STUDY ON CREDIT RISK MODELING SYSTEM USING MACHINE LEARNING TECHNIQUES

BY

Sima Akter

ID: 152-15-5812

This Report Presented in Partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Ahmed Al Marouf (AAM)
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Shah Md. Tanvir Siddique (SMTS)
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

SEPTEMBER 2018

APPROVAL

This Project titled **"STUDY ON CREDIT RISK MODELING SYSTEM USING MACHINE LEARNING TECHNIQUES"** , submitted by Sima Akter to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 30 July 2018.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Sheak Rashed Haider Noori — **Internal Examiner**
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Zahid Hasan — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Ahmed Al Marouf (AAM), Lecturer, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

_____

Ahmed Al Marouf (AAM)
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

_____

Shah Md. Tanvir Siddique (SMTS)
Senior Lecturer
Department of CSE
Daffodil International University

Submited by:

_____

Sima Akter
ID : 152-15-5812
Department of CSE
Daffodil International University

ACKNOWLEDGEMENTS

First I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year thesis successfully.

I really grateful and wish my profound my indebtedness to **Ahmed Al Marouf (AAM)**, **Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of *"Machine Learning"* to carry out this project. His endless patience, scholarly guidance , continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to Head, Department of CSE, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Every lender's organization such as banks and credit card companies use credit score system to determining the creditworthiness of their clients. Currently, they are using numerical scoring system in where the score determined by the compering new customer vs. existing customer profile. This does not capture the exact behavior of certain individual entities or more optimal ways to segment scoring models for which few loan trends to classify in a result organization are deprive of profit and lead to the loss. Now it analyzed that the problem can be optimized using Machine Learning technique and possible to forecast the behavior of the customer. In this study, we applied various machine learning technique to predict the classified loans, minimize credit risk and maximize the profit of the lender's organization. Hence, this study intended to find the best modeling with best performance and accuracy by the comparing their results.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The process of deciding to accept or reject a clients credit by banks commonly executed via judgmental techniques and/or credit scoring models. From our observation, a good credit risk model helps financial institutions to sanction loan application for creditworthy applicants, thus increasing the profit of financial institution; it also declines non-creditable application which decreases losses. In recent years it is noticeable that the credit scoring system is one of the primary measurement technics to a financial institution for determining the credit risk and increasing the cash flow of the organization. It helps a financial organization to reduce possible risk and take managerial decisions. The factors involved in determining this likelihood are complex, and extensive statistical analysis and modeling are required to predict the outcome of each individual case. In this modern computerized world, this process of deciding can be optimized using statistical methods in machine learning.

In business terms, we try to minimize the risk and maximize of profit for the bank. To minimize loss from the banks perspective, the bank needs a decision rule regarding who to give approval of the loan and who not to. An applicants demographic and socio-economic profiles are considered by loan managers before a decision is taken regarding his/her loan application.

Here we are going to design a Credit Risk Modelling System for lender's Organization. In our country, due to some unknown reasons, few of loans are get classified. Therefore, a good credit risk modeling system can help to forecast the credit risk. Our study is to develop a credit risk modeling system which can minimize the risk of Credit given to people and maximize bank profit.

Hence, this study intended to find the best modeling with best performance and accuracy. In this project, we will sort out different challenges during data collection, filtering, and a suitable classification algorithm to find the best possibility of a loan. Later with evaluation metrics, we will evaluate our data and will try to find the best solutions for the Provided

Dataset.

1.2 Motivation

Currently, the lender's organization are using judgmental techniques to determine the credit-worthiness ability of customers. However, sometimes a lot of customers that pose a high risk are still approved due to this judgmental techniques as a result of many accounts are getting be classified, and another problem is potential customer get a lower score. As such, a dynamic behavioral credit risk model that can at an early stage identify customers that pose a higher risk of future delinquency could be a valuable tool for the lender's organization. Therefore, it is required for building a predictive model to classify accounts as either high risk or not. The models will build using several different machine-learning algorithms trained on the historical credit datas. Machine learning is a sub-field of computer science that has the ability to find patterns, generalize and learn without being explicitly programmed. Machine learning techniques are therefore highly suitable for a problem such as this.

1.3 Rationale of the Study

Previously, lot's of studies have been conducted to predict the defaulter in where decision is taken by the analyzing the demographic and socio-economic profiles of an applicant. There are several limitations in this methods, in particular actual comparison are not possible to perform between the existing customer behavior with a new potential customer. In this process, few variables get missing from comparison for which delinquency rates are rising over the period. This possess a real challenge for banks and other lending financial institutions, as they are now more than ever are in need of a robust risk prediction model to generalize the economic behavior of their current and potential future clients. In this research we used machine learning technique to build an predictive models in where several important features are identified and used to predict the behavior of customer.

1.4 Research Questions

In order to understand further objectives of this study we introduced the following research questions:

1. What early behavior best predicts a high risk customer?
2. Which Machine Learning Algorithms are suitable for credit risk modeling?

3. Which feature is really required to predict the defaulter?

4. How do different algorithms perform on credit risk prediction?

1.5 Expected Output

Machine learning and artificial intelligence have been changing our world over the past two decades, computers are trained to learn behavior of people in order to provide them different types of facility from personal to financial sectors. Computers can even be trained to drive cars. As the learning algorithms become more sophisticated and advanced they are applied in a wider range of fields. To find added value, this research will focus on different types of modeling approach and model performance of machine learning and **identify suitable machine learning algorithm for credit risk system**.

1.6 Report Layout

- Chapter 2 provides the research background behind the credit risk modeling system.
- Chapter 3 discussed about the research methodology, different types algorithm used in this research, data preprocessing and statistical analysis.
- Chapter 4 perform the evaluation of different experimental result and discussed about the findings.
- Chapter 5 summarize key contributions of this work and highlights opportunities for future research.

Chapter 2

Background

2.1 Introduction

This chapter serves to highlight and review work relevant to this paper. In it we look at several papers that are highly relevant to this thesis.

There are much literature on credit and risk scoring models, but very few use machine learning methods or use credit card data. One explanation may be the lack of credit datasets, as such data cant be published given its sensitive nature.

2.2 Related Works

In retail banking, the credit risk of an applicant was evaluated in a subjective manner based upon underwriters' experiences. Typically, information on the customer was obtained through personal relationships between the customer and staff at the lender, which curtailed the movement of customers between lenders. Lending was often a judgmental process where an underwriter (typically the bank manager) assessed applications based on criteria known as the 5Cs:

- Character - is the applicant or any of their family known to the organization?;
- Capital - how much of a deposit is the applicant offering and what is the loan amount being requested?;
- Collateral - what security is the applicant offering?;
- Capacity - what is the repaying ability of the applicant?;
- Condition - what are the general conditions of the economy at present?

Thus banks and financial institutions to improve the process of assessing creditworthiness of an applicant during the credit evaluation process develop Credit scoring models. Credit scoring is a system creditors (banks, insurance companies) use to assign credit applicants to either a good credit group the one that is more likely to repay the debt or a bad credit group the one who has a high possibility of defaulting on debt or any financial obligation i.e. not

paying within the given deadline.

Construction of credit scoring models requires data mining techniques. Using, demographic characteristics, historical data on payments and statistical techniques, these models can help in identifying the important demographic characteristics, which is related to credit risk, and assign a score to each customer. The probability of an applicant will be defaulter was calculated from the information of applicant which was provided by the applicant at the time of application filing and thus information was used for the identification of his/her creditworthiness.

2.3 Research Summary

In this research, we have reviewed four different machine learning algorithm for the purpose of predictions of the loan class with the supervised model for classification. Our goal is to predict a class level of loan which is a choice of the predefined list of possibilities and make accurate predictions for new, never-before-seen data. This research deals with the design aspects related to financial fraud detection. The main aim of important feature selection from data sets is to enhance the computational performance and the overall prediction result of this research.

This dataset has 111107 total instances and 19 relevant client information features including their year of job experience, home ownership ,annual income, credit history,purpose of loan, delinquent history etc. The ultimate goal here is to train and compare multiple predictive models using supervised learning techniques and finally selecting an optimal model that best classifies defaulters and non-defaulters in the defaulters column accurately.

Performance metrics are used as small increase in performance can lead to large economic benefits. In Classification method accuracy, sensitivity, specificity, precision, false positive rate are the performance measure.

2.4 Scope of the Problem

When a business applies for a loan, the lender must evaluate whether the business can reliably repay the loan principal and interest. Lenders use different types of dimensions to assess credit risk. There are lots of complexity arising when banks incorporate the many dimensions during the credit risk assessment. These dimensions typically include financial information

such as liquidity ratio, or behavioral information such as loan/trade credit payment behavior. Summarizing all of these various dimensions into one score is challenging, but machine learning techniques help achieve this goal.

The common objective behind machine learning and traditional statistical learning tools is to learn from data. Both approaches aim to investigate the underlying relationships by using a training dataset. Typically, statistical learning methods assume formal relationships between variables in the form of mathematical equations, while machine learning methods can learn from data without requiring any rules-based programming. As a result of this flexibility, machine learning methods can better fit the patterns in data.

2.5 Challenges

Hence this study is intended to find the best modeling with best performance and accuracy. In this research, we sort out different challenges during data collection and filtering and a suitable classification algorithm to find the best possibility of a loan. Later with evaluation metrics, we evaluated our data and find the best solutions for the Provided Dataset.

Chapter 3

Research Methodology

3.1 Introduction

In this chapter we will illustrate our methodology used for this research. The methodology we used as breakdown below:

- **Literature Research:**

  The literature research will discus the following concepts.

  1. Retail credit risk current developments which already discussed in background research section.
  2. Brief general machine learning introduction and Theory on the models that will be put in practice.
  3. Data analysis and transformation.
  4. Evaluation methods

- Data

  Publicly available credit data from Microsoft Corporation will be used for this research.

- Apply Machine Learning Model to Data Set

  Different algorithms will be used, and for each algorithm different settings will be tested. the best performing model of every algorithm on every credit risk quantity is kept for evaluation.

- Assessing model performance:

  Different performance metrics will be evaluated for the developed models.

3.2 Research Subject and Instrumentation

In this section we will illustrate the tools and technique used for this research.

Python Language and Library: In this research we used python as programming language and NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn as library for analyzing the data and building Machine Learning Algorithms.

Python: Python is a very powerful programming language used for many different applications. Over time, the huge community around this open source language has created quite a few tools to efficiently work with Python. In recent years, a number of tools have been built specifically for data science. As a result, analyzing data with Python has never been easier.

NumPy: NumPy is an open source extension module for Python. It provides an abundance of useful features for operations on n-arrays and matrices in Python. The library provides vectorization of mathematical operations on the NumPy array type, which ameliorates performance and accordingly speeds up the execution.

Pandas: Pandas is a Python module that contains high-level data structures and tools designed for fast and easy data analysis operations. Pandas is built on NumPy and makes it easy to use in NumPy-centric applications, such as data structures with labelled axes. Explicit data alignment prevents common errors that result from misaligned data coming in from different sources.

Matplotlib: Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits. Matplotlib allows us to quickly make line graphs, pie charts, histograms and other professional grade figures.

Seaborn: Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.

Scikit-learn: Scikit-learn is Machine Learning library written in Python programming language. It is used to implement different types of machine learning algorithm implementation. This library is designed to interoperate with Python analytic libraries such as NumPy, Pandas, and SciPy.

Machine Learning: Machine learning is a arena of computer science that involves the learning of pattern identification and computational learning theory in AI. Machine learning generally refers to the changes in systems that carry out tasks linked with artificial intelligence (AI). Such tasks include recognition, analysis, planning, robot control, forecasting, etc. It explores the study and construction of algorithm that can make prediction on data. Machine Learning is used to build programs with its tuning parameters that are adapted consequentially so as to increase their functioning by adapting to earlier data.

To predict the class of the loan we use four classification algorithm (Logistic Regression, k-Nearest Neighbours, Decision Trees and Random Forest). Our goal is to predict class level which is a choice of the predefined list of possibilities and make accurate predictions for new, never-before-seen data.

Logistic Regression: Logistic regression is inherited from the field of statistics. It is used in the binary classification problem in where the problem is divided into two classes. Logistic regression used an equation are very much like the linear regression in where input values (x) are combined linearly using weights or coefficient values to predict an output value (y).

$$y = \frac{e^{B0+B1\times x}}{1+e^{B0+B1\times x}} \tag{3.1}$$

Finally after getting the result (y) it divides the result into two class, 0 or 1. If y is greater then 0.5 then 1 otherwise 0.

The things need to consider at the time of data preparation :

- Remove Noise: Logistic regression assumes no error in the output variable (y), consider removing outliers and possibly misclassified instances from your training data.
- Remove Correlated Inputs: Like linear regression, the model can over-fit if you have multiple highly-correlated inputs. Consider calculating the pairwise correlations between all inputs and removing highly correlated inputs.
- Fail to Converge: It is possible for the expected likelihood estimation process that learns the coefficients to fail to converge. This can happen if there are many highly correlated inputs in your data or the data is very sparse (e.g. lots of zeros in your input data).

k-Nearest Neighbours : The K Nearest Neighbors (k-NN) is a simple Machine Learning Algorithm which are used in classification and regression problem. It is most widely used in classification problem as like we used to predict the defaulter. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set its nearest neighbors.

Choosing the factor of K: We can illustrate from the below figure the boundary becomes mellifluous over the increasing the value of K. With the increasing value of K the result of test becoming elegant. The error rate for training and testing depend on the value of parameter K.

We started with k=1 and later check for k up to 40.

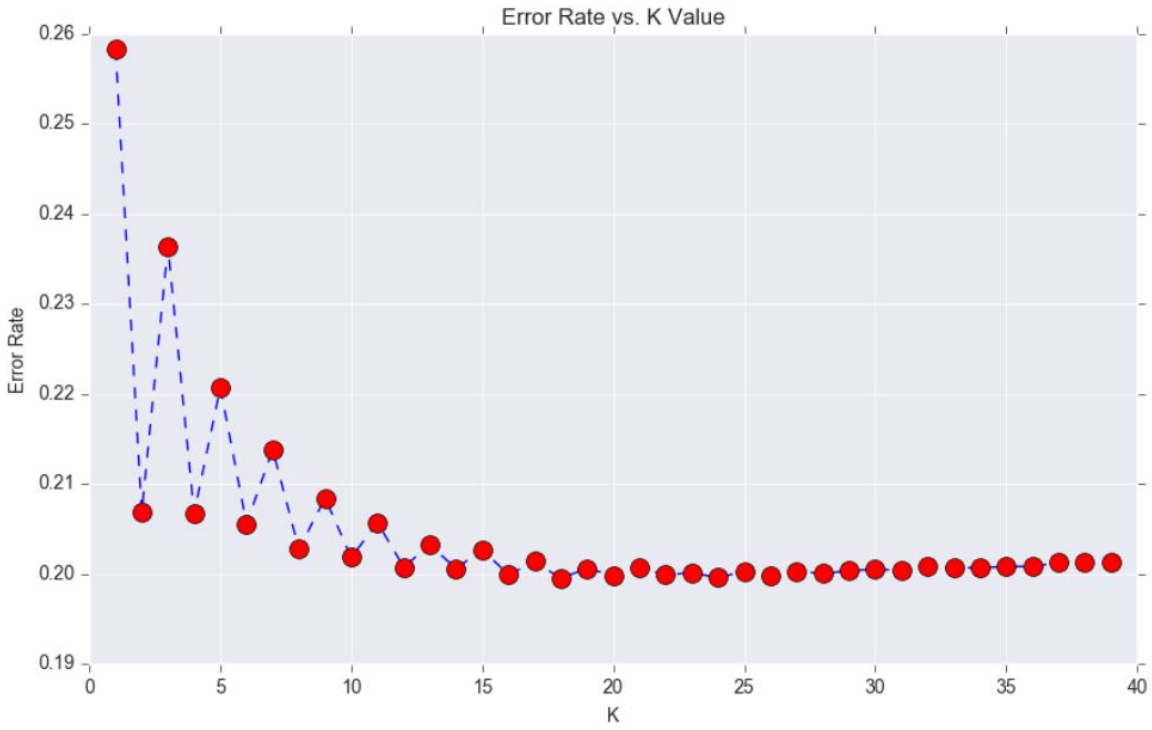

Figure 3.1: Error Rate vs. K Value

From the above diagram we can see that that after around K¿20 the error rate just tends to hover around .20-.23. To finalize the K value for our model we will retrain the model with that and check the classification report and we will use K=20.

Strengths, weaknesses, and parameters of kNN : The below two important parameter has significant impact in KNN algorithm :

- Total Number of Neighbors
- Measurement of distance between data point: Most of the time shortest value for K like four of six can give a better result, but we should adjust the parameter with our test result.

There are several advantages of Knn : The model is very easy to understand, and often gives reasonable performance without a lot of adjustments .Using this algorithm is a good baseline method to try before considering more advanced techniques. On the other hand though it is fast but for larger training set prediction may be slow .It does not perform well when data set have many features. In these later method we will try to reduce or overcome the limitation of Knn.

Decision Trees : Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

Strengths, weaknesses, and parameters: As discussed earlier, the parameters that control model complexity in decision trees are the pre-pruning parameters that stop the building of the tree before it is fully developed. Usually, picking one of the pre-pruning strategiessetting either max_depth, max_leaf_nodes, or min_samples_leafis sufficient to prevent over-fitting.

Advantages includes : the resulting model can easily be visualized and understood by non-experts (at least for smaller trees), and the algorithms are completely invariant to scaling of the data. As each feature is processed separately, and the possible splits of the data dont depend on scaling, no preprocessing like normalization or standardization of features is needed for decision tree algorithms. In particular, decision trees work well when you have features that are on completely different scales, or a mix of binary and continuous features.

The main downside of decision trees is that even with the use of pre-pruning, they tend to overfit and provide poor generalization performance. Therefore, in most applications, the ensemble methods we discuss next are usually used in place of a single decision tree.

Random Forest: Random Forest: Random Forest is an algorithm which can be used in classification and regression problem. It is a Supervised Classification Algorithm, from the name it gives us the information that the algorithm creates the forest using the number of trees. The algorithm is getting more robust with more trees in the forest. The higher number of trees in the forest the higher accuracy in the results. It is a tree-based algorithm which builds several trees and then combines their output to ameliorate the ability of the model. The method of consolidating trees is called an ensemble method. Ensembling is a technic which combines weak learners for building a strong learner.

Strengths, weaknesses, and parameters: This algorithm can solve both type of problems i.e. classification and regression and does a decent estimation at both fronts. One of benefits of Random forest which excites me most is, the power of handle large data set with higher dimensionality. It can handle thousands of input variables and identify most significant variables so it is considered as one of the dimensionality reduction methods. Further, the model outputs Importance of variable, which can be a very handy feature (on some random data set).

3.3 Data Collection Procedure

Data collection and preprocessing is a vital part of machine learning algorithm evaluation and implementation. Pre-processing refers to the transformations applied to the data before feeding it to the algorithm. Data Preprocessing, Includes Data Collection, Cleaning, Conversion of features, Imputing missing data, features standardization, feature scaling and identify new potential features.

Exploratory Data Analysis : The first step in exploratory analysis is reading in the data and then exploring the variables. It is important to get a sense of how many variables and cases there are, the data types of the variables and the range of values they take on. Initially From the Source Data There are 111107 observations and 20 features. Our features list are:

- **Loan ID:** A unique Identifier for the loan information.
- **Customer ID:** A unique identifier for the customer. Customers may have more than one loan.
- **Loan Status:** A categorical variable indicating if the loan was paid back or defaulted.
- **Current Loan Amount:** This is the loan amount that was either completely paid off, or the amount that was defaulted.
- **Term:** A categorical variable indicating if it is a short term or long term loan.
- **Credit Score:** A value between 0 and 800 indicating the riskiness of the borrowers credit history.
- **Years in current job:** A categorical variable indicating how many years the customer has been in their current job.
- **Home Ownership:** Categorical variable indicating home ownership. Values are ”Rent”, ”Home Mortgage”, and ”Own”. If the value is OWN, then the customer is a home owner with no mortgage.
- **Annual Income:** The customer’s annual income.
- **Purpose:** A description of the purpose of the loan.
- **Monthly Debt:** The customer’s monthly payment for their existing loans.
- **Years of Credit History:** The years since the first entry in the customers credit history.
- **Months since last delinquent:** Months since the last loan delinquent payment.

- **Number of Open Accounts:** The total number of open credit cards.
- **Number of Credit Problems:** The number of credit problems in the customer records.
- **Current Credit Balance:** The current total debt for the customer.
- **Maximum Open Credit:** The maximum credit limit for all credit sources.
- **Bankruptcies:** The number of bankruptcies.
- **Tax Liens:** The number of tax liens.

Removing Unimportant Variables: Getting rid of unnecessary variables is a good first step when dealing with any data set, since dropping variables reduces complexity and can make computation on the data faster. Loan id is random variable which does not make any sense . As we are building the model on loan level not on customer level ,we are considering the attributes of loans but not the customer. So we are removing the customer id.

3.4 Statistical Analysis

Categorical Variables Transformation: It is necessary to convert categorical features to dummy variables using pandas. Otherwise, our machine learning algorithm won't be able to directly take in those features as inputs. Before the transformation, we perform impact analysis for each categorical variables.

Categorical Variables in the Data Set : There are four categorical Variables in the data

set . Term of loans, Purpose of loans, Home ownership and years in the current job.

Term of Loans: There are two types of loan, short term and long term. The following graph show that Term loan has more vulnerable approximately term loan given to 25 percent of people among them 48 percent are defaulter which is 12 percent of given data set . It seems Short term loan are more safe as it has 20 percent defaulter which is 26 percent of the given short term loan . It is provided that 32 percent among provided data set are defaulter.

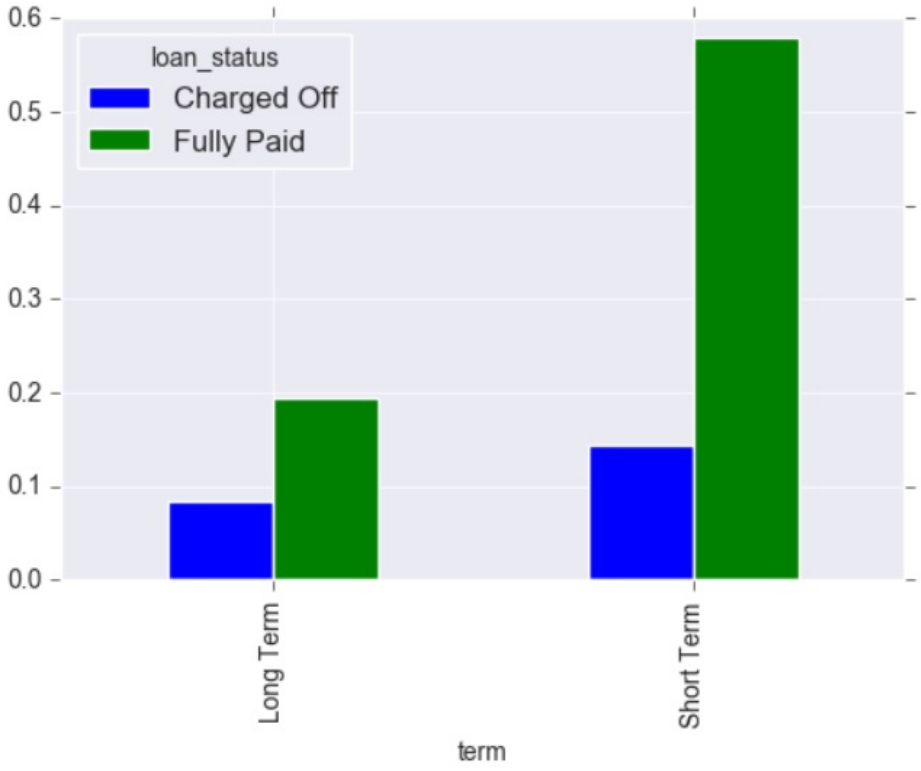

Figure 3.2: Term Loan

Purpose of Loans : There are sixteen types of purpose in where approximately 80 percent of loans are used for debt consolidation and 25 percent are defaulter among them. Other loans are very nominal as home improvements and others loans are in second and third positions respectively.

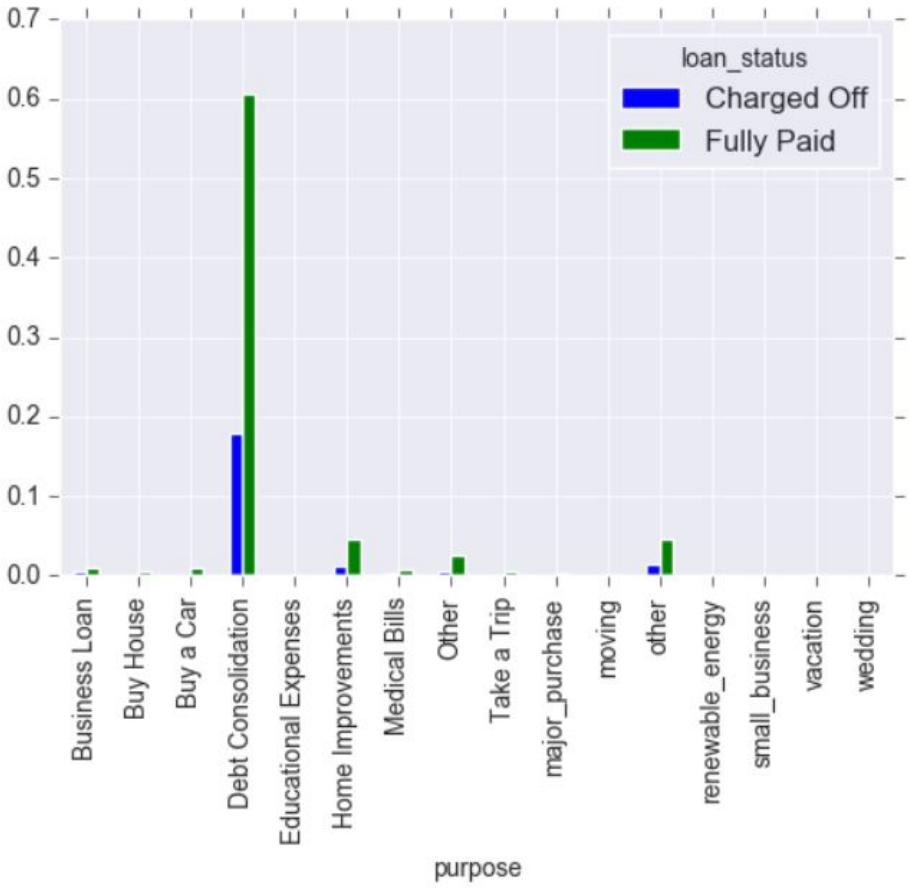

Figure 3.3: Purpose of Loan

House Ownership : Though most loans have been provided to them who has house mortgage, most defaulters are those who live in house with Rent. 15 percent are defaulter which is more than 34 percent of that loan which has been provided to people with rent house.

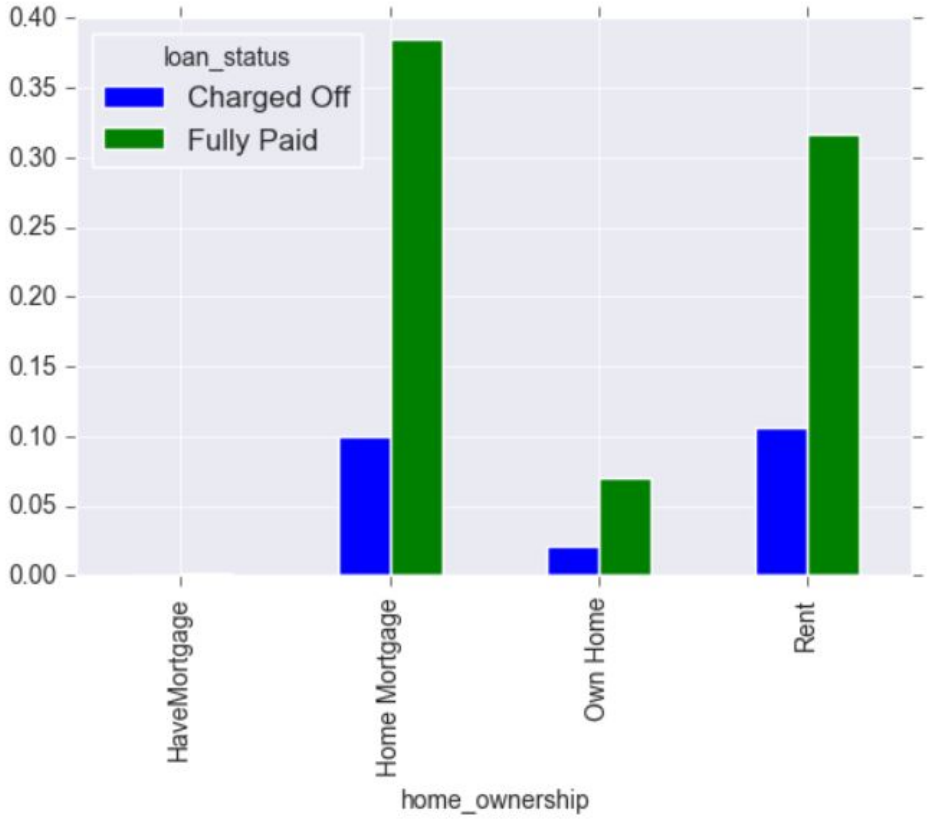

Figure 3.4: House Ownership

Year in Current Job: From the bellow graph it is found that most of the loans given to the employee who has more than 10 years experience. It is a numerical variable we will convert

it to numeric values.

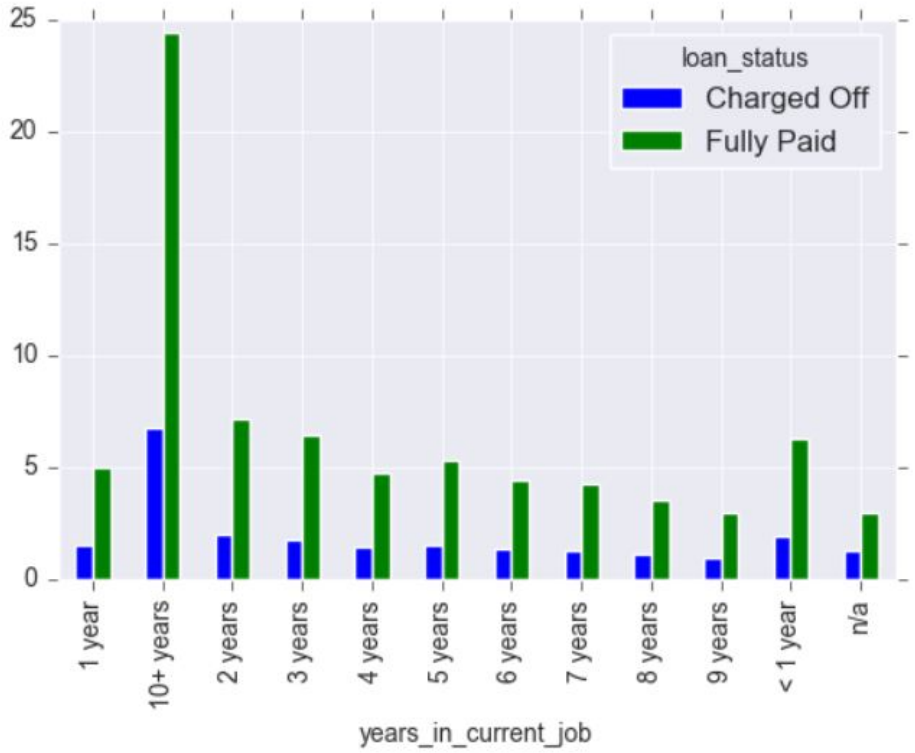

Figure 3.5: Year in Current Job

Categorical Features Encoding: We identified that there are four categorical features Purpose, Term , Home Ownership and year in current job. We convert them into distinct features using one hot encoding method and later we will drop the parent column.

Missing Values Imputation: Data sets are often littered with missing data, extreme data points called outliers and other are strange values. Missing values, outliers, and strange values can negatively affect statistical tests and models and may even cause certain functions to fail.

Detecting missing values is the easy part: it is far more difficult to decide how to handle them. In cases where we have a lot of data and only a few missing values, it might make sense to simply delete records with missing values present. On the other hand, if we have more than a handful of missing values, removing records with missing values could cause to get rid of a lot of data. Missing values in categorical data are not particularly troubling because we can simply treat NA as an additional category. Missing values in numeric variables are more troublesome since we can't just treat a missing value as a number.

Here are a few ways we could deal with them:

1. Replace the null values with 0s
2. Replace the null values with some central value like the mean or median
3. Impute values (estimate values using statistical/predictive modeling methods).

We used seaborn to create a simple heatmap to see where we are missing data!

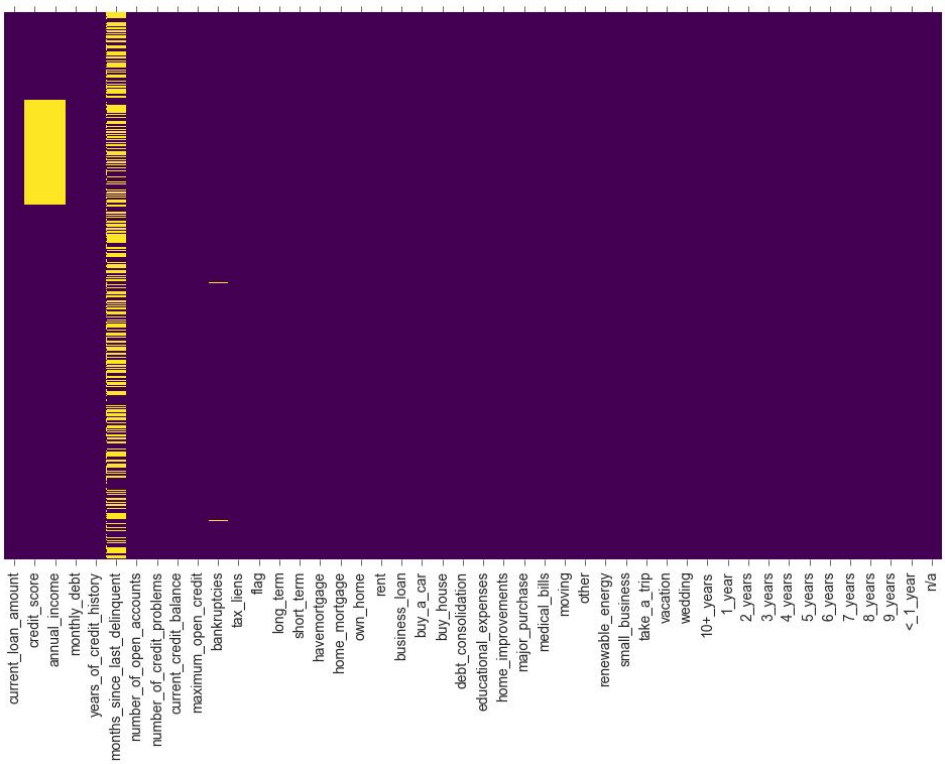

Figure 3.6: Missing Value Representation

Correlation Coefficient: Before building a model and evaluating its results, we need to examine the correlations between the variables in the data set. The goal is to identify those variables which have a strong linear relationship and is done by developing a correlation matrix which takes each continuous variable and finds the correlation coefficient for every pairing in the data set. The correlation coefficient is calculated using Pearson or Spearman measurements, with the values ranging from -1 (negative correlation) to 1 (positive correlation).

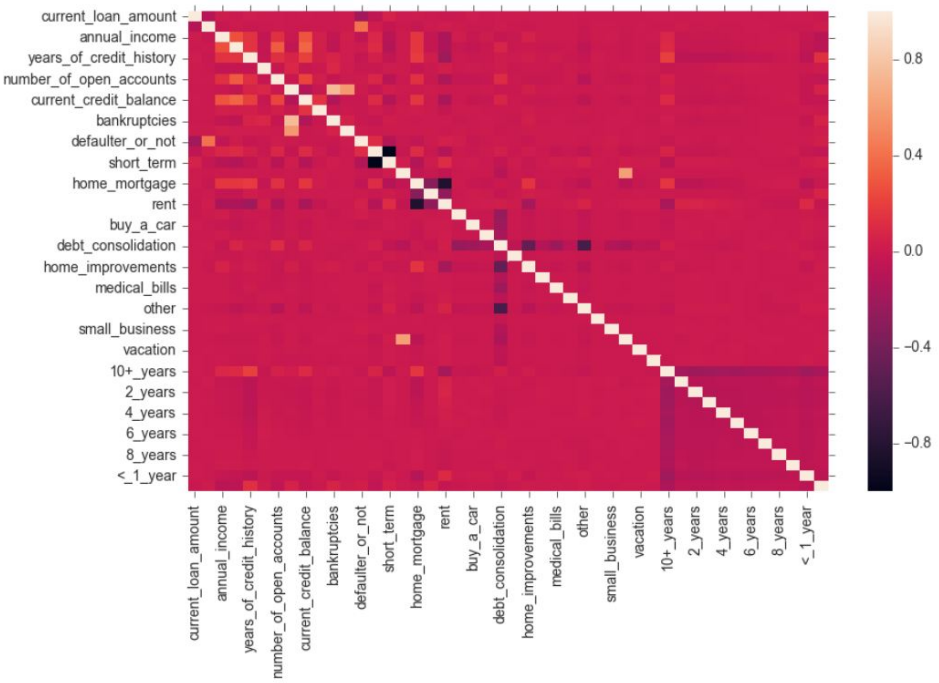

Figure 3.7: Correlation Coefficient

3.5 Implementation Requirements

For implementation, the Python programming language was used. In addition, a number of Python libraries were used for various tasks.

- All programming had done in Python 3.5.
- For machine learning and preprocessing, the Python machine learning library Scikit-Learn is required.
- For data manipulation, exploration and preprocessing, the Python library Pandas is required.
- To visualize data, the Python 2D plotting library Matplotlib is required.

Chapter 4

Experimental Results and Discussion

4.1 Introduction

This section reports the results of the experiments conducted in this research. A total of 4 different machine learning algorithms (Logistic Regression, k-NN, Random Forests and Decission tree) have been done. The results are reported for each machine learning algorithms. The best performing model was decided by the evaluation of confusion matrix, f1-score, precision and recall.

4.2 Experimental Results

To reach the final decision we compare the result of each algorithm and plot as below :

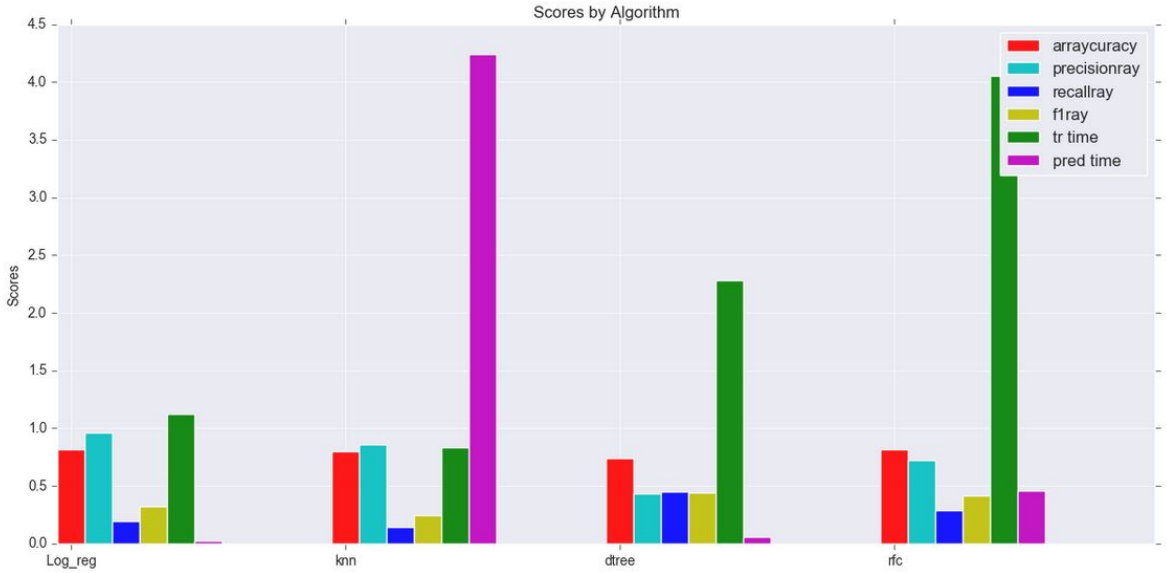

Figure 4.1: Final Result Comparison

Here After analyzing the bar chart with 4 categories we find that,

- Logistic Regression has the maximum accuracy and precision with minimum timing value

- Knn has the maximum train and test time but less accuracy and precision.

So from this above bar graph we can reach into the decision that Logistic Regression has less training time and test time despite the fact for our data set it shows better accuracy and performs better in other evaluation metrics .

4.3 Descriptive Analysis

This section describes measures to evaluate performance in binary classification, that is when there are only two classes, two possible outcomes.

Confusion Matrix: Supervised machine learning classifiers have several evaluation metrics to choose from. Many of them come from a confusion matrix which records correctly and incorrectly classified samples from both classes. Table 4.1 present a confusion matrix. The metrics following will make use of the confusion matrix in the definitions.

Table 4.1: Confusion Matrix

	1 (Predicted)	**0 (Predicted)**
1 (Actual)	True Positive	True Negative
0 (Actual)	False Positive	False Negative

Accuracy: One of the most common metrics is accuracy, which gives the ratio of correctly classified samples to misclassified samples. Accuracy does not take class distribution into account, which makes it poor measure for evaluating performance on imbalanced data.

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn}$$

Precision: It indicates how many values, out of all the predicted positive values, are actually positive. It is formulated as :

$$precision = \frac{tp}{tp+fp}$$

Recall: It indicates how many positive values, out of all the positive values, have been correctly predicted. The formula to calculate the true positive rate is

$$recall = \frac{tp}{tp+fn}$$

F Score: F score is the harmonic mean of precision and recall. It lies between 0 and 1. Higher the value, better the model. It is formulated as :

$$fScore = \frac{2((precision * recall)}{precision + recall}$$

Result Evaluation: In this section we sort out the result of each model in where we used 40 percent of data for testing set and 60 percent for training set and found that Logistic Regression is best performing model.

Table 4.2: Results for experiment

Algorithm	Precision	Recall	F-Measure	Accuracy
Logistic Regression	0.95	0.19	0.32	0.82
kNN	0.86	0.14	0.24	0.80
Decision Trees	0.43	0.45	0.44	0.74
Random Forest	0.72	0.29	0.42	0.81

4.4 Summary

In this research, we perform analysis for 111107 observations and 20 features using four different classification algorithm and it has been observed that the linear regression algorithm is the best fit with the smallest error.

Chapter 5

Summary, Conclusion, Recommendation and Implication for Future Research

5.1 Summary of the Study

In this study, we used dataset provided by Microsoft Corporation for research purpose. To achieve our desired result from this dataset we perform the data validation, data cleaning, imputation and different types of statistical analysis. It is very important to prepare well-structured data for getting intended result from obscure dataset. After preparing the dataset we divide the dataset into two different group one for training and another for testing purpose and perform the different types of model evaluation for predicting feature defaulter. Finally, perform the comparison the result of different types of model.

5.2 Conclusions

The basic purpose of the study contains credit risk modeling with a machine learning algorithm. Basically, we have tried to establish a solid comparison between different classification algorithm and improve the accuracy of the prediction by increasing accuracy and minimizing errors, bias and variance. As a result, we got Logistic Regression as the best performing model in terms of accuracy, precision and timing and Decision Tree is the worst. However, due to Hardware resource limitation, we have failed to test a lot of others machine learning algorithm with a large amount of data. Maybe if we incorporate better learning algorithm and better and improved data cleaning and collection and handling missing values properly the result would have been different.

5.3 Recommendations

In this research project, we explored commonly used machine-learning algorithms used in building models for consumer credit risk. We assessed the quality of the methods on our training and testing sets. By the looking at the metric, logistic regression seems to perform best on our datasets. We recommend the Bank to look into further adjusting parameters in the random forest algorithm, or building a customized version to fill its specific needs. In addition, further fine tune can be performed in the existing parameters to achieve more accurate models.

5.4 Implication for Further Study

The next step in machine learning research on retail credit risk data such as this data would be to evaluate the added value of online learning algorithms. These algorithms are constantly updated when new data becomes available. This would be valuable because when models adjust themselves, there is no need to invest in creating new models when a lot of new data is available or macro economic circumstances have changed.

Future work could also look at other machine learning algorithms for this type of problem. Artificial neural networks have shown promise for imbalanced dataset problems, for instance. A recurrent neural network architecture called Long short-term memory (LSTM) could be interesting to apply to this type of forecasting problem. This type of network could also be used to develop more dynamic models.

References

[1] *Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., Siddique, A., 2016. Risk and risk management in the credit card industry. Journal of Banking & Finance*

[2] *Jasson Brownlee: Master Machine Learning Algorithm*

[3] *Peter Harrington; Machine Learning in Action*

[4] *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani; An Introduction to Statistical Learning*

[5] *Introduction to Statistical Learning; "http://www-bcf.usc.edu/ gareth/ISL/" last access on 01 March 2018, 08:45 PM*

[6] *LaTeX guides : "https://www.sharelatex.com" last access on 28 July 2018, 08:00 PM*

[7] *German Credit Data Analysis ; "https://onlinecourses.science.psu.edu/stat857/node/215" last access on 12 June 2018, 02:00 PM*

[8] *Andreas C. Mller, Sarah Guido : Introduction to Machine Learning with Python ,classification and Regression pages 39–57*

[9] *Data Source ; "https://gallery.azure.ai/Competition/1ad7a6df99794816b9bc071e27d46b10" last access on 05 May 2018, 01:00 PM*

[10] *Credit Scoring using Machine Learning Techniques: International Journal of Computer Applications (0975 8887) Volume 161 No 11, March 2017*

[11] *Yap, Bee Wah, Seng Huat Ong, and Nor Huselina Mohamed Husain. "Using data mining to improve assessment of creditworthiness via credit scoring models." Expert Systems with Applications*

[12] *Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. "Consumer credit-risk models via machine-learning algorithms." Journal of Banking & Finance*

[13] *Siddiqi, Naeem. Credit risk scorecards: developing and implementing intelligent credit scoring*

Appendix A

Research Reflection

This is my first research project in machine learning area. After a lot of research I finally nailed my interest to credit risk modelling as this is an important problem to deal with and it affects all major financial institutions including insurance companies , banks etc. Additionally, this particular dataset gave me opportunity to play around with different machine learning algorithms. Which is exactly what I wanted to grow up my knowledge in the area of Machine Learning and data science field.

Appendix B

Related Issues