Customer Churn Analysis and Prediction

Badrun Nahar Bristy

Student Id: 012201027

A Project

in

The Department

of

Computer Science and Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Science in Computer Science and Engineering

United International University

Dhaka, Bangladesh

January, 2022

Approval Certificate

This project titled "**Customer Churn prediction and Analysis**" submitted by **Badrun Nahar Bristy**, Student ID: **012201027**, has been accepted as Satisfactory in fulfillment of the requirement for the degree of Master of Science in Computer Science and Engineering on 31-01-2022.

Board of Examiners

1.

_____ Supervisor

Prof. Dr. Mohammad Nurul Huda
Professor, CSE and Director, MSCSE
Department of Computer Science and Engineering (CSE)
United International University (UIU)
Dhaka-1212, Bangladesh

2.

_____ Examiner

Mr. Suman Ahmmed
Assistant Professor
Department of Computer Science and Engineering (CSE)
United International University (UIU)
Dhaka-1212, Bangladesh

3.

_____ Ex-Officio

Prof. Dr. Mohammad Nurul Huda
Professor, CSE and Director, MSCSE
Department of Computer Science and Engineering (CSE)
United International University (UIU)
Dhaka-1212, Bangladesh

Declaration

This is to certify that the work entitled "**Customer Churn Analysis and Prediction**" is the outcome of the research carried out by me under the supervision of **Dr. Mohammad Nurul Huda, Professor, CSE and Director -MSCSE, UIU**.

_____

Badrun Nahar Bristy
Student ID: 012201027
MSCSE Program
Department of Computer Science and Engineering (CSE)
United International University (UIU)
Dhaka-1212, Bangladesh

In my capacity as supervisor of the candidate's project, I certify that the above statements are true to the best of my knowledge.

_____

Prof. Dr. Mohammad Nurul Huda
Professor, CSE and Director, MSCSE
Department of Computer Science and Engineering (CSE)
United International University (UIU)
Dhaka-1212, Bangladesh

Abstract

One of the biggest data domain and most demanding use cases of recent time is Customer churn prediction. For a healthy and growing business churn prediction is an important indicator. This project aims to develop a churn prediction for banking sector. For predicting customer churn I have chosen hyper parameters of deep learning. I have collected a dataset from kaggle, which have 10000 rows and 14 columns. I divided the dataset into two parts. One is train data which have contains 75% data and another is test data which contains 25% data of the whole dataset. I have done some analysis on the data set. I have used deep learning hyper parameter. Using deep learning the trained model gives 79% accuracy. I have also used some machine learning algorithm such as Random Forest, Decision Tree, K-nearest neighbor (KNN) and Logistic regression. Among this four algorithm Random Forest has given better accuracy which is about 85%.

Acknowledgement

First and foremost, I would like to express my heartiest gratefulness to the Almighty Allah for establishing me to complete this project.

Then I would like to express my special thanks to my supervisor Dr. Mohammad Nurul Huda sir for his enormous support, skilled guidance, invaluable advice and motivation during this project. He has been my role model in the last couple of months and his advices helped me to overcome all the difficulties.

I would like to convey my special thanks to my examiner Suman Ahmed sir for his assistance and advises for new features of this project.

I thank my parents for giving birth to me at the first place and supporting me spiritually throughout my life.

I am also grateful to our Vice Chancellor Prof. Dr. Chowdhury Mofizur Rahman sir & Head of Department Prof. Dr. Salekul Islam sir for giving me the opportunity to complete my project.

Last but not least, to all people that I forgot to mention here, who have contributed towards the completion of this project either directly or indirectly. Your kindness and cooperation in completion of this project paper are very much appreciated. Thank you very much in advance and May Allah bless all of you.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1
Introduction

1.1 Introduction

The most valuable asset of an organization is their customer. A basic requirement for any organization is customer retention. Banks are also included within those organization. In this competitive market, the growing number of operators and increasing number of scientific progresses increase the level of oppositions. Companies are surviving because of customer churn and because of customer churn companies are facing various types of losses. In many kind of research, we can see that machine learning and deep learning technologies are highly efficient to predict customer churn.

The main goal of churn prediction is to detect the customers who intended to leave a service provider. Gaining new customer is more costly than retaining one customer. To identify the possible churners predictive models can provide correct identification. It can identify the churners in the near future and that can helps an organization a retention solution.

This project presents a prediction model using hyper parameter of deep learning model. I have also used some machine learning algorithm such as Random Forest, Decision Tree, K-nearest neighbor (KNN) and Logistic regression. The proposed model composed with 5 steps. Such as data selection, data analysis, data preprocessing, data dividing and model training. I have used a data set having 10000 instances. 75% of data is used for model training and 25% is used for testing. And I have done this project in google Colaboratory and Jupyter Notebook.

1.2 Motivation

The motivation of doing such a project is to learn Machine Learning and Deep Learning through an effective way. In this era, Machine Learning and Deep

learning algorithms are like shining star. These algorithms are solution of very critical problems. Machine learning and Deep Learning can predict anything for an organization. In this project These algorithms are used to predict churn rate of a banking sector.

1.3 Project Goals

The main goal of this project is to predict customer churn so that the banking sector can establish a healthy system to keep their customer long time and can reduce churning rate of customers. There are many variables, which are influencing churning rate of customer. In this project, I have tried to highlight these variables. The objective is to build prediction model, find out the accuracy and performance of the model using Machine Learning and Deep Learning. Finally, I have tried to find out and develop the best model.

1.4 Possible Applications

The churn rate of any organization needs to keep as low as possible. So that Customer Churn Prediction is a very important system for any organization. All organization's authority needs to know which customers are going to churn and why are they going to churn. Mostly the customer churn prediction project is applied in Banking Sector, Telecommunication Sector and Mobile Financial Organizations. In figure 1.1 we can see some possible applications of churn predictions.

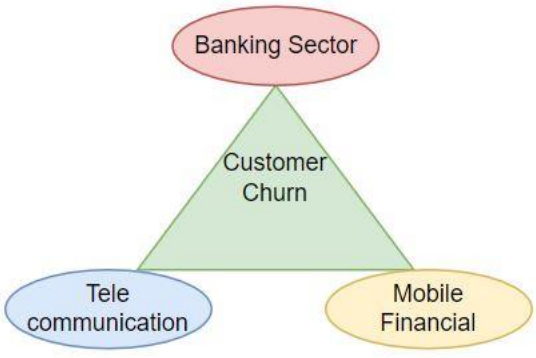


Figure1.1: All possible applications of Churn Predictions

1.5 Software Architecture

In this project work some analysis on the dataset is done. Before model building, I have done some data preprocessing and data set splitting. After that I have the model training part. And, finally by getting the best result I have shown the churn modeling is a User Interface. In Figure 1 we can the software architecture model of the project. In which we can see that, at first the data collection and data preparation is done. After that the Model testing, training and developing part has done. And, finally the result is shown by a user interface.

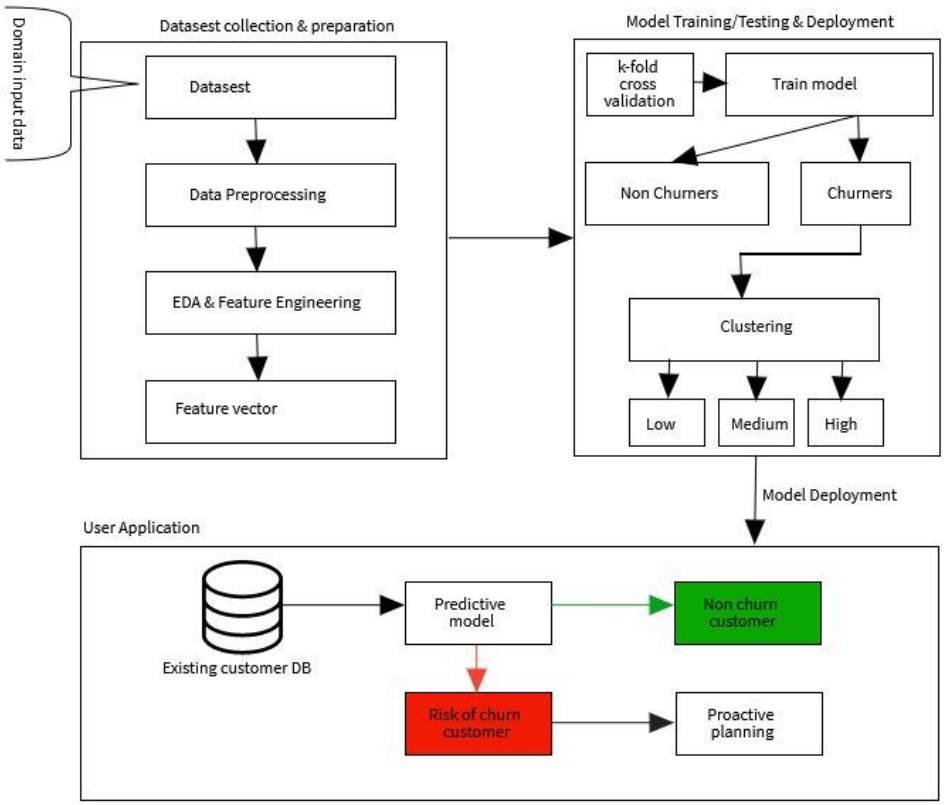


Figure1.2 : Software Architecture

1.6 Organization of the Report

In this section, we can see a short view of the overall report.

In Chapter 2, the Background and Literature Review parts are written. In This section, we can get a clear idea about all the Machine Learning and Deep Learning

algorithms those are used to build up the model for this project. Also, we can see some previous works review which are already done in different years.

In chapter 3, the methodology of the whole project is shown. You can see the Data Analysis, Data Preprocessing, Model Building and Performance Measure part of this project clearly in this chapter.

In chapter 4, the result and output part is clearly described. You can get idea about the performance and accuracy of each, and every algorithm used here. Also, you can find out the best algorithm for this project.

In chapter 5, this report has come to an end. This is the conclusion part.

Chapter 2

Literature Review and Background

2.1 Literature Review

Another name of customer churn is customer attrition. Customer churn can predict who are at risk of leaving any company. So, we can say that customer churn plays a vital role at customer retention. Mostly customer churn happens when customer could not be able to find any success with your company. Customer churn can give you a mathematical representation of how churn impact your business.

In a study 2020, V Kavitha et. al.[1] use various machine learning algorithm like Random Forest, Logistic Regression and XGBoost to predict the churn of customer. In 2017 Philip et. al. [2]apply deep learning (unsupervised) to find accurate values which help to predict the churn of customer. For predicting customer churn one study in 2015 Thanasis Vafeiadis et. al.[3] used Artificial Neural Network (ANN), Support Vector Machine (SVM) , Decision Tree (DT), Naïve Bayes and Logistic Regression. In 2021 another study which is done by Praveen Lalwani[4]. In this study, they apply different machine learning algorithm to predict customer churn. They have used logistic regression, naive bayes, support vector machine, random forest, decision trees, etc. Another study in 2016 which is done by Preeti K. Dalvi[5]. In this study they tried to analysis churn prediction for a Telecom Industry. They used Logistic Regression to get best result. They also used Decision Tree. In 2020 Ishpreet Kaur has done a study on customer churn in banking system[6]. IN this study, they have used Decision Tree, KNN, Random Forest, Linear Regression to predict model accuracy.

Table 2.1: Summary of Literature Review

Year	Author	Using Algorithms
2020	V Kavitha	Random Forest, Logistic Regression and XGBoost
2017	Philip	Deep Learning
2015	Thanasis Vafeiadis	ANN, SVM, DT,NB,LR

2021	Praveen Lalwani	SVM, DT,NB,LR
2016	Preeti K. Dalvi	Logistic Regression, DT
2020	Ishpreet Kaur	DT, KNN, RF, LR

In Table 2.1 You can see a summary of Literature review.

2.2 Background Study

In this project, I tried to predict customer churn rate of a banking system. The main objective of this project is to build and efficient model to predict whether a account holder is going to churn or not in near future. In banking sector, customer churn is depended on verity of things. Such as customer age, gender, geography, salary, credit card holder, high interest rate etc.

I used deep learning hyper parameter and some machine learning tool such as Random Forest, Decision Tree, K-nearest neighbor (KNN) and Logistic regression. In the following section you can see the algorithms in a nutshell.

2.2.1 Hyper Parameter

The key to machine learning is hyper parameter tuning. For a learning algorithm hyper parameter chooses a set of optimal hyper parameter. Before the beginning of learning process, hyper parameter's value is already set. Perform hyper parameter tuning is by using an exhaustive grid search from Scikit learn is one traditional and popular way. Every possible combination of hyper parameter is tried by this method. This method can find out the best set of values. Hyper parameters are eventual because they governance the in-total dealing of a machine learning model. The main objective is to bring an appeasement summation of hyper parameters that reduces a default loss function to provide better conclusions.

2.2.2 Random Forest

A very useful supervised learning algorithm is random forest algorithm. Random forest is used for classification. Also, it is used for regression. But the Random Forest is basically used for classification problem. Random forest is an ensemble learning. This ensemble learning made of various

decision trees and can predict a better accuracy by majority voting. Figure 2.1 shows Random Forest block diagram.

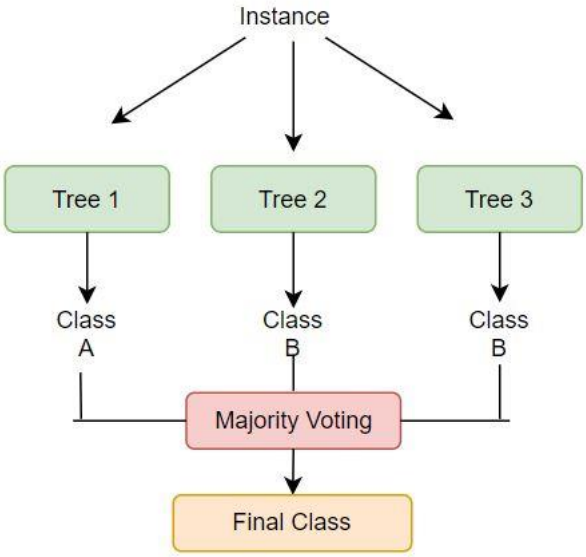


Figure 2.1: Random Forest Algorithm

2.2.3 Decision Tree

Decision Tree algorithm is very powerful and popular for classification and prediction. A Decision tree is like tree structure, where each internal node denotes a test on an attribute, each branch represents outcome of the test, and each leaf node holds a class label. In Figure 2.2 I stated a basic block diagram for Decision Tree Algorithm.

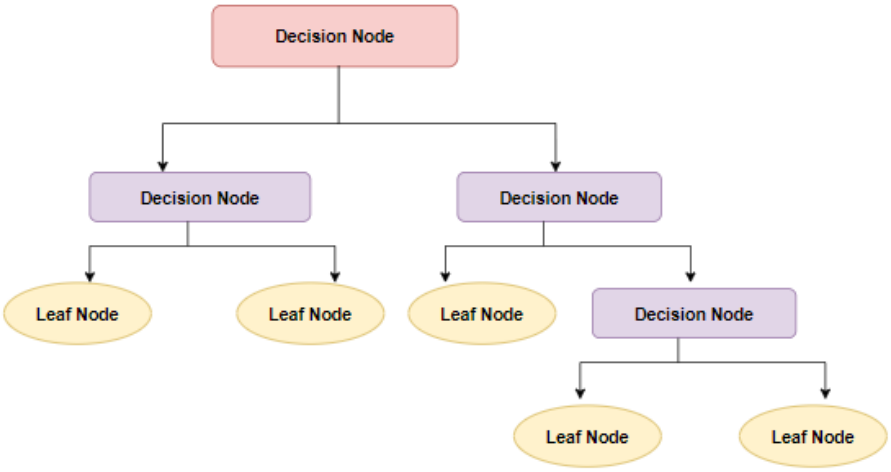


Figure 2.2: Decision Tree Algorithm

2.2.4 KNN

Another simple and supervised machine learning algorithm is k-Nearest Neighbor. Basically, this algorithm is used for classification and regression problems. The implementation and understanding of this algorithm is very easy. Based on similarity measure KNN algorithms use data and classify new data point. Classification is happened by a majority voting of neighbors. There are some areas where KNN can also be used such as speech recognition, Handwriting detection, image, and video recognition. In Figure 2.3 we can see a block diagram for KNN algorithms.

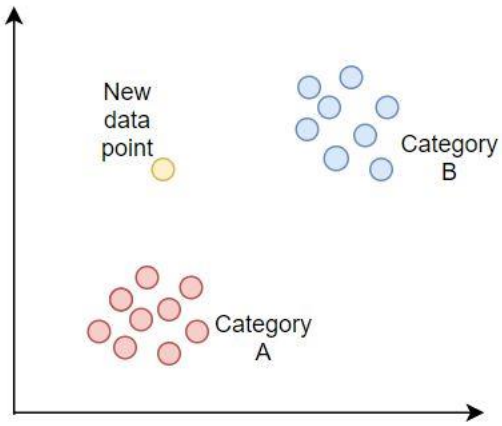


Figure 2.3: KNN Algorithm

2.2.5 Logistic Regression

Logistic regression is a simple Machine Learning algorithm. It is a supervised learning classification algorithm. This algorithm is used for predicting the probability of a target variable. In various classification problems, Logistic Regression can be used. The classifications problem can be spam detection, Diabetes prediction, cancer detection etc. Logistic regression is very important for Machine Learning. It can predict the output of a categorical dependent variable. This algorithm has the ability to provide probabilities. Using continuous and discrete dataset, it can also classify new data. That is why it is a significant machine learning algorithm.

Chapter 3

Methodology

3.1 Data Analysis

In this section, I have done some analysis on the whole dataset. There are 14 features as column and 10000 rows in the dataset. In figure 3.1 we can see 1st 5 rows of the whole data set. RowNumber, CustomerID, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProduct, HasCrCard, IsActiveMember, EstimatedSalary, Exited these are the 14 features. Here Exited represents Churn.

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Figure 3.1: 1st 5 rows of the Data Set

I have found some correlation between all features. In figure 3.2 we can see the correlations. Negative (-) sign means negatively proportional and positive sign represents positively proportional. For example, if we consider the correlation between age and exit. Then we can see that it is positively proportional. The customers who have more ages the churn possibility of them are higher. And the ratio is 0.29.

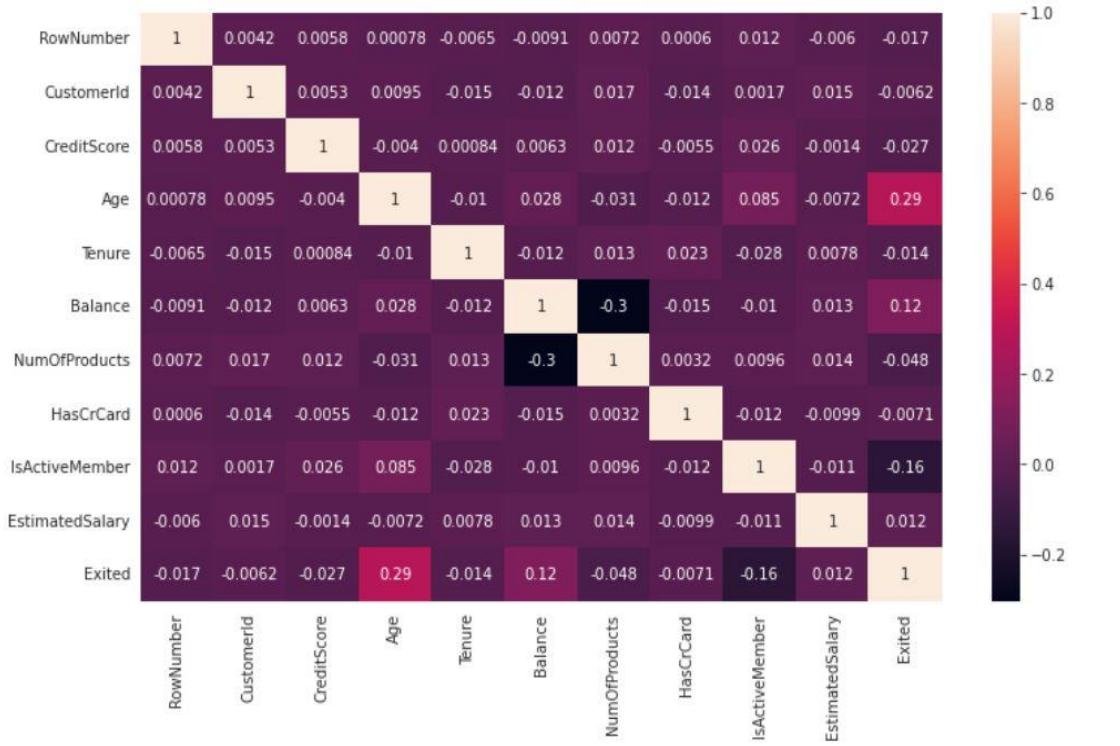


Figure 3.2: Correlation Between Features

In Figure 3.3 we can see the count of customer according to geography. In this dataset there are three country such as France, Spain, Germany. In France there are 5014 customers, in Germany 2509 customers and in Spain 2477 customers.

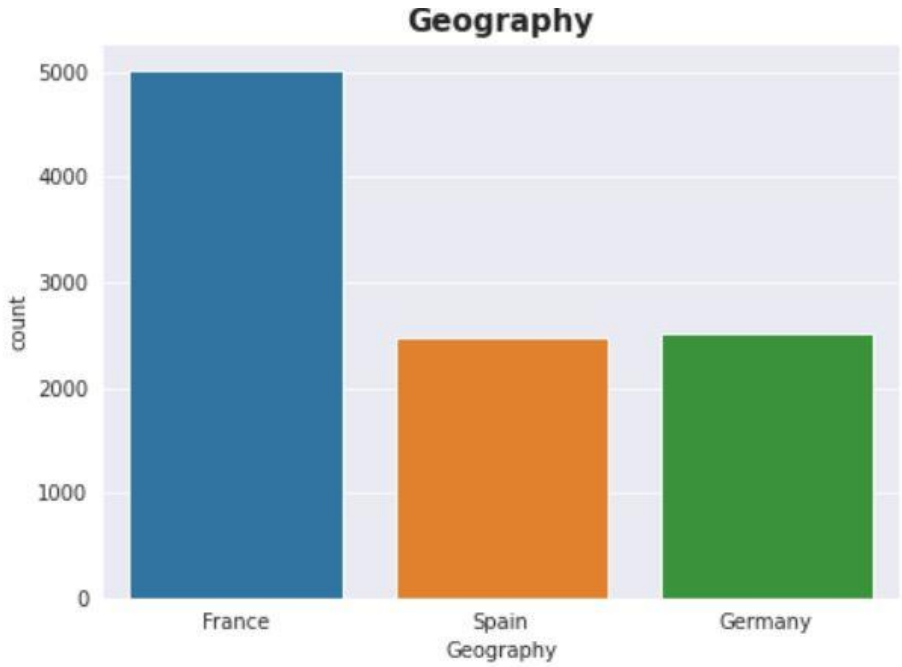


Figure 3.3: Customer Count According Geography

Figure 3.4 indicates the churn rate according to geography. The blue line represents not churner, and the orange line represents churner.

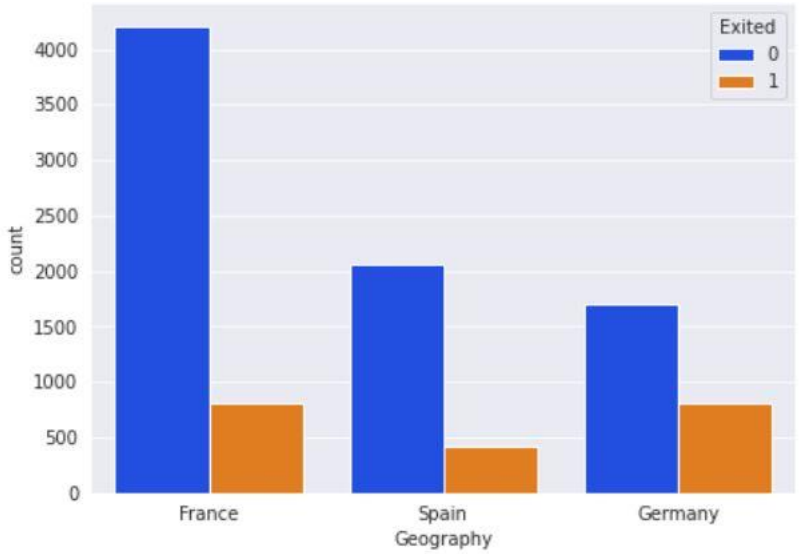


Figure 3.4: Churn Rate According to Geography

Figure 3.5 shows the churn rate according to gender. Here the blue line represents not churners and orange line represents churner. The churn rate of female is higher than male.

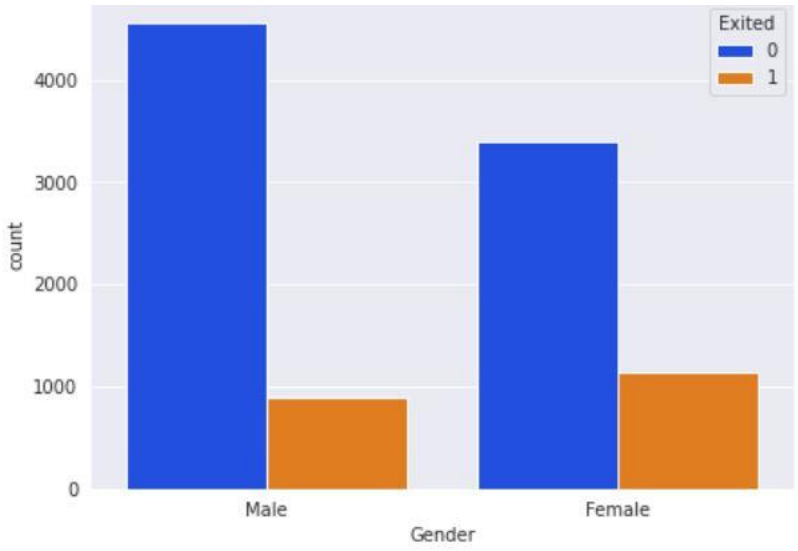


Figure 3.5: Churn Rate According to Gender

In Figure 3.6 we can see the churn rate according to age. Here the blue line represents not churners and orange line represents churner. In this figure we can find out that the age between 40 to 50 are more churn than others.

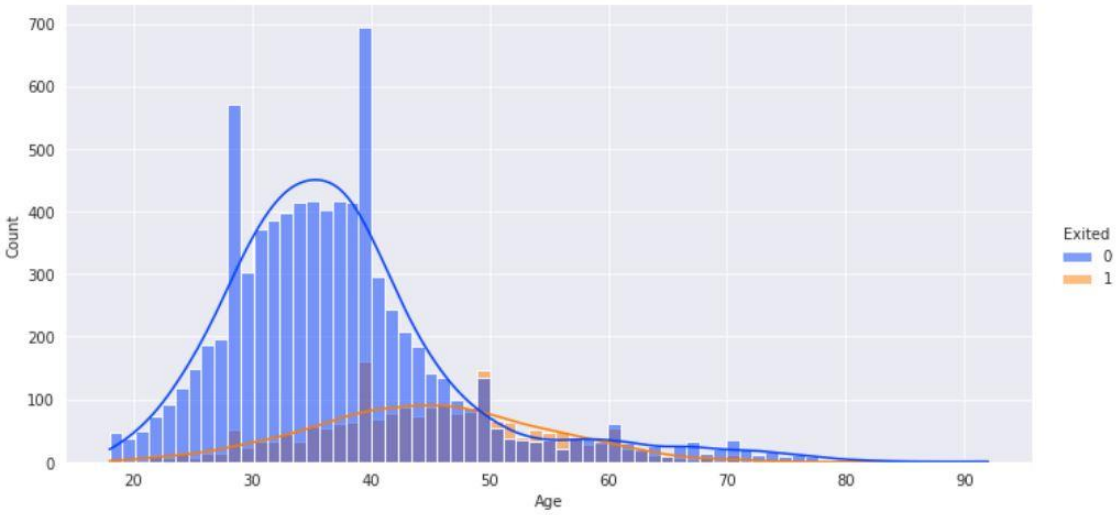


Figure 3.6: Churn rate according to Age

Figure 3.7 represents the number of products that used by the customers. We can see in this figure that the count of using one product customer is higher. Using two products of customer is also high. A very little group of customers are using three products and the number of using four products customer is very less.

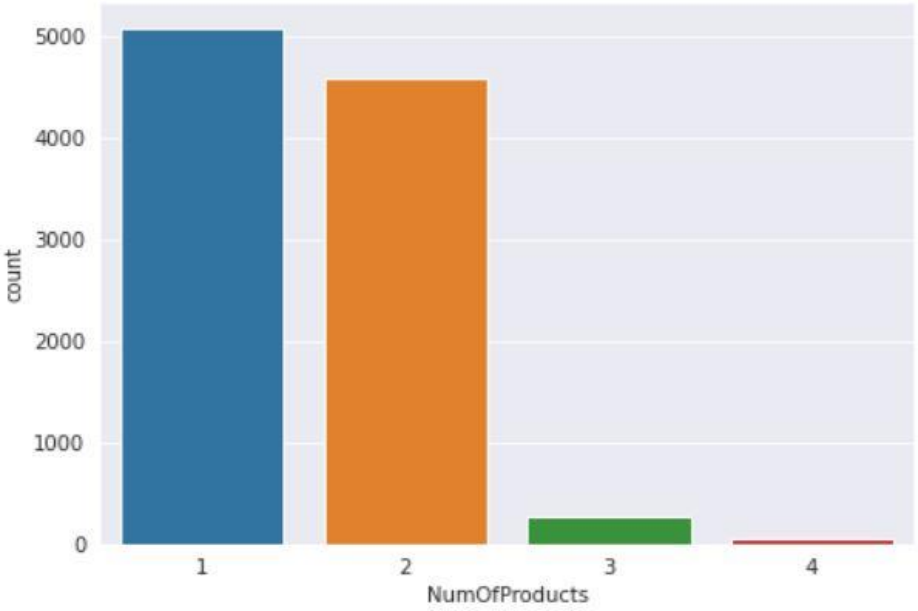


Figure 3.7: Number of Using Products

Figure 3.8 represents the churn rate according to using number of total products. Customers who are using 4 products their churning rate is 100%. Using 3 products customer's churning rate is higher than not churn.

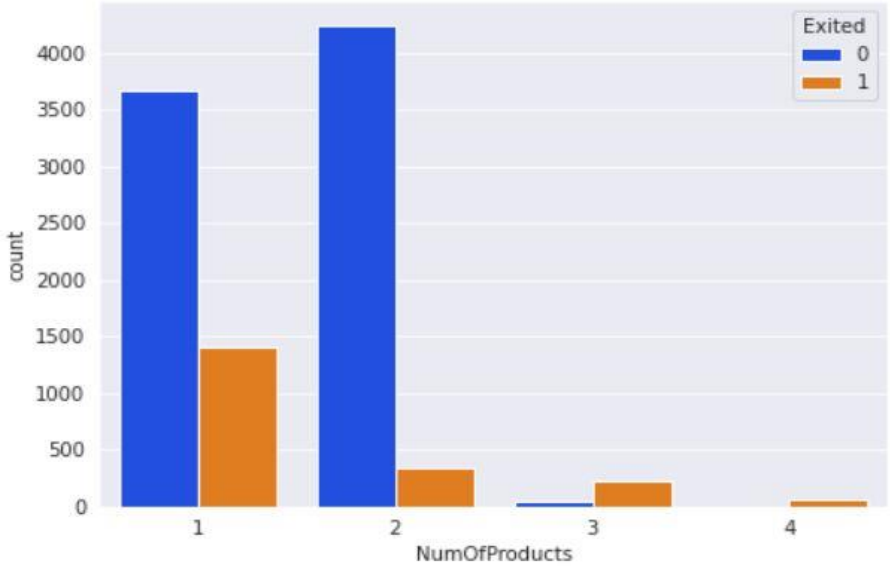


Figure 3.8: Churn Rate According to Using Number of Product

In Figure 3.9 we can see that the tenure count. The number of customers is less who are using products less than one year tenure.

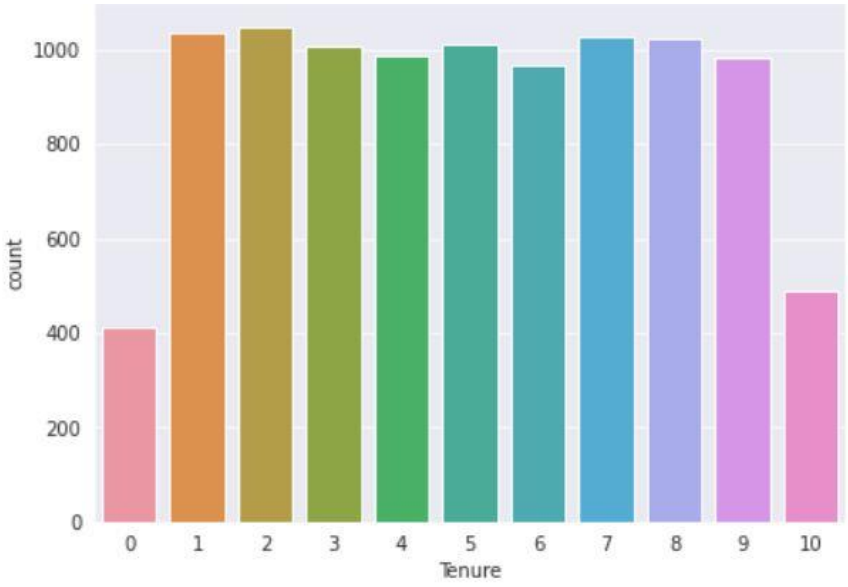


Figure 3.9: Count customer according to Tenure

Figure 3.10 represents the count of churner and non-churner according to tenure. Here the blue line represents not churners and orange line represents churner.

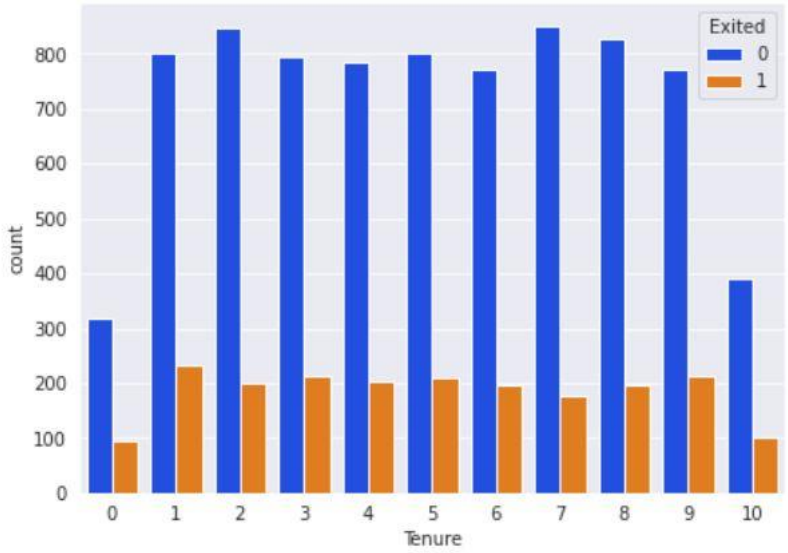


Figure 3.10 : Count Churner According to Tenure

Figure 3.11 shows the count of churner and non-churner according to Active member. Here the blue line represents not churners and orange line represents churner. In this figure '1' means active member and '0' means not active. We can see that the rate of non-churner of active member is higher than active member. And the churn rate of non-active member is higher than active member.

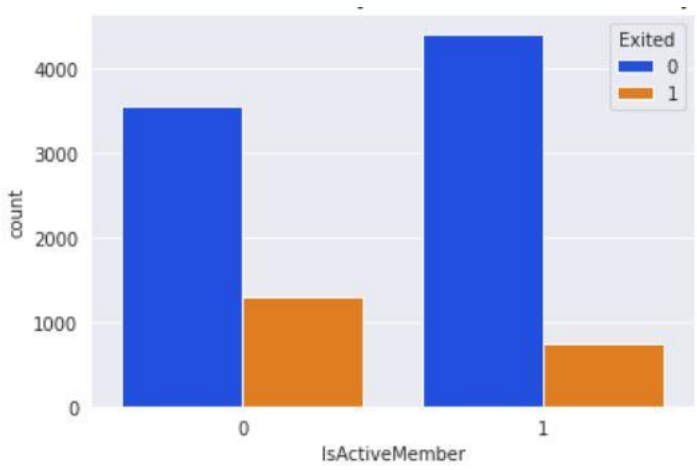


Figure 3.11: Count Churner According to Active Member

In Figure 3.12 we can see that the churn rate according to credit score. Here the blue line represents not churners and orange line represents churner. We can see that the people who has credit score 600 to 700 has high chance to churn.

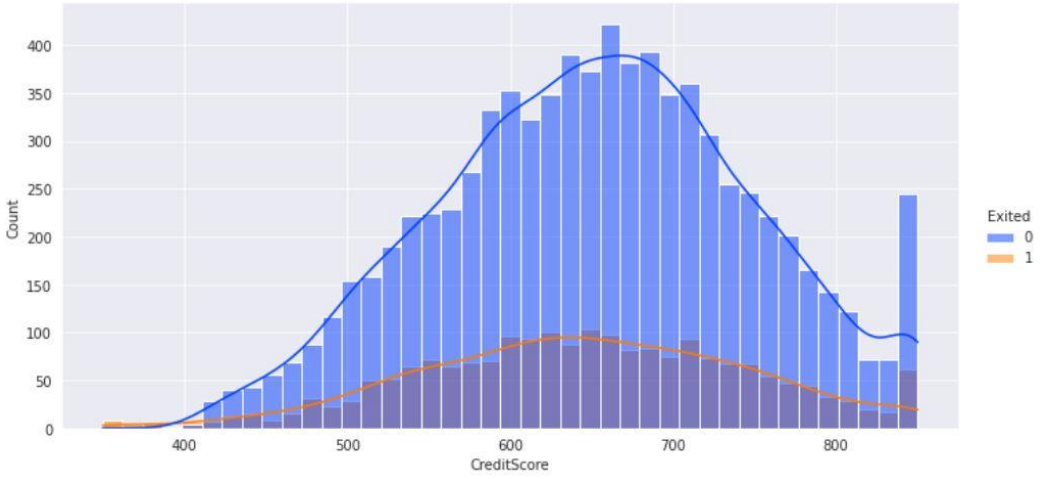


Figure 3.12 : Churn Rate According to Credit Score

Figure 3.13 represents the churn rate according to balance. Here the blue line represents not churners and orange line represents churner. The figure shows that the customers who have balance between 100000 to 150000 have high chance to churn.

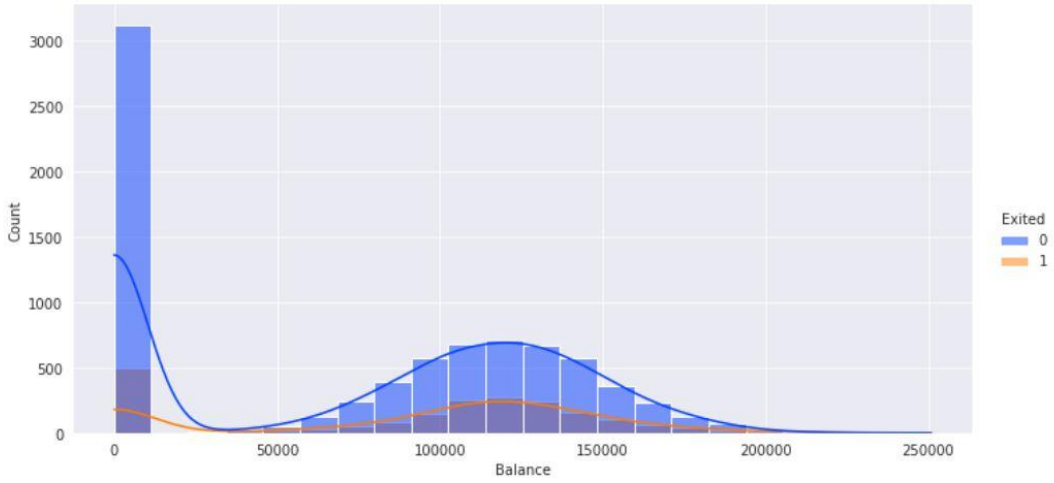


Figure 3.13 : Churn Rate According to Balance

Figure 3.14 shows the churn rate according to estimated salary. Here the blue line represents not churners and orange line represents churner.

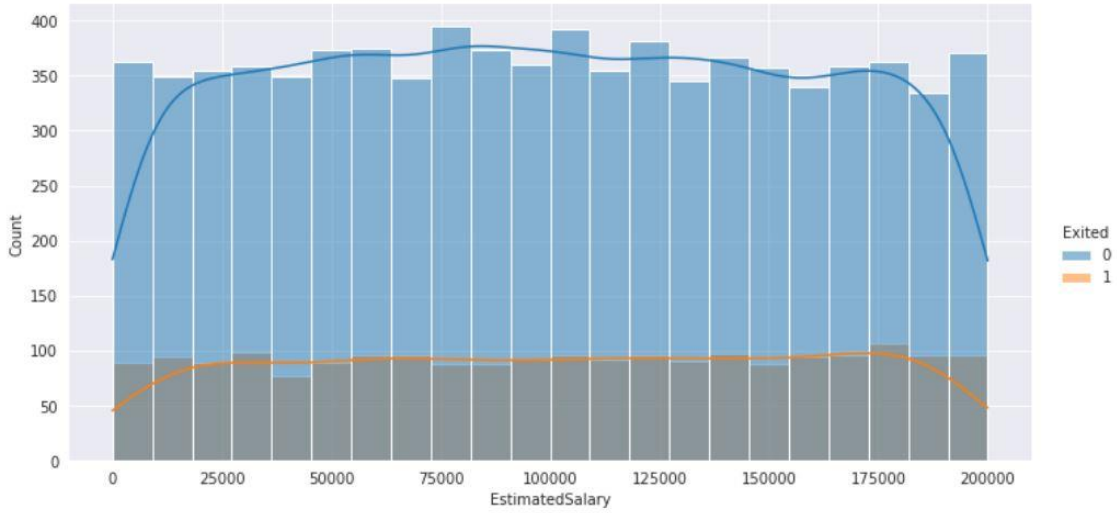


Figure 3.14 : Churn Rate According to Estimated Salary

3.2 Data Preprocessing

In this section, I have done data preprocessing. Data pre-processing is a very important part for any project because the success rate of the project is directly dependent on the data set. The data set get more suitable by doing preprocessing and it will give better accuracy.

In data preprocessing part I have dropped some unnecessary features, divide independent and dependent variables, sampling and feature scaling. Now I am

trying to show all of these parts briefly. Figure 3.15 shows a visual representation of data preprocessing.

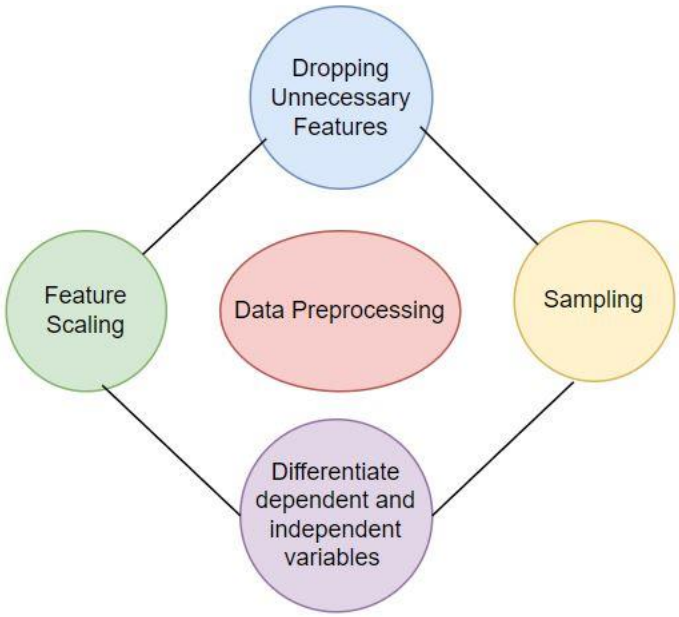


Figure 3.15: Data Preprocessing

3.2.1 Dropping Unnecessary Features

For getting better accuracy, we need a suitable dataset. To make the dataset suitable for this project I have dropped some features. Earlier the dataset has 14 features. I have dropped 3 features. These three features are RowNumber, CustomerId and Surname. After dropping these three features the data set now have 11 features and 10000 columns. In Figure 3.16 we can see the 1st 5 rows of these new preprocessed dataset.

	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	619	0	1	42	2	0.00	1	1	1	101348.88	1
1	608	1	1	41	1	83807.86	1	0	1	112542.58	0
2	502	0	1	42	8	159660.80	3	1	0	113931.57	1
3	699	0	1	39	1	0.00	2	0	0	93826.63	0
4	850	1	1	43	2	125510.82	1	1	1	79084.10	0

Figure 3.16 : 1st Five Rows of Preprocessed Dataset

3.2.2 Dividing Independent and Independent Variables

Dependent variable usually depends on another variable. Mostly they depend on independent variable. On the other hand, independent variable does not depend on other variables. In this project the exits feature is dependent variable. Exits feature represents churner. This feature is dependent variable because it depends on the other features. A customer will be churned or not churned it totally depends on the other independent features. Here I have divided the independent and dependent variables and put the independent variables in ‘x’ variable and in ‘y’ variable put the independent variables.

3.2.3 Sampling

In statistical analysis, sampling is an important process. In this process from a larger population, a predetermined number of observations are taken. Depending on the type of analysis the methodology used to sample from a larger population. The analysis can include simple random sampling or systematic sampling.

Sampling is useful in machine learning. Specially sampling is important for analyzing too large dataset. In big data analytics applications analyzing the whole dataset is too much costly. On the other hand, analyzing a representative sample is cos effective and more efficient.

In this project, I used over sampling to get better accuracy. From the minority class oversampling involves randomly selecting examples and add them to the training dataset.

After that, I have split the data set into two parts. These two parts are train dataset and test dataset. Train dataset contains 75% data and test dataset 25% data.

3.2.4 Feature Scaling

Feature scaling is also known as normalization. Usually feature scaling is performed in the data preprocessing steps. It is an important method. The range of feature data or independent variables are normalized by this method. Feature scaling can be done unit of the values. If we did not perform feature scaling, then machine learning algorithm considers greater values as higher values and tends to weigh smaller values as lower values. Feature scaling is necessary for any machine-learning algorithm. Because it can calculate distance between data. In raw data the range of values varies widely. For this reason, in some machine learning algorithms, without feature scaling the objective function do not work properly.

In Neural Network Algorithms, feature scaling gives better error surface shape. Feature scaling prevents the chances from getting stuck in local minima. It can also make the training faster. So that I have used feature scaling in the entire dataset of this project.

3.3 Model Building

In this project, model building is done by Deep Learning Hyper Parameter and different famous Machine learning algorithm. Random Forest, Decision Tree, KNN, Logistic Regression these are the Machine Learning algorithm which are used to build the model. I split the dataset into training and testing data. 75% of total data are in training data and rest of the 25% data are in testing data for performance evaluation of the model. In chapter 2 I have already described all these Machine Learning algorithms briefly. In the next section I will show the result and output part briefly.

3.4 Performance Measure

For measuring performance of a project there are various performance measures such as Accuracy, Precision, Recall and F1 Score. In this project, I have also used this performance measures. In any Machine Learning or Deep Learning project, Accuracy is a measurement which is used to determine the best model. Based on

the input or training data which model is best at identifying relationships and patterns between variables in a dataset is measured by accuracy. Accuracy is calculated by the following rule: Accuracy = (TP+TN)/(TP+FP+FN+TN).

Precision is another performance measurement. Precision refers to the ratio between the number of true positive and the total number of true predictions. It follows the rule; Precision=TP/TP+FP.

Another important performance measure is Recall. Another way to say that how many of the true positives are recalled. Recall literally measures the ratio between the number of correct results and the number of results that should have been returned. Recall follows the rule; Recall=TP/TP+FN.

F1-score is another performance measure, which is the weighted average of precision and recall. When you have an uneven class distribution, the F1 Score is more useful than accuracy. Both false positive and false negative are considered by F1 Score. It follows the rule; F1-score=2*(Recall * Precision) / (Recall + Precision).

Chapter 4

Results and Analysis

4.1 Result

The implementation part of this project has been done in python language. And the project is done in Google Collab and Jupyter Notebook. All the machine learning algorithms are done in Jupyter Notebook. And the Deep Learning Hyper Parameter is done in Google Collab.

Now I am going to show about the performance of Hyper Parameter. The Accuracy score of Hyper Parameter is 83.32%. Which is good enough. Using Hyper Parameter the Precision score is 0.817. The recall score is 0.849. And the F1 Score is 0.833. In Figure 4.1 we can see the confusion matrix of the model using Hyper Parameter.

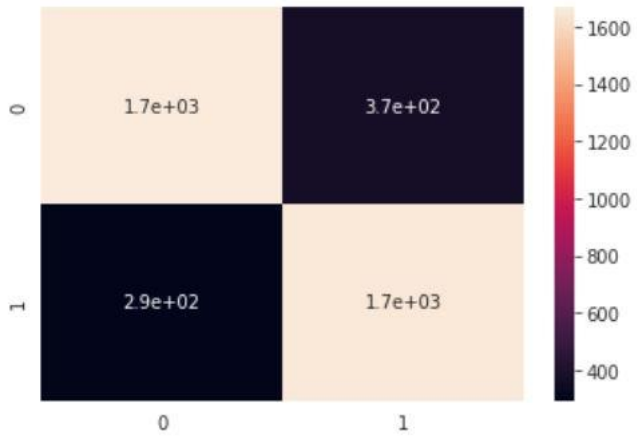


Figure 4.1: Confusion Matrix of Hyper Parameter

When I applied Logistic Regression then I did not get a good accuracy. The accuracy was 0.752. The precision score was 0.735. The Recall score of the model was 0.769 and the F1 score was 0.752. The Confusion Matrix of this model using Logistic Regression is shown in Figure 4.2.

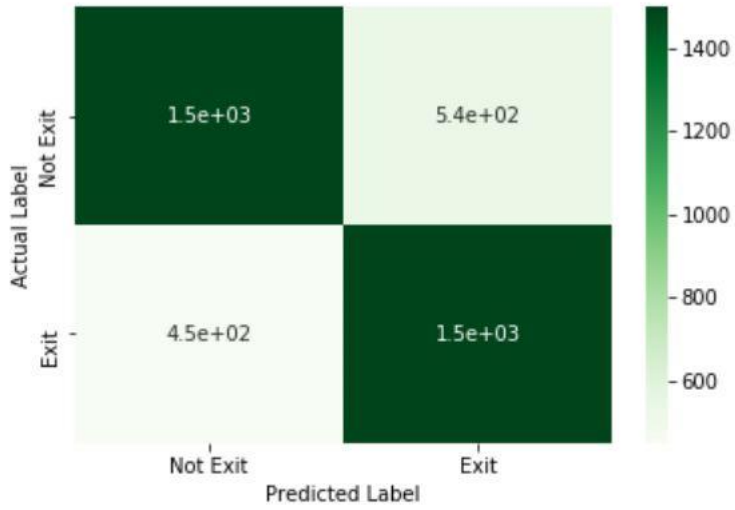


Figure 4.2 : Confusion Matrix of Logistic Regression

By Using the Decision Tree algorithm, I got a better accuracy than Logistic Regression. But it is not a good accuracy overall. The accuracy of the model was 0.790. The Precision score of the model was 0.775. The Recall score was 0.805 and the F1 score is 0.790. Figure 4.3 shows the confusion matrix of the model.

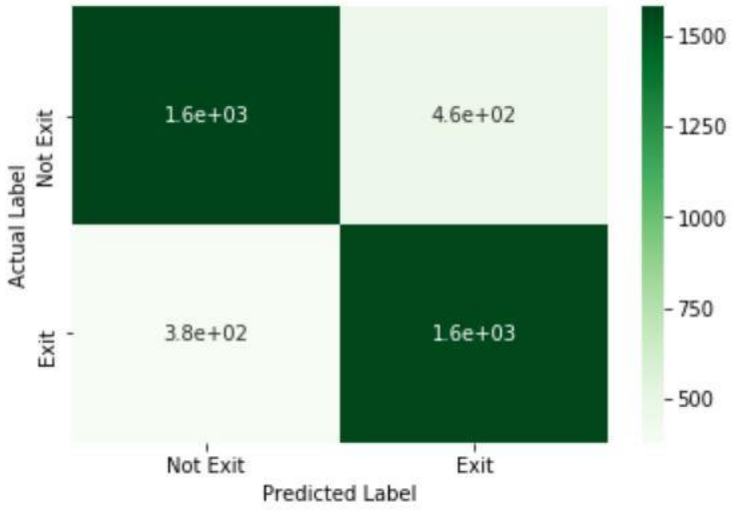


Figure 4.3: Confusion Matrix of Decision Tree

k-Nearest-Neighbor (KNN) gives almost same accuracy as Decision Tree. The accuracy of the model using KNN algorithm is 0.795. The Precision score is 0.781. The Recall score is 0.806. And the F1 score is 0.793. The Confusion Matrix of this model is shown in Figure 4.4.

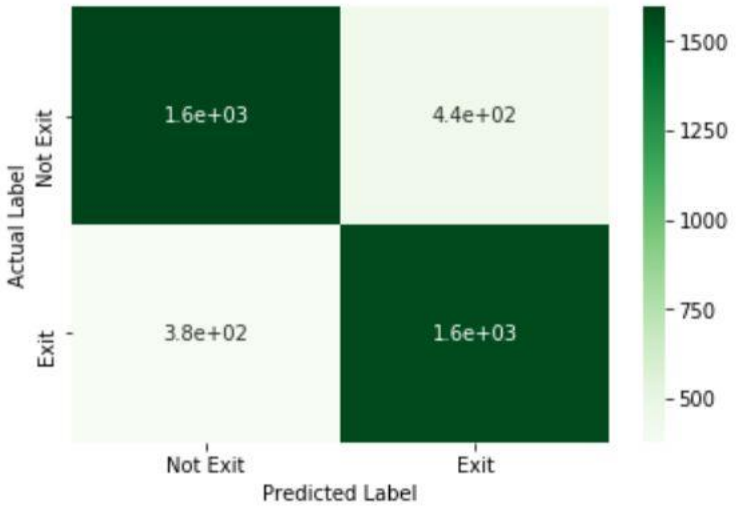


Figure 4.4 : Confusion Matrix of KNN

Using Random Forest algorithm I have got the best result among all the algorithm. The accuracy of the model after applying Random Forest algorithm is 0.849. The precision score is 0.837. The Recall score is 0.859 and the F1 score is 0.848. The Confusion Matrix of this model is shown in Figure 4.5.

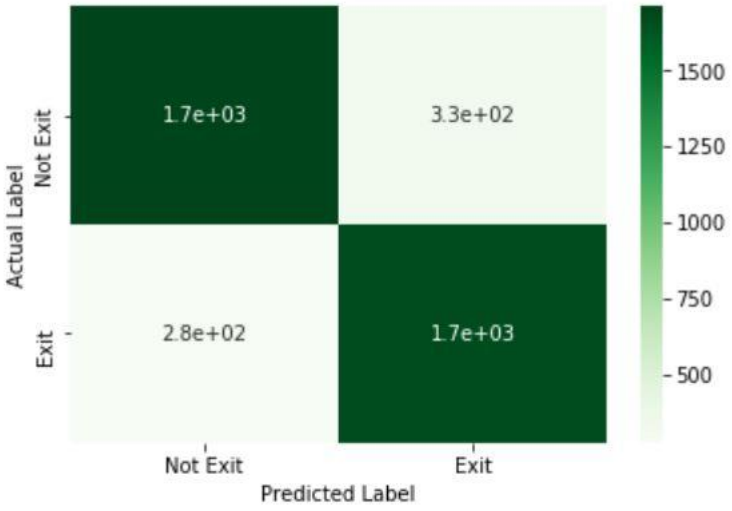


Figure 4.5 : Confusion Matrix of Random Forest

Table 4.1 shows the comparison between all those algorithms performance. We can differentiate between algorithms and can be able to identify which algorithm is best for this project through this table.

Table 4.1: Comparison between Classifiers

Classifier	Precision	Recall	F1 Score	Accuracy
Logistic Regression	0.735	0.769	0.752	75.16%
Decision Tree	0.775	0.805	0.790	79.03%
KNN	0.781	0.806	0.793	79.45%
Random Forest	0.837	0.859	0.848	85%

4.2 Output Analysis

We already found the best classifier for this project. Random forest gives best accuracy for this project, which is almost 85%. As Random Forest is the best classifier, I add this classifier to the User Interface to visualize the output. For designing the User Interface, I have used HTML and CSS. Figure 4.6 shows the home page of the project. Here we can see a form for predict customer churn. When we fill up the form and submit it, we get another page as output. The output page shows the result as ‘churn’ or ‘not churn’.

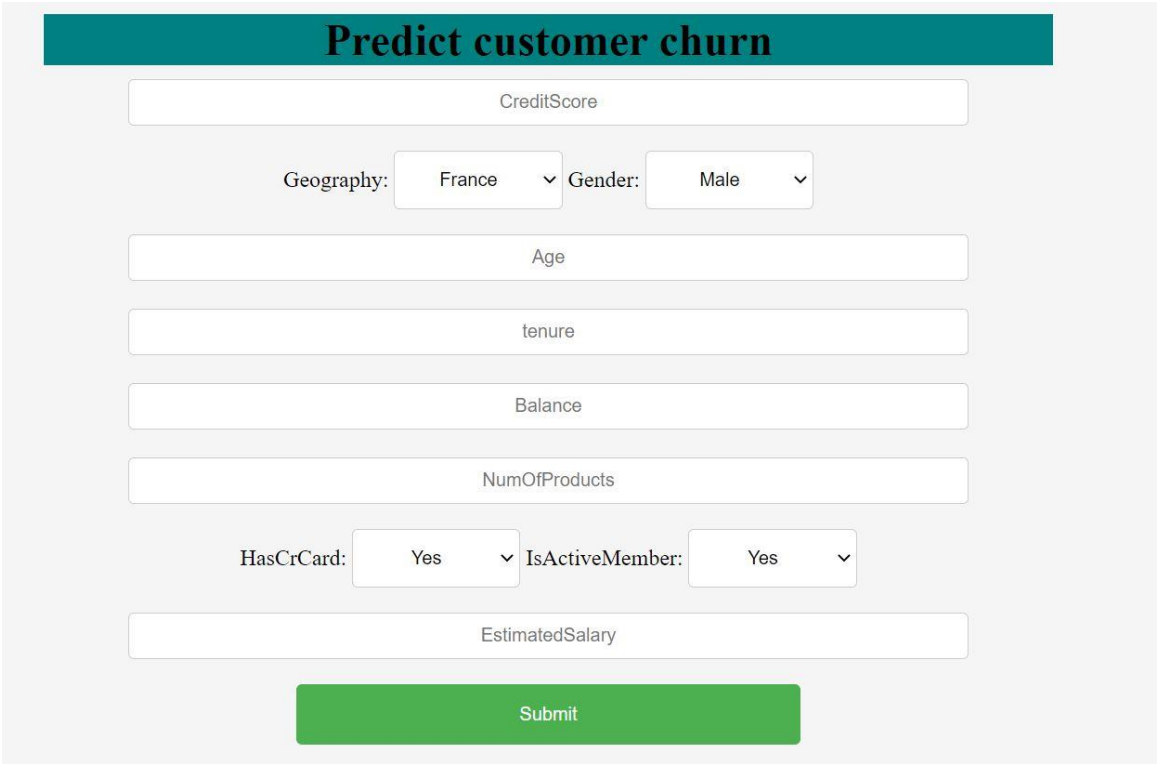
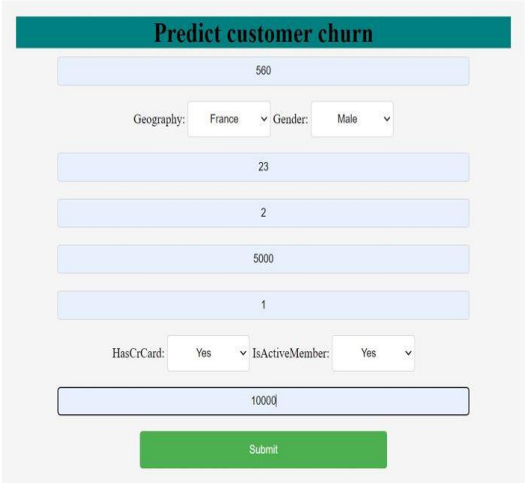
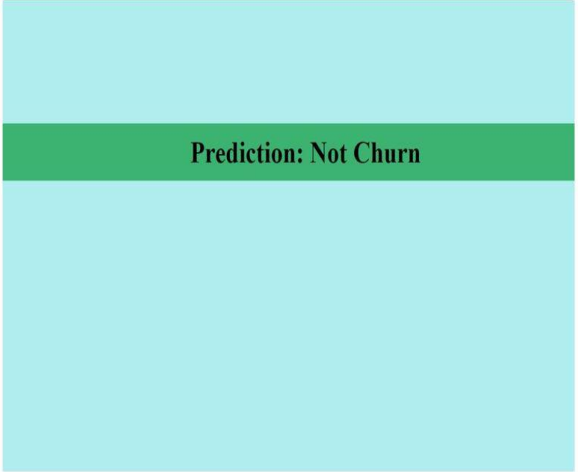
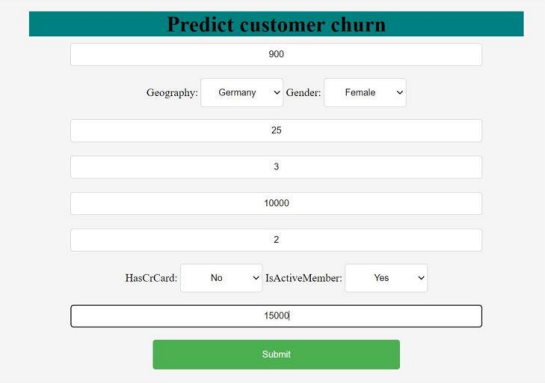


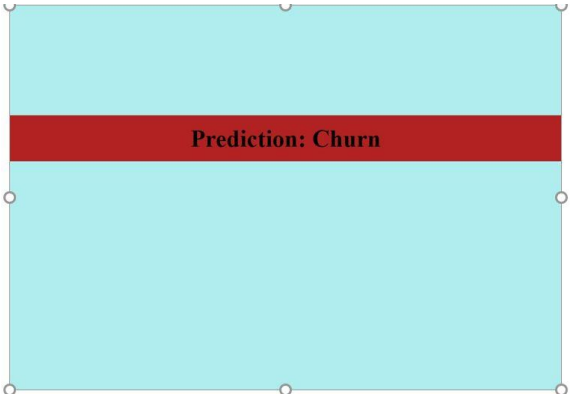
Figure 4.6: Home page of UI

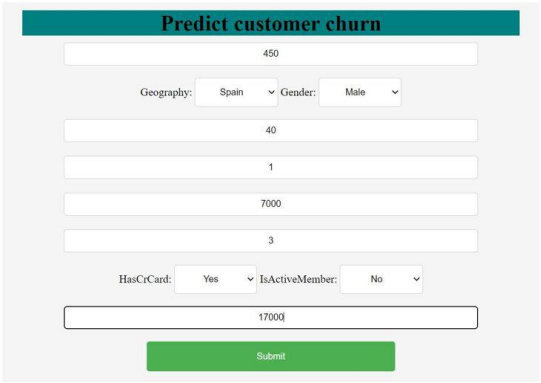
I have done data analysis in Chapter 3.1 briefly. As per the data analysis, churning rate depends on the features. For different type of feature rate, the churn rate is different. For example, the people who are using 4 products has more chance to churn. Another example is customer who are active their churning rate is less than not active people. Bellow for five different inputs what are the five different outputs is shown visually.

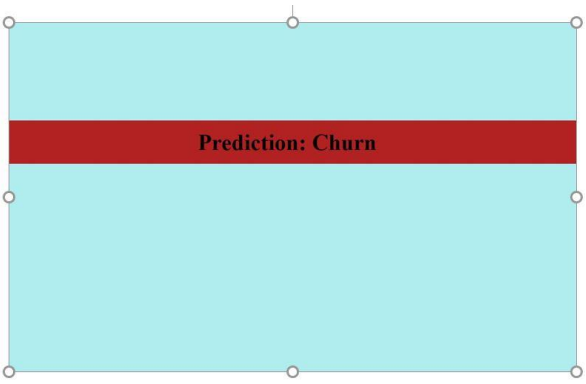


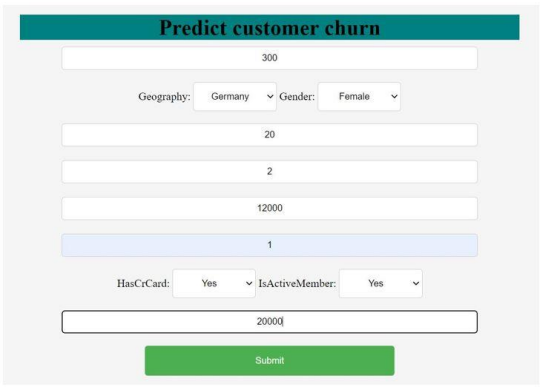


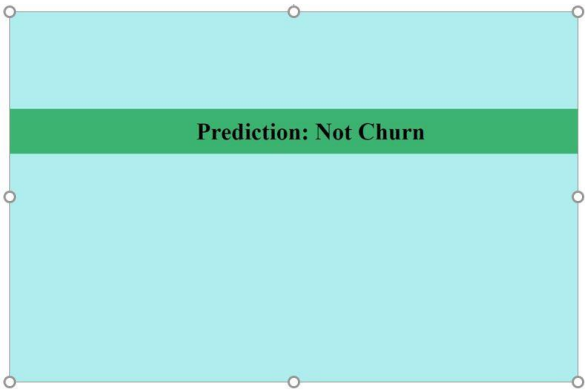


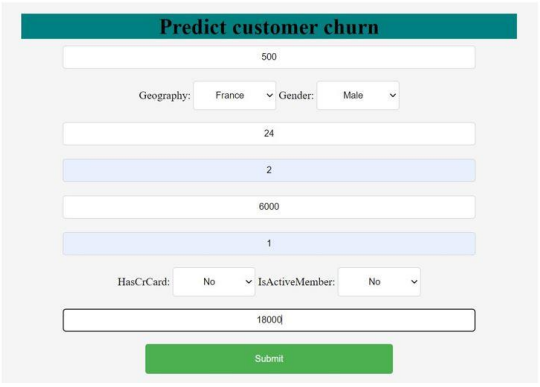


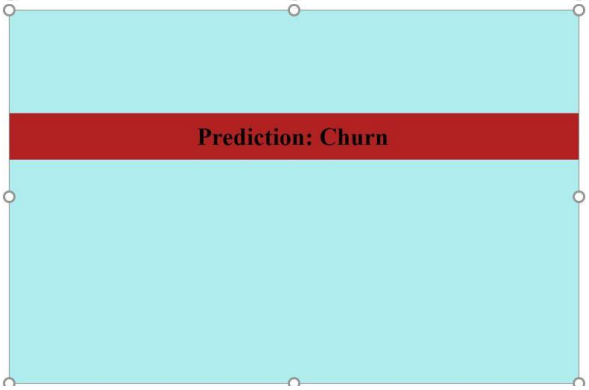





Predict customer churn
300
Geography: Germany Gender: Female
20
2
12000
1
HasCrCard: Yes IsActiveMember: Yes
20000
Submit


Prediction: Not Churn


Predict customer churn
500
Geography: France Gender: Male
24
2
6000
1
HasCrCard: No IsActiveMember: No
18000
Submit


Prediction: Churn

Chapter 5
Conclusion and Limitations

5.1 Conclusion

This study represents customer churn in Banking Industry using Machine Learning and Deep Learning. At first some analysis on the data set has been done. After data pre-processing is done as the data set is imbalance, I need to balance the dataset. After that model building is also done. For model building I used such Machine Learning algorithm such as Linear Regression, Decision Tree, KNN and Random Forest. And as Deep Learning method I used Hyper Tuning. Among all these algorithms Random Forest gives best accuracy. The performance is shown in terms of accuracy, recall value, precision and F1 score.

5.2 Limitations

The main limitation of this project is in this project I did not use any dataset of Bangladeshi bank. The using data set belongs to Germany, France, and Spain. I did not find any suitable dataset of Bangladeshi bank for this project.

5.3 Future Works

In future I want to improve this model accuracy by using more advanced machine learning algorithms like support vector machine (SVM), also with some advanced assembling techniques such as boosting, bagging.

Also, I will try to find Bangladeshi bank dataset to implement this project.

References

[1] V. Kavitha, G. Hemanth Kumar, S. V Mohan Kumar, and M. Harish, “Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms,” *Int. J. Eng. Res.*, vol. V9, no. 05, pp. 181–184, 2020, doi: 10.17577/ijertv9is050022.

[2] P. Spanoudes and T. Nguyen, “Deep learning in customer churn prediction: Unsupervised feature learning on abstract company independent feature vectors,” *arXiv*, pp. 1–22, 2017.

[3] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, “A comparison of machine learning techniques for customer churn prediction,” *Simul. Model. Pract. Theory*, vol. 55, no. June, pp. 1–9, 2015, doi: 10.1016/j.simpat.2015.03.003.

[4] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, “Customer churn prediction system: a machine learning approach,” *Computing*, no. February, 2021, doi: 10.1007/s00607-021-00908-y.

[5] P. K. Dalvi, S. K. Khandge, A. Deomore, A. Bankar, and V. A. Kanade, “Analysis of customer churn prediction in telecom industry using decision trees and logistic regression,” *2016 Symp. Colossal Data Anal. Networking, CDAN 2016*, 2016, doi: 10.1109/CDAN.2016.7570883.

[6] I. Kaur and J. Kaur, “Customer churn analysis and prediction in banking industry using machine learning,” *PDGC 2020 - 2020 6th Int. Conf. Parallel, Distrib. Grid Comput.*, pp. 434–437, 2020, doi: 10.1109/PDGC50313.2020.9315761.