A PROJECT REPORT

ON

Loan Approval Model

Submitted To

SSJ IT Solutions Private Limited

Submitted By

Ritik Singh

Anand Gupta

Prakhar Gupta

Suryansh Tiwari

Swarn Pallav Bhaskar

Under the Supervision of

Mr. Kushagra Srivastava

Date of Submission

July 10, 2021

Declaration

I the undersigned solemnly declare that the project report **"Loan Approval Model"** is based on our team work carried out during the course of our internship under the mentorship of **Kushagra Srivastava**.

I assert the statements made and conclusions drawn are an outcome of my research work I further certify that

1. The work contained in the report is original and has been done by me under the general supervision of our supervisor.
2. The work has not been submitted to any other Institution for any other degree/diploma/certificate in this university or any other University of India or abroad.
3. We have followed the guidelines provided by the university in writing the report.
4. Whenever we have used materials (data, theoretical analysis, and text) from other sources, we have given due credit to them in the text of the report and giving their details in the references.

Team Members

Swarn Pallav Bhaskar

Suryansh Tiwari

Ritik Singh

Prakhar Gupta

Anand Gupta

Date: July 10, 2021

Certificate

This is to certify that the Project report entitled **"Loan Approval Model"** done by **Swarn Pallav Bhaskar, Suryansh Tiwari, Ritik Singh, Prakhar Gupta** and **Anand Gupta** is an original work carried out by them under my guidance. The matter embodied in this project work has not been submitted earlier for the award of any degree or diploma to the best of my knowledge and belief.

Date: July 10, 2021

Mr. Kushagra Srivastava
Signature of the Mentor

Acknowledgement

The merciful guidance bestowed to us by the almighty made us stick out this project to a successful end. We humbly pray with sincere heart for his guidance to continue forever.

We pay thanks to our project guide Mr. Kushagra Srivastava who has given guidance and light to us during this project. His versatile knowledge has cased us in the critical times during the span of this project.

We pay special thanks to **Mr. Suraj Jaiswal,** founder of **SSJ IT Solutions Private Limited**, who has been always present as a support and help us in all possible way during this project.

We also take this opportunity to express our gratitude to all those people who have been directly and indirectly with us during the completion of the project.

We want to thanks our friends who have always encouraged us during this project.

At the last but not least thanks to all the mentors of **SSJ IT Solutions Private Limited** who provided valuable suggestions during the period of project.

Abstract

In today's world, Loan approval is a very important process for banking organizations. The model predicts whether the application will be approved or rejected. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. In recent years many researchers worked on loan approval prediction systems. Everyday a large number of people make application for loans, for a variety of purposes. But all these applicants are not reliable and everyone cannot be approved.

Every year, we read about a number of cases where people do not repay bulk of the loan amount to the banks due to which they suffers huge losses. The risk associated with making a decision on loan approval is immense. So the idea of this project is to gather loan data from multiple data sources and use various machine learning algorithms on this data to extract important information. This model can be used by the organizations in making the right decision to approve or reject the loan request of the customers. Machine Learning (ML) techniques are very useful in predicting outcomes for large amount of data.

In this report we have implemented six machine learning algorithms, Logistic Regression (LR), Support Vector Classifier (SVC), k-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF) and XG Boost (XGB) are applied to predict the loan approval of customers. The experimental results conclude that the accuracy of Decision Tree machine learning algorithm is better as compared to Logistic Regression, Random Forest, Support Vector Classifier, k-Nearest Neighbor and XG Boost machine learning approaches.

TABLE OF CONTENT

LIST OF FIGURES

Chapter 1

Introduction

1.1 Problem Statement:

With increase in population and people becoming more aware of the loan-borrow system, it has become difficult for the banks and private lending organizations to process the applications manually, there is an urgent requirement of automation in this field.

1.2 The ML World:

ML helps banks more confidently issue credit to those who pass system checks. For this, programs and algorithms analyze all available information about a potential borrower, study their credit history, changes in their level of wages, and on this basis determines the reliability of the client and the security of the loan. Moreover Chinese banks have already gone further and decided not to limit themselves to analyzing the data exclusively.

They began to introduce facial micro-expressions recognition technology. This allows them to find out if customers are ling about their financial situation when they come to take out loans. To do this, they developed AI systems that, with the help of smartphone cameras, detect minimal changes in facial expressions that are invisible to the naked eye. Thus, banks identify potential fraudsters, and they have already educed their losses from unpaid loans by 60%.

This is a global task that is successfully solved through AI & ML in Financial Services. When an algorithm can analyze all of the available structured and unstructured data (both internal from the company's business processes and external such as customer requests and their actions on social media), a financial institution can discover both useful and potentially dangerous trends. It helps assess risk levels and allow people to make the most informed decisions.

Banks and payment systems have already been developing models to identify ad block most fraudulent transactions. These models are built on the clients transaction history as well as the client's behavior o the Internet. Systems based on ML that detect online frauds have been developed from Big Data technologies.

Chapter 2

Literature Survey

As per various literature surveys it is found that for implementing this project four basic steps are required to be performed.

i. Preprocessing
ii. Missing Data Treatment
iii. Plotting values on graph for visualization purpose
iv. Apply different Machine Learning Algorithms to compare their accuracy.

Description about all these processes is given below-

2.1 Preprocessing

Preprocessing of the data includes data cleaning, data integration, data transformation, data reduction, missing values imputation among other tasks. Below are some of the data transformations that were done to the Loan Approval dataset before we apply any EDA techniques.

Number of Records: This dataset only has 614 observations, limiting us to come to a conclusion.

Missing data: The missing values constitute to 1.05% of the entire dataset. And the missing values are represented by “NaN”.

2.2 Missing data Treatment:

The missing values are found to exist in attributes Sex, Marital_Status, Dependents, SE, CPL_Amount Out, CPL_Term and Credit_His. Out of these, CPL_Amount and CPL_Term are continuous variable. There are different methods to impute missing value, ranging from deleting the observations, deleting the attribute if of no importance, zero them out or plug the mean/median/mode value from all the values.

Here we imputed the values by using the median value for Numerical fields. For remaining

attributes with categorical values, the missing values are imputed using the frequency count of the observations. The Class group with highest frequency was used.

2.3 Plotting values on graph for visualization purpose:

Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. There is just something extraordinary about a well-designed visualization. The colors stand out, the layers blend nicely together, the contours flow throughout, and the overall package not only has a nice aesthetic quality, but it provides meaningful insights to us as well.

seaborn.barplot() method:

A barplot is basically used to aggregate the categorical data according to some methods and by default it's the mean. It can also be understood as a visualization of the group by action. To use this plot we choose a categorical column for the x-axis and a numerical column for the y-axis, and we see that it creates a plot taking a mean per categorical column.

2.4 Approaches:

2.4.1 Logistic Regression:

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems.

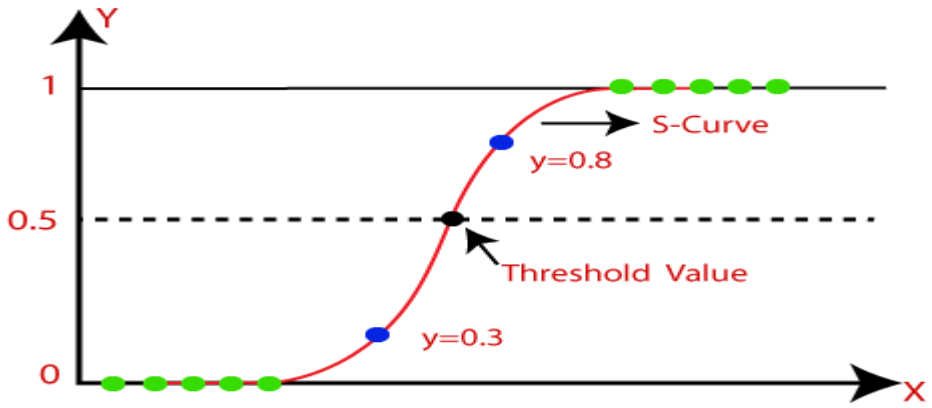


Fig 2.1

2.4.2 Logistic Regression Using K fold:

The k-fold cross-validation procedure is a standard method for estimating the performance of a machine learning algorithm or configuration on a dataset.

A single run of the k-fold cross-validation procedure may result in a noisy estimate of model performance. Different splits of the data may result in very different results.

Repeated k-fold cross-validation provides a way to improve the estimated performance of a machine learning model. This involves simply repeating the cross-validation procedure multiple times and reporting the mean result across all folds from all runs.

2.4.3 Decision Tree:

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

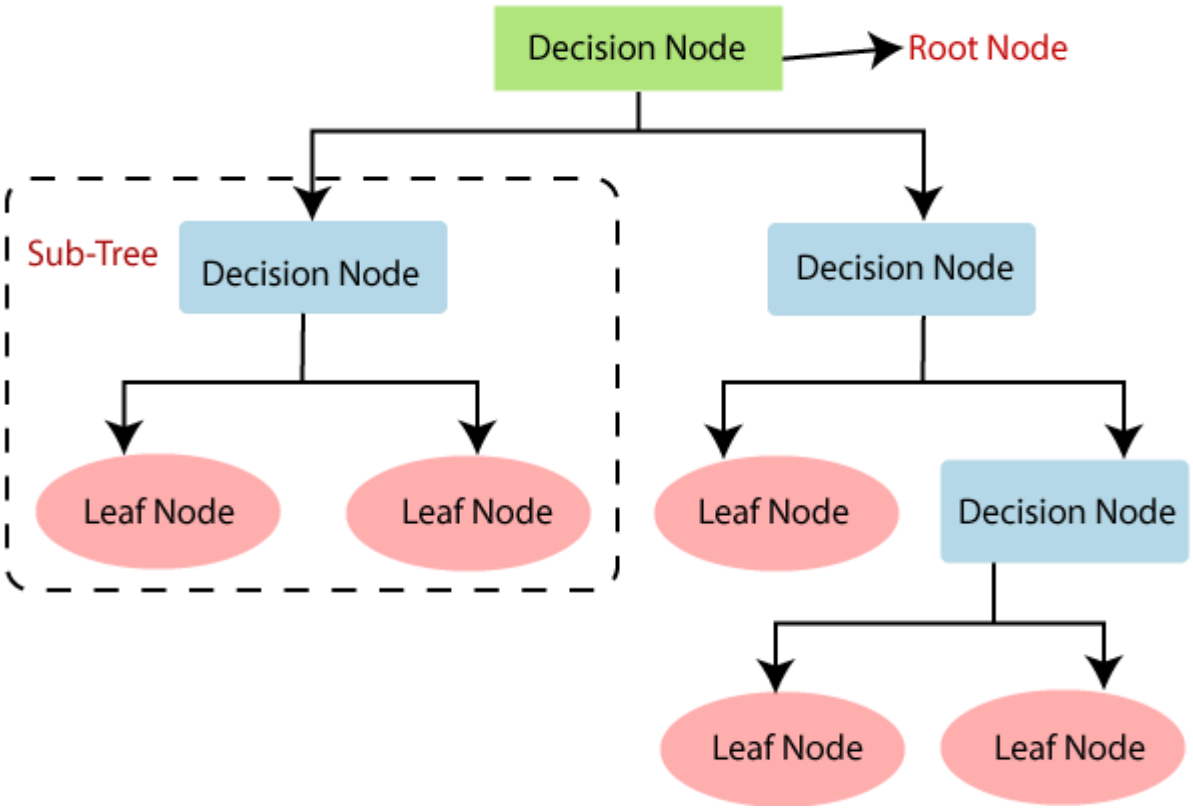


Fig. 2.2

2.4.4 Random Forest:

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

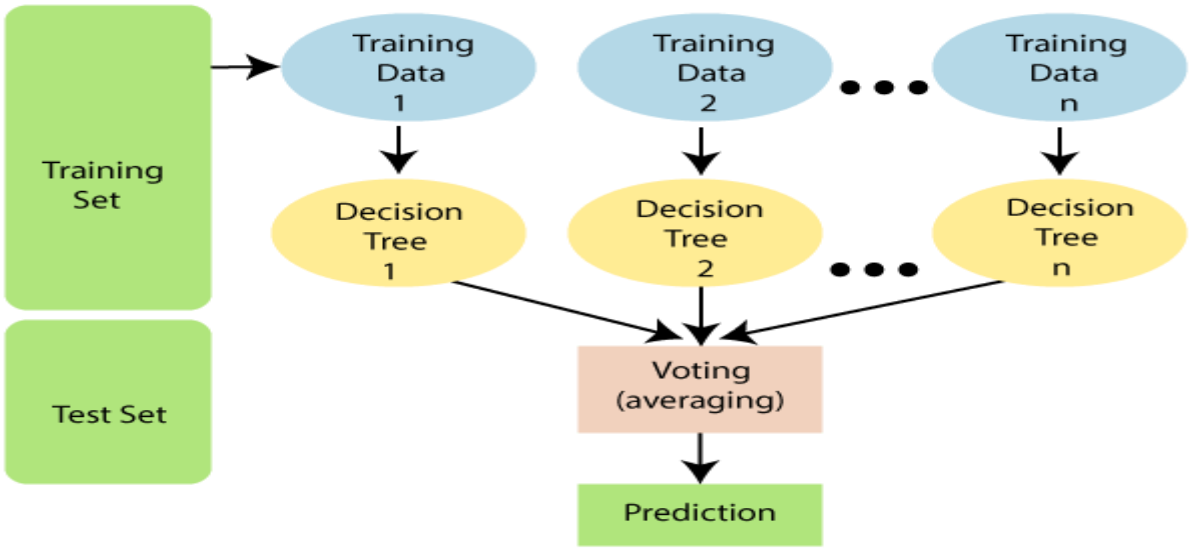


Fig 2.3

2.4.5 XGB Classifiers:

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

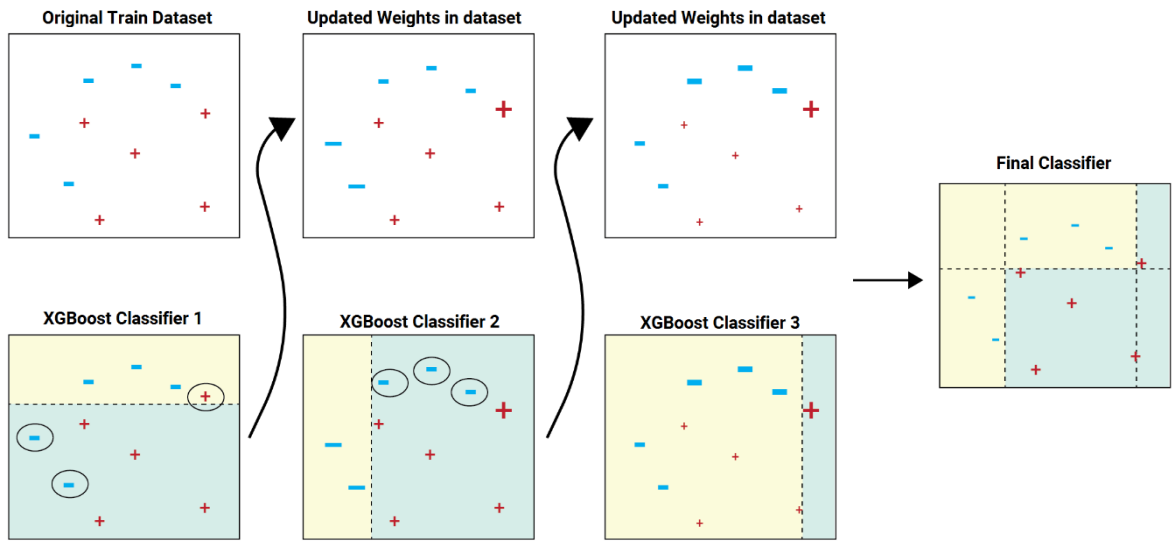


Fig 2.4

2.4.6 K Nearest Neighbours:

The **k**-nearest **neighbors** (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both **classification** and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

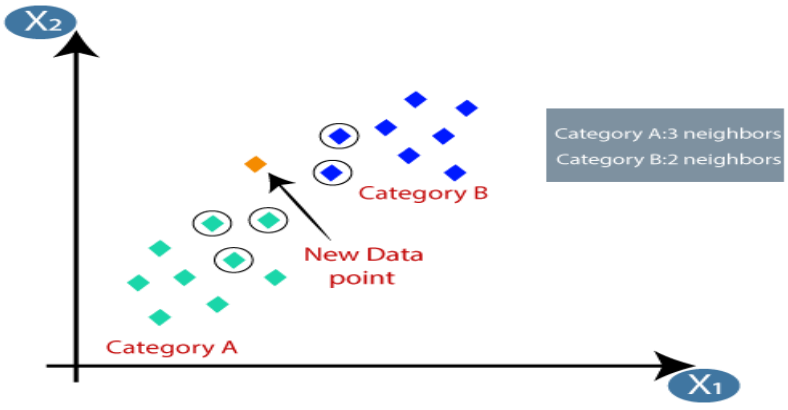


Fig 2.5

2.4.7 Support Vector Machine:

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary model (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

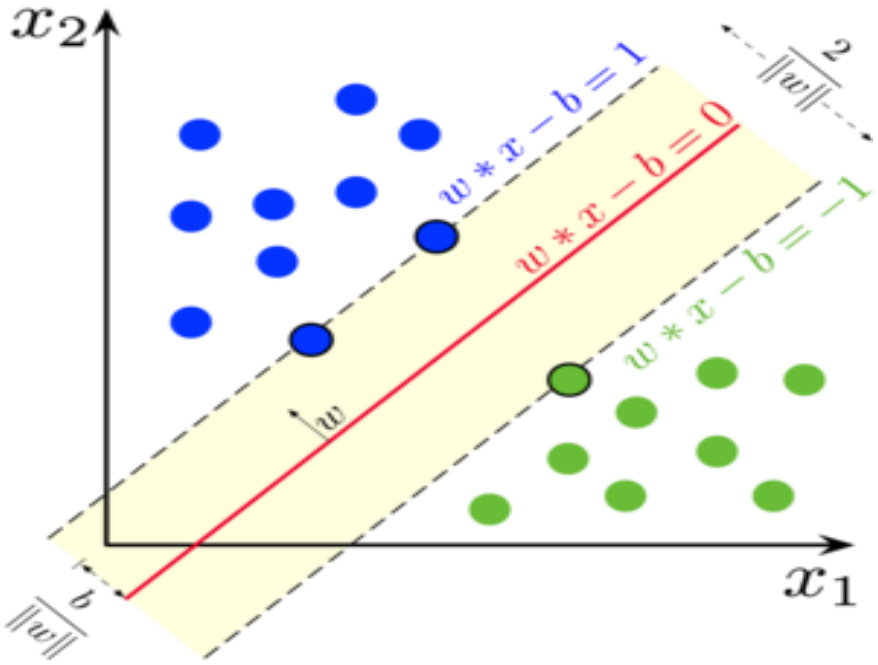
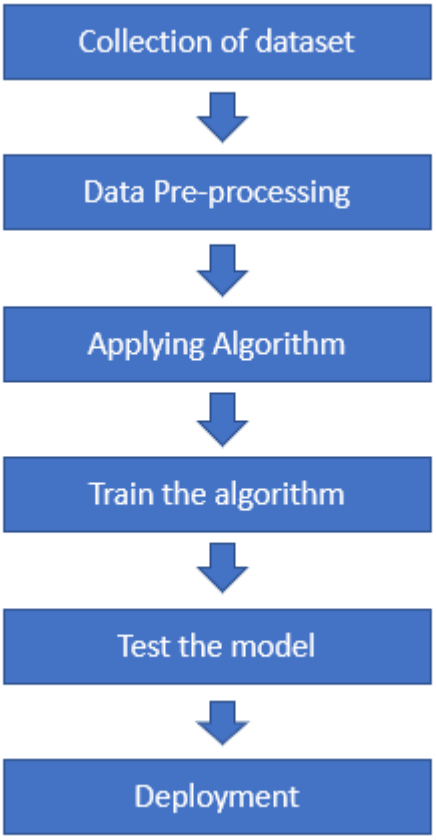


Fig 2.6

Chapter 3

System Analysis & Design

3.1 Flowchart :



3.2 Analysis:

Seven machine learning approaches are applied on the test data to predict the loan approvals of loan requests. Python programming language is used to implement machine learning algorithms. For training 75 percent data is used and 25 percent data is used for testing. The prediction accuracy for different ML approaches is calculated and compared.

Chapter 4

Result & Discussion

4.1 OUTPUT:

- **SURYANSH TIWARI**

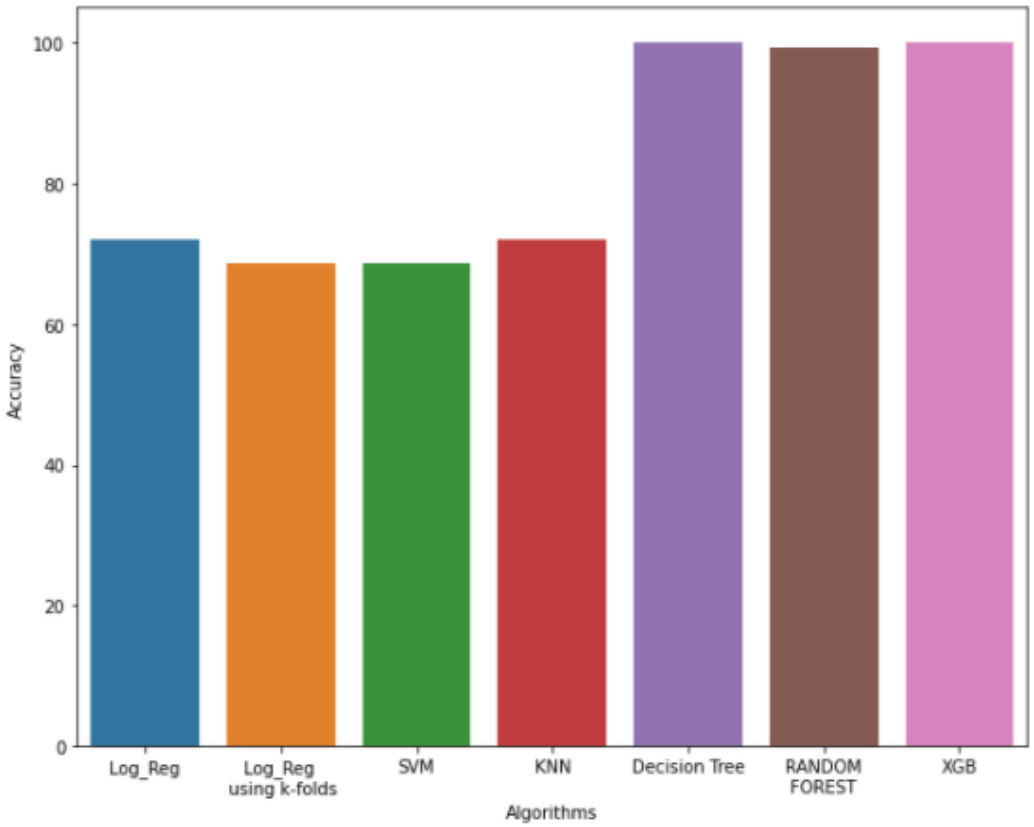


Fig 4.1

- **SWARN PALLAV BHASKAR**

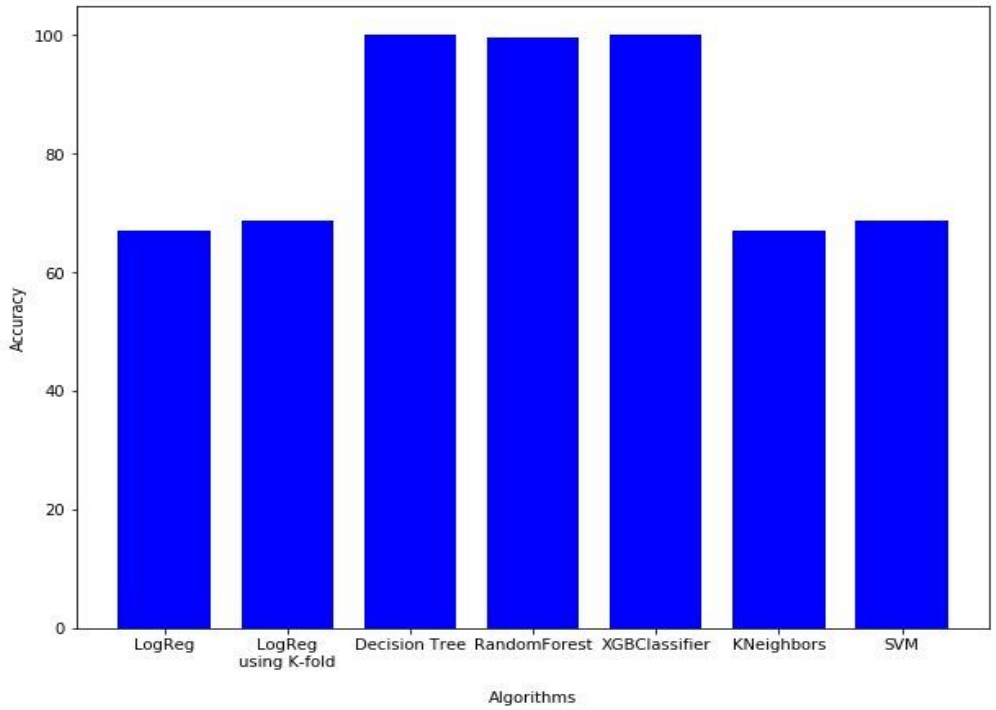


Fig 4.2

- **ANAND GUPTA**

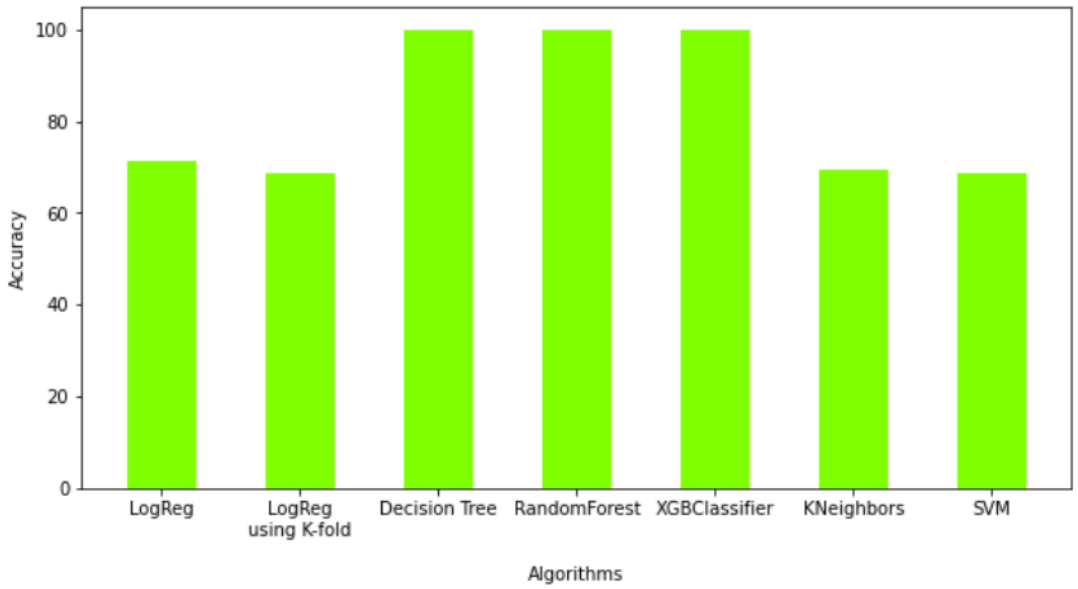


Fig 4.3

- **RITIK SINGH**

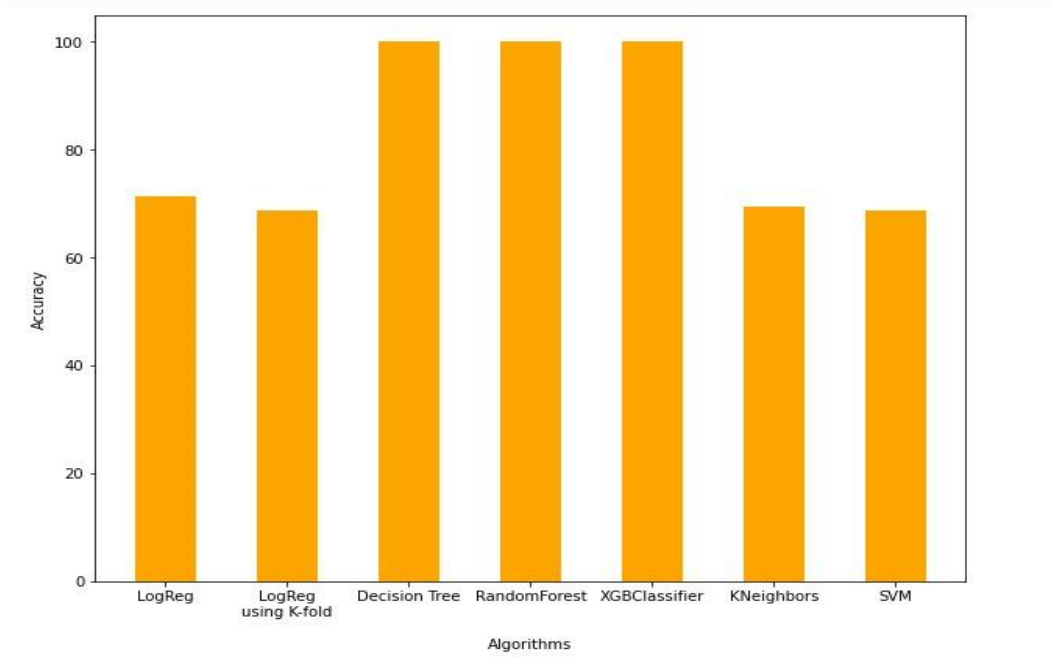


Fig 4.3

- **PRAKHAR GUPTA**

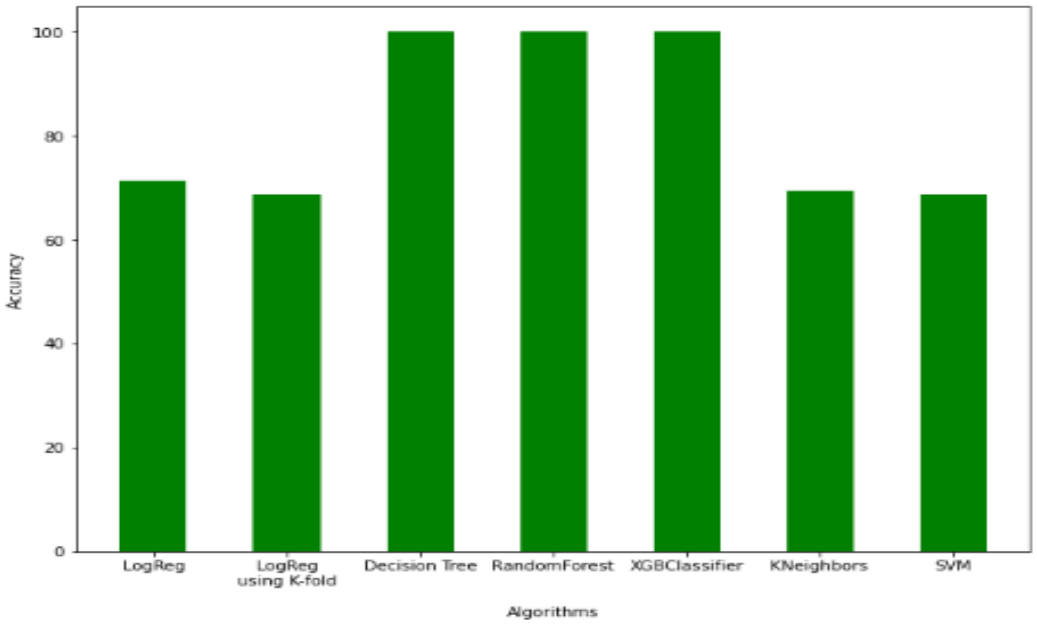


Fig 4.5

Chapter 5

Conclusion, Limitation & Future Scope

Banks fundamental business model rely on financial intermediation by raising finance and lending (mortgage, real estate, consumer and company's loans), the latter is the major source of credit risk composed from 2 main points loan approval and fraud. Granting loans to both retail n corporate customers based on credit scoring is key risk assessment tool that allow optimally managing, understanding and quantifying a potential obligator's credit risk through creditworthiness sore, which represent a more robust and consistent evaluation technique comparing to judgmental scoring.

Credit scoring in retail portfolios reflects the default risk of a customer at the moment f loan application, it helps to decide whether to accept or reject credit application base on few main input data: Customer information, Credit information, Credit History, Bank account behavioral etc.

Machine learning increases understanding by showing which factors most affect specific outcomes: Correlation matrix helps to dismiss correlated variable and feature selection methods (particularly Multivariate correlations) like stepwise regression are used to filter irrelevant predictors; it adds the best feature (or deletes the worst feature) at each round.

The only major limitation which remains is the infrastructure required for ML, however these resources are not huge still our banks lack it, with the advancements in technology we must gear ourselves up for all the roles and opportunities.

Still after so much improvement in technology, very personalized observation of each application is still a long way to go, which will be possible in years to come.

By more personalisation, it will even automate the fixing of interest rate, moratorium period etc. for public and hence it will help them in repaying the load more easily and will decrease the bank risk exponentially.

By predicting the loan defaulters, the bank can reduce its Non-Performing Assets.

Distribution of loans is almost core business part of all the banks. The main portion the bank's assets is directly came from the profit earned from the loans distributed by the banks. The prime

objective in banking environment is to invest their assets in safe/hands where it is. Today many banks/financial companies approve loan after a regress process of verification and validation but still there is no surety whether the chosen applicant is the deserving right applicant out of all applicants. Through this system we can predict whether that particular applicant is safe or not and the whole process of validation of features is automated by machine learning techniques.

Gone are the days when people use to wait for their application no. to come and their loan application be considered by the bank employee and the approval/disapproval used to depend on the persons will, not necessary on applicant's credit score which always promotes corruption.

With up skilling of digitalization in India, more and more sectors of the society will get its benefit. Lower sections of society who lack basic amenities will be able to get loans approved depending upon their credit score and this will contribute in uplifting their lifestyle.

References

- **https://towardsdatascience.com/data-visualization-using-matplotlib-16f1aae5ce70**
- **https://seaborn.pydata.org/tutorial/categorical.html**
- **https://www.kdnuggets.com/2020/07/easy-guide-data-preprocessing-python.html**
- **https://www.saedsayad.com/missing_values.htm#:~:text=A%20missing%20value%20can%20signify,way%20they%20treat%20missing%20values.**
- **https://towardsdatascience.com/data-preprocessing-with-python-pandas-part-1-missing-data-45e76b781993**
- **https://stackoverflow.com/questions/18689823/pandas-dataframe-replace-nan-values-with-average-of-columns**