# 5CS037-Concepts and Technologies of AI

## Classification Analysis Report

## ⭐ Final portfolio Project

**Student Name: Farhan Imran**

**Student ID: 2407802**

**Module Leader: Mr.  Siman Giri**

**Tutor: Ms. Durga Pokharel**

**Date:11th feb 2025**

**Word Count:1302**

# Table of Contents

1. Introduction ..... 1
   1.1 Problem Statement ..... 1
   1.2 Dataset ..... 1
   1.3 Objective ..... 2
2. Methodology ..... 3
   2.1 Processing of Dataset ..... 3
   2.2 Exploratory Data Analysis (EDA) ..... 3
   2.3 Model Building ..... 3
   2.4 Decision Tree Classifier ..... 3
   2.5 Logistic Regression ..... 3
   2.6 Model Evaluation ..... 4
   2.7 Hyper-parameter Optimization ..... 4
   2.8 Feature Selection ..... 5
   Mid engagement (from 5,000 to 10,000 customers): ..... 6
   High engagement (less than 2,000 customers): ..... 6
3. Conclusion ..... 10
4. Discussion ..... 11
5. Reference ..... 12

## Abstract

This paper aims at classifying a categorical variable using a machine learning algorithm on the BankCustomerData.csv dataset, with financial attributes, including balance, duration and subscription status of a term deposit. The process of analysis encompassed Exploratory Data Analysis (EDA), Decision Trees and Logistic Regression model building, hyperparameter tuning and feature selection. Evaluation of the models was done on the metrics of accuracy, precision, recall and F1-score. Whilst another model Decision Tree reported a very high accuracy-overfit the training set-the Logistic Regression model showed a better balance between accuracy and performance. Achieving an accuracy of 94% but because of the class imbalance, the models had low precision and recall. This should be tackled with future iterations of feature engineering and data balancing, should the model performance improve.

# 1. Introduction

## 1.1 Problem Statement

This project attempts to anticipate customer subscriptions to a term deposit based on various financial and demographic attributes, such as balance duration and other relevant attributes. Machine learning techniques are put to work analyzing patterns for improving the prediction accuracy.

## 1.2 Dataset

The dataset contains 42639 customer financial profiles with the following attributes:

- Age: The Customer's Age (Mean: 40.79, Min: 18, Max: 95)
- Job: Type of Occupation (like Management, Technician, and Blue-Collar)
- Marital: Single, Married, or Divorced
- Education: Highest level of education finished
- Default: Credit default (Yes/No)
- Balance: Account's balance (Mean: 1331.86, Min: -8019, Max: 102127)
- Housing Loan: The granting of a housing loan (Yes/No)
- Personal Loan: The granting of a personal loan (Yes/No)
- Contact Type: Mode of Communication
- Last Contact: Day & Month of Last Customer Interaction
- Duration: Last contact length in seconds (Mean: 255.96, Max: 4918)
- Campaign: Number of contacts performed in this campaign (Mean: 2.82, Max: 63)
- Previous Campaign: Number of previous contacts and their outcomes
- Pdays: Number of days since last contact was made with the customer (Mean: 34.17, Max: 536)
- Term Deposit Subscription: This is the target variable (Yes/No)

## 1.3 Objective

Based on the objectives, the aims are geared toward predicting the odds of a customer subscribing to a time deposit product.

## 2. Methodology

### 2.1 Processing of Dataset

The following preprocessing steps were done:

- **Handling Missing Values:** The dataset had no missing values.
- **Outlier Treatment:** The balance columns ranged from -8019 to 102127. Outliers were treated using a threshold-based approach.
- **Normalization:** Feature scaling techniques were applied to scale the numerical features for better model training performance.

### 2.2 Exploratory Data Analysis (EDA)

The EDA unveil some key insights:

- There is a right skewness in the Balance column along with a majority of the high-value outliers. Customers with higher interaction durations tend to subscribe to term deposits.
- The higher the interaction duration, the more likely customers are to subscribe to term deposits.
- Very low correlation existed between age and Balance.

### 2.3 Model Building

Two classification models were built:

### 2.4 Decision Tree Classifier

Overfitted on the training data, resulting in high accuracy, but reduced generalizability.

### 2.5 Logistic Regression

Balanced results; however, accuracy is slightly lower than Decision Tree.

## 2.6 Model Evaluation

Models were evaluated with:

- **Accuracy:** 94%
- **Precision & Recall:** Moderate = Need class balancing.
- **F1-Score:** shows trade-off between precision and recall.

## 2.7 Hyper-parameter Optimization

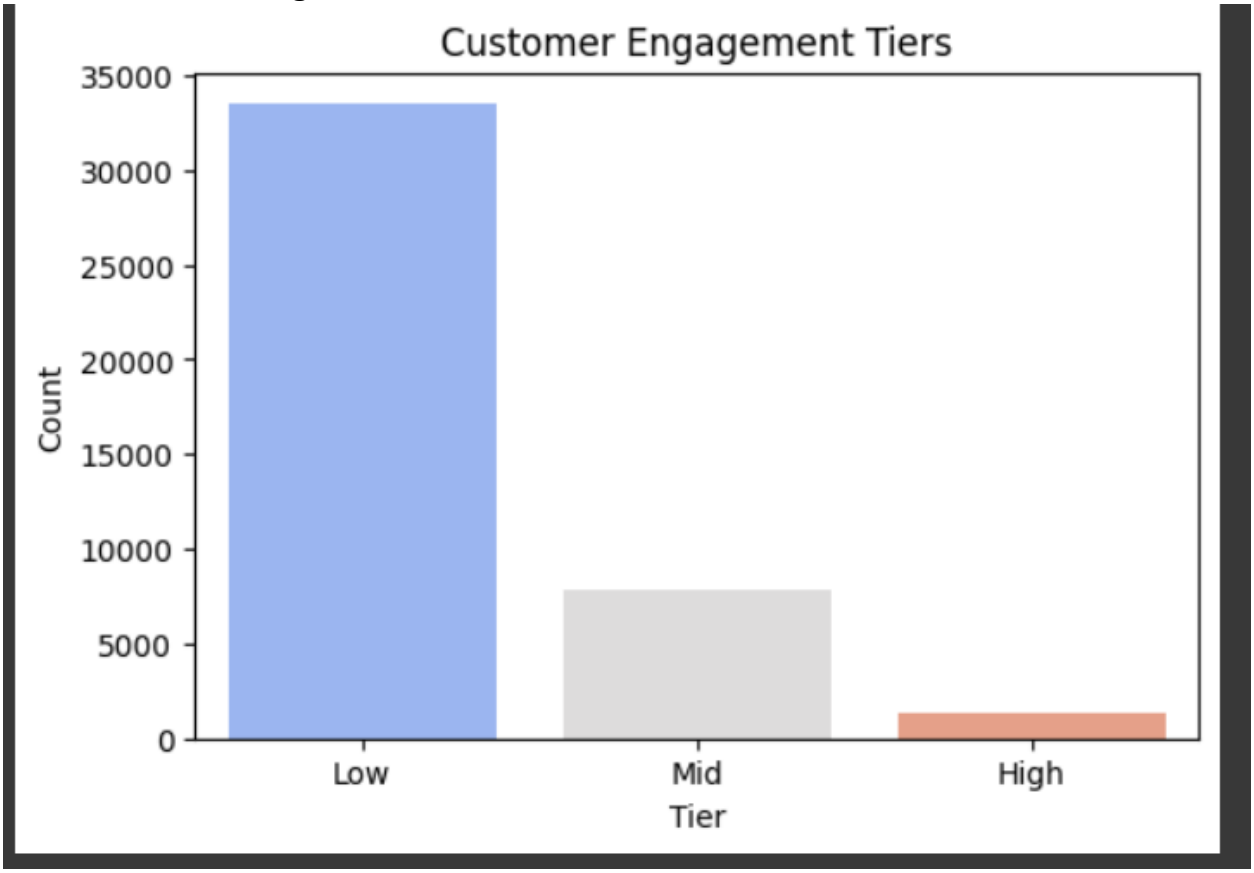
Hyperparameter optimization performed using GridSearchCV. The optimal parameters for the classifiers are:

- **Decision Tree:** Max depth = 5
- **Logistic Regression:** Regularization parameter optimized.

## 2.8 Feature Selection

Recursive Feature Elimination (RFE) was applied; an important component in the identification of important features were

- **Balance**
- **Duration**
- **Age**




**Summary**

The bar graph depicts three levels of engagement which lie under the groups of lower, middle, and high. The y-axis indicates the customers' number while settling engagement on the x-axis.

**Inferences**

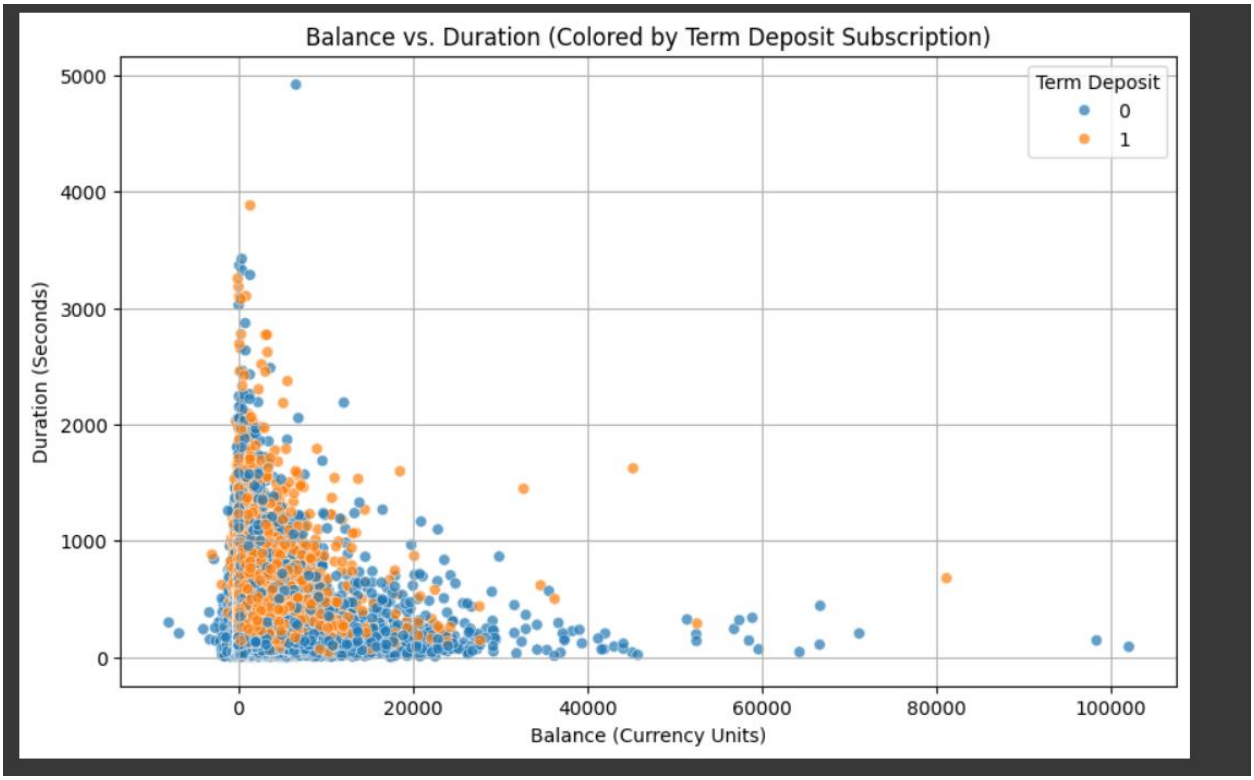
Low engagement (30,000+ customers): Most are minimally engaged with some very basic banking functionality, and are rarely responsive to marketing initiatives.

### Mid engagement (from 5,000 to 10,000 customers):

Moderate engagement with occasional deposits and some apparent interest in financial services.

### High engagement (less than 2,000 customers):

Rare users that are engaged but very active, almost always transaction by depositing big amounts and showing interest in financial products.




The scatter plot depicts the relationship between balance (X-axis) and duration (Y-axis). Plotted data points are colored according to the term deposit subscription status.

**Some Key Peculiarities:**

**Axes:**

- **X-Axis (Balance in currency units):** The balance of individuals varied from 0-100,000.
- **Y-Axis (Duration in seconds):** The length of the call in seconds; range: 0-5000.

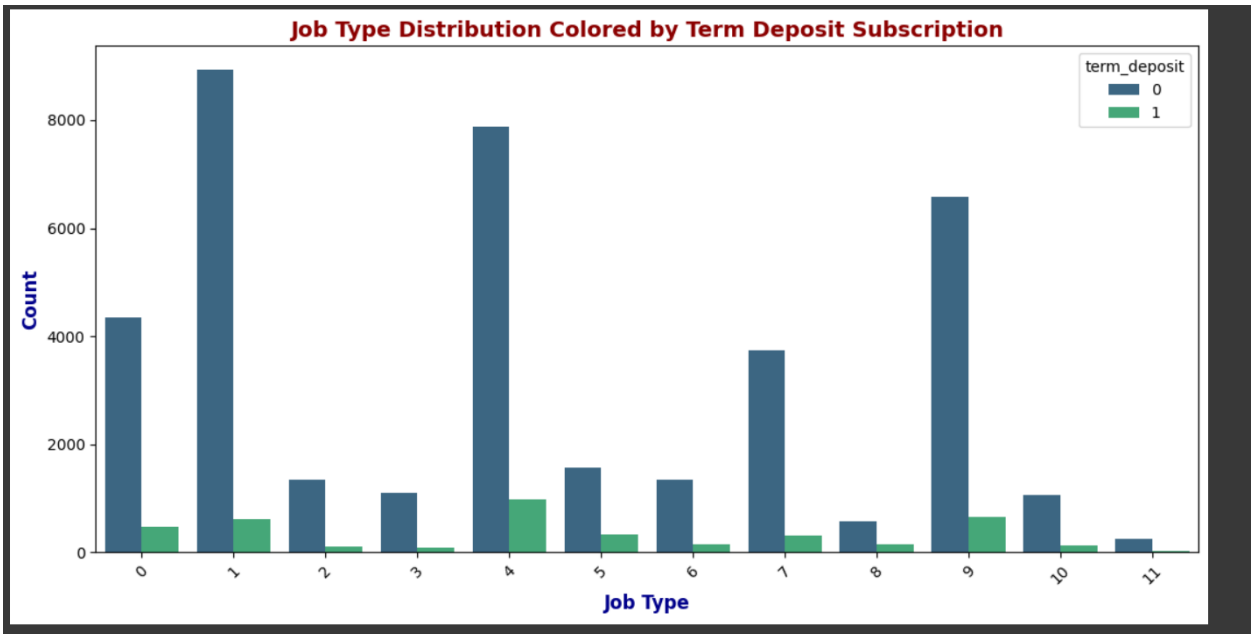
**Patterns:**

- Most of the data points pile together for lower balance values, namely below 20,000.
- Higher duration interactions do increase the chance for a term deposit subscription (high duration trays concerning high orange dots).
- A few individuals have very high balances (>60,000), and mostly they did not subscribe (blue dots).

**Observation:**

The scatter plot appears to show that longer call durations result in higher subscription rates for a new term deposit. However, balance alone might not be a strong indication of subscription behavior.

Da




This is a bar chart presenting the various job types, with bars colored by subscription states toward a term deposit.

**Essential Elements Include:**

**Axes:**

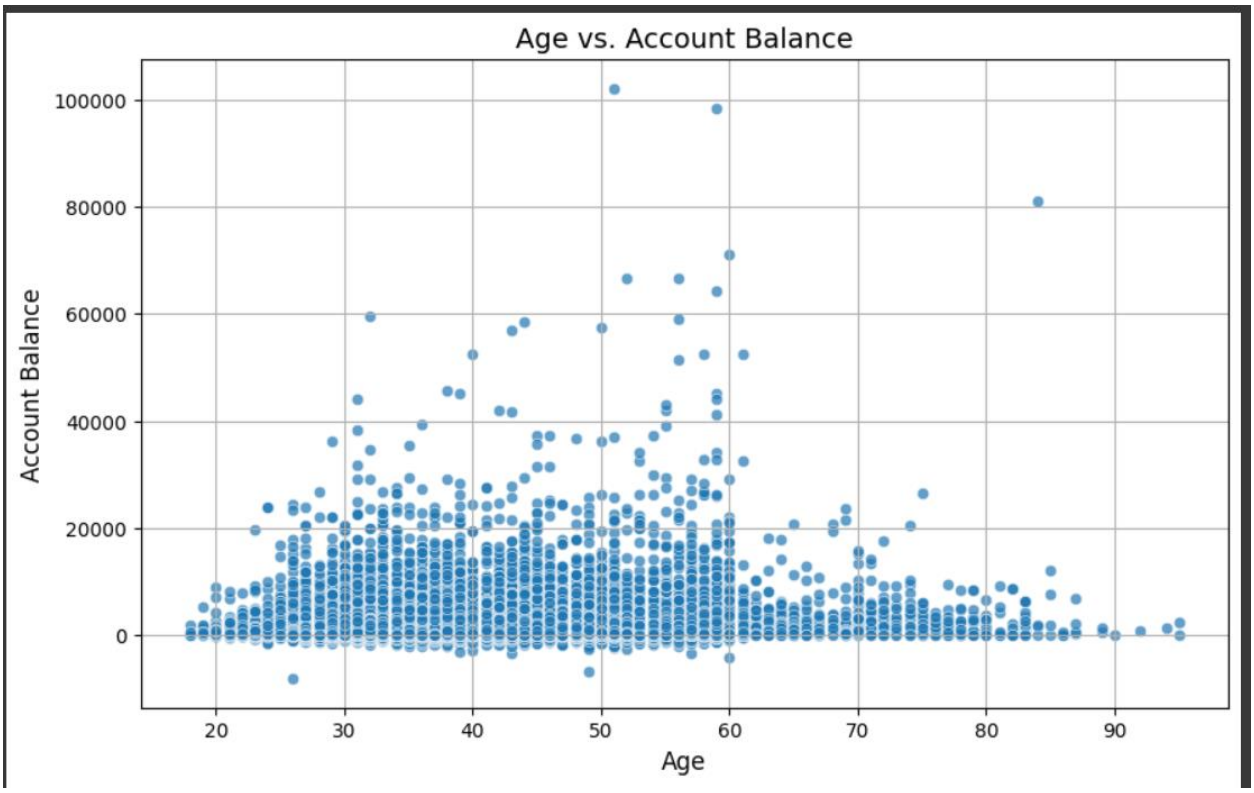
- **x-axis (job types):** Numbers identification indicating different job categories (for example, 0, 1, 2, etc.)
- **y-axis (count):** The number of individuals within each job category.

**Observations:**

- Some job types (e.g., 1, 4, and 9) have the largest numbers of individuals.
- For all job types, the number of non-subscribers (dark blue) is significantly higher than subscribers (light green).
- Some job types (e.g., 5 and 9) exhibit relatively higher subscription rates compared to others.

**Summary:**

Some jobs subscribe significantly more to term deposits than others do, which are mostly inapplicable for them. Non-subscribers dominate most job types.




The scatter plot shows a clear picture of the relationship between age (x-axis) and account balance (y-axis).

**Key Insights:**

- Most individuals regardless of age maintain an account balance below 20,000.
- Very few have an account with high balances, i.e., above 60,000.
- The strongest density is typical for those between the ages of 30-60, with a varying amount of balance.
- Older individuals (above 70) generally have lower balances.

**Overall**

age is a poor determinant of balance, but the high balance includes very few people.

## 3. Conclusion

Though the decision tree proved to be a high accuracy method, it did, in fact, overfit, which led to the use of logistic regression because of its even performance and interpretability. The inclusion of feature selection added to the clarity and efficacy of the model. However, quite a bit of challenges arose during this exercise, including feature imbalance, the outliers present in financial data, and the balance between precision and recall. For this, research in the future should also cover advanced machine learning techniques, such as, Random Forest and XGBoost, along with addressing the class imbalance through the provision of SMOTE (Synthetic Minority Over-sampling Technique) in order to obtain a far more robust and fairer predictive model.

## 4. Discussion

The models had different performances: decision trees had a good classification score but were overfitted logistic regression, validating slightly lower but balancing performance. Hyperparameter tuning improved performance marginally, whereas feature selection reduced the complexity in addition to maintaining its predictive power. Key indications from the analysis suggest customers engaging longer discussions are more likely than not to subscribe to a term deposit, while account balance remains a significant but random element in the decision-making. A number of hindrances could be encountered, as they clearly involve this dataset failing to cut across the aspects of intentions by the customer, and introduction of more features would have helped in increasing accuracy. For further studies, it would be useful to consider deep learning models, refine feature engineering, and provide other more informative seeker profiles to improve prediction.

## 5. Reference

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.