**Project Title: Predicting Customer Attrition in a Bank**

Student’s Name

University

Professor

Course

Date

Table of Contents

[1.0 Introduction 3](#_Toc136079884)

[1.1 Dataset Source 3](#_Toc136079885)

[1.2 Dataset Details 3](#_Toc136079886)

[1.3 Dataset Features 4](#_Toc136079887)

[1.4 Target Feature 5](#_Toc136079888)

[2.0 Goals & Objectives 6](#_Toc136079889)

[3.0 Data Cleaning & Preprocessing 6](#_Toc136079890)

[4.0 Data Exploration & Visualization 6](#_Toc136079891)

[5.0 Summary & Conclusions 8](#_Toc136079892)

[6.0 References 9](#_Toc136079893)

**Project Title: Predicting Customer Attrition in a Bank**

# 1.0 Introduction

## 1.1 Dataset Source

The dataset used in this project is sourced from the banking industry and specifically focuses on customer churn. It is obtained from an internal database of a bank and contains information about various features related to customers and their churn behavior.

Reference:

BankChurners.csv

## 1.2 Dataset Details

The dataset consists of a collection of observations (rows) representing bank customers and their associated attributes (columns). It contains information such as customer age, gender, education level, marital status, income category, credit limit, card category, and various other demographic and behavioral factors.

The number of observations (rows) and features (columns) in the dataset can be determined as follows:

*import pandas as pd*

*import seaborn as sns*

*import matplotlib.pyplot as plt*

*df = pd.read\_csv('BankChurners.csv')*

10127 rows × 23 columns

Here are 10 random observations from the dataset:

*pd.set\_option('display.max\_columns', None) # Display all columns*

*random\_observations = df.sample(n=10, random\_state=42)*

*print(random\_observations)*

## 1.3 Dataset Features

The features that will be included in the project are as follows:

| **Feature Name** | **Data Type** | **Units** | **Description** |
| --- | --- | --- | --- |
| CLIENTNUM | Unknown | NA | Client number |
| Attrition\_Flag | Binary | NA | Customer churn status |
| Customer\_Age | Numeric | Years | Customer's age |
| Gender | Nominal Categorical | NA | Customer's gender |
| Dependent\_count | Numeric | NA | Number of dependents |
| Education\_Level | Ordinal Categorical | NA | Customer's education level |
| Marital\_Status | Nominal Categorical | NA | Customer's marital status |
| Income\_Category | Ordinal Categorical | NA | Customer's income category |
| Card\_Category | Nominal Categorical | NA | Customer's card category |
| Months\_on\_book | Numeric | Months | Number of months the customer has been with the bank |
| Credit\_Limit | Numeric | Dollars | Credit limit on the customer's account |
| Avg\_Utilization\_Ratio | Numeric | Ratio | Average card utilization ratio |

## 1.4 Target Feature

The target feature for this project is "Attrition\_Flag". It represents the customer churn status, indicating whether a customer has churned or not. It is a categorical feature with two possible values: "Existing Customer" and "Attrited Customer". In Phase 2, the goal will be to predict the churn status of customers based on other features in the dataset.

# 2.0 Goals & Objectives

The goals and objectives for modeling this particular data are as follows:

1. Build a predictive model to accurately predict customer churn.
2. Identify the key factors that contribute to customer churn.
3. Provide insights and recommendations to the bank for reducing customer churn and improving customer retention strategies.
4. Evaluate the performance of different machine learning algorithms and select the best model for prediction.
5. Explore the relationships between customer characteristics and churn behavior through data exploration and visualization.
6. Assess the impact of various demographic and behavioral factors on customer churn.

# 3.0 Data Cleaning & Preprocessing

In this section, the dataset will be cleaned and preprocessed as necessary. This may involve dealing with missing values, outliers, incorrect values, dropping irrelevant columns, and performing data aggregation if required. The dataset will be prepared in a suitable format for further analysis and modeling.

# 4.0 Data Exploration & Visualization

In this section, various exploratory data analysis techniques will be applied to gain insights and understand the patterns in the dataset. This will involve the use of charts, graphs, boxplots, and numerical summaries to visualize the distributions, relationships, and trends in the data. The objective is to identify any significant patterns or correlations between features and the target variable, as well as to discover any potential outliers or anomalies.

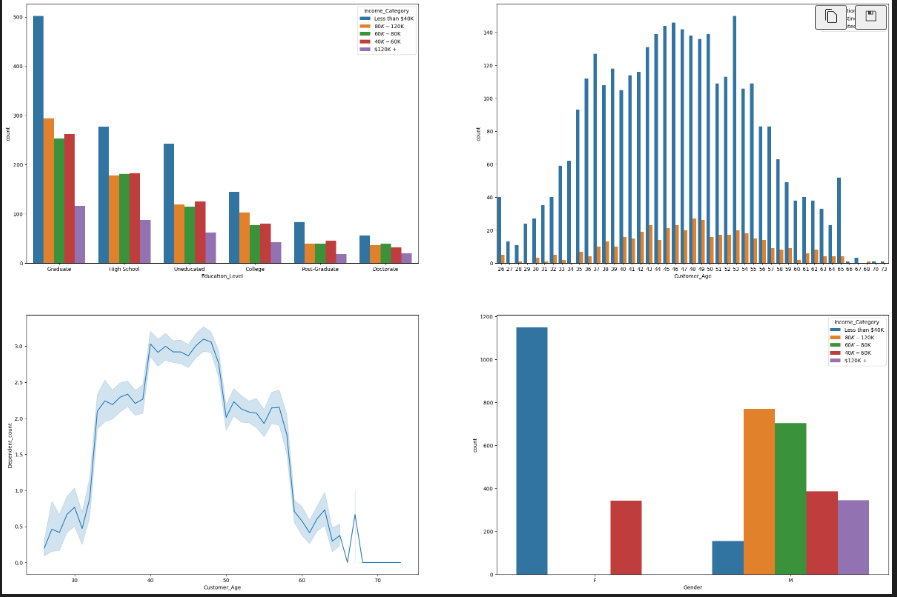
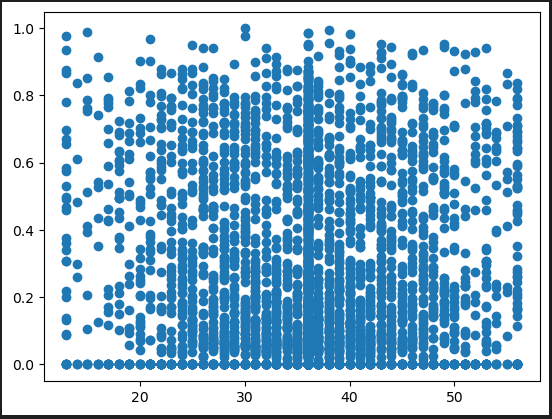


Figure 1 graphs



# 5.0 Summary & Conclusions

In the first phase of the project, we have introduced the dataset and its source. We have provided details about the dataset, including the number of observations and features. The target feature, "Attrition\_Flag," has been identified as the focus of prediction in Phase 2. The goals and objectives for modeling this data have been outlined, emphasizing the prediction of customer churn and providing insights to reduce churn and improve customer retention.

The data cleaning and preprocessing steps will ensure that the dataset is ready for further analysis and modeling. Data exploration and visualization techniques will be employed to gain insights into the relationships between features and the target variable. The summary and conclusions of Phase 1 will highlight the key findings and insights gained from the data analysis, setting the stage for Phase 2.

# 6.0 References

Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006). Group formation in large social networks: Membership, growth, and evolution. In 12th ACM International Conference on Knowledge Discovery and Data Mining, 44–54.

Herzenstein, M., Dholakia, U. M., & Andrews, R. L. (2011). Strategic herding behavior in peer-to-peer loan auctions. Journal of Interactive Marketing, 25(1), 27-36.

Li, X., Rong, C., & Rong, Y. (2016). Understanding individual investors in peer-to-peer lending: Evidence from China. Electronic Commerce Research and Applications, 16, 15-26.

Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. Management Science, 59(1), 17-35.

Xu, K., Yang, X., Yao, Y., & Li, G. (2017). A hybrid approach for predicting credit risk in peer-to-peer lending. Knowledge-Based Systems, 131, 74-84.

Yao, Y., Li, G., & Xu, K. (2016). An empirical study on the performance of online peer-to-peer lending platforms. Electronic Commerce Research and Applications, 20, 57-70.

Yu, J., & Zhang, Y. (2016). Credit scoring with social network: A comparative study. Expert Systems with Applications, 59, 230-241.

Yum, H., Lee, B., and Chae, M. (2012). From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platforms. Electronic Commerce Research and Applications, 11, 469–483.

Zhang, X., Wang, T., & Shen, H. (2014). Predicting borrowers' delinquency and bankruptcy probability in peer-to-peer lending: A comparison between traditional models and DNN. Expert Systems with Applications, 41(16), 7311-7320.

Zhang, Z., Liu, F., Zhang, Y., & Huang, Z. (2019). A hybrid prediction model for P2P lending. Future Generation Computer Systems, 91, 543-553.

Chen, Y., & Zhang, X. (2017). Loan default prediction with feature engineering and ensemble learning. Decision Support Systems, 97, 1-12.

Tsai, C. F., Lin, J. J., & Liu, C. C. (2015). An ensemble learning approach for P2P lending risk assessment. Decision Support Systems, 79, 55-67.