A

PROJECT REPORT

ON

"Bank Customer Churn Prediction"

BY

Mr. Dadaso Shivaji Patil

Submitted in Partial fulfillment of

Post Graduation Diploma in Data Science and AI

Savitribai Phule Pune, University

For the Academic Year

2023-2024

UNDER THE GUIDANCE OF
Prof. Chhaya Kadam

Department of Technology,
Savitribai Phule Pune University,
Pune-411017

SAVITRIBAI PHULE PUNE UNIVERSITY
Pune-411017

SAVITRIBAI PHULE PUNE UNIVERSITY

A

PROJECT REPORT

ON

“Bank Customer Churn Prediction”

BY

Mr. Dadaso Shivaji Patil

Submitted in Partial
fulfillment of

Post Graduation Diploma in Data Science and AI

Savitribai Phule Pune,University

For the Academic Year

2023-2024

UNDER THE GUIDANCE OF
Prof. Chhaya Kadam

Department of Technology,
Savitribai Phule Pune University,
Pune-411007

SAVITRIBAI PHULE PUNE UNIVERSITY

DEPARTMENT OF TECHNOLOGY

CERTIFICATE

This is to certify that **Mr. Dadaso Shivaji Patil (PGD23DS49)** has successfully completed her project on

"Bank Customer Churn Prediction"

In partial fulfillment of 2nd Semester work for their Post Graduation Diploma in Data Science and AI under Savitribai Phule Pune University, for the academic year 2023-2024

Prof. Chhaya Kadam
(Project Guide)

Dr. Aditya Abhyankar
(HOD)

Signed By
(External Examiner)

Dr. Manisha Bharati
(Course Co-ordinator)

Place: Pune
Date: /05/2024

STUDENT DECLARATION

We undersigned a student of the Department of Technology, Savitribai Phule Pune University, Pune PGD Data Science and AI - 2nd semester, declare that the summer internship project "**Bank Customer Churn Prediction"** is a result of our own work and our indebtedness to other work publications, references, if any, have been duly acknowledged. If we are found guilty of copying any other report or published information and showing it as our original work. We understand that we shall be liable and punishable by the Institute or University, which may include Fail in the examination, repeat study and resubmission of the report or any other punishment that Institute or University may decide.

Name of Student: Dadaso Shivaji Patil
Enrolment No.: PGD23DS49
Signature:

ACKNOWLEDGEMENT

We are deeply grateful to all those who have played a pivotal role in shaping our Capstone project and enriching our learning experience. We extend my heartfelt gratitude to Prof. Chhaya Kadam (Project Guide from the Department of Technology, SPPU) For their scholarly guidance, insightful suggestions, and continuous encouragement throughout the course of my project.

The entire faculty of the Department of Technology, Savitribai Phule Pune University, for fostering an environment of academic excellence and nurturing my curiosity.

We are profoundly thankful to everyone mentioned above for their selfless contributions to my journey. Their mentorship and support have been instrumental in shaping my skills and fostering personal growth. Thank You.

Mr. Dadaso Shivaji Patil [PGD23DS49]

ABSTRACT

The project titled "Bank Customer Churn Dataset" aims to analyze and predict customer churn in a banking context. Customer churn, which refers to the loss of clients or customers, is a significant issue for banks as it directly impacts profitability. By understanding the factors that contribute to customer churn, banks can develop strategies to retain customers and improve their services.

This project involves several stages, including data acquisition, cleaning, exploration, and modeling. The dataset, which contains various features such as customer demographics, account information, and transaction history, is first preprocessed to handle missing values and ensure data quality. Exploratory data analysis is then performed to gain insights into the patterns and relationships within the data.

Feature engineering techniques are applied to create new relevant features that can improve the predictive power of the model. Following this, feature selection methods are employed to identify the most significant features. The dataset is then split into training and testing sets to evaluate the model's performance.

Several machine learning algorithms are applied to the dataset, including logistic regression, decision trees, random forests, and gradient boosting. Hyperparameter tuning is conducted to optimize the model's performance. The models are evaluated using various metrics such as accuracy, precision, recall, and the F1-score to determine the best-performing model.

The best model is then validated on a separate test dataset to ensure its generalizability. Upon achieving satisfactory results, the model is deployed for real-time prediction of customer churn. Post-deployment, the model's performance is continuously monitored to ensure it remains accurate and effective over time.

The findings from this project provide valuable insights into the factors influencing customer churn and offer actionable recommendations for banks to enhance customer retention strategies. By leveraging advanced data analytics and machine learning techniques, this project demonstrates the potential of data-driven decision-making in addressing critical business challenges.

INDEX

CHAPTER 1

INTRODUCTION AND RATIONALE OF THE STUDY

1.1 INTRODUCTION TO THE TITLE

In today's competitive banking landscape, retaining customers is crucial for sustainable growth and profitability. Customer churn, the phenomenon where customers discontinue their relationship with a bank, poses a significant challenge to financial institutions. Identifying and predicting potential churners before they leave is paramount for banks to proactively address customer needs, enhance satisfaction, and maintain loyalty.

This project focuses on employing machine learning techniques to predict customer churn in a bank setting. By analyzing historical data of customer interactions, transactions, and demographic information, we aim to build a predictive model that can forecast the likelihood of a customer churning in the near future.

The insights gained from this predictive model can empower banks to take targeted retention actions, such as offering personalized incentives, improving customer service, or adjusting product offerings, to mitigate churn and foster long-term customer relationships.

Through this project, we endeavor to demonstrate the effectiveness of data-driven approaches in helping banks anticipate and prevent customer churn, ultimately contributing to enhanced customer retention and satisfaction.

1.2 SIGNIFICANCE OF THE STUDY

Understanding why customers leave a bank, known as churn, is crucial for the banking industry. This study's importance lies in several key areas:

- **Customer Retention:** Predicting churn helps banks retain valuable customers. By identifying at-risk customers early, banks can take proactive measures to address their concerns and prevent them from leaving.

- **Cost Reduction:** Acquiring new customers is more expensive than retaining existing
- ones. Predictive churn analysis allows banks to focus resources on retaining current customers, reducing overall operational costs.
- **Service Improvement:** Insights from churn prediction models can highlight areas for service improvement. Addressing these issues not only reduces churn but also enhances overall customer satisfaction.

- **Strategic Decision-Making:** Churn prediction provides valuable insights for strategic planning. Banks can tailor their offerings and marketing strategies based on the characteristics of potential churners, improving their competitive position in the market.

- **Enhanced Customer Experience:** By anticipating customer needs and concerns, banks can deliver a more personalized and satisfying experience. This fosters long-term relationships and strengthens customer loyalty.

In conclusion, this study on bank customer churn prediction is significant as it enables banks to retain customers, reduce costs, improve services, make strategic decisions, and enhance the overall customer experience, thereby contributing to the long-term success and competitiveness of the bank.

CHAPTER 2

LITERATURE REVIEW

In the realm of customer churn prediction in the banking sector, various theoretical frameworks and studies have provided valuable insights into understanding and predicting customer behavior.

2.1 BACKGROUND

Customer Churn Prediction is important in all businesses because it helps to gain a better understanding of your customers and of future expected revenue. It can also help your business identifies and improve upon areas where customer service is lacking. A lot of work has been done on this and still, and there are a lot of industries customer data.

2.2 CUSTOMER CHURN PREDICTION

The term Customer Attrition refers to the customer leaving one business service to another. Customer Churn Prediction is used to identify the possible churners in advance before they leave the company. This step helps the company to plan some required retention policies to attract the likely churners and then to retain them which in turn reduces the financial loss of the.

Customer churn is a concern for several industries, and it is particularly acute in the strongly competitive industries. Losing customers leads to financial loss because of reduced sales and leads to an increasing need for attracting new customers. Customer retention is crucial in a variety of businesses as acquiring new customers is often more costly than keeping the current ones. Due to the unpredictable nature of customers, it is quite a daunting task to predict whether the customer will quit the company or not. For financial institutes, it is even more complex to identify the customer churn due to the sparsity of the data as compared to another domain. This requires longer investigation periods for churn prediction.

The economic value of customer retention is widely recognized (Poel & Lariviere, 2004):

(1) Successful customer retention allows organizations to focus more on the needs of their existing customers instead of seeking new and potentially risky ones.
(2) Long term customers would be more beneficial and, if satisfied, may provide new referrals.
(3) Long term customers tend to be less sensitive towards a competitive market.

(4) Long term customers become less expensive to serve due to the bank's knowledge
(5) Losing customers leads to reduced sales, and increased sales to attract new customers.

Customer Churn has become a major problem in all industries including the banking industry and banks have always tried to track customer interaction so that they can detect the customers who are likely to leave the bank. Customer Churn modelling is mainly focusing on those customers who are likely to leave and so that they can take steps to prevent churn. In an era of the competitive world, more and more companies do realize that their most precious asset is the existing customer base and their data. We mainly investigate the predictors of churn incidence as part of customer relationship management. Churn Management is an important task to retain valuable customers. Business organizations, such as banks, insurance companies, and other service providers are changing their employees to be more customer-and service oriented and they are setting strategies to ensure customer retention. The best core marketing strategy for the future is to retain existing customers and avoiding customer churn.

2.3 DATA EXPLORATION AND PRE-PROCESSING

Data Exploration is required to gain further understanding of the data and business problem. The most important stage of Data Analysis is the Data Preparation. In general, the data cleaning and pre-processing take approximately 80% of the time. The data preparation is more challenging and time-consuming part.

The Real-world data can be noisy, incomplete and inconsistent. The data preparation stage deals with – incomplete data where some attribute values were missing, where certain important attributes were missing. In the data preparation stage the outliers and errors in data were also handled, even the data discrepancies were handled in the data preparation.

Data preparation generates a smaller dataset than the original one. This task includes selecting relevant data, attribute selection, removing anomalies, eliminating duplicate records. This stage also deals with filling the missing values, reducing ambiguity and removing outliers.

This stage is of high importance due to the following:
(1) the real data is impure;
(2) high-performance mining requires quality data;
(3) quality data yields high-quality patterns

2.4 MACHINE LEARNING TECHNIQUES

Previous research has explored the effectiveness of machine learning algorithms such as logistic regression, decision trees, random forests, support vector machines, and neural networks in predicting customer churn. These techniques leverage historical data on customer demographics, transaction history, and interactions to identify patterns indicative of potential churn.

2.5 FEATURE SELECTION

Scholars have investigated the significance of different features or variables in predicting churn, emphasizing the importance of feature selection and engineering to enhance model performance. Variables such as customer demographics, account activity, product usage, and customer service interactions have been identified as influential factors in predicting churn.

2.6 CUSTOMER SEGMENTATION

Segmentation approaches have been employed to group customers based on similar characteristics or behavior’s. By segmenting customers, banks can tailor retention strategies to meet the specific needs and preferences of different customer groups, thereby improving the effectiveness of churn prediction models.

2.7 EVALUATING MODELS

Finally, researchers use metrics to see how well their predictions work. They use numbers to judge accuracy, and also look at how much money a bank could lose if a customer leaves.

By synthesizing insights from existing literature and theoretical frameworks, this project aims to develop a robust churn prediction model tailored to the banking industry's unique challenges and requirements.

CHAPTER 3

OBJECTIVES AND SCOPE OF PROJECT

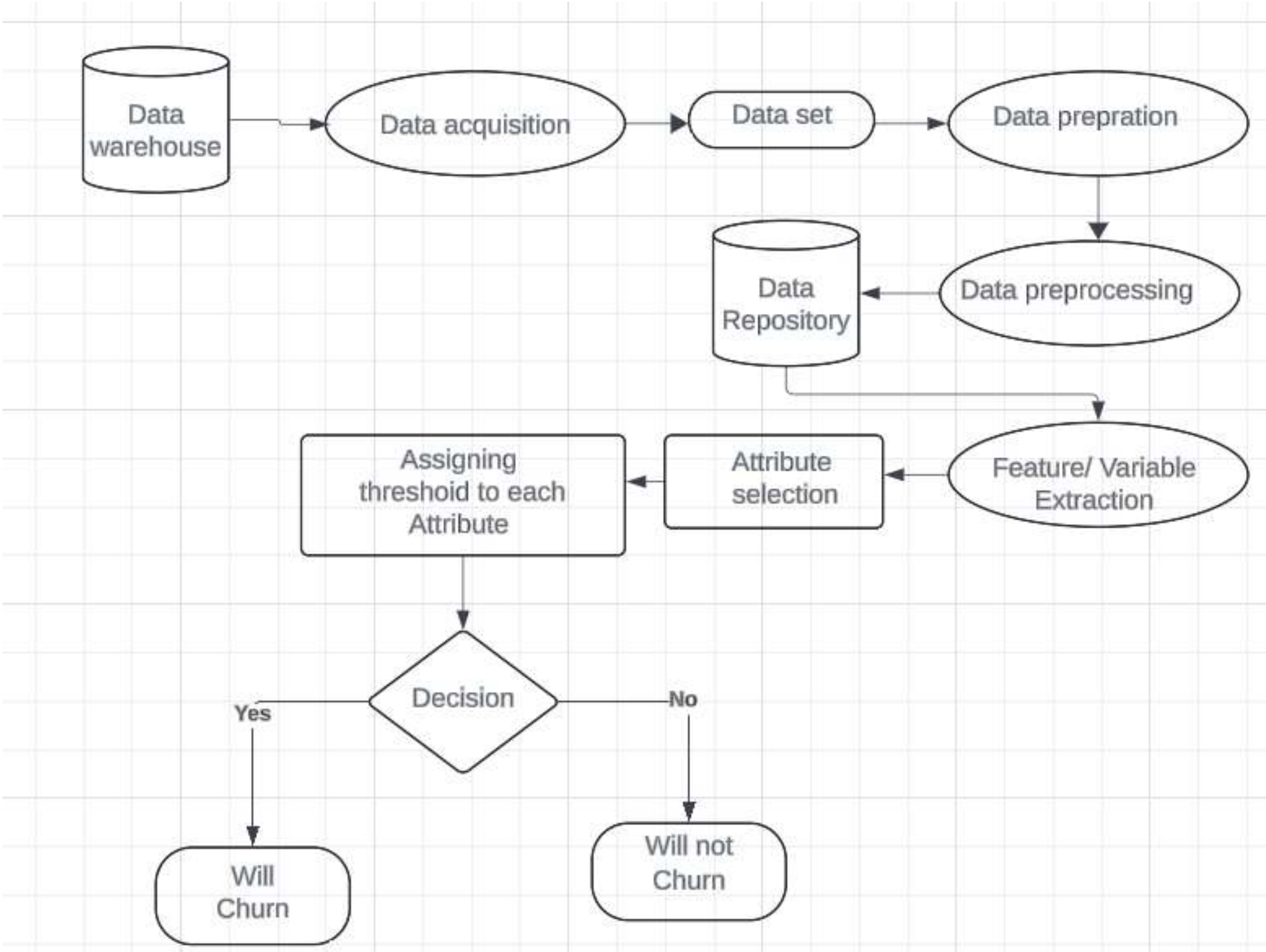
3.1 OBJECTIVES

- **Develop Prediction Model**: Create a model to forecast when bank customers might leave based on their behavior.
- **Identify Key Factors:** Figure out what factors influence customer churn the most, like age, account activity, or product usage.
- **Evaluate Model:** Test how accurate the prediction model is and make sure it works well with real data.
- **Recommend Retention Strategies:** Suggest actions banks can take to keep customers from leaving, based on the predictions.
- **Assess Impact:** See if the strategies recommended actually help reduce churn rates and improve customer satisfaction over time.

3.2 SCOPE

- **Data Collection:** Gather past customer data, like transactions and interactions with the bank.
- **Analysis:** Look into the data to find patterns and understand what might cause customers to leave.
- **Model Building:** Create a model that can predict churn based on the data patterns found.
- **Testing:** Check how accurate the model is by comparing its predictions to what actually happened.
- **Strategy Suggestions:** Based on the model's predictions, suggest different strategies banks can use to keep custome

3.3 Architecture



CHAPTER 4

RESEARCH METHODOLOGY

4.1 RATIONAL FOR THE STUDY

The study aims to address the need for banks to proactively manage customer churn, which impacts their profitability and sustainability. By predicting churn, banks can implement targeted retention strategies and improve customer satisfaction.

4.2 STATEMENT FOR PROBLEM

The problem we're tackling is the challenge of reliably forecasting whether a startup will succeed or fail. This uncertainty makes it difficult for entrepreneurs, investors, and policymakers to make informed decisions, hindering the growth of promising ventures. We need a straightforward approach to predict

4.3 SIGNIFICANCE OF THE PROBLEM

Customer churn affects a bank's revenue, reputation, and customer base. Predicting and preventing churn is essential for banks to maintain profitability, retain valuable customers, and stay competitive in the market.

4.4 RESEARCH OBJECTIVES

1. Develop a predictive model for customer churn.
2. Identify key factors influencing churn behavior.
3. Evaluate model performance and provide actionable insights for retention strategies.

4.5 SCOPE OF THE STUDY

The study focuses on analyzing historical customer data from a specific bank to predict churn. It includes developing a predictive model, evaluating its performance, and providing recommendations for retention strategies.

4.6 RESEARCH HYPOTHESIS

H0: There is no significant relationship between customer attributes and churn behavior.

H1: Customer attributes significantly influence churn behavior.

4.7 RESEARCH DESIGN (RESEARCH TYPE)

The research design employs a quantitative approach, utilizing statistical analysis and machine learning techniques for churn prediction.

4. 8 DATA SOURCE (SECONDARY DATA)

The dataset used for this project was obtained from Kaggle, specifically the “Bank Customer Churn” dataset, which can be accessed at (https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn). This dataset contains information on bank customers, including transaction history, demographics, account activity, and customer interactions, and other relevant features. The dataset comprises 10,000 rows and 18 columns.

This dataset uses the following 14 variables:

1. **Row Number:** row number index
2. **Customer-Id:** bank customer id
3. **Surname:** surname of bank customer
4. **Credit-Score:** credit score of bank customer
5. **Geography:** country of bank customer
6. **Gender:** Gender of bank customer
7. **Age:** Age of bank customer
8. **Tenure:** how long does a customer have a bank account.
9. **Balance:** bank balance of customer
10. **Num Of Products:** number of products
11. **HasCrCard:** Whether the customer has a credit card or not.
12. **IsActiveMember:** Whether the customer is active or not.
13. **Estimated Salary:** estimated salary of customer
14. **Exited:** Is customer churn or not

4.9 OUTLINE OF ANALYSIS

1.Data preprocessing: Cleaning, transformation, and feature engineering.
2.Model development: Building and fine-tuning predictive models.
3.Model evaluation: Assessing model performance using appropriate metrics.
4.Insights generation: Interpreting model results and deriving actionable
5.recommendations.

4.10 LIMITATION OF THE PROJECT

1.Limited availability of historical data for analysis.
2.Constraints in accessing specific customer information due to privacy regulations

CHAPTER 5

IMAGES (CODE SNIPPET)

5.1 IMPORT LIBRARIES:

```
[ ]  import pandas as pd
     import numpy as np
     import seaborn as sns
     %matplotlib inline
     import matplotlib.pyplot as plt
     from sklearn.preprocessing import LabelEncoder
     from sklearn.model_selection import train_test_split
     import warnings
     warnings.filterwarnings("ignore")
```

5.2 READIND THE DATA (Explore the data to get the number of rows & columns):

Reading the data

```
[ ]  data=pd.read_csv('bank churn dataset.csv')
```

Explore the data and get the number of rows & columns :

```
[ ]  data.head()
```

	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

➢ **Check shape & info of data:**

```
[ ]  data.shape

     (10000, 14)
```

```
[ ]  data.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 10000 entries, 0 to 9999
     Data columns (total 14 columns):
      #   Column           Non-Null Count  Dtype
     ---  ------           --------------  -----
      0   RowNumber        10000 non-null  int64
      1   CustomerId       10000 non-null  int64
      2   Surname          10000 non-null  object
      3   CreditScore      10000 non-null  int64
      4   Geography        10000 non-null  object
      5   Gender           10000 non-null  object
      6   Age              10000 non-null  int64
      7   Tenure           10000 non-null  int64
      8   Balance          10000 non-null  float64
      9   NumOfProducts    10000 non-null  int64
      10  HasCrCard        10000 non-null  int64
      11  IsActiveMember   10000 non-null  int64
      12  EstimatedSalary  10000 non-null  float64
      13  Exited           10000 non-null  int64
     dtypes: float64(2), int64(9), object(3)
     memory usage: 1.1+ MB
```

CHAPTER 6

DATA ANALYSIS

6.1 DATA PREPARATION:

Data preparation includes the following processes:

- Checking for data duplication. The result is that there are no duplicate data
- Checking for missing values. There are no missing values in this data
- Feature Engineering, extracting age group features, and transforming the data into the desired form.
- Encoding, converting categorical data into numerical. The encoding method used in this case is ordinal encoding.

➢ **To check the number of missing data for each column:**

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

```
RowNumber          0
CustomerId         0
Surname            0
CreditScore        0
Geography          0
Gender             0
Age                0
Tenure             0
Balance            0
NumOfProducts      0
HasCrCard          0
IsActiveMember     0
EstimatedSalary    0
Exited             0
dtype: int64
```

- ➢ **Descriptive statistics of the data set:**

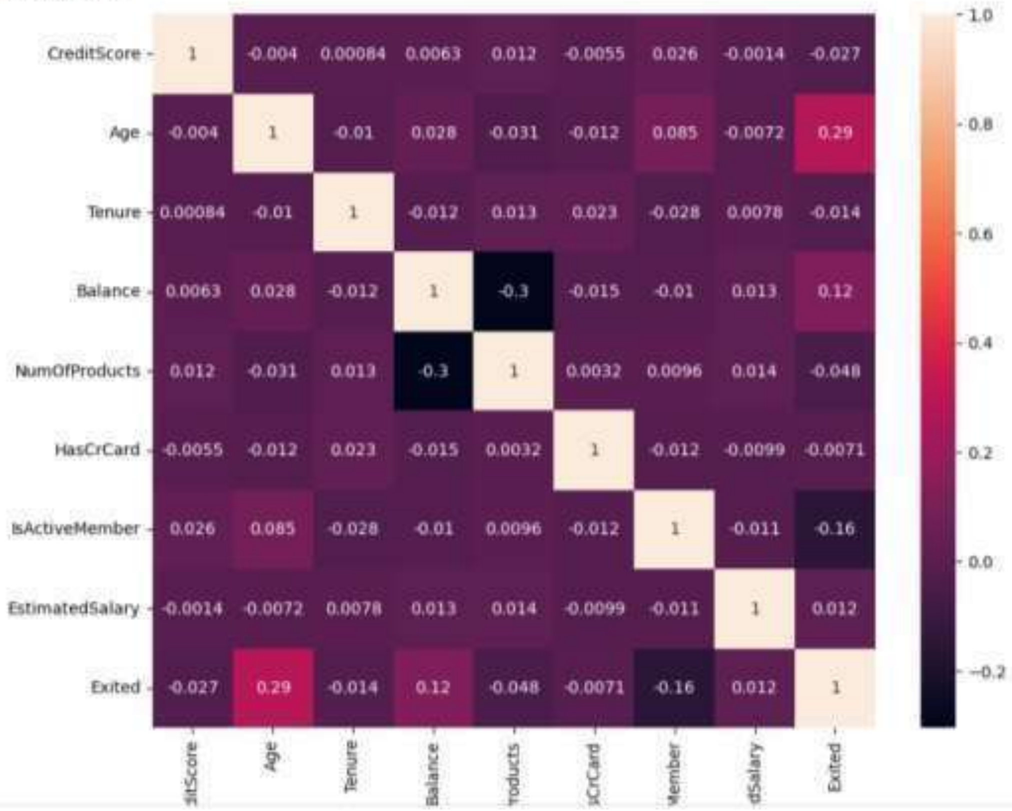
```
[ ] data.describe()
```

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

- ➢ **Correlation of data:**

```
plt.subplots(figsize=(10,8))
sns.heatmap(data.corr(), annot=True)
```

<AxesSubplot:>

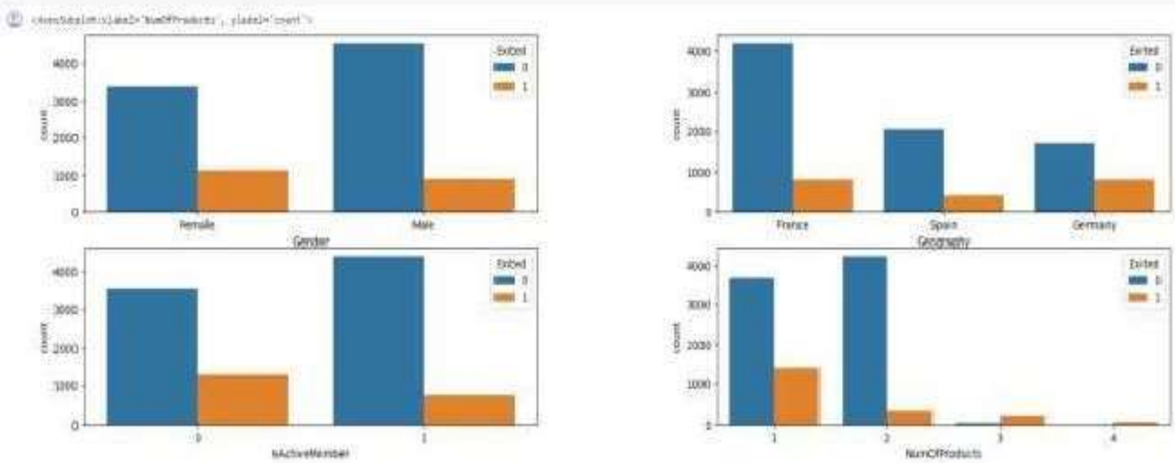


6.2 EXPLORATORY DATA ANALYSIS (EDA)

Data Visualization plays an important role in understanding the data as it gives an overview of the data before the modelling process. Exploratory Data Analysis (EDA) is a method of analyzing the data and summarizing the outcomes or insights using visual techniques. EDA is used for initial investigation of the data to produce useful outcomes and visual representations which can be helpful in learning the data before building Machine Learning algorithms.

- **Countplot of categorical variables (Gender, Geography, IsActiveMember, NumOfProducts) who get exited or not**

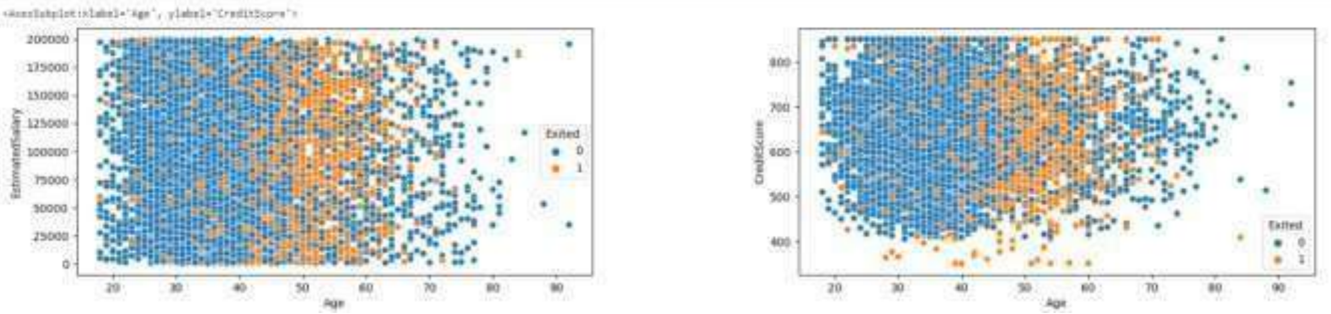
```
fig, ax = plt.subplots(2,2, figsize=(20, 6))

plt.subplots_adjust(wspace=0.4)

sns.countplot(x='Gender',hue='Exited', data=data, ax=ax[0][0])
sns.countplot(x='Geography', hue='Exited', data=data, ax=ax[0][1])
sns.countplot(x='IsActiveMember', hue='Exited', data=data, ax=ax[1][0])
sns.countplot(x='NumOfProducts', hue='Exited', data=data, ax=ax[1][1])
```



From above fig. Customers who have num of product 3 or 4 have higher chances get exited.

- **Scatterplot of age against EstimatedSalary and credit score who get exited or not**

```python
fig, ax = plt.subplots(1,2, figsize=(20, 4))

plt.subplots_adjust(wspace=0.4)

sns.scatterplot(x='Age', y='EstimatedSalary', hue='Exited', data=data, ax=ax[0])
sns.scatterplot(x='Age', y='CreditScore', hue='Exited', data=data, ax=ax[1])
```



From above fig. Customers whose age between 45 to 65 have higher chances get exited.

- **creating dummy variables for the categorical features**

```python
[ ] encoder = LabelEncoder()
    data["Geography"] = encoder.fit_transform(data["Geography"])
    data["Gender"] = encoder.fit_transform(data["Gender"])
```

6.3 PREDICTION WITH ML MODELS

- **Splitting the Dataset into Dependent and Independent Variables**

```python
[ ]  x = data.drop("Exited", axis=1)
     y = data["Exited"]
```

- **Splitting the dataset into Training and Testing Data**

```python
[ ] x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2, random_state = 42)
```

- **Training and Evaluation of Machine Learning Models**

We divide data into training and test set. Now we use the three most commonly used machine learning algorithms: random forest, support vector machines, logistic regression

Random Forest Classifier:

```
[ ] from sklearn.ensemble import RandomForestClassifier
    rfc=RandomForestClassifier(n_estimators=200,criterion='entropy',random_state=7)
    rfc.fit(x_train,y_train)
    rfcpred=rfc.predict(x_test)

    from sklearn.metrics import accuracy_score
    print("Random_Forest :", accuracy_score(y_test,rfcpred)*100)

    Random_Forest : 86.95
```

Support Vector Machines:

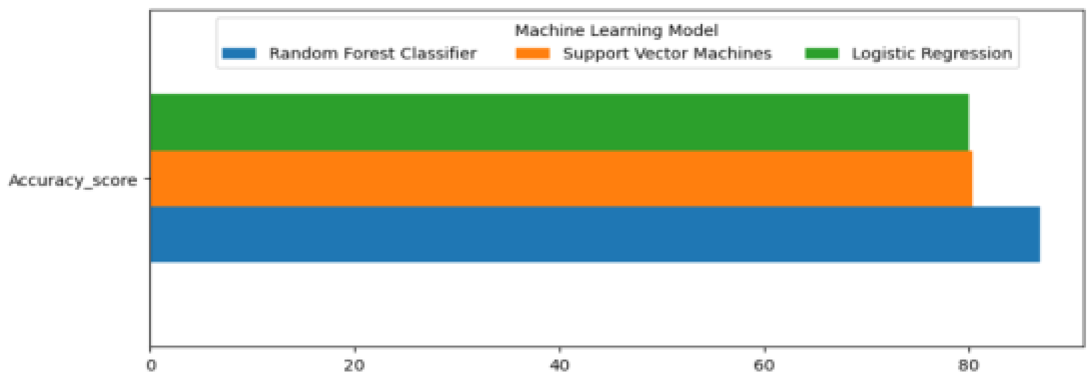
```
from sklearn.svm import SVC
svm=SVC()
svm.fit(x_train,y_train)
svmpred=svm.predict(x_test)

print("Support_Vector :", accuracy_score(y_test,svmpred)*100)

Support_Vector : 80.35
```

Logistic Regression:

```
[ ] from sklearn.linear_model import LogisticRegression
    lr= LogisticRegression()
    lr.fit(x_train,y_train)
    pred=lr.predict(x_test)

    print("Logistic :", accuracy_score(y_test,pred)*100)

    Logistic : 80.05
```

- **Gathering accuracy score for each model**

```
Accuracy_score = { 'Random Forest Classifier ': { 'Accuracy_score': (accuracy_score(y_test,rfcpred)*100)},
                   'Support Vector Machines': {'Accuracy_score':  (accuracy_score(y_test,svmpred)*100)},
                   'Logistic Regression': {'Accuracy_score': (accuracy_score(y_test,pred)*100)}}

# Plotting comparsion of each model
Accuracy_score = pd.DataFrame(Accuracy_score)
Accuracy_score.plot(kind="barh",figsize=(10,4)).legend(loc='upper center', ncol=3, title="Machine Learning Model")
```



Conclusion: Random Forest Classifier is best fitted which gives accuracy score 86.95%.

CHAPTER 7

CONCLUSION

- Random Forest Classifier is best fitted which gives accuracy score 86.95%.
- Bank customers in the 45–65 age group have a higher churn percentage than other age groups
- Female bank customers churn the most compared to males
- The most important features include the estimated salary, whether the customer is active member or not, whether the customer has a credit card or not, the number of products the customer has bought through the bank and the balance in their account.
- By utilizing this predictive power of the model, the bank can take proactive actions to prevent customer attrition and improve overall customer retention rates.

CHAPTER 8

BIBLIOGRAPHY

1. Hemlata Dalmia, Ch V S S Nikil and Sandeep Kumar,”
Churning of Bank Customers Using Supervised Learning”, January 2020, DOI:10.1007/978-981-15-3172-9_64, In book: Innovations in Electronics and Communication Engineering (pp.681-691),
https://www.researchgate.net/publication/340855263_Churning_of_Bank_Customers_Using_Supervised_Learning

2. Elyusufi, Y., & Ait Kbir, M. (2022). Churn prediction analysis by combining machine learning algorithms and
best features exploration. International Journal of Advanced Computer Science and Applications, 13(7), 615–622.
https://doi.org/10.14569/IJACSA.2022.0130773

3. Prashant Verma1,” Churn Prediction for Savings Bank Customers: A Machine Learning Approach”, J. Stat. Appl. Pro. 9, No. 3, 535-547 (2020)
https://www.naturalspublishing.com/files/published/yt9r868jnrv116.pdf

Reference Books:

R. Nageswara Rao, “Core Python Programming”, Dreamtech Press; Second edition ISBN10: 938605230X, ISBN-13: 978-9386052308 ASIN: B07BFSR3LL

3. Kevin Murphy, Machine Learning: A Probabilistic Approach, MIT Press, 1st Edition, 2012, ISBN No.: 978-0262-30616-4

L. Breiman, J. Friedman, R. Olshen, and C. Stone (1984). Classification and Regression Trees. Boca Raton, FL: Chapman & Hall/CRC (orig. published by Wadsworth).

Python Data Science Handbook - Essential Tools for Working with Data (Jake VanderPlas) 2. DATA SCIENCE AND ANALYTICS WITH PYTHON (JESUS ROGEL - SALAZAR)

CHAPTER 9

GLOSSARY OF TERMS

1. **Churn:** Churn, in the context of banking, refers to the rate at which customers discontinue their relationship with a bank by closing their accounts or ceasing to use its services**.**

2. **Customer Churn Prediction:** Customer churn prediction is the process of identifying and forecasting which bank customers are likely to terminate their relationship with the bank within a specified future time period.

3. **Feature:** A feature refers to any characteristic or attribute of a bank customer or their behavior that is used as input to a churn prediction model. Examples include transaction frequency, account balance, customer demographics, etc.

4. **Model:** A model is a mathematical representation or algorithm that is trained on historical data to make predictions about future events, such as whether a bank customer will churn or not.

5. **Supervised Learning:** Supervised learning is a type of machine learning where the model is trained on labeled data, meaning that each input sample is associated with a corresponding target label (e.g., churn or no churn).

6. **Classification:** Classification is a type of supervised learning task where the goal is to assign a category or label to each input sample. In the context of churn prediction, it involves classifying customers as churners or non-churners.

7. **Algorithm:** An algorithm is a set of rules or procedures used by a computer program to solve a specific problem. In churn prediction, various machine learning algorithms (e.g., logistic regression, random forest) can be used to build predictive models.

8. **Training Data:** Training data is the portion of the dataset that is used to train a machine learning model. It consists of input features and their corresponding target labels (e.g., churn or no churn).

9. **Testing Data:** Testing data is a separate portion of the dataset that is used to evaluate the performance of a trained machine learning model. It helps assess how well the model generalizes to unseen data.

10. **Evaluation Metrics:** Evaluation metrics are measures used to assess the performance of a churn prediction model. Common metrics include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).

11. **Feature Importance:** Feature importance refers to the degree to which each input feature contributes to the predictive performance of a churn prediction model. It helps identify which features are most influential in determining churn.

12. **Deployment:** Deployment refers to the process of integrating a trained churn prediction model into a bank's operational systems or software platforms so that it can be used to make real-time predictions on new customer data.