Department of Electrical and Computer Engineering
North South University

Senior Design Project

CUSTOMER CHURN PREDICTION IN BANKING INDUSTRY

FARIHA MOAZZEMA RAISA **ID# 1811295042**

SUMAYA SARWAR SARA **ID# 2013961642**

Faculty Advisor:

Mr. Intisar Tahmid Naheen

Lecturer

ECE Department

Spring, 2023

LETTER OF TRANSMITTAL

20 June, 2023

To

Dr. Rajesh Palit
Chairman,
Department of Electrical and Computer Engineering
North South University, Dhaka

Subject: **Submission of Capstone Project Report on "Customer Churn Prediction in Banking Industry"**

Dear Sir,

With due respect, we would like to submit our **Capstone Project Report** on **"Write you Customer Churn Prediction in Banking Industry"** as a part of our BSc program. The report deals with prediction of occurrences of customer churn in the banking sector. This project was very much valuable to us as it helped us gain experience from practical fields and apply in real life. We tried to the maximum competence to meet all the dimensions required from this report.

We will be highly obliged if you kindly receive this report and provide your valuable judgment. It would be our immense pleasure if you find this report useful and informative to have an apparent perspective on the issue.

Sincerely Yours,

.....

Fariha Moazzema Raisa
ECE Department
North South University, Bangladesh

.....

Sumaya Sarwar Sara
ECE Department
North South University, Bangladesh

APPROVAL

Fariha Moazzema Raisa (ID # 1811295042), Sumaya Sarwar Sara (ID # 2013961642) from Electrical and Computer Engineering Department of North South University, have worked on the Senior Design Project titled "Customer Churn Prediction in Banking Industry" under the supervision of Mr. Intisar Tahmid Naheen partial fulfillment of the requirement for the degree of Bachelors of Science in Engineering and has been accepted as satisfactory.

Supervisor's Signature

……

Mr. Intisar Tahmid Naheen

Lecturer

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

Chairman's Signature

……

Dr. Rajesh Palit

Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

DECLARATION

This is to declare that this project is our original work. No part of this work has been submitted elsewhere partially or fully for the award of any other degree or diploma. All project related information will remain confidential and shall not be disclosed without the formal consent of the project supervisor. Relevant previous works presented in this report have been properly acknowledged and cited. The plagiarism policy, as stated by the supervisor, has been maintained.

Students' names & Signatures

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

1. Fariha Moazzema Raisa

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

2. Sumaya Sarwar Sara

ACKNOWLEDGEMENTS

The authors would like to express their heartfelt gratitude towards their project and research supervisor, Mr. Intisar Tahmid Naheen, Lecturer, Department of Electrical and Computer Engineering, North South University, Bangladesh, for his invaluable support, precise guidance and advice pertaining to the experiments, research and theoretical studies carried out during the course of the current project and also in the preparation of the current report.

Furthermore, the authors would like to thank the Department of Electrical and Computer Engineering, North South University, Bangladesh for facilitating the research. We would also like to thank my friends X, Y, and Z for helping us in this project. The authors would also like to thank their loved ones for their countless sacrifices and continual support.

ABSTRACT

Customer Churn Prediction in Banking Industry

This report presents a comparative study on predicting customer churn in the banking industry. Customer churn, or the rate at which customers discontinue their services, poses a significant challenge for banks. By analyzing historical customer data and employing various machine learning algorithms, this study aims to identify the most effective predictive models for customer churn. The report evaluates and compares the performance of different algorithms in terms of accuracy, precision, recall and F1-score, providing insights into the potential factors influencing customer churn and enabling banks to implement proactive measures for customer retention.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Background and Motivation

Customer churn, also known as customer attrition, refers to the tendency of customers to discontinue their business relationship with a company within a specific time frame. Effective customer relationship management (CRM) strategies play a vital role in building and nurturing long-lasting customer relationships. CRM has gained widespread recognition and adoption across various sectors such as telecommunications, banking and insurance, and the retail market. A primary objective of CRM is customer retention, considering that the cost of acquiring new customers far exceeds the cost of retaining existing ones. Consequently, the development and application of customer retention models, specifically churn models, have become essential components of Business Intelligence applications.

1.2 Purpose and Goal of the Project

Churn prediction, a crucial aspect of customer retention, involves forecasting customers' likelihood to churn based on their historical data and activities. In recent years, churn prediction has emerged as a highly debated research area. This study focuses on predicting the probability of bank customers leaving, utilizing a dataset comprising their transactional information.

1.3 Organization of the Report

Chapter 2 presents the literature reviews related to this project.

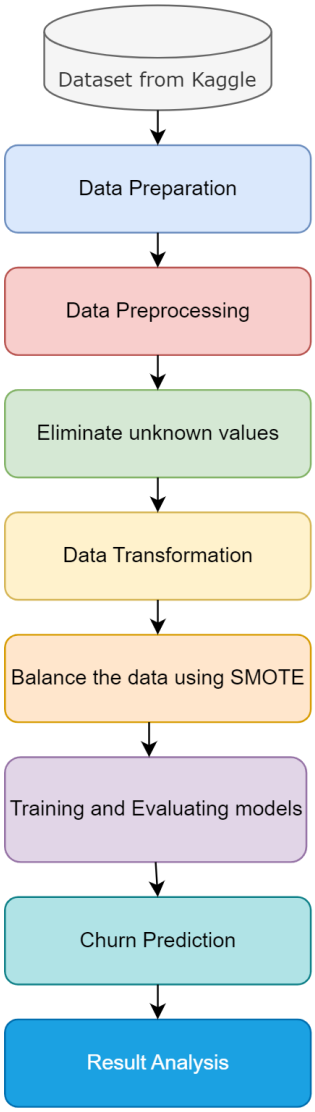
Chapter 2 Research Literature Review

2.1 Existing Research and Limitations

The paper aims to provide an overview of the study titled “Customer Churn Prediction in the Banking Industry.” The investigation focuses on predicting customer churn, a crucial concern for banks, by employing comparative analysis techniques. The research aims to contribute to the existing knowledge by evaluating and comparing different predictive models utilized in the banking industry. By reviewing previous studies on customer churn prediction and its significance in the banking sector, this study seeks to highlight the importance of accurate churn prediction for banks to develop effective customer retention.

Chapter 3 Methodology

3.1 System Design



The customer churn prediction analysis system is designed to identify and predict potential customer churn within a business. It involves collecting relevant customer data such as purchase history, demographic information, and customer interactions. This data is then processed and

analyzed using machine learning algorithms to identify patterns and indicators of churn. The system generates predictive models that can accurately forecast customer churn, which can be found in result analysis, enabling businesses to take proactive measures such as targeted marketing campaigns or personalized retention strategies to mitigate churn and maximize customer retention.

3.2 Hardware and/or Software Components

The respective research is an AI-related project. To implement the project, dataset of this particular topic is collected from Kaggle. As for EDA and preprocessing techniques, the missing values or inconsistencies are either removed or inserted, the null values are searched for here and the errors are corrected, ensuring uniformity in data format. Jupyter notebook is used as the coding platform her, which is a browser-based IDE and python is used in this platform to code for the data analysis. Numerical and categorical variables are identified in the data to determine appropriate analysis techniques. Analyzing individual variables, a pair of variables or more than two variables through, examining mean, median, standard deviation, minimum, maximum, determining frequency distribution, scatter plots, correlation analysis, cross-tabulation, clustering, dimensionality reduction, or advanced visualizations to gain insights into complex interactions. Then identifying patterns, outliers, and data distributions by creating a visual representation of the data using graphs, charts and histograms is done. For numerical variables, histogram and boxplots are used for visual representation and for categorical values, pie-chart, boxplots and bar-chart are shown. Data is normalized in standardized form and the imbalanced ratio was balanced through oversampling using SMOTE for the data to be trained and tested.

Table I. List of Tools Used for Data Analysis

Tool	**Functions**	**Other similar Tools (if any)**	**Why selected this tool**

Kaggle	**Data collection**		**Commonly known for project-based datasets**
Jupyter Notebook	**For exploratory data analysis, training and testing of data**		**Browser-based IDE and user friendly for ML-based projects**

3.3 Hardware and/or Software Implementation

As for statistical modeling or evaluation of models, machine learning algorithms such as, decision tree, random forest, KNN, gradient boosting, SVC, and logistic regression are applied. These statistical techniques are used to test hypotheses, build predictive models, or uncover underlying patterns of the data, such as evaluate and compare the performance of different algorithms used in terms of accuracy, precision, recall values, and F1-score, providing more insights into the potential factors influencing the customer churn.

```
data.isnull().sum()
```

```
customer_id          0
credit_score         0
country              0
gender               0
age                  0
tenure               0
balance              0
products_number      0
credit_card          0
active_member        0
estimated_salary     0
churn                0
dtype: int64
```

Figure 1. Null value determined

```
In [8]: data.columns
```

```
Out[8]: Index(['customer_id', 'credit_score', 'country', 'gender', 'age', 'tenure',
       'balance', 'products_number', 'credit_card', 'active_member',
       'estimated_salary', 'churn'],
      dtype='object')
```

```
In [9]: data=data.drop(['customer_id'],axis=1)
```

```
In [10]: data.head()
```

Out[10]:

	credit_score	country	gender	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	churn
0	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Figure 2. Data cleaning - unnecessary feature removed: customer ID dropped

```
In [25]: data['country'].unique()
Out[25]: array(['France', 'Spain', 'Germany'], dtype=object)

In [26]: from sklearn.preprocessing import LabelEncoder
         enc=LabelEncoder()

In [27]: country=enc.fit_transform(data['country'])

In [28]: gender=enc.fit_transform(data['gender'])

In [29]: country
Out[29]: array([0, 2, 0, ..., 0, 1, 0])

In [30]: gender
Out[30]: array([0, 0, 0, ..., 0, 1, 0])

In [31]: data['gender']=gender
         data['country']=country
```

Figure 3. Conversion of categorical variables to numerical variables

Out[16]: <Axes: >

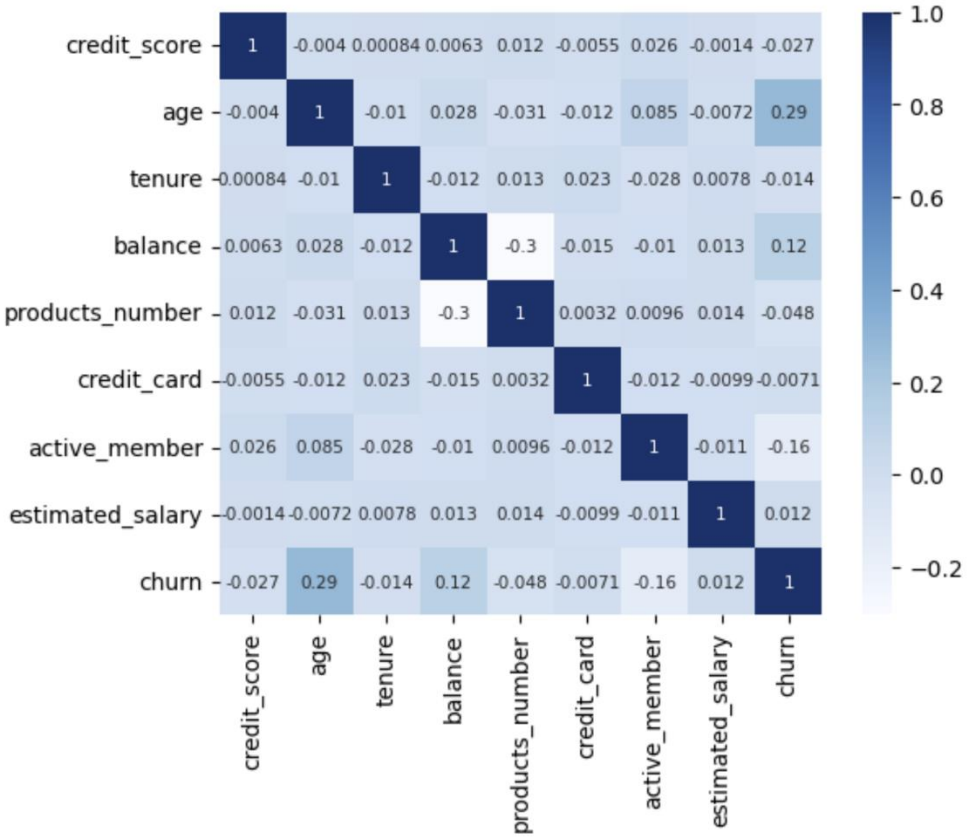


Figure 4. Correlation matrix

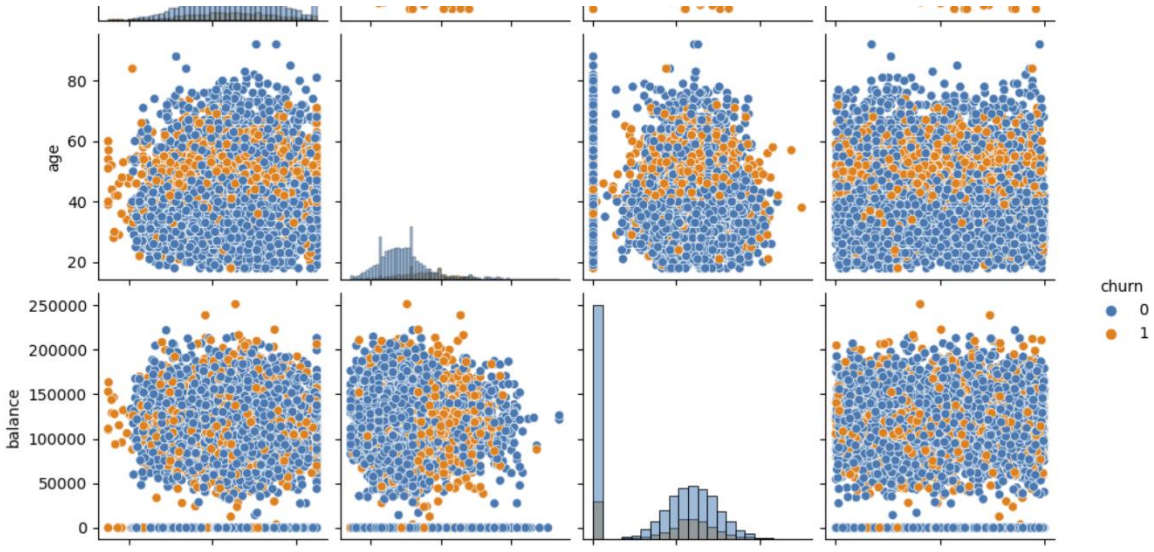


Figure 5. Pair-plot of churn

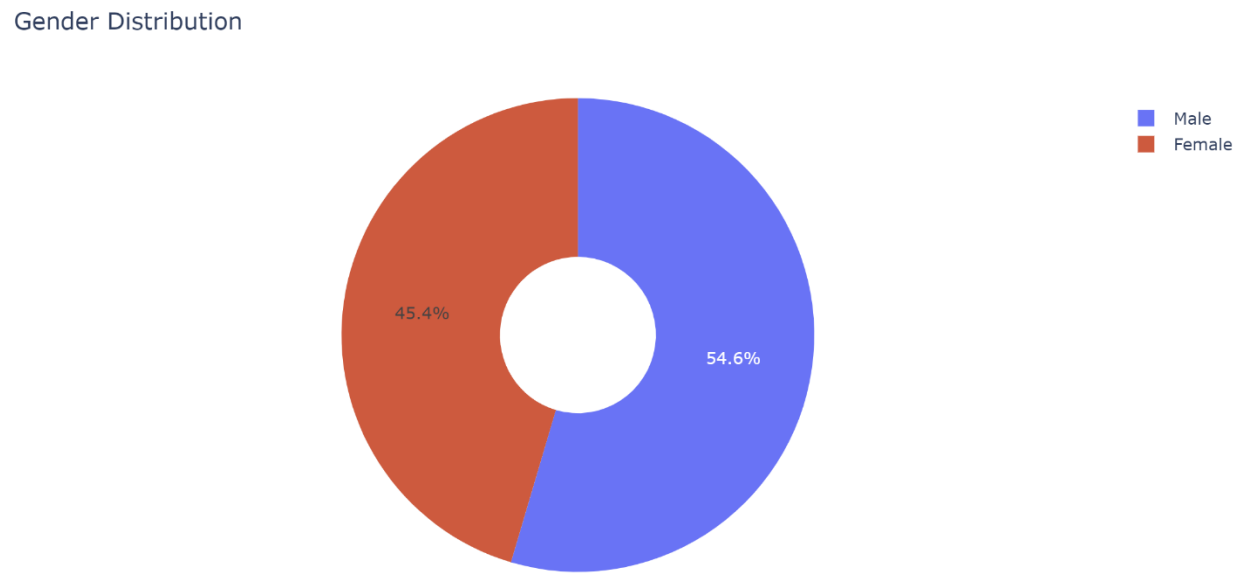


Figure 6. Pie-chart of gender distribution

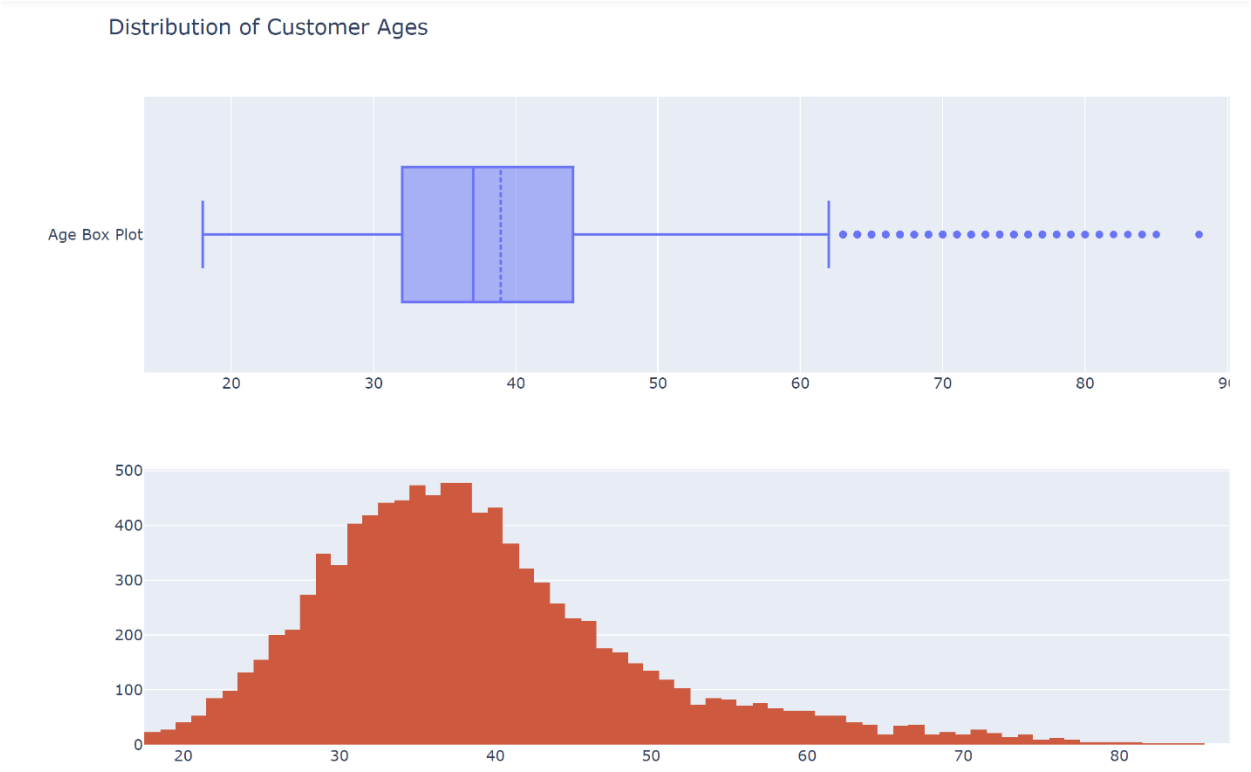


Figure 7. Box plot and Histogram of customer's ages

Chapter 4 Investigation/Experiment, Result, Analysis and Discussion

In the beginning of the EDA, null values are determined in the data. The unnecessary values are discarded. To determine appropriate analyzing techniques, numerical and categorical variables are identified. For analyzing individual variables, mean, median, standard deviation, minimum and maximum are examined for the numerical variables, and for categorical variables, frequency distribution of the occurrences of each category is determined by counting. Correlation analysis is carried out to explore the relationship between pairs of variables. Histograms and boxplots are shown in case of numerical variables for visual representation and pie-charts, bar-chart, or box plots in case of categorical variables. The ratio of the data is balanced through oversampling using SMOTE bring it to a number of about 7000 rows to train and test the data more appropriately. Machine learning algorithms are applied for model evaluation, such as, decision tree, logistic regression, KNN, random forest and so on. Random forest turns out to be the most suitable algorithm with great performance for the project after much careful evaluation of the data in terms of determining accuracy, precision, recall, and F1-score.

```
In [127]: sns.barplot(x=final_data_ac['Models'],y=final_data_ac['Accuracy'])
Out[127]: <Axes: xlabel='Models', ylabel='Accuracy'>
```

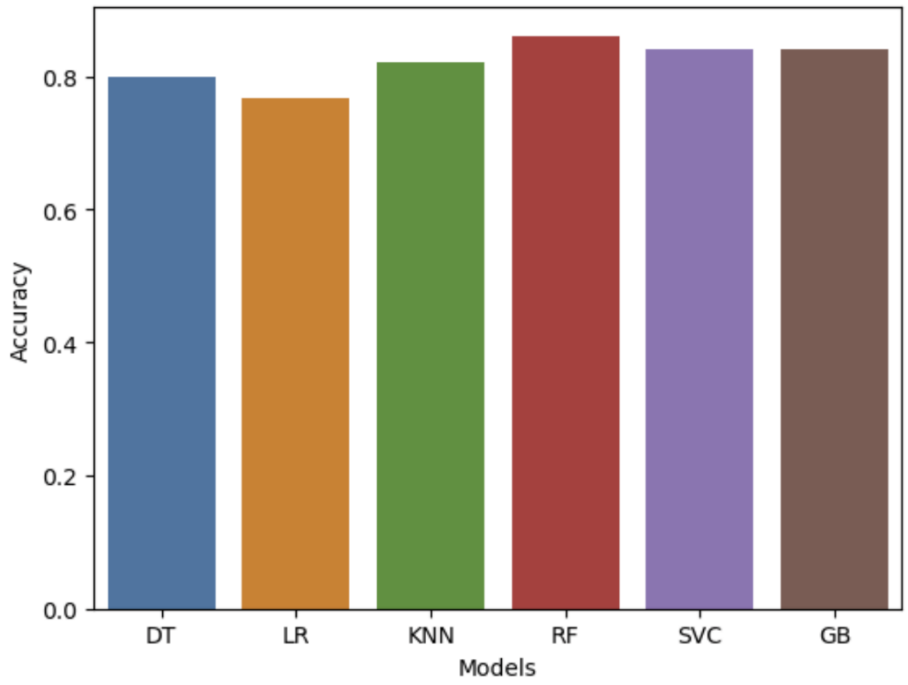


Figure 8. Bar chart of Accuracy

```
In [128]: sns.barplot(x=final_data_pc['Models'],y=final_data_pc['Precision'])
Out[128]: <Axes: xlabel='Models', ylabel='Precision'>
```

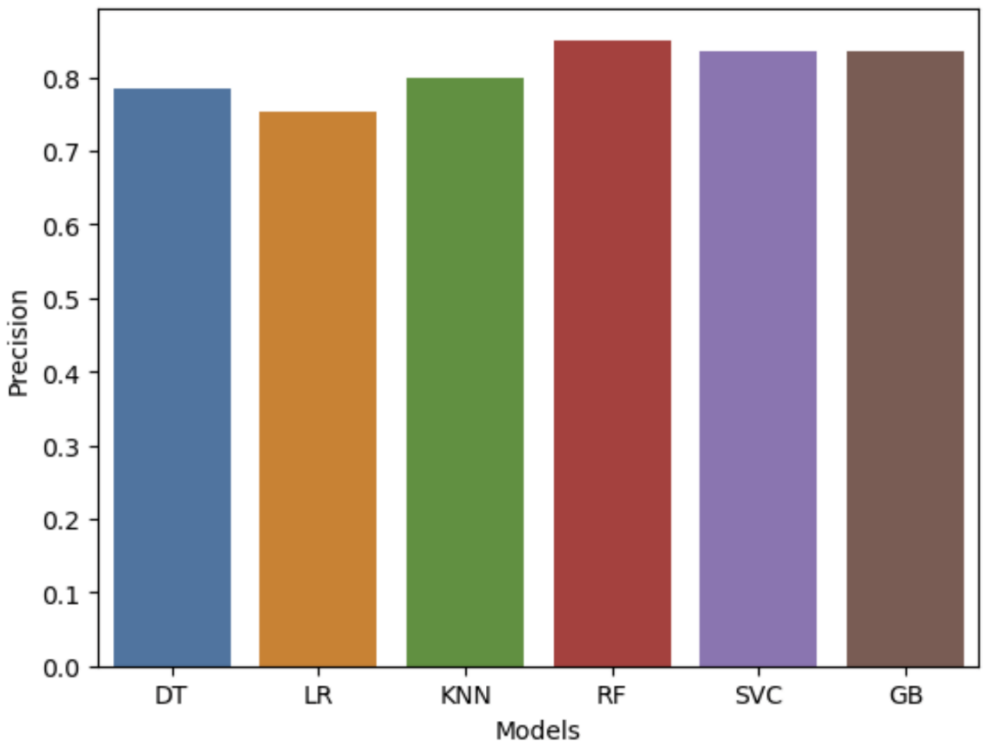


Figure 9. Bar chart of precision

```
In [129]: sns.barplot(x=final_data_rc['Models'],y=final_data_rc['Recall'])
```

```
Out[129]: <Axes: xlabel='Models', ylabel='Recall'>
```

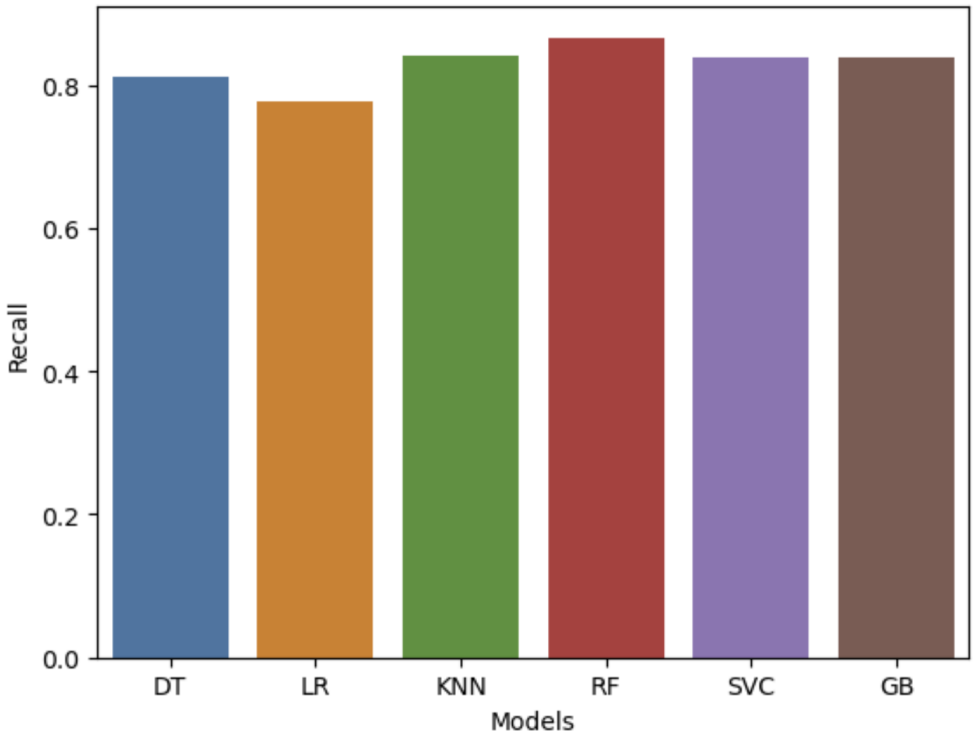


Figure 10. Bar chart of Recall

```
In [133]: sns.barplot(x=RF['Results'],y=RF['RandomForest'])

Out[133]: <Axes: xlabel='Results', ylabel='RandomForest'>
```

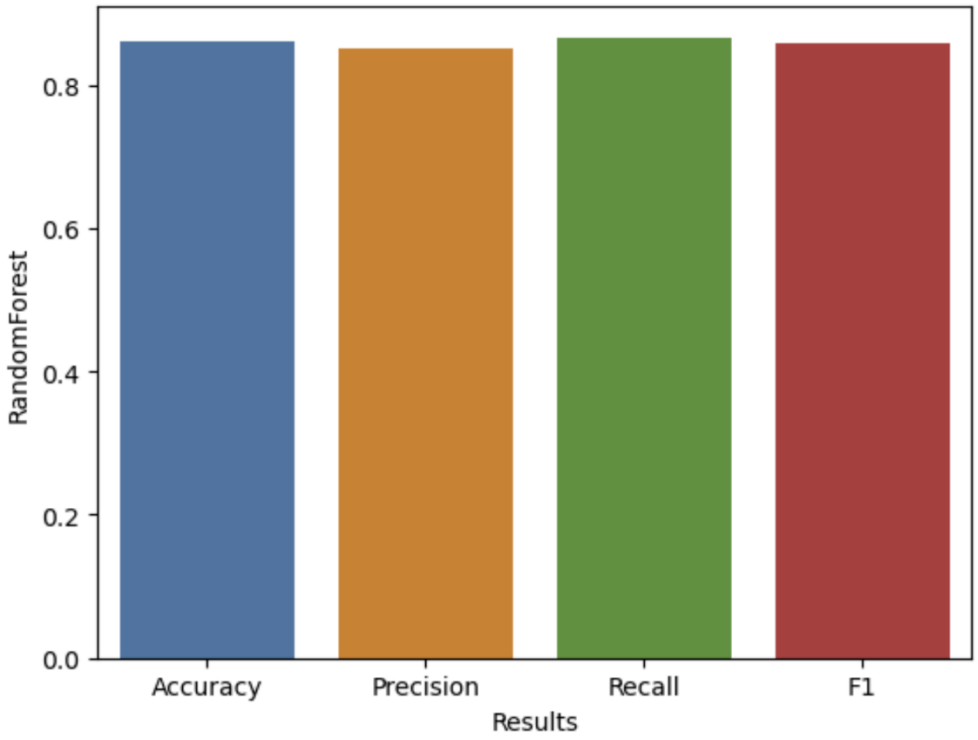


Figure 11. Bar chart of Random Forest showing its accuracy, precision, recall and F1-score

```
In [133]: sns.barplot(x=final_data_f1['Models'],y=final_data_f1['F1'])
Out[133]: <Axes: xlabel='Models', ylabel='F1'>
```

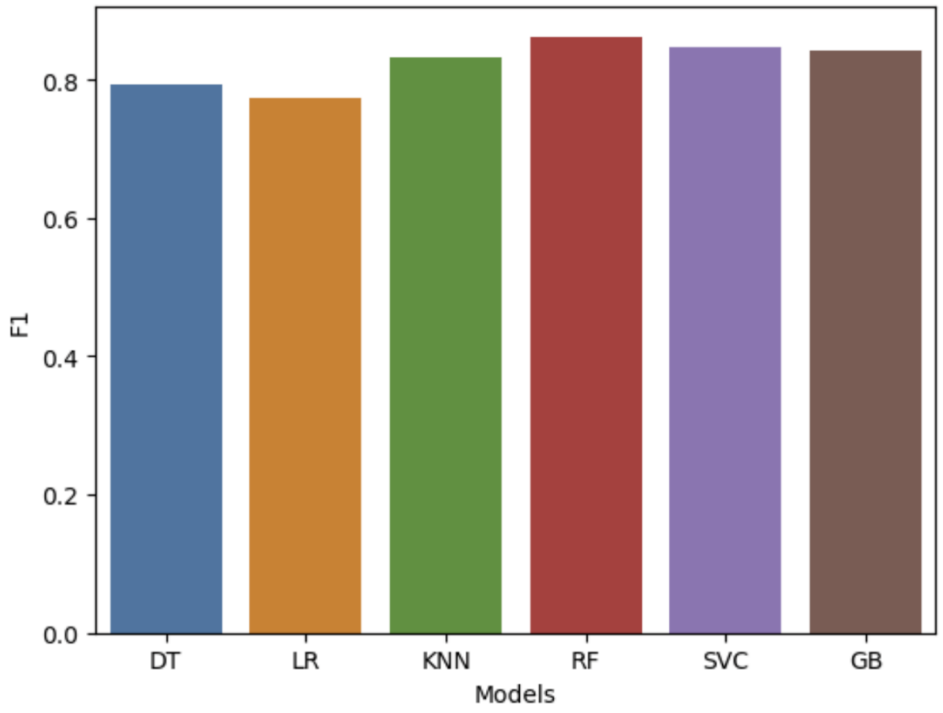


Figure 12. Bar plot of F1-score

The paper provides a thorough analysis of different prediction models that have been employed in the banking industry to track client turnover. In order to help banks increase customer retention and profitability, the study examines various approaches, assesses and effectiveness of three machine learning models, namely K-Nearest Neighbors (KNN), Random Forest, and Logistic Regression, and provides helpful insights for establishing effective ways to decrease customer turnover. The research and comparison of these models on the dataset show that Random Forest balances accuracy and efficiency while KNN achieves the maximum accuracy but with a higher level of computational complexity. Despite being simpler, logistic regression offers acceptable performance.

Chapter 5 Impacts of the Project

5.1 Impact of this project on societal, health, safety, legal and cultural issues

It is crucial to remember that, even while the application of churn prediction might have favorable effects, it should be done with due regard for data privacy, the moral use of customer data, and openness in the decision-making procedures.

Societal Impact:

- **Boosted client happiness:** By precisely forecasting client attrition, banks can take proactive steps to retain customers, boosting customer happiness in the banking sector as a whole.
- **Increased Financial Inclusion:** By identifying probable causes of client attrition, banks can provide better banking products and services that cater to a variety of consumer segments, including underserved communities.
- **Economic Growth:** Banks can contribute to the general growth and stability of the banking industry, which in turn has a favorable effect on the economy, by stabilizing their client base and lowering customer churn.

Health Impact:

- **Lessened Financial Stress:** Banks can identify customers who may leave owing to financial troubles by using churn prediction. Banks may be able to lessen the financial stress experienced by such consumers by proactively providing individualized financial aid or counseling.

Safety Impact:

- **Fraud Detection:** By examining customer behavior and transactional data, churn prediction models can assist banks in locating patterns of fraudulent behavior. This can

improve the safety and security of banking services by helping to prevent and identify fraudulent activity.

Legal Implications:

- **Regulation Compliance:** Banks must abide by a number of legal and regulatory regulations. The ability to predict customer churn can help banks identify clients who may be impacted by regulatory changes, allowing them to provide timely information and counsel to maintain compliance.

Cultural Impact:

- **Personalized Customer Experience:** Churn prediction helps banks better comprehend the preferences and behavior of their customers. Banks may offer a more culturally aware and individualized banking experience that caters to a variety of cultural backgrounds and needs by customizing products and services for each individual consumer.

5.2 Impact of this project on environment and sustainability

The impact of a "Customer Churn Prediction in the Banking Industry" project on the environment and sustainability may not be direct, as the primary focus of such a project is on customer retention and business operations rather than environmental aspects. However, there are indirect ways in which the project can contribute to sustainability.

- **Reduced paper usage:** Banking industry studies that predict customer turnover frequently analyze vast amounts of customer data. Without largely depending on physical papers, banks can acquire insights into client behavior, preferences, and satisfaction levels by applying modern data analytics techniques and machine learning algorithms. This may result in less paper being used, enhancing environmental sustainability.

- **Efficient resource allocation:** Customer churn prediction algorithms assist banks in identifying clients who are most likely to depart and taking preventative actions to keep them. Banks can avoid pointless marketing campaigns or incentives that might have a wider environmental impact by effectively allocating money to targeted retention initiatives. This promotes waste minimization and resource optimization.

- **Enhanced customer experience:** Improving overall customer experience is a major goal of churn prediction initiatives. Banks can improve customer happiness and loyalty by identifying prospective churners and understanding their demands. clients who are happy with a bank are more likely to stick around, which reduces the need for acquiring new clients, which might have environmental consequences owing to marketing and onboarding efforts.

- **Digital transformation:** Churn prediction initiatives frequently need financial institutions making investments in cutting-edge analytics software, data infrastructure, and digital

capabilities. The firm may undergo a wider digital transformation as a result, which would streamline operations, minimize the need for physical resources, and boost operational effectiveness. Initiatives in automation and digitization reduce resource consumption and increase energy efficiency, which helps to promote sustainability.

- **Behavioral insights for sustainability initiatives:** Churn prediction tools produce useful information on the preferences and behavior of customers. Banks can use this data to customize their goods and services to fit with environmentally friendly procedures. For instance, banks can create and market eco-friendly banking options or sustainable investing opportunities if customers express a preference for them, so promoting ecologically responsible purchasing decisions.

While the direct impact of a churn prediction project on the environment may be limited, its contribution to overall sustainability lies in optimizing resource allocation, promoting digital transformation, and leveraging customer insights to support environmentally conscious practices. It is crucial for banks and organizations to consider sustainability as a broader goal and integrate it into their overall business strategies.

Chapter 6 Project Planning

The planning of project, selection of the research topic, proposal presentation all took place from the end of January to around 20th February, 2023, as shown in the chart. Findings of research paper, their literature reviews and paper reviews, all kinds of research study required about less than 20 days from the month of February to March. Even though it shows that the EDA started a little bit late after the data collection, the whole process of data analysis and preprocessing altogether started around the beginning of April from the moment of final selection of a dataset, including drafts, reviewed drafts, final drafts, many approvals and disapprovals, more guidance, more meetings and more suggestions. As EDA is an iterative process, the model was being trained as well along the process, along with the evaluation and testing, to gain better insight of the complex interactions and finally around after 31st May, all of the training, testing and data analyzing process came to an end for the project to be presented. As shown in the graph, the report writing was in process along with the continuation of data analysis and came to be ready to be submitted on 20th June, 2023.

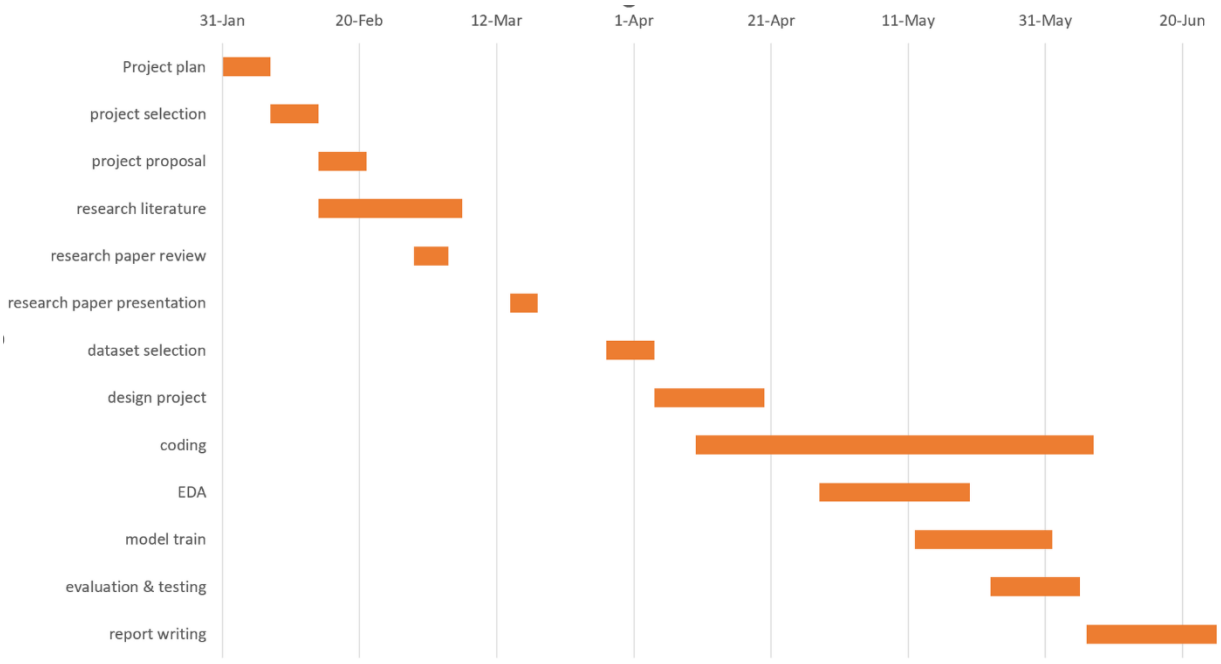


Figure 13. A Gantt chart of the Churn Prediction Analysis

Chapter 7 Conclusions

7.1 Summary

Customer churn prediction analysis in the banking industry aims to forecast the likelihood of customers ending their relationship with a bank. By utilizing various data mining and machine learning techniques, banks can identify patterns and indicators that suggest a customer may churn. These indicators may include changes in transactional behavior, declining account activity, or customer demographics.

The analysis enables banks to proactively address customer dissatisfaction, tailor retention strategies, and minimize the loss of valuable customers. By predicting churn, banks can implement targeted marketing campaigns, personalized offers, and improved customer service to mitigate the risk of customer attrition and maintain long-term customer relationships. Ultimately, churn prediction analysis empowers banks to make data-driven decisions, enhance customer retention efforts, and optimize overall business performance.

7.2 Limitations

Limitation of dataset can lead to several limitations in case of ML based projects. For lack of enough resources, results cannot be found accurate during training and testing at times, which is why the EDA needs to be iteratively processed more times for the evaluation of models to get more exact accuracy.

7.3 Future Improvement

To carry out such a huge project such as, surveys related to banking industries, more resources are required. Hence, more datasets are necessary for the data analyzing process to bring more accurate results. As a result, churn can be predicted more precisely through analyzing transactional behavior, account activity and customer services in depth.