PH 3022

Machine Learning and Neural Computation

Group Project- Initial Report

Name: Dileka Ratnayaka

Index Number: S15436

Group 04

1. Describe the Dataset

The Bank Customer Churn prediction dataset comprises 10,006 rows and 12 columns, each offering unique insights into customer behavior within a banking context.

The dataset features diverse attributes;

- customer_id: Unique identifier for each customer.
- credit_score: Numerical representation of the customer's creditworthiness.
- country: Categorical variable indicating the country of the customer.
- gender: Categorical variable denoting the gender of the customer.
- age: Integer value representing the age of the customer.
- tenure: Number of years the customer has been with the bank.
- Account balance: Continuous variable indicating the balance in the customer's account.
- products_number: Integer value representing the number of products the customer has with the bank.
- credit_card: Binary variable indicating whether the customer has a credit card (1 for yes, 0 for no).
- active_member: Binary variable indicating whether the customer is an active member of the bank (1 for active, 0 for inactive).
- estimated_salary: Continuous variable representing the estimated salary of the customer.
- churn: Binary variable indicating the churn status of the customer (1 for churn, 0 for not churned).

The dataset features a variety of data types, including integers for numerical features like `credit_score` and `age`, floats for continuous variables such as `Account balance` and `estimated_salary`, and objects for categorical attributes like `country` and `gender`.

This comprehensive dataset enables the exploration of factors influencing customer churn and the development of predictive models using machine learning techniques such as Random Forest and Decision Trees. Understanding these dynamics can inform targeted retention strategies, ultimately contributing to improved customer satisfaction and business performance in the banking sector.

2. Explanation of Machine Learning Model and Feature Engineering Techniques

For this project, I have chosen to utilize Random Forest and Decision Tree machine learning techniques due to their compatibility with the characteristics of our dataset.

Random Forest: Our dataset contains multiple features with varying importance levels, and Random Forest is well-suited for handling such heterogeneous data. By constructing numerous decision trees during training, Random Forest can effectively capture complex relationships and interactions between the features. Moreover, its ability to handle large datasets with high dimensionality makes it an ideal choice for our dataset, which consists of over 10,000 rows and 12 columns. Additionally, Random Forest tends to perform well even without extensive hyperparameter tuning, which can expedite the model development process while still achieving robust results.

Decision Tree: Decision Tree models are particularly advantageous for datasets like ours, which contain a mix of numerical and categorical features. Decision Trees partition the input space based on feature values, enabling straightforward interpretation of the decision-making process. This interpretability is valuable for understanding the underlying patterns in the data, especially in the context of customer churn prediction where actionable insights are crucial. Furthermore, Decision Trees are adept at handling both classification and regression tasks, making them versatile for analyzing our dataset and predicting customer churn status accurately.

For feature engineering, I plan to explore techniques such as:

- Handling Missing Values: I will investigate if there are any missing values in the dataset and decide on the appropriate strategy for handling them, such as imputation or removal.

- Feature Scaling: If necessary, I will scale the numerical features to ensure that they have a similar scale, which can improve the performance of certain machine learning algorithms.

- One-Hot Encoding: I will encode categorical variables like `country` and `gender` using one-hot encoding to convert them into a numerical format that can be fed into the machine learning models.

3. Progress of the Project

I initiated the project by conducting exploratory data analysis to understand the structure and characteristics of the dataset. Using Python's pandas library, I loaded the dataset and examined its first few rows, followed by obtaining information about the dataset's attributes using the `info()` function. Additionally, I calculated summary statistics using the `describe()` function to gain insights into the distribution of numerical features.

```python
# Data Exploration
import pandas as pd

# Load the dataset
df = pd.read_csv('Bank Customer Churn Prediction Classification Dataset.csv')

# Display the first few rows of the dataset
print(df.head())

# Display information about the dataset
print(df.info())

# Summary statistics of numerical features
print(df.describe())
```

Figure 1 - Python Code for Data Exploration

Subsequently, I identified missing values in the dataset using the `isnull()` function, and by summing the missing values for each column, I assessed the extent of missing data. I then addressed missing values and cleaned the dataset by removing rows with missing values and eliminating duplicate entries using the `dropna()` and `drop_duplicates()` functions, respectively

```python
# Check for missing values
print(df.isnull().sum())
```

```
customer_id                                                    0
credit_score                                                   0
country                                                       14
gender                                                         0
age                                                            0
tenure (From how many years he/she is having bank acc in ABC Bank)     0
Account balance                                                2
products_number (Number of Product from bank)                  0
credit_card (Is this customer have credit card ?)              0
active_member (Is he/she is active Member of bank ?)           0
estimated_salary                                               3
churn (Churn Status)                                           0
dtype: int64
```

Figure 2 - Python Code for Checking missing values and the obtained results after analyzing the dataset

```python
# Handle missing values (for example, by dropping rows with missing values)
df.dropna(inplace=True)  # Assuming dropping rows is an appropriate strategy

# Remove duplicates if any
df.drop_duplicates(inplace=True)
```

Figure 3 - Python code for handling missing values and removing duplicate rows