Project Report

On

Credit Card Approval System

Submitted in partial fulfillment for the award of

Post Graduate Diploma in Big Data Analytics (PG-DBDA)

From Know-IT(Pune)

Guided by:

Mr. Prasad Deshmukh.

Mr. Anay Tamhankar.

Submitted By: -

Swapnil Joshi (230343025022)

Kartik Hinge (230343025025)

Vinayak Mandlik (230343025032)

Shivam Soam (230343025046)

CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

This is to certify that

Swapnil Joshi (230343025022)
Kartik Hinge (230343025022)
Vinayak Mandlik (230343025022)
Shivam Soam (230343025022)

Have successfully completed their project on

Credit Card Approval System

Under the guidance of Mr. Prasad Deshmukh & Mr. Anay Tamhankar

ACKNOWLEDGEMENT

This project **"Credit Card Approval System"** was a great learning experience for us and we are submitting this work to CDAC Know-IT (Pune).

We all are very glad to mention the name of **Mr. Prasad Deshmukh** and **Mr. Anay Tamhankar** for his valuable guidance to work on this project. His guidance and support helped us to overcome various obstacles and intricacies during project work.

We are highly grateful to **Mr. Vaibhav Inamdar** Manager (Know-IT), C-DAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS, Pune.

Our most heartfelt thanks go to **Mrs. Bakul Joshi** (Course Coordinator-DBDA) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility and extra Lab hours to complete the project and throughout the course up to the last day here in C-DAC Know-IT, Pune.

ABSTRACT

This Project is focusing on application of machine learning techniques to predict customer eligibility for a credit card to mitigate possible future credit risk which may affect the bank's financial stability and credit performance. Credit card is a credit facility given for a customer by banks and finance companies around the globe. The credit facility has a credit risk for the banks and financial companies. The repayments are least assured and it often ends up as a non-performing credit facility (NPL). To mitigate credit risk banks are assessing applicant's creditworthiness and checking the eligibility before granting a credit facility. The decision is mostly based on traditional credit scoring models and credit worthiness will not always be accurate. This project aims to help banking and financial institutions to identify and interact with creditworthy customers by using predictive models. We used Logistic Regression, Decision Tree Classifier, Random Forrest Classifier, Support Vector Mechanism (SVM) and XGBoost Model to develop models. The XGBoost algorithm achieved the highest accuracy of 82.30 % in predicting credit card approval, followed by Random Forest Algorithm and Decision Tree Model with an accuracy of 72.00% and 66.71%, respectively. SVM had the lowest accuracy of 50.26%. Feature importance analysis revealed that the most important features in predicting credit card approval were gender, total income, education type and years employed. These features had a significant impact on the credit card approval decision, and should be considered when evaluating creditworthiness of applicants. This project uses a combination of data analytics, machine learning, and predictive modelling techniques to analyses credit data and identify patterns that can improve the accuracy of credit card approval predictions. We also realized that customer behavior might be different from country to country and application of several real banking datasets not limited to customer demographic and socio- cultural but also other credit facility features including COVID-19 impact to be an area of concern for researchers. Furthermore, whether there is a relationship between Nonlinearity in highly imbalanced class problems with SMORTE application is another area of concern for researchers.

TABLE OF CONTENTS

1. INTRODUCTION

This research is focusing on application of machine learning (ML) techniques to predict customer eligibility for a credit card. One of key objective of the bank is to increase the returns. When increasing the returns there is an increase of risk. Banks are faced with various risks such as interest rate risk, market risk, credit risk, off-balance-sheet risk, technology and operational risk, foreign exchange risk, country or sovereign risk, liquidity risk, liquidity risk and insolvency risk. Effective management of these risks is key to a bank's performance. Credit can be defined as the risk of potential loss to the bank if a borrower fails to meet its obligations (interest, principal amounts). Continuously monitoring of customer payments could reduce the probability of accumulating non-performing assets (NPA). Whether to grant or not to grant a loan to a customer is one of key decisions of banks use to reduce probable NPA at the first hand. Credit card as a credit facility instruments banks need to effectively managed credit risk of the facility. The Basel Accord allows banks to take the internal ratings-based approach for credit risk. Banks can internally develop their own credit risk models for calculating expected loss. There are several manual steps involving when granting a credit card to a customer. Assessing applicant's creditworthiness and checking the eligibility are the key factors and decisions the bank would take about a credit worthiness will not always be accurate. Application of machine learning techniques can eliminate manual paperwork, time-consuming processes and most importantly data driven decision making before granting a credit card to a customer. In this research, different supervised machine learning algorithms were used to develop models and follow the steps in cross-industry standard process for data mining (CRISP-DM) life cycle. Accuracy of models was validated by using different validation techniques.

1.1 Motivation

In times of yore, when providing a credit card to a customer, banks had to rely on the applicant's background and the history to understand the creditworthiness of the applicant. The process includes scrutinization of application data with reference documents and this process was not always accurate and customers and the bank had to face difficulties in approving the credit card. But with the digital transformation, there is a growth in Artificial Intelligence & Machine Learning Technology in the past two decades. Therefore, ML techniques being used to evaluate credit risk and automate credit scoring by predicting the customer eligibility correctly using customer demographic data and historical transactional data. Furthermore, ML helps banks to make smarter data –driven decisions for customers; use banking data in a more productive and efficient way; streamline customer interaction by removing manual and lengthy processes.

1.2 Background of the Study

Commercial banks contribute to economic growth in various aspects. One of the biggest revenue streams of any banking or financial institution would be from the interest charged from the lending. Banks must face the biggest credit risk in all their lending. There are various lending products the banks are offering to the customers. However, Credit cards are one of the key lending products any bank would ever have. Almost all the financial institutions across the globe are going through challenging time and credit risk in offering credit facilities to their end customers. The repayments are least assured and it often ends up as a non-performing credit facility (NPL). This will in return affect banks cash flow and leads to build up backlogs in balance sheet which will not look good if the bank is a listed organization. Banks and financial institutions are critically assessing eligibility for a credit facility before granting facility to the customer due to the credit risk factor the credit card involved in. This process involves verification, validation, and approval and may cause delay of granting a facility which will be disadvantageous for the applicant as well as for the bank. Credit officers determine whether the borrowers can fulfill their requirements to being eligible for a facility and these judgments and predictions are always not accurate. Credit scoring is a traditional method assessing the credibility of a customer / entity applying for a bank credit facility. How much ever the banks and financial institutions are doing the background check of the individual customers by analyzing their eligibility, the bank most of the time end up in making wrong decisions. The study determines whether an Artificial Intelligence system using Machine Learning Technology can assist the industry in overcoming from this risk.

1.3 What Is a Credit Card?

Credit card is a credit facility given for a customer by banks and finance companies. It has a higher annual percentage rate (APR) than other consumer loans. By law, card issuers must provide 21 days of grace period before interest on purchases and begin to accrue. When customers paying off balance before the grace period expired consider as a good practice. Interest charges will begin for any unpaid balance typically after one month of purchase is made. In case of any unpaid balance left it had been carried forward from a previous month and for new charges there is no grace period provided. Interest will be accruing daily or monthly according to issuer interest and the country's financial policies. Credit card will be entered to delinquent state if the customer failed to paid minimum monthly amount for 30 days from original due date. Most of financial institutes start to reaching customers when customer card status become past due. After 60 days or more delinquent status become overdue and most companies involve in taking legal actions to start debt collection (Fernando, 2021).

1.4 Component of a Credit Card

Figure 1.1 - Component of a Credit Card

1.5 Types of Credit Cards

Most popular credit card networks/brands are Visa, MasterCard and American Express. These cards were issued by banks and financial institutions. Different types of credit cards categories are in a particular brand as well such as for low net worth, medium net worth, and high net worth customers. To attract more customers, different incentives are offering such as airline miles, hotel room booking, restaurant dine-in, super market grocery buying, gift certificates to major retailers and cash back on purchases. Furthermore, in some banks have established rewards system for credit card usage. At the end of year these rewards points can be redeemed. Branded versions of credit cards are issued to generate customer loyalty with store's name/ organization name emblazoned on the face of the cards. These credit cards called co-branded credit cards.

1.6 Credit Line

A line of credit (LOC) is a stipulated amount of money that a card issuer has agreed to lend for a customer at the beginning of credit card account opening. Until the limit is reached, the borrower can draw money from the credit card and as money is repaid, it can be borrowed again in the case of

an open line of credit. Credit line can be increase after evaluating customers' repayment capacity later.

1.7 Credit Card Issuing Process

Before providing a credit card to the customer there is a process to establish a relationship with customer and the bank. Applying for a credit card for first time can be time consuming. Filling out an application form is mandatory and most bank nowadays allow to apply online by filling an application form. Choosing of suitable card can be done after self-studying or consulting sales executives. Figure 1.2 illustrated credit card issuing process as below.

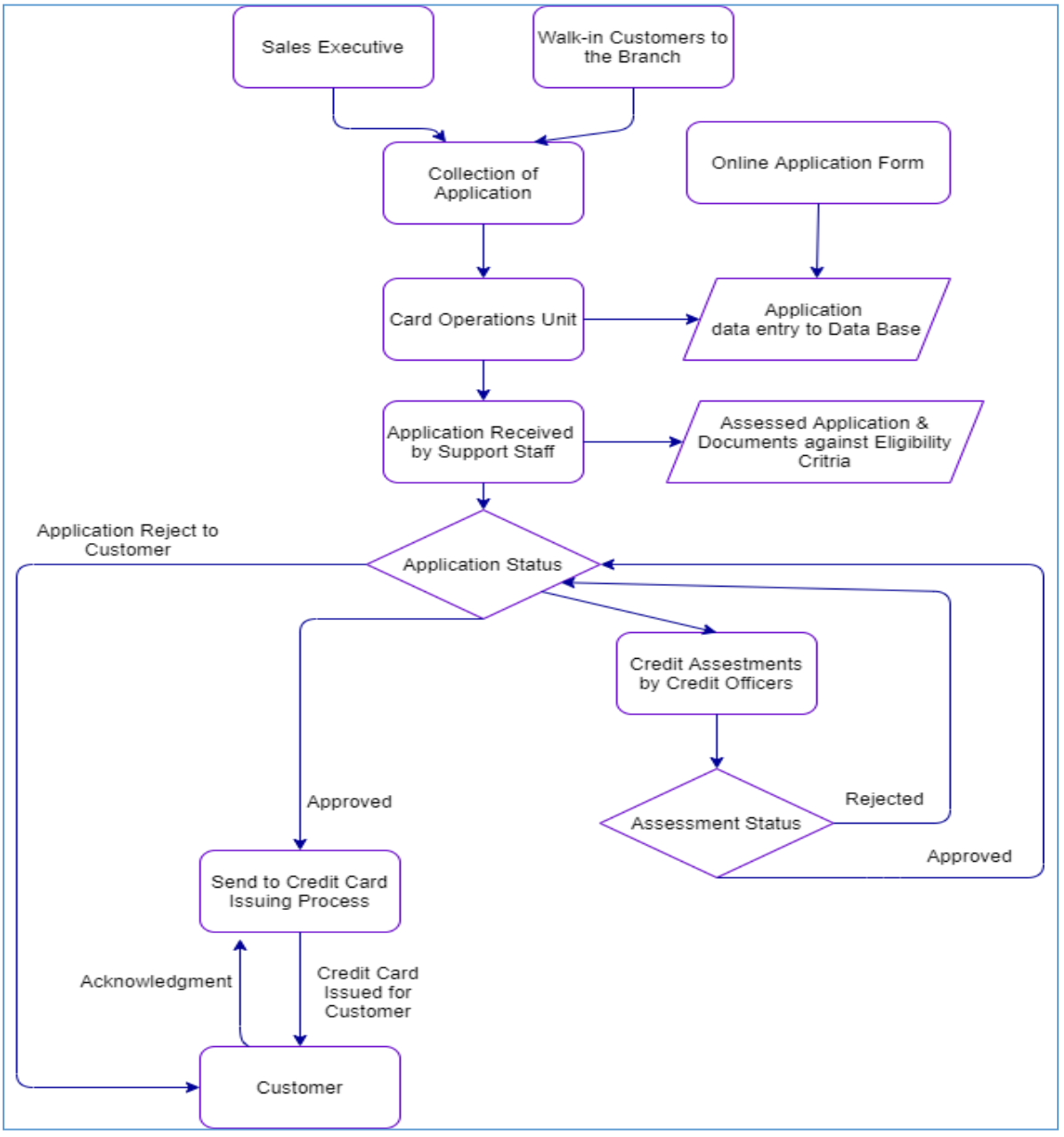

Figure 1.2 - Credit Card Issuing Process

2. METHODOLOGY

2.1 Systematic Approach

To carry out the project, CRISP-DM frame work was used as shown in Figure 3.1 and detail discussion of each phase relevant to application for project is listed below.

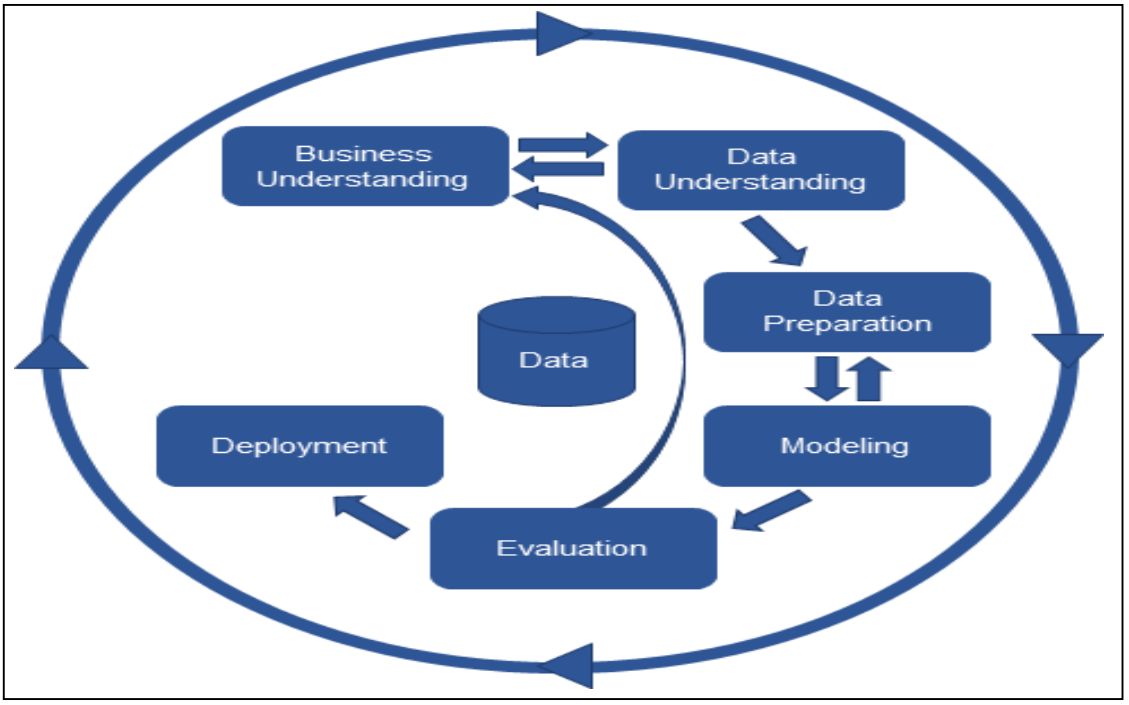

Figure 2.1 - CRIP –DM Model

CRISP-DM (Cross-industry standard process for data mining) data mining process was published in 1999 to standardize. There 6 phases, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and deployment. Brief description of each phase is listed below (Data Science Process Alliance, n.d.).

- **Business Understanding** - Understanding of objectives and requirements and produce of detail plan for project focus in here.
- **Data Understanding** - Focusing on identify, collect and analyze data. Data format/fields identification, identify relationships by visualization; verify data quality (clean/dirty) are the main activities carried during this phase.
- **Data Preparation** – This phase often called ‘data munging or wrangling. Selection of data, clean data, construct data, integrated data and format data are basic activities carried out under this phase.

- **Modeling** – Determine selection of algorithms, generate test design, build model and asses model are main activities carried out in this phase.
- **Evaluation** – Focusing on identification of which model best fit the business requirement. Evaluate results, review process, determine next step are key activities in here. By determining whether to proceed to deployment or iterate further will be judge in here.
- **Deployment** – Focusing on accessible methods for developed model output/results. Deployment plan, monitoring and maintenance, produce final report and review project are key activities in here.

2.2 Business Understanding

Credit card is one of the key lending product facilities given for a customer by a bank. The repayments of credit card are always not guaranteed and it often ends up as non-performing credit facility (NPL). Banks are assessing the background check of the individual customers by analyzing their eligibility, yet the bank sometime end up in making wrong selections. The credit card has a higher annual percentage rate (APR) and by law, card issuers must provide 21 days of grace period before interest on purchases and begin to accrue. When customers paying the balance before the grace period expired consider as a good practice. For any unpaid balance normally after one month of purchase is made Interest charges will apply. Any un paid balance carried forward from previous month and for new charges grace period will not be provided. According to country's financial policy interest will be accruing daily or monthly.

3.FUNCTIONAL REQUIREMENTS

3.1 Tool Used

- **Python 3: -** Python is a general purpose and high-level programming language. It is use for developing desktop GUI applications, websites, and web applications. Python allows to focus on core functionality of the application by taking care of common programming tasks. Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, Small Talk, and Unix shell and other scripting languages.
- **Spark: -** Spark is an open-source distributed computing framework that provides a simple and efficient way to process large datasets. In this project, Spark was used to handle and analyze large amounts of data related to credit card applications. Spark was used for tasks such as data cleaning, data transformation, and feature engineering.
- **MLlib: -** MLlib is a library for machine learning in Spark that provides scalable implementations of common machine learning algorithms. In this project, MLlib was used to build and train machine learning models to predict credit card approval.
- **Tableau: -** Tableau is a data visualization and business intelligence tool that allows you to create interactive dashboards, reports, and charts. In this project, Tableau was used to create visualizations and reports to communicate the project's progress and results.

3.2 Data Understanding

The data set has been taken from kaggle.com data repository. This data set is publicly available data set. Hence information security is not a concern in here.

URL - https://www.kaggle.com/rikdifos/credit-card-approval-prediction/tasks?taskId=1416

There are two data set and detail of data shown in Table 3.1 and 3.2.

- application_record.csv for applicant information – No of records 4,38557
- credit_record.csv for credit record information – No of records 1,048575

Application record file has nine categorical variables and nine numerical variables as shown in figure 3.2. According to the figure there are null data in occupation type column. The data set does not contain direct class variable.

Credit Card Application Data			
Feature Name	**Explanation**	**Data Type**	**Possible Values**
ID	Client number	Numerical	
CODE_GENDER	Gender of the client	Categorical	M, F
FLAG_OWN_CAR	Is there a car	Categorical	N , Y
FLAG_OWN_REALTY	Is there a property	Categorical	N, Y
CNT_CHILDREN	Number of children	Numerical - Integer	
AMT_INCOME_TOTAL	Annual income	Numerical - float	
NAME_INCOME_TYPE	Income category	Categorical	Commercial associate Pensioner, State servant Student, W o r k i n g
NAME_EDUCATION_TYPE	Education level	Categorical	Academic degree, Higher education, Incomplete higher, Lower secondary, Secondary / secondary special
NAME_FAMILY_STATUS	Marital status	Categorical	Civil marriage, Married, Separated, Widow Single / not married,
NAME_HOUSING_TYPE	Way of living	Categorical	Co-op apartment, House / Apartment, Municipal apartment, Office apartment Rented apartment, With parents
DAYS_BIRTH	Birthday	Numerical - Integer	Count backwards from current day (0), -1 means yesterday
DAYS_EMPLOYED	Start date of employment	Numerical - Integer	Count backwards from current day (0). If positive, it means the person currently unemployed.
FLAG_MOBIL	Is there a mobile phone	Numerical - Integer	1, 0
FLAG_WORK_PHONE	Is there a work phone	Numerical - Integer	1, 0
FLAG_PHONE	Is there a phone	Numerical - Integer	1, 0
FLAG_EMAIL	Is there an email	Numerical - Integer	1, 0
OCCUPATION_TYPE	Occupation	Categorical	Several occupations
CNT_FAM_MEMBERS	Family size	Numerical - Float	

Figure 3.1- Detail Information about application data set

Credit Card Record (Payment History Data)			
Feature Name	**Explanation**	**Remarks**	
ID	Client number	Numerical - Integer	
MONTHS_BALANCE	Record month	Numerical - Integer	The month of the extracted data is the starting point, backwards, 0 is the current month, -1 is the previous month, and so on
STATUS	Status	Categorical	0: 1-29 days past due 1: 30-59 days past due 2: 60-89 days overdue 3: 90-119 days overdue 4: 120-149 days overdue 5: Overdue or bad debts, write-offs for more than 150 days C: paid off that month X: No loan for the month

Figure 3.2 - Detail Information about payment history data set

3.3 Data Preparation Methods

Data preparation phase, which is often referred to as "data munging" or "Data Preprocessing" prepares the final data set(s) for modeling. Python programing with libraries /packages use to prepare the data set.

Key main areas related to data preparation phase considered in the project as follows.

- Data Preparation with Explanatory Data Analysis (EDA) under each preparation activity
- Feature Selection from finally prepared data set

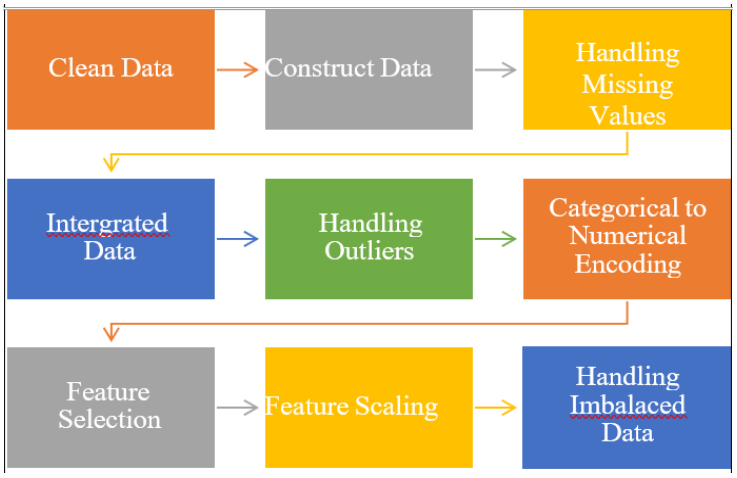

Figure 3.3 - Activities in Data Preparation Phase

3.4 Clean Data

Data set might contain erroneously entered data. These erroneous values need to correct, impute, or removed from the data set

Figure: -3.4 Data Cleaning Process

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data. In our dataset ID column contains white spaces. Joining two data set with white spaces does not give correct aggregation. Removed white spaces from ID column in both data set. ID column converted as string column. DAYS_EMPLOYED column count backwards from present day (0). Values contains negative and positive both. Positive mean the person currently unemployed. Positive data values are set to 0 and convert negative values to positive value by multiplying -1 to bring into standard format.

3.5 Handling Missing Value

Missing value occurred may be due to many reasons. By handling missing values, it will increase performance of the model. Common methods are replacing missing values with mean or median of entire column (imputation) or deleting rows/ columns. In this data set there are missing values in OCUPATION_TYPE column and hence this is a categorical column replaced missing values with 'Other.'

3.6 Construct Data

In general, constructing data involves creating a structured and organized set of information that can be used for analysis, processing, and other purposes. Here we derive new attributes from exiting data set. This data set does not contain direct class label. Derived a variable from applicant information data set as customer is good or bad by using credit card payment history details. By using customer's credit history data dependent variable was generated. There are 8 different payment status and they are listed below.

- C: paid off that month
- X: No loan for the month
- 0: 1-29 days past due
- 1: 30-59 days past due
- 2: 60-89 days overdue
- 3: 90-119 days overdue
- 4: 120-149 days overdue
- 5: Overdue or bad debts, write-offs for more than 150 days

After 60 days or more customers' delinquent status become overdue and most companies involve in taking legal actions to start debt collection. Hence, the overdue status codes 3, 4, and 5 consider as bad customer status. As shown in table 4.1, customers who are paid off that month, those who have not taken loan for the month, those who are not paid and past due for 1 to 29 days and 30-59 days past due consider as good customers. Payments are past due from 60^{th} day and above consider as defaulters to minimize class imbalance problem. After selecting default codes new column "Final Label" created as 0 is for a good customer and 1 is for bad customer considering below mentioned logic. Further derived two new columns from DAYS_BIRTH and DAYS_EMPLOYED as AGE_IN_YEARS and EMPLOYED_IN_YEARS by dividing 365 (making to years) because of days are make less sense in business domain.

3.7 Integrating Data

Integrating data refers to the process of combining and merging data from multiple sources to create a unified and consistent view of the data. The purpose of integrating data is to make it easier to access and analyze the data and to provide a completer and more accurate picture of the information. Integrating data phase basically combined data from multiple sources.

In here we have two data set combined. Below figure 3.7 illustrated combining of two data set.

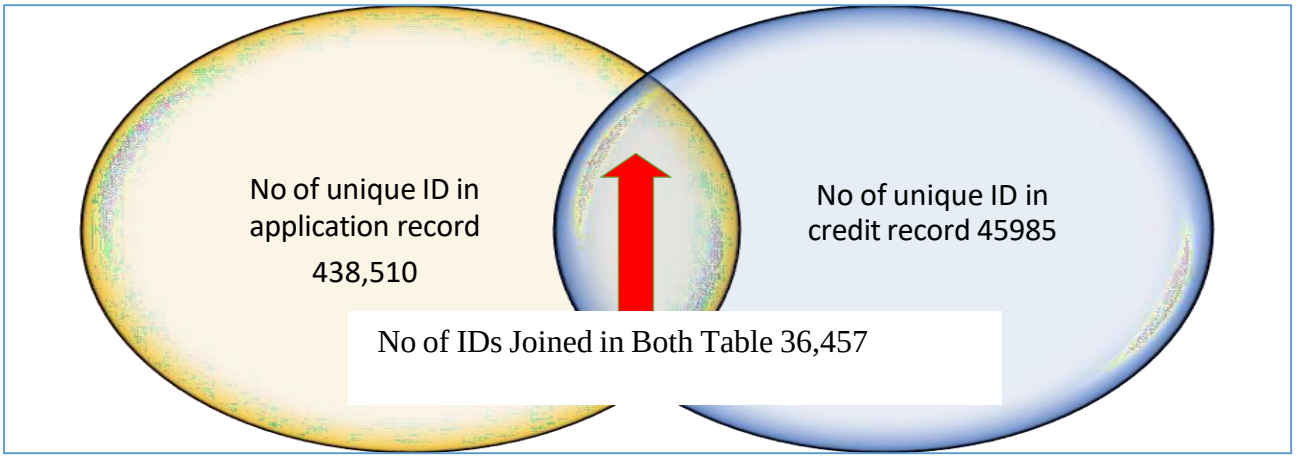

Figure 3.5 - Selected Data Set

Credit record data sets and application record data sets are merged to generate final data set. After identification of class label there are 615 bad customers and 35841 good customers. Final data set consider for model building is 35841.

3.8 Outlier Removals

Outlier is a data point that differs significantly from other observations. To remove outliers, we can use statistical method Inter Quartile Range (IQR) and removed outliers from the data set. Final data set is 33,140. Figure 3.5 shown main component of IRQ.

below 25th percentile – 1.5 * IQR, or above 75th percentile + 1.5 * IQR

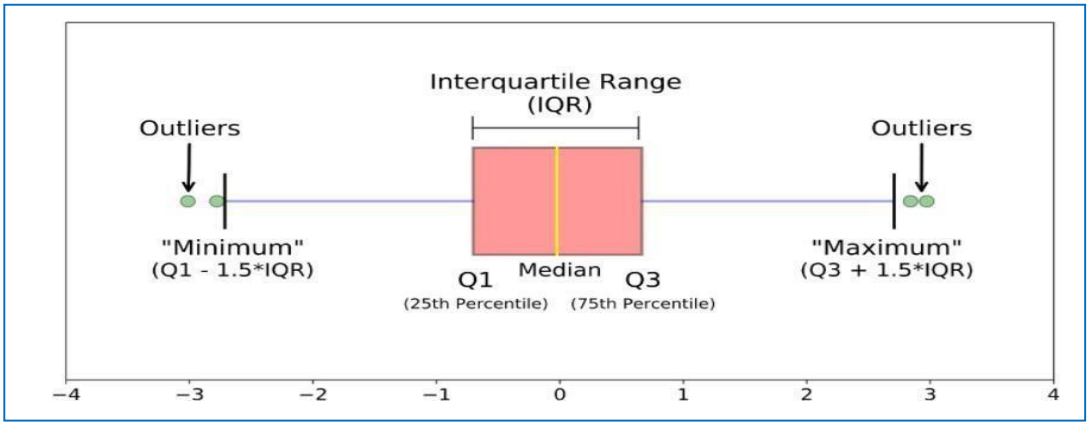

Figure 3.6 - Interquartile Range

3.9 Encoding Categorical Data

Categorical data cannot be use at mathematical equations. Such as ‘Male’ and ‘Female’ in gender column. These columns need to convert to numerical values. There are varies methods can apply for categorical encoding by considering categorical feature is ordinal or nominal.

3.10 Feature Selection

Feature selection, also known as variable selection, is a process of selecting a subset of relevant and important features (or variables) from a larger set of features in a dataset. The goal of feature selection is to reduce the dimensionality of the data and improve the accuracy and efficiency of the analysis.

3.11 Correlation Based Feature Selection

Correlation is a bi-variate investigation and its asset association between two variables and the way of the relationship. Correlation coefficient value varies between +1 and -1. A value of + 1 shows a perfect relationship between the two variables. Relationship of two variables will weaker when correlation coefficient value goes towards 0. The + sign indicates a positive relationship and a - sign indicates a negative relationship. Main difference between each correlation methods explained as follows (Statistics Solutions, n.d.)

- **Pearson Correlation Coefficient –** To measure relationship between linearly related variables Pearson r correlation is the most widely used. Both variables should be normally distributed (normally distributed variables have a bell shape curve).
- **Spearman Rank Correlation** - To measure the degree of association between two variables Spearman rank correlation is use. Spearman rank correlation is a non-parametric test. (Statistics Solutions, n.d.) mentioned that “It does not carry any assumptions regarding the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal.”
- **Information Gain** - Information Gain sometimes denoted as Mutual Information measure the dependence between the two variables. It measures information value of each independent variable respect to dependent variable and select the one has most information gain. The variable considers as more dependent when the information gain value is high.

3.12 Handling Imbalance Data

Class imbalance is common problem in machine learning which is inherited by nature for default prediction, customer churning etc. Class imbalance is number of observations belong to one class is significantly lower than the other class.

To balanced data, we can use over sampling or under sampling. Both methods might be cause model over fitting unless if we use correct technology.

- Oversampling happened when adding more copies of the minority class. Oversampling can be a correct choice when you have less data.
- Under sampling happened when removing some observations of the majority class. Under sampling can be a correct choice when you have large data set such as millions of rows. Figure 3.7 shown oversampling and under sampling.

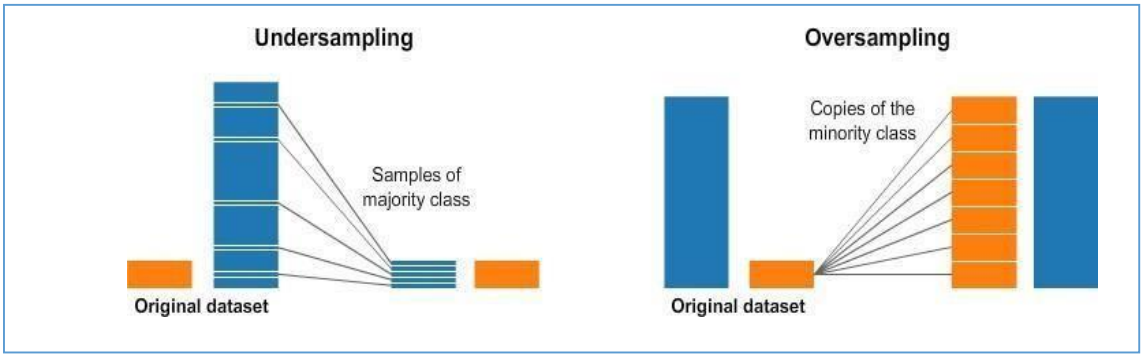

Figure 3.7 - Oversampling and Under Sampling.

4. Modeling

We have acquired relevant data set and data preparation with feature selection was done and finalized our data set. Then applied standard scaler to numerical data for data scaling and apply SMORTE for finalized data set. Next step is to divide the data set as a training and test into a ratio of 80:20. Training data set is used to train the model by applying SVM. In here use linear and nonlinear SVM both models. Python programming and its libraries have been used to develop the models. Finally evaluate the predicted results of Random Forest Classifier, Decision Tree Classifier, Logistic regression & SVM. Then compare the accuracy of these models by using Mean Squared Error and Confusion Matrix to choose the most accurate model. Test data set used to test the model and evaluate the outcome. Workflow of the modeling process shown in figure 4.1.

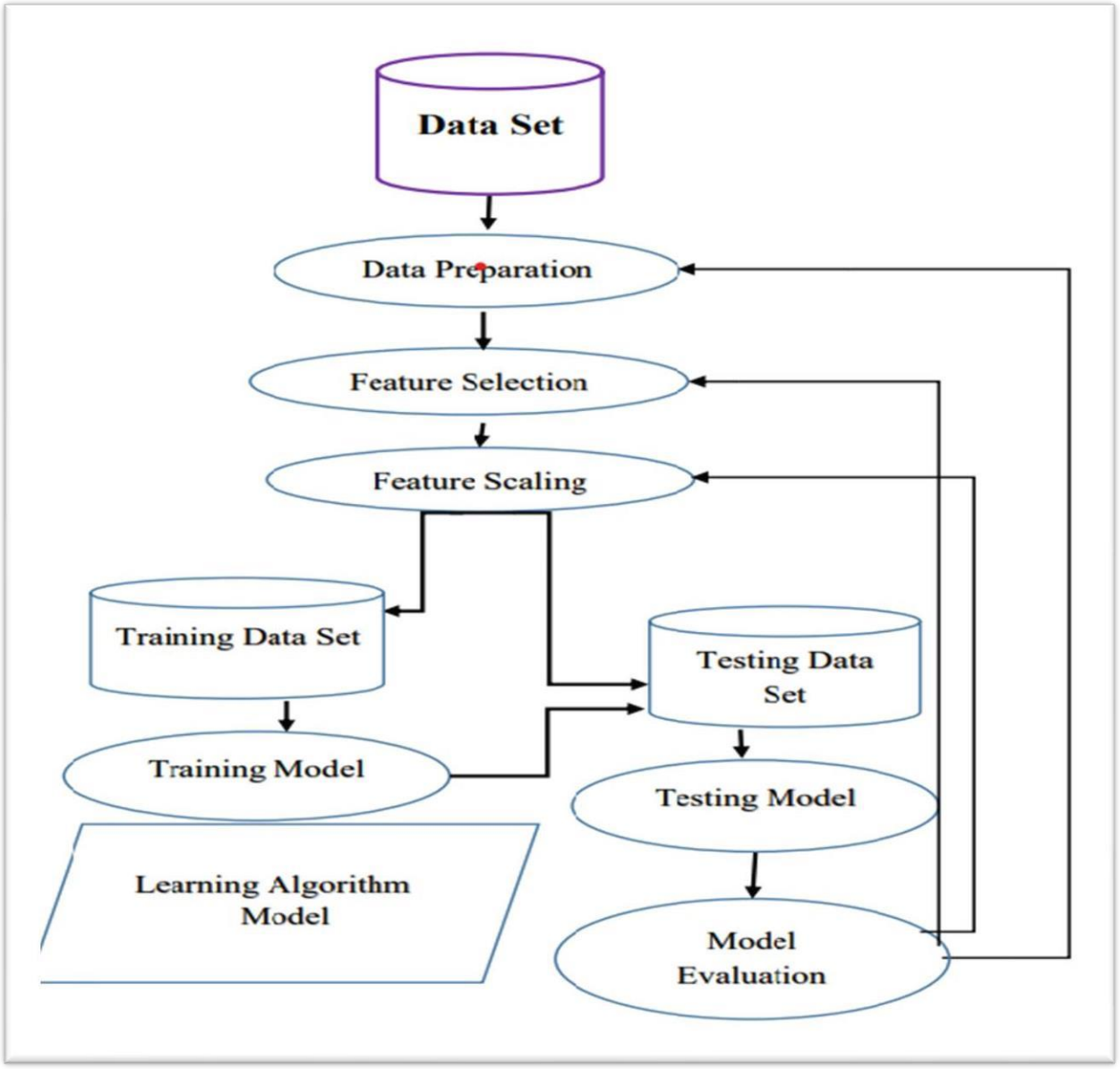

Figure 4.2 – Modeling Work Flow

4.1 Logistic regression: -

Logistic regression is a machine learning algorithm used for binary classification tasks. It models the probability of a binary output variable based on one or more input features, using a logistic function to transform a linear equation into a probability value between 0 and 1.

4.2 Decision Tree Classifier: -

The basic idea of a decision tree is to recursively split the data into subsets based on the most informative feature, until a stopping criterion is met, such as a maximum depth or a minimum number of samples per leaf. The splits are chosen to minimize the impurity of the resulting subsets, using a criterion such as Gini impurity or information gain

4.3 Random Forest Classifier: -

Random forest is a machine learning algorithm used for classification and regression tasks. It is an ensemble method that combines multiple decision trees to improve the accuracy of predictions. The basic idea of random forest is to create many decision trees, each trained on a random subset of the data and a random subset of the features. This randomness helps to reduce overfitting and improve the generalization performance of the model.

4.4 Support Vector Machines (SVM): -

Support Vector Machines (SVMs) are a machine learning algorithm used for classification and regression tasks. SVMs aim to find the hyperplane that separates the data into two classes, with the largest possible margin between the two classes. SVMs can handle non-linearly separable data by using kernel functions to transform the data into a higher- dimensional space where a linear boundary can separate the classes. “Support Vector Machine” (SVM) is a supervised machine learning algorithm. SVM can be used for classification and regression problems. SVM plots each data item as a point in n- dimensional space.

4.5 XGBoost:

XGBoost (Extreme Gradient Boosting) stands as a formidable machine learning algorithm extensively employed for classification and regression tasks. Its core mission involves constructing an ensemble of decision trees that iteratively correct errors made by preceding trees. By optimizing a loss function through gradient boosting, XGBoost excels in capturing intricate data relationships, ultimately leading to highly accurate predictions. XGBoost achieves predictive prowess by assembling a collection of trees that work in concert to refine results. XGBoost dynamically adapts its tree ensemble to enhance its performance.

5. DATA VISUALIZATION AND REPRESENTATION

In this project visualization of data is done using tableau. Tableau is a data visualization tool used for creating interactive and visually appealing dashboards, reports, and charts. It allows users to connect to various data sources and transform raw data into meaningful insights through drag-and-drop functionalities. Tableau offers a wide range of visualization options, including bar charts, line charts, scatter plots, heat maps, and geographic maps. Tableau enables users to create interactive dashboards that allow for exploration of data and easy sharing of insights with stakeholders. Tableau also offers features such as filtering, sorting, and grouping, which make it easier for users to analyze data and uncover patterns and trends. By using Tableau for data visualization, users can make better-informed decisions and communicate insights more effectively to their audience. Overall, Tableau is a powerful tool for data visualization that can help users to explore and understand data in a more intuitive and interactive way.

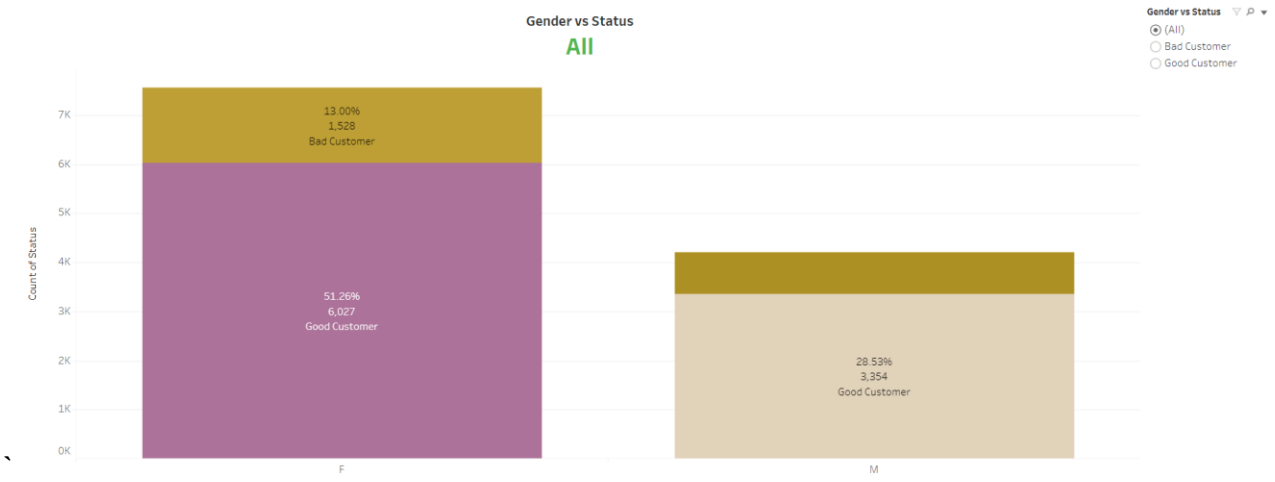

Fig 5.1 Gender vs Status

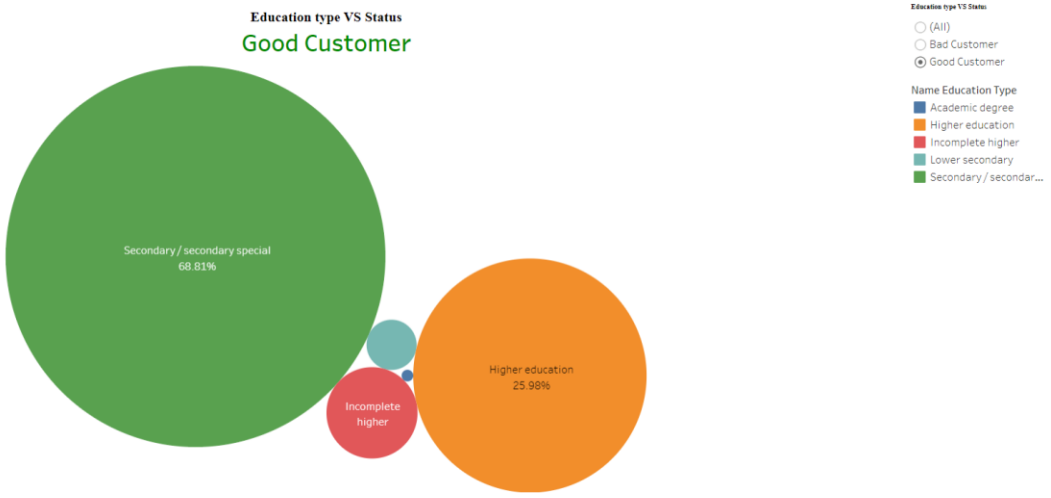

Fig 5.2 Education vs Status

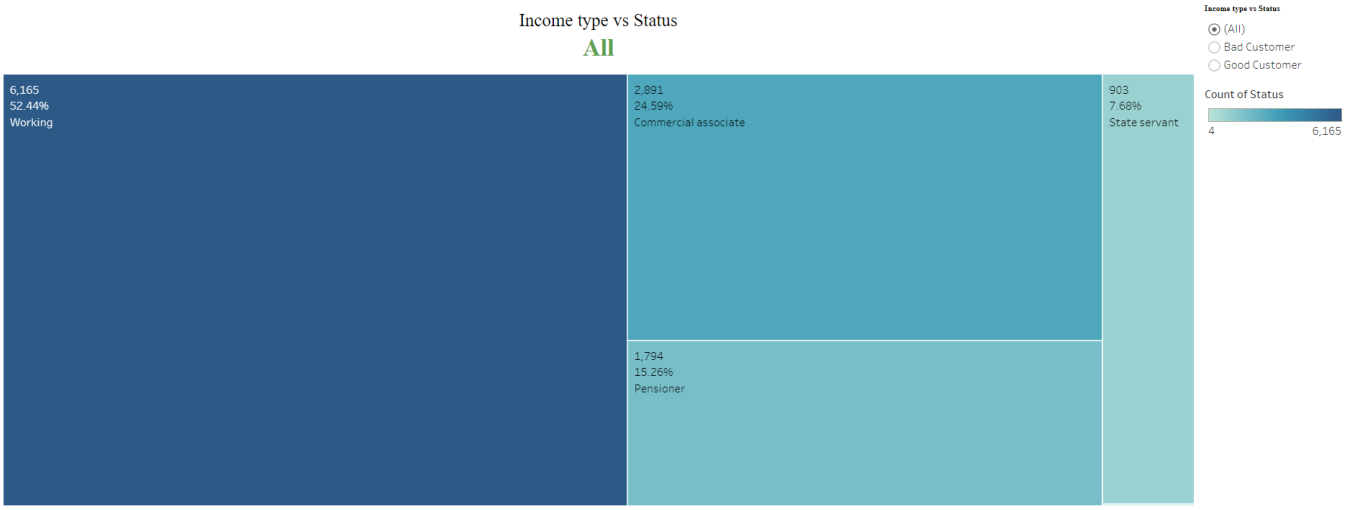

Fig 5.3 Income Type vs Status

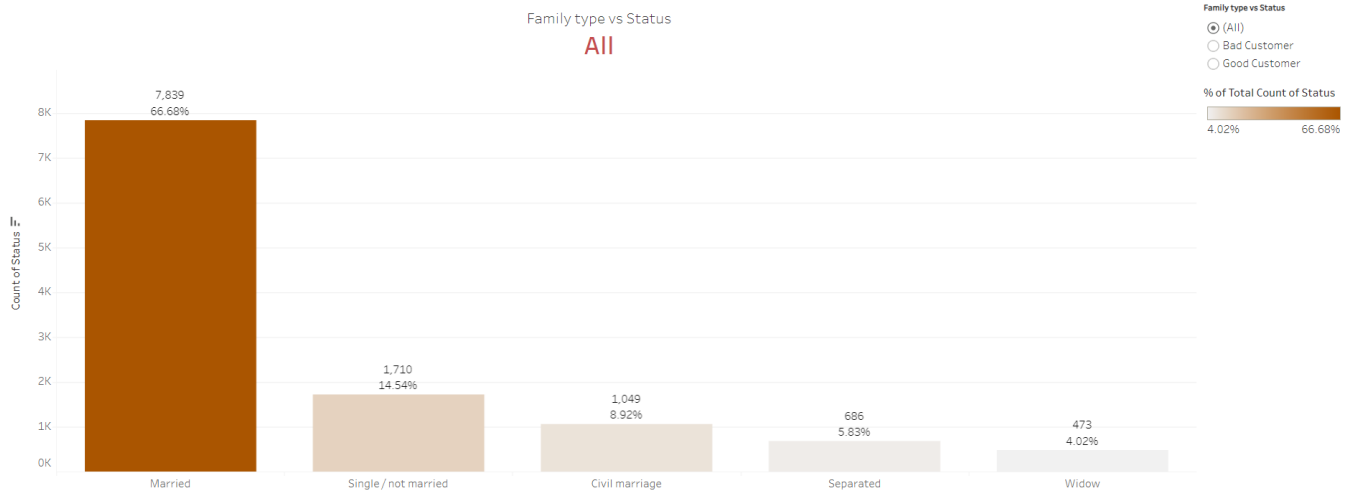

Fig 5.4 Family Type vs Status

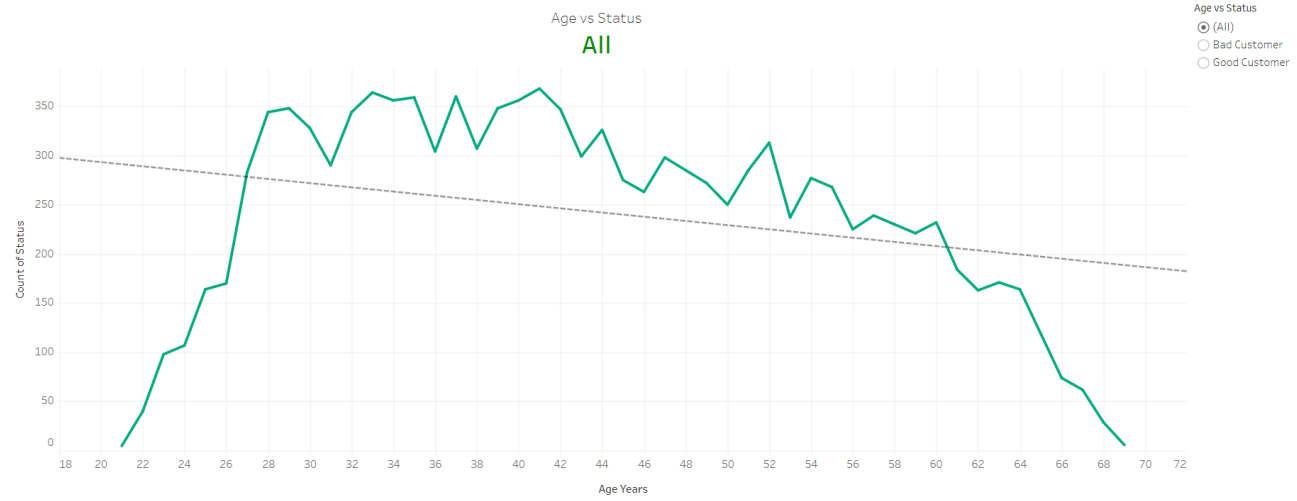

Fig 5.5 Age vs Status

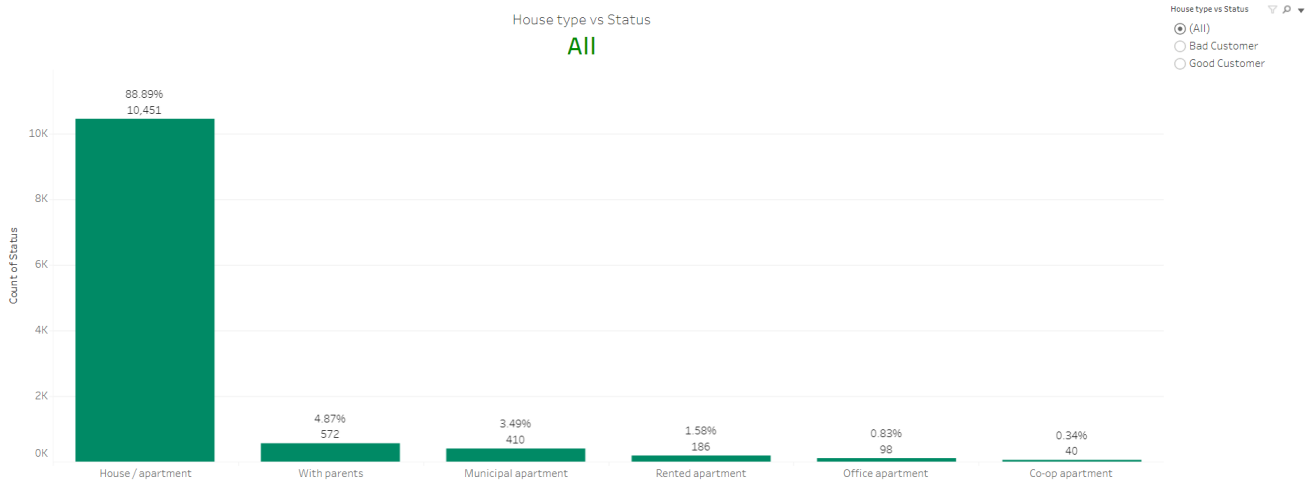

Fig 5.6 House type vs Status

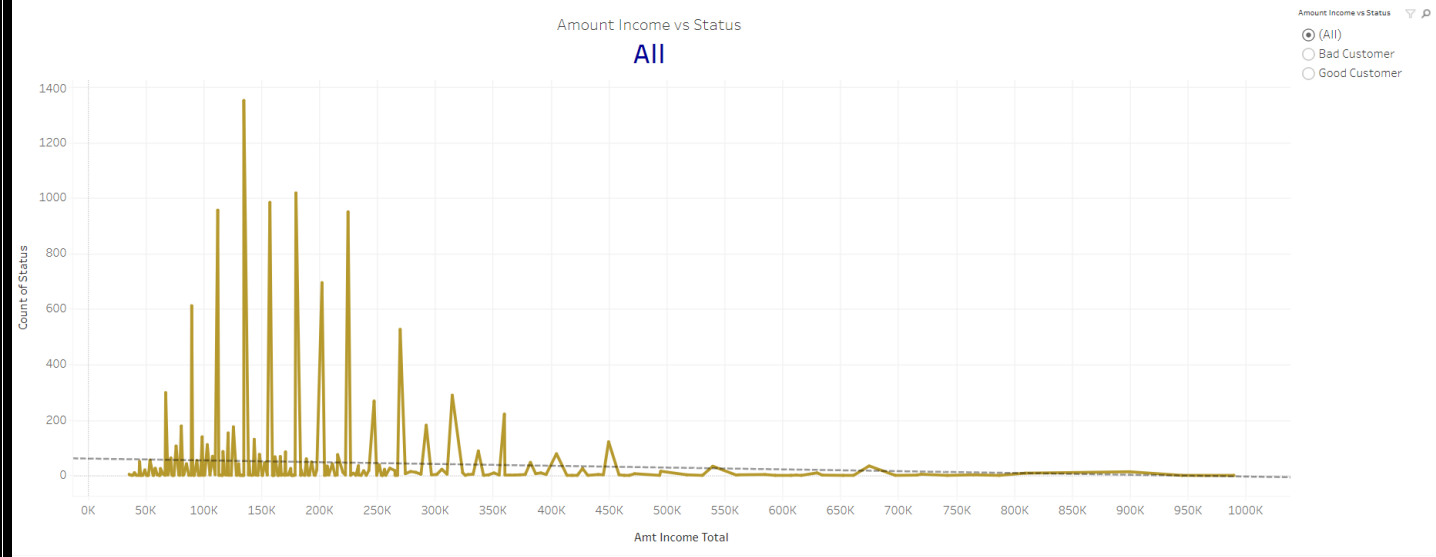

Fig 5.7 Amount Income vs Status

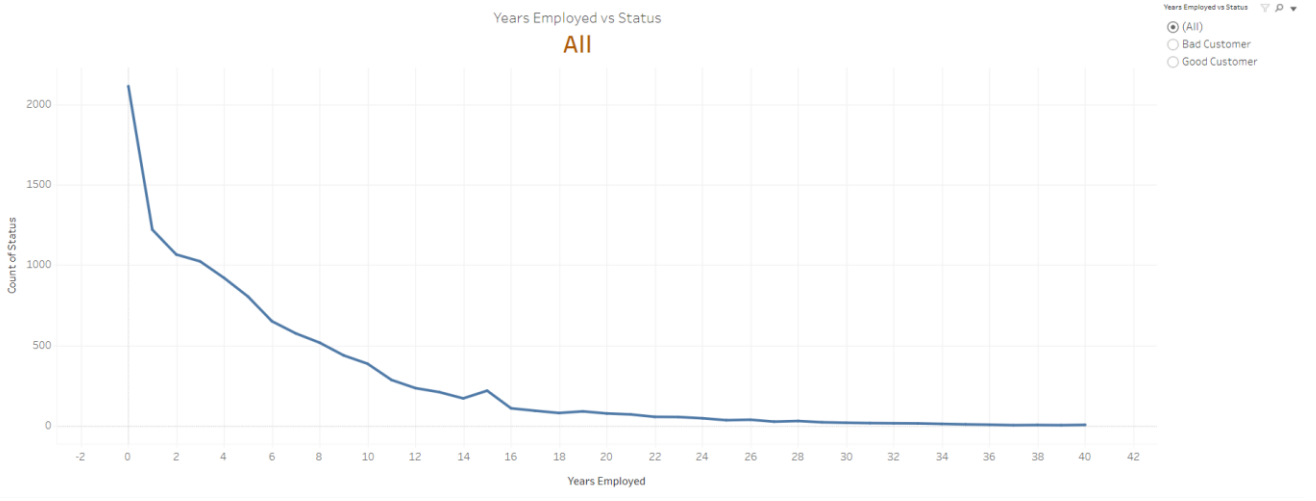

Fig 5.8 Years employed vs Status

Age and Employment Duration Analysis:

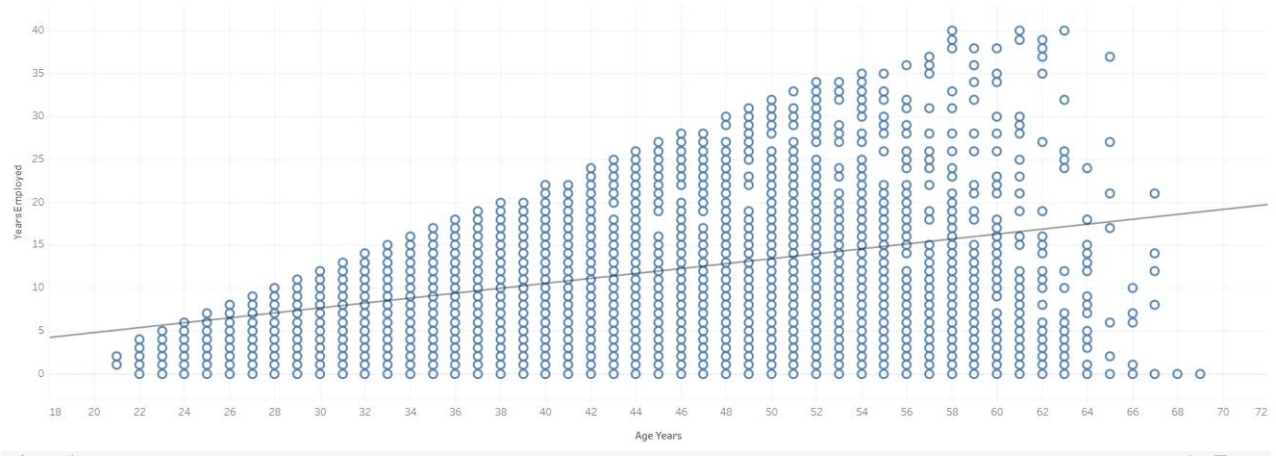

Fig 5.9 Age vs Years Employed

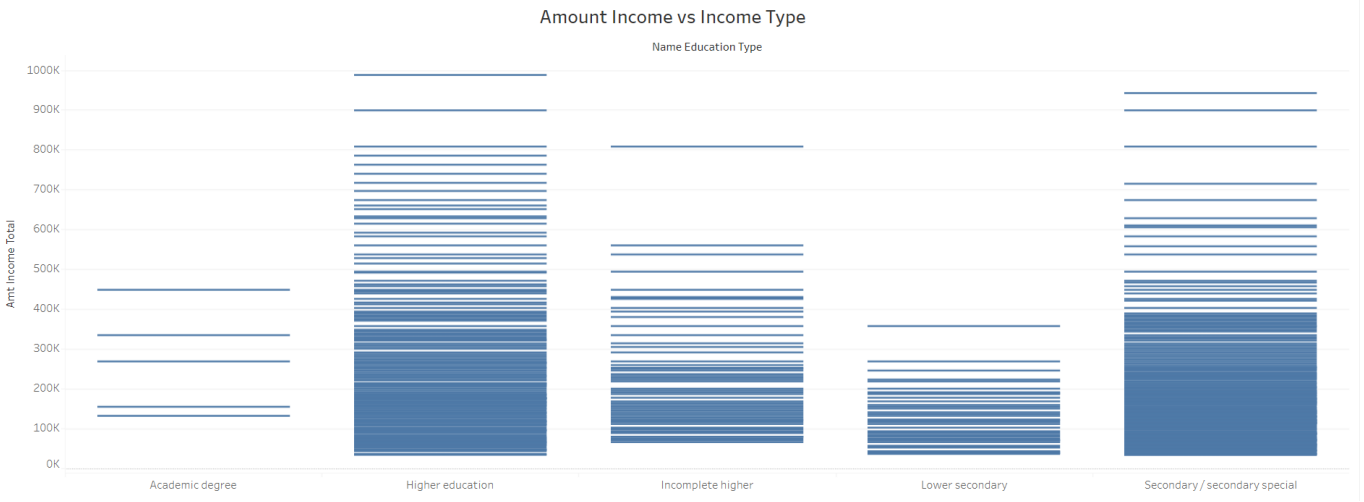

Fig 5.10 Education vs Income

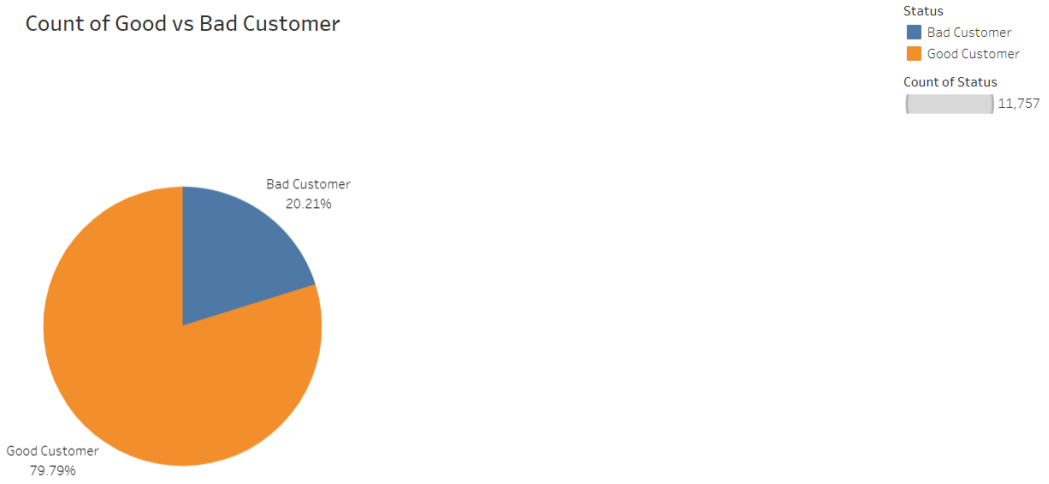

Fig 5.11 Status type pie chart

CONCLUSION

- We have obtained the publicly available data set and explanatory analysis was carried out to understand the data set. Then conducted several activities related to data preparations such as data preprocessing, feature selections and feature scaling. To achieve a desired outcome, it is very important to carry out these activities accurately. We have divided the data set into two parts as a training and test data set and the intended purpose is to validate the accuracy of the model. From this analysis we concluded that the XGBoost algorithm achieved the highest accuracy of 82.30% in predicting credit card approval, followed by Random Forrest and Decision Tree with an accuracy of 72.00% and 66.71%, respectively. SVM had the lowest accuracy of 50.26 %. Feature importance analysis revealed that the most important features in predicting credit card approval gender, total income, education type and years employed. These features had a significant impact on the credit card approval decision, and should be considered when evaluating creditworthiness of applicants.

FUTURE SCOPE

We realized that customer behavior might be different country to country and application of several real banking datasets can be considered for further studies. To consider default customers not only their demographic and socio-cultural data but also other existing credit facilities information such as other loans can be taken as features to get more accurate results. This data set was generated before the pandemic situation. After the COVID-19 pandemic there have been change in defaulters and their payment methods but things are slowly returning to normal. Application of data sets including COVID-19 impact under new normal to be an area of concern for researchers. Furthermore, the data set is a highly imbalanced data set and we have applied SMORTE for balancing. Whether there is a relationship between Nonlinearity in highly imbalanced class problems with SMORTE application is another area of concern for researchers.

REFERENCES

Agarwal, A., Rana, A., Gupta, K., Verma, N., 2020. A Comparative Study and enhancement of classification techniques using Principal Component Analysis for credit card dataset, in: 2020 International Conference on Intelligent Engineering and Management (ICIEM). Presented at the 2020 International Conference on Intelligent Engineering and Management (ICIEM), IEEE, London, United Kingdom, pp. 443–448. https://doi.org/10.1109/ICIEM48762.2020.9160230

Antonakis, A.C., Sfakianakis, M.E., 2009. Assessing naïve Bayes as a method for screening credit applicants. Journal of Applied Statistics 36, 537–545. https://doi.org/10.1080/02664760802554263

Banasik, J., Crook, J., Thomas, L., 1999. Not if but when will borrowers' default. Journal of the Operational Research Society 50, 6.

Bhatore, S., Mohan, L., Reddy, Y.R., 2020. Machine learning techniques for credit risk evaluation: a systematic literature review. Journal of Banking and Financial Technology 4, 111–138. https://doi.org/10.1007/s42786-020-00020-3

Blagus, R., Lusa, L., 2013. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 14, 106. https://doi.org/10.1186/1471-2105-14-106

Chornous, G., Nikolskyi, I., 2018. Business-Oriented Feature Selection for Hybrid Classification Model of Credit Scoring, in: 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). Presented at the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), IEEE, Lviv, pp. 397–401. https://doi.org/10.1109/DSMP.2018.8478534

Data Science Process Alliance, n.d. What is CRISP DM? What is CRISP DM? URL https://www.datascience-pm.com/crisp-dm-2/ (accessed 11.20.20).

Elreedy, D., Atiya, A.F., 2019. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. Information Sciences 505, 32–64. https://doi.org/10.1016/j.ins.2019.07.070

Fernando, J., 2021. Delinquent Account Credit Card [WWW Document]. Loan Basis. URL https://www.investopedia.com/terms/d/delinquent-account-credit-card.asp (accessed 4.28.21).

Galarnyk, M., 2018. Understanding Boxplots. Understanding Boxplots. URL https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51 (accessed 1.20.21).

Hussein, A.S., Li, T., Yohannese, C.W., Bashir, K., 2019. A-SMOTE: A New Pre-processing Approach for Highly Imbalanced Datasets by Improving SMOTE. International Journal of Computational Intelligence Systems 12(2), 11.