### 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**

June 07, 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 my 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: June 07, 2021

## Certificate

This is to certify that the Project report entitled **“Credit Card Approval Model”** doneby **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: June 7, 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

The credit card was first used in the united states of America in 1920s. credit cards are issued by banks after the consumer opens the bank account and requested for the card. The credit card has its personal identification number (PIN) for payment by signing a receipt with card details electronic verification system allow to verify the credit card validity and customer should maintain sufficient amount of credit in his account then only he can have accessing permissions. Credit card processing is very efficient when compared to other cards.

With the increasing number of credit card applications, banks are opting towards the use of prediction-based algorithms as opposed to manual approval methods. Data analysis has exhibited a strong correlation between several financial and personal factors of a client and the likelihood of said client complying with their respective bank's credit policies.

A credit approval system requires deciding approval or rejection the application for supply credit cards based on some personal data of the applicant. In this paper, based on a data sample of 1000 applications from the past for credit card requests, we built a model to visualize the number of applicants who are either approved, rejected or their application is pending. In addition to this we have also visualized the booking amount of the corresponding applicant.

We have used matplotlib library of python to visualize the information from the data provided.

#### TABLE OF CONTENT

Declaration.................................................................................................. (ii)

Certificate ................................................................................................................. (iii)

Acknowledgement .......................................................................................................... (iv)

Abstract ................................................................................................................. (v)

Table of Content................................................................................................................. (vi)

List of Figures ……………………………………………………………………………...(vii)

**Chapter 1. Introduction ..........................................................................................** (viii)

* 1. Context…………………………………………………………………(viii)
  2. Motivations…………………………………………………..................(viii)
  3. Idea……………………………………………………………………..(viii)

**Chapter 2. Literature Survey…………………………………………………….. (ix)**

2.1 Preprocessing…………………………………………………………..(ix)

2.2 Missing Data Treatment……………………………………………….(ix)

2.3 Plotting Values on Graph for Visualization Purpose…………………..(x)

**Chapter 3. System Analysis & Design……………………………………………(xi)**

3.1 Flowchart………………………………………………………………(xi)

**Chapter 4. Result & Discussion…………………………………………………..(xii)**

4.1 Output………………………………………………………………….(xii)

**Chapter 5. Conclusion, Limitation & Future Scope……………………………(xxii)**

**References……………………………………………………………………………(xxv)**

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
|  |  | **Page No.** |
| **Figure 3.1** |  | (xi) |
| **Figure 4.1** |  | (xiii) |
| **Figure 4.2** |  | (xv) |
| **Figure 4.3** |  | (xvii) |
| **Figure 4.4** |  | (xix) |
| **Figure 4.5** |  | (xxi) |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# Chapter 1

## Introduction

* 1. **Context:**

This project has been done as a part of my Internship at SSJ IT Solutions Private Limited. Supervised by Mr. Kushagra Srivastava, I had three weeks to fulfill the requirements in order to succeed the module. Every week, a meeting was organized to show and report my progress and fix the next objectives.

* 1. **Motivations:**

Being extremely interested in everything having a relation with the Machine Learning, the independent project was a great occasion to give me the time to learn and confirm my interest for this field. The fact that we can make estimations, predictions and give the ability for machines to learn by themselves is both powerful and limitless in term of application possibilities. We can use Machine Learning in Finance, Medicine, almost everywhere. That’s why I decided to conduct my project around the Machine Learning.

* 1. **Idea:**

The idea is to build a model to visualize the number of applicants who are either approved, rejected or their application is pending based on a data sample of 1000 applications from the past for credit card requests. In addition to this we have also visualized the booking amount of the corresponding applicant.

# 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.

1. Preprocessing
2. Missing Data Treatment
3. Plotting values on graph for visualization purpose

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.



**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.



**2.4.4 Random Forest:**

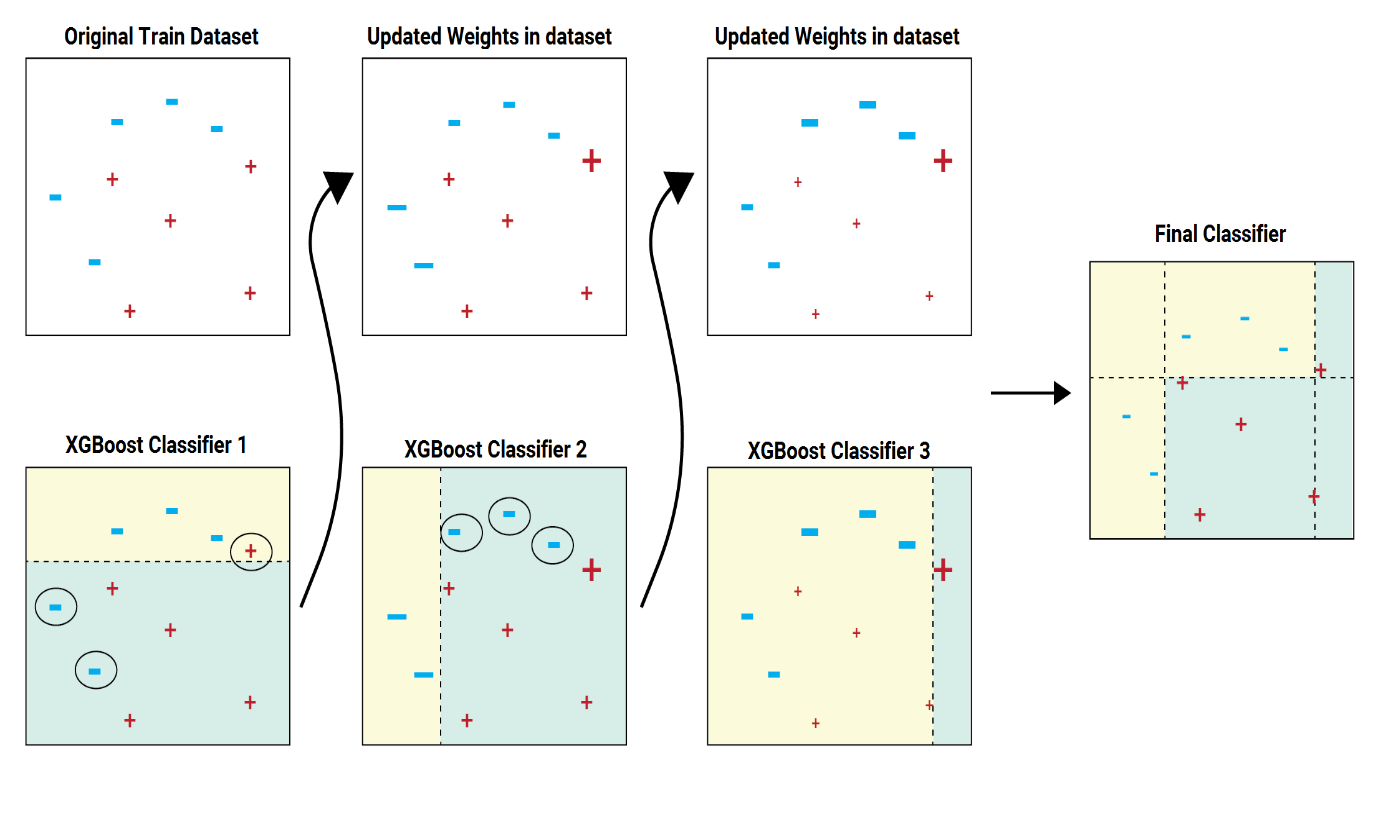
A random forest is a machine learning technique that’s used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

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.



# 2.4.5 XGB Classifiers:

XGBoostis 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.



**2.4.6 K-Neighbours:**

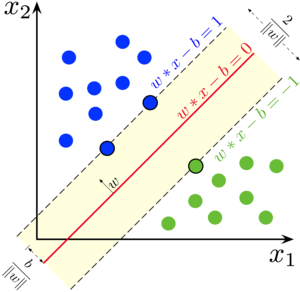
The **k**-nearest **neighbors** (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both **classification** and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use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

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



**2.4.7 Support Vector Machine:**

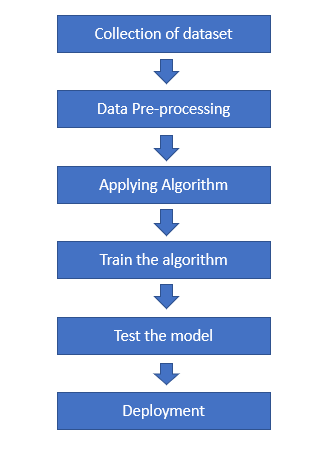
In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two 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.



# Chapter 3

## System Analysis & Design

**3.1 Flowchart :**



**3.2 Analysis:**

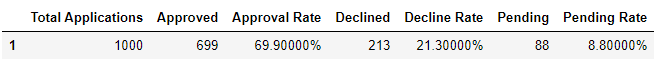
Seven machine learning approaches are applied on the test data to predict the loan approvals of loan requests. Python programming language is used to implement machine learning algorithms. For training 75 percent data is used and 25 percent data is used for testing. The prediction accuracy for different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

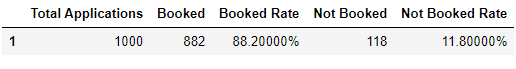
# Chapter 4

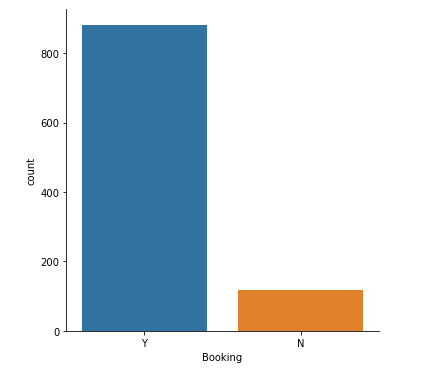
## Result & Discussion

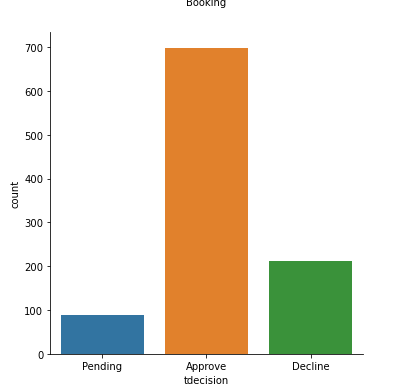
**4.1 OUTPUT:**

* **SURYANSH TIWARI**









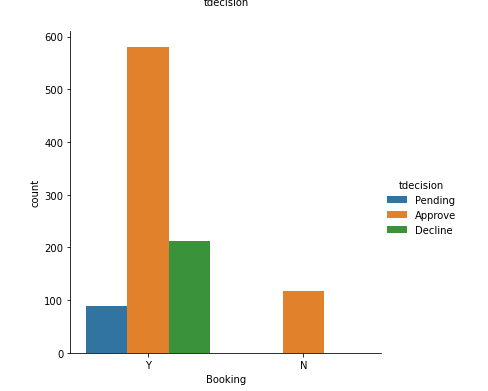
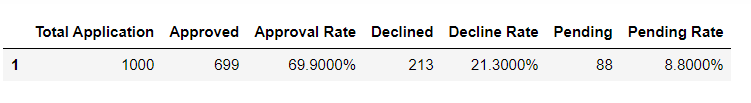
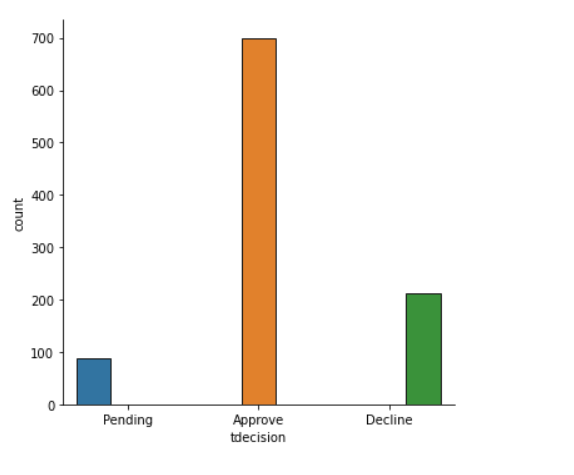
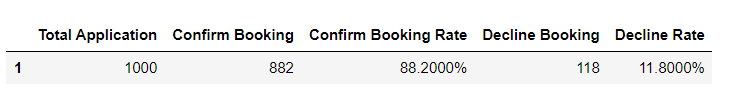


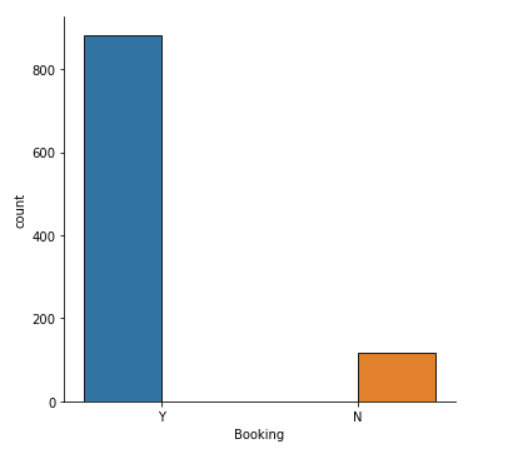
Fig 4.1

* **ANAND GUPTA:**









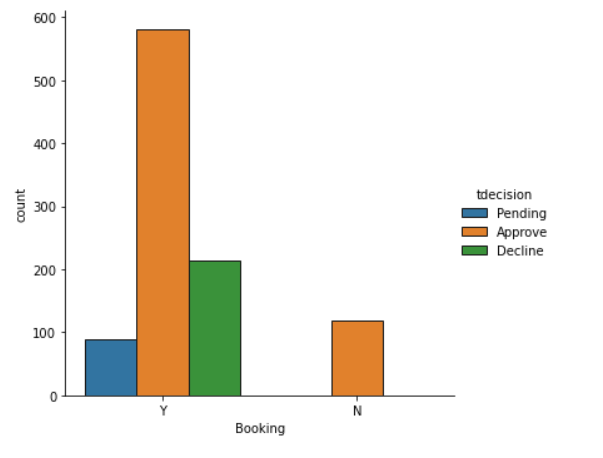
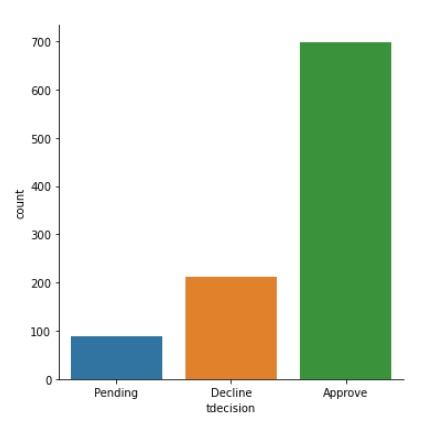


Fig 4.2

* **RITIK SINGH**







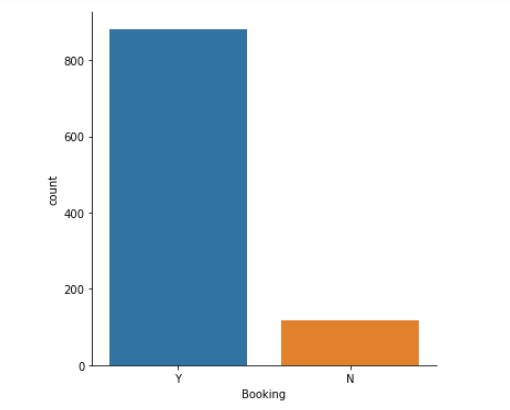
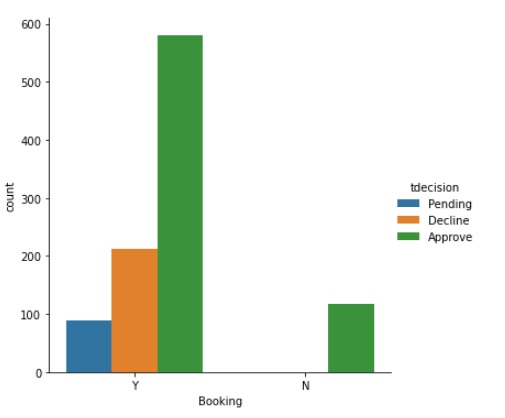
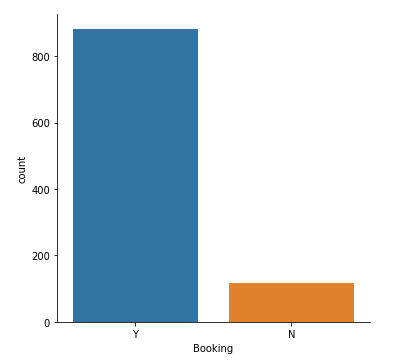


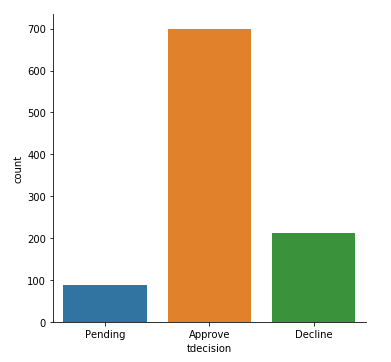
Fig 4.3

* **SWARN PALLAV BHASKAR**









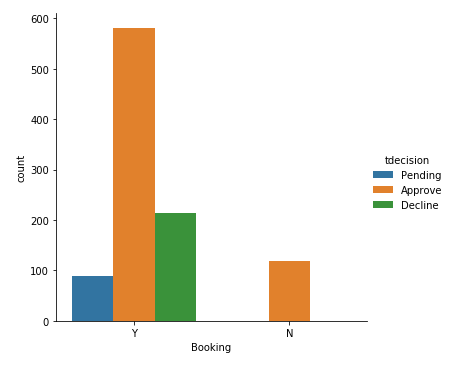
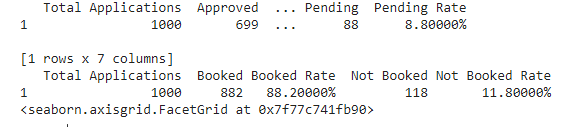
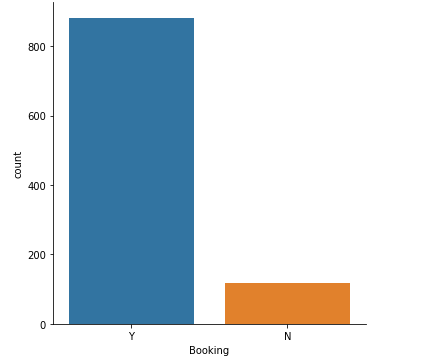
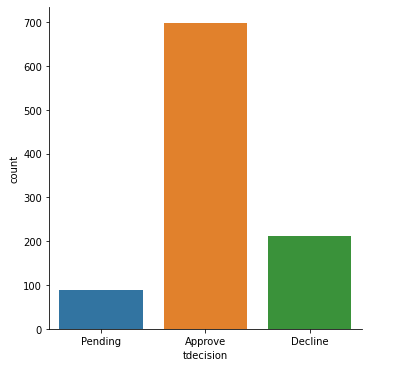


Figure 4.4

* **PRAKHAR GUPTA**







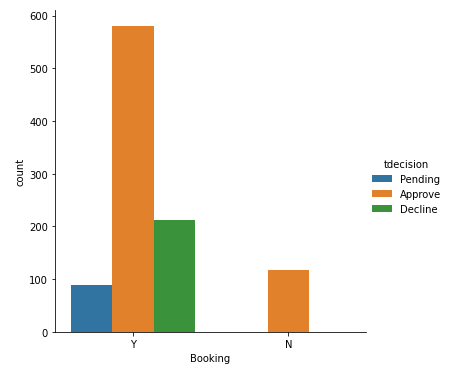


Figure 4.5

# Chapter 5

## Conclusion, Limitation & Future Scope

Using machine learning algorithm to preprocess data and automated decision taking can help credit card companies to a greater extent as there will be no need for human intervention. This will go a long way in making the most of the growth today as well as in redefining the future of approval system.

Fully automated system will make the process more transparent and will nullify the presence of bribery.

During the times of early computers i.e. 90s manual computation was the traditional method but now it has changed for good as involvement of human brain in decision making tasks may lead to corruption but in the age of computers that can be minimized.

**What happens today? -** Consumers have a lot of option for credit card companies and refusing the approval to a genuine request due to human error causes loss to company, now there are various machine learning algorithms which are upgrading day by day to help judge applicants on more accurate basis and decide the boolean answer.

The advancements in machine learning algorithms are helping such applications become more contextual.

Implementing machine learning is more feasible today because of the vast improvements in computing power, the abundance of data as well as storage and algorithmic advances. Like most other emerging technologies though, it is no panacea, and to get the most benefit, issuers need to build the right capabilities.

The only basic need that obstructs various capabilities of machine learning is the absence of robust digital foundation, especially big data analytics and cloud capabilities.

According to a Harvard Business review analysis, companies that have strong experience in digitization have a 50% higher chance of generating profit from AI implementation.

**Be cautious about bias:**

Decisions of machine learning algorithms can be prone to bias, one because they are designed by humans and second, because of the limited available test data. Any resulting bias, for example in the form a denial of credit card or inappropriate messaging may lead to regulatory sanctions and invite consumers’ backlash. Choosing the appropriate data samples that minimize bias is essential. Moreover, the insights generated by the system should be regularly monitored, especially because such systems learn over time and may offer unfavorable results in future.

In developing countries like India, consumers have made strong growth in credit card adoption while some consumers see credit cards as a way to extend their purchasing capacity.

The requests that banks receive are rejected for various reasons like high-loan balances, low-income levels, or too many inquiries on an individual’s credit report. Manually analyzing these applications is error-prone and a time-consuming process. Luckily, this task can be automated with the power of machine learning and pretty much every bank does so nowadays.

The best thing about machine learning algorithms is that it’s good to make better sense of exponentially increasing data like in our case and have greater opportunities in future as reserve because developments and research work for same is being carries out 24/7 in different countries so upcoming future will witness much more accurate algorithms and hence will directly affect the precision in automated decision making.

**New Opportunities-**

Machine learning capabilities can help banks open-up new monetization avenues. “Predictive spending insights", built on transaction and third-party data is one such example. Case in point is American Express’s AmexAdvance, that combines transactional and third-party data to brand marketers and media partners deliver personalization services.

**Roadmap to future-**

Develop an AI roadmap: Credit card issuers should clearly define the purpose, context and scope of why they would be using machine learning. In the short-term, issuers can focus on proven application areas such as fraud detection and over a period of time look at relatively more advanced use cases such as automated rewards recommendation. Moreover, issuers can also work with academia and technology partners to identify more futuristic use-cases.

**The Problem-**

Issuing credit card involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting.

This is a very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated.

This problem is particularly challenging from the perspective of learning, as it is characterized by various factors such as class imbalance. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over the course of time.

These are not the only challenges in the implementation of a real-world credit card approval system, however. In real world examples, the massive stream of such requests is quickly scanned by automatic tools that determine which transactions to authorize.

Overall, there is a need for research works in this field to maintain the benefit of automation in times where fraudulent ways to hamper the true decision making are increasing day by day.

This automated setup has boosted the efficiency of banks and other private companies to a much greater extent.

**References**

* [**https://towardsdatascience.com/data-visualization-using-matplotlib-16f1aae5ce70**](https://towardsdatascience.com/data-visualization-using-matplotlib-16f1aae5ce70)
* [**https://seaborn.pydata.org/tutorial/categorical.html**](https://seaborn.pydata.org/tutorial/categorical.html)
* [**https://www.kdnuggets.com/2020/07/easy-guide-data-preprocessing-python.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://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://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**](https://stackoverflow.com/questions/18689823/pandas-dataframe-replace-nan-values-with-average-of-columns)