Date	Event
2000 BC	First use of credit in Assyria, Babylon, and Egypt
1100s	First pawnshops in Europe established by charitable institutions; by 1350 they were being run on commercial concerns
1536	Charging of interest deemed acceptable by the Protestant church
1730	First advertisement for credit placed by Christopher Thornton of Southwark, London. He offered furniture that could be paid off weekly
1780s	First use of cheques in England
1803	First consumer reports by the Mutual Communications Society in London
1832	First publication of the American Railroad Journal
1841	Establishment of Mercantile Agency, an American credit reporting agency
1849	Harrods established as one of the world's first department stores
1851	First use of credit ratings for trade creditors by John M. Bradstreet
1856	Singer Sewing Machines offer consumer credit
1862	Poor's Publishing publishes a Manual of the Railroads of the United States
1869	First American consumer bureau: Retailer, Commercial Agency (RCA) in Brooklyn
1886	Seam established and launches its catalogue in 1893
1906	National Association of Retail Credit Agencies established in the USA
1909	John M. Moody publishes the first credit rating grades for publicly traded bonds
1913	Henry Ford uses production lines to produce affordable automobiles
1927	Establishment of Schufa Holdings AG, the first credit agency in
1934	Germany
1936	First Public Credit Registry (PCR) established in Germany
	R.A Fisher uses statistical methods to discriminate between iris species
1941	David Durand writes a report suggesting statistics could be used to make credit decisions
1942	Henry Wells uses credit scoring at Spigel Inc.
1950	Diners Club and American Express launch the first charge cards
1950s	Sears uses propensity scorecards for catalogue mailings
1956	FI consultancy established in California, USA
1958	First use of application scoring by American Investments
1960s	Widespread adoption of credit scoring by credit card companies
1966	Credit Data Corp. becomes the first automated credit bureau
1970	Fair Credit Reporting Act governs credit agencies
1974	Equal Credit Opportunity Act causes the widespread adoption of credit
1975	scoring
	FI implements the first behavioural scoring system for Wells Fargo
1978	Stannic implements the first vehicle finance scorecards in South Africa
1982	CCN offers Credit Account Information Sharing (CAIS), its consumer credit bureau service
1984	FI develops the first bureau scores used for pre-screening
1987	MDS develops the first bureau scores used to predict bankruptcy
1995	Mortgage securitisers Freddy Mac and Fannie Mae adopt credit scoring

2000	Moody's KMV introduces RiskCalc for financial ratio scoring (FRS)
2000s	Basel II implemented by many banks

[8]

Credit scoring is beneficial to both lenders and borrowers. For example, credit scores help reduce discrimination as credit scoring models afford a more objective analysis of a consumer's creditworthiness. In return, this allows providers of credit facilities to focus mainly on information related to credit risk and avoid the personal subjectivity of a credit analyst or an underwriter [9]. In the United States, under the Equal Credit Opportunity Act, variables of overt discrimination such as race, sex, religion, and age cannot be included in credit scoring models. Rather, the models compose of only information that is non-discriminatory in nature and that has over time been proven to be predictive of payment performance.

Secondly, credit scoring helps in increasing the speed and consistency of the loan application processes and allows lending firms to automate their lending processes [10]. In this case, credit scoring significantly reduces human involvement in credit evaluation and lessens the cost of delivering credit [11]. Moreover, by using credit scores, financial institutions are able to quantify risks associated with granting credit to a particular applicant in a shorter period of time. According to Leonard [12], a study done by a Canadian bank found that the time it took to process a consumer loan application was shortened from nine days to three after credit scoring was used. As such, the optimisation of the loan processing time means that time saved on processing could be utilised to address more complex aspects in the firm. Banaslak and Kiely [13] concluded that with the help of credit scores, financial institutions are able to make faster, better and higher-quality decisions. Furthermore, it also implies that credit scoring can help improve the allocation of resources toward the 'first equilibrium' [14].

Additionally, credit scores can aid financial organisations in determining the interest rate which the firms should charge their consumers as well as the pricing of portfolios [15]. Understandably, in line with the basic financial tenet of risk and return, customers bearing a higher risk are charged a higher interest rate and vice versa. Based on the consumer's credit scores, the financial institution is also able to determine the credit limits to be set for the consumer [16]. This helps the financial institution manage their accounts more

effectively and profitably. Again, profit scoring can be used to maximise profits across a range of products [17], [18].

In line with the above, credit scoring models have fuelled the development of the sub-prime lending industry, where sub-prime customers have poor credit records and fall short of credit acceptance and risk. For instance, these clients may fall short of the requirements set for traditional financing due to credit impairment, missing data in their credit histories, or difficulty in validating their income [19]. One of the main drivers of the progress of sub-prime lending has been automated underwriting. Automated underwriting permits sub-prime mortgage loans to be packaged and sold as investment securities. The initial success of specialised financial institutions in this market has driven more financial institutions to enter the sub-prime lending market. Moreover, the growth trend is expected not to dissipate as technology in credit scoring advances [20].

Technology advancement has led to the development of smarter credit scoring models. Thanks to the technological advancements, credit card issuers are able to utilise the information generated from the models to formulate strategies for collecting credit and hence use their resources more effectively [21].

Finally, the insurance industry has employed the use of credit scoring in streamlining the process of applying for and renewing insurance contracts. In particular, credit scores can aid insurance firms in making better predictions on claims. Therefore, these firms are capable of controlling risk more effectively. As such, insurance companies are also able to price their products more accurately. Moreover, they are able to offer additional insurance coverage to a larger number of consumers at a cost that is more equitable, as well as be able to react and adjust to market changes quickly and gain a competitive edge [22].

Credit scoring is used for:

1. Shifting to high-tech. A structural change in the market has seen a shift of volumes, values, and profits from traditional (relationship) to high-tech (transactional) lenders. Those who were quickest and best at updating their systems forced a shift of higher-risk applicants to lenders less able, increasing the latter’s chances of adverse selection [8].

2. Organisational instability. Companies became unsettled as new processes were implemented. There was a move away from bricks and mortar to travelling salespeople with laptops, from over-the-counter service to Internet banking, and from clerks shuffling snail-mail to PCs and email. As new ways of doing things evolve, the old ways become obsolete [8].

3. Changed skills requirements. The investment is not once-off, but ongoing, and requires the development of a completely different set of skills than those required previously by lenders. The labourers that once shovelled the coal are replaced by technicians that watch the gauges and turn the valves [8].

4. Credit market growth. Credit scoring and decision automation have significantly lowered the cost of extending credit, improved lenders' capacity to service smaller loans, and generally increased service levels. Infrastructure investments are huge, and there has been much industry consolidation to gain economies of scale [8]. This includes credit bureaux, as smaller operators were either swallowed by bigger fish, or beached [23].

Most people think that a scorecard is a piece of paper used in sporting by a scorekeeper, a spectator, or a participant to record the competitors' performance. The results are then used to determine the winner. In credit scoring, the concept remains the same, except to win is to be approved for a loan. The only difference lies in the methods of deriving and applying the scorecards. Credit scoring is the application of predictive models also called algorithms to rank customers by their probability of either being 'good' or 'bad' customers at a future date based on their past performance. The reasoning is clear: a 'good' customer is low-risk and a 'bad' customer high-risk. Lenders welcome low-risk customers with open arms but turn away high-risk customers [8].

Table 2. Example of an application scorecard.

Characteristic	Attributes					Points
Years @ address	<3 years 30	3-6 years 36	>6 years 36		Blank 35	38
Years @ employer	<2 years 30	2-8 years 39	9-20 years 43	>20 years 64	Blank	43
Home phone	given 47				Not Given 30	30
Accom. status	Own 41	Rent 30	Parents 9		Other 36	41
Bankers	Us 42	Then 42			Blank 42	42
Credit card		Bank or travel 75		Retail or garage 43	Blank 43	43
Judgments on bureau	Clear 16	1-16	2-30	3-34		20
Past experience	None 3	New 13	Up-to-date 36	Arrears -1	Write-off Reject	3
						Final score 267

[8]

The series of statements above and Table 2 clearly explain the final scorecard. The table comprises of characteristics (rows) and attributes (columns). An attribute is a set of values or a non-overlapping range of numbers, and the points derived from that attribute are assigned for each case where it holds true. The points are then summed; the higher the score, the lower the risk [8].

As shown in the example, the attributes for a hypothetical customer have been derived, the points assigned, and the final score calculated. The precise cut-off mark for this scorecard is unknown, but it was usually 200 or less, so it is almost certain that a customer with a score of 267 would have been approved. The main advantage of a traditional scorecard is transparency. That is the reason why it has remained the most preferred format since the introduction of credit scoring. Other scorecard formats have been used, but with mixed results. Some formats are valid but have their own merits and demerits [8].

The sector is of great economic importance. For instance, in the European Union the amount of consumer loans and mortgages given to individuals is higher than corporate loans. This indicates that lenders require formal tools to determine ‘bad’ and ‘good’ customers [8].

A credit score is an estimate of the probability calculated using a predictive model that shows the borrowers’ detrimental behaviour in the future. In practice, lenders use predictive models called scorecards to determine the probability of a customer defaulting on payment. The probability of default scorecards are usually developed using a classification of algorithms [24].

In spite of numerous studies, there is no literature on recent advancements in predictive learning. For example, the development of selective classifiers systems capable of pooling different algorithms and optimising their weighting using empirical search signifies an important milestone in machine learning [25]. However, no one has made an attempt to study the usefulness of such an approach in credit scoring. In general, recent advancement focuses on three dimensions: novel classification algorithms, novel performance measures, and statistical hypothesis tests. The first dimension concerns the development of scorecards, for example an extreme learning machine, the second dimension assesses scorecards, for example an H-measure, and the third dimension compares scorecard performance [26].

Numerous literature review focus on the development, application and evaluation of predictive models used in the credit sector [27].

These models determine the creditworthiness of an applicant based on a set of descriptive variables. Corporate risk models use data from a statement on financial position, financial ratios or macro-economic pointers, while retail risk models use data captured in the application form such as the customer’s transaction history [18]. The difference between variables used in corporate and retail models indicates that more challenges arise in consumer than corporate credit scoring. Thus, this thesis focuses on the retail business.

2.1 Mobile data for credit scoring

To what extent can one tell your personality by simply looking at how you use your mobile phone? The use of standard carrier logs to determine the personality of a mobile phone user is a hot topic, which has generated tremendous interest. The number of mobile phone users has reached 6 billion worldwide and [2] service providers are allowing increasing access to phone logs to researchers and [6] commercial partners [7]. If predicted accurately, mobile phone datasets could provide a valuable and cost-effective method of surveying personalities. For example, marketers and phone manufacturing companies might seek to access dispositional information about their customers so as to design customised offers and promotions [28]. The human-computer interface field uses personality. Thus, it benefits from the appraisal of user dispositions using automatically collected data. Lastly, the ability to extract personality and other psychosocial variables from a large population might lead to unparalleled discoveries in the field of social sciences [29].

The use of mobile phones to predict people's personalities is a result of advancement in data collection, machine learning and computational social science which has made it possible to infer various psychological states and traits based on how people use their cell phones daily. For example, some studies have shown that people's personality can be predicted based on the pattern of how they use social media such as Facebook or Twitter [30], [31], [32]. Other researchers have used information about people's usage of various mobile applications such as YouTube, Internet Calendar, games and so on to make conclusions about their mood and personality traits [33], [34], [35], [36] [37]. While these approaches are remarkable, they require access to a wide-ranging information about a person's entire social network. These limitations greatly weaken the use of such classification methods for large-scale investigations [38].

2.2 Related works of credit scoring based on mobile data

There are some studies about the use of mobile phone data in credit scoring globally. In open sources there are only three research papers of mobile data usage for credit scoring:

1. "Behaviour Revealed in Mobile Phone Usage Predicts Loan Repayment", authors: Björkegren and Grissen, 2017.

2. “Mobile phone-based Credit Scoring”, authors: Skyler Speakman, Eric Mibuari, Isaac Markus, Felix Kwizera, 2017.

3. “MobiScore: Towards Universal Credit Scoring from Mobile Phone Data”, authors: Jose San Pedro, Davide Proserpio and Nuria Oliver, 2015.

Björkegren and Grissen use behavioural signatures in mobile phone data to predict default with an accuracy almost similar to that of credit scoring methods that use financial history. The approach was validated using call records matched to loan results for a sample of borrowers in a Caribbean country. Applicants in the highest quartile of risk according to the authors’ measure were six times more likely to default in payment than those in the lowest quartile. They used two different algorithms, Random Forest and Logistic regression. The result obtained with the Random Forest algorithm was 0.710 AUC (area under the curve) and with Logistic regression 0.760 AUC. The dataset included information on 7,068 customers from a South-American country [39].

Jose San Pedro et al. developed MobiScore, a methodology used to build a model of the user’s financial risk using data collected from mobile usage. MobiScore was using data on 60,000 real people obtained from telecommunication companies and financial service providers in a Latin American country. They used gradient boosting, support vector machine and linear regression models to solve the problem. AUC results with different combinations were between 64.1 and 72.5 [40].

Speakman showed how to use boosted decision trees to create a credit score for under-banked populations, enabling them to access a credit facility that was previously denied due to the unavailability of financial data. Their research result was a 55% reduction in default rates while simultaneously offering credit opportunities to a million customers that were given a 0 credit limit in the bank’s original model. The dataset contained 295,926 labelled examples with over 30 categorical and real-valued features. AUC results with the boosted decision trees algorithm were 0.764 and with logistic regression 0.74 [41].

2.3 Overview of different credit scoring methods

Looking at credit scoring history, it is still peculiar to see how the concept came about. It started in 1936 when the English statistician Sir Ronald Aylmer Fisher published an article that detailed the utilisation of a technique called 'linear discriminant analysis' in classifying different species of irises. As years went by, Fisher used the technique in the classification of skulls (he utilised only their physical measurements); he dwelled on differentiating the origins of skulls using their physical requirements. Although Fisher's contribution was based primarily on sciences, his works provided the basis for predictive statistics being used in other disciplines. In 1941, David Durand, another interesting researcher, showed that Fisher's techniques could still be used in the discrimination of good and bad business. According to Johnson [42] Durand's study examined 7,200 reports on good and bad instalment loans made by 27 organisations. The data detailed from Durand's case was based on age, gender, stability (time at address and employment), occupation and industry, and major assets (bank accounts, real estate, life policies) [8].

In this case, Durand in 1947 became the first person to understand that an individual could use Fisher's techniques in discriminating between good and bad loans. Unfortunately, Durand's research project for the US National Bureau of Economic Research was not used in making any prediction [8].

The only people who could make decisions on whether to give loans or send merchandise were the credit analysts. Still these professionals were not sufficient as a large number of them were absorbed in to the military, leading to an acute shortage of credit analysts at that time. To address the problem, organisations had to seek the services of analysts in drafting the procedures and rules on whom to give loans to and whom not to [42].

With the help of analysts, these rules were used by everyone in the organisation whenever they had to make a credit decision. After the war ended, people were able to connect these two events and decipher the benefit of statistically derived models in lending decisions. The beginning of the 1950s saw the formation of the first consultancy in San Francisco by Bill Fair and Earl Isaac. The company's client base at the time was mainly made up of finance houses' retailers and mail order firms [8].

With the arrival of credit cards in the 1960s banks and other credit card issuers started to appreciate the usefulness of credit scoring. This realisation led to high demand for credit

cards by the people, forcing the banks and other credit card issuers to automate the lending decision. Credit scoring became a very useful concept to these institutions as it was possible then to predict as default rates would drop by 50% or more – see Myers and Forgy [43] for an early report on such success [8].

Towards the 1980s, the success of credit scoring in credit cards implied that banks would start using credit scoring in issuing personal loans, unlike in the recent past where they would use credit scoring to issue home loans and small business loans. Furthermore, the progressive growth in direct marketing in the 1990s led to the adoption of scorecards which improved the response rate to advertising campaigns [8].

Table 3. History of credit scorecards.

Name	Year	Notes
Fair Isaac (FI)	1956	Founded San Francisco,
FI		CA, by Bill Fair and Earl Isaac
	1958	First scorecard development for American Investments
	1984	Develops first bureau score for pre-screening
	1995	First use of scoring by mortgage securities
Experian Scorex		
Management	1974	Founded by John
Decision		Coffman and Cary Chandler
		Systems
System (MDS)	1982	MDS purchased by CCN
Scorex	1984	Founded in Monaco by
MDS	1987	Jean-Michel Trousse
ExperianScorex	2003	Created as subsidiary of Experian, after purchase of Scorex

[8]

Literature proposes a wide range of classifications techniques for scoring credit data sets. The issue of good or bad distribution is the most appropriately solved during the classification of datasets and the literature on machine learning and data mining have discussed it. According to Weiss and Provost [44], the naturally arising class distribution

in the 25 datasets examined did usually not give the best performing classifiers. More precisely, based on the AUC measure, it was revealed that the best class distribution should comprise between 50% and 90% minority class examples in the training set. Provost, Jensen and Oates [45] proposed a progressive adaptive sampling strategy for selecting the best class distribution. Whereas this approach of adjusting class can be very useful in large datasets, with sufficient observations in the smaller class of defaulters, in some low default portfolios there is a very small number of defaulters in the first place.

Literature has compared several kinds of techniques in an attempt to establish the most effective method to overcome a large class imbalance. Chawla, Bowyer, Hall, and Kegelmeyer [46] came up with a synthetic, minority over-sampling technique (SMOTE), which was used in example datasets in fraud, telecommunications management, and detection of oil spills using satellite images. Japkowicz [47] compared over-sampling and downsizing with his own method of 'learning by recognition' in order to establish the most effective technique. The results, however, were not conclusive but established that both over-sampling the minority class and downsizing the majority class can be effective. Consequently, Batista [48] presented ten alternative methods in addressing class imbalance and exampled them on thirteen datasets. The methods chosen involved a variety of under-sampling and over-sampling techniques. Findings showed that in general, the over-sampling technique gives more accurate results than under-sampling. Similarly, a combination of either SMOTE and ENN (a nearing neighbour cleaning rule) or SMOTE and Tomek links was recommended [46].

According to an experimental comparison of classification algorithms for imbalanced credit scoring datasets we describe in this work the same classification algorithms, as they are the main classification algorithms for credit scoring, and compare how would they work for credit scoring based on mobile data.

Logistic regression is one of the most important techniques widely used in credit scoring across sectors all over the world [24]. This technique is also effective tool applied in social sciences for categorical data analysis. Siddiqi [49] argued that in credit scoring a customer's default probability is a function of specific social-economic variables such as income, employment, marital status and other behavioural characteristics as shown in their default history. Hence, credit scoring can be viewed as a twofold classification of problems using logistic regression as a suitable technique. Jentzsch [50] observed that

logistic regression requires fewer assumptions as compared to discriminant analysis. Indeed, the only constraint when using logistic regression is that the output variable must be binary and that there is no multicollinearity among the independent variables.

At face value, logistic regression can be seen as linear regression. However, what differentiates logistic regression from linear regression is the outcome. For logistic regression, the outcome variable is dichotomous, while with linear regression, the outcome can be any real number [51].

This study focuses on a twofold response of whether a borrower turns out to be a good or bad customer (non-defaulter vs defaulter). For this binary response technique, the dependent variable y can take either of the two forms; y=0 bad payer and y=1 good payer. Now let us assume that x is a column vector of M descriptive variables and the response probability is modelled as P=Pr(y=1/x). N denotes the number of observations. The logistic regression model formula (1) is described below.

$$logit(\pi) \equiv log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta^T x, \quad (1)$$

According to the author's previous experience, logistic regression is one of the most widely used techniques for scoring and it usually yields the most stable result. Logistic regression results can be easily interpreted, implemented and presented.

The recent method used to develop a credit score is the decision tree technique [24]. A decision tree is made up of a set of sequential binary splits on the history of the customer's dataset with the aim of realising binary classification. Thus, a dichotomous tree is developed by splitting the customer's histories at each node based on a function of a single input [52]. There are diverse opinions concerning the use of the decision tree technique. Breiman, Friedman, Olshen, and Stone [53] were instrumental in the widespread and popularity of the decision tree. However, Rosenberg and Gleit [54] claim that Harvard Business School scholars Raiffa and Schlaifer [55] were the first to use the CART model (Classification and regression tree).

A decision tree technique is basically a fixed acyclic graphical model. To be of use in a binary classification as a fixed (rooted) tree, the model must meet the following criteria. There must be only one node without edges entering it (this node is called the root), every

node excluding the root has to have only a single edge entering it, and there must be an exclusive path from the root to each node.

In addition, the decision tree model is a non-parametric technique. This simply means that the system is unable to learn parameters upon which to score the borrower's attributes. Instead, the model memorises specific key characteristics about data. According to Arminger, Enache, and Bonne, [56] and Hand and Jacka, [57] the goal of the recursive partitioning technique was to reduce cost. Therefore, the method considers all possible splits to establish the optimal, and the best sub-tree fitted is selected based on the total bias rate or the lowest cost of miscalculation [58]. The classification of a new borrower is established based on the classification of the subsequent leaf node once traversing the tree model using the borrower's attributes as input. The decision tree model formula (2) is described below.

$$Entropy\ (S) = \ -p_1\ log_2(p_1) - p_0\ log_2\ (p_0) \tag{2}$$

According to [59], [60] gradient boosting is an ensemble algorithm that enhances the correctness of a predictive function using incremental minimisation of the bias term. After the early base learner is grown, each tree in the sequence is fitted to the alleged 'pseudo residuals' of the prediction from the initial trees with the aim of minimising the bias. This gives the following model the gradient boosting model formula (3).

$$F(x) = G_0 + \beta_1 T_1(x) + \beta_2 T_2(x) + \cdots + \beta_n T_n(x), \tag{3}$$

The author's previous practical tests using gradient boosting for credit scoring have not yielded results that were good could be interpreted easily.

This is defined as group of un-pruned classification or regression trees, trained on bootstrap samples of the training data using a random feature section in the course of tree generation. Once a large number of trees has been produced, the most popular class is voted by each tree. These tree-voting processes are jointly defined as random forests. For the random forest classification method, two parameters need tuning: the number of trees and the number of attributes used to grow each tree [61].

Support vector machines are a set of authoritative controlled learning methods applied in classification and regression. Their primary standard is to build a maximum-margin splitting hyperplane in some transformed feature space. Instead of calling for one to stipulate the precise transformation, they apply the principle of kernel substitution to change them to a general (non-linear) model. The LS-SVM (support vector machine) proposed by Suykens, Van Gestel, DeBrabanter, De Moor, and Van Dewalle [62] is an advanced use of Vapnik's original SVM method which enables us to solve linear KKT (Karush-Kuhn-Tucker conditions) systems.

The optimisation problem for LS-SVM takes the form (4):

$$\min_{w,b,c} J(w,b,e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2, \quad (4)$$

Subject to these equality limitations (5):

$$y_i[w^T \varphi(x_i) + b] = 1 - e_i, i = 1, \dots, N, \quad (5)$$

With w representing the weight vector in primal space, the normalisation parameter is denoted by c and yi=+1 or 1 for good customers [62]. The answer can then be found after constructing the Lagrangian and selecting a certain kernel function that calculates inner products in the transformed space, depending on which classifier of the following method is arrived at (6).

$$y(x) = sign\left[\sum_{i=1}^{N} \alpha_i \, y_i \, K(x, x_i) + b\right] \quad (6)$$

According to Bishop, [63] neutral networks are mathematical formulas modelled on how human brains function. Neutral networks have the added advantage of flexibility; the technique is flexible in modelling almost any non-linear relationship between dependent and independent variables. Even though numerous architectures have been suggested, this thesis concentrates on the most widely used type of neutral networks called the multilayer perceptron. A multilayer perceptron is normally comprised of an input layer, a hidden layer and an output layer. Each neuron in the layer process inputs and gives output value to the next layer. Such connections are allotted a weight during training. In the logistic function, it takes the form:

$$h_i = f^{(1)}\left(b_i^{(1)} + \sum_{j=1}^{M} w_{ij}\, x_j\right), \quad (7)$$

Where W is a weight matrix, and the weight connected input j to the hidden neuron i is denoted by Wij. We will make a binary prediction for the analysis to be done in this thesis, therefore, for the activation function in the output layer (8).

$$f^{(2)}(x) = \frac{1}{1+e^{-1}} \quad (8)$$

We will be applying the logistic sigmoid activation function to derive a response probability (9).

$$\pi = f^{(2)}\left(b^{(2)} + \sum_{j=1}^{n_h} v_j\, h_j\right) \quad (9)$$

Where nh is the number of hidden neurons, v the weight vector, vj the weight connecting the hidden neuron j to the output neuron.

In the course of model estimation, first we randomly initialised and iteratively adjusted the weight of the network in order to minimise an objective function, for example totalled squared errors accompanied by the regularisation term to avoid over-fitting. Simple gradient descent learning or complex optimisation techniques such as Quasi-Newton or Levenberg-Marquardt are the basis for the iterative procedure. A grid search grounded on validation set performance can be used to determine the number of hidden neurons [63].

2.4 Measures

Harris [64] notes that in the process of developing and reporting the credit scoring models it is pragmatic to differentiate between the training and the reporting phase. This is due to the need of the person to provide clarity on the type of the metric that was initially applied in the selection of model parameters. When denoting the metric adopted, it would be sensible to use the term evaluation metric in the training process. On the other hand, to report the model performance during the performance phase, the term performance metric will be adopted [64].

In this analysis, both the performance metric and the primary model evaluation metric are represented by the region under the ROC (Receiver Operating Characteristic) curve called AUC. The ROC curve, often adopted by the AUC, illustrates a two-component aspect of differential performance where the sensitivity (10) (i.e. the relative amount of the actual positives which is forecasted as positive) and the specificity (11) (i.e. the proportion of actual negatives that are forecasted as being negative) are plotted on the Y and X axis, respectively. Normally, the AUC figure is demonstrated as in (12) the figure below where S1 illustrates the total sum of the customer's creditworthiness rank. In this, a score of 100% shows that the person classifying can impeccably differentiate between the classes, and a score of fifty percentage shows a classifier possessing a minor quality of differentiation [65].

$$\text{Sensitivity} = \frac{true\ positive}{true\ positive + false\ negative} \tag{10}$$

$$\text{Specificity} = \frac{true\ negative}{true\ positive + true\ negative} \tag{11}$$

$$\text{AUC} = \frac{(S_{1-sensitivity}) + [(sensitivity + 1) + 0.5]}{sensitivity + specificity} \tag{12}$$

Different metrics can also be applied and used to produce the working of the categories used herein. For instance, to check for the correctness (13) it can also be taken to be the measure of how correct those applying for credit on a held back data test are classified.

Several performances are often applied when reporting the performance of the classifier developed in this analysis. For instance, the test accuracy below has also been reported to be a measure of how precise the applicants of credit is. Moreover, the balanced accuracy data represented in the figure (14) gives the entire meaning of the classifier performance. The quantifier circumvents the ambiguous impact on the accuracy which is brought about by uneven datasets by illustrating the arithmetic average of specificity and sensitivity. Subsequently, the slanted datasets are familiar, similar to what is happening with actual world credit, which scores the datasets making it irrelevant [65].

$$\text{Test accuracy} = \frac{True\ positive}{True\ positive + True\ negative} + \frac{True\ nagative}{false\ negative + true\ negative} \quad (13)$$

$$\text{BAC} = \frac{Specificity + Sensitivity2}{2} \quad (14)$$

3 Empirical research

3.1 Data

The dataset comprises of information on 2,503 customers who have obtained a consumer loan, and allows one to understand their previous payment behaviour. Any means of identification have been entirely removed from the data and consequently anything personal has been scrapped off. Information was initially obtained with the consent of the customers. The dataset was collected from customers of a European consumer lending company who use the mobile application of this lending company. The customers gave their permission to use the data for credit scoring.

Using their payment behaviour we are able to separate the trustworthy customers from the untrustworthy ones. Our target variable identifies untrustworthy customers as those, who have got a 90-day delay in payment of their instalments. Additionally, the dataset will include about 1,516 trustworthy customers without debts that exceed the 90-day limit. Conclusively, this will result in the percentages of the trustworthy and untrustworthy customers being 60.57% and 39.43%, respectively.

Android phone users can be requested to yield the following data about their device (see Table 4). For this research we did not use phone numbers, calendar body texts or text messages (SMS).

Table 4. Raw data from Android phones.

Data group	Data description
Device	Device ID
Device	OS (operating system) version
Device	SDK (software development kit) version
Device	Release version
Device	Device
Device	Model
Device	Product
Device	Brand
Device	Display
Device	Hardware
Device	Manufacturer
Device	Serial
Device	User
Device	Host
Network	Network ID

Network	Carrier
Network	Operator
Network	Subscriber
Calendar	Calendar ID
Calendar	Title
Calendar	Date
Calendar	Body
Call info	Caller ID
Call info	Receiver (contact/unknown)
Call info	Type (incoming/outgoing/missed/unanswered)
Call info	Number
Call info	Date
Call info	Duration
Contact info	Contact ID
Contact info	Contact number
Installed apps	App ID
Installed apps	Package name
Installed apps	Label
Installed apps	Version name
Installed apps	Version code
Installed apps	Install date
SMS info	SMS ID
SMS info	Receiver (contact/unknown)
SMS info	Type (incoming/outgoing/missed/unanswered)
SMS info	Conversation
SMS info	Number
SMS info	Message length
SMS info	SMS date
Images	Image ID
Images	Image date
Images	Image location
Data storage	Data storage ID
Data storage	Path
Data storage	Last modified

From among all the varying parameters, 22 variables were selected to be used in the experiments necessary for the research (the variables are shown in Table 5). The variables were chosen by using manual review and statistical analysis of dependencies. We chose variables that were less dependent on each other. Using these variables, one of them is a categorical variable while others are numerical. In some experiments we calculated some numerical variables into bins so that their data type changed to categorical.

Table 5. Variables for experiments.

Data group	Calculated data points	Data type
Call info	Average number of calls per month.	Numerical
Call info	Average number of incoming calls per month.	Numerical
Call info	Average number of outgoing calls per month.	Numerical
Call info	Average number of missed calls per month.	Numerical
Call info	Average number of unanswered calls per month.	Numerical
Call info	Average call duration.	Numerical
Call info	Average outgoing call duration.	Numerical
Call info	Average incoming call duration.	Numerical
Call info	Maximum outgoing call duration.	Numerical
Call info	Maximum incoming call duration.	Numerical
Images	Average number of images per month.	Numerical
Images	Average number of images made in distinct places per month.	Numerical
SMS info	Average number of SMSs per month.	Numerical
SMS info	Average number of incoming SMSs per month.	Numerical
SMS info	Average number of incoming SMSs per month from contacts.	Numerical
SMS info	Average number of incoming SMSs per month from an unknown number.	Numerical
SMS info	Average number of outgoing SMSs per month.	Numerical
SMS info	Average number of outgoing SMSs per month from a contact.	Numerical
SMS info	Average number of outgoing SMSs per month from an unknown number.	Numerical
SMS info	Average number of SMS conversations per month.	Numerical
Contacts	Number of contacts.	Numerical
Device	SDK version.	Categorical

3.2 Experiment design

We carried out four experiments with five different classification methods and considered AUC to be the performance parameter. As the author's previous experiences have illustrated, there are no specific rules for working with alternative data. Accordingly, we carried out four experiments based on different pre-processing techniques.

In the first experiment we included all the calculated variables. The SDK variable, which is categorical, needs to be encoded. The SDK data has to be converted into numbers to make them comparable. The SDK version comprises six different values (19, 20, 21, 22, 23, 24, 25), for which we generated dummy variables. The values of these parameters are either 1 or 0. As a result, there can be no missing information in the dataset. The second

step in the data pre-processing is to scale all variables to make them comparable with each other.

In the second experiment we used the same pre-processing techniques as in the first experiment, but we added backward elimination. The principle of Occam's razor states that the [66] model needs to be as simple as possible until it achieves an acceptable level of performance on training data. This will help to avoid over-fitting the model. With backward elimination we can throw out variables with p-value (probability value) >0.05 and the highest p value. After that we can calculate a new combination of p values and continue the same process until we have a set of variables, all with p lower than 0.05.

In the third experiment we used the same pre-processing method as before but modified the variables. We used the optimal binning technique to group the variables. Optimal binning is a method of pre-processing categorical predictors where we set values for variables by grouping them into optimal bins. Its purpose is to reduce the impact of statistical noise [67].

In order to choose the classifier methods for the experimental part we used three parameters:

- How have they functioned in previous credit scoring research?
- How have they functioned in previous credit scoring research using mobile data?
- How have they functioned in the author's practical work in credit scoring models?

According to these three parameters we chose for our experiments the following methods: logistic regression, decision tree, random forest, SVM and neural networks.

When organising benchmarks in pattern recognition, there is often the problem of determining the size of the test set that would give statistically significant results. The commonly adopted ratio is 8:2 according to the Pareto principle. According to research by Isabelle Guyon and the formula she found we can determine the example test size. The fraction of patterns reserved for the validation set should be inversely proportional to the square root of the number of free adjustable parameters. The ratio of the validation set (v)

to the training set (t) is v/t, and the scales are ln(N/h-max), where N is the number of families of recognisers, and h-max is the largest complexity of those families. Each family of recognisers is characterised by its complexity, which may or may not be related to the VC-dimension (Vapnik–Chervonenkis dimension), the description length, the number of adjustable parameters, or other measures of complexity [68].

According to a small sample size of customers we chose three different test size examples for this research. The test sizes we chose were 10%, 25% and 40%.

Testing any combination of variables first results in all variables. We then chose only the variables with p>=0.05 and finally binned the variables with p>=0.05. The intervals of the variables can be determined in a variety of ways. For example, by using prior knowledge on the data. The boundaries of the intervals are usually defined beforehand.

3.3 Experiments

The experiments described in this chapter were done using the Python programming language and the Spyder environment. We also used numpy, matplotlib, panda and sklearn Python libraries for statistical analyses.

Tables 6, 7 and 8 show a representation of the performance of classification methods using mobile data. The results in Table 6 show the classifiers' performances with all variables. The results in Table 7 show the classifiers' performance with only the variables whose p value is higher than 0.05. Table 6 shows binned variables whose p value is higher than 0.05. Table 8 shows manually binned variables whose p value is higher than 0.05. The tables suggest the models created for the prediction of creditworthiness as illustrated by AUC on the suppressed datasets. To determine the importance of the variation in performance between the models we can take AUC as the main parameter to see which model had the best performance. Tables 6, 7 and 8 can also be compared for training accuracy, test accuracy and training time(s).

The target variable chosen was 0 for a performing customer and 1 for a non-performing customer. A non-performing customer in this research is set as one who is 90 or more days overdue in paying their debt. According to Barisitz, [69] the rule of being 90 days overdue is most common in the European country from which the data for this research were collected.

Table 6. Showing comparative classifier performances with all variables.

Classifier	Training accuracy	Test accuracy	AUC	Training time (s)
Logistic regression				
Test size= 0.10	0.62	0.62	0.51	0.005
Test size= 0.25	0.61	0.62	0.55	0.005
Test size= 0.40	0.63	0.61	0.56	0.001
Decision tree				
Test size= 0.10	1.0	0.57	0.54	0.093
Test size= 0.25	1.0	0.56	0.54	0.074
Test size= 0.40	1.0	0.56	0.54	0.052
Random forest				
Test size= 0.10	0.98	0.63	0.62	0.103
Test size= 0.25	0.98	0.60	0.52	0.076
Test size= 0.40	0.98	0.61	0.58	0.059
SVM				
Test size= 0.10	0.61	0.65	0.56	3.33
Test size= 0.25	0.60	0.59	0.56	2.15
Test size= 0.40	0.60	0.59	0.57	1.22
Neural networks				
Test size= 0.10	0.69	0.60	0.59	100,63
Test size= 0.25	0.67	0.59	0.57	69,79
Test size= 0.40	0.69	0.61	0.55	51.71

Table 7. Showing comparative classifier performances with variables were p>=0.05.

Classifier	Training accuracy	Test accuracy	AUC	Training time (s)
Logistic regression				
Test size= 0.10	0.68	0.62	0.54	0.004
Test size= 0.25	0.76	0.56	0.53	0.003
Test size= 0.40	0.77	0.55	0.50	0.003
Decision tree				
Test size= 0.10	1.0	0.56	0.53	0.032

Test size= 0.25	1.0	0.55	0.53	0.029
Test size= 0.40	1.0	0.58	0.55	0.023
Random forest				
Test size= 0.10	0.98	0.66	0.56	0.064
Test size= 0.25	0.98	0.62	0.59	0.056
Test size= 0.40	0.97	0.60	0.58	0.045
SVM				
Test size= 0.10	0.60	0.65	0.53	1.305
Test size= 0.25	0.60	0.59	0.56	1.094
Test size= 0.40	0.60	0.59	0.57	0.678
Neural networks				
Test size= 0.10	0.64	0.61	0.57	84.36
Test size= 0.25	0.63	0.60	0.55	89.72
Test size= 0.40	0.65	0.60	0.58	62.62

Table 8. Showing comparative classifier performances with variables were p>=0.05.

Classifier	**Training accuracy**	**Test accuracy**	**AUC**	**Training time (s)**
Logistic regression				
Test size= 0.10	0.68	0.62	0.54	0.004
Test size= 0.25	0.76	0.56	0.53	0.003
Test size= 0.40	0.77	0.55	0.50	0.003
Decision tree				
Test size= 0.10	0.74	0.59	0.53	0.06
Test size= 0.25	0.76	0.56	0.53	0.003
Test size= 0.40	0.78	0.55	0.50	0.003
Random forest				
Test size= 0.10	0.73	0.58	0.54	0.22
Test size= 0.25	0.75	0.53	0.52	0.022
Test size= 0.40	0.76	0.54	0.50	0.021
SVM				

Test size= 0.10	0.60	0.65	0.49	0.007
Test size= 0.25	0.59	0.59	0.48	0.238
Test size= 0.40	0.60	0.59	0.48	0.168
Neural networks				
Test size= 0.10	0.61	0.63	0.51	84.40
Test size= 0.25	0.61	0.62	0.54	64.33
Test size= 0.40	0.62	0.58	0.45	53.82

3.4 Results

The empirical results consist of the performance estimates of five classifiers with four different combinations. The tables on the previous page report the AUCs of all five classifiers with all four experiment combinations.

Random forest provides the best average AUC level across experiments with different test sizes. Random forest also ranks the best AUC at 0.62 with all variables and a test size of 10%. The second-best method was neural networks with the highest AUC and all variables using 10% for the test size.

According to the author's previous knowhow as regards choosing a test size for a small dataset of 2,503 customers, we were able to take the most stable results at a 40% test size. With the test size being 40%, we gained the best result from the first experiment with the random forest algorithm AUC=0.58, and the same result from the neural networks algorithm in the second experiment with only the variables where p >=0.05.

The weakest result overall was obtained from the SVM algorithm, which yielded very poor results in the second and third experiment, where AUC was below 0.50. The decision tree algorithm shows the most stable results across experiments and test sizes, having AUC between 0.50 and 0.55.

Comparing the results with related works in Table 10, it is apparent that in this research, the AUC results are lower than in others. There is high correlation between the test size and AUC, and seeing how our sample size is only 2,503 customers compared to 7,068, 60,000 and 295,926 we can consider our results to be good.

Table 9. Comparing related works.

Work	Test set size	Method	AUC
"Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment", authors: Björkegren and Grissen, 2017	7,068	Random Forest Logistic regression	0.710 0.760
"Mobile phone-based Credit Scoring", authors: Skyler Speakman, Eric Mibuari, Isaac Markus, Felix Kwizera, 2017	60,000	Linear regression	0.725
"MobiScore: Towards Universal Credit Scoring from Mobile Phone Data", authors: Jose San Pedro, Davide Proserpio and Nuria Oliver, 2015	295,926	Boosted decision trees algorithm Logistic regression	0.74 0.76
Current research	2,503	Logistic regression Decision tree Random forest SVM Neural networks	0.54 0.55 0.62 0.57 0.59

This study has two important theoretical contributions. First, based on the use of mobile data for credit scoring research, we can see that all the tested methods with all variables yielded a better result than in a random study.

Secondly, we empirically demonstrate that the best method for credit scoring based on mobile data is the random forests classification method with AUC 0.62.

Our research on mobile data scoring will make it possible for other financial companies to use mobile data for their credit scoring. While prior three researches on this subject showed that mobile data is only useful with big datasets, we maintain that it can yield positive results even with a small dataset. Thus, this knowledge can now be used in small or medium-sized companies as well.

4 Conclusion

For the past three decades, financial risk forecasting has been one of the main areas of growth in statistics and probability modelling. People often think of the term ‘financial risk’ in relation with portfolio management when it comes to pricing of options among other financial instruments. The main challenge for consumer loan firms over the past years has been reaching the huge sector of unbanked customers. There are more than 2 billion people in world who do not have a bank account [1] and the number of mobile phone users has reached 6 billion worldwide [2]. Few conceptual works have been posited with a research subject that brings together credit scoring and mobile data.

This thesis is based on a synthesis of earlier academic research and states that mobile data can give positive results for credit scoring even with a small dataset. Our findings also reveal that the best model in terms of mobile data usage for credit scoring is the decision tree method.

If finance companies want to have more accurate data on those customers who are more likely to pay back their loans, they need to find alternative data sources such as mobile phone data. This will a give huge advantage to finance companies in third world countries where most people do not have any bank history – the only data they have is their mobile phone data.

We hope this study opens up further discussion and advances theory to generate a more accurate understanding of how we can use mobile data to make predictions and added value. This thesis could be a point of discussion not only for financial sector companies but also for example in the field of insurance or fraud prevention, where mobile data can help make predictions.

There are many ways in which future studies could elaborate on this subject. One way is to look at algorithms in more depth and try to come up with more accurate models. Making predictions on mobile data can be used in other sectors as well, not only in finance. It is very probable that if we can predict customers’ payment behaviour based on mobile data, we could also predict their insurance or fraud risk. There are multiple research possibilities in the field of alternative data sources such as mobile data that could

add value for businesses. In the modern world we have many technical solutions at our disposal that create and gather data every day.

References

[1] C. Hodgson, “The world's 2 billion unbanked, in 6 charts,” 30 08 2017. [Online]. Available: http://uk.businessinsider.com/the-worlds-unbanked-population-in-6-charts-2017-8/#the-vast-majority-94-of-adults-in-oecd-high-income-countries-said-they-had-a-bank-account-in-2014-while-only-54-of-those-in-developing-countries-did-the-middle-east-had-the-l.

[2] L. Whitney, “2011 ends with almost 6 billion mobile phone subscriptions,” 04 01 2012. [Online]. Available: https://www.cnet.com/news/2011-ends-with-almost-6-billion-mobile-phone-subscriptions/. [Accessed 03 02 2018].

[3] Black, F., & Scholes, M, “The pricing of options and corporate liabilities,” *Journal of Political Economy 81,* p. 637–654, 1973.

[4] R. C. Merton, “On the pricing of corporate debt: the risk structure of interest rates,” *Journal of Finance 29,* pp. 449-470, 1974.

[5] Bellotti, T., & Crook, J, “Support vector machines for credit scoring and discovery of significant features,” *Expert Systems with Applications,* pp. 36, 3302–3308., 2009.

[6] de Montjoye, Y.-A., Hidalgo, C., Verleysen, M., Blondel, V, “ Unique in the Crowd: The privacy bounds of human mobility,” *Nature Sci. Rep,* 2013.

[7] D. Goldman, “Your phone company is selling your personal data,” 01 11 2011. [Online]. Available: http://money.cnn.com/2011/11/01/technology/verizon_att_sprint_tmobile_privacy/index.htm.

[8] R. Anderson, The Credit Scoring Toolkit, Theory and Practice for Retail Credit Risk Management and Decision Automation, New York: Oxford University Press Inc, 2017.

[9] A. Fensterstock, “Credit scoring and the next step,” *Business Credit 107(3),* pp. 46-49, 2005.

[10] J. Rimmer, “Contemporary changes in credit scoring,” *Credit Control 26(4),* pp. 56-60, 2005.

[11] Wendel, C., and M. Harvey., “Credit scoring: Best practices and approaches,” *Commercial Lending Review 18(3),* pp. 4-7, 2003.

[12] K. J. Leonard, “The development of credit scoring quality measures for consumer credit applications,” *International Journal of Quality & Reliability Management, Vol. 12 Issue: 4,* 1995.

[13] Banaslak, M. J. and Kiely, G. L., “Predictive collection score technology,” *Business Credit, vol.102(2),* pp. 18-20, 2000.

[14] Jacobson, T., and K. Roszback, “Bank lending policy, credit scoring and value-at-risk,” *Journal of Banking and Finance 27(4),* pp. 615-633, 2003.

[15] Avery, R.B., R.W. Rostic, P.S. Calem, and G.B. Canner, “Credit scoring: Statistical issues and evidence from credit-bureau files,” *Real Estate Economics 28(3),* pp. 524-526, 529-539, 544, 2000.

[16] A. L. S. M. a. J. B. Sandler, “Fair lending scrutiny of credit score-based underwriting systems,” *ABA Bank Compliance (March/April),* p. 38, 2000.

[17] S. Park, "Solving the mystery of credit scoring models," *Business Credit 106(3),* pp. 43-47, 2004.

[18] L. Thomas, "A survey of credit and behavioral scoring – forecasting financial risk of lending to consumer," *International Journal of Forecasting 16(2),* pp. 163-167, 2000.

[19] J. Quittner, "Credit cards: sub-prime's tech dilemma: with delinquencies and charge-offs on the rise, the industry examines the role of automated decisioning," *Bank Technology 16(1),* pp. 19-23, 2003.

[20] M. Perin, "Risky business: sub-prime market growth attracts host of new players," *Houston Business Journal (August),* 1998.

[21] K. Cundiff, "Closing the loop: How credit scoring drives performance improvements along the financial value chain," *Business Credit 106(3),* pp. 38-42, 2004.

[22] Kellison, B., and P. Brockett, "Check the score: credit scoring and insurance losses: Is there a connection?," *Texas Business Review Special Issue,* pp. 1-6, 2003.

[23] F. M, "An Overview and History of Credit Reporting," *Discussion Paper 02–07, Issued by the Payment Cards Center of the FRB of Philadelphia.,* 2002.

[24] Hand, D. J., & Henley, W. E., "Statistical classifica- tion methods in consumer credit," *Journal of the Royal Statistical Society, Series A 160,* p. 523–541, 1997.

[25] Partalas, I., Tsoumakas, G., & Vlahavas, I, "An ensemble uncertainty aware measure for directed hill climbing ensemble pruning," *Machine Learning, 81,* p. 257–282, 2010.

[26] García, S., Fernández, A., Luengo, J., & Herrera, F, "Advanced nonparametric for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," *Information sciences,* pp. 180, 2044–2064, 2010.

[27] Jonathan N.Crook David B.Edelman Lyn C.Thomas , "Recent developments in consumer credit risk assessment," 2007.

[28] R. de Oliveira, "Towards a psychographic user model from mobile phone usage," *In: Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems,* 2011.

[29] Arteaga, S.M., Kudeki, M., Woodworth, A, " Combating obesity trends in teenagers through persuasive mobile technology," *ACM SIGACCESS Accessibility and Computing 94,* p. 17–25, 2009.

[30] M. Back, "Facebook profiles reflect actual personality, not self-idealization," *Psychological Science 21(3),* p. 372–374, 2010.

[31] Stecher, K., Counts, S, "Self-presentation of personality during online profile creation," *In: Proc. AAAI Conf. on Weblogs and Social Media (ICWSM),* 2009.

[32] Stecher, K., Counts, S, "Spontaneous inference of personality traits and effects on memory for online profiles," *In: Proc. Int. AAAI Conference on Weblogs and Social Media (ICWSM),* 2008.

[33] Chittaranjan, G., Blom, J., Gatica-Perez, D, "Mining large-scale smartphone data for personality studies," *In: Personal and Ubiquitous Computing ,* 2012.

[34] Do, T.M.T., Gatica-Perez, D, "By their apps you shall understand them: mining large-scale patterns of mobile phone usage," *In: Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia,* 2010.

[35] H. Verkasalo, "Analysis of users and non-users of smartphone application," *Telematics and Informatics 27(3),* p. 242–255 , 2010.

[36] J. Staiano, "Friends dont Lie–Inferring Personality Traits from Social Network Structure," 2012.

[37] F. Pianesi, "Multimodal recognition of personality traits in social interactions," *In: Proceedings of the 10th International Conference on Multimodal Interfaces,* 2008.

[38] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic3, and Alex (Sandy) Pentland, "Predicting Personality Using Novel Mobile Phone-Based Metrics," 2013.

[39] Daniel Björkegren, Darrell Grissen, "Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment," *Netmob ,* 2013.

[40] D. P. N. O. Jose San Pedro, "MobiScore: Towards Universal Credit Scoring from Mobile Phone Data," pp. 195-207, 2017.

[41] Skyler Speakman, Eric Mibuari, Isaac Markus, Felix Kwizera, " Mobile –phone based Credit Scoring," *NetMob2017,* 2017.

[42] R. Johnson, "Legal, Social and Economic Issues in Implementing Scoring in the United States," 2004.

[43] E.W., Myers J.H. and Forgy, "The Development of Numerical Credit Evaluation Systems," *Journal of American Statistical Association 58(303),* p. 779–806, 1963.

[44] G. M. &. P. F. J. Weiss, "Learning when training data are costly: The effect of class distribution on tree induction," *Journal of Artificial Intelligence Research, 19,* p. 315–354, 2003.

[45] D. J. T. O. Foster Provost, "Efficient progressive sampling," *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining ,* pp. 23-32, 1999.

[46] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research, 16,* p. 321–357, 2002.

[47] N. Japkowicz, "Learning from imbalanced data sets: A comparison of various strategies," *In AAAI workshop on learning from imbalanced data sets,* p. pp. 10–15, 2000.

[48] G. Batista, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter, 6(1),* p. 20–29, 2004.

[49] N. Siddiqi, "Credit risk scorecards: Developing and implementing intelligent credit scoring (Vol. 3)," *Wiley.com,* 2005.

[50] N. Jentzsch, "The economics and regulation of financial privacy [electronic resource]: An international comparison of credit reporting systems," *Springer,* 2006.

[51] Hosmer, D. W., & Lemeshow, S, "The multiple logistic regression model," *Applied Logistic Regression, 1,* p. 25–37, 1989.

[52] J. Y. Coffman, “The proper role of tree analysis in forecasting the risk behavior of borrowers,” *Management Decision Systems, Atlanta, MDS Reports, 3(4),* p. 7, 1986.

[53] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J, “Classification and regression trees,” *Monterey, CA: Wadsworth & Brooks,* p. 1984.

[54] Rosenberg, E., & Gleit, A, “Quantitative methods in credit management: A survey,” *Operations Research, 42(4),* p. 589–613, 1994.

[55] Raiffa, H., & Schlaifer, R, “Applied statistical decision theory,” *Harvard Business School Publications,* 1961.

[56] Arminger, G., Enache, D., & Bonne, T, “Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis, and,” *Computational Statistics, 12,* p. 2, 1997.

[57] Hand, D. J., & Jacka, S. D, “Statistics in finance,” *Arnold London,* 1998.

[58] Zekic-Susac, M., Sarlija, N., & Bensic, M, “Small business credit scoring: A comparison of logistic regression, neural network, and decision tree models,” *In Paper presented at the 26th international conference on information technology interfaces,* 2004.

[59] J. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics,* p. 1189–1232, 201.

[60] J. Friedman, “Stochastic gradient boosting,” *Computational Statistics & Data Analysis,* p. 367–378, 2002.

[61] L. Breiman, “Random forests,” *Machine Learning, 45(1),* p. 5–32, 2001.

[62] Suykens, J. A.K.,Van Gestel, T., De Brabanter, J., De Moor,B., &Vandewalle, J, “Least squares support vector machines,” *Singapore: World Scientific,* 2002.

[63] C. M. Bishop, “Neural networks for pattern recognition,” *Oxford, UK: Oxford University Press,* 1995.

[64] T. Harris, “Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions,” *Expert Systems with Applications,* p. 4404–4413, 2013.

[65] T. Harris, “Credit scoring using the clustered support vector machine,” *Expert Systems with Applications,* 2014.

[66] A. E. D. M. K. AnselmBlumer, “Occam's Razor,” *Information Processing Letters,* vol. 24, pp. 377-380, 1987.

[67] V. T. Pavel Mironchyk, “Monotone optimal binning algorithm for credit risk modelling,” 2017.

[68] I. Guyon, “A scaling law for the validation-set training-set size ration,” 1996.

[69] S. Barisitz, “ Nonperforming Loans in CESEE–What Do They Comprise,” *Focus on European Economic Integration,* pp. 46-48, 2011.

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

[71] Ross, M. B. Yobas J. N. Crook P., "Credit scoring using neural and evolutionary techniques," *IMA Journal of Management Mathematics, Volume 11, Issue 2,* pp. 11-125, 2000.

[72] Desai, V. S., Crook, J. N., & Overstreet, G. A. Jr, "A comparison of neural networks and linear scoring models in the credit union environment," *European Journal of Operational Research, 95(1),* pp. 24-37, 1996.

[73] Provost, F., Jensen, D., & Oates, T, "Efficient progressive sampling," *In Proceedings of the fifth international conference on knowledge discovery and data mining. ACM Press.,* 1999.

Appendix 1 – Logistic Regression Classifier

```
# Importing the libraries

from time import time

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

# Importing the dataset

dataset = pd.read_csv(mobile_data.csv')

X = dataset.iloc[:, :-1].values

y = dataset.iloc[:, 21].values

# Encoding categorical data

from sklearn.preprocessing import LabelEncoder, OneHotEncoder

labelencoder_X_1 = LabelEncoder()

X[:, 20] = labelencoder_X_1.fit_transform(X[:, 20])

onehotencoder = OneHotEncoder(categorical_features = [20])

X = onehotencoder.fit_transform(X).toarray()

# Splitting the dataset into the Training set and Test set, we run code with test_size= 0.10
; 0.25 and 0.40

from sklearn.cross_validation import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.10, random_state =
0)
```

```
# Feature Scaling

from sklearn.preprocessing import StandardScaler

sc = StandardScaler()

X_train = sc.fit_transform(X_train)

X_test = sc.transform(X_test)

# Fitting Logistic regression to the Training set

from sklearn.linear_model import LogisticRegression

classifier = LogisticRegression()

classifier.fit(X_train, y_train)

t0 = time() # current time

classifier.fit(X_train, y_train)

print('training time: {}s'.format(round(time()-t0, 3)))

# import accuracy_score from sklearn

from sklearn.metrics import accuracy_score

# training set accuracy

y_pred_train = classifier.predict(X_train) # training set predictions

training_set_accuracy = accuracy_score(y_train, y_pred_train)

print('Training Set Accuracy: {}'.format(training_set_accuracy))

# Predicting the Test set results

t1 = time() # current time

y_pred = classifier.predict(X_test) # test set predictions
```

```
print('predicting time: {}s'.format(round(time()-t1, 3)))

# test set accuracy

test_set_accuracy = accuracy_score(y_test, y_pred)

print('Test Set Accuracy: {}'.format(test_set_accuracy))

# balanced accuracy score

# from sklearn.metrics import balanced_accuracy_score # needs sklearn version 0.20 (dev version)

# BAC = balanced_accuracy_score(y_test, y_pred)

# print('Balanced Accuracy Score: {}'.format(BAC))

from sklearn.metrics import roc_curve, auc

# get class probabilities

scores = classifier.predict_proba(X_test)

scores = scores[:, 1] # choose probabilities from only positive class (1)

fpr, tpr, thresholds = roc_curve(y_test, scores, pos_label=1) # pos_label = 1 means 1 is positive class

roc_auc = auc(fpr, tpr) # fpr = false_positive_rate, tpr = true_positive_rate

print('ROC Area Under Curve: {}'.format(roc_auc))

plt.title('Receiver Operating Characteristic')

plt.plot(fpr, tpr, 'b',

label='AUC = %0.2f'% roc_auc)

plt.legend(loc='lower right')

plt.plot([0,1],[0,1],'r--')
```

```
plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.ylabel('True Positive Rate')

plt.xlabel('False Positive Rate')

plt.show()

# Making the Confusion Matrix

from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, y_pred)
```

Appendix 2 – Decision Tree Classification

```
# Importing the libraries

from time import time

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

# Importing the dataset

dataset = pd.read_csv('mobile_data.csv')

X = dataset.iloc[:, :-1].values

y = dataset.iloc[:, 30].values

# Encoding categorical data

from sklearn.preprocessing import LabelEncoder, OneHotEncoder

labelencoder_X_1 = LabelEncoder()

X[:, 29] = labelencoder_X_1.fit_transform(X[:, 29])

onehotencoder = OneHotEncoder(categorical_features = [29])

X = onehotencoder.fit_transform(X).toarray()

# Splitting the dataset into the Training set and Test set

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.1, random_state = 0)

# Feature Scaling

from sklearn.preprocessing import StandardScaler
```

```
sc = StandardScaler()

X_train = sc.fit_transform(X_train)

X_test = sc.transform(X_test)

# Fitting Decision Tree Classification to the Training set

from sklearn.tree import DecisionTreeClassifier

classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)

t0 = time() # current time

classifier.fit(X_train, y_train)

print('training time: {}s'.format(round(time()-t0, 3)))

# import accuracy_score from sklearn

from sklearn.metrics import accuracy_score

# training set accuracy

y_pred_train = classifier.predict(X_train) # training set predictions

training_set_accuracy = accuracy_score(y_train, y_pred_train)

print('Training Set Accuracy: {}'.format(training_set_accuracy)

# Predicting the Test set results

t1 = time() # current time

y_pred = classifier.predict(X_test) # test set predictions

print('predicting time: {}s'.format(round(time()-t1, 3)))

# test set accuracy

test_set_accuracy = accuracy_score(y_test, y_pred)
```

```
print('Test Set Accuracy: {}'.format(test_set_accuracy))

# balanced accuracy score

# from sklearn.metrics import balanced_accuracy_score # needs sklearn version 0.20 (dev version)

# BAC = balanced_accuracy_score(y_test, y_pred)

# print('Balanced Accuracy Score: {}'.format(BAC))

from sklearn.metrics import roc_curve, auc

scores = classifier.predict_proba(X_test) # get class probabilities

scores = scores[:, 1] # choose probabilities from only positive class (1)

fpr, tpr, thresholds = roc_curve(y_test, scores, pos_label=1) # pos_label = 1 means 1 is positive class

roc_auc = auc(fpr, tpr) # fpr = false_positive_rate, tpr = true_positive_rate

print('ROC Area Under Curve: {}'.format(roc_auc))

plt.title('Receiver Operating Characteristic')

plt.plot(fpr, tpr, 'b',

label='AUC = %0.2f'% roc_auc)

plt.legend(loc='lower right')

plt.plot([0,1],[0,1],'r--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.ylabel('True Positive Rate')

plt.xlabel('False Positive Rate')
```

```
plt.show()

# Making the Confusion Matrix

from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, y_pred)
```

Appendix 3 – Random forest Classification

```
# Importing the libraries

from time import time

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

# Importing the dataset

dataset = pd.read_csv('

mobile_data.csv')

X = dataset.iloc[:, :-1].values

y = dataset.iloc[:, 21].values

# Encoding categorical data

from sklearn.preprocessing import LabelEncoder, OneHotEncoder

labelencoder_X_1 = LabelEncoder()

X[:, 20] = labelencoder_X_1.fit_transform(X[:, 20])

onehotencoder = OneHotEncoder(categorical_features = [20])

X = onehotencoder.fit_transform(X).toarray()

# Splitting the dataset into the Training set and Test set, we run code with test_size= 0.10
; 0.25 and 0.40

from sklearn.cross_validation import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.10, random_state = 0)

# Feature Scaling

from sklearn.preprocessing import StandardScaler

sc = StandardScaler()

X_train = sc.fit_transform(X_train)

X_test = sc.transform(X_test)

# Fitting Random Forest Classification to the Training set

from sklearn.ensemble import RandomForestClassifier

classifier = RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_state = 0)

classifier.fit(X_train, y_train)

# Predicting the Test set results

y_pred = classifier.predict(X_test)

t0 = time() # current time

classifier.fit(X_train, y_train)

print('training time: {}s'.format(round(time()-t0, 3)))

# import accuracy_score from sklearn

from sklearn.metrics import accuracy_score


# training set accuracy

y_pred_train = classifier.predict(X_train) # training set predictions
```

```
training_set_accuracy = accuracy_score(y_train, y_pred_train)

print('Training Set Accuracy: {}'.format(training_set_accuracy))

# Predicting the Test set results

t1 = time() # current time

y_pred = classifier.predict(X_test) # test set predictions

print('predicting time: {}s'.format(round(time()-t1, 3)))

# test set accuracy

test_set_accuracy = accuracy_score(y_test, y_pred)

print('Test Set Accuracy: {}'.format(test_set_accuracy))

# balanced accuracy score

# from sklearn.metrics import balanced_accuracy_score # needs sklearn version 0.20 (dev version)

# BAC = balanced_accuracy_score(y_test, y_pred)

# print('Balanced Accuracy Score: {}'.format(BAC))

from sklearn.metrics import roc_curve, auc

scores = classifier.predict_proba(X_test) # get class probabilities

scores = scores[:, 1] # choose probabilities from only positive class (1)

fpr, tpr, thresholds = roc_curve(y_test, scores, pos_label=1) # pos_label = 1 means 1 is positive class

roc_auc = auc(fpr, tpr) # fpr = false_positive_rate, tpr = true_positive_rate

print('ROC Area Under Curve: {}'.format(roc_auc))

plt.title('Receiver Operating Characteristic')
```

```
plt.plot(fpr, tpr, 'b',

label='AUC = %0.2f'% roc_auc)

plt.legend(loc='lower right')

plt.plot([0,1],[0,1],'r--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.ylabel('True Positive Rate')

plt.xlabel('False Positive Rate')

plt.show()

# Making the Confusion Matrix

from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, y_pred)
```

Appendix 3 – Support vector machine Classification

```
# Importing the libraries

from time import time

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

# Importing the dataset

dataset = pd.read_csv('mobile_data.csv')

X = dataset.iloc[:, :-1].values

y = dataset.iloc[:, 21].values

# Encoding categorical data

from sklearn.preprocessing import LabelEncoder, OneHotEncoder

labelencoder_X_1 = LabelEncoder()

X[:, 20] = labelencoder_X_1.fit_transform(X[:, 20])

onehotencoder = OneHotEncoder(categorical_features = [20])

X = onehotencoder.fit_transform(X).toarray()

# Splitting the dataset into the Training set and Test set, we run code with test_size= 0.10
; 0.25 and 0.40

from sklearn.cross_validation import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.10, random_state =
0)
```

```
# Feature Scaling

from sklearn.preprocessing import StandardScaler

sc = StandardScaler()

X_train = sc.fit_transform(X_train)

X_test = sc.transform(X_test)

# Fitting SVM to the Training set

from sklearn.svm import SVC

# kernel = 'linear' can only model fully separable functions because it uses a straight line

# as its decision boundary. If you want more flexibility use kernel = 'rbf'

classifier = SVC(kernel = 'linear', random_state = 0, probability=True) # set probability
= True because we use predict_proba later

classifier.fit(X_train, y_train)

t0 = time() # current time

classifier.fit(X_train, y_train)

print('training time: {}s'.format(round(time()-t0, 3)))

# import accuracy_score from sklearn

from sklearn.metrics import accuracy_score

# training set accuracy

y_pred_train = classifier.predict(X_train) # training set predictions

training_set_accuracy = accuracy_score(y_train, y_pred_train)

print('Training Set Accuracy: {}'.format(training_set_accuracy))

# Predicting the Test set results
```

```
t1 = time() # current time

y_pred = classifier.predict(X_test) # test set predictions

print('predicting time: {}s'.format(round(time()-t1, 3)))

# test set accuracy

test_set_accuracy = accuracy_score(y_test, y_pred)

print('Test Set Accuracy: {}'.format(test_set_accuracy))

# balanced accuracy score

# from sklearn.metrics import balanced_accuracy_score # needs sklearn version 0.20 (dev version)

# BAC = balanced_accuracy_score(y_test, y_pred)

# print('Balanced Accuracy Score: {}'.format(BAC))

from sklearn.metrics import roc_curve, auc

# get class probabilities, only available because we set probability = True during training

# This is peculiar for SVM

scores = classifier.predict_proba(X_test)

scores = scores[:, 1] # choose probabilities from only positive class (1)

fpr, tpr, thresholds = roc_curve(y_test, scores, pos_label=1) # pos_label = 1 means 1 is positive class

roc_auc = auc(fpr, tpr) # fpr = false_positive_rate, tpr = true_positive_rate

print('ROC Area Under Curve: {}'.format(roc_auc))

plt.title('Receiver Operating Characteristic')

plt.plot(fpr, tpr, 'b',
```

```
label='AUC = %0.2f'% roc_auc)

plt.legend(loc='lower right')

plt.plot([0,1],[0,1],'r--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.ylabel('True Positive Rate')

plt.xlabel('False Positive Rate')

plt.show()

# Making the Confusion Matrix

from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, y_pred)
```

Appendix 3 – Artificial Neural Network

```
# Artificial Neural Network

# Installing Theano

# pip install --upgrade --no-deps git+git://github.com/Theano/Theano.git

# Installing Tensorflow

#Install Tensorflow from the website: https://www.tensorflow.org/versions/r0.12/get_started/os_setup.html

# Installing Keras

# pip install --upgrade keras

# Data Preprocessing

# Importing the libraries

from time import time

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

# Importing the dataset

dataset = pd.read_csv('mobile_data_old.csv')

X = dataset.iloc[:, :-1].values

y = dataset.iloc[:, 21].values

# Encoding categorical data

from sklearn.preprocessing import LabelEncoder, OneHotEncoder
```

```
labelencoder_X_1 = LabelEncoder()

X[:, 20] = labelencoder_X_1.fit_transform(X[:, 20])

onehotencoder = OneHotEncoder(categorical_features = [20])

X = onehotencoder.fit_transform(X).toarray()

# Splitting the dataset into the Training set and Test set, we run code with test_size= 0.10 ; 0.25 and 0.40

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.10, random_state = 0)

# Feature Scaling

from sklearn.preprocessing import StandardScaler

sc = StandardScaler()

X_train = sc.fit_transform(X_train)

X_test = sc.transform(X_test)

# Importing the Keras libraries and packages

import keras

from keras.models import Sequential

from keras.layers import Dense

# Initialising the ANN

classifier = Sequential()

# Adding the input layer and the first hidden layer

classifier.add(Dense(units = 6, kernel_initializer = 'uniform', activation = 'relu', input_dim = 35))
```

```
# Adding the second hidden layer

classifier.add(Dense(units = 6, kernel_initializer = 'uniform', activation = 'relu'))

# Adding the output layer

classifier.add(Dense(units = 1, kernel_initializer = 'uniform', activation = 'sigmoid'))

# Compiling the ANN

classifier.compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics = ['accuracy'])

# Fitting the ANN to the Training set

t0 = time() # current time

classifier.fit(X_train, y_train, batch_size = 10, epochs = 300) # large epochs could lead to overfitting

print('training time: {}s'.format(round(time()-t0, 3)))

# Making the predictions and evaluating the model

# Predicting the Test set results

t1 = time() # current time

scores = classifier.predict(X_test)

print('predicting time: {}s'.format(round(time()-t1, 3)))

# Keras returns probability scores between 0 and 1, so to get our predictions

# we set a threshold of 0.5

y_pred = (scores > 0.5)

# import accuracy_score from sklearn

from sklearn.metrics import accuracy_score

test_set_accuracy = accuracy_score(y_test, y_pred)
```

```
print('Test Set Accuracy: {}'.format(test_set_accuracy))

from sklearn.metrics import roc_curve, auc

fpr, tpr, thresholds = roc_curve(y_test, scores, pos_label=1) # pos_label = 1 means 1 is positive class

roc_auc = auc(fpr, tpr) # fpr = false_positive_rate, tpr = true_positive_rate

print('ROC Area Under Curve: {}'.format(roc_auc))

plt.title('Receiver Operating Characteristic')

plt.plot(fpr, tpr, 'b',

label='AUC = %0.2f'% roc_auc)

plt.legend(loc='lower right')

plt.plot([0,1],[0,1],'r--')

plt.xlim([0.0, 1.0])

plt.ylim([0.0, 1.05])

plt.ylabel('True Positive Rate')

plt.xlabel('False Positive Rate')

plt.show()

# Making the Confusion Matrix

from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, y_pred)
```