Faculty of Natural and Mathematical Sciences
Department of Informatics

Bush House, 30 Aldwych
Strand Campus
London WC2B 4BG
United Kingdom
https://www.kcl.ac.uk/nms/depts/informatics

7CCSMDPJ

Individual Project Report 2017/18

Student Name: Kevin Jolly

Student Number: 1767397

Project Title: Predicting loan defaults for LendingClub

Degree Programme: MSc Data Science

Date of Submission: 24 August, 2018

Supervisor Name(s): Dr. Sophia Tsoka

Word Count: 12,065

This dissertation is submitted for the degree of MSc in Data Science.

☒ I **agree** to the release of my project

□ I **do not** agree to the release of my project

Signature: Kevin Jolly **Date:** 24/08/2018

ABSTRACT

LendingClub is one of the world’s largest peer to peer lending institutions that matches potential borrowers to investors through an online platform. Defaults on these loans cost the investor a large amount of money that can otherwise be saved with appropriate information that characterize bad performing loans. Predictive analytics can provide value to the investor by helping them identify the type of loans that are most likely to default. The primary objective of this project is to perform a detailed analytical study on the loans that are offered by LendingClub in order to understand the characteristics that differentiate loans that have defaulted from the loans that have not and to identify and evaluate different machine learning techniques that can be used to predict these defaults based on the needs of the investor.

The project has identified the key behavioral characteristics of loans that have defaulted based on the loan and borrower characteristics along with the geographical region of origination using visual analytics. The project has also implemented a wide array of feature selection techniques in order to understand the top predictors of default. Finally, the project implements eight highly interpretable classification algorithms in order to compare and evaluate the results of each algorithm. This provides the investor with a wide range of models to choose from based on their investing style.

Novel contributions include the in-depth visual analytics deployed using statistical plots that provide a holistic view of the defaulted loans, best performing loan grades and geographical analysis of the states with the highest proportion of defaults. A new approach to feature selection that implements filter selection techniques in combination with recursive feature elimination in order to identify the best predictors of default along with the interpretation of the best performing predictive model makes the project unique in the field of peer-to-peer lending.

ACKNOWLEDGMENTS

I would like to thank my project supervisor, Dr. Sophia Tsoka, for her valuable feedback & guidance throughout the duration of the project.

TABLE OF CONTENTS

GLOSSARY

AUC	Area Under the Curve	44
CART	Classification And Regression Tree	17
DTI	Debt To Income Ratio	54
K-NN	K-Nearest Neighbors	47
ROC	Receiver Operator Characteristics	22
URL	Uniform Resource Locator	31

LIST OF FIGURES AND TABLES

FIGURES:

TABLES:

1 INTRODUCTION

1.1 Overview & domain

The peer to peer lending industry is a relatively new alternative/non-bank credit industry that has become one of the fasting growing lending sectors in the world with an estimated compound annual growth rate of 51.5% between the years of 2016 to 2022 [1].

Some of the fundamental reasons for the growth of the peer to peer lending industry can be attributed to the lower interest rates, ease of securing credit and transparency that benefits both the lender and the borrower.

As of 2017, LendingClub is the world's largest peer to peer lending institution [2]. The data about the loans issued by LendingClub is provided on their official website. This helps investors make informed decisions about the loans that they decide to invest in.

Being an investor, avoiding defaults on these loans is of utmost importance in order to minimize any financial loss. Therefore, providing investors with the analytical framework that can enable them to differentiate between the types of loans that tend to default and the types of loans that don't default provides immense value.

Interpretable predictive models will enable the investor to understand what percentage of loans from a pool of invested loans are likely to default. This will enable the investor to make more informed decisions about the way in which they should allocate their money, thereby maximizing their financial returns while avoiding losses at the same time.

Therefore, this project uses a combination of visual analytics and predictive models in order to help the investor make the best possible investment decisions.

1.2 Motivation

Witnessed the impact and value an analytical framework offers while pursuing an internship at a company that provides advanced analytics on the loans that are issued by peer to peer lenders to institutions/investors and a keen interest in making data driven investments that can maximize returns while minimizing loss are the key driving motivators behind the project.

1.3 Aims & Objectives

The main aims and objectives of the project are as follows:

- To identify the behaviour/characteristics that differentiate the defaulted loans from the loans that have been repaid.
- To identify the grade of loans that performed the best and the associated reasons.
- To determine if the geographic region of origin of the loans affect the rate at which the loans were defaulted.
- To identify the features that best predict if a loan was likely to default.
- To evaluate and determine the machine learning classification model that best predicts if a loan will default or not with a high amount of interpretability.

1.4 Summary of conclusions

Subdividing the loans issued by LendingClub into four key characteristics – loan characteristics, borrower characteristics, borrower indebtedness and credit history, the attributes under each characteristic were analysed using visual analytics and the key traits and patterns that differentiate the loans that have defaulted and the loans that have been fully repaid was established.

Identified the grade of loans that had the highest proportion of defaulted and fully repaid loans and determined the potential reasons as to why these loan grades performed this way by using the loan and borrower characteristics.

Performed geographic analysis of the loans in order to determine why certain states in the United States had a higher proportion of defaulted loans by using the different characteristics of the loans.

Utilized a wide range of feature selection methods in order to determine the top ten features that best predicted if a loan would default or not and used these features in order to implement, evaluate and compare the performance of eight highly interpretable machine learning classification models and identified the model that best predicts if a loan would default or not.

2 BACKGROUND

2.1 Characteristics of loans issued in peer to peer lending

The fundamental aim of this section is to highlight the key characteristics of the loans issued by LendingClub and the attributes that fall under each characteristic.

There are five characteristics [3] under which the different attributes of the loans fall under. The first characteristic is the **borrower assessment** which includes attributes such as the interest rate, loan grade and loan subgrade [3]. The second characteristic is the **loan characteristics** which includes attributes such as loan purpose and loan amount [3]. The third characteristic is the **borrower characteristics** which includes attributes such as the annual income of the borrower, home ownership status and length of employment [3].

The fourth characteristic is the **credit history** which includes attributes such as the credit history length, delinquency in the last two years, inquiries in the last six months, derogatory public records, revolving utilization rates, number of open credit lines and months since last delinquency [3]. The fifth and last characteristic is the **borrower indebtedness** which includes the attribute – debt to income ratio. [3].

2.1.1 Borrower Assessment

A brief description about the attributes that fall under the borrower assessment characteristic is provided in this section. The first attribute is the **interest rate** [3] which is the value of the interest that is assigned to the loan issued by LendingClub.

The second attribute is the **loan grade** [3]. The loans in LendingClub are categorized into grades that go from A to G. LendingClub uses external credit grading agencies in order to give each loan a grade [3]. Grade A loans are considered risk averse while grade G loans are considered high risk.

The third attribute is the **loan subgrade** [3]**.** Each grade of loan is subdivided into five sub grades. Grade A loans are divided into A1, A2, A3, A4 and A5. Therefore, there are a total of 35 sub grades that go from A1 to G5. Sub grades help in making the categorization of loans much more specific compared to the generic nature of the seven grades. Once again loans having the sub grades A1 to A5 are considered risk averse with A1 being the safest while loans having the sub grades G1 to G5 are considered high risk with G5 having the highest risk.

2.1.2 Loan Characteristics

A brief description about the attributes that fall under the loan characteristics are provided in this section. The first attribute is the **loan purpose** [3]**.** This attribute is used to specify the purpose or reason for which the loan was borrowed by the borrower. The different types of loan purposes are – debt consolidation, credit card, home improvement, major purchase, small business, car, medical, moving, wedding, house, vacation, educational, renewable and others. The second attribute is the **loan amount** [3]. This attributes specifies the amount of money borrowed by the borrower in U.S dollars.

2.1.3 Borrower Characteristics

A brief description about the attributes that fall under the borrower characteristics are provided in this section. The first attribute is the **annual income** [3]**.** This attribute is used to specify the annual income of the borrower in U.S dollars.

The second attribute is the **home ownership status** [3]**.** This attribute is used to specify the type of housing the borrower is presently living in. The different types of housing are – rent, mortgage, own, none, any and other [3]. The third and final attribute is the **employment length** [3]**.** This attribute is used to specify the number of years the borrower has been employed for.

2.1.4 Credit History

A brief description about the attributes that fall under the credit history characteristics are provided in the section below. The first attribute is the **credit history length** [3]**.** The credit history length is an attribute that specifies the number years it has been since the borrower opened their earliest credit line from the current year.

The second attribute is the **delinquency in the last two years** [3]**.** This attribute specifies the number of delinquencies the borrower has had in the past two years. A delinquency is when the borrower has had a payment that has been due for over thirty days. The third attribute is the **inquiries in the last six months** [3]**.** This attribute specifies the number of inquiries that the borrower has had about their file by creditors in the last six months.

The fourth attribute is the **derogatory public records** [3]**.** This attribute specifies the number of derogatory public records that are present in the borrower's file. Derogatory public records can range from minor to major acts of offenses that may or may not be criminal in nature.

The fifth attribute is the **revolving utilization rates** [3]**.** The revolving utilization rate is the amount of credit the borrower is using relative to all available credit.

The sixth attribute is the **number of open credit lines** [3]. This attribute specifies the number of credit lines that are presently open concurrently in the borrower's file. The seventh and final attribute is the months since last delinquency**.** This attributes specifies the number months it has been since the borrower's last recorded delinquency.

2.1.5 Borrower Indebtedness

A brief description about the attribute that falls under the borrower indebtedness is provided in this section. The first attribute is the **debt to income ratio** [3]**.** The debt to income ratio is the total debt that is paid by the borrower on a monthly basis excluding mortgage and the LendingClub loan itself to the total monthly income of the borrower.

2.2 Feature selection

The fundamental aim of this section is to highlight the theory behind some of the most popular feature selection methodologies that are used in predictive analytics. The project makes use of two methods of feature selection – Filter and Wrapper.

2.2.1 Filter Method

The filter method of feature selection is composed of a generic set of rules that are used to remove features that provide no predictive value. This method of feature selection is computationally less expensive [4].

The filter method of feature selection can be implemented in order to filter features that are constant or quasi-constant in nature by filtering out the features that have zero or near zero variances.

Removing features that are exactly alike in every manner aids in dimensionality reduction and is an important technique in the filter method of feature selection in order to remove features that are redundant to the predictive model.

Using correlation as a measure to select features is an important technique in the filter method of feature selection. Features provide predictive value when they are correlated with the target attribute but are uncorrelated with each other [5]. Therefore, using a correlation matrix in order to filter out the features that are highly correlated with each other removes features that offer no predictive power.

Finally, univariate feature selection techniques such as the implementation of a decision tree algorithm to understand the predictive power of each feature with respect to the target attribute is useful to filter out the features that perform substantially worse than random guessing

2.2.2 Wrapper Method

The wrapper method of feature selection is composed of classification algorithms such as the decision trees and random forests that are used to select a subset of features that provide the most predictive value for that classification algorithm alone. As a result, the wrapper method of feature selection is more specific in nature and is computationally more expensive. A performance metric such as the accuracy score is used to evaluate the best subset of features [4].

2.3 Machine learning models

The fundamental aim of this section is to highlight the theory behind the classification models used to predict defaults.

2.3.1 K-Nearest Neighbors

The K-Nearest Neighbors algorithm is a classification algorithm that uses a distance metric such as the Euclidian distance in order to classify an unknown data point. Implementing a K-Nearest Neighbors algorithm with the number of neighbours equal to N would imply that the classification algorithm would search for the closest N neighbors to the unknown data point based on the Euclidian distance and classify this data point as a majority class among these N neighbors [6].

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Equation 1: Euclidian distance formula

In the equation above, x_2, x_1, y_2, y_1 are the co-ordinates of the unknown data point and the data point that is closest to it along the x and y axis.

An example of such an implementation would be when the number of neighbours is set to four and three out of the four neighbors of the unknown data point belong to class ‘A’, while one out of the four neighbors of the unknown data point belong to class ‘B’ , the unknown data point would be then classified as belonging to class ‘A’.

2.3.2 Logistic Regression

The logistic regression is mathematically centred around the logit function which transforms the equation that predicts the outcome as a function of the input features into a probability that is between zero to one [7].

Therefore, using logistic regression the model returns a probability that describes how likely a particular loan is going to default.

2.3.3 Linear Support Vector Machines

A linear decision boundary divides a two dimensional feature space in two halves. Each half of the boundary consists of data points that belong to a unique category or class in the most ideal case. In a feature space having higher number of dimensions, the linear decision boundary is called a hyperplane.

Finding the ideal hyperplane that can perfectly divide the feature space such that each side of the hyperplane consists of data points that belong to a unique category is the fundamental theory behind the Linear Support Vector Machines [8].

2.3.4 Decision Tree

The decision tree splits data by selecting an attribute that minimizes the amount of impurity as the root of the tree. This value of this impurity is commonly referred to as the Gini Index and is used by the Classification And Regression Tree (CART) algorithm to select attributes at each node of the tree. This Gini Index is given by the formula below [9]:

$$Gini\ (D) = 1 - \sum_{i=1}^{m} p_i^2$$

Equation 2: Gini Index formula

In the equation above D is the attribute or node at which a set of data points are to be further partitioned into further branches, p_i is the probability that a tuple at D will be classified as a particular class [9].

Once the root of the tree is selected using the Gini Index, the tree builds itself in a recursive manner until no more attributes can be formulated into a branch of the tree. Therefore, when an unknown example has to be classified using the decision tree, it will follow the rules that the tree has built from the root of the tree all the way down to the branches until it has been classified into a particular class.

2.3.5 Random Forests

The fundament concept behind the Random Forests is that of ‘bagging’. Bagging is when multiple decision trees are used together in order to make a prediction [9]. Each decision tree is trained on a subset of the data and the average prediction performance is extracted across all the trees.

2.3.6 Gradient Boosted Trees (AdaBoost)

The theoretical foundation of the AdaBoost algorithm is based on making weak learners into strong learners. A weak learner is a classifier or a rule that performs only slightly better than guessing at random. A strong longer is a classifier or a rule that predicts the outcome or target with a high level of accuracy [10].

By combining many weak learners together the AdaBoost improves the overall prediction accuracy of the classifier.

2.3.7 Naive Bayes

The Naive Bayes classifier is based on Bayes theorem [11]. The equation for Bayes theorem is given below:

$$P(Target \mid Features) = \frac{P(Features \mid Target) \times P(Target)}{P(Features)}$$

Equation 3: Bayes Theorem

In the equation above the probability of predicting the target is computed given a set of features by independently evaluating the probabilities of each feature with respect to the target which is then multiplied with the probability of the target and divided by the probability of the features [11].

2.3.8 Ensemble Classifier

The ensemble classifier uses several machine learning models of the same or different type in order to make a prediction based on majority voting [12].

2.4 Performance Evaluation Methods

The fundamental aim of this section is to provide a brief description about the different evaluation methods that are used to evaluate and compare the performance of different machine learning models.

2.4.1 Confusion Matrix

The confusion matrix is a contingency table that represents the predicted classes and the actual classes [13]. A loan that has defaulted is said to be the target or positive class.

If a loan was predicted as a default and the loan was actually a default it is known as a true positive. If a loan was predicted as a default but the loan was not actually a default it is known as a false positive.

If a loan was predicted as not defaulted and the loan was not actually a default it is known as a true negative. If a loan was predicted as not defaulted but the loan was actually defaulted it is known as a false negative.

Some of the performance metrics that can be extracted from the confusion matrix are accuracy, precision, recall, F-1 score, false positive and false negative rates [13].

The accuracy can be extracted from the confusion matrix by using the formula given below [13]:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Positives + True\ Negatives + False\ Negatives}$$

Equation 4: Accuracy Formula

A high value of precision implies that not many of the loans that have not defaulted are predicted as defaulted. Precision is a metric that provides the accuracy of the predicted loans that are defaults [13]. The precision can also be extracted from the confusion matrix by using the formula given below [13]:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Equation 5: Precision Formula

A high value of recall implies that the machine learning model predicted most of the defaulted loans correctly. Recall is a metric that provides the proportion of the true positives to the sum of the true Positives and the false Negatives [13]. The recall can also be extracted from the confusion matrix by using the formula given below [13]:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Equation 6: Recall Formula

The F-1 score is a single measure that can evaluate both the precision and recall as it is the harmonic mean between the two [13].

The false positive rate is the ratio of the loans that are predicted as defaulted but have actually not defaulted to the sum of the loans that are predicted as defaulted but actually have not and the loans that are predicted as not defaulted and have actually not defaulted. The false positive rate is also known as the fall-out rate [13] as it is the rate at which the model predicts the loans as defaulted when they actually have not. The false positive rate can be extracted from the confusion matrix by using the formula given below [13]:

$$False\ Positive\ Rate = \frac{False\ Postives}{False\ Positives + True\ Negatives}$$

Equation 7: False Positive Rate formula

The false negative rate is the ratio of the loans that have been predicted as not defaulted when they actually have defaulted to the sum of the loans that have been predicted as not defaulted when actually have and the loans that have been predicted as defaulted when they actually have. The false negative rate is also known as the Miss Rate [13] as it is the rate at which the classifies fails to predict the defaulted loans. The false negative rate can be extracted from the confusion matrix by using the formula given below [13]:

$$False\ Negative\ Rate = \frac{False\ Negatives}{False\ Negatives + True\ Positives}$$

Equation 8: False Negative Rate formula

2.4.2 Lift Curve

The lift is a ratio that measures how likely the target class is going to be present in a sample of the total data when a machine learning model is used to when the machine learning model is not used.

A high value of lift typically indicates that a particular machine learning model is going to have a very high probability of accurately predicting that the target class is present for a particular percentage of the total sample [14].

2.4.3 Cumulative Gains Curve

The cumulative gains curve captures the total percentage of the target class for a particular percentage of the sample population [14].

Cumulative gains curve provide a way to choose a model that captures that maximum percentage of the target class from a large percentage of the sample.

2.4.4 Area under the curve

The Receiver Operator Characteristic or the ROC curve is a plot that is constructed between the True Positive Rate on the y-axis and the False Positive Rate along the x-axis [13]. The optimal ROC curve is one in which the value of the True Positive Rate is 1 [13] indicating that the classifier predicts the defaulted loans with an accuracy close to 100% and the False Positive Rate is 0 [13] indicating that the classifier predicts very few loans as defaulted when they actually have not.

When such an optimal ROC curve is constructed, the area under the curve attains the maximum possible value of 1 indicating that the model performs extremely well.

3 RELATED WORK

This section provides a description about the recent advancements and work done in the application of building and evaluating predictive models in order to predict defaults in peer to peer lending.

3.1 Feature Selection in peer to peer lending

Selecting features that provide a high predictive value is an integral part in any project that involves the building and evaluation of predictive models. A greater number of features requires an equally greater number of examples. This is known as the curse of dimensionality.

A predictive model performs much better at finding patterns in the data when the number of features or attributes are lower. It also makes the process computationally less expensive.

Work done in predicting defaults in peer to peer lending has used a combination of a correlation matrix and random forests in order to pick the best predictors of defaults [15].

The random forests are first implemented in order to select features with a high importance. Once the best predictors of default were selected, a correlation matrix was used to remove features that were highly correlated to another feature among the best predictors [15].

This approach of feature selection is extremely good at establishing how important a feature is when it comes to predicting defaults as the feature importance scores are determined in order to provide the best performance for the random forest model.

The correlation matrix is a technique that is used to filter out features that are highly correlated to each other since having two variables that are highly correlated provide no predictive value.

The drawback to this method of feature selection is that the best number of predictors of default are usually picked arbitrarily without analytically analysing what the best number of predictors should be.

Additionally, evaluating the performance of the top predictors by only implementing the random forest model may give biased results as the feature selection is done by evaluating the predictors that best perform on the random forest algorithm.

Finally, the feature selection does not implement filter selection techniques such filtering out features with zero or almost zero variance and univariate feature importance filtering using decision trees.

A lack of general analytical depth and a heavy dependence on algorithms such as random forests reduces the interpretability of the reason behind why the best predictors were selected.

The work presented in this project aims to address this issue with feature selection performed in recent work done in predicting defaults in peer to peer lending by providing an analytical base for evaluating what the best number of features should be.

Biases in feature selection are addressed in this project by evaluating the performance of the selected features using an algorithm that is inherently different from the one that is used for the process of feature selection.

Finally, filter methods to remove features that provide no predictive value are done before deploying the random forests in order to make the process computationally more efficient.

3.2 Comparing and evaluating multiple predictive models

Interpretable predictive models are the ones in which the behaviour of each variable in predicting the target can be explained and visualized. Such models find utility in the financial and lending markets because of several regulations by government bodies to make the behaviour of such models interpretable in nature.

Investors also find value in interpretable models because they can understand why certain attributes behave the way they do in a predictive model and make their investment decisions accordingly.

Recent work in building predictive models in order to predict defaults have used a combination of several highly interpretable algorithms and have compared and contrasted the performance of each [15][16][17].

One of the recent works in predicting defaults proposes to provide a solution to help investors avoid investing into bad loans by predicting the probability of a default thus helping them maximize their returns while keeping their losses low. The work explains the data, the features, feature selection, the different machine learning models used to predict the probability of default and the metrics that were used to evaluate the different models [16]. The work is critical at explaining why certain features were picked over the others in the data using ablative analysis methods by implementing a logistic regression model to evaluate the performance of the model with and without features of interest [16]. The work is also critical at explaining the accuracy the model should provide with respect to the probability of default that the dataset inherently contains. Additionally, a wide range of machine learning models such as logistic regression, support vector machines and Naive Bayes were used to predict the probability of default. A nice comparison of the results using a wide variety of metrics that are not only limited to the accuracy such as precision, sensitivity and specificity was done. However, this work lacks the inclusion of K-Nearest neighbors, random forests, decision trees, gradient boosted trees and ensemble methods to predict the defaults.

Another recent work on the topic of predicting defaults in the domain of peer to peer lending proposes a solution to predict the probability of default on loans using a traditional set of supervised machine learning algorithms discussed above along with tree based models such as the decision tree and random forests while evaluating the performance of all the models comparatively using the area under the curve metric [17]. This work however lacks the inclusion of supervised machine learning algorithms such as the K-Nearest neighbors and gradient boosted trees and broad set of evaluation metrics such as precision, recall, false positive and negative rates along with metrics such as the cumulative gains and lift curve.

The last and final piece of work done on the topic of evaluating predictive models to predict defaults in peer to peer lending proposes comparing random forests, decision trees and neural networks [15]. The work compares the performance of these models by using the lift curve and the accuracy by employing a ten-fold cross validation to obtain the results [15]. This work provides an alternative approach to evaluating the performance of classifiers by making using of the lift curves – an approach that has not been used in previous work.

The major drawback in this piece of work is the lack of interpretability for the neural networks employed as they are essentially 'black-boxes' that do not explain how the different features have interacted with respect to the target variable in order to make predictions.

3.3 Key differences in implementation

This section highlights the key differences that make this project unique to the related work that has been done in the domain of predicting defaults in peer to peer lending.

While most of the recent work proposes a solution that is solely focused on the use of building, comparing and evaluating multiple predictive models, this project takes an analytical approach by utilizing data visualization as a key foundation for providing investors with appropriate information that will allow them to characterize loans that default and the loans that do not, based on geographical location, loan and borrower characteristics.

Furthermore in recent work, feature selection has been shallow by relying on algorithms such as random forests which may result in biased results as features that are selected are best suited for random forests. Therefore, this project addresses this problem by using one algorithm, the random forest for picking the best features and evaluating the results with a different classification algorithm – the K-Nearest Neighbors. The lack of analytical depth when picking the best number of features/predictors is also addressed by making use of visual analytics.

Filter methods of feature selection such as the correlation matrix has been implemented after the best features are selected using random forests in recent work. In this project however, filter methods of feature selection are implemented initially in order to remove features that offer no predictive value, thereby improving the overall computational efficiency.

4 APPROACH

This section describes the methodology that was used to answer each of the five research questions in detail.

4.1 Data description

The data used in the project is a reviewed dataset by the popular competitive machine learning website – Kaggle, about the loans issued by the peer to peer lending institution – LendingClub from the year 2007 to 2015 [18]. The dataset consists of 74 features/attributes and 887,379 rows and is 104 megabytes in size.

The attributes present in the dataset can be broadly classified into four categories. The first category of attributes is the **loan identifiers** which can uniquely identify each loan. The second category of attributes is the **monetary attributes** which provide information about the monetary value of the loan.

The third category of attributes is the **borrower information** which provide information about the borrower. The fourth and final category of attributes are the **miscellaneous attributes.** These attributes give no specific information about the loan or the borrower but instead is used to verify if the data about the loan was taken from the official LendingClub website or not.

4.2 Data cleaning

Preparing the data for building predictive models is a vital step in the process of making the dataset as computationally efficient as possible through dimensionality reduction and imputing missing data. This section describes the approach taken in cleaning the data.

4.2.1 Importing the data

The data is imported into a Jupyter Notebook by using the *read_csv* module of the Pandas package [19] in python.

The first five rows of the data along with information about the features and the data types of each feature is viewed by using the *head()* and *info()* functions respectively. This provides a high level overview of what the data looks like and what the various features are.

4.2.2 Dropping redundant features

Redundant features are those features that provide no specific value to the predictive model. In this sub-section, a brief description of why each of these redundant attributes/features were dropped is provided. The first feature to be dropped is the **id.** Each loan is uniquely identified by the 'id'. Since each loan is uniquely identified no particular patterns emerge that are useful to the predictive model.

The second feature to be dropped is the **member_id.** Each loan is also uniquely identified by the 'member_id' feature. Same as with the 'id', this feature provides no predictive value as it is an identifier. The third feature to be dropped is the **url.** The 'url' attribute provides information about the URL link from which the loan data about each borrower was taken from. All of the URLs point to the LendingClub's official website. This attribute only serves as a means to verify if the data was accurate or not and hence is a good way to verify if the data was sourced from a reliable source.

The fourth feature to be dropped is the **desc.** The 'desc' attribute provides a brief description about the reason for borrowing the loan according to the borrower. Since there is another attribute called 'purpose' which gives the purpose of the loan in a categorical manner this attribute can be discarded. The fifth feature to be dropped is the **zip_code.** Since the data is anonymized, the 'zip_code' attribute which provides information about the zip code of the borrower is hidden from the public. This attribute cannot be used as a result and is hence discarded.

The sixth feature to be dropped is the **loan_status.** The loan status provides information about the status of the loan. Since this information has been encoded as the target attribute called – 'default', this feature is now redundant. The seventh feature to be dropped is the index**.** The index gives each row a unique number starting from 0. Since this feature is just a unique identifier for the number of rows it provides no value to the predictive model.

The seventh feature to be dropped is the **title.** The 'title' is a feature that provides a unique title for the reason why each loan was borrowed. Since there is already a feature called 'purpose' that describes the reason as to why each loan was borrowed, this feature now becomes a duplicated one.

The final feature to be dropped is the **emp_title.** The 'emp_title' feature provides over fifty unique titles for the employment status of each of the borrowers. Similar titles have different income groups based on the company, years of experience and industry sector and cannot inherently be grouped. A better feature that can be utilized to predict 'defaults' is the annual income of each borrower. Thus the employment title is a redundant feature that only provides textual information without numerical value to the predictive model.

4.2.3 Handling attributes with missing values

Missing values represent a loss of information either by human error or simply because the data was not available at that time or for that example. Attributes with missing values have to be handled carefully in order to avoid biases introduced into the predictive model based on the type of imputation strategy used.

Imputation of missing values is necessary, since one of the constraints of implementing predictive models with the scikit-learn package in python is that it cannot handle missing values [20].

In this project, the first step taken is to identify the attributes with missing values along with the total number and percentage of missing values for each attribute. This information is then transformed neatly into a table which is provided below:

Attribute	Number missing	Percent missing
dti_joint	886870	99.94264006698378
annual_inc_joint	886868	99.94241468414286
verification_status_joint	886868	99.94241468414286
il_util	868762	97.90202382522011
mths_since_rcnt_il	866569	97.65489154014237
total_cu_tl	866007	97.59155896184156
inq_fi	866007	97.59155896184156
all_util	866007	97.59155896184156
max_bal_bc	866007	97.59155896184156
open_rv_24m	866007	97.59155896184156
open_rv_12m	866007	97.59155896184156
total_bal_il	866007	97.59155896184156
open_il_24m	866007	97.59155896184156
open_il_12m	866007	97.59155896184156
open_il_6m	866007	97.59155896184156
open_acc_6m	866007	97.59155896184156
inq_last_12m	866007	97.59155896184156
mths_since_last_record	750326	84.55530275113566
mths_since_last_major_derog	665676	75.01597400885078
mths_since_last_delinq	454312	51.19706461387975
next_pymnt_d	252971	28.50766132622025
total_rev_hi_lim	70276	7.919502264534094
tot_coll_amt	70276	7.919502264534094
tot_cur_bal	70276	7.919502264534094
emp_title	51462	5.799325879922783
emp_length	44825	5.051392922302647
last_pymnt_d	17659	1.9900177939752912
revol_util	502	0.05657109307297107
title	152	0.017129095910541042
collections_12_mths_ex_med	145	0.016340255967292442
last_credit_pull_d	53	0.005972645284596547
total_acc	29	0.0032680511934584885
open_acc	29	0.0032680511934584885
acc_now_delinq	29	0.0032680511934584885
inq_last_6mths	29	0.0032680511934584885
earliest_cr_line	29	0.0032680511934584885
delinq_2yrs	29	0.0032680511934584885
pub_rec	29	0.0032680511934584885
annual_inc	4	0.00045076568185634325

Table 1: Percentage and number of missing values for each attribute

The second step is to remove attributes that have a large percentage of their values that are missing. Attributes that have over 50% of their values missing will introduce a bias into the predictive model when imputing values since a majority of the data imputed into such attributes are artificial and not real. Therefore, a plot is constructed between the percentage of missing values along the x-axis and the number of attributes that will be dropped as a result along the y-axis. This plot is illustrated below:

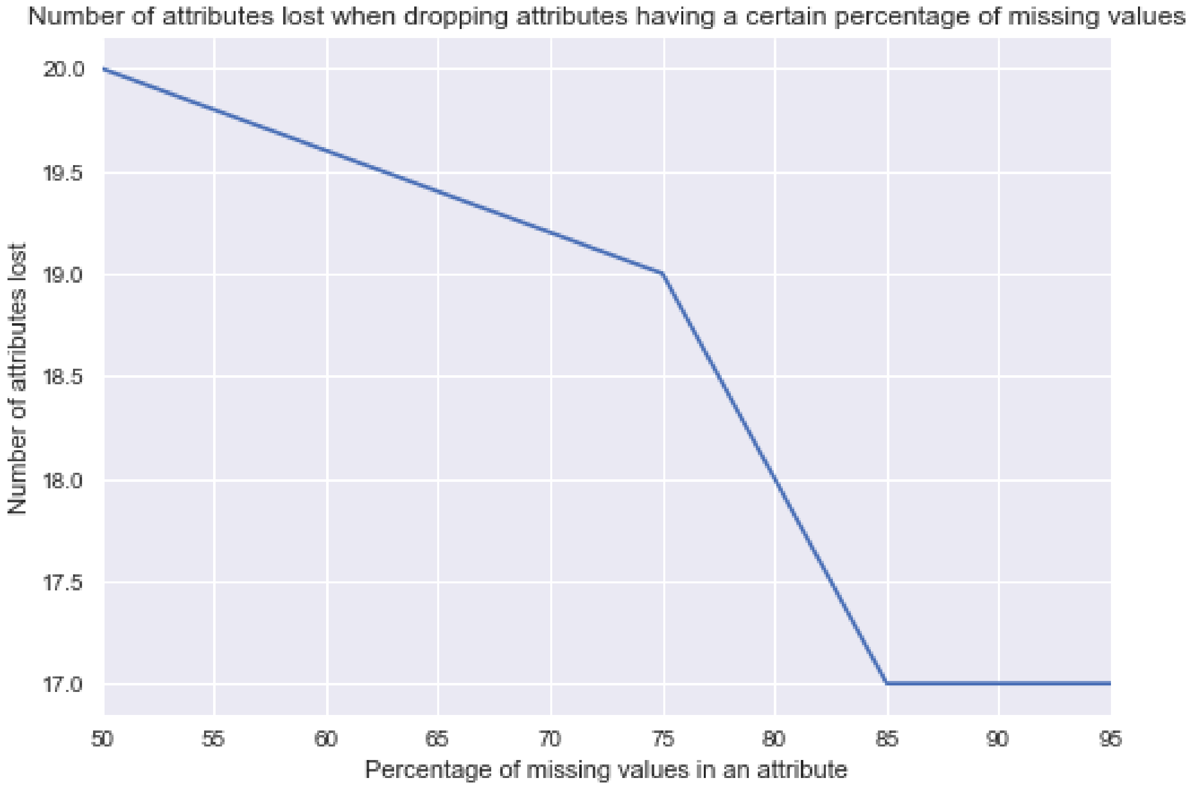

Figure 1: Number of Attributes lost versus percentage of missing values in an attribute

From the plot illustrated above it is clear that, dropping attributes that have over 85% of their values missing results in the loss of 17 attributes. Dropping attributes that have over 75% of their values missing results in the loss of 19 attributes. Dropping attributes that have over 50% of their values missing results in the loss of 20 attributes.

As a result, dropping attributes that have over 75% of their values is an optimal choice as one additional variable which is, ‘the months since last delinquency’ is retained. The months since last delinquency provides information about the number of months since it has been since the borrower’s last delinquency. This is an important predictor of defaults intuitively.

Broadly speaking, missing values in an attribute can be categorized into - missing completely at random, missing at random and not missing at random [21].

All of the attributes in this data have values missing completely at random with no real or prior knowledge as to why they are missing. Therefore appropriate imputation strategies are implemented for features that have data missing completely at random. The first set of attributes with values that are missing completely at random are identified. Some of the strategies that are used to impute the missing values for such attributes are the deletion of the missing values from the attribute, imputation techniques such as mean and/or median imputation and imputation using common sense/domain knowledge [21].

Application of the imputation strategies for the attributes in this project involved the imputation of the majority value in an attribute. Attributes that have a value that occurs for the majority of the values in that attribute can be used to impute the missing values if the percentage of missing values is small.

Imputing missing values with the mean if there is no majority value and if the distribution of values in that attribute is normal. This approach is used because when the distribution of data is normal in nature the majority of the values occur at the mean of the distribution.

Finally, imputing random values if there is no majority values and if the distribution of values is not normal in nature provided that only a small percentage of the data is missing.

4.2.4 Features with dates

Features with dates do not inherently make sense to a predictive model unless they represent a numeric value. Therefore, the features with dates are converted into a number which represent the number of years from the present year which is 2018. The **earliest credit line** is converted into an integer which represents the number of years the earliest credit line was opened for the borrower from the present year.

The **issue date** is converted into an integer which represents the number of years the loan was issued to the borrower from the present year. The **last payment date** is converted into an integer which represents the number of years it has been since the borrower last made a payment from the present year and the **last credit pull date** is converted into an integer which represents the number of years it has been since the borrower last pulled credit from their credit lines from the present year.

4.2.5 Creating the target attribute

The problem here is that of a binary classification in which the objective is to predict if a loan is going to default or not. Therefore, the target attribute called ‘default’ is created which has two values – 1 if a loan was defaulted and 0 if a loan was fully repaid.

The ‘loan_status’ attribute provides the different categories of loan statuses. This attribute has ten categories – Current, Fully Paid, Charged Off, Late (31 – 120) days, Issued, In Grace Period, Late (16 – 30) days, Does not meet credit policy, status: Fully Paid, Does not meet credit policy, status: Charged off and Default.

The ‘Fully Paid’ and the ‘Does not meet credit policy, status: Fully Paid’ categories are the loans that have been fully paid and are grouped together and is given a value of 0.

The ‘Charged Off’ loans are the loans that have gone past the stage of default and as such are written off by LendingClub. Such loans are considered to have been defaulted.

Therefore, the ‘Charged Off’, ‘Default’ and ‘Does not meet credit policy, status: Charged Off’ are grouped together as the defaulted loans and is given a value of 1.

4.3 Behavior of defaulted loans

This section details the approach that was taken in order to answer the first research question which aims to identify how the behavior of defaulted loans differ from the behavior of the loans that have been fully repaid.

The percentage of loans that have defaulted and the percentage of loans that have been fully paid are identified by using the bar plot from the Seaborn package in Python which is built on top of Matplotlib [22]. This is plotted in order to provide the investor with a basic understanding of the percentage of loans from LendingClub that are likely to default.

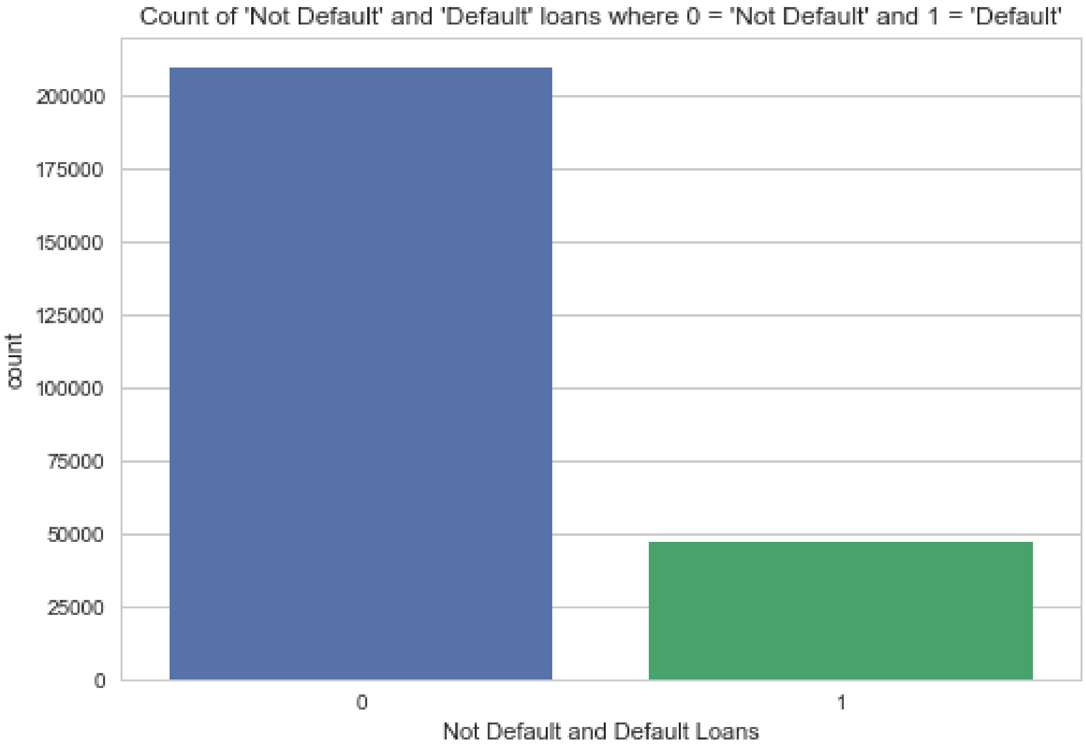

Figure 2: Counts of the defaulted and fully repaid loans

From the plot generated above it is clear that 18% of the loans in the dataset have defaulted. And 72% of the loans in the dataset have been fully repaid. Next, the different attributes are grouped into 4 characteristics – loan characteristics, borrower characteristics, credit history and borrower indebtedness and are analysed. The attributes under the borrower assessment and loan characteristics are grouped into one characteristic called the loan characteristics since attributes such as the loan grade, subgrade and interest rate are inherently characteristics of the loan itself.

4.3.1 Distinguishing the behaviour of defaulted loans for continuous and nominal attributes

In this sub-section the behaviour of the defaulted loans are compared and contrasted with the behaviour of the loans that have been fully repaid by employing different visual analytical techniques for continuous and nominal attributes. The attributes under the loan characteristics are – interest rate, loan purpose, loan amount, loan grade and loan subgrade.

The loan grade and subgrade are analysed in depth in research question two, therefore the loan characteristics analysed in the first research question are interest rate, loan purpose and loan amount.

For the continuous variables such as the interest rate and loan amount, the distribution data in each variable is first visualized by using a density plot from the Seaborn package in python [22]. Density plots are a useful way to analyse the density or concentration of data points over a particular range of values along the horizontal axis. An example of the density plot used to visualise the distribution of interest rates is illustrated below:

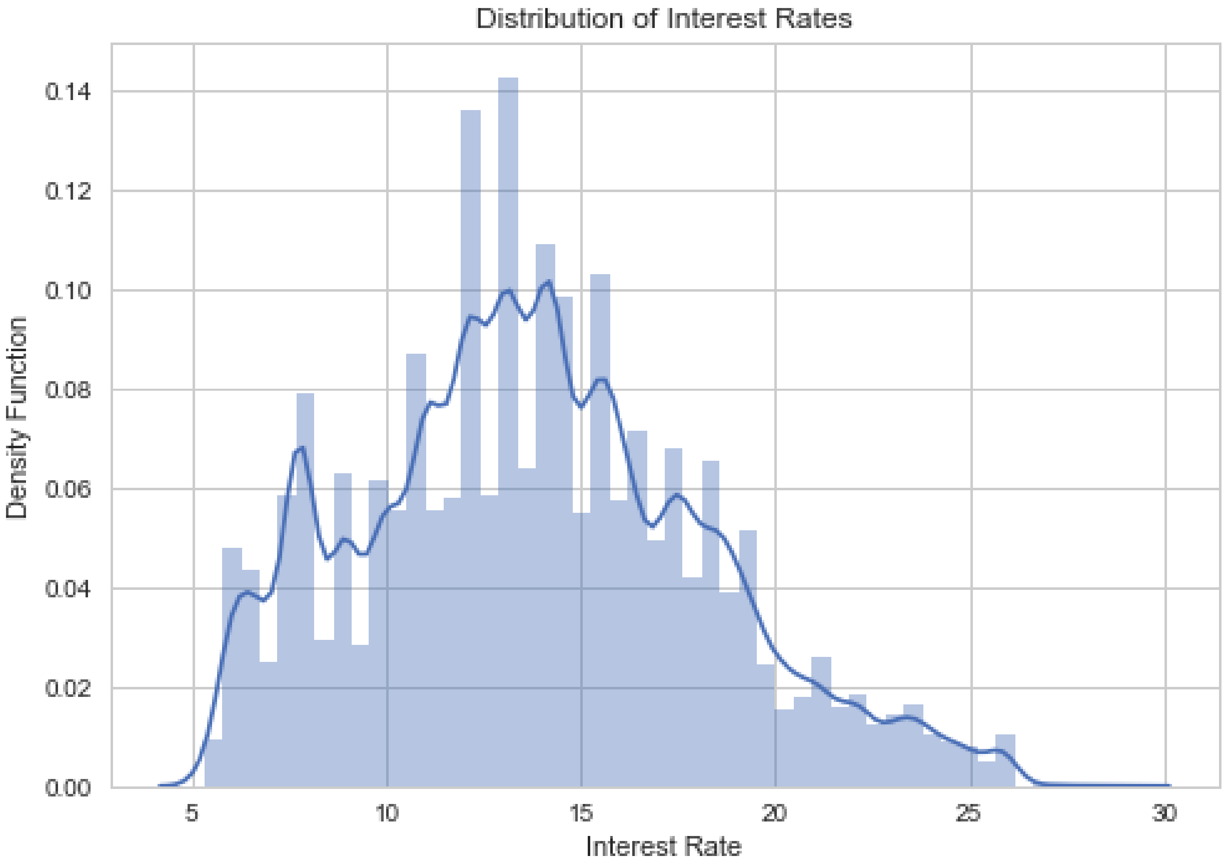

Figure 3: Distribution of interest rates

The figure above provides the investor with information about how the interest rates are distributed for both the defaulted and fully paid loans together so that when separate density plots are plotted for the defaulted and fully paid loans it becomes easier to differentiate the behaviour of the defaulted loans from the distribution of the entire data.

Overlapping the density plots of the defaulted loans and the fully repaid loans facilitates the comparison of how the behaviour of the defaulted loans differ from the loans that are fully repaid with respect to the distribution of the data.

In order to analyse the nominal variables a different approach was used. Here, horizontal bar plots [22] were used to analyse the counts of the different categories under each attribute for both the defaulted and fully repaid loans together. This is illustrated in the image below for the home ownership status of the borrower:

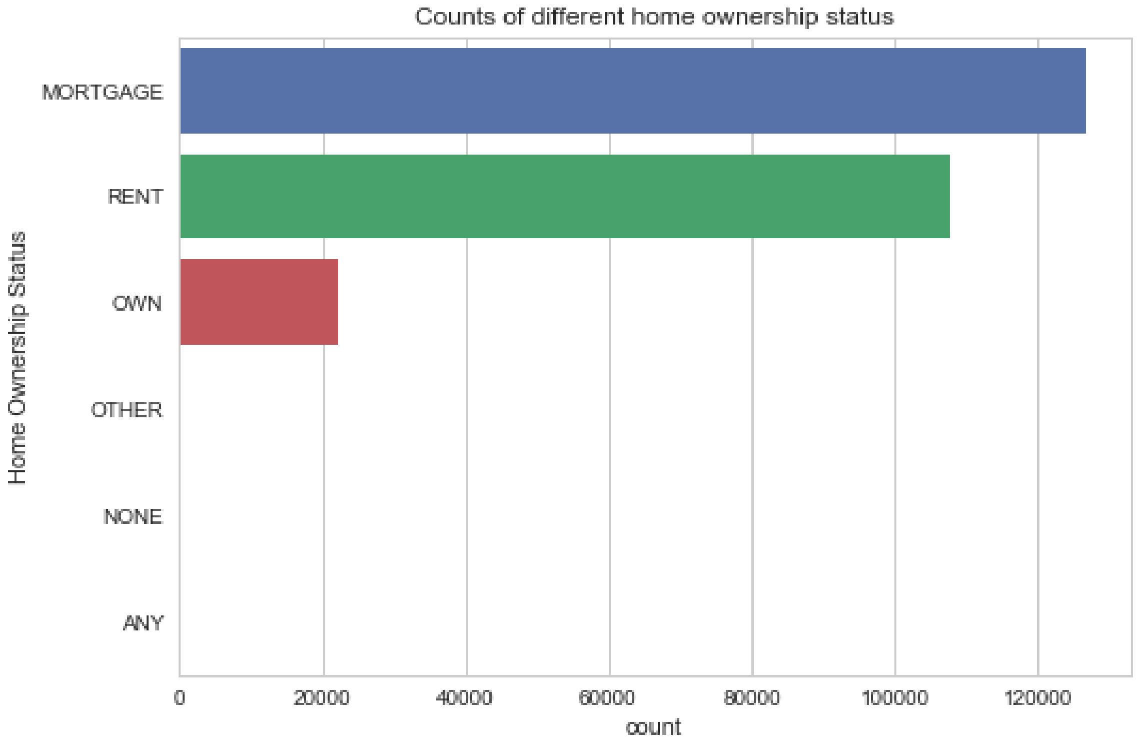

Figure 4: Counts of the different home ownership status

This provides the investor with information about how the different categories of an attribute are distributed in terms of a general count for both the defaulted and fully repaid loans together. Using this, two separate bar plots are plotted for the defaulted and fully repaid loans which enables the investor to compare and contrast the behaviour of loans.

An additional step taken to further characterise the behaviour of the fully repaid and defaulted loans is the relative proportion of the counts of each category. For the home ownership status a category such as 'RENT' for the defaulted loans is first counted and then divided by the total number of loans within the defaulted pool of loans and is then multiple by 100 in order to obtain a percentage that indicates the percentage of a particular category under the defaulted loans. This is done for the fully repaid loans as well. This provides the investor with a relative percentage for each category under a nominal attribute that can help them clearly distinguish how the behaviour of defaulted loans differ from that of the fully repaid loans.

4.4 Best performing loan grade

The aim of this research question is to identify the grade of loans that performed the best, i.e. had the highest number of fully paid loans and the potential reasons as to why it did perform well.

The loan grade is the information that LendingClub provides to potential investors when they decide to invest in loans. Loans with the grade A are considered to be the safest while loans with the grade G present the highest levels of risk. Therefore, by using visual analytics this research question will provide information to the investor that will accurately inform them about the best and worst performing grades.

The grade and subgrade of the loan are nominal attributes. Therefore, a horizontal bar plot is used to identify the counts of the different grades and subgrades of loans in order to identify the most popular and the least popular grades and subgrades of loans.

Next, the distribution of loan grades are identified for the defaulted and fully paid loans using two different horizontal bar plots along with the relative percentage of each grade in the defaulted and fully paid loans. This helps the investor compare and contrast the number and relative percentage of grades in the defaulted and fully paid pool of loans.

Once the grades with the highest number of defaulted and fully paid loans are identified, these grades are analysed in detail with respect to the loan and borrower characteristics in order to understand why these grades were the worst and best performing grades respectively.

4.5 Defaulted loans and geographic location of loan origination

The third research question aims to understand if the geographic location of origination of the loans affect the probability of defaults in any way or not. Performing a geographic analysis of the loans provides the investor with information about the locations that they should avoid investing from as they might be associated with a higher proportion of defaults. A horizontal bar plot is used to identify the states in the United States of America that have borrowed the most number of loans. This is illustrated in the image below:

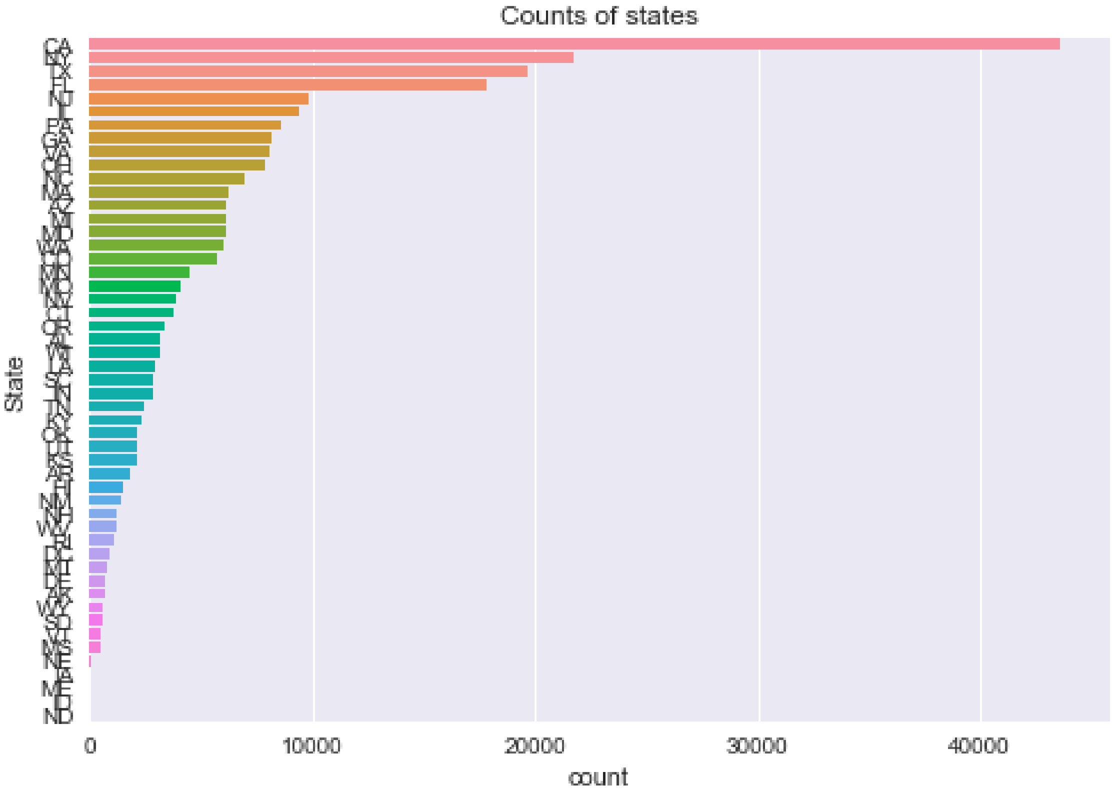

Figure 5: Number of loans across different states

A quick look at the plot above indicates that most of the loans originate from California, New York, Texas and New Jersey. Next, the software Tableau is used to visualize the number of defaults in each state across a geographic map of the United States. Such a plot makes it easier to interpret the results of the visual analysis in the form of a heat map with the colour red indicating a large number of defaults while blue indicating a small number of defaults. This plot is illustrated in the image below:

Total number of defaults in each state

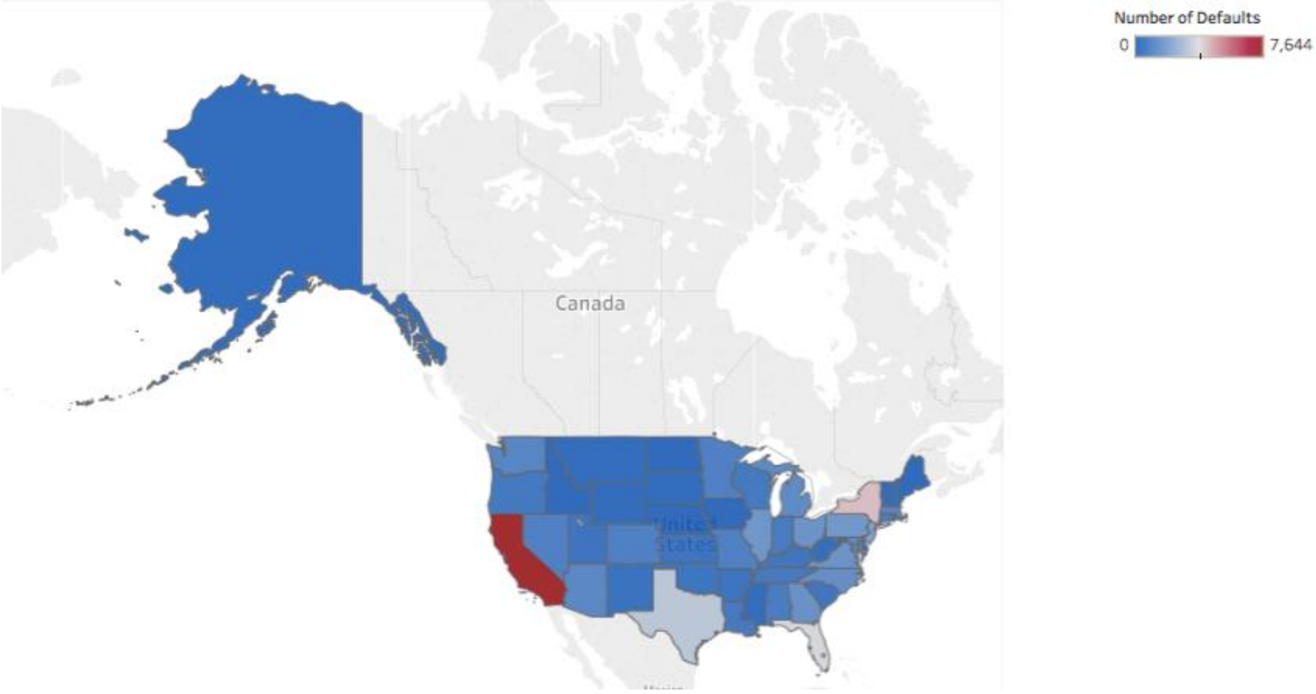

Figure 6: Total number of defaults in each state

A closer look at this plot indicates that the results are biased because from the bar plot constructed earlier, it was clear that California has the highest number of loans and in turn the highest number of defaults.

Therefore, a plot consisting of the proportion of defaults in each state is constructed by dividing the number of defaults in each state with the total number of loans in that particular state alone. The following equation is used to calculate the proportion of defaults in each state:

$$Proportion\ of\ defaults\ in\ a\ state = \frac{Number\ of\ defaults\ in\ that\ state}{Total\ number\ of\ loans\ issued\ in\ that\ state}$$

Equation 9: Formula for calculating the proportion of defaults in each state

4.6 Identifying the best predictors of default

The fundamental aim of this research question is to identify the features/attributes that best predict if a loan would default or not.

Filter methods of feature selection are first used to drop features that provided no predictive value. Constant features and quasi-constant features that had only one or two values across the entire feature with almost zero variance provide no predictive value to a predictive model and as such are filtered out using the variance threshold function from scikit-learn. Nominal features with only one majority category that occurred for over 99% of the values in the feature were manually identified and dropped.

Duplicated features are those pair of features in which each observation between the two features are identical in every possible way. Thus, the second feature in the pair of features was dropped which provides value in terms of dimensionality reduction and removing redundancy. The features that are highly correlated with each other are visualized by using the correlation matrix from Seaborn [22]. This correlation matrix is illustrated below:

Correlation Matrix

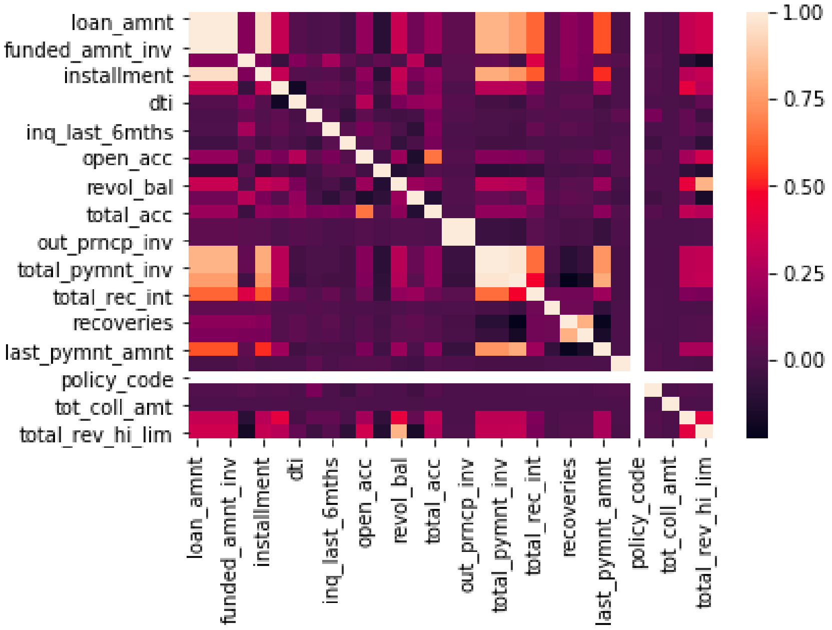

Figure 7: Correlation matrix of all the features

In the correlation matrix above, the main diagonal is light red/white which indicates a high correlation as they have a value of -1 of the Pearson's correlation coefficient. This is because along the diagonal each variable/attribute is correlated with the same variable itself in which case they have a high value of correlation.

There are however, variables outside the main diagonal that are correlated with other variables and have a very high negative value of Pearson's correlation coefficient. Therefore, features that are highly correlated with another variable is filtered and dropped from the data. Next, a decision tree is fit for each of the remaining features in the dataset and the area under the curve score is extracted for each feature. This univariate analysis returns a value between 0 and 1 for each variable. A score of 1 indicates that the feature is highly influential in predicting defaults while a score that is lesser than 0.5 indicates that the feature performs worse than random guessing when it comes to predicting defaults. The score obtained for each feature is plotted using a vertical bar plot as illustrated below:

Univariate AUC scores for each attribute

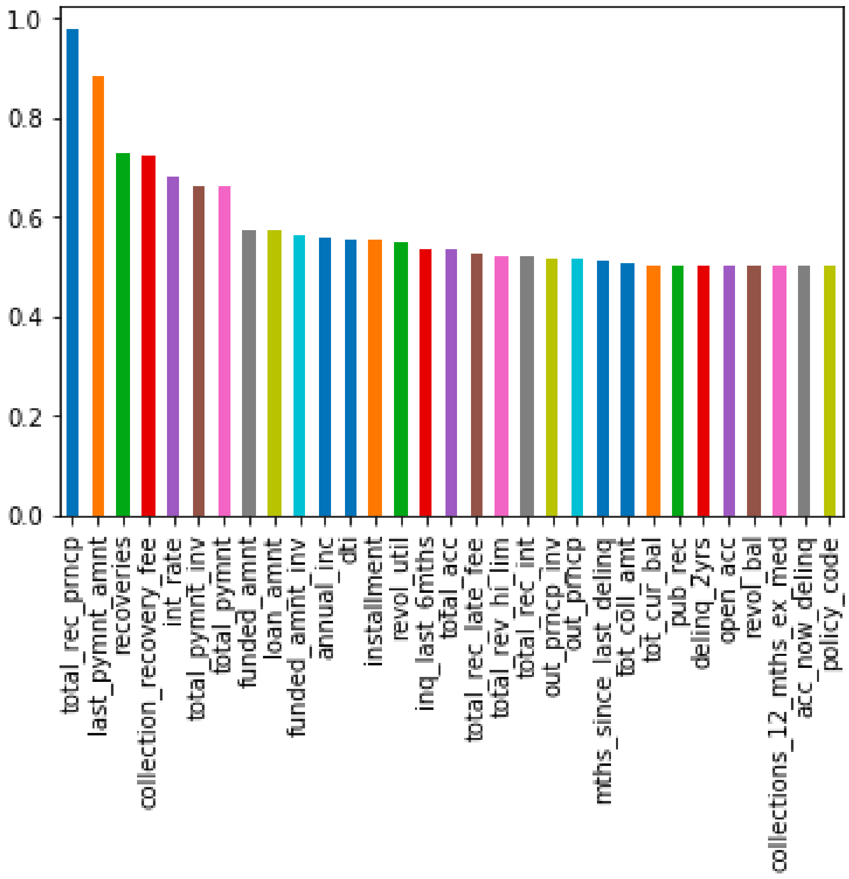

Figure 8: AUC scores for each variable

Features that have an univariate score lesser than 0.5 are dropped from the dataset as they perform worse than random guessing.

One of the main drawbacks about building predictive models with scikit-learn is that the categorical variables have to be encoded into a numeric value. Therefore the ordinal variables, the variables that have a natural order to them are integer encoded. Features such as the Loan Grade and the Loan Subgrade have an order to them as they go in alphabetical order from A to G and LendingClub associates grade A loans as the safest while grade G loans are the most risky in terms of investment. These ordinal variables are given a value from 1 to the number of categories present in that variable, for example a loan with the grade A is given a value of 1 and another loan with the grade B is given a value of 2. Nominal variables do not have an inherent order to them and therefore cannot be integer encoded. These variables are one-hot encoded instead. This means that if a variable has three categories, each of the three categories is converted into a new feature. If the category is present in the feature it is given a value of 1 otherwise if the category is absent in the feature it is given a value of 0.

Finally, the best features are selected recursively using recursive feature elimination with the random forests classifier to determine the feature importance [20].

The recursive feature elimination, eliminates features that have a low value of importance with respect to the random forest algorithm and then iterates over the remaining features in a process that continues until the best features are obtained.

Implementing the recursive feature elimination in order to find to the top 10 to the top 100 features in steps of ten and evaluating how accurate each set of features are by implementing a K-Nearest Neighbors classifier, results in a plot that provides information about how many features to use when building the predictive models.

The reason for choosing the K-Nearest Neighbors in order to evaluate the best number of features to use as input to the predictive model is because the random forest classifier implemented in the recursive feature elimination provides a set of features that are the best predictors for the Random Forest model alone. In order to avoid any kind of bias when picking the best number of features the area under the curve score is extracted from the K-Nearest Neighbors classifier. This results in a plot as illustrated below:

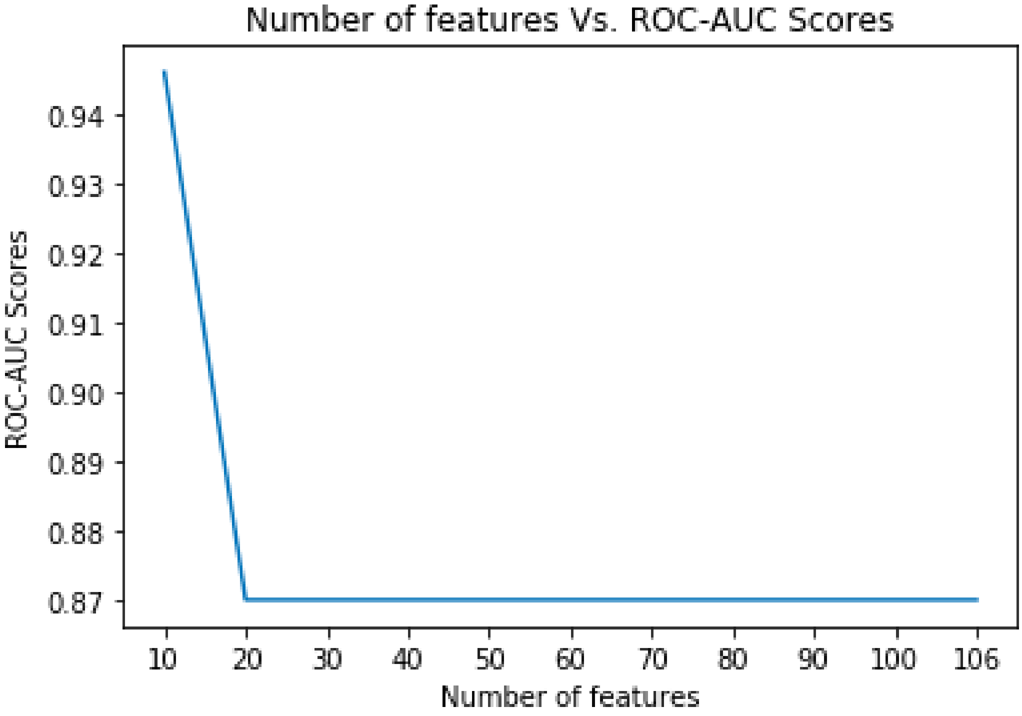

Figure 9: AUC scores for different number of features

From the plot above between the AUC scores and the number of features it is clear that building predictive models with the top 10 features results in the best prediction performance as it has an AUC score of 0.94 which is the highest.

4.7 Identifying the best machine learning model to predict defaults

There are a wide range of machine learning algorithms available in order to build predictive models. However, when it comes to the financial and lending industry models that are highly interpretable work the best as the data is sensitive in nature. Financial regulators and investors alike would like to understand how the models are producing results for the purpose of transparency. Applying domain expertise along with the models will also prevent over-reliance on models alone if the models are interpretable. Therefore, eight highly interpretable machine learning algorithms are used in order to compare and evaluate the model that provides the best results when it comes to predicting defaults.

The eight machine learning algorithms used in this project are – K-Nearest Neighbors, Logistic Regression, Linear Support Vector Machines, Decision Tree (CART), Random Forests, Gradient Boosted Trees, Gaussian Naive Bayes Classifier and the Ensemble model.

The data after feature selection consists of 10 attributes and 1 target attribute. This data is then separated into training and test sets [20]. The machine learning algorithms are trained on 70% of the data and the output of these algorithms are evaluated on the remaining 30% of the data that the machine learning algorithm has not seen yet. This ensures that the results are not biased in any way.

Each of the eight algorithms are fit to the training data and the accuracy scores are evaluated on the test sets [20]. This provides an initial representation of how the base classifier performs without any kind of fine tuning or optimization.

4.7.1 K-Nearest Neighbors

After the base classifier is built, the hyper-parameters of the algorithm are fine tuned in order to extract the value of the number of neighbors that produces the best performing classifier. This is done by using the GridSearchCV algorithm [20]. A grid of possible values of the number of neighbors is constructed. This grid contains values from 1 to 24. The algorithm then fits each value present in the grid to the K-NN classifier and evaluates the accuracy score. Additionally, GridSearchCV uses cross validation in order to evaluate the accuracy scores. This means that entire dataset is used for training and testing, instead of a single split. This ensures additional reliability on the results that GridSearchCV produces.

In order to increase reliability, a plot is created between the number of neighbors along the x-axis and the training/testing accuracy along the y-axis. This plot can be used to help pick hyper-parameters in combination with the GridSearchCV algorithm. An illustration of this plot that was implemented in the dataset to identify the most optimal number of neighbors is illustrated below:

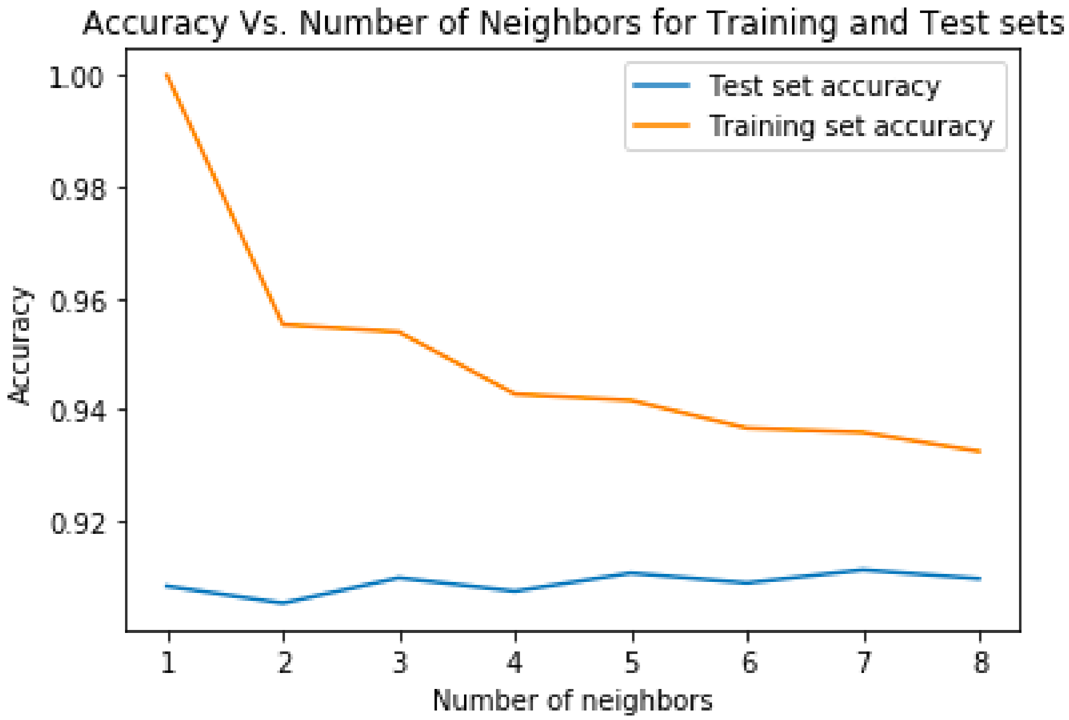

Figure 10: Chart displaying accuracy for training and test sets with respect to the number of neighbors

From the plot above it is clear that when 1 neighbor is used to construct a K-NN model the model is highly overfit. This means that the model performs extremely well on the training data but poorly on the test data. However, when the number of neighbors is 5 the model performs considerably better on both the training and test sets. The GridSearchCV also produced 5 as the optimal number of neighbors. Thus using this plot in combination with the GridSearchCV algorithm provides a robust way to perform hyper-parameter optimization. After the model was retrained with the most optimal value of the nearest neighbor, the accuracy of the model improved as a result.

Finally, the model was retrained on a dataset that was scaled using the *StandardScaler()* method from Scikit-learn [20]. Scaling the data implies subtracting each value in an attribute with the mean of entire attribute and dividing each value by the variance of the entire attribute. This ensures that the data is standardized, making pattern detection for the predictive model much more computationally simpler. Scaling the data saw an improvement to the overall accuracy score of the K-NN model.

4.7.2 Logistic Regression

After building the base classifier for the logistic regression model, the hyper-parameters are optimised by using the GridSearchCV algorithm [20]. The hyper-parameter for the logistic regression model is the inverse regularization strength. After obtaining the most optimal value of the inverse regularization strength, a plot between the classification error and the different values of the inverse regularization strength is plotted to verify if this value is optimal or not. This plot is illustrated below:

Classification error for different values of inverse regularization strength

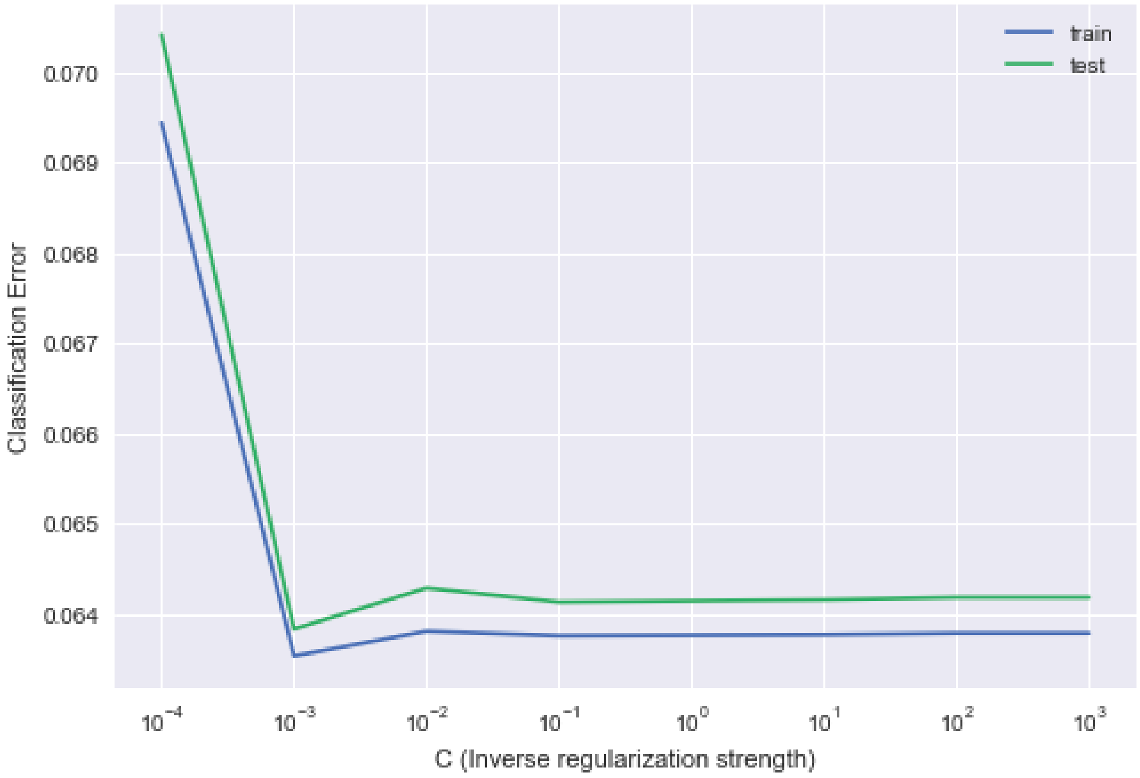

Figure 11: Classification errors for the hyper-parameters of the logistic regression model

From the plot above, the classification errors for the training and test sets is the lowest at an inverse regularization strength of 0.01.

Fitting the new logistic model with the optimal value of the hyper-parameter resulted in an increase in the accuracy score. Scaling the data and fitting the logistic regression model with the optimal hyper-parameter value along with the standardized data also resulted in the increase of the accuracy score.

4.7.3 Linear Support Vector Machines

After building the base classifier, the hyper-parameters are optimised using the GridSearchCV algorithm [20]. In order to verify if the result given by the GridSearchCV is accurate, a plot between the classification errors and the inverse regularization strength is plotted. The resulting plot is illustrated below:

Classification error for different values of inverse regularization strengths

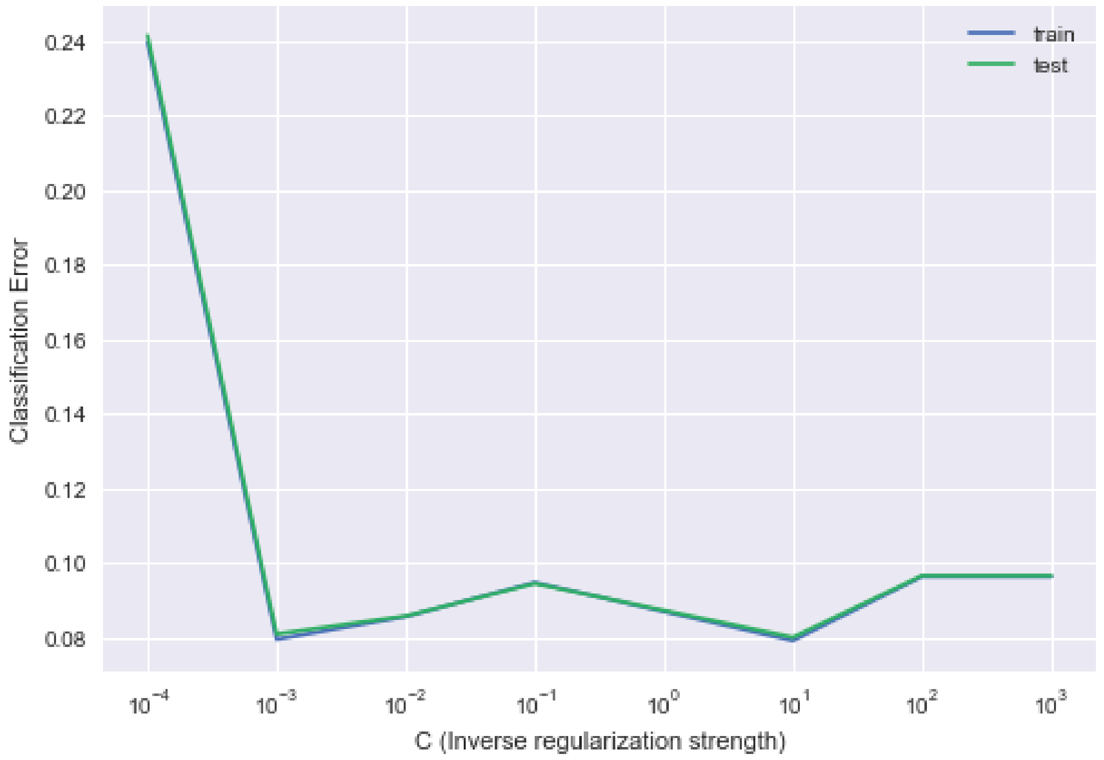

Figure 12: Classification error for the hyper-parameters of the linear SVM model

From the plot above it is clear that a value of 10 for the inverse regularization strength gives the lowest value of the classification error which is the same value given by the

GridSearchCV algorithm. Implementing the new model with the optimal value of the hyper-parameter and scaling the data resulted in an increase in the accuracy score of the Linear Support Vector Machine.

4.7.4 Decision Tree (CART)

The Classification and Regression Tree (CART) decision tree is first fit to the training data and the accuracy score of the base classifier is extracted from the test data. GridSearchCV is used to optimise the hyper parameters of the tree [20]. There are two main hyper-parameters that affect the performance of the tree – Maximum Depth and Minimum samples per leaf. The maximum depth is the how deep the tree is allowed to grow to, while the minimum samples is the minimum number of samples that have to be present in each node. Optimizing the hyperparameters of the tree resulted in the decrease in the accuracy of the classifier. This is because GridSearchCV uses cross-validation to in order to prevent generalization errors caused due to bias and variance. This fundamentally suggests that the base model, with its high accuracy score of 97% was overfit. Hence, the model with the optimal hyper-parameters from GridSearchCV will be used to build the final decision tree classifier.

Tree based classifiers do not require the data to be standardized as they work well even if the data is not scaled as they split data based on individual features and only consider the values present in the particular feature.

4.7.5 Random Forests

The Random Forest classifier is first fit into the training data and the accuracy score of the base classifier is extracted from the test data. The base classifier resulted in a very high accuracy of 98% indicating that the model was probably overfit. Therefore, the GridSearchCV algorithm is used to optimize the hyper-parameters of the tree [20].

Since the Random Forest algorithm is based on building multiple decision trees it uses the maximum depth and minimum samples per leaf along with the total number of estimators as hyper-parameters. The total number of estimators is the total number of decision trees to be constructed. After optimizing the hyper-parameters, the accuracy score of the model decreased to 92.6%.

Since the random forest classifier is a tree based algorithm scaling the data is not implemented.

4.7.6 Gradient Boosted Trees (AdaBoost)

Since the gradient boosted trees are based on the idea of converting a weak learner into a strong learner, the base classifier is built using a decision tree with a maximum depth of 1. The hyper-parameters are optimized using the GridSearchCV algorithm [20]. This resulted in an increase in the accuracy score. Since the gradient boosted trees is a tree based algorithm scaling is not implemented.

4.7.7 Gaussian Naive Bayes Classifier

The base classifier is built on the training data and the accuracy is evaluated on the test data. Hyper-parameter optimization is not implemented as there are no hyper-parameters available to optimize for the Gaussian Naive Bayes Classifier apart from the prior probabilities of default which is not available.

The data is then scaled [20] which resulted in very little improvement in the accuracy score.

4.7.8 Ensemble Model

The base classifier is implemented on the training data and the accuracy is evaluated on the test data. Since there are no hyper-parameters to optimize in the ensemble model, the GridSearchCV algorithm is not implemented. Scaling the data did not improve the accuracy of the ensemble model.

4.7.9 Evaluation of results

In order to evaluate the results of all the classifiers a table is constructed with the accuracy scores of each classifier along with the scores of each classifier plotted using a horizontal bar plot. Next, the accuracy of each model is extracted after performing ten-fold cross-validation on each model. A box plot of the cross validated scores of each model is plotted in order to understand the range of the accuracy scores of each model [23].

Tables of the precision, recall, false positive and negative rates are constructed and the associated bar plots are plotted in order to compare these metrics across all models.

After picking the best model, based on the metrics discussed above the lift chart and cumulative gains chart are plotted for the model in order to provide the investor with further interpretability of how the model performs.

5 RESULTS

This section provides the answers to each of the research questions that this project seeks to answer.

5.1 Behaviour of defaulted loans

The first research question seeks to understand how the behaviour of the defaulted loans differ from the loans that have been fully repaid.

5.1.1 Defaulted loans and interest rates

A higher density of loans enter into default when the interest rates of the loans are greater than 15% while a lower density of loans enter into default when the interest rates of the loans are lesser than 10%.

Interest rates of defaulted and not defaulted loans

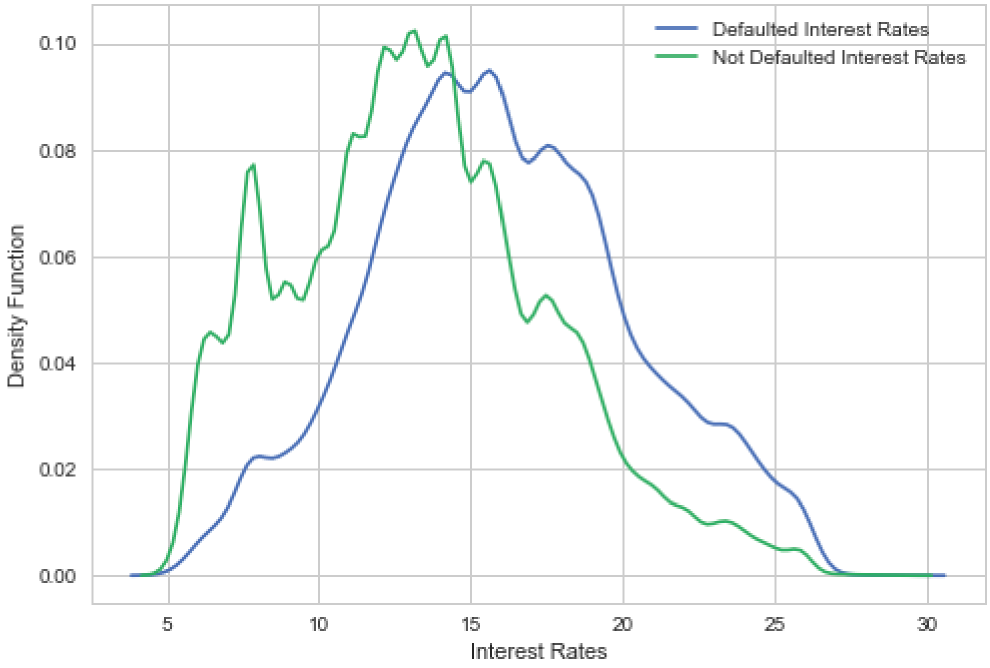

Figure 13: Behaviour of defaulted loans with respect to the interest rates

Higher interest rates are generally associated with loans that are riskier in nature. However, knowledge of the threshold of the interest rates at which the loans tend to default is valuable information as investors can now focus on loans that have interest rates that are slightly lower than 15% for safer returns.

5.1.2 Defaulted loans and home ownership status

Borrowers who rented their homes had a higher chance of their loans entering into default compared to borrowers who had to pay a mortgage for their homes.

	RENT	**MORTGAGE**
DEFAULTED	47%	43%
FULLY REPAID	40%	50%

Table 2: Proportion of home ownership status for the defaulted and fully repaid loans

Borrowers who rent homes might generally be people who have just entered the workforce and therefore have a higher probability of defaulting on loans due to a lack of financial stability.

5.1.3 Defaulted loans and Debt-to-Income Ratio

There is a higher density of loans that have entered into default when the debt-to-income ratio of the borrowers is greater than 15%. This is illustrated in the plot shown below:

DTI ratio of the defaulted and fully repaid loans

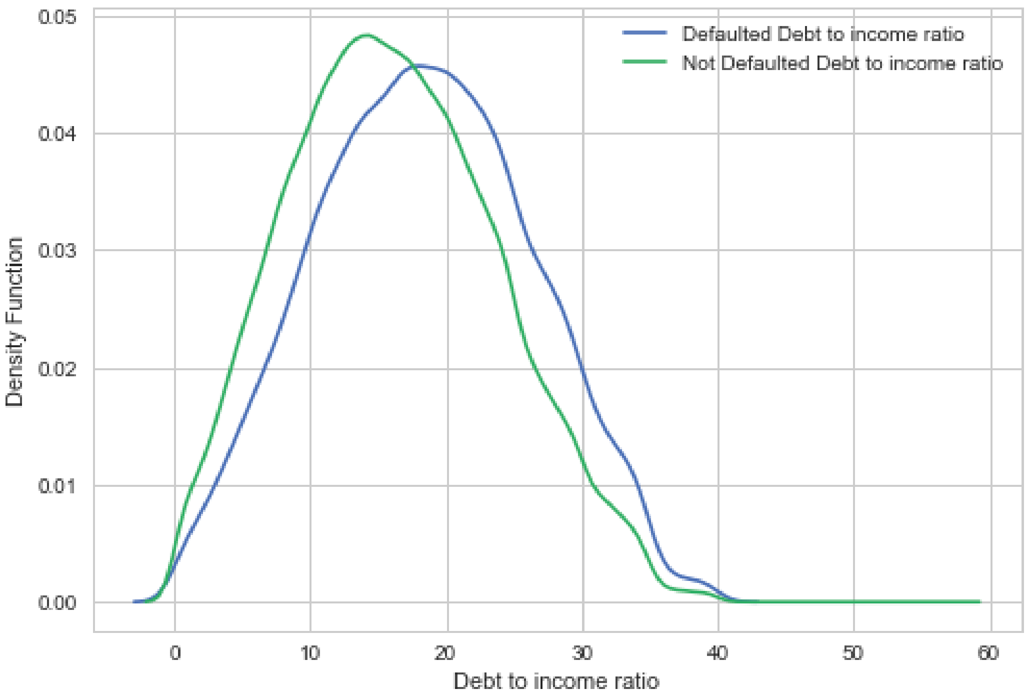

Figure 14: Behaviour of defaulted loans with respect to the DTI ratio

Higher debt-to-income ratios signify that the borrower has a higher monthly debt payment to make compared to the income that they are bringing in every month. The threshold of 15% is valuable information to the investor as they can avoid investing into loans when the borrower has a debt-to-income ratio greater than 15%.

5.1.4 Defaulted loans and revolving utilization rates

Borrowers who had revolving utilization rates between 50 to 100% had a higher density of loans that defaulted. This is illustrated in the plot shown below:

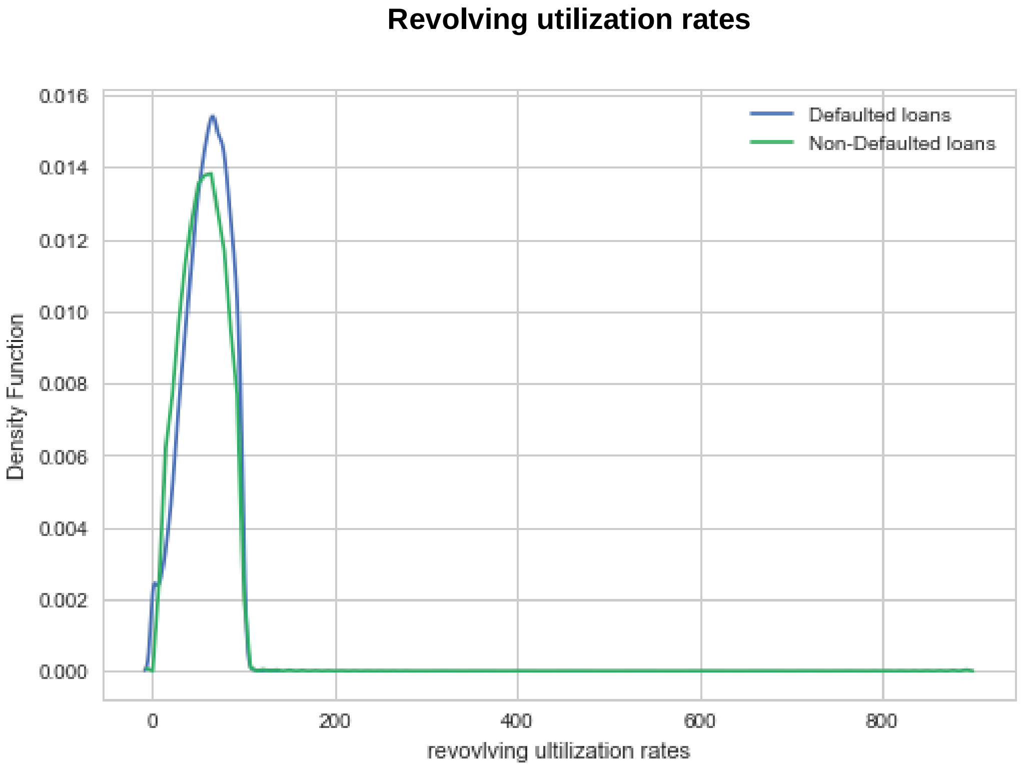

Figure 15: Behaviour of defaulted loans with respect to the revolving utilization rates

A higher revolving utilization rate indicates that the borrower is using a higher amount credit relative to all available credit lines and hence is exhausting their supply of credit fast resulting in a higher density of loans defaulting.

5.2 Best performing loan grade

The second research questions aims to identify the grade of loan that performed the best by having the highest number of fully repaid loans and the potential reasons for the good performance.

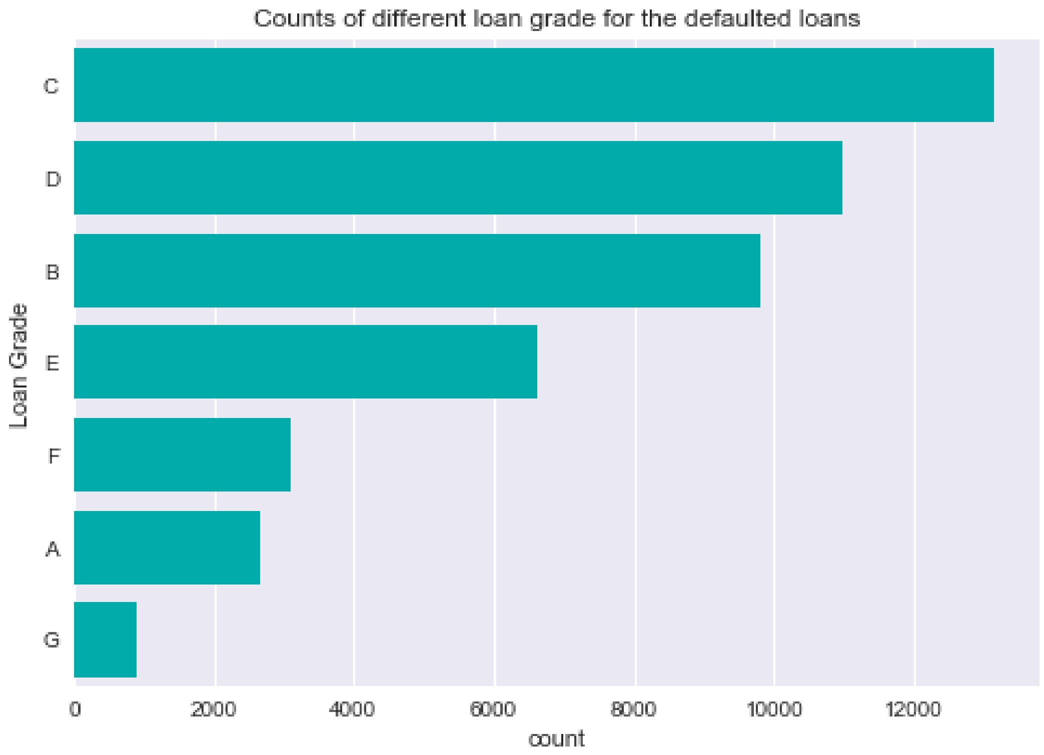

Figure 16: Loan grade counts for defaulted loans

A quick glance at the figure indicates the loans with the grade – C had the highest number of defaults. The proportion of each loan grade within the defaulted loans is also calculated and is summarised in the table below:

Loan Grade	Proportion (%)
C	27.8
D	23.2
B	20.7
E	14.0
F	6.5
A	5.6
G	1.8

Table 3: Proportion of each grade within the defaulted loans

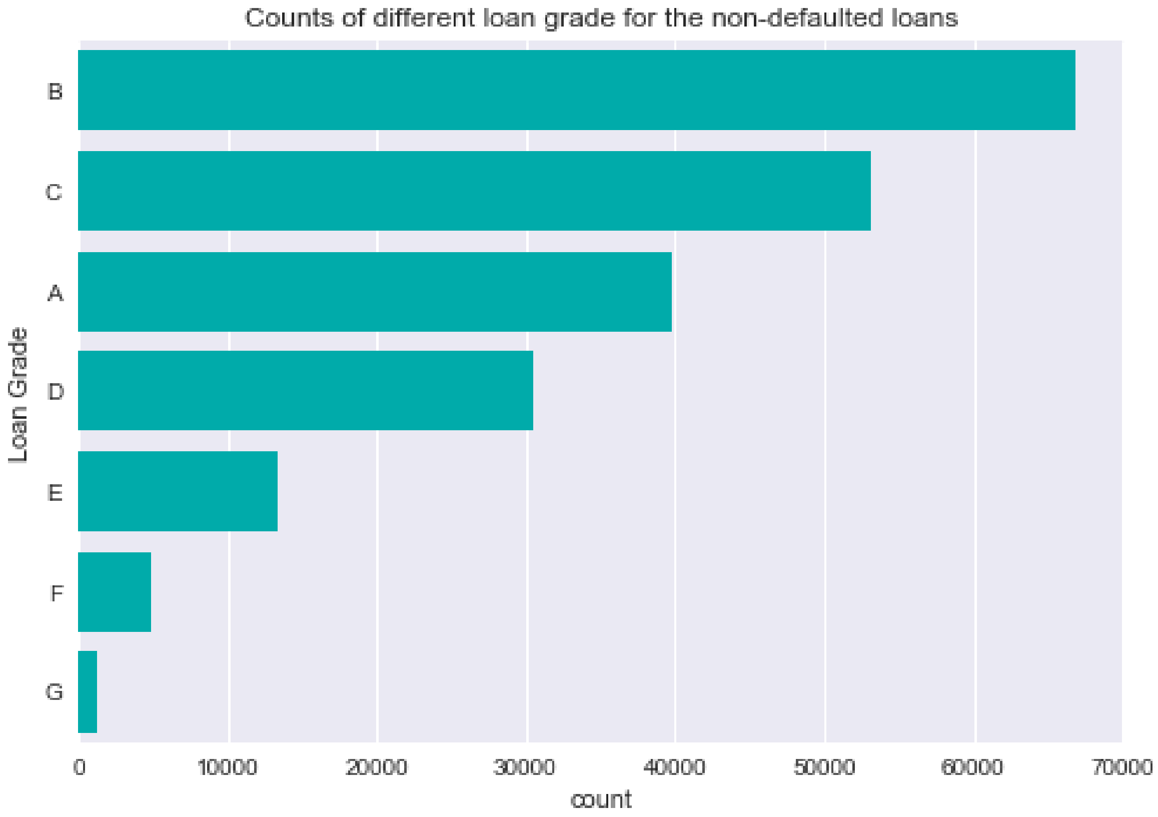

Figure 17: Loan grade counts for fully repaid loans

A quick glance at the figure indicates the loans with the grade – B has the highest number of fully paid loans. The proportion of each loan grade within the fully repaid loans is also calculated and is summarised in the table below:

Loan Grade	**Proportion (%)**
B	31.8
C	25.3
A	18.9
D	14.5
E	6.3
F	2.3
G	0.6

Table 4: Proportion of each grade within the fully repaid loans

The tables indicate that the proportion of loans with the grade – B is lower among the pool of defaulted loans, while it is the highest among the pool of fully repaid loans making it the best performing loan grade.

Loan grade – C on the other has the highest proportion among the defaulted loans hence making it the worst performing loan grade.

In order to explore the potential reasons for the success of grade – B and the bad performance of grade – C further analysis is done.

5.2.1 Loan grades and interest rates

The first research question indicated that loans with interest rates greater than 15% was associated with a higher density of defaults. Comparing the density plots of the interest rates of loan grades B and C resulted in the plot shown below:

Interest rates of grades B and C

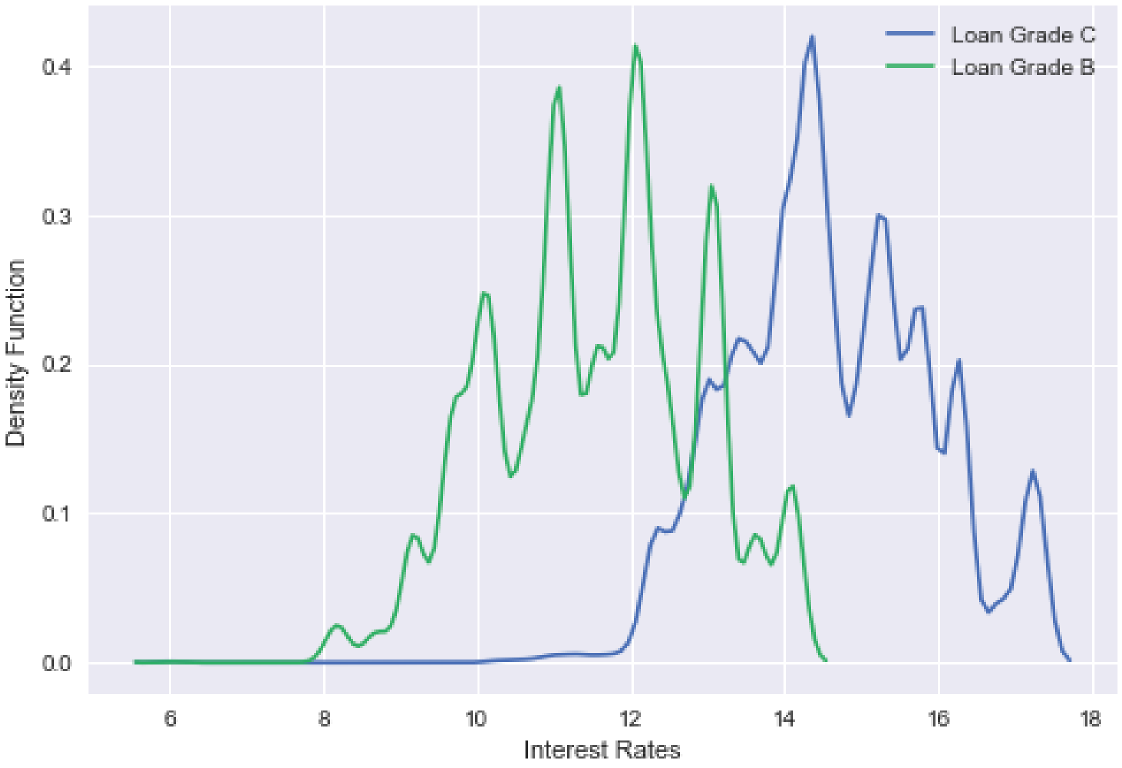

Table 5: Density plot of interest rates of different loan grades

From the plot above it is clear that the loan grade B has a majority of its loans with interest rates lesser than 15% while the loan grade C has a majority of its loans with interest rates greater than 15%. This explains why a large percentage of the loans with the loan grade C entered into default while loans with the grade B did not.

5.3 Defaulted loans and geographic location

The third research questions aims to understand if the geographic location of origination of the loan affected the probability of loans entering into default.

5.3.1 Proportion of defaults in each state

The proportion of defaults in each state is illustrated below:

Proportion of defaults in each state

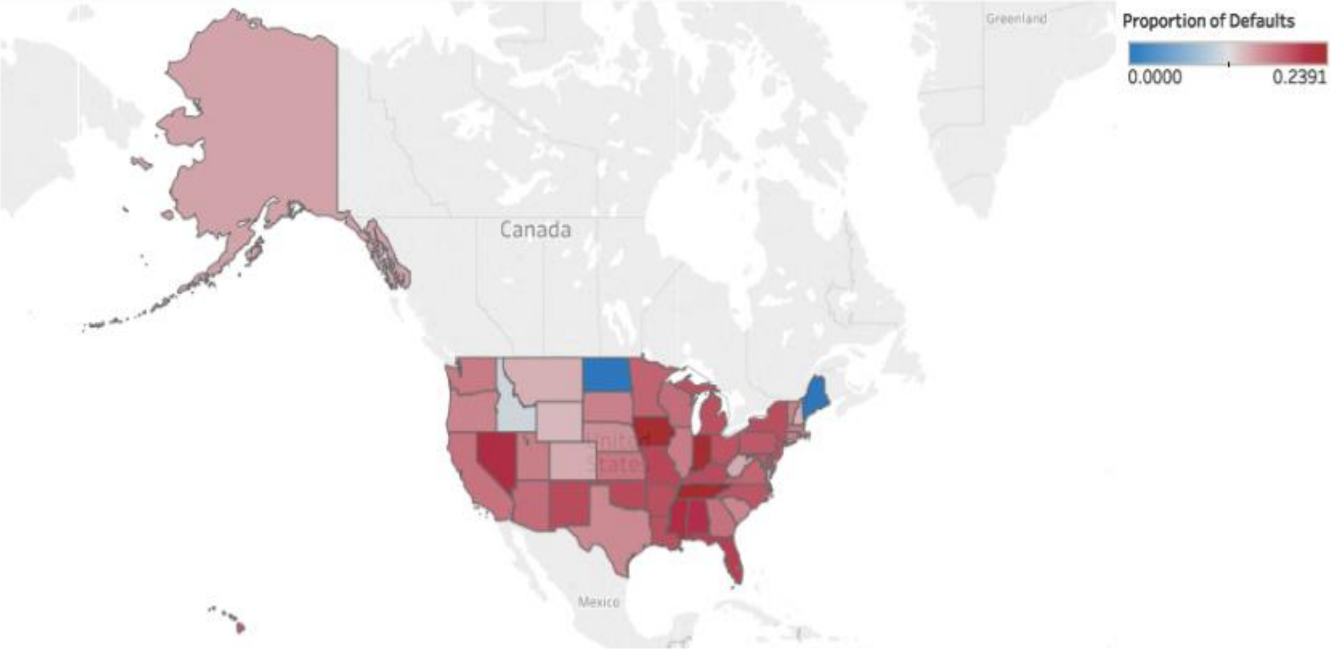

Figure 18: Proportion of defaults in each state

A quick look at this plot instantly paints a new picture about how the defaulted loans are distributed across the different states that is not biased in any way based on the counts of the defaults alone.

Thus, the states that have the highest proportion of defaults are now Tennessee, Iowa, Indiana and Nevada. In order to understand why these four states have the highest proportion of defaults the loan and borrower characteristics are visually analysed across a geographic map using Tableau. Interesting and meaningful insights were then extracted that explained why these states had the highest proportion of defaults.

5.3.2 Geographic location and interest rates

The plot that illustrates the average interest rates in each state is illustrated below:

Average interest rate in each state

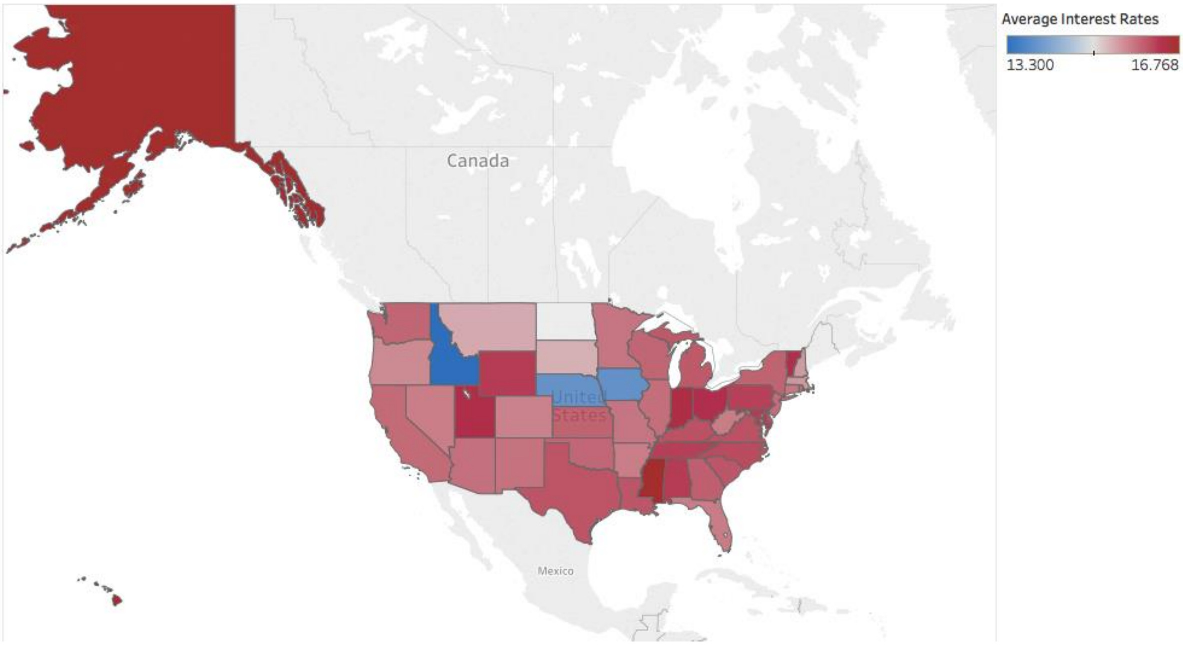

Figure 19: Average interest rate across each state

The states of Tennessee, Indiana & Iowa which are associated with the high proportion of defaults are also associated with a higher average interest rate of greater than 15%.

5.3.3 Geographic location and debt-to-income ratio

The plot that illustrates the average debt-to-income ratio is illustrated below:

Average debt-to-income ratio in each state

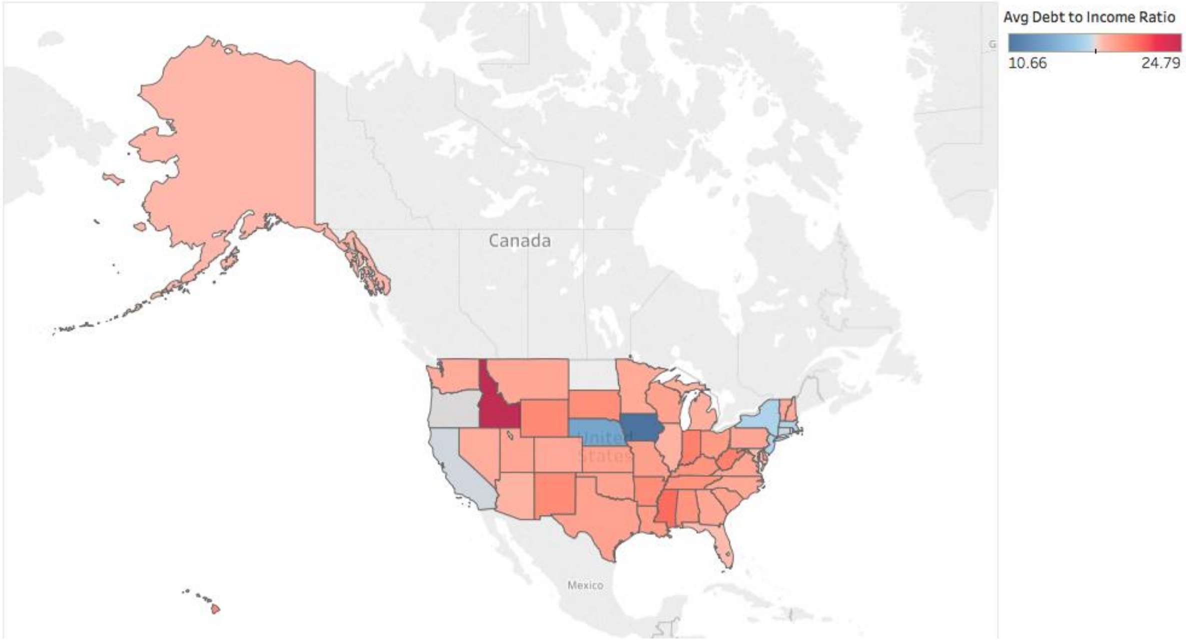

Figure 20: Average debt-to-income ratio across all states

From the plot above, the states of Tennessee, Indiana and Nevada had a debt-to-income ratio greater than 18.5% on average which confirms the answer from the first research question which associated higher debt-to-income ratios with more defaults.

5.3.4 The state of Iowa

The state of Iowa seems to be an outlier that has a high proportion of defaults but not particularly higher interest rates or debt-to-income ratios. One possible explanation is that Iowa is a state with the second lowest average annual income across all the states. However, this is only an hypothesis that has not been investigated into further.

5.4 The best predictors of defaults

The fundamental aim of this research question is to identify the best predictors of default and to use these predictors to build the predictive models. The table below provides the results of the feature selection process that helped identify the top ten predictors of default:

Feature Name
Loan Amount
Interest Rate
Loan Subgrade
Loan Issue Date
Interest received till date
Post charge off gross recovery
Last payment date
Last payment amount
The most recent month LendingClub pulled credit for the loan
The total current balance of all accounts of the borrower

Table 6: Top ten predictors of default

5.5 The best performing predictive model

The fundamental aim of the last research question is to identify the machine learning algorithm that best predicts if a loan would default or not. All the metrics that are used to evaluate the different predictive models are implemented on the test data which the predictive models have not seen yet. This ensures that the results obtained in this section are not biased in any way and are robust.

5.5.1 Accuracy Scores

The table below provides the accuracy scores of every model built on the test data:

Classifier	Accuracy (%)
Gradient Boosted Trees	97.0
Ensemble Model	96.2
Linear Support Vector Machines	95.7
K-Nearest Neighbors	95.2
Logistic Regression	95.1
Decision Tree	94.4
Random Forests	92.6
Naive Bayes	91.3

Table 7: Accuracy percentages of all the models

The accuracy scores are then plotted visually using a horizontal bar plot which is illustrated below:

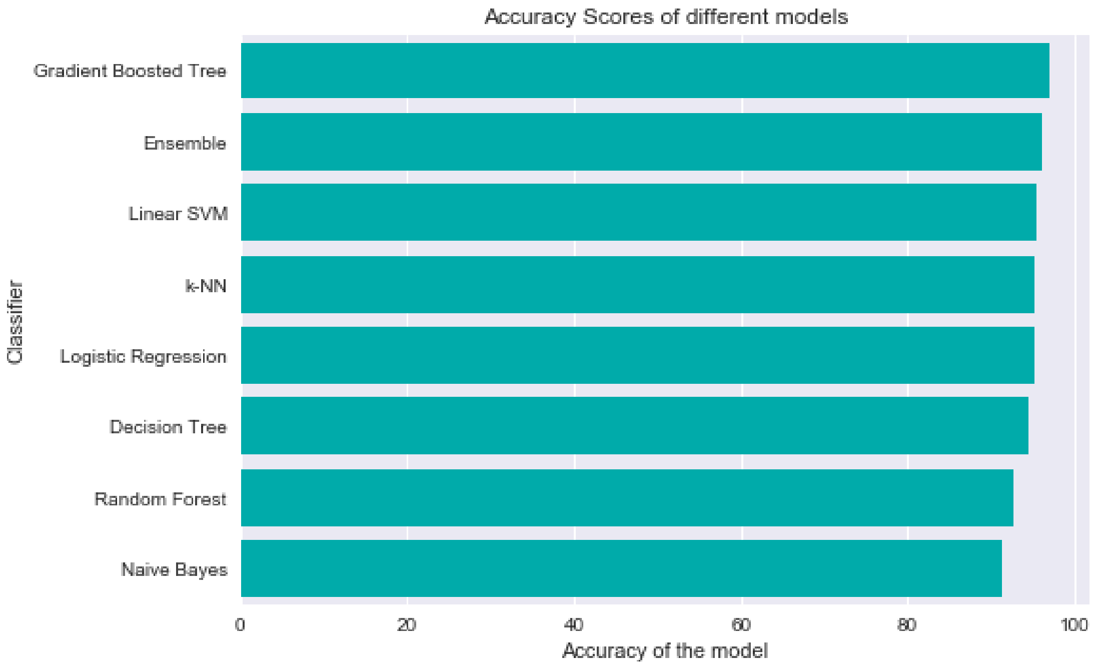

Figure 21: Bar plot of the accuracy scores

From the table and the plot it is clear that the gradient boosted trees (AdaBoost) model has the highest accuracy of 97% while the Naive Bayes model has the lowest accuracy of 91.3%

The boxplot of the ten-fold cross-validated accuracy scores of each model [23] is illustrated below:

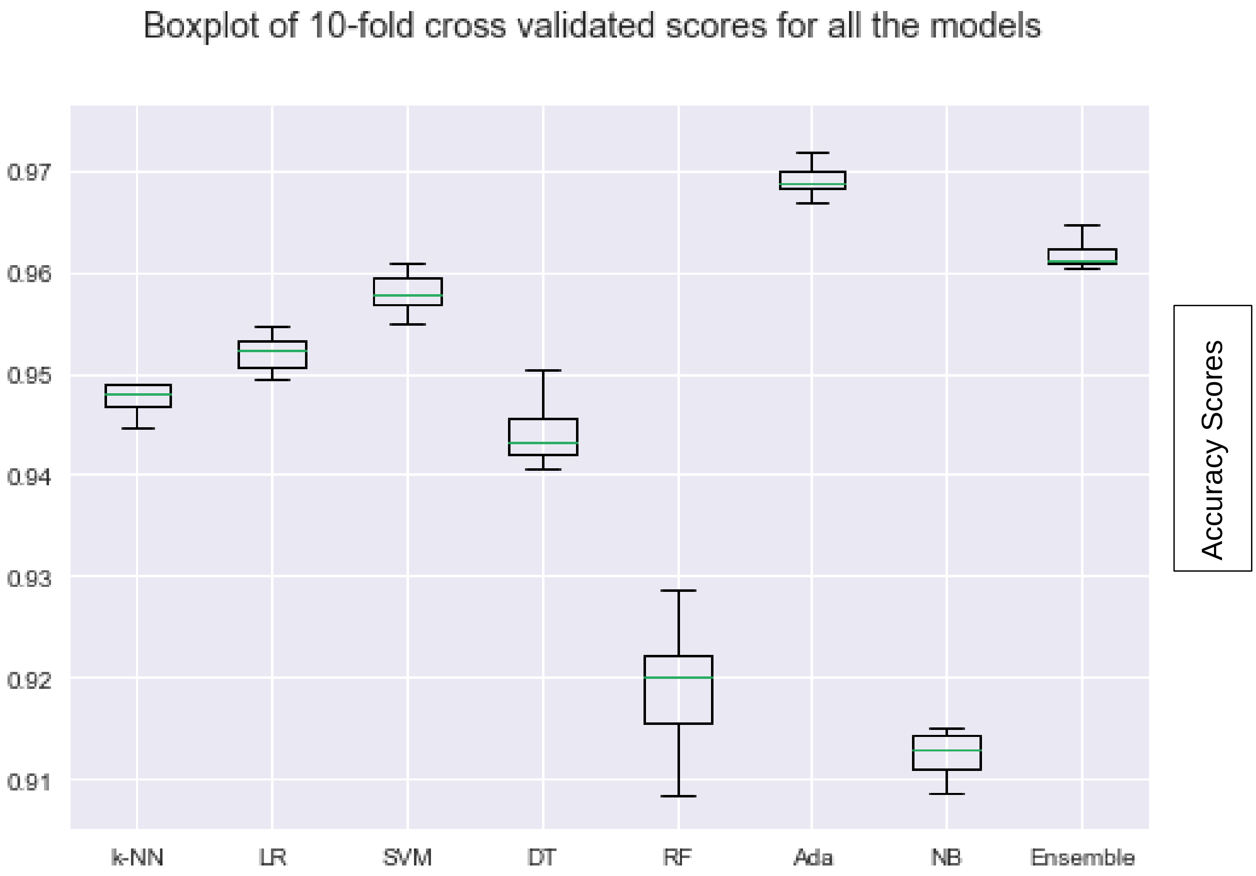

Figure 22: Boxplot of cross-validated scores

From the plot above it is clear that the Gradient boosted trees (AdaBoost) and the Ensemble model have the highest accuracy scores along with the smallest ranges indicating the difference between the lowest and highest values of accuracy recorded for these models do not differ by much. On the other hand, the random forest model and the Naive Bayes classifier have very large ranges along with the lowest accuracy scores.

5.5.2 False Positive Rates

The table below provides the false positive rates of all the classifiers:

Classifier	False Positive Rate (%)
K-Nearest Neighbors	2.9
Logistic Regression	2.3
Decision Tree	2.3
Linear Support Vector Machines	2.2
Gradient Boosted Trees	1.6
Ensemble Model	1.1
Random Forests	0.25
Naive Bayes	0

Table 8: False Positive Rates

The false positive rates are then plotted visually by using a horizontal bar plot as illustrated below:

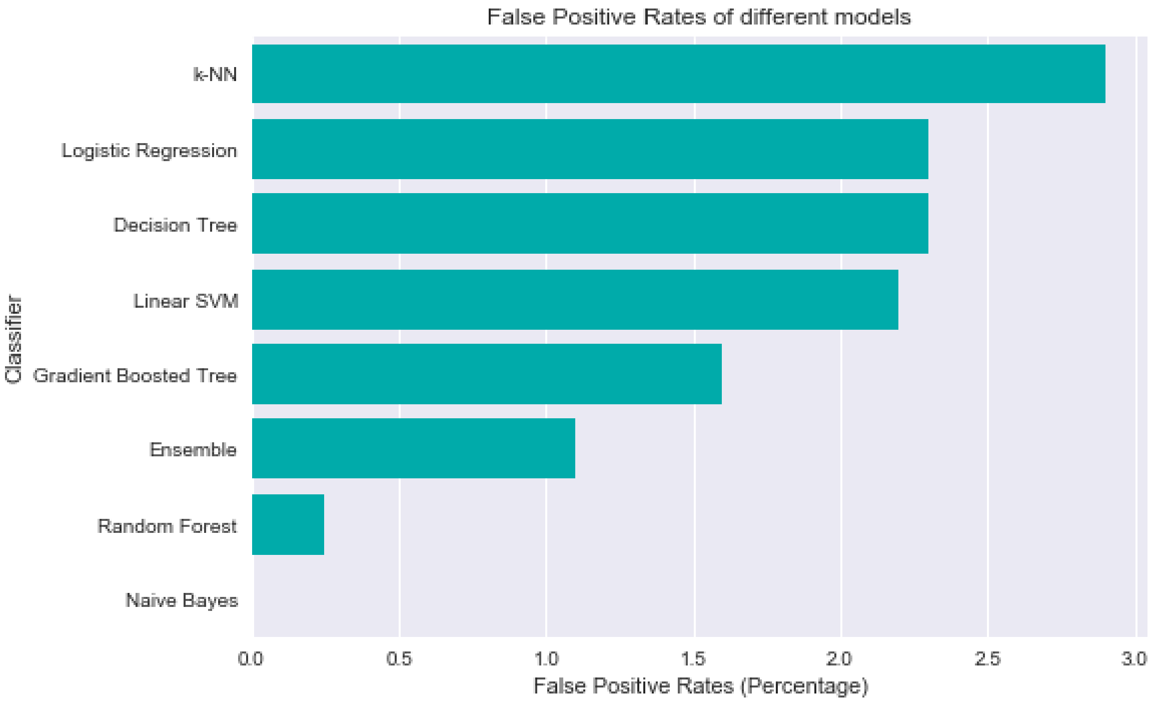

Figure 23: Bar plot of false positive rates

From the table and the figure it is clear that the K-Nearest Neighbors and Logistic Regression have the highest False Positive Rates. This means that these classifiers have a high probability of classifying a fully paid loan as a defaulted loan. This means that using classifiers with a high false positive rate will cost the investor a valuable investment as they might not invest in a loan that these classifiers predict as potentially being a default when they are actually going to be fully paid.

The Gaussian Naive Bayes, Gradient Boosted Trees, Random Forests and the Ensemble model have the lowest False Positive Rates.

5.5.3 False Negative Rates

The table below provides the false negative rates of all the classifiers:

Classifier	False Negative Rate
Naive Bayes	47.6
Random Forests	39.4
Decision Tree	20.2
Logistic Regression	16.1
Ensemble Model	15.6
Linear Support Vector Machines	13.4
K-Nearest Neighbors	12.9
Gradient Boosted Trees	9

Table 9: False Negative Rates

The false negative rates are then plotted visually by using a horizontal bar plot as illustrated below:

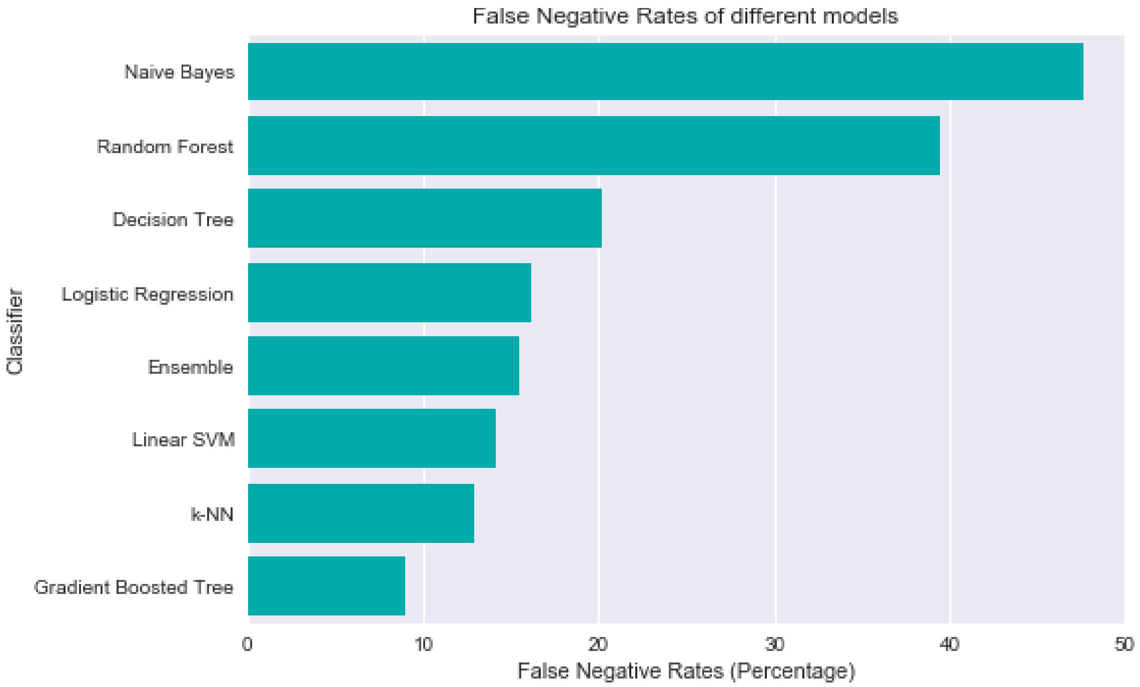

Figure 24: Bar plot of false negative rates

From the data presented in the table and figure it is clear that the Gaussian Naive Bayes classifier has the highest false negative rate while the Gradient Boosted Trees has the lowest false negative rate. A high false negative rate indicates that a loan that has defaulted is classified as fully paid. This is the most expensive mistake a classifier for predicting defaults can make as this costs the investor a large sum of money. Therefore, a classifier that minimizes the false negative rate as much as possible, i.e. the gradient boosted trees in this case is the most optimal classifier for predicting defaults.

5.5.4 F-1 Scores

The table below gives the F-1 score of all the models:

Classifier	F-1 Scores (%)
Gradient Boosted Trees	91.6
Ensemble Model	89.0
Linear Support Vector Machines	87.2
K-Nearest Neighbors	86.9
Logistic Regression	86.2
Decision Tree	83.8
Random Forests	74.9
Naive Bayes	68.7

Table 10: F-1 Scores

The F-1 scores are plotted using a horizontal bar plot and is illustrated below:

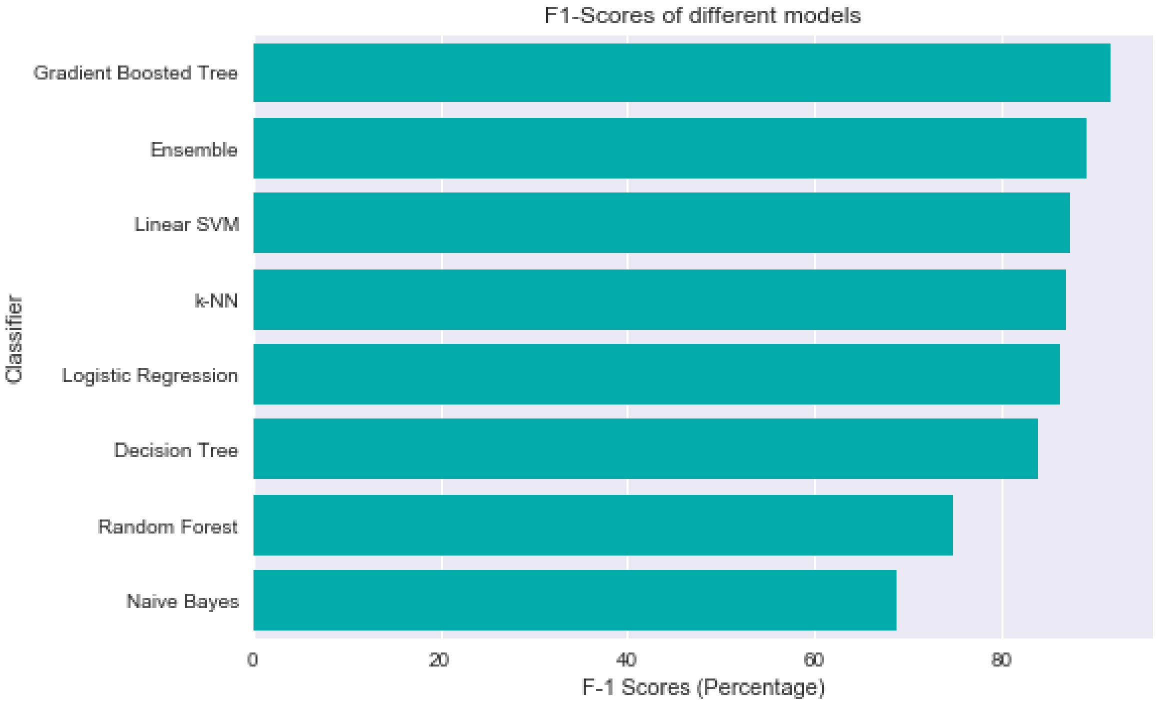

Figure 25: Bar plots of F-1 scores

From the table and the bar plot above it is clear that the Gradient Boosted Trees has the highest F-1 scores while the Gaussian Naive Bayes classifier has the lowest F-1 score. A high value of the F-1 score indicates that not many of the fully repaid loans are predicted as defaulted and that the model predicted most of the defaulted loans accurately. Therefore, the Gradient Boosted Trees with the highest F-1 score is the most optimal model.

5.5.5 Interpreting the best predictive model

From the analysis done above it is clear that the Gradient Boosted Trees with its high accuracy and F-1 scores along with its low false positive and negative rates is the most optimal predictive model for predicting defaults in peer to peer lending.

However, making the model interpretable to the investor who has no prior knowledge about machine learning is the key strength of predictive analytics. Therefore, the lift curve and the cumulative gains curve is plotted for the model.

The lift curve of the Gradient Boosted Trees model is illustrated below:

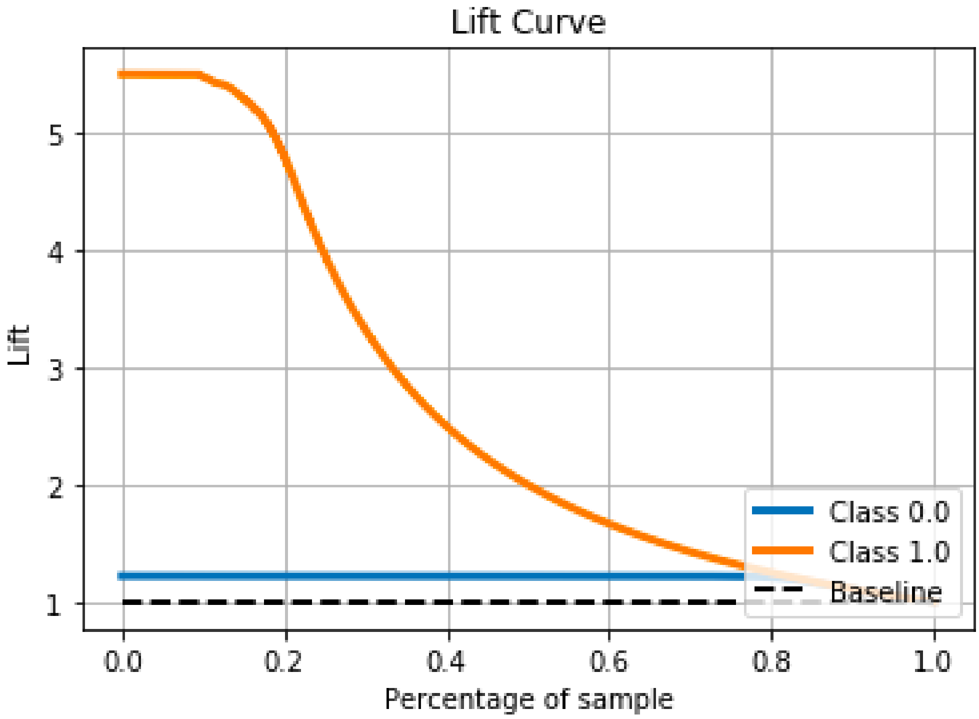

Figure 26: Lift curve

From the lift curve illustrated above the following inference can be made - 18% of the population of loans that are likely to default have a high value of lift that is greater than 5. This means that 18% of the population that are likely to default will give very accurate results when it comes to predicting defaults using the Gradient Boosted Trees model compared to not using any predictive model at all. In other words, using the predictive model is almost 5 times as effective at predicting defaults.

Next, the cumulative gains curve of the Gradient Boosted Trees model is illustrated below:

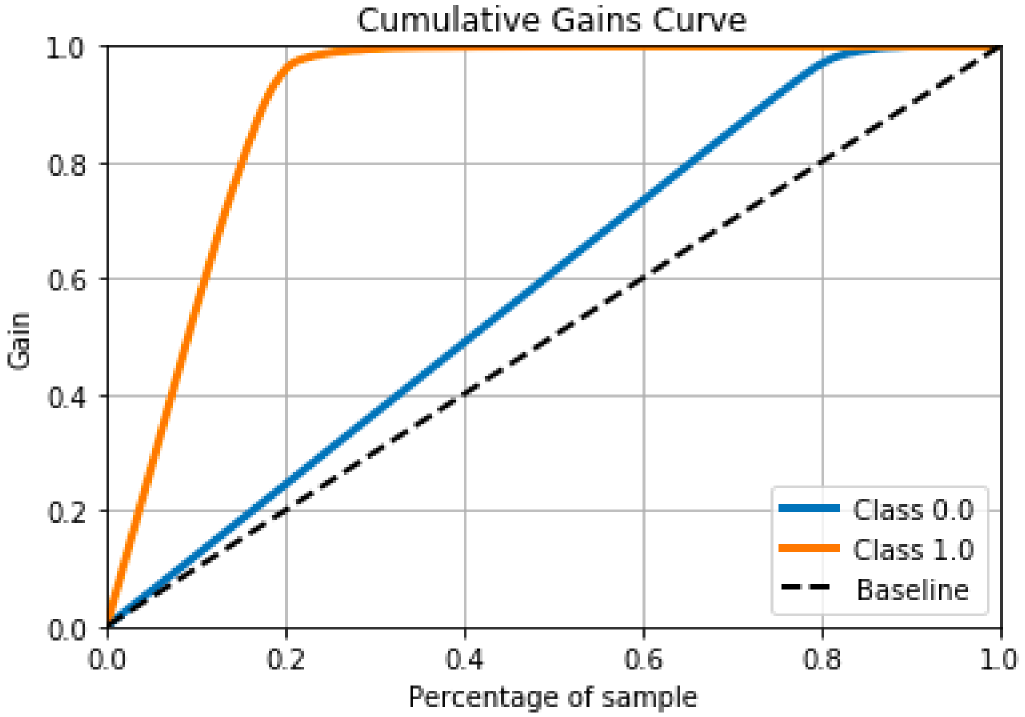

Figure 27: Cumulative gains curve

From the cumulative gains curve above the following inferences can be made - The Gradient Boosted Trees model for the target class – Defaults (1.0) captures/predicts 100% of the defaulted loans if there is 67% of the population of loans that have defaulted in the pool of loans. The model also captures/predicts 95% of the defaulted loans if there is 80% of the population of loans that have defaulted in the pool of loans.

6 CONCLUSION

The concluding section of the project summarises the investigation and key answers derived from the execution of the project.

6.1 Key Findings

The key findings in the project can be summarized as follows – The project has identified the behaviour that differentiate loans that have defaulted from the loans that have been fully repaid using visual analytics .

The project has also identified that the grade of loans that have the highest number of loans that have been fully repaid. This loan grade is grade – B. The grade of loans that have the highest number of loans that have been defaulted is grade – C. The potential reasons for the performance of these grades have also been identified.

Discovered that the geographic location of origination of the loan does indeed affect the probability of default and can be associated with the behaviour of defaulted loans characterised by the first research question.

The project has Identified the top ten predictors of default using a wide array of feature selection methods and using these predictors the project has identified that the most optimal predictive model that predicted defaults is the gradient boosted trees and provided an interpretation of how the model works.

6.2 Lessons Learned

The process of building the foundation for this predictive analytics project had a large amount of research in order to ensure that the eventual predictive models that was used to predict defaults is of substantial quality. As a result, building the dashboard application in order to predict defaults could not be fit into the timeline and as a result, the results had to be built into a Jupyter Notebook and Tableau.

Acknowledging that the process of building predictive models is not a light task and giving it adequate time in the future will help prevent overshooting the limits of what can be implemented within the given timeframe.

6.3 Future Work

This project holds a lot of potential for future work that can be carried out. The steps involved in building a predictive analytics project such as predicting defaults for loans in the peer to peer lending domain can be extended to a wide array of financial and credit industries.

The pipeline built for this project can be automated and the predictive models can be further refined with new data. This presents the opportunity of building a Dashboard using a web development framework and presenting these results in the form of an application to potential investors.

7 REFERENCES

[1] Shoaib Iqbal, “Global Peer to Peer Lending Market by End-user (Consumer Credit Loans, Small Business Loans, Student Loans, and Real Estate Loans) and Business Model Type (Alternate Marketplace Lending and Traditional Lending) – Global Opportunity Analysis and Industry Forecast, 2014-2022”, *alliedmarketresearch.com,* para. 1, Mar., 2017. [Online]. Available: https://www.alliedmarketresearch.com/peer-to-peer-lending-market. [Accessed: May. 12, 2018].

[2] Olivier Garret and David Galland, “The 4 Best P2P Lending Platforms For Investors In 2017 – Detailed Analysis”, *forbes.com,* para. 5, Jan. 29, 2017. [Online]. Available: https://www.forbes.com/sites/oliviergarret/2017/01/29/the-4-best-p2p-lending-platforms-for-investors-in-2017-detailed-analysis/ - 3374c6d152ab. [Accessed: May. 12, 2018].

[3] Serrano-Cinca C, Gutierrez-Nieto B and Lopez-Palacios L, “Determinants of Default in P2P Lending”, *PLoS One,* vol. 10, no. 10, Oct. 2015, [Online Serial]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4591266/. [Accessed: June. 1, 2018]

[4] Shweta Srivastava, Nikita Joshi and Madhvi Gaur, “A Review Paper on Feature Selection Methodologies and Their Applications”, *International Journal of Engineering Research and Development,* vol. 7, no. 6, pp. 57-61, Jun. 2013, Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.446.9202&rep=rep1&type=pdf. [Accessed: Jul. 16, 2018]

[5] Mark A. Hall, “Correlation-based Feature Selection for Machine Learning”, Doctor of Philosophy thesis, The University of Waikato, Hamilton, NZ, 1999.

[6] Sadegh Bafandeh Imandoust and Mohammad Bolandraftar, “Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background”, *International Journal of Engineering Research and Applications,* vol. 3, no. 5,

pp. 605-610, Sep. 2013, Available: https://www.ijera.com/papers/Vol3_issue5/DI35605610.pdf. [Accessed: Jun. 5, 2018]

[7] Peng, Chao-Ying Joanne, Kuk Lida Lee and Gary M. Ingersoll, “An introduction to logistic regression analysis and reporting”, *The journal of educational research,* vol. 96, no. 1, pp. 3-14, Sep. 2002, Available: https://datajobs.com/data-science-repo/Logistic-Regression-[Peng-et-al].pdf. [Accessed: Jun. 5, 2018]

[8] Evgeniou Theodoros and Massimiliano Pontil, “Support vector machines: Theory and applications”, *Advanced Course on Artificial Intelligence, Springer,* pp. 249-257, Jul. 1999

[9] Bhumika Gupta, Aditya Rawat, Akshay Jain, Arpit Arora and Naresh Dhami, “Analysis of Various Decision Tree Algorithms for Classification in Data Mining”, *International Journal of Computer Applications,* vol. 163, no. 8, April 2017, Available: https://pdfs.semanticscholar.org/fd39/e1fa85e5b3fd2b0d000230f6f8bc9dc694ae.pdf. [Accessed: Jun. 6, 2018]

[10] Robert E. Schapire, “Explaining adaboost”, *Empirical Inference, Springer,* pp. 37-52, 2013, Available: http://rob.schapire.net/papers/explaining-adaboost.pdf. [Accessed: Jun. 6, 2018]

[11] Irina Rish, “An empirical study of the naive Bayes classifier”, *IJCAI 2001 on empirical methods in artificial intelligence,* vol. 3, no. 22, pp. 41-46, Aug. 2001, Available: https://www.cc.gatech.edu/~isbell/reading/papers/Rish.pdf. [Accessed: Jun. 7, 2018]

[12] Thomas G. Dietterich, “Ensemble Methods in Machine Learning”, *International workshop on multiple classifier systems, Springer,* pp. 1-15, Jun. 2000. [Abstract]. Available: https://link.springer.com/chapter/10.1007/3-540-45014-9_1. [Accessed: Jun. 7, 2018]

[13] David M. W. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.”, School of Informatics and Engineering Flinders University, Adelaide, Australia, Tech. Report, SIE-07-001, 2011.

[14] Tariq Jaffery and Shirley X. Liu, “Measuring Campaign Performance by Using Cumulative Gain and Lift Chart”, *SAS Global Forum 2009,* p. 196, 2009. Available:

http://support.sas.com/resources/papers/proceedings09/196-2009.pdf. [Accessed: Jun. 8, 2018]

[15] Jin Yu and Yudan Zhu, “A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P) lending.”, *Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference, IEEE,* pp. 609-613, Apr. 2015.

[16] Chang Shunpo, Simon Dae-oong Kim and Genki Kondo, “Predicting Default Risk of Lending Club Loans”, *CS229, Machine Learning,* Project Report, Stanford University, 2015. Available: http://cs229.stanford.edu/proj2015/199_report.pdf. [Accessed: Jun. 9, 2018]

[17] Jitendra Nath Pandey and Maheshwaran Srinivasan, “Predicting Probability of Loan Default.”, *CS229, Machine Learning,* Project Report, Stanford University, Dec. 2011. Available: http://cs229.stanford.edu/proj2011/PandeySrinivasan-PredictingProbabilityOfLoanDefault.pdf. [Accessed: Jun. 9, 2018]

[18] Wendy Kan, “Lending Club Loan Data.”, *kaggle.com, 2016.* [Online]. Available: https://www.kaggle.com/wendykan/lending-club-loan-data. [Accessed: Jan. 16, 2018]

[19] Wes McKinney, “Data Structures for Statistical Computing in Python”, *Proceedings of the 9th Python in Science Conference,* pp. 51-56, 2010.

[20] Fabian Pedregosa et al., “Scikit-learn: Machine Learning in Python.”, *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[21] Judi Scheffer, “Dealing with missing data”, *Research Letters in the Information and Mathematical Sciences,* vol. 3, pp. 153-160, 2002.

[22] John D. Hunter, “Matplotlib: A 2D Graphics Environment.”, *Computing in Science & Engineering,* vol. 9, pp. 90-95, 2007.

[23] Jason Brownlee, “How To Compare Machine Learning Algorithms in Python with scikit-learn.”, Jun. 2016. [Online]. Available: https://machinelearningmastery.com/compare-machine-learning-algorithms-python-scikit-learn/. [Accessed: Jul. 19, 2018]

8 APPENDIX

8.1 Appendix A: General Information

Additional plots and analysis that have been implemented for the purpose of data cleaning and the research questions are in the form of Jupyter Notebooks. A detailed ‘README’ file is provided along with the source code files for interpretation.