Support Vector Machine Vs. Logistic Regression for predicting mortgage defaults

Aram Olvbo

Bachelor's thesis
2021:K43

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Abstract

Mortgage loan providers estimate the credit risks it caries when approving a mortgage loan to their clients. Further, defaulting a mortgage loan is a risk that has been calculated through decades using statistical models. By using entries at the time of a mortgage application, the goal of the thesis is to compare the accuracy between logistic regression and Support Vector Machine in predicting a mortgage loan default. For this purpose, Fannie Mae 30-year-fixed-rate single-family mortgage loans are used for the years; 2000, 2005 and 2010. The models aim is to predict probability of default during five years period from the loan acquiring date. While the result showed that logistic regression was both faster and less complex to implement, SVM proved to have a marginally better prediction with the drawback of a longer computational time. The forecast accuracy to compare the two models at hand was ROC and Precision-recall, although precision-recall was favored due to the unbalanced data.

Acknowledgement

A huge thank you to Professor Erik Lindström for his astonishing patience and support, I truly could not have done it without your exceptional understanding. Additionally, I want to also thank my dear friends Halwest Mohammed and Jonas Eyob for being my moral support. Last but not least, I am grateful for my wife Sara and rest of the family for keeping my calm. To the reader, better late than never.

Contents

Dictionary

Default - the failure to fulfill an obligation to repay a loan

PD - Probability of Default

SVM - Support Vector Machine Learning

NN - Neural Network

OTC - Over the Counter

SVM - Support Vector Machine

PRC - Precision Recall Curve

ROC - Receiving Operating Characteristics

AUC - Area Under Curve

SMOTE - Synthetic Minority Oversampling Technique

PRC - Precision Recall Curve

RBF - Radial Basis Function

FICO - Fair Isaac Corporation, is the company providing with credit-risk model with a score, i.e. borrowers credit score.

DTI - Debt to Income

LTV - Loan to Value

1 Introduction

1.1 Background

The news were out on September 15 2008, that the well know financial institute Lehman Brothers had collapsed and a full blown financial crisis had started, the so called sub-mortgage market crisis. The financial screens were flashing red and the new era of regulations on the financial and mortgage market was to be imposed. As a result banks had to scrutinize their risk exposures and mortgage takers more in-depth.

The extraordinary instability in the financial and real estate market conditions during the financial crisis lead to the housing crash in the late 2000s where there was a dramatic increase in mortgage defaults. In 2011, 14.2 % of mortgages were in difficulty to repay their mortgages, total mortgage debt was roughly USD 10 trillion which is a big figure of potential mortgage defaults. The potential loss to mortgage lenders stresses the financial system and was one of the main factors contributing to the larger economic downturn of the financial crisis. The event have underscored the importance of understanding household incentives to default on mortgages. Not to exclude the cost arising for a mortgage lender when a mortgage loan defaults. (Copeland 2017) Artificial Intelligence and Machine Learning are two buzzwords often used. Both of the terms crop up in topics such as Big Data, analytics and other waves of technological changes which are sweeping through our world. For the purpose of clarification the two terms are explained before going further on the topic.

Artificial Intelligence is the broader concept of machines being able to carry out simulation of intelligent behavior or the capability to imitate intelligent human behavior. Machine Learning is an application of AI based on giving machines access to data and let them learn for themselves. The great American pioneer, Professor Arthur Samuel, in the field of artificial intelligence coined the term "machine learning" in 1959, "field of study that gives computers the ability to learn without being explicitly programmed", Samuel (1959)

However, mortgage lenders have their own way of processing and modelling the probability of an individual not fulfilling obliged repayments. There is not a uniform calculation which gives the same probability of default, every bank has their own way of modelling this probability. Hence, an individual who applies for a mortgage loan could have a different probability of default figure, depending on the commercial bank's model. Hence, this thesis is aimed at calculating the probability of default for a mortgage borrower, using machine learning techniques.

1.1.1 Credit Quality

Credit quality is a way of informing the risk of default of an individual or entity. Investors want to have an accurate view on the risk they are taking upon themselves, to not end up losing a portion of their investment. Furthermore, since the financial crisis, credit risk has become more important and new regulatory laws have been and are being implemented. For instance after the financial crises the government requires that the final terms of a mortgage have to be shown to borrowers at least three business days before the closing date, to prevent mortgage takers from making hurried decisions or signing off on a mortgage without fully understanding the terms. Moreover, additional rules have been implemented to tighten the lending standards, "ability to repay" rules. Under these rules, lenders get greater legal protections if they make so-called "qualified mortgages", in which borrowers' monthly debt payments do not exceed 43 per cent of their income. General rules (§ 1026.31) Consumer Financial Protection Bureau

In the over the counter (OTC) market, where derivatives and securities are being traded off-exchange, counterparties need to minimize the risk of a counterparty not being able to pay its obligations. An example is if two counterparties, A and B, are transacting a derivative. Derivatives are leveraged and volatile in the sense that the derivative prices can move dramatically. Counterparties have to post collateral when they transact, this collateral is changing depending on the derivatives position, that is if counterparty A or B is making money. Hence, the collateral required by the counterparties is to cover future obligations with a high degree of certainty.

In the commercial space, banks will always see risk with lending money to a borrower due to the uncertainty of modelling and calculating the risk of a borrower not being able to pay its obligations in the future. Therefore, the bank will charge the interest, risk-free interest plus a premium for the default risk. The premium is composed of the degrees of presumed risk. The riskier a grade, the higher the actual interest rate. The presumed risk can be calculated and represented in many ways, such as;

- PD model (Probability of default) - A model that predicts the probability of a particular borrower's loan to default.
- LGD model (Loss given default) - If the borrower default, the lender can recover some amount by selling the collateral. A common example is: the borrower defaults on the mortgage, so the lender forecloses the house and the sale of the house decreases the actual loss. Hence, if the borrower defaulted while owing a mortgage balance of 100k, and the house was sold through foreclosure for 50k, then LGD = 50/100 = 50%, since the lender lost 50% of the balance.

- EAD model (Exposure at default) - A model that tries to predict what the amount owed will be at the time of default, such as the above example.

Probability of default is the risk measure that will be covered in this thesis. The risk of a default (or probability of default) is not always the easiest to calculate. It is highly complex and therefore undertaken by specialised financial institutions ("rating agencies") such as Moody's, Fitch and Standard & Poor's which rate debt securities in several market segments related to public and commercial securities, including the government, municipal and corporate bond space. The quality of a debt being repaid is rated from highest quality and the lowest quality.

However, the mortgage loan space is different in the sense that commercial banks lending mortgage loans have their own way of processing and modelling the probability of an individual not fulfilling the obliged repayment. There is not a uniform calculation which gives the same probability of default, every bank has their own way of modelling this probability. Hence, an individual who applies for a mortgage loan would have a probability of default depending on the commercial bank's model which can vary, as mentioned earlier.

1.1.2 Artificial Intelligence

Artificial intelligence as we know it gained its popularity in Professor Samuel Arthur's paper written (1959), where his work on game of checkers was one of the earliest examples of non-numerical computation and were adopted by many computer designers. In his paper it was verified that a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program. It was also shown that the principals of machine learning were applicable to many other situations. Professor Samuel's work is still worth reading and have been the foundation for Machine Learning. As Professor Samuel wrote in his paper:

... it can learn to do this in a remarkably short period of time (8 or 10 hours of machine-playing time) when given only the rules of the game, a sense of direction, and a redundant and incomplete list of parameters which are thought to have something to do with the game, but whose correct signs and relative weights are unknown and unspecified.

Even though the concept of Artificial Intelligence is several decades old the Artificial Intelligence market is $ 3.6 billion and predicted to grow to $ 36 billion in 2020, that is a growth taking into account hardware, software and services. The technology might be best known by an example of famously self driven Tesla. However, many more successful applications exists such as finding frauds across transactions, detecting anomalous behavior, speeding up drug discoveries, and voice recognition. Many firms such as IBM, Google,

Microsoft and Facebook are offering their machine learning platforms for developers to create products and enhance existing ones, Bughin (2017)

Technological innovations in computer science and storage has slowly but surely reached levels which have made it practical to implement complex algorithms, especially neural networks to mimic the human brain. Since Professor Arthur's first paper on machine learning, many methods have evolved. However, this thesis is not intended to give an in-depth explanation of all existing methods. Nevertheless, a in-depth analysis of Logistic Regression and Support-Vector-Machine Learning will be carried out in the thesis.

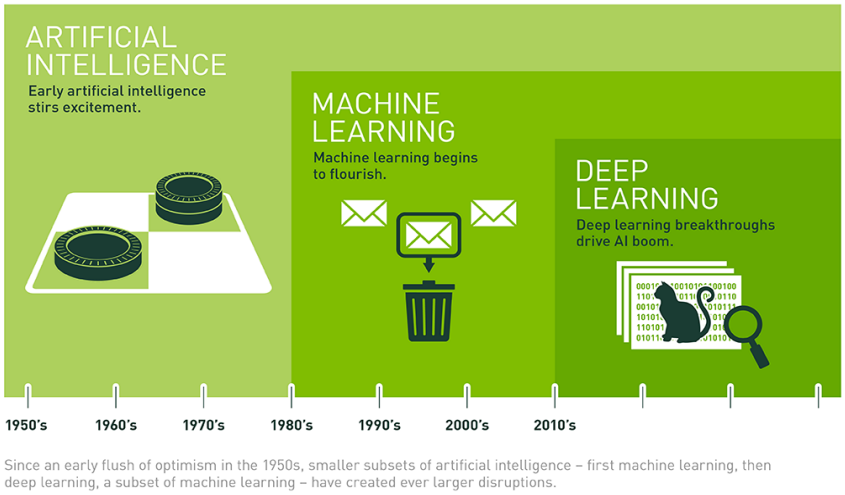


Figure 1: *The evolution of AI*

1.2 Literature Study

The thesis will have its foundation from one of the earliest papers on Machine Learning, Arthur L (1959), which was mentioned earlier in the introduction. Arthur L (1959) used programming to train a computer to play a game of checkers, the program learned by rewarding and punishing moves, to play and even become better than the programmer. The more game time the better the program became.

1.2.1 Probability of Default

Early on, Von Furstenberg (1969) and (1970) established the influence on home mortgage default rates of variables such as income, loan age, and loan-to-value ratio. Von Furstenberg (1969) concluded that it was the financing

characteristics of the mortgage loan which accounted for much of the observed variation in default risk. Moreover, Von Furstenberg (1970) reemphasized that characteristics of the mortgagor, such as age and income, are statistically far less powerful explanatory variables of the level of default rates. Rather Loan-To-Value was the only mortgage characteristic, correlated with income, whose causal relation to risk could firmly be established.

Additional variables were investigated by Gau (1978) and Vandell (1978). Gau (1978) investigated through application of factor analysis, 64 interrelated variables describing the financial, property, and borrower characteristics of residential mortgages were transformed into a smaller number of dependent factor dimensions. The model derived could then be utilized to identify the relative risk of default of conventional mortgages. Vandell (1978) found that default risk is predicted to increase when a household income is lowered. The empirical results were derived by fitting a model with relating variables associated with the borrower, property, and mortgage instrument to the probability of default over time.

1.2.2 Machine Learning

Two pattern recognition approaches are investigated in Vapnik (1995) who combined three ideas to support-vector network;

1. The solution technique from optimal hyperplanes (that allows for an expansion of the solution vector on support vectors).
2. The idea of convolution of the dot-product (that extends the solution surfaces from linear to non-linear).
3. The notion of soft margins (to allow for errors on the training set).

The algorithm developed by Vapnik (1995) was tested and compared to the performance of other classical algorithms at that point of time. Despite the simplicity of the design in its decision surface the new algorithm exhibited a adequate performance in the comparison study. Schölkopf (2000) built upon Vapnik's (1995) work and proposed a new class of support vector algorithms for regression and classification.

Schölkopf (2000) let effectively a parameter to control the number of support vectors. The parameterization had the benefit of enabling one to eliminate one of the other free parameters of the algorithm, i.e. the accuracy parameter in the regression case, and the regularization constant in the classification case.

Chang, Dae-oong Kim and Kondo (2016) predicted default risk of lending club loans by analyzing different methods such as logistic regression, Naive Bayes and SVM. Within these methods further improvements were made by using different sample sizes, looking for high bias or variances. From the comparisons and fine tunings to predict default rates(optimizing specificity), Naive Bayes with Gaussian performed well with independent feature sets. However, most classifiers show significant deterioration in performance when dealing with skewed data sets which is the case here. He and Ghodsi (2010), showed that two proposed special modified SVM methods had consistent improvement over ordinary SVM performances.

1.3 Problem Formulation

As mentioned earlier in the introduction, defaulting mortgages is costly for mortgage providers. Currently all mortgage providers use proprietary calculations to generate the risk profile for the mortgagor. However, due to the mechanics of the global economic markets banks are highly inter-connected which was evident during the financial crisis with many mortgage loans defaulting and leading the financial system into distress.

The aim of the thesis is to help mortgage providers to classify a clients probability of defaulting a mortgage. A simplified scenario could be a client applying for a mortgage loan for a specific property and enters their personal information to hopefully be accepted by the lender. The lender should be able to calculate the risk, including probability of the client not being able to repay the mortgage, given the client's information and financial background check. The risk calculated is translated into a mortgage interest rate.

This thesis will help predicting the probability of default, the probability of a household not being able to repay their mortgage. Additionally, comparing the two common known methods, SVM and Logistic Regression from the computational complexity, strength and predictive stand points.

The thesis will address the following questions;

- How will SVM perform predicting probability of default on mortgage loan's taken pre 2007 financial crisis, 2000-2005?
- How will SVM perform predicting probability of default of mortgage loan's over the period of 2007, 2005-2010?
- How will SVM performs in predicting probability of default on mortgage loans post financial crisis era, 2010-2015?
- How does the above result stand compared to a Logistic Regression model?

2 Data

Figure 2: *The steps in the process of this thesis. Selecting the best subset from Fannie Mae was carried out with help of previous studies.*

Data is downloaded from Fannie Mae, Single-Family loans guaranteed or owned by Fannie Mae. The loans are 30-year, fixed-rate mortgage, fully amortized and documented. The attributes chosen is selected from previous work, the loan to value (LTV) a mortgage characteristic correlated with Income, where the causal relation to risk could firmly be established as per George M. and Von Furstenbergs (1970) work. A Raising LTV increases default rates far more drastically and consistently than lowering a mortgage takers income. It was established that characteristics of the mortgagor, such as age and income, are statistically far less powerful explanatory variables of the level of default rate.

Kerry D and Vandell (1978) stated in their work that interest rate has implications on the mortgage default risk and can be predicted to increase or decrease roughly, however moderately and still within an acceptable range

Von Frustenberg (1969) also concluded that LTV had a major significance for defaults. The loan-value ratio is the variable governing the level of default rates over the life of the mortgage. Reducing the downpayment in the highest LTV range by as little as 1 % of home value can cause default rates to rise by 50 % . Moreover, neither age nor income, but rather the financing characteristics of the mortgage correlated therewith, which account for much of the observed variation. Income definitely cannot compete with LTV as the principal variable explaining the higher default rates for lower income group. Hence, Debt-to Income should account the factor of not only the income itself rather how much indebted the borrower is. This factor believed to be reflected at by the FICO Score, which will be explained later

Public available data for the period 2000-2015 was downloaded from The Federal National Mortgage Association, known as Fannie Mae, database. Fannie Mae is a United States government-sponsored enterprise and provide access to mortgage financing. The dataset used includes Fannie Mae's 30-year fixed-rate, fully documented, single-family amortizing loans that the company owned on or after January 1 2000. All the available variables when acquiring a house loan is seen in the table below.

The explanatory variables for the analysis was narrowed down to only numerical values, leaving only 10 variables of the original 22. The 10 variables were

POSITION	FIELD NAME	TYPE
1	LOAN IDENTIFIER	ALPHA-NUMERIC
2	CHANNEL	ALPHA-NUMERIC
3	SELLER NAME	ALPHA-NUMERIC
4	ORIGINAL INTEREST RATE	NUMERIC
5	ORIGINAL UNPAID PRINCIPAL BALANCE (UPB)	NUMERIC
6	ORIGINAL LOAN TERM	NUMERIC
7	ORIGINATION DATE	DATE
8	FIRST PAYMENT DATE	DATE
9	ORIGINAL LOAN-TO-VALUE (LTV)	NUMERIC
10	ORIGINAL COMBINED LOAN-TO-VALUE (CLTV)	NUMERIC
11	NUMBER OF BORROWERS	NUMERIC
12	DEBT-TO-INCOME RATIO (DTI)	NUMERIC
13	BORROWER CREDIT SCORE	NUMERIC
14	FIRST-TIME HOME BUYER INDICATOR	ALPHA-NUMERIC
15	LOAN PURPOSE	ALPHA-NUMERIC
16	PROPERTY TYPE	ALPHA-NUMERIC
17	NUMBER OF UNITS	ALPHA-NUMERIC
18	OCCUPANCY STATUS	ALPHA-NUMERIC
19	PROPERTY STATE	ALPHA-NUMERIC
20	ZIP (3-DIGIT)	ALPHA-NUMERIC
21	MORTGAGE INSURANCE PERCENTAGE	NUMERIC
22	PRODUCT TYPE	ALPHA-NUMERIC
23	CO-BORROWER CREDIT SCORE	NUMERIC

Figure 3: *Mortgage loan acquisition variables from Fannie Mae database.*

lastly narrowed down to 4 variables, to only use continuous data at the point of acquiring a mortgage loan. Loans which had been altered, that is a new mortgage taken on an existing house was not considered.
The four explanatory variables left was, Original Interest Rate, Loan-To-Value ratio (LTV), Debt-to-Income ratio (DTI) and Borrower credit score (FICO). Below will give a throughout explanations of each one of the explanatory variables.

- The initial Loan-To-Value (LTV) ration, that is loan to purchase price of the property of a mortgage reflects the amount of equity a borrower have invested in the purchased property. A default would cause the borrower to lose this equity. A common sense view would be a lower LTV, more equity invested, should repel a investor of defaulting, as more is at stake for the borrower.

- Debt-to-Income ration (DTI) is the total amount of borrowers monthly gross income going towards paying obligations at the time of acquiring the mortgage loan. A high DTI would most probably restrict the borrowers mortgage amount.

- Interest rates of each mortgage loan is taken into account. As mentioned before, interest rates also reflects riskiness of a loan. The higher the interest rate the higher the risk of the specific borrower.

- Lastly, borrower Credit Score (FICO score) is a numerical value used throughout the financial industry to evaluate borrowers creditworthiness. The FICO mortgage score is between 300 and 850, a higher number indicates a lower credit risk of the borrower. The exact model used to compute the FICO score is not public, however, the percentage of each component is public, as per figure 4.

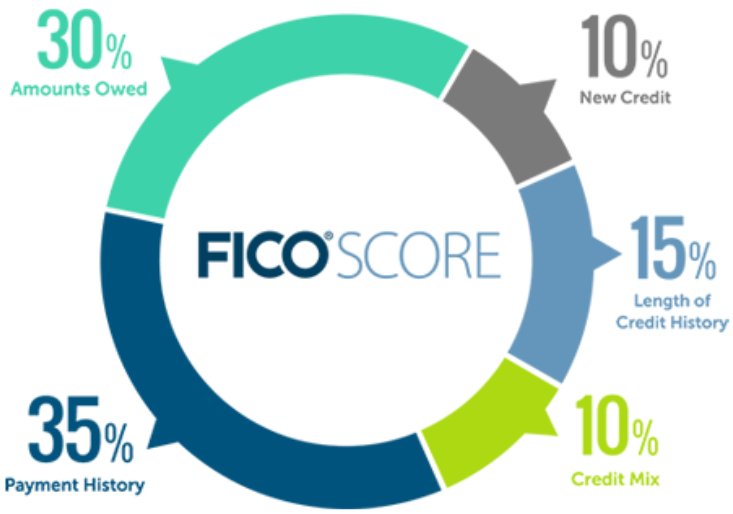


Figure 4: *The calculation of Mortgage FICO Score*

Finally, for each specific loan with the explanatory variables the delinquency status was investigated during a 5 year period, 2000-2005, 2005-2010 and 2010-2015. That is, mortgage loans undertaken during 2000 and investigate if the borrower has defaulted during a 5 years period, i.e. up to 2005. Likewise for 2005-2010 and 2010-2015 time periods. Note that when during these 5 years the borrower has defaulted in not considered. There is no consensus on

the definition of a default and after how many days of delinquency a mortgage loan is considered defaulted. However, according to FICO, a payment delayed by more than 30-days is flagged, consequently resulting in a decreased score, and stays on your FICO credit report for seven years. Therefore, a heuristic used within this thesis is to consider a delinquency status of more than 30 days as defaulted.

All the explanatory variables was scaled [0 ,1] and delinquency status fields were altered. Any repayment delayed more than 30 days was labeled defaulted, a binary approach is taken where 1 is considered defaulted and 0 not defaulted, i.e repaid or still repaying according to agreement.

Data		
Variable	Description	Value
Interest Rate	The interest on a mortgage loan in effect for the periodic installment due.	Percentage
Loan- To - Value (LTV)	A ratio calculated at the time of origination from a mortgage loan. The original LTV reflects the loan-to-value ratio of the loan amount secured by a mortgage property on the origination date of the underlying mortgage loan.	Percentage (0% - 97%)
Debt- To - Income (DTI)	A ratio calculated at origination derived by dividing the borrower's total monthly obligations (including housing expenses) by his or her stable monthly income. This calculation is used to determine the mortgage amount for which a borrower qualifies	Percentage (1% - 64%)
Borrower Credit Score	A numerical value used by financial services industry ti evaluated the quality of borrower credit. Credit scores are typically based on a proprietary statistical model that is developed for use by credit data repositories. The credit repositories apply the model to borrower credit information to arrive at a credit score. When this term is used by Fannie Mae, it is referring to the "classic" FICO score developed by Fair Isaac Corporation.	Numerical (300-800)
Delinquency Status	The number of days, represented in months, the obligator is delinquent as determined by the governing mortgage documents	0 = Current, or less than 30 days past due 1 = 30 - 59 days 2 = 60 - 89 days 3 = 90 - 119 days

Table 1: *Variables chosen for this thesis.*

2.1 Unbalanced Data

The data set contained 1.3 - 1.7 million observations for each time period with a high non-default rate. On average approximately 12 percent of the observations were defaulted mortgages. The classifiers will have hard time predicting the skewed underrepresented observations. There is multiple methods to balance the data, oversampling, undersampling and Synthetic Minority Oversampling Technique (SMOTE), Haibo He (2009). SMOTE creates new synthetic observations of the minority class so that both defaulted and non-defaulted mortgages has equal representation. A new synthetic observations is generated

by finding the nearest neighbor of a minority observation. The thesis will use SMOTE to balance the uneven data.

3 Method

3.1 Logistic Regression

The base case and golden standard for probability prediction when it comes to machine learning is Logistic Regression and many previous work has been done on probability predictions with Logistic Regression as a benchmark, referring to one of the many previous work in the field, Sirignano, J., Sashwani, A. Giesecke, K. (2016). Logistic Regression is a machine learning algorithm that can be used for binary classification problems like predicting the probability of default or no-default.
In this case logistic regression models the probability of default using a sigmoid transform of linear function f features. Letting $y = 1$ denote default and $y = 0$ denote non-default, our assumption can be written:

$$Pr(y_i = 1) = \sigma\left(\mathbf{w}^t \mathbf{x}_i\right), \tag{1}$$

where σ denotes the sigmoid (or logistic) function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \tag{2}$$

and $\mathbf{w}$ are parameters of the model.
If we predict that $y_i = 1$ with probability p_i, it follow that the likelihood of observing y_i is $p^{y_i}(1-p)^{1-y_1}$. This means that the likelihood of a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$ is:

$$L(\mathbf{w}) = \prod_{i=1}^{n} \sigma\left(\mathbf{w}^T \mathbf{x}_i\right)^{y_i} \left(1 - \sigma\left(\mathbf{w}^T \mathbf{x}_i\right)\right)^{1-y_i}, \tag{3}$$

and the log-likelihood is given by:

$$\ell(\mathbf{w}) = \sum_{i=1}^{n} \left(y_i \log \sigma\left(\mathbf{w}^T \mathbf{x}_i\right) + (1 - y_i) \log(1 - \sigma\left(\mathbf{w}^T \mathbf{x}_i\right))\right). \tag{4}$$

It can be shown that ℓ is actually a concave function, meaning that any local maxima are also global maxima. Convex minimization problems (or equivalently, concave maximization) are well-studied and there are many highly effective approaches to solving them; Boyd, Vandenberghe (2009)

It is often desirable to add l^2 regularization to regression problems like this. This is used to reduce overfitting and to somewhat increase the stability of the problem. With the l^2 term, the function becomes strictly concave, guaranteeing a unique global maximum. The form of the regularized objective function is:

$$\ell_{reg}(\mathbf{w}) = \ell(w) - \frac{1}{c}\|\mathbf{w}\|^2, \tag{5}$$

where c is a hyper-parameter determining the strength of the regularization and can be fit using for instance cross-validation.

3.2 Support Vector Machines

SVMs are supervised Machine Learning algorithms which are capable of performing both linear and non-linear classification. The objective of the algorithm is to find an optimal hyperplane that separates the data. The optimal hyperplane is the one which has the biggest margin between the data points, where the margin is defined as the smallest distance between the hyperplanes that bound each class. Each point lying on these hyperplanes are called support vectors since they support the position of the hyperplanes and due to being vectors in an n-dimensional space. Data points that are not lying on the hyperplanes are called non-support vectors since they will not affect the position of our separating hyperplane. Figure 5 shows two cases whereas the first shows data that is linearly separable while the other is linearly non-separable, i.e., some points are on the wrong side of the margin.

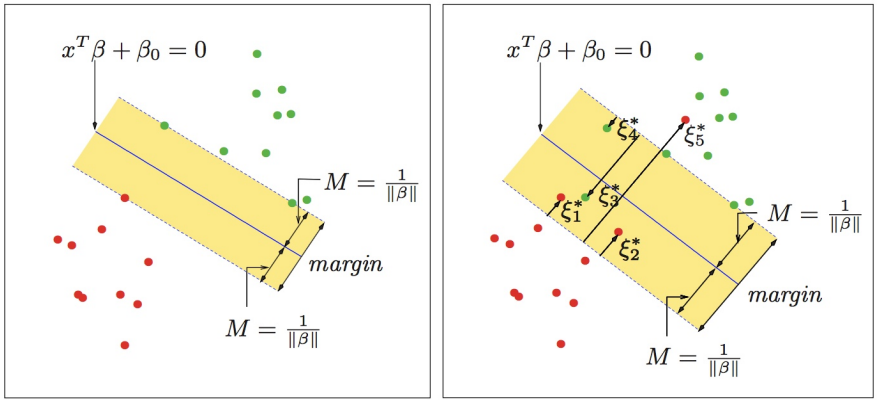


Figure 5: *Linearly separable vs. linearly non-separable. The color of the data points shows which class they belong to. The left figure depicts the case of the data being linearly separable. None of the observations are on the wrong side of the margin, i.e., there is no misclassification. The right figure shows the case of linearly non-separable data and thus allowing some of the observations to be misclassified.*

3.2.1 Linearly Separable Data

The linearly separable dataset $\mathfrak{D}$ is composed of n vectors $\boldsymbol{x}_i$, where $\boldsymbol{X}$ is called the input space. $\boldsymbol{x}_i$ is the i^{th} input vector, or observation, in that input

space. Each input vector $\boldsymbol{x}_i$ is paired with an output value y_i which indicates if the element belongs to the class or not. The dataset can be defined as:

$$\mathfrak{D} = \{(\boldsymbol{x}_i, y_i) | \boldsymbol{x}_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^n. \tag{6}$$

As mentioned above, the objective is to find a seperating hyperplane that maximizes the margin between the classes in the dataset. From figure 5, the separating hyperplane is defined as:

$$f(\boldsymbol{x}) = \mathbf{w}^T \boldsymbol{x} + b, \tag{7}$$

where $\mathbf{w}$ is the normal to the hyperplane and b is the bias. The equations (6) and (7) yields that the hyperplane will have the following properties:

$$\mathbf{w}^T \boldsymbol{x} + b > 0, \text{when } y_i = 1, \tag{8}$$

and

$$\mathbf{w}^T \boldsymbol{x} + b < 0, \text{when } y_i = -1. \tag{9}$$

Since $y_i \in \{-1, 1\}$, the lower and upper bound of the support vectors are defined as:

$$\mathbf{w}^T \boldsymbol{x}_l + b = -1, \tag{10}$$

and

$$\mathbf{w}^T \boldsymbol{x}_u + b = 1, \tag{11}$$

where $\boldsymbol{x}_l$ and $\boldsymbol{x}_u$ are the input vectors yielding the lower and upper bound. These two equations together with (6) can be combined into one single constraint:

$$y_i(\mathbf{w}^T \boldsymbol{x}_i + b) \geq 1, \text{for } 1 \leq i \leq n. \tag{12}$$

The margin, M, that bound each class, is obtained by taking the projection of the vector $(\boldsymbol{x}_u - \boldsymbol{x}_l)$ onto the normal vector to the separating hyperplane. This yields the following equation:

$$M = \frac{2}{\|\mathbf{w}\|_2} \tag{13}$$

The equations (12) and (13) result in the following maximization problem to solve:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{maximize}} \quad & \frac{2}{\|\mathbf{w}\|_2} \\ \text{subject to} \quad & y_i(\mathbf{w}^T \boldsymbol{x}_i + b) \geq 1, \text{for } 1 \leq i \leq n. \end{aligned} \tag{14}$$

The maximization problem above can instead be turned into a minimization problem (by the simple fact that maximization of M in (13) is the same as minimization of $\|\mathbf{w}\|$):

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \|\mathbf{w}\|_2 \\ \text{subject to} \quad & y_i(\mathbf{w}^T \boldsymbol{x}_i + b) \geq 1, \text{for all } 1 \leq i \leq n, \end{aligned} \tag{15}$$

which is a quadratic minimization problem with linear inequality constraints. Finally, (15) is equivalent to the following problem:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{subject to} \quad & y_i(\mathbf{w}^T \boldsymbol{x}_i + b) \geq 1, \text{for all } 1 \leq i \leq n. \end{aligned} \tag{16}$$

(16) can be solved through convex optimization, more specifically, by forming the Lagrangian $\mathfrak{L}$ of the problem:

$$\begin{aligned} \mathfrak{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i (y_i(\mathbf{w}^T \boldsymbol{x}_i + b) - 1), \\ \alpha_i \geq 0, \forall i, \end{aligned} \tag{17}$$

where α_i are Lagrangian multipliers that are necessary for solving these kind of optimization problems. Moving on, minimization of $\|\mathbf{w}\|_2$ and b requires partial differentiation of the Lagrangian primal problem (17) and putting them equal to zero:

$$\nabla_{\mathbf{w}} \mathfrak{L}(\mathbf{w}, b, \alpha) = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i = 0, \tag{18}$$

$$\nabla_b \mathfrak{L}(\mathbf{w}, b, \alpha) = \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{19}$$

The two partial derivatives above in combination with equation (17) results in:

$$\mathfrak{L}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j, \tag{20}$$

which is the Lagrangian dual problem. Maximization of the dual together with the Karush-Kuhn Tucker conditions gives the following:

$$\underset{\alpha}{\text{maximize}} \quad \mathfrak{L}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j, \tag{21}$$

$$\alpha_i \geq 0, \forall i, \tag{22}$$

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i, \tag{23}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \tag{24}$$

.

$$\alpha_i (y_i (\mathbf{w}^T \boldsymbol{x}_i + b) - 1) = 0, \forall i, \tag{25}$$

$$y_i (\mathbf{w}^T \boldsymbol{x}_i + b) - 1 \geq 0, \forall i. \tag{26}$$

Equations (22)-(26) results in the solution to both the primal and dual problem. The observations, $\boldsymbol{x}_i$, where the Lagrangian multipliers, α_i, are nonzero are called support vectors since, due to (23), $\mathbf{w}$ is characterized by α_i only. These support vectors determine the normal to the optimal separating hyperplane:

$$\hat{\mathbf{w}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{x}_i. \tag{27}$$

These results builds on the assumption that the data is linearly separable, which is often not the case in more realistic data sets. Instead, a more realistic scenario is that the data is not linearly separable, leading to some of the observations being wrongly classified.

3.2.2 Linearly Non-Separable Data

By introducing slack variables, ξ_i, for $1 \leq i \leq n$, some of the data points are allowed to be on the other side of the margin, i.e., allowing misclassification. The second case in figure 5 shows the principle of having some observations on the wrong side of the margin.

Adding the slack variables to equation (12) yields:

$$y_i (\mathbf{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \text{for } 1 \leq i \leq n. \tag{28}$$

The slack variables are also used as penalty variables for the optimization problem. That is, adding the sum of slack variables multiplied with a cost parameter, C, to (16) gives:

$$\begin{aligned} \underset{\mathbf{w},b,\xi}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & y_i (\mathbf{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \forall i \\ & \xi_i \geq 0. \end{aligned} \tag{29}$$

The Lagrangian primal for the optimization problem above becomes:

$$\mathfrak{L}(\mathbf{w}, b, \alpha, \beta) =$$
$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(y_i(\mathbf{w}^T\boldsymbol{x}_i + b) - (1-\xi_i)) - \sum_{i=1}^{n}\beta_i\xi_i, \tag{30}$$
$$\alpha_i \geq 0, \beta_i \geq 0, \forall i,$$

where α_i and β_i are Lagrangian multipliers. As for the linearly separable data, setting the partial derivatives of the Lagrangian equal to zero results in:

$$\nabla_{\mathbf{w}}\mathfrak{L}(\mathbf{w}, b, \alpha, \beta) = \mathbf{w} - \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i = 0, \tag{31}$$

$$\nabla_b\mathfrak{L}(\mathbf{w}, b, \alpha, \beta) = \sum_{i=1}^{n}\alpha_i y_i = 0, \tag{32}$$

$$\nabla_\xi\mathfrak{L}(\mathbf{w}, b, \alpha, \beta) = C - \alpha_i - \beta_i = 0. \tag{33}$$

The Lagrangian dual is obtained by inserting the equations (31)-(33) into (30):

$$\mathfrak{L}(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j. \tag{34}$$

Maximization of (34) together with the Karush-Kuhn Tucker conditions yields the following optimization problem:

$$\underset{\alpha}{\text{maximize}} \quad \mathfrak{L}(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j, \tag{35}$$

$$\mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i, \tag{36}$$

$$\sum_{i=1}^{n}\alpha_i y_i = 0, \tag{37}$$

$$\beta_i\xi_i = 0, \forall i, \tag{38}$$

$$C - \alpha_i - \beta_i = 0 \Leftrightarrow \xi_i(C - \alpha_i) = 0, \forall i, \tag{39}$$

$$\alpha_i(y_i(\mathbf{w}^T\boldsymbol{x}_i + b) - (1-\xi_i)) = 0, \forall i, \tag{40}$$

$$y_i(\mathbf{w}^T\boldsymbol{x}_i + b) - (1-\xi_i) \geq 0, \forall i, \tag{41}$$

$$\alpha_i \geq 0, \beta_i \geq 0, \forall i. \tag{42}$$

Equations (35)-(42) results in the solution to both the primal and dual. The same way as in the case of linearly separable data, the observations where the coefficients α_i are nonzero are called support vectors since, due to (36), $\mathbf{w}$ is characterized by α_i alone. The normal to the optimal separating hyperplane for non-linearly separable data is:

$$\hat{\mathbf{w}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{x}_i. \tag{43}$$

Both these methods are not always applicable, it all depends on the data. Some datasets are not as easily separable and requires other representation before classification can be made properly, thus introducing the Kernel Trick.

3.2.3 The Kernel Trick

To get round the problem with the data being non-separable, i.e., it is not possible to divide the data with a hyperplane, the data is mapped non-linearly into a feature space $\mathbf{x}_i \rightarrow \varphi(\mathbf{x}_i)$, which is a higher dimensional space. The mapping of the data will result in the data being separable by a hyperplane.

The feature space, $\varphi(\mathbf{x}_i)$, is defined as:

$$\varphi(\mathbf{x}_i) = (\varphi_1(\mathbf{x}_i), \ldots, \varphi_m(\mathbf{x}_i), \ldots, \varphi_N(\mathbf{x}_i)), \text{for } 1 \leq i \leq n, 1 \leq m \leq N. \tag{44}$$

There will be a reduction in the dimensionality if $N < n$. By mapping the observations from (35) into feature space and using the kernel function where $K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j)$, the following Lagrangian is obtained:

$$\mathfrak{L}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{45}$$

The solution to the Lagrangian above is given from equation (36) and results in:

$$f(\mathbf{x}) = \varphi(\mathbf{x})^T \mathbf{w} + b = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b, \tag{46}$$

In (45), α_i and b can be computed by solving $y_i f(\mathbf{x}_i) = 1$ for all $\mathbf{x}_i$ where $0 < \alpha_i < C$. C, as before, is the cost parameter.

As can be seen, $K(\mathbf{x}_i, \mathbf{x}_j)$ is the inner product for the mapped pairs of points in feature space. Both (44) and (45) contains $\varphi(\mathbf{x})$ through inner products. The reason for it being called a "trick" is that it is only required to know the Kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ and not the transformation $\varphi(\mathbf{x})$, thus, saving computational memory. However, a problem when working with kernels is not knowing the validity of a certain kernel. By use of Mercer's Theorem, the legitimacy of the mapping into feature space of a kernel is given, Minh, Partha and Yuan (2006). The theorem says that the function $K(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel if and only if it is a symmetric positive semi-definite function. A few common kernels are polynomial kernel, Radial basis function (RBF) kernel and neural network kernel, Abello, Pardalos, Resende (2002). Usually the choice of kernel depends on the problem at hand. This thesis will use RBF while it can handle situations when there is non-linear relationships between the class labels and attributes. In addition, contrast to polynomial kernel, the number of hyperparameters in RBF kernel is easier to control which reduces the model complexity. An interesting point with polynomial kernel is that it can combine two or more features together up to the order of the polynomial.

The value of the cost parameter C depends on how much of the maximization of the margin can be given up to minimize the misclassification, i.e., a trade-off between the both. A smaller value of C encourages a greater margin and thus allowing more misclassification, which in turn makes the surface smoother. For a larger value of C the opposite holds; optimization will choose a hyperplane with a narrower margin, resulting in fewer points on the wrong side of the classification.

3.2.4 The Radial Basis Function Kernel

As mentioned previously, a common kernel to use for Support Vector Machine classification problems is the Radial Basis Function (RBF) kernel. A general radial basis function is defined by:

$$K(\mathbf{x}, \mathbf{x}') = g(\|\mathbf{x} - \mathbf{x}'\|), \tag{47}$$

for some function g. Going forward, we will only consider the Gaussain RBF, which is one of the most widely used radial basis functions. This is given by the kernel:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right), \tag{48}$$

where we define $\gamma = \frac{1}{2\sigma^2}$ as the spread of the RBF, i.e., a low value on γ can be interpreted as the reach of the influence of a single training point is far and a high value on γ means that the reach of the influence of a single point is close. In other words, a small γ yields a model that might be too constrained and thus the ability to explain the complexity of the data will be affected negatively.

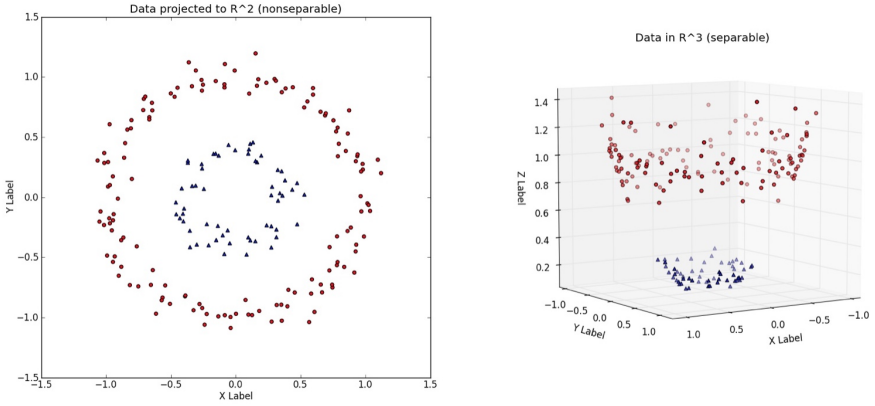


Figure 6: *An illustration of how the kernel trick manipulates the data and brings it to a higher dimension. The data points marked with a circle belongs to a class and the data points marked with a triangle belongs to another class. The observations in the left figure can not be separated with a hyperplane due to its shape. However, the kernel trick maps the data into a feature space, making it easier to separate, which can be seen in the right figure.*
Source: https://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html

To find the most suitable model for the data, the two parameters for the RBF kernel, C and γ, must be chosen wisely. Unfortunately, the procedure of finding the optimal parameters is both time-consuming and computationally costly (especially for large datasets). An approache of optimizing C and γ is through Cross-Validation and Grid-Search.

3.2.5 Cross-Validation

A common problem when trying to fit training data is overfitting, which can occur by having too many explanatory variables. Tendencies of overfitting can also arise when tweaking the cost parameter C, since information about the test set may be exposed into the model. A way of overcoming the issue is by splitting the dataset into another set, a so called validation set, Arlot and Celisse (2016). The procedure of the machine learning algorithm will begin with training on the training set. Afterwards, evaluation is performed on the validation until the analysis is successful. Lastly, final evaluation is performed on the test set.

A typical way of cross-validating data is by using the k-fold cross-validation method. The method is initiated by dividing the data into k equally large data sets. $k-1$ of the subsets are used as the training data, which will yield a model that can be tested against the last subset. The procedure is repeated so that each of the k subsets are used as test data. The performance of the

cross-validation is measured by the percentage of the data that is predicted correctly.

3.2.6 Grid-Search

The method of cross-validation can be used to conduct a so called grid-search. The search will be used to find the pair of parameters (C, γ) that yields the highest cross-validation score. The method is computationally costly due to many different pairs of (C, γ) that are cross-validated Sayrif, Prugel-Bennett, Wills (2016). At first it can be performed on a smaller grid, to avoid a time-consuming search. The first search on the coarser grid can be used as an indication of which region of the grid that results in a high cross-validation accuracy. Afterwards, a second search on the region which gave a high score can be performed to find an even better pair of parameters. For Logistic Regression there is only one parameter, C, regularization parameter, which needs to be optimized.

3.3 Scoring

Scoring functions keeps track of the performance of a certain model. More explicitly, they indicate how well a model predicts new data points. After the training of the dataset the different scoring methods will be used as a comparison of the performance.

There are different cases of classification, which, can be used for different scoring methods. The labels that will be used are:

- **True Positive (TP):** a positive prediction, true label is positive.
- **False Positive (FP):** a negative prediction, true label is positive.
- **True Negative (TN):** a negative prediction, true label is negative.
- **False Negative (FN):** a positive prediction, true label is negative.

3.3.1 Accuracy Classification Score

The Classification Accuracy is defined as the number of correct predictions made of the whole set of predictions. The scoring method is straightforward and is calculated as the following ratio:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{49}$$

which, obviously, gives a number between 0 and 1, where the performance of the model improves as the score gets closer to 1.

3.3.2 Precision, Recall and Fall-Out Score

The Precision Score, also known as Positive Predictive Value, is the number of True Positives divided with the number of examples with the true value positive:

$$Precision = \frac{\#\ of\ TP}{\#\ of\ TP + \#\ of\ FP}. \tag{50}$$

The Precision Score measures a classifiers ability to label data points that are positive as positive. If the Precision Score value is low it implies that the number of False Positives is high.

The Recall Score is defined as the number of True Positives divided with the number of True Positives and False Negatives:

$$Recall = \frac{\#\ of\ TP}{\#\ of\ TP + \#\ of\ FN}. \tag{51}$$

The Recall Score measures a classifiers ability to find all the predictions that are labeled as positive. A low value implies that the number of False Negatives is high.

The Fall-Out Score, or False Positive Rate, is the number of False Positives divided with the number of False Positives and True Negatives:

$$Fall\text{-}Out = \frac{\#\ of\ FP}{\#\ of\ FP + \#\ of\ TN}. \tag{52}$$

The Fall-Out Score measures the probability of labeling data points as positive when the true value is negative.

3.3.3 F_1-Score

In the F_1-Score method both the Precision Score and Recall Score are taken into consideration. It is a weighted average between those two scores, thus creating a balance between them. The value is calculated as:

$$F_1\text{-}Score = 2 * \frac{Precision * Recall}{Precision + Recall}. \tag{53}$$

Equation (53) shows that both the Precision and Recall contribute to the score with the same amount, which is useful when there is an uneven class distribution between False Positives and False Negatives.

3.3.4 Receiver Operating Characteristic - ROC

The ROC curve, Receiver Operating Characteristic, plots the two parameters: Recall and Fall-Out at different classification thresholds between 0 and 1. Figure 7 shows a typical ROC curve.

The ROC curve will show how good a classification model's performance is, i.e., how well the model is at labeling a data point correctly. To measure the performance of a model, the Area Under the Curve, AUC, of the ROC curve is used. The AUC will range between 0 and 1, where a model that is able to perfectly distinguish between classes has an AUC of 1. A model with an AUC of 0.5 has no ability to distinguish between classes and will always label a data point with a random class. In the case of the AUC being 0, the classification model will always incorrectly predict positives as negatives and negatives as positives.

However, when dataset is imbalanced where majority one of the two classes is very few as in the case of the Fannie Mae dataset, > 5 % defaulted. Clearly the dataset is skewed and ROC will not be able alone to present if the model is predicting well. Therefore, it is good to calculate additional measures that collect more specific aspects of the evaluation which takes us to Precision recall curve.

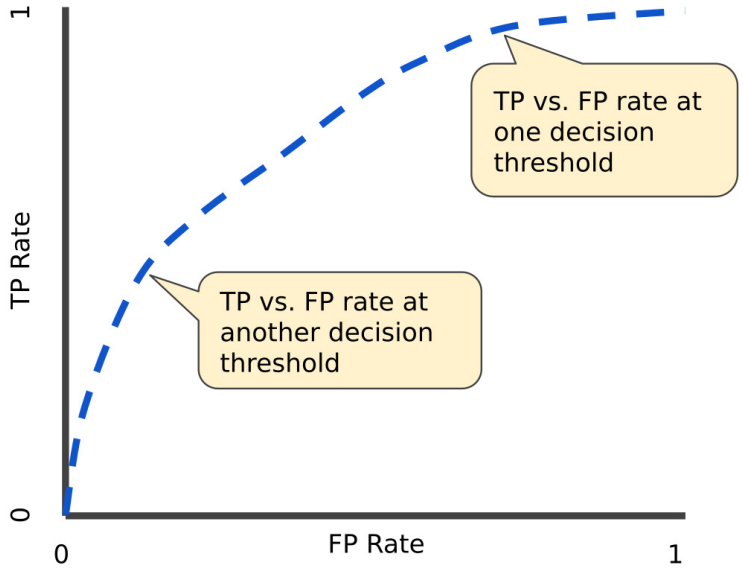


Figure 7: *Figure showing an ROC curve at different classification thresholds. The plot shows Recall on the y-axis versus the Fall-Out on the x-axis. A higher value on the y-axis means more True Positives and fewer False Negatives, while a lower value on the x-axis implicates fewer False Positives and more True Negatives.*

3.3.5 Precision-Recall Curve - PRC

Using equation (50) and (51), both precision and recall is useful where there is imbalance in the dataset and massive skew in the observations. What typically happens is a high number of true negatives. A precision-recall curve is a plotted below Figure 8, precision (y-axis) and the recall (x-axis) for different thresholds, similar to the ROC curve. The precision recall does not take into account the true negatives and only concerned with the correct prediction of the minority class, in this case defaults.

The no-skilled model would not be able to separate between the defaults and non-defaults, it would be a random classifier.

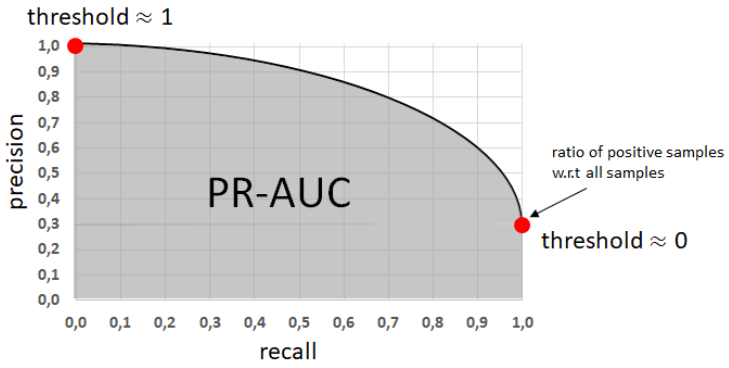


Figure 8: *Precision Recall Curve - The more skilled a model is the closer to the (1, 1) point the curve is, further away from the horizontal line of the no-skilled model.*

3.3.6 Bootstrapping for confidence intervals

To get confidence intervals for the resulting AUC scores, we use bootstrapping. This means that we resample (with replacement) our training data several times and, for each sampled dataset, we fit a model and calculate the resulting scores on the test data. This gives us a distribution of estimates of the AUC score which we can use to calculate various statistics and confidence intervals.
One can get a 95%-confidence interval from the distribution of scores by looking at the 2.5% quantile and 97.5%-quantile of the distribution. A different alternative is to estimate the standard deviation of the distribution and use a Gaussian approximation.

4 Results

This section gives account of the results obtained from the experiments. That is, the results from the pre-processed Fannie Mae mortgage loan data using Support Vector Machine Algorithm and Logistic Regression. The first set, pre-financial crisis, data consists of mortgage loans undertaken during 2000 and processed the performance up to 2005, whether a default incurred. Similarly, the second set is made up of house loans undertaken 2005 and performance investigated up to 2010, and the third set loans undertaken 2010 and performance investigated up to 2015, crisis and post crises sets respectively.

The whole first set consists of 1 264 975 samples, where each sample is made up of 5 variables. In other words, the first set is a matrix with the dimension 1 264 975$\times$5. However, since the set is too large we chose only a fraction of the set for our results. The same was done for the crisis and post crisis sets. Due to computational constrains only 400 000 data points were randomly selected for each time period. Furthermore, the data was normalized for each column, [0,1], as the attributes varies substantially. Borrower credit score (FICO) spans between 300-800 while interests rates are in single digit range.

Further, the whole set is divided into 3 subsets:

- Training data: 50% of the whole subset.
- Validation data: 25% of the whole subset.
- Test data: 25% of the whole subset.

4.1 Models

Running the entire algorithm with all its different stages on a quad-core PC 12 GB RAM was very time-consuming, including test feedback cycle taking several hours. The procedure includes cleaning of the data, picking out fragments and shuffling random observations from the whole data set. For logistic regression a grid search is used to cross-validate the data to find the correct regularization strength, C (the inverse of regularization strength). Same process for the SVM algorithm is used to optimize C and γ. As mentioned above, 400 000 out of 1.4-1.7 million data points were used for the whole process, training, testing and validation. The Logistic regression has only one parameter, C, the regularization to be optimized. However, for SVM there is the RBF kernel parameters C, regularization parameter, and γ, how much influence a training point has. To find the optimized parameters for C and γ was very time consuming.

The following results were obtained by randomly selecting a subset of the data, where the ratio between observations with delinquency status 0 or 1

were around 1:5. Due to the limited time the various ways of approaching imbalanced data set was not investigated and SMOTE was used, shortly mentioned in the data section.

To put in perspective, the analyses was carried out with the full dataset, 1.3 million data points, modelling SVM and finding the optimal C and γ, the code was running for over 9 hours using COLAB PRO+, 40 CPUs and 52Gb RAM memory.

Looking at Figure 9, borrower credit score (FICO) is clearly negatively correlated to the rest of the variables which is not a surprise. To measure forecast accuracy, the Receiver Operating (ROC) and Precision Recall Curve (PRC) is investigated. Due to the unbalanced data F1 score was optimized, finding the highest F1 score.

ROC-AUC for Logistic Regression result are 0.72 , 0.72 and 0.69, 2000-2005, 2005-2010, 2010-2015. Although, the more interesting score is the PRC-AUC and the highest value was found for the 2005 time period, 0.389, while 2010 had the worse 0.129. Each of the periods a C value was found using grid-search algorithm, this was not much time consuming for logistic regression and took minimal computational time.

SVM results, as mentioned, took more computational resources and as seen the results were only marginally better. 0.72, 0.72 and 0.70 for the periods 2000-2005, 2005-2010 and 2010-2015 respectively. Similar to Logistic regression, highest PRC-AUC for SVM was 2005 time period while lowest was for 2010 data, observed in table 2. Worth mentioning to the reader, the data set for 2010-2015 was very large and computational error was encountered multiple times due to not enough RAM memory. A larger number of data points was selected for the time period 2010-2015 due to larger skewness in the data set.

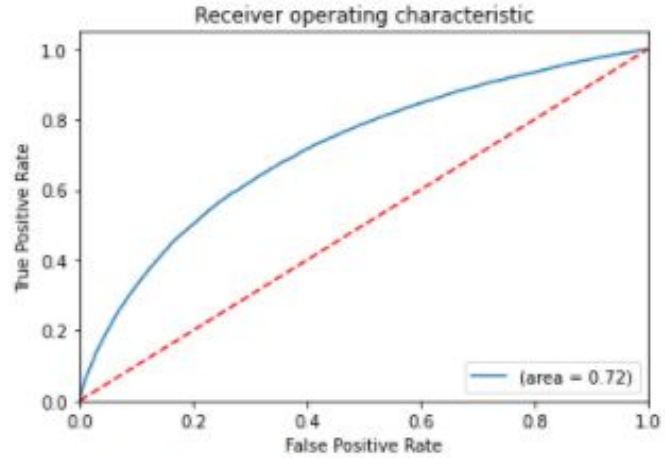
The confidence interval is tighter for logistic regression models in contrast to SVM models, although all is within 95% significance for both models.

Table 2: ROC- and PRC-AUC values including 95% confidence interval. The confidence interval is tighter for logistic regression compared to SVM models.

Model and Year	ROC-AUC	Conf.I	Precision-Recall AUC	Conf.I
LogReg 2000	0.719524	± 0.000054	0.316956	± 0.000028
LogReg 2005	0.718352	± 0.000143	0.388074	± 0.000256
LogReg 2010	0.697259	± 0.000276	0.132291	± 0.000229
SVM 2000	0.722620	± 0.000159	0.317464	± 0.000359
SVM 2005	0.723311	± 0.000400	0.389054	± 0.000757
SVM 2010	0.701747	± 0.000856	0.137313	± 0.000609

Figure 9: *Variable correlation matrix heatmap were identical for all time periods 2000-2005, 2005-2010-and 2010-2015. Borrower credit score have a negative correlation. A lower DTI, LTV and interest rate translates into a higher borrower credit score, i.e. better financial health.*

	precision	recall	f1-score	support
0	0.92	0.68	0.78	87757
1	0.26	0.64	0.37	15304
accuracy			0.68	103061
macro avg	0.59	0.66	0.58	103061
weighted avg	0.82	0.68	0.72	103061



Model: f1=0.371 auc=0.317

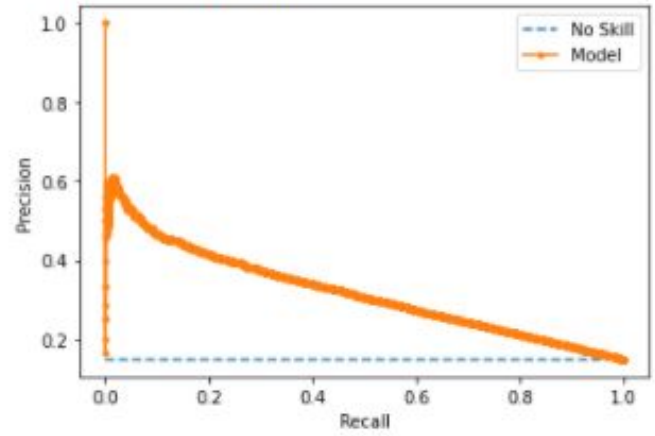
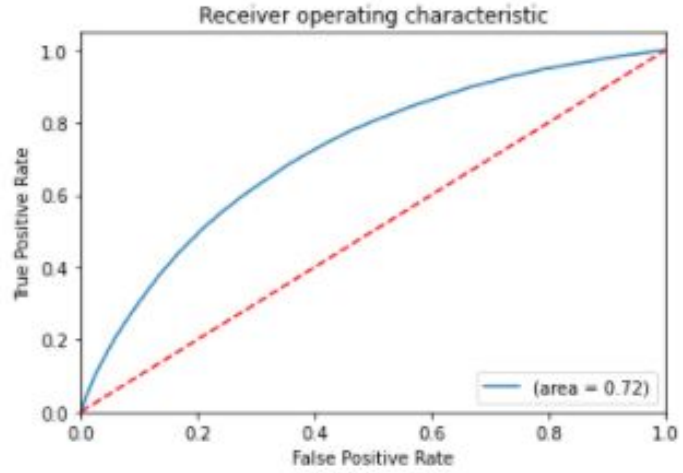


Figure 10: *Results for Logistic Regression year 2000 acquired data with Performance Metrics (Accuracy, Sensitivity vs Specificity, Precision vs Recall, and F1 Score), ROC curve and Precision-Recall Curve.*

	precision	recall	f1-score	support
0	0.88	0.67	0.76	85411
1	0.34	0.66	0.45	22437
accuracy			0.67	107848
macro avg	0.61	0.66	0.61	107848
weighted avg	0.77	0.67	0.70	107848



Model: f1=0.450 auc=0.388

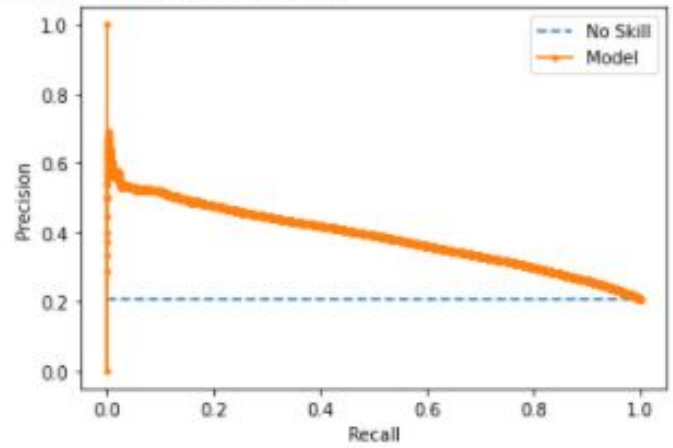
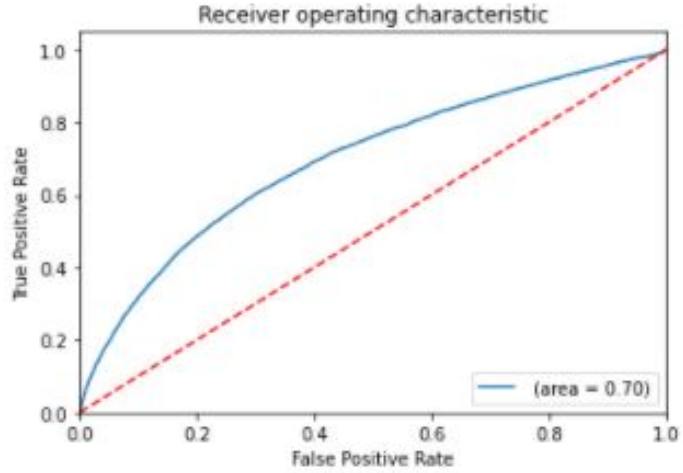


Figure 11: *Results for Logistic Regression year 2005 acquired data with Performance Metrics (Accuracy, Sensitivity vs Specificity, Precision vs Recall, and F1 Score), ROC curve and Precision-Recall Curve.*

	precision	recall	f1-score	support
0	0.97	0.71	0.82	190640
1	0.10	0.59	0.18	10695
accuracy			0.71	201335
macro avg	0.54	0.65	0.50	201335
weighted avg	0.92	0.71	0.79	201335



Model: f1=0.175 auc=0.130

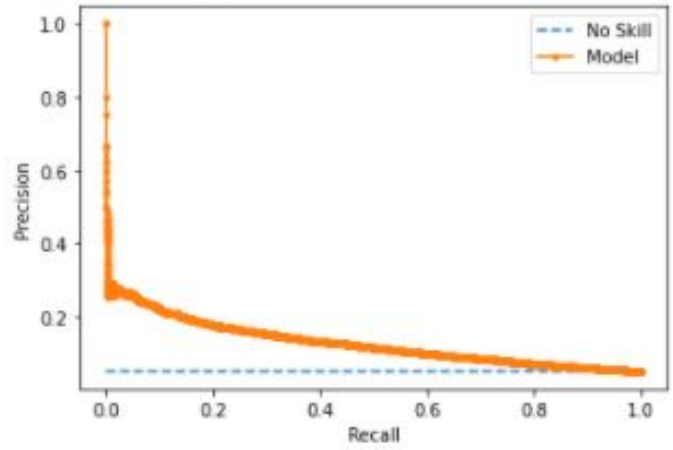
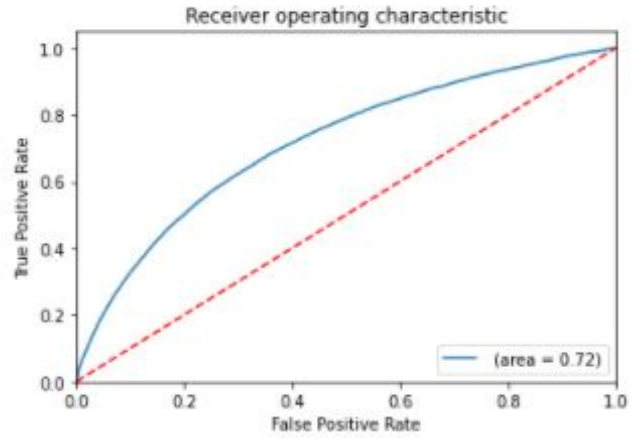


Figure 12: *Results for Logistic Regression year 2010 acquired data with Performance Metrics (Accuracy, Sensitivity vs Specificity, Precision vs Recall, and F1 Score), ROC curve and Precision-Recall Curve. Precision score for defaults is much lower than other data sets.*

	precision	recall	f1-score	support
0	0.92	0.67	0.77	87757
1	0.26	0.65	0.37	15304
accuracy			0.67	103061
macro avg	0.59	0.66	0.57	103061
weighted avg	0.82	0.67	0.71	103061



Model: f1=0.368 auc=0.317

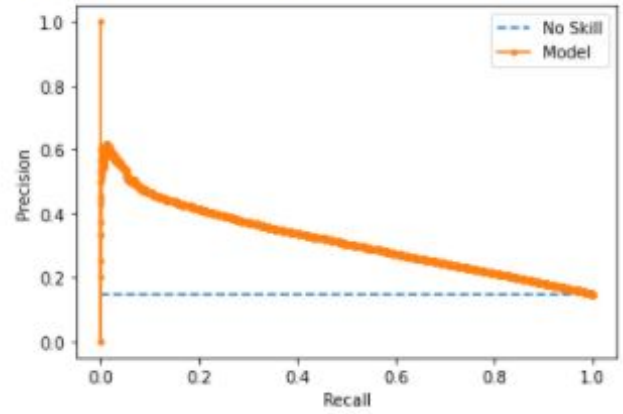
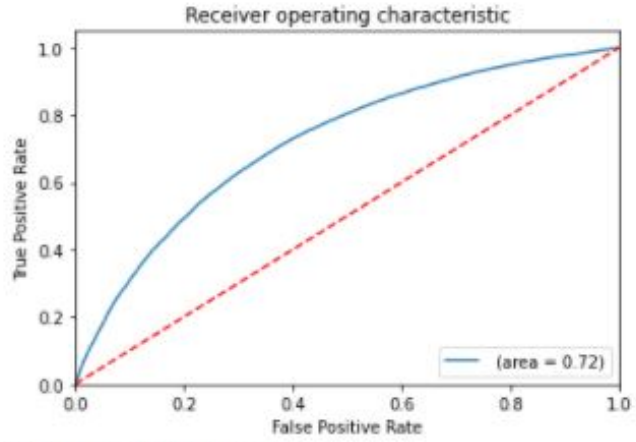


Figure 13: *Results for SVM year 2000 acquired data with Performance Metrics (Accuracy, Sensitivity vs Specificity, Precision vs Recall, and F1 Score), ROC curve and Precision-Recall Curve.*

	precision	recall	f1-score	support
0	0.88	0.69	0.77	85411
1	0.35	0.64	0.45	22437
accuracy			0.68	107848
macro avg	0.62	0.66	0.61	107848
weighted avg	0.77	0.68	0.71	107848



Model: f1=0.453 auc=0.389

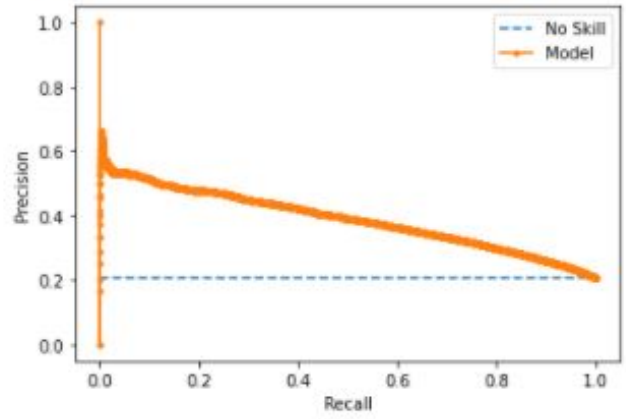
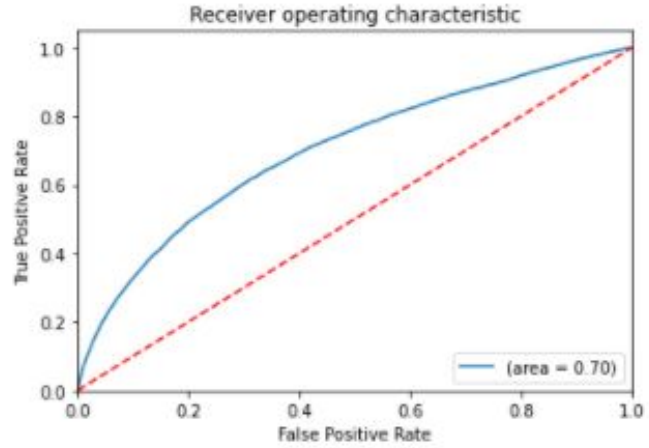


Figure 14: *Results for SVM year 2005 acquired data with Performance Metrics (Accuracy, Sensitivity vs Specificity, Precision vs Recall, and F1 Score), ROC curve and Precision-Recall Curve.*

	precision	recall	f1-score	support
0	0.97	0.65	0.78	190640
1	0.09	0.65	0.16	10695
accuracy			0.65	201335
macro avg	0.53	0.65	0.47	201335
weighted avg	0.92	0.65	0.74	201335



Model: f1=0.164 auc=0.137

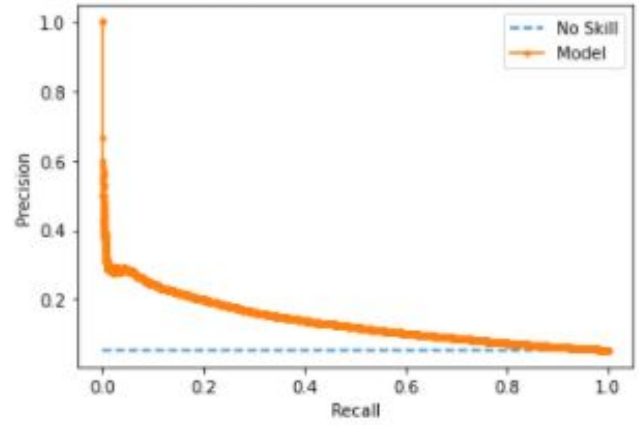


Figure 15: *Results for SVM year 2010 acquired data with Performance Metrics (Accuracy, Sensitivity vs Specificity, Precision vs Recall, and F1 Score), ROC curve and Precision-Recall Curve. Here too similar to logistic regression the precision for defaults is significantly lower than for the other period.*

5 Discussion

The results show that the predictors seem to predict the responder reasonably well. While SVM seems to be a somewhat better performer, the improvement over logistic regression is quite small. Two strengths of SVMs are being able to handle high-dimensional feature spaces well and being able to create non-linear decision boundaries using the kernel trick, neither of which is leveraged much in this problem. On the contrary, the predictors are few and seem to have fairly straight-forward linear relationships to the responder, indicating that a logistic regression might be sufficient.

Another advantage of logistic regression is that the output of the model can be interpreted as a probability (and the coefficients as change in log-odds per change in predictor), making ROC and precision/recall curves easy to calculate. For a SVM, on the other hand, some calibration is needed to turn the model outputs into probabilities.

This difference might partly explain the narrower confidence intervals for logistic regression. The calibration for SVM (in sklearn) uses a split of the training data, where the model is trained on one of the splits and then calibrated to output probability estimates by looking at the other split. This step both reduces the amount of data used to train the model, and introduces some random noise in the choice of splits. Both of these effects could make the confidence intervals larger.

In general, the confidence intervals are quite narrow. This is likely due to the large sizes of the dataset compared to the complexity of the models (which basically fit 4 parameters). It would be interesting to also investigate how the SMOTE oversampling interacts with bootstrapping. As done now, the bootstrapped samples are drawn from the dataset after SMOTE. This might cause odd effects where some samples in the original datasets are basically duplicated in the over-sampled dataset, leading to lower variance in the bootstrapped estimates. One might want to try using SMOTE after bootstrapping instead.

A clear benefit of logistic regression is the fast computational time while it obtains results very close compared to SVM, due to the kernel computation of the SVM model. If more variables were chosen, the results of SVM could likely yield a slightly better model but at a cost of even longer computational time. The results appeared to be same for all time periods investigated. Another way of overcoming the lengthy execution time is to use approximation methods to find the optimal pairs of C and γ, which was not considered due to time constraints.

A further research could have been done by using other methods to balance the unbalanced data. Since the data set is skewed towards non-defaulted mort-

gages, this has the implication that one class (default) will be under represented compared to the other by the model. Arguably, it does not come without problematic consequences, such as losing valuable information from the removed subset Haibo He(2009)

There is three additional ways to investigate the models by looking at the time-variable parameters 1) interaction variables, 2) more attributes included and lastly 3) categorize variables. Due to time constrains a thorough analysis was not carried out.

The emphasis for future work would be to add more variables such as: unemployment by zip code, FED rate, US Treasury, DTI change during the mortgage life time, and more models such as Random Forest and Neural Network. Lastly, as mentioned above to investigate the different ways to process unbalanced data.

6 Bibliography

Abello, J., M., P. and G.C., M., (2013) *Handbook of Massive Data Sets. Springer US. P.445-459*

A. L. Samuel (1959). "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development, vol. 3, no. 3, pp. 210-229, July 1959* Retrieved from `http://doi.org/10.1147/rd.33.0210`

Bengio, Y. (2009).Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning.* Retrieved from `https://www.iro.umontreal.ca/~bengioy/papers/ftml_book.pdf`

Bughin, J. et al (2) Artifical Intelligence The Next Digital Frontier. *Mckinsey Co.* Retrieved from `https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx`

B.J. Copeland *(2017)* Artificial intelligence. *Britannica* Retrieved from `www.britannica.com/technology/artificial-intelligence`

Chang, S. Dae-oong, S. Kondo, G. (2016) Predicting Default Risk of Lending Club Loans. *Stanford University* Retrieved from `http://cs229.stanford.edu/proj2015/199_report.pdf`

Cortes, C., & Vapnik, V. (1995). Support vector networks, *Machine Learning, 20, 273-297* Retrieved from `http://image.diku.dk/imagecanon/material/cortes_vapnik95.pdf`

CFPB (1) Regulatory laws in USA for consumers. *Consumer Financial Protection Bureau* Retrieved from `www.consumerfinance.gov`

Gau (1978). A taxonomix model for the risk-rating of residential mortgages, *The Journal of Business 51(4), 687-706.* Retrieved from `http://www.jstor.org/stable/2352656`

Ha Quang, Minh; Niyogi, Partha; Yao, Yuan. (2006) Mercer's Theorem, Feature Maps, and Smoothing *COLT'06: Proceedings of the 19th annual conference on Learning Theory June 2006 Pages 154–168* Retrieved from `https://doi.org/10.1007/11776420_14`

Haibo He, Member, IEEE, and Edwardo A. Garcia (2009) Learning from Imbalanced Data *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* Retrieved from `http://www.ele.uri.edu/faculty/he/PDFfiles/ImbalancedLearning.pdf`

He, He, and Ali Ghodsi. (2010) Rare class classification by support vector machine *Pattern Recognition (ICPR)* Retrieved from `https://hhexiy.github.io/docs/presentation/rcsvm_poster.pdf`

Vandell, Kerry D. (1978). Default risk under alternative mortgage insturments, *The Journal of Finance, vol. 33, no. 5, [American Finance Association, Wiley], 1978, pp. 1279–96,* Retrieved from `https://doi.org/10.2307/2327266`

Von Furstenberg (1969). Default risk on fha-insured home mortage as a function of the terms of finance: A quantitiative analysis, *Journal of Finance*

24, 459-477. Retrieved from https://doi.org/10.2307/2325346

Von Furstenberg (1970). Risk structures and the ditribution of bentifs withtin the fha home mortgage insurance program, *Journal of Money and Banking 2, 303-322.* Retrieved from https://doi.org/10.2307/1991011

Samuel, Arthur L. (1959) Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.368.2254&rep=rep1&type=pdf

Schölkopf, B., A. Smola, R. C. Williamson, and P.L. Bartlett (2000). New Support Vector Algorithms. *Neural Comput 2000; 12 (5): 1207–1245.* Retrieved from https://doi.org/10.1162/089976600300015565

Sirignano, J., Sashwani, A. Giesecke, K. (2016).Deep Learning for Mortgage Risk. *Cornell University Library.* Retrieved from https://arxiv.org/abs/1607.02470

Stephen Boyd, Lieven Vandenberghe (2009) *Convex Optimization. Cambridge University Press*

Bachelor's Theses in Mathematical Sciences 2021:K43
ISSN 1654-6229

LUTFMS-4012-2021

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/