A CAPSTONE PROJECT ON

Prediction of Credit Card Approvals.

SUBMITTED BY

KAMINI RANA

(22MBA20255)

In partial fulfilment for the award of the degree of

MASTER OF BUSINESS ADMINISTRATION
IN
BUSINESS ANALYTICS

CHANDIGARH UNIVERSITY

Discover. Learn. Empower.

APEX INSTITUTE OF TECHNOLOGY-MANAGEMENT

CHANDIGARH UNIVERSITY
GHARUAN, MOHALI(PUNJAB)

APRIL 2024

CHANDIGARH UNIVERSITY

Discover. Learn. Empower.

BONAFIDE CERTIFICATE

Certified that this project report **"Prediction of Credit Card Approvals."** is the bonafide work of **" KAMINI RANA (22MBA20255)"** who carried out the project work under my/our supervision.

SIGNATURE
(Signature of the HOD)

SIGNATURE
(Signature of the supervisor)

Mr. Gagan Vibhu
HEAD OF THE DEPARTMENT

Prof. Shailja Gera
SUPERVISOR

Submitted for the project viva-voce examination held on

INTERNAL EXAMINER

EXTERNAL EXAMINER

ACKNOWLEDGEMENTS

By the grace of the God Almighty, we are grateful to be able to have patience and courage to complete this project. I submit my sincere gratitude to my project supervisor. **Ms. Shailja Gera**, for allowing us to write on our field of interest. We would like to thank her for her constant support and patience throughout this research project. We are deeply indebted to him for his precious advice and help. Special thanks to our parents for their constant help, encouragement and never-ending support in everything we do. Last but not least , I would also seize the opportunity to thanks our friends for constantly encouragement and assistance in different ways. With the invaluable help from them, the completion of this project would not have been possible.

ABSTRACT

The prediction of credit card approvals stands as a seminal endeavor at the intersection of data analytics and financial decision-making, aimed at revolutionizing the credit assessment process within the banking industry. This project embarks on a journey to develop and deploy advanced predictive modeling techniques to forecast credit card approvals with unprecedented accuracy and efficiency. Leveraging a rich repository of historical data encompassing various socio-economic, demographic, and financial parameters, this study endeavors to unveil the underlying patterns and trends that influence credit card approval decisions. Through the iterative refinement of predictive algorithms, this project seeks to empower financial institutions with the tools and insights necessary to optimize risk management practices, streamline application processes, and enhance overall customer satisfaction levels. Central to the success of this project is the meticulous exploration and analysis of diverse data sources, ranging from traditional credit bureau data to alternative data streams such as transaction histories and social media profiles. By synthesizing these disparate datasets and harnessing the power of machine learning algorithms, this study endeavors to uncover nuanced correlations and predictive signals that transcend conventional credit scoring methodologies. Moreover, this project emphasizes the paramount importance of ethical data usage, privacy protection, and regulatory compliance in the development and deployment of predictive models within the financial sector.

The implications of this project extend far beyond the realm of credit card approvals alone, permeating through various facets of financial decision-making and risk management. From loan origination to portfolio optimization, the methodologies and insights derived from this study hold the potential to revolutionize industry practices and reshape strategic frameworks. Furthermore, as financial institutions navigate the ever-evolving landscape of regulatory requirements and consumer expectations, the predictive models developed in this project serve as invaluable assets for driving informed decision-making and fostering sustainable growth.

In essence, the prediction of credit card approvals project embodies a paradigm shift in the way financial institutions leverage data analytics to drive business outcomes and enhance customer experiences. As the project unfolds, it illuminates a pathway toward a future defined by precision, agility, and customer-centricity, where predictive insights serve as guiding beacons for navigating the complexities of modern finance.

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 Project Overview:-

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.

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.

1.2 Statement of the problem

Many researchers have conducted machine learning applications on credit scoring and customer default predictions. Researchers' have concluded that SVM (support vector machine) and ANN (Artificial Neural Network) performed better than other classifiers. However, it is important to study how these two algorithms behave differently with filter based feature selection and balancing imbalanced data which is inherited by nature using Synthetic Minority Oversampling Technique (SMORTE).

"To examine two algorithms and identify best classification algorithm to predict customer eligibility for a credit card and to minimize possible credit loss "

1.3 Research Aims and Objectives

The primary focus of the research is expressed under aims and objectives as follows.

1.4 Aim

This research supports the decision making process while speeding up the process to give a benefit for the bank as well as for the applicant and to attract on time paying customers by using banking data for smarter data–driven decision making. This research is highly applicable for Sri Lankan banking industries as most of the banks are granting credit card facilities to the customers. Hence the application of the model to local context to be considered.

1.4.1 Objectives

Research objectives of the project as follows:

1.4.1.1 To predict the customer eligibility for a credit card to minimize possible future credit loss by using supervised machine learning techniques.

1.5 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 (Thomas J. Catalano, 2020).

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).

1.5.1 Component of a Credit Card

Figure 1.1 - Component of a Credit Card

Figure 1.1 illustrated components of a credit card and details of components were listed below.

- **Issuer Logo:** In front of the credit card, credit card network logo (e.g. visa, master) and issuing bank logo displayed.
- **EMV Chip:** The chip stores card data in an encrypted way to prevent stealing of credit card number easily.
- **Magnetic Strip:** The magnetic strips are readable through some specific machines used for monetary transactions. Also it contains account data.

- **Card Holder Name and Card Number:** Card holder name & Credit Card number appeared in front side of the card.
- **Credit Card Expiration Date**: Card has an expiration date. The date shows the month and the year and helps merchants to identify the validity of the card.
- **Signature Box:** Signature box is the place cardholders are supposed to place their signature.
- **CVV Code:** In back side of the card there is CVV number. It is three-digit combination and used to protect customers' financial transaction from fraud and theft.
- **Hologram:** In the backside of card unique three-dimensional hologram display of credit card network. (E.g. Visa uses a dove hologram, MasterCard – a world map)

1.5.2 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.

1.5.3 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.

1.5.4 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.

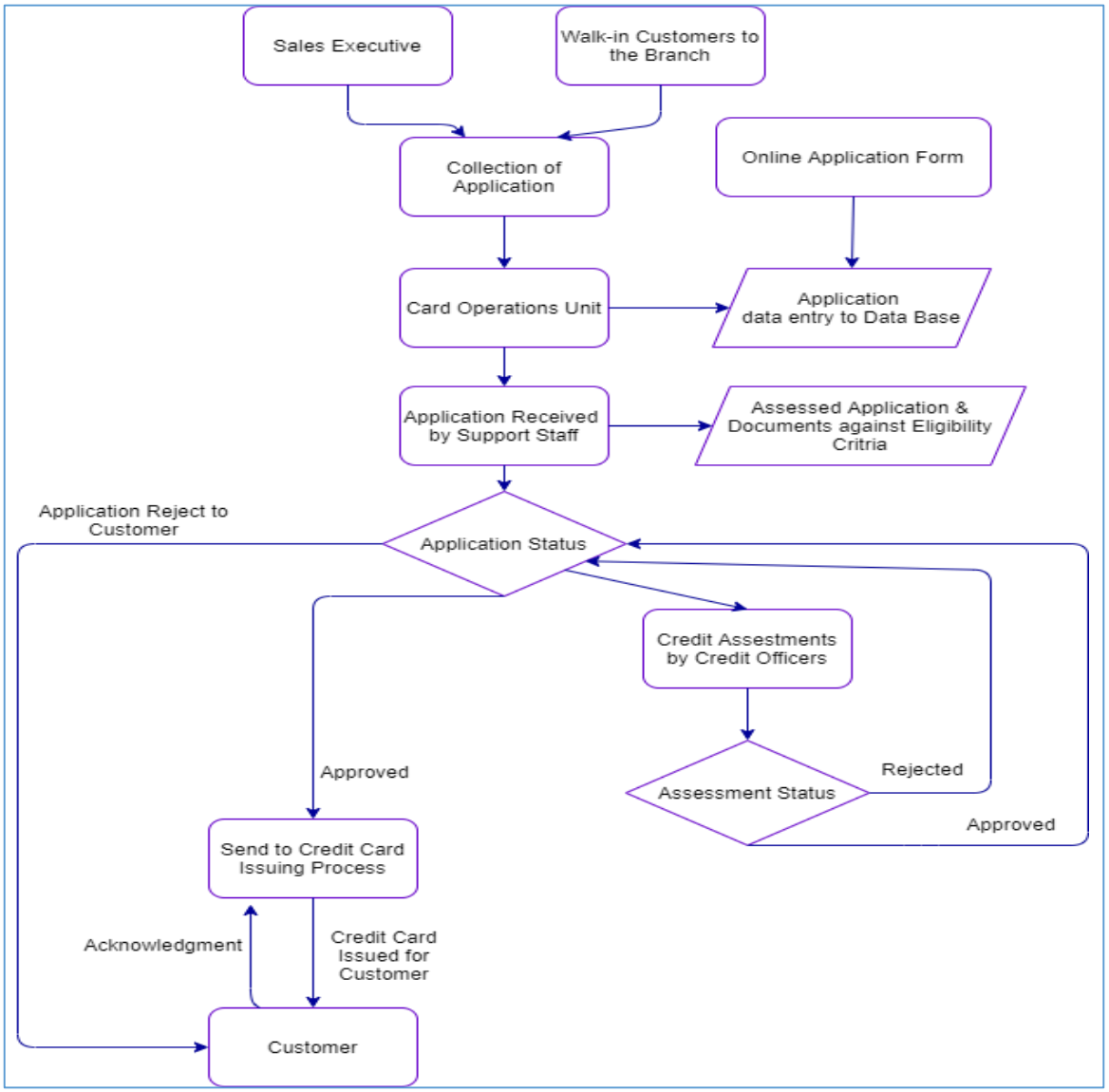


Figure 1.2 - Credit Card Issuing Process

Credit card application form with required supportive documents are handover by sales executive or walking customer to the branch. All applications will be handover to credit card operations unit. Application data entered to the Card Application Database.

CHAPTER 2

LITERATURE REVIEW

2.1 Risk Management at Banks

Saunders and Cornett (2014) states that the bank management's intendent to increase returns for shareholders which come with increased in risk. There are many risk faced by the banks. They are credit risk, interest rate risk, market risk, off-balance-sheet risk, technology and operational risk, foreign exchange risk, country or sovereign risk, liquidity risk, liquidity risk and insolvency risk. By effectively managing these risks, the banks can perform better. Furthermore, these risks and subject to regulatory attention due to the major role playing by banks in financial system. In (Leo et al., 2019) mentioned that "credit can be defined as the risk of potential loss to the bank if a borrower fails to meet its obligations (interest, principal amounts)". Credit risk is the single largest risk banks face.

2.2 Machine Learning & AI Implementation on Risk Management

Bhatore et al., (2020) carried out a comprehensive literature review on currently available research methods and machine learning techniques for credit risk evaluation. They have selected three major factors that create credit risk, careful examination and inspections while giving loans (credit scoring), continuous monitoring of customer payments and any other behavior patterns to decrease the probability of generating frauds (fraud detection) and non-performing assets (NPA). Further they have analyzed model evaluation techniques, current studies and research trends. Team have reviewed a total of 136 papers published between 1993 and March 2019 and concluded that Ensemble and Hybrid models with neural networks and SVM are more adaptive and mentioned that lack of complete public datasets will be cause for concern for researcher. Following figure summarized their findings about application of different ML techniques.

Leo et al., (2019) have been conducted a Literature Review on Machine Learning in Banking Risk Management. According to them (Bellotti and Crook, 2009; Huang et al., 2007; Li et al., 2017, Harris 2013) use SVM to develop scoring model for consumer credit management. Further they have mentioned (Yeh and Lien 2009; Galindo and Tamayo 2000; Keramati and Yousefi 2011) use Neural Network to build credit scoring model.

Banasik et al., (1999) mentioned that credit scoring systems were built to answer whatlikelihood of applicant of the credit facility to be received will be default in the future. Differentmodeling techniques use previous customers credit details and classified the customer as 'goodand 'bad 'considering their payment settlement pattern over a specified period. Furthermore, SriLaxmi et al., (2020) states that there are multiple criteria and factors considered when approving of a credit card. Mainly demographic, income, credit bureau data of the customer. Using credit card customer's past data can identify key factors affect in credit risk by using models such as Logistic Regression and Random Forest. As a methodology they have Cross- Industry Standard Process for Data Mining (CRISP-DM). Moreover, Sariannidis et al., (2019)modeled seven classification methods, KNN, Logistic Regression, Naïve Bayes, Decision Trees, Random Forest, SVC, and Linear SVC and compared the prediction accuracy of these models. They have stated that in terms of lending decisions except few, most of the characteristic variables used can satisfactorily analyze default features. Additionally, it is important to have a better understanding of borrowers' behavior with accounting, demographic and historical characteristics.

Karthiban et al., (2019) proposed a hybrid model which includes a novel 16-layer genetic cascade ensemble of classifiers, normalization techniques and two types of SVM classifiers. He used kernel functions, parameter optimizations, and stratified 10-fold cross-validation for feature extraction methods. The model achieved 97.39% prediction accuracy and concluded that proposed method can be applying in the banking domain to assess the bank credits of the applicants and aid the decision making process. Furthermore, Pristyanto et al., (2019) applied information gain, gain ratio, and correlation based feature selection (CBFS) and as a classifier used K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), Artificial Neural Network. He concluded that feature selection does not always advance classifier accuracy but be subject to the characteristics and algorithms. Moreover, Munkhdalai et al., (2019) performed a broad comparison between the machine-learning approaches and a human expert-based model FICO credit scoring system by using a Survey of Consumer Finances (SCF) data. To reduce the computation cost and to choose the most informative variables they have applied two variable-

selection methods for feature-selection. In this study, they present TSFFS (two-stage filter feature selection) algorithm and use the NAP method for variable-selection. As ML techniques logistic regression, support vector machines, an ensemble of gradient boosted trees and deep neural networks used. They have concluded subset selected by NAP from the deep neural networks and XGBoost algorithms trained on achieve the very best accuracy and area under the curve (AUC).

Comparative evaluation of the performances of five popular classifiers namely, Naive Bayesian Model, Logistic Regression Analysis, Random Forest, Decision Tree, and K-Nearest Neighbor Classifier use in credit scoring used to carried out by (Wang et al., 2020). They have concluded that each individual classifier has its own strength and weakness. However, the results of this experiment discover that Random Forest achieves better results than others in terms ofprecision, recall, AUC (area under curve) and accuracy. However, Karthiban et al., (2019) applied the Regression, Naive Bayes, Generalized Linear model, Deep learning(DL), Decision tree, Random Forest and Gradient Boosted trees were for Bank Loan Approval data set. They used confusion matrix to evaluate models. In there they have consider Accuracy, Sensitivity orTrue Positive Rate or Recall, Specificity or True Negative Rate, Precision, F measure, Classification Error, AUC, ROC Curve for evaluation. Moreover, Antonakis and Sfakianakis (2009) benchmark two data sets NBR against linear discriminant analysis, logistic regression analysis, k-nearest neighbors, classification trees and neural networks. He concluded that considering all measures used, NBR is found to have lower predictive power than other five classifiers in each data set.

To accurately identify loan defaulters (Shoumo et al., 2019) applied support vector machine, extreme gradient boosting, logistic regression and random forest classifiers to a loan data set. Dimensionality reduction carried out by using Recursive Feature Elimination with Cross Validation and Principal Component Analysis. To model evaluation metrics such as F1 score, AUC score, prediction accuracy, precision and recall have been used. They have concluded that support vector machines can outperform other tree-based models or regression models. Furthermore, they have concluded that the model has shown that recursive feature elimination with cross-validation can outperform models based on principal component analysis. However, Agarwal et al., (2020) use different classification algorithms were evaluated such as Logistic Regression, Decision Tree, K-Nearest Neighbor and Naive Bayesian for credit card dataset. The dataset is obtained from UCI Repository credit card defaulter. Main objective to compare the performance measures between the original dataset and original dataset with the principal

component is applied. Benchmarked before and after applying the principal component. Different algorithms are compared on the basis of various metrics such as Accuracy, Precision, F1-Score, Recall, ROC. He concluded that Logistic regression is performed better in this particular data set in accuracy and precision measures. The other performance measures such as ROC, F1-Score showed good results for naïve Bayesian. K-Nearest neighbor showed acceptable performance in terms of recall.

Kumar Gupta and Goyal (2018) applied Artificial Neural Network (ANN) to predict the creditworthiness of an application. Data set has been taken from kaggle.com (lending club loan data). Dependent Variable is loan status (0 and 1). Scoring system develop by using discriminant analysis. They have concluded that results of both the systems have shown an equal outcome on the dataset. The classifier is very effective with the accuracy of 97.68% in artificial neural network. The system classifies the predicted variable correctly with a very low error. Hence, both models can be used to identify credit default with equal accuracy. However, Lee and Chen (2005) evaluate the performance of credit scoring using two-stage hybrid modeling methods with artificial neural networks and multivariate adaptive regression splines (MARS). They have used n-fold cross-validation to reduce the possible bias linked with the random sampling of the training and testing samples and the entire dataset is randomly split into mutually exclusive n numbers. To build the two-stage hybrid model, a single-layer BPN model again applied. Important independent variables gained from the MARS were input to the input layer of the hybrid model. He concluded that the proposed hybrid method outperformsthe results using discriminant analysis, logistic regression, artificial neural networks andMARS.

Wang et al., (2011) Carried out comparative assessment of the performance of three ensemble methods; Bagging, Boosting, and Stacking with four classifiers namely Logistic Regression Analysis (LRA), Decision Tree (DT), Artificial Neural Network (ANN) and Support Vector Machine (SVM). They have discovered that the three ensemble methods can significantly improve individual base learners. Precisely Bagging performs better than Boosting across all credit datasets. In terms of average accuracy, type I error and type II error; Stacking and Bagging DT get the best performance in their experiments. Furthermore, Marqués et al., (2012) use two resampling-based ensembles (bagging and AdaBoost) and two attribute-based algorithms (random subspace and rotation forest) in various sequences. To compare the performance of the rotation forests with other classifier ensembles six real-world credit data sets used. Fivefold cross-validation method has been adopted and to evaluate accuracy, error

rate, Gini coefficient, Kolmogorov– Smirnov statistic, mean squared error, area under the ROC curve, and Type-I and type-II errors used. Their experimental results and statistical tests disclosed that new two-level classifier ensemble based approaches are a suitable solution for credit scoring problems performing better than the traditional single ensembles and individual classifiers.

Chornous and Nikolskyi (2018) proposed an ensemble-based classification model with business related feature selection to increase accuracy of classification of credit scoring. The data set was collected from Vidhya loan prediction hackathon and contains of 614 observations. He has selected Information Gain, Chi-Squared and Mean Decrease Gini as feature selection methods. He concluded that a hybrid approach for user-defined variables can be more effective in ensemble binary classification models. Furthermore, Oreski et al., (2012) propose a feature selection technique for finding an optimum feature subset which makes neural network classifiers high in accuracy. The feature selection techniques used here is Genetic algorithm, Forward selection, Information gain, Gain ratio, Gini index and Correlation. Credit dataset collected at a Croatian bank used to conduct the experiment. They have concluded that discovering the most important features in defining the risk of a default, hybrid system with a genetic algorithm can be used as feature selection techniques.

Madyatmadja and Aryuni (2005) study to discover an appropriate data mining method for credit scoring credit card application in a Bank and improve the performance. Their proposed model of classification applies Naïve Bayes and the ID3 algorithm. The class variable in the data set is classified into two class labels as ‘approve’ and ‘reject’. By using the credit experts’ knowledge, the class label is determined. They have got 82% Accuracy on Naïve Bayes classifier and 76% accuracy on ID3. They have concluded that the Naïve Bayes classifier performed better with high accuracy than the ID3 classifier. Furthermore, Hamid and Ahmed (2016) build a new model for categorizing loan risk in the banking sector by using data mining to predict the status of loans. Three algorithms have been used to build the proposed model: j48, bayesNet and Naïve Bayes. The developments were carried out with Weka application. They have concluded that J48 was selected as the best algorithm based on its high accuracy and low mean absolute error as shown in the result.

CHAPTER 3

METHODOLOGY

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.

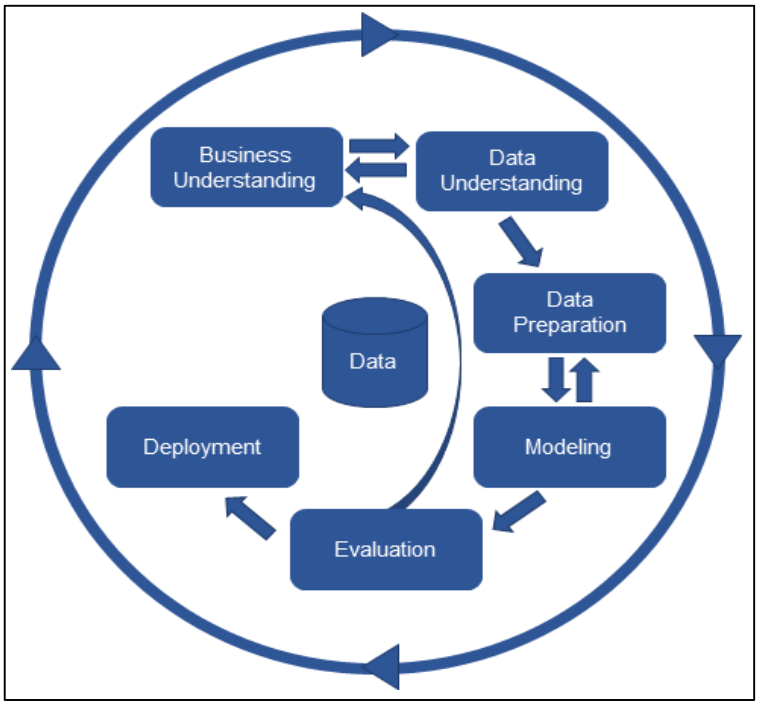


Figure 3.1 - CRIP –DM Model *(Taylor, 2017)*

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 phases are 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.

3.1 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.

3.2 Data Understanding

The data set has been taken from kaggle.com data repository (Song, 2019). This data set is publically 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

Table 3.1- Detail Information about application data set

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 ,Working
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 occupation
CNT_FAM_MEMBERS	Family size	Numerical - Float	

Credit card records has one categorical column and two numerical columns as shown in figure 3.3. This data set contain duplicate data for ID column.

```
In [53]: credit_record.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 3 columns):
 #   Column          Non-Null Count    Dtype
---  ------          --------------    -----
 0   ID              1048575 non-null  int64
 1   MONTHS_BALANCE  1048575 non-null  int64
 2   STATUS          1048575 non-null  object
dtypes: int64(2), object(1)
memory usage: 24.0+ MB
```

Figure 3.2 - Information of credit record data

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 3.4 shows identified different data preparation activities related to our project and each activity will discuss separately below.

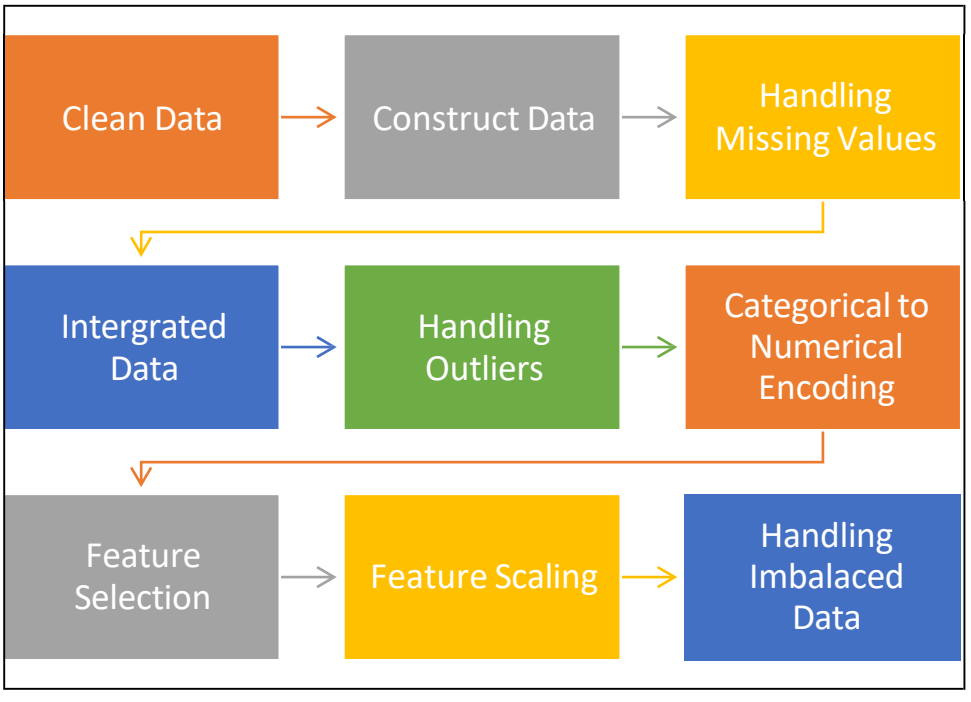


Figure 3.3 - Activities in Data Preparation Phase

3.3.1 Clean Data

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

3.3.2 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.

3.3.3 Construct Data

Derive new attributes from exiting data set.

3.3.4 Integrated Data

Integrating data phase basically combined data from multiple sources.

3.3.1 Encoding Categorical Data

Categorical data can't 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

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.

3.3.2 Feature Scaling

We have applied Standard scaler for numerical features. Standardization is highly used in SVM and ANN. In here transformed set of numerical values making mean equal to 0 and standard deviation equal to 1.

3.3.3 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 observation belong to one class is significantly lower than the other class.

3.4 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 ANN and 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 ANN and SVM, compare the accuracy of two 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 3.9.

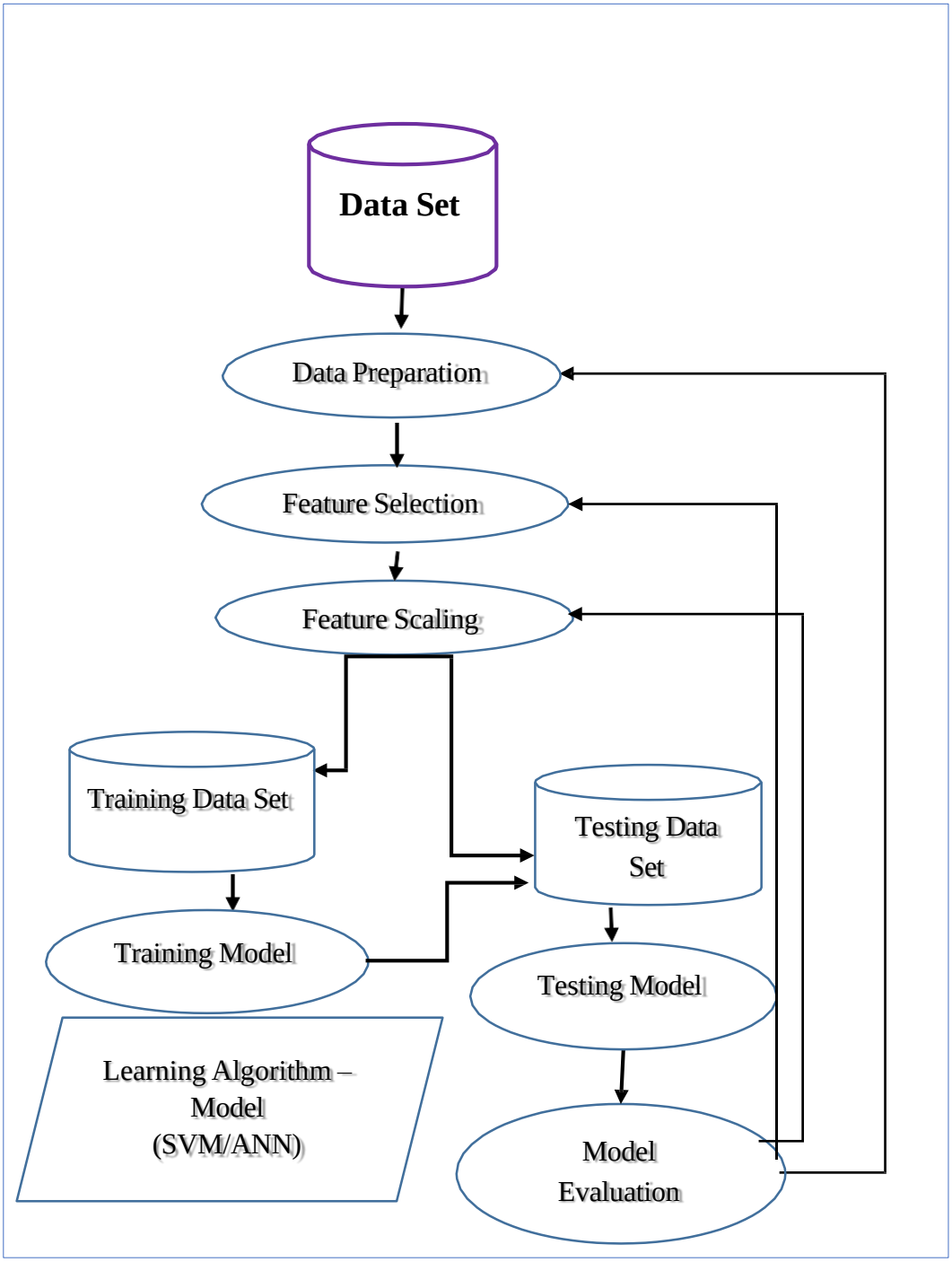


Figure 3.4– Modeling Work Flow

3.4.1 Artificial Neural Network

Artificial Neural Network (ANN) is evolved from biological neural network of human brain. It is deep learning algorithm and use as information processing technique. We can use ANN not only for a classification problem but also regression. Neural network may contain 3 layers as follows:

- Input Layers – Raw information feed as input to the network
- Hidden Layer – Input unit and weight. There can be many hidden layers.
- Output Layer – This layer depends on hidden layer and weights or input layer. Prediction related to response variable return in output layer.

3.4.2 Support Vector Machine

"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. Here n is the number of features. Value of each feature belongs to a particular coordinate. Classification was performed by finding hyperplanes that divide two classes (Ray, n.d.).

Hyperplane

Hyperplane is the best decision boundary. Features penetrated in the data set decide the dimension of the hyperplane. Hyperplane will be a straight line if we have two features. If we have more features, the hyperplane will be a 2D plane. Maximum distance between data points calculate maximum margin. Hyperplane created by considering maximum margin.

Support Vector

SVM selects extreme points known as support vectors. The extreme points are the points closest to the hyperplane. These vectors support hyperplanes and hence we call them support vectors.

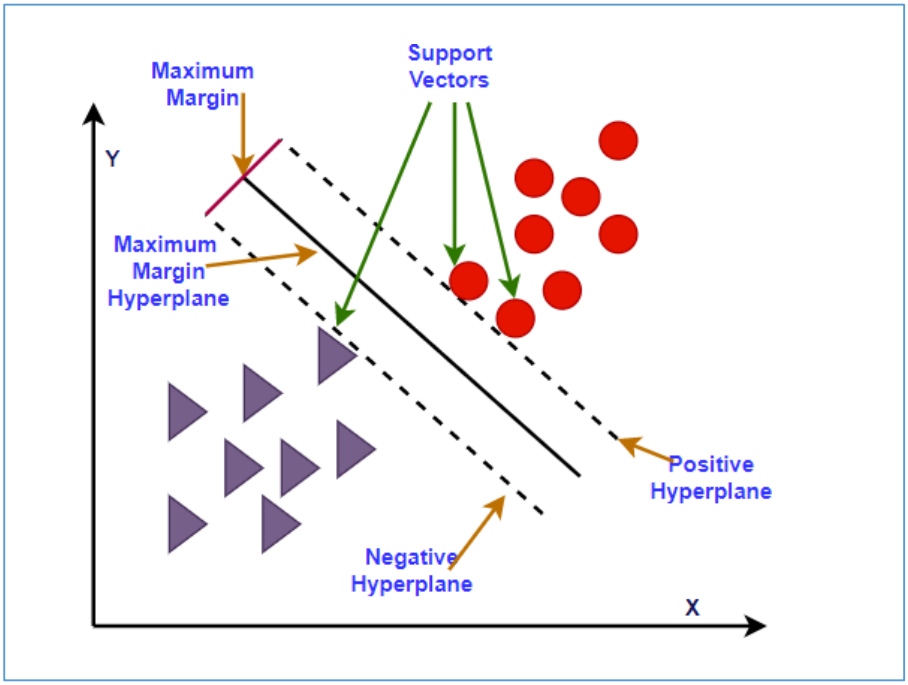


Figure 3.5 - SVM decision boundary or hyperplane

SVM are categorized into two types as linear and nonlinear. Using a straight line, we can divide data into two dimensions in linear SVM. We cannot separate if the data are arranged in non-linearly. Linear data used two dimensions X and Y. For nonlinear data Z is added additionally. We can use both linear and nonlinear methods and select the more accurate model. Figure 3.16 shows an example for Nonlinear.

3.4.3 Model Validation

Confusion Matrix - Confusion Matrix is widely used performance measurement in classification problems.

Figure 3.6 – Confusion Matrix

- True Positive – Predicted value is positive and equal to actual positive value
- True Negative - Predicted value is negative and equal to actual negative value
- False Positive (Type 1 Error) – Predicted as positive and actual value is not positive
- False Negative (Type 2 Error) - Predicted as negative and actual value is not negative.
- Recall – Considering positive classes how many correct predictors.
  Recall = TP /(TP+FN) (2)
- Accuracy - Considering all classes (positive and negative), how many correct predictors.
  Accuracy = (TP+TN) / Total (3)
- Precision – Considering all positive predicted classes how many correct positive actuals.
  Precision = TP/(TP+FP) (4)
- Recall, Accuracy, precision values should be high for better performance.
- F1 Measure – It is difficult to compare precision and recall when they are in high and low or vice versa. Precision and recall can be measure at same time by using F-score.
  F1 Measure = (2 * Recall * Precision) / (Recall + Precision) (5)

CHAPTER 4

IMPLEMENTATION AND RESULTS

This chapter describes implementation and results evaluation activities in detail. List of activities discussed in here are explanatory analysis of data, data preparation activities, models building, evaluation of models and deployments.

4.1 Explanatory Data Analysis

Graphical and numerical representation of data provide better insight about particular data set. Graphical representation of our data set is described below.

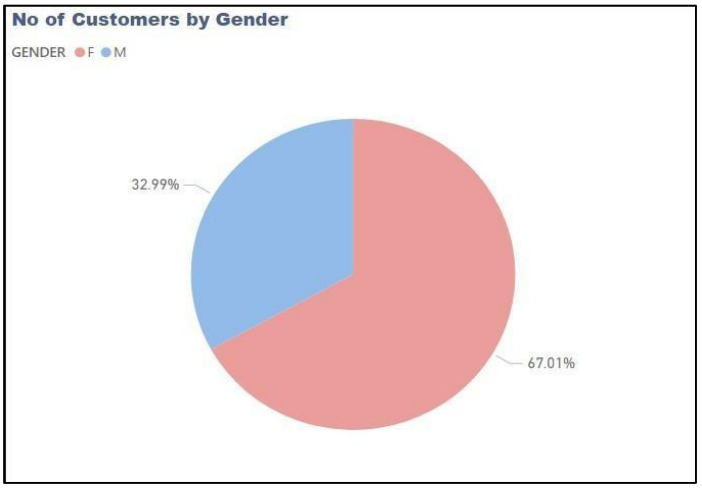


Figure 4.1– No of Customers by Gender

According to the figure 4.1 distribution of gender of the customers are 67 % female and 32% are male.

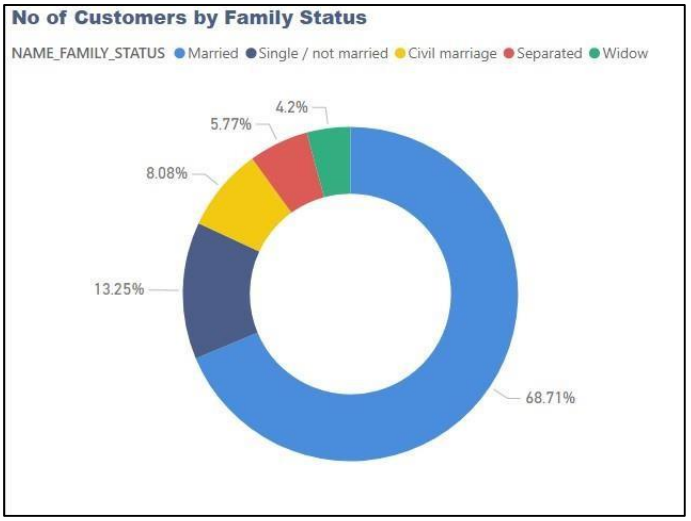


Figure 4.2– No of Customers by Family Status

According to the 4.2 graph No of Customers by family status 69% are married, 13% are single, 8% are civil marriage, 5% are separated and 4% are widows.

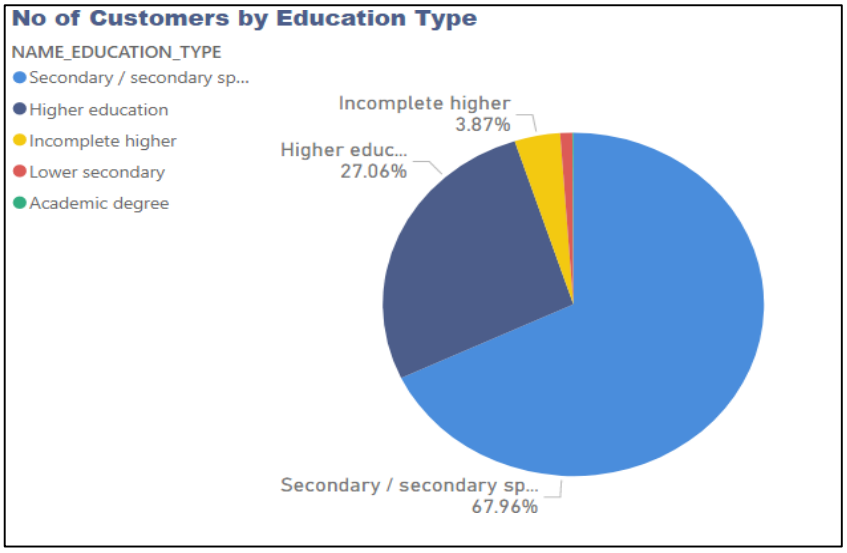


Figure 4.3– No of Customers by Education Type

As shown in figure 4.3 No of Customers by Education Type figure 68% have secondary education, 27% are have higher education and 4% are have incomplete higher education.

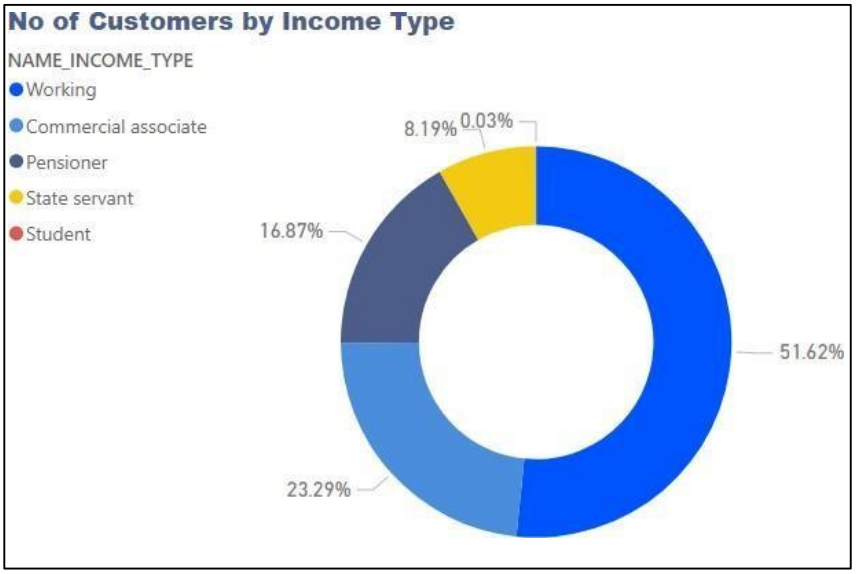


Figure 4.4 – No of Customers by Income Type

According to the graph 4.4 No of Customers by Income Type, 62% are working, 23% are commercial associates, 16% are pensioner, 8% State servant.

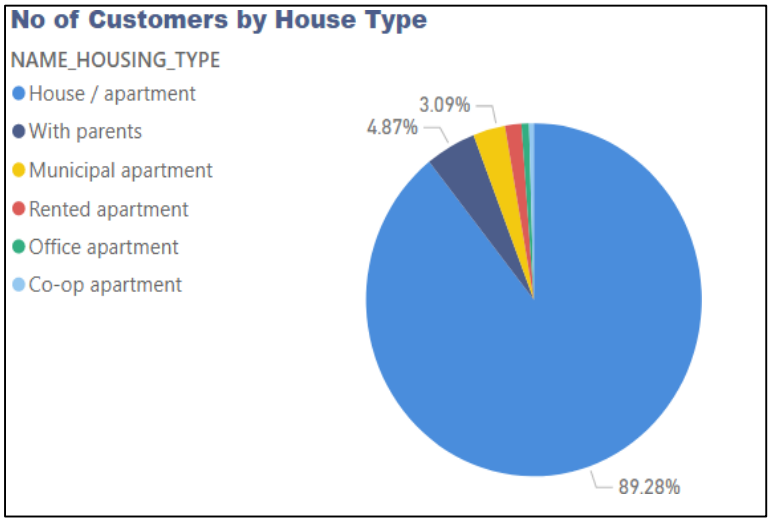


Figure 4.5– No of Customers by House Type

According to the graph 4.5 No of Customers by House Type, 89% are have house, 5% living with parents, 3% are live in municipal apartment.

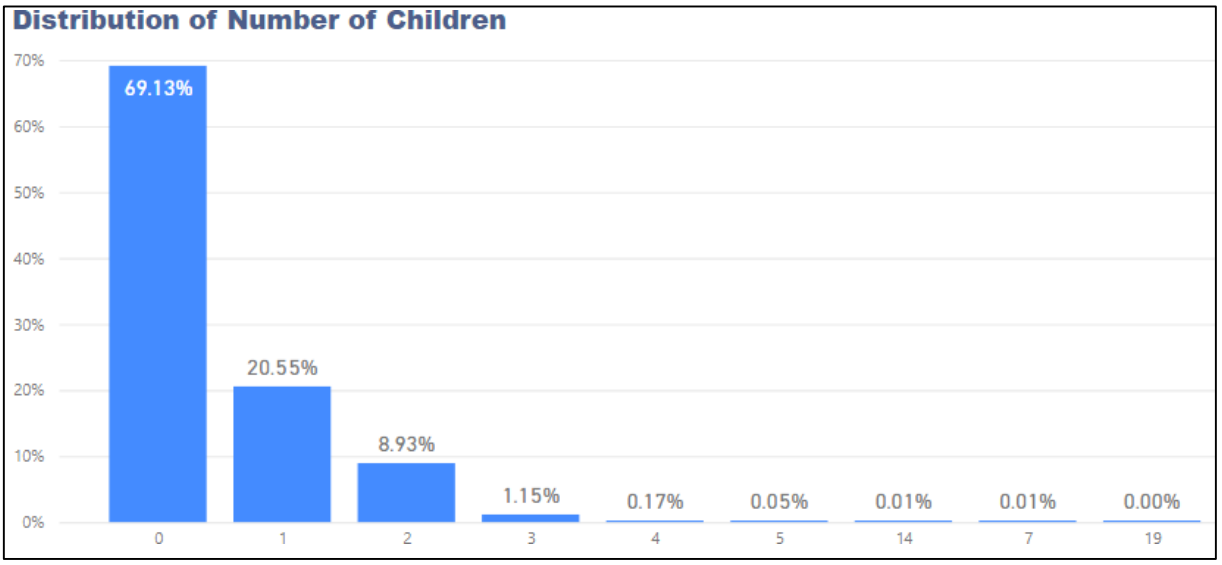


Figure 4.6– Distribution of Number of Children

As shown in 4.6 graph Distribution of Numbers of Children, 69 % are not have a child. 20% have one child, 9% have two children and 1% have 3 children. Remaining 0.5% have more than 3 children.

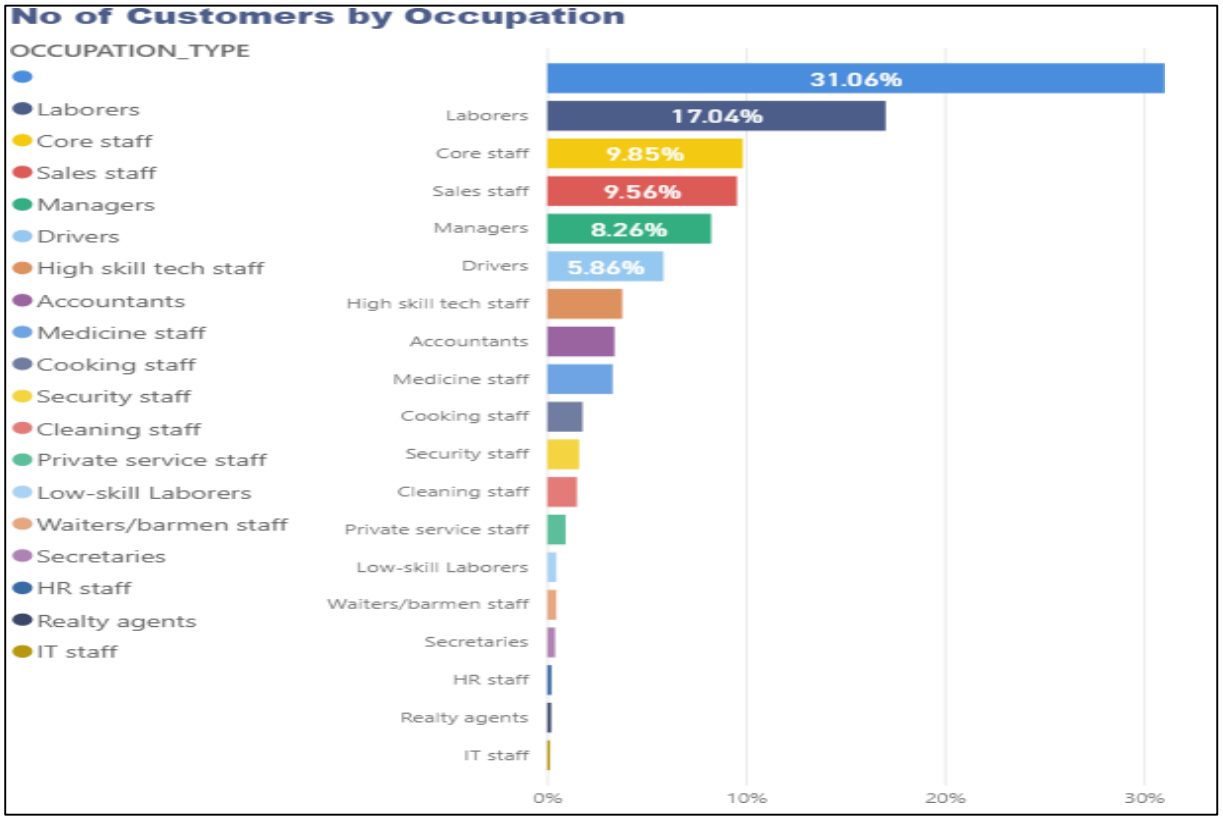


Figure 4.7– No of Customers by Occupation Type

According to the above graph 4.7 Distribution of occupation, 31% have missing values (occupation not mentioned), 17 % are labors, 10% are core staff, 10% sales staff, 8% are managers and 6% are high skill tech staff.

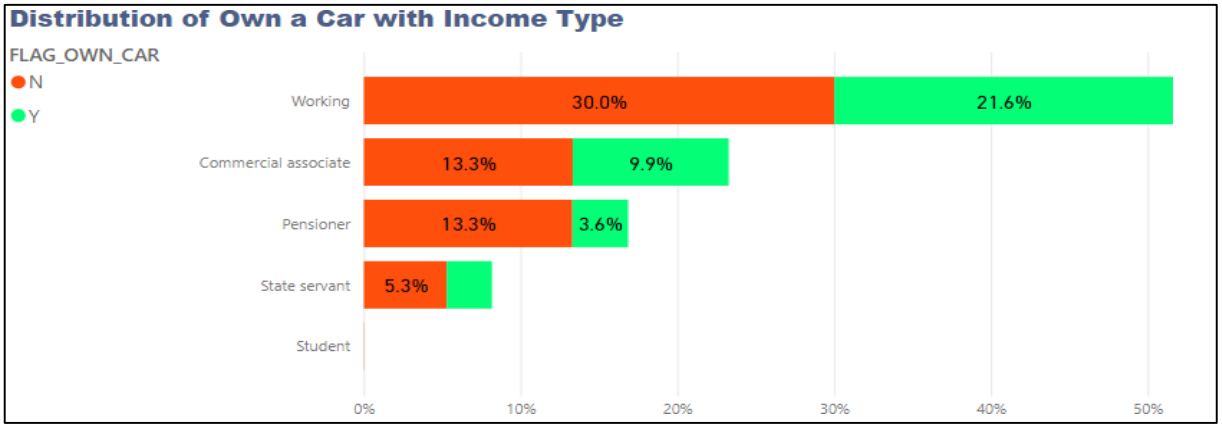


Figure 4.8– Distribution of Own a Car with Income Type

According to the above graph 4.8 Distribution of own a car with income type, 50 % are working customers and 30 % out of them does not own a car and 20% own a car. 20 % are commercial associates and 13 % out of them does not own a car and 10% own a car. 15 % are pensioner and 13 % out of them does not own a car and 4% own a car.

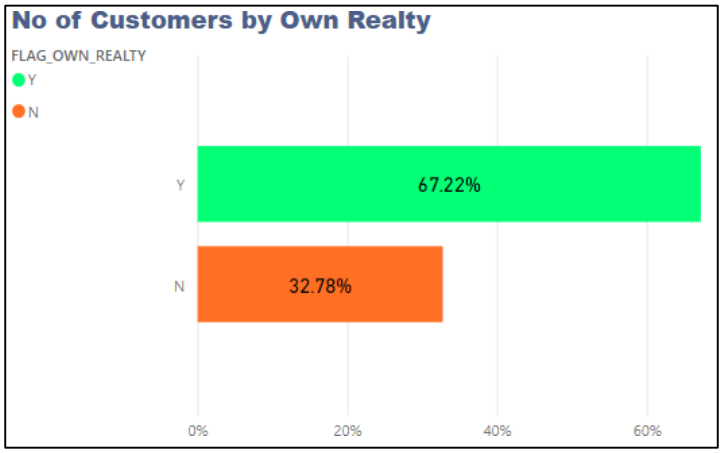


Figure 4.9– No of Customers by Own Realty

As shown in 4.9 graph No of Customers by Own Realty, 67% are own a realty and 33% are not own a realty.

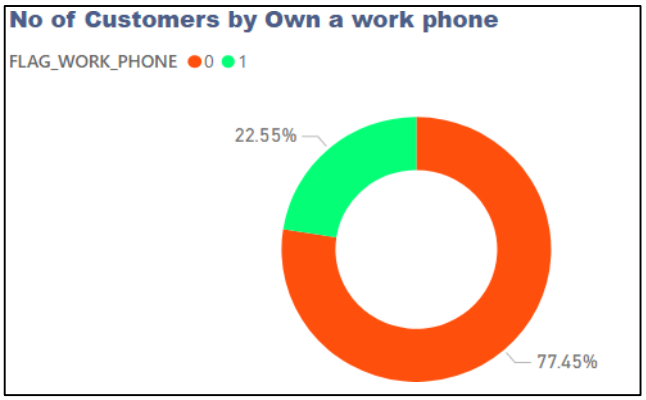


Figure 4.10– No of Customers by own a work phone

According to the graph 4.10 No of Customers by Own a work phone, 77% are not own a work phone and 23% are own a work phone.

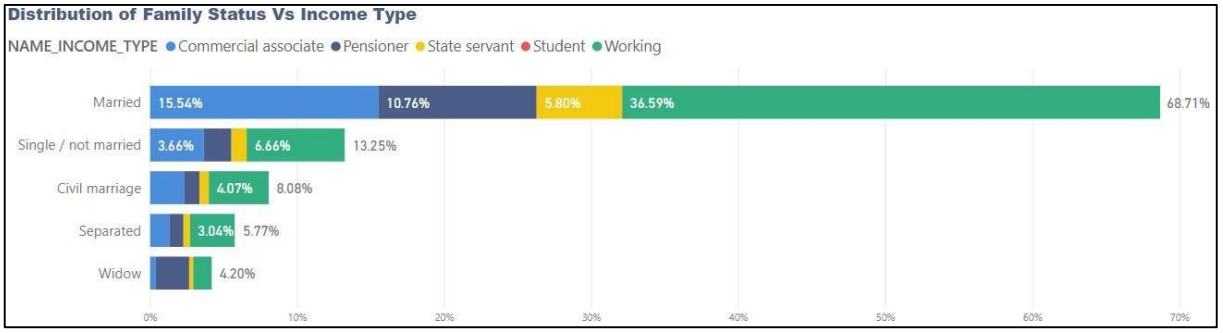


Figure 4.11– Distribution of Family Status Vs Income Type

As shown in the graph 4.11 68% are married and 37% of them are working, 15% are commercial associates, 11% are pensioner and 6% are state servants.

4.2 Data Preparation Activities

Data preparation activities related to our project discuss separately below.

4.2.1 Clean Data

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.

4.2.2 Handling Missing Value

In this data set there are missing values in OCCUPATION_TYPE column and hence this is a categorical column replaced missing values with ‘Other’.

4.2.3 Construct Data

This data set doesn’t 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.

Methodology of Deriving a Dependent Variable (Class Variable)

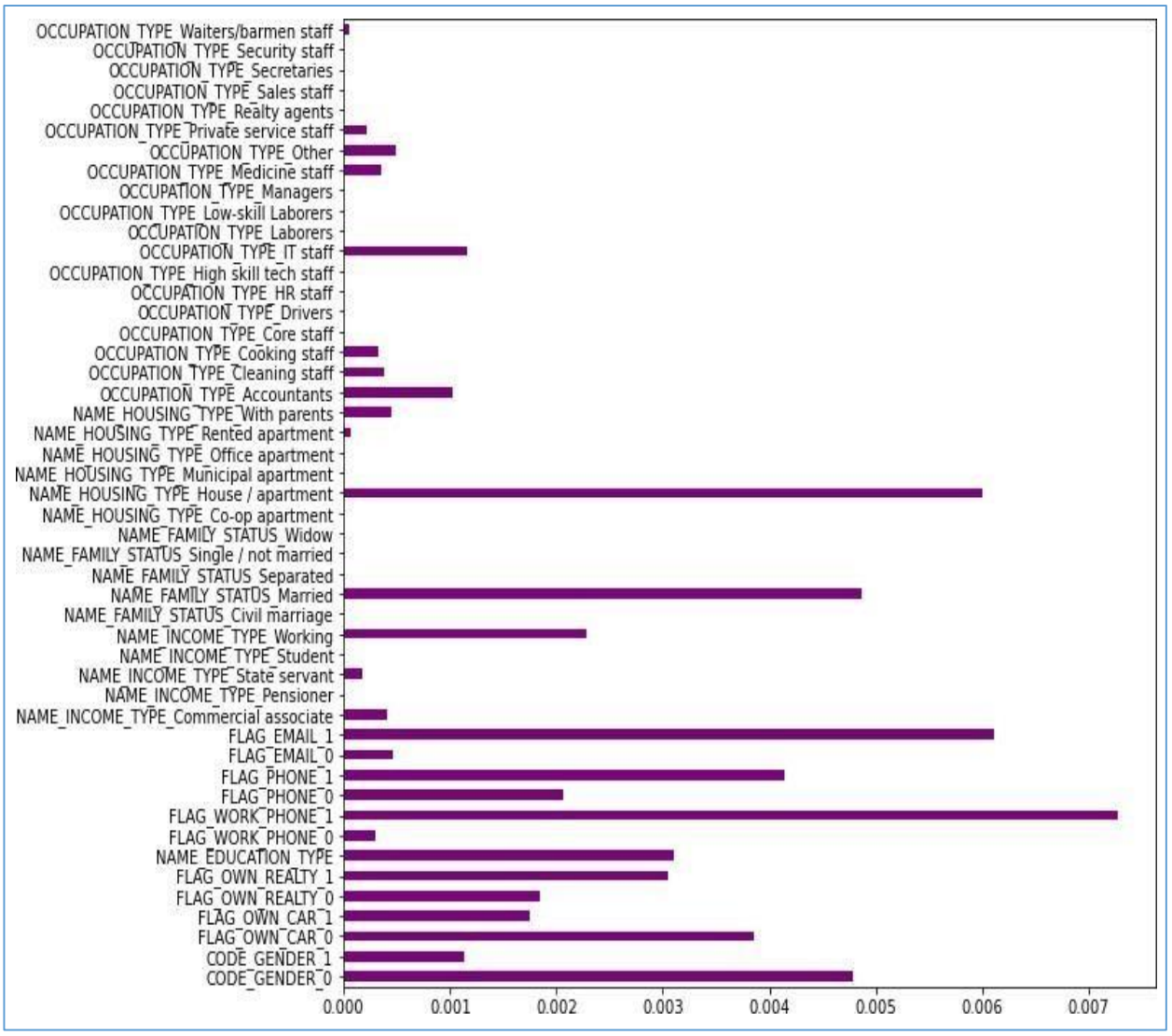
4.2.4 Integrated Data

In here we have two data set combined together. Below figure 4.12 illustrated combining of two data set

Information Gain

sklearn.feature_selection import mutual_info_classif use to calculate information gain and figure 4.23 shown categorical features.

Figure 4.12 - Categorical features with Mutual Information gain



4.2.5 Feature Scaling

We have applied Standard scaler for numerical features. Below numerical features have been transformed.

- AMT_INCOME_TOTAL
- NAME_EDUCATION_TYPE
- CNT_FAM_MEMBERS
- EMPLOYED_IN_YEARS
- AGE_IN_YEARS
- CNT_CHILDREN

4.3.1 Application of Support Vector Machine

- LinearSVC – This is similar to SVC with parameter kernel=’linear’. But more flexible and scalable to large number of samples.
- Tol – This is tolerance for stopping criteria. Default is 1e-5
- Verbose - Enable verbose output.
- Max_iter - The maximum number of iterations to be run.
- Kernel – This is to specify the kernel type to be used in the algorithm. Default is Radial Basis Function (RBF)

Confusion Matrix and Classification Report of Validation (Test) and Training Data – High Learning Rate (0.001)

The validation results we got for the predictions in validation and training data set by keeping parameters as epoch = 100, batch size = 100 and learning rate = 0.001 discussed below. Figure 4.15 shows confusion matrix. We can see false negative predictions are 1770 and false positive predictions are 1400 on testing data set. There are 6844 false negative predictions and 5771 false positive predictions in training data set.

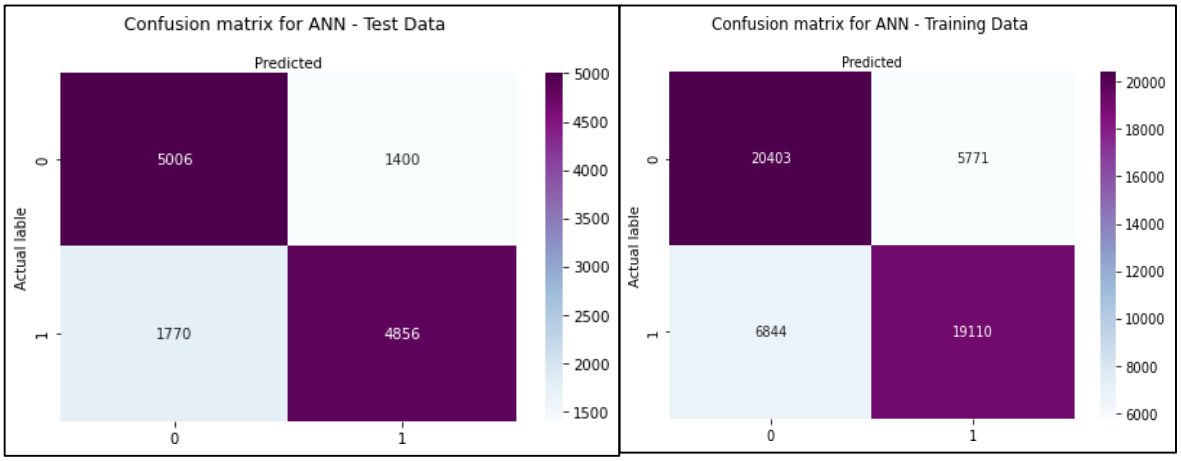


Figure 4.15 - Confusion Matrix on Validation and Training Data Set – ANN

Classification report - ANN Training Data

	precision	recall	f1-score	support
0	0.75	0.78	0.76	26174
1	0.77	0.74	0.75	25954
accuracy			0.76	52128
macro avg	0.76	0.76	0.76	52128
weighted avg	0.76	0.76	0.76	52128

Accuracy: 0.7579995395948435
Precision: 0.768055946304409
Recall: 0.736302689373507

Classification report - ANN Test Data

	precision	recall	f1-score	support
0	0.74	0.78	0.76	6406
1	0.78	0.73	0.75	6626
accuracy			0.76	13032
macro avg	0.76	0.76	0.76	13032
weighted avg	0.76	0.76	0.76	13032

Accuracy: 0.7567526089625537
Precision: 0.7762148337595908
Recall: 0.7328705101116813

Figure 4.16 - Classification Report on Validation and Training Data Set – ANN

Figure 4.17 shows classification report of predictions in validation and testing data set. We can see Accuracy is 0.76, Precision is 0.76 and Recall is 0.74 in validation data set. In training data set Accuracy is 0.76, Precision is 0.77 and Recall is 0.73. In our model validation accuracy and training accuracy are almost same.

Area Under the Curve (AUC) shown in figure 4.17 for validation and training data set under different learning rates.

4.5 Deployment

After going through the model validation process, it was highlighted that Nonlinear SVM performs better than others. Therefore, we have decided to deploy the prediction model by using Nonlinear SVM. Firstly, we saved the nonlinear classification model to a pickle file. Here, we are saving our training model and will be using this for deploying the model. Use ‘streamlit’ for model deployment. Streamlit is an open-source python library which is easy to use and we can create beautiful web apps. There are two options. To predict a single customer entry and to predict a bulk set of data. Python script with streamlit created to implement the application. Samples of screen images of the application discuss in below.

CHAPTER 5

CONCLUSION

5.1 Conclusion

In conclusion, the prediction of credit card approvals project offers substantial insights and benefits for the financial industry. Through the utilization of historical data and advanced predictive modeling techniques, institutions can enhance their risk management practices, streamline application processes, and improve customer satisfaction levels. Moreover, the ability to tailor marketing strategies to specific customer segments, detect and prevent fraudulent activities, ensure regulatory compliance, and optimize card portfolios further underscores the significance of this project. By leveraging predictive analytics in credit card approval processes, financial institutions can make more informed decisions, minimize financial risks, and ultimately create a more resilient and customer-centric banking environment. As technology continues to advance and data-driven approaches become increasingly prevalent, the insights gained from this project will undoubtedly serve as a valuable foundation for future endeavors in the ever-evolving landscape of financial services.

Moreover, the project underscores the growing importance of responsible and ethical data usage in the financial industry. As institutions leverage vast amounts of customer data for predictive purposes, ensuring data privacy, security, and transparency becomes paramount. By adhering to industry best practices and regulatory guidelines, financial institutions can foster trust with their customers and stakeholders, thereby strengthening their brand reputation and credibility in the market.

5.2 Future Work

Predicting credit card approvals through data analysis and predictive modeling offers multifaceted benefits for future endeavors in the financial sector. By leveraging historical data and sophisticated algorithms, institutions can more accurately assess the creditworthiness of potential cardholders, thereby enhancing risk management practices and reducing potential losses from unpaid debts. Moreover, streamlined application processes resulting from these predictive models can significantly improve the overall customer experience, leading to higher satisfaction levels among applicants. Additionally, insights gleaned from credit card approval predictions can inform targeted marketing strategies, enabling institutions to tailor their offerings to specific customer segments effectively. Furthermore, these models play a crucial role in fraud detection, helping to safeguard against fraudulent activities and minimize financial risks. Compliance with regulatory requirements is also bolstered through predictive analytics, ensuring adherence to stringent anti-money laundering and know your customer regulations. Lastly, the analysis of credit card approval data aids in portfolio management, allowing institutions to optimize their card offerings and better cater to the diverse needs of their clientele. Overall, investing in credit card approval prediction projects lays a solid foundation for informed decision-making, risk mitigation, and improved customer strategies.

REFERENCES

[1]Anna B. Holm. E-recruitment: "Towards an ubiquitous recruitment process and candidate relationship management. German" Journal of Human Resource Management, 26(3):241– 259, 2012.

[2] Lori Foster Thompson, Phillip W. Braddy, and Karl L. Wuensch. "E-recruitment and the benefits of organizational web appeal. Computers in Human Behaviour", 24(5):2384 – 2398, 2008.

[3] Peter Kuhn and Hani Mansour. "Is internet job search still ineffective?" The Economic Journal, 124(581):1213–1233, 2014 .

[4] Constantin Mang. "Online job search and matching quality. Technical report, Ifo Working Paper, 2012. 5. Raquel Campos, Mara Arrazola, and Jos de Hevia. Online job search in the spanish labor market." Telecommunications Policy, 38(11):1095 – 1116, 2014.

[5] Linda Barber. "E-recruitment Developments." Institute for Employment Studies, 2006.

[6] Amit Singh, Catherine Rose, Karthik Visweswariah, Vijil Chenthamarakshan, and Nandakishore Kambhatla. Prospect:" A system for screening candidates for recruitment. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management", CIKM '10, pages 659–668, New York, NY, USA, 2010. ACM.

[7]Evanthia Faliagka, Athanasios Tsakalidis, and Giannis Tzimas. "An integrated e-recruitment system for automated personality mining and applicant ranking." 51

[8] Evanthia Faliagka, Konstantinos Ramantas, Athanasios K Tsakalidis, Manolis Viennas, Eleanna Kafeza, and Giannis Tzimas. "An integrated e-recruitment system for cv ranking based on ahp. In WEBIST," pages 147–150, 2011.

[9] V Senthil Kumaran and A Sankar. "Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping (expert). International Journal of Metadata, Semantics and On national Journal of Metadata, Semantics and Ontologies," 8(1):56–64, 2013.

[10] Bradford Heap, Alfred Krzywicki, Wayne Wobcke, Mike Bain, and Paul Compton. PRICAI 2014: "Trends in Artificial Intelligence: 13th Pacific Rim International Conference on Artificial Intelligence", Gold Coast, QLD, Australia, December 1- 5, 2014. Proceedings, chapter "Combining Career Progression and Profile Matching in a Job Recommender System," pages 396–408. Springer International Publishing, Cham,2014.

[11] Kush R. Varshney, Vijil Chenthamarakshan, Scott W. Fancher, Jun Wang, Dongping Fang, and Aleksandra Mojsilovic. “Predicting employee expertise for tal-´ ent management in the enterprise.” In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’14, pages 1729–1738, New York, NY, USA, 2014. ACM.

[12] Wenxing Hong, Siting Zheng, Huan Wang, and Jianchao Shi. “A job recommender system based on user clustering.” Journal of Computers, 8(8):1960–1967,2013.

[13] Vapnik, V. N. “The Nature of Statistical Learning Theory (2nd Ed.),” Springer Verlag, 2000.

[14] Susmita Ray," A Quick Review of Machine Learning Algorithms," 2019 International Conference on Machine Learning, 52 Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019 https:/ / doi.org/ 10.1109/ comitcon.2019.8862451 .

[15] S. Marsland, Machine learning: an algorithmic perspective. CRC press, 2015. L. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176,Oct., 2015 https:/ / doi.org/ 10.1109/ comst.2015.2494502