Mini Project 1 – Banking Campaign Output Prediction

Introduction

This project aims to predict whether a client will respond positively or negatively to a marketing campaign conducted by a bank. The campaign utilizes phone calls as its communication channel. The dataset provided contains information related to direct marketing campaigns and encompasses various client attributes and campaign-specific features. The success of the campaign hinges on the ability to identify potential subscribers.

In this endeavor, I aim to leverage machine learning models to predict client responses to these marketing campaigns. By analyzing the attributes of clients and features associated with the campaigns, I intend to develop predictive models capable of accurately forecasting client behavior. To achieve this, I have explored various machine learning algorithms, including random forest and decision tree models.

Data Overview

The given data of direct marketing campaigns is a common approach used by a banking institution to promote various banking products and services. For this project, the marketing campaign is done for the banking product “bank term deposit”. Term deposits are a type of investment where funds are deposited for a fixed period at a fixed interest rate, often offering returns compared to standard savings accounts. The data set provided contains a comprehensive set of features related to both clients and campaign interactions, offering valuable insights into the characteristics and behaviors of customers.

To be specific, the dataset includes the following key attributes:

Client Attributes:

- Age: The age of the client, represented as a numeric value.
- Job: The occupation or job type of the client, categorized into various professions such as 'admin', 'blue-collar', 'entrepreneur', and more.
- Marital Status: The marital status of the client, categorized as 'divorced', 'married', 'single', or 'unknown'.
- Education: The educational background of the client is categorized into different levels such as 'basic', 'high school', 'professional course', and others.
- Default: Indicates whether the client has credit in default, with possible values of 'no', 'yes', or 'unknown'.
- Housing: Specifies whether the client has a housing loan, with values of 'no', 'yes', or 'unknown'.
- Loan: Indicates whether the client has a personal loan, with values of 'no', 'yes', or 'unknown'.

Campaign-related Attributes:

- Contact: The method of communication used in the campaign, categorized as 'cellular' or 'telephone'.
- Month: The month in which the last contact was made during the campaign, represented as abbreviated month names ('jan', 'feb', 'mar', etc.).
- Day_of_week: The day of the week on which the last contact was made, categorized as 'mon', 'tue', 'wed', 'thu', or 'fri'.
- Duration: The duration of the last contact in seconds, which can influence the outcome but is known only after the call is made.
- Campaign: The number of contacts made during the campaign for a particular client.
- Pdays: The number of days that have passed since the client was last contacted by a previous campaign, with '999' indicating that the client was not previously contacted.
- Previous: The number of contacts made before the current campaign for a particular client.
- Poutcome: The outcome of the previous marketing campaign, categorized as 'failure', 'nonexistent', or 'success'.

Social and Economic Context Attributes:

- Emp.var.rate: Employment variation rate, a quarterly indicator represented as a numeric value.
- Cons.price.idx: Consumer price index, a monthly indicator represented as a numeric value.
- Cons.conf.idx: Consumer confidence index, a monthly indicator represented as a numeric value.
- Euribor3m: The Euribor 3-month rate, a daily indicator represented as a numeric value.
- Nr.employed: The number of employees, a quarterly indicator represented as a numeric value.

This rich data set provides a comprehensive overview of the factors influencing the outcomes of direct marketing campaigns in the banking sector. It offers a valuable opportunity to analyze and model client behavior, ultimately facilitating the development of predictive models to enhance campaign effectiveness and optimize resource allocation.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
 10  duration        41188 non-null  int64
 11  campaign        41188 non-null  int64
 12  pdays           41188 non-null  int64
 13  previous        41188 non-null  int64
 14  poutcome        41188 non-null  object
 15  emp.var.rate    41188 non-null  float64
 16  cons.price.idx  41188 non-null  float64
 17  cons.conf.idx   41188 non-null  float64
 18  euribor3m       41188 non-null  float64
 19  nr.employed     41188 non-null  float64
 20  y               41188 non-null  object
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
```

Figure: Information of data

Data Processing

In this step, various steps were taken to prepare a dataset for model training.

Checking Missing Values: In exploring the campaign dataset, first I checked for missing values within it. Hence, there were no missing values in the dataset.

Checking and dropping duplicate data: After that, I checked for duplicate data within the dataset to identify the duplicate rows based on their exact match across all the columns.

```
The number of duplicate data:
12
```

Figure: Duplicate data found

The found duplicate data were then removed from the dataset to ensure the data integrity and avoid redundancy. By cleaning and preparing the dataset for later project phases, this data processing step enhances the overall quality of the data and guarantees that the analysis's conclusions are supported by a trustworthy and clean dataset. Removing duplicate rows improves the finding's robustness within the campaign dataset and is in line with the best standards in data preprocessing.

Identifying numerical and categorical data columns: Initially the dataset is examined to distinguish between numerical and categorical columns. Differentiating this dataset helps to convert the data for representing unique categories.

Using One-hot encoding for Categorical Variables: Categorical variables are encoded using one-hot encoding to represent them as binary. This is required as most machine learning algorithms require numerical input. Each categorical variable is converted into multiple binary variables.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 53 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   age                             41188 non-null  int64
 1   campaign                        41188 non-null  int64
 2   pdays                           41188 non-null  int64
 3   previous                        41188 non-null  int64
 4   emp.var.rate                    41188 non-null  float64
 5   cons.price.idx                  41188 non-null  float64
 6   cons.conf.idx                   41188 non-null  float64

 48  day_of_week_thu                 41188 non-null  uint8
 49  day_of_week_tue                 41188 non-null  uint8
 50  day_of_week_wed                 41188 non-null  uint8
 51  poutcome_nonexistent            41188 non-null  uint8
 52  poutcome_success                41188 non-null  uint8
dtypes: float64(5), int64(4), object(1), uint8(43)
memory usage: 4.8+ MB
```

Figure: Data after using one-hot encoding

Scaling Features with Min-Max Scaling: Min-max scaling is applied to the feature variables to ensure that all features are within the same scale range, typically between 0 and 1. This normalization technique helps to prevent features with larger magnitudes from dominating the learning process and ensures a more balanced model performance.

Feature Selection: Feature selection is performed using the SelectKBest method with the chi-squared scoring function. It selects the top ‘k’ features that are the most relevant to the target variable.

Oversampling with ADASYN: The dataset is oversampled using the Adaptive Synthetic Sampling Approach for an imbalanced learning technique.

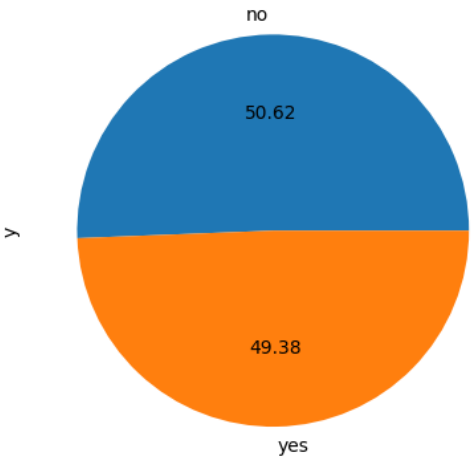

Figure: Dataset with yes/no after oversampling

It generates synthetic samples for the minority class to address class imbalance, ensuring a more balanced distribution of target classes. This step helps improve the model’s ability to capture patterns in the minority class and enhances the overall predictive performances.

Exploratory Data Analysis (EDA)

Exploratory data analysis is a crucial step in the data analysis process. The objective of this is to gain insights and an understanding of the underlying patterns and characteristics of the dataset. A distribution analysis was carried out of a few important features like age, marital status, job, and so on.

Distribution of Client's Age

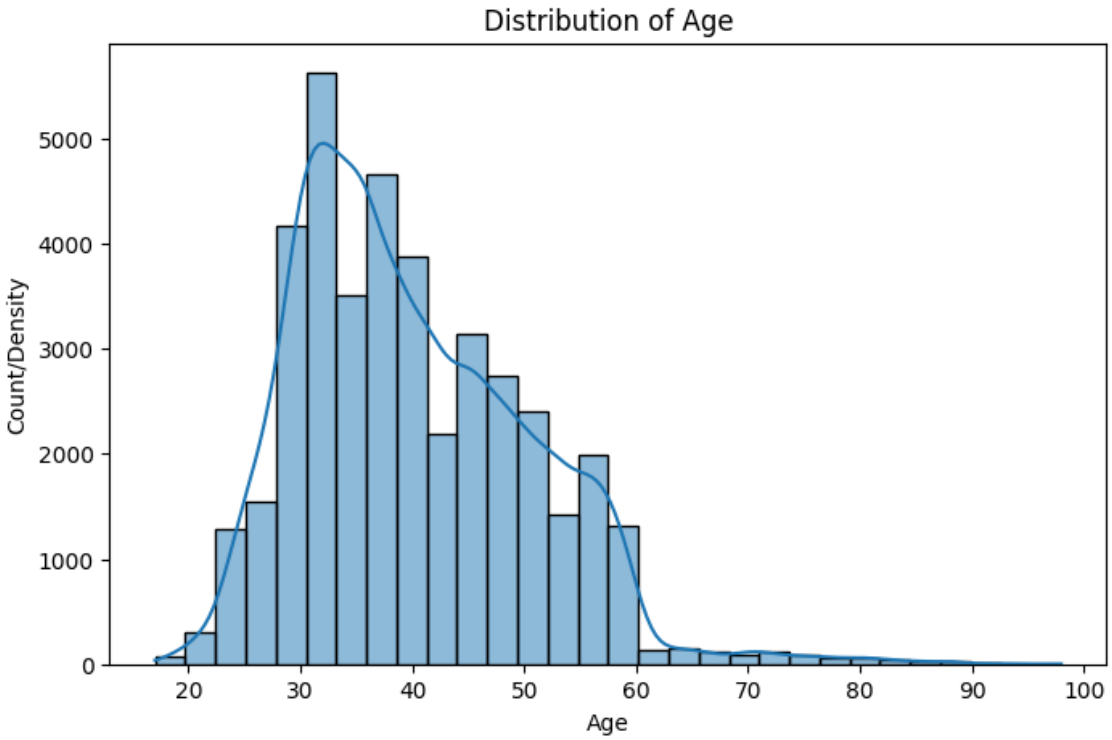

Figure: Distribution of Client's Age

These steps involve visualizing the distribution of client ages in the dataset. Understanding the age demographics helps identify the demographic composition of the clients and potential age-related trends or patterns. Here, the clients between the age group of 30 to 40 are highest in numbers. While people aged above 60 are very few and likely to reduce after that.

Distribution of Client's Job

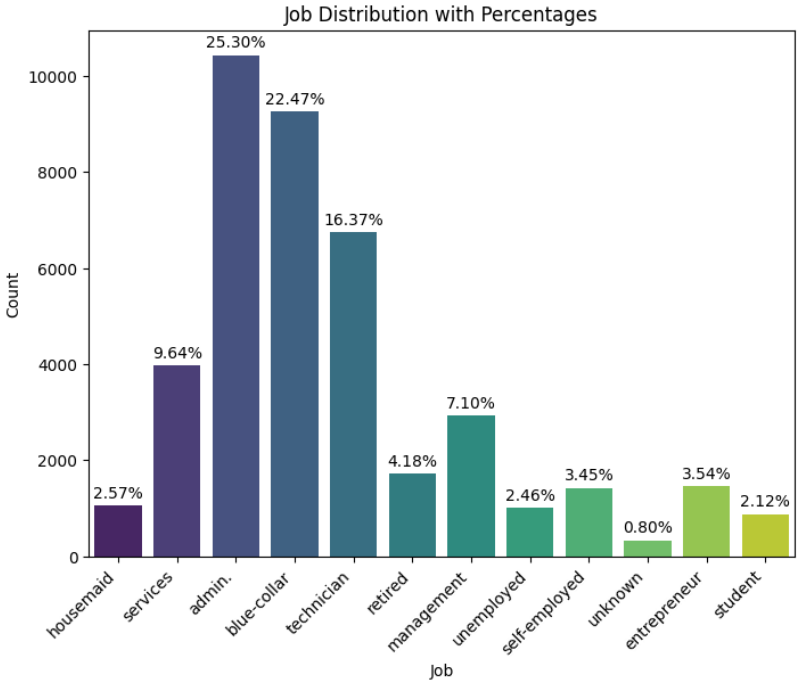

Figure: Distribution of Job

Analyzing the distribution of job categories among clients gives information about the occupational diversity within the dataset. Among different job categories, most of the client's occupation is found to be admin with 25.30% followed by 22.47% blue collar jobs and 16.37% technician.

Distribution of Client's Marital Status

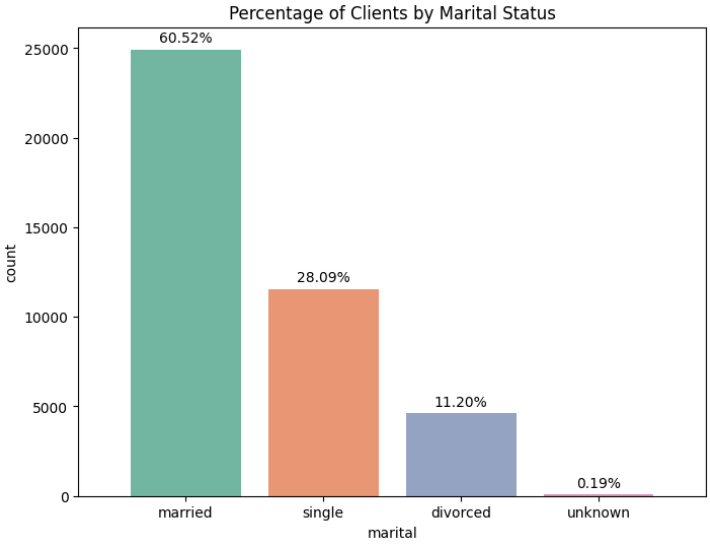

Figure: Distribution of Client's Marital Status

Marital status distribution helps to understand the relationship status of the target audience. It gives clear information about the marital composition within the dataset. As given in the figure above, among all clients most of the clients are married in status with 60.52% followed by single and divorced respectively.

Distribution by Clients' education level

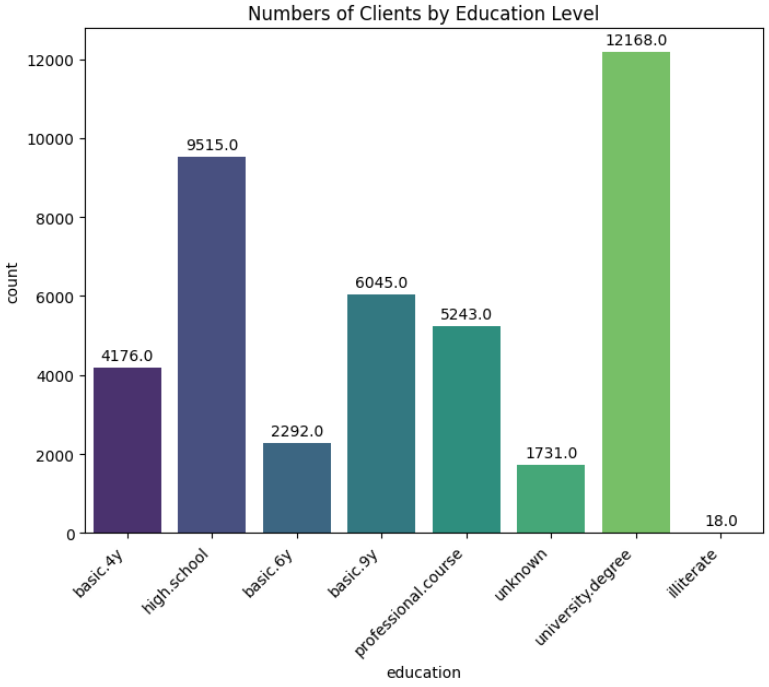

Figure: Distribution by education level

Understanding the education level distribution helps in tailoring marketing strategies and communication styles. 12168 of the clients are found to have a university degree.

Distribution of Term Deposit Subscription Response

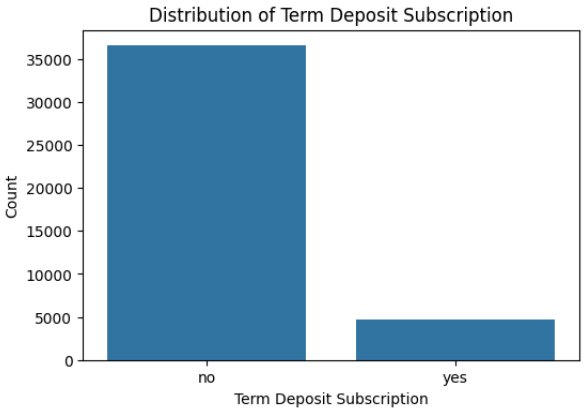

Figure: Distribution of Client Response

Understanding the subscription distribution helps in evaluating campaign effectiveness and informing future campaign strategies. As shown in the figure, most of the client's response towards the product of the campaign is 'No'.

Distribution by Default status, Housing loan status, and Personal loan status

Figure: Distribution by default, housing loan, and personal status

This analysis provides information about the financial and credit history of the clients and informs about the risk profiles of the target audience. 52.30% of clients with housing loan status are found to have positive responses to the banking campaign product. In the same way, most of the clients with default and personal loan status have negative responses to the campaign product.

Correlation between numerical columns

The correlation heatmap provides a visual representation of the pairwise correlations between numeric columns in the dataset. It helps to identify how different features relate to each other in the dataset. With highest positive correlation value of 0.97 between the employment variation rate and euribor 3-month rate suggests a strong positive linear relationship between these two variables. It focuses on the importance of considering macroeconomic variables when analyzing any banking data and making strategic decisions in the financial sector.

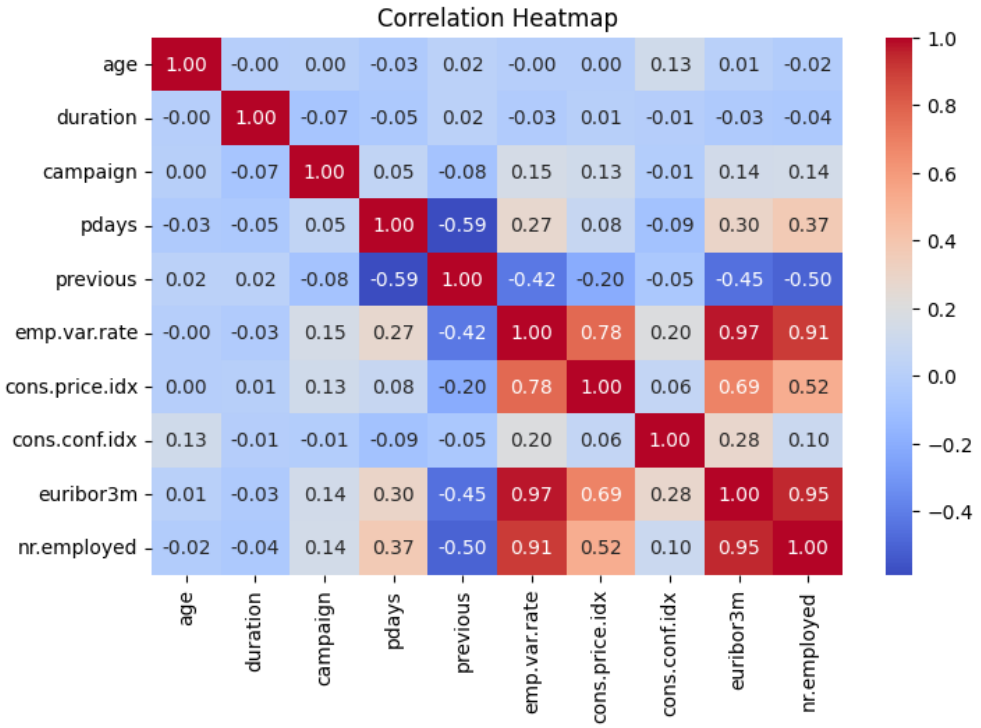

Figure: Correlation Heatmap between numerical variables

Modelling

For this project, I have used three different models to evaluate the performance, namely Decision Tree Classifier, Random Forest Classifier, and Gradient Boosting Classifier. Among this, I will explain the performance of two models only, i.e Decision tree Classifier and Random Forest Classifier. Both models exhibit high precision, recall, F1-score, and accuracy, indicating strong predictive performance. Before training these models, I had to oversample the dataset using the Adaptive Synthetic Sampling Approach because of the imbalanced dataset.

The reason behind using the decision tree classifier and random forest classifier is because of their scalability features to handle large datasets, feature importance, and so on. It offers a balance between interpretability, predictive performance, and scalability, making them suitable choices for analyzing banking campaign data and predicting client's responses.

The dataset is split into train sets and test sets with 20% for testing and the rest for training.

Model 1: Decision Tree Classifier

Decision trees can capture non-linear relationships between input and features and the target variables. In a banking campaign like this, generally, customer behavior and response patterns may not always follow linear trends. So, using decision trees is well-suited for capturing complex decision boundaries.

Classification report for Decision Tree Classifier Model

	precision	recall	f1-score	support
no	0.90	0.88	0.89	7284
yes	0.88	0.90	0.89	7158
accuracy			0.89	14442
macro avg	0.89	0.89	0.89	14442
weighted avg	0.89	0.89	0.89	14442

Figure: Classification report for Decision Tree Classifier

As shown in the figure above, the accuracy of the decision tree classifier is approximately 0.89, indicating 89% of the predictions made by the model are correct. With a precision value of 89%, it indicates that the model's performance is accurate for both positive and negative responses. The F1-score for both classes is approximately 0.89, indicating a good balance between precision and recall.

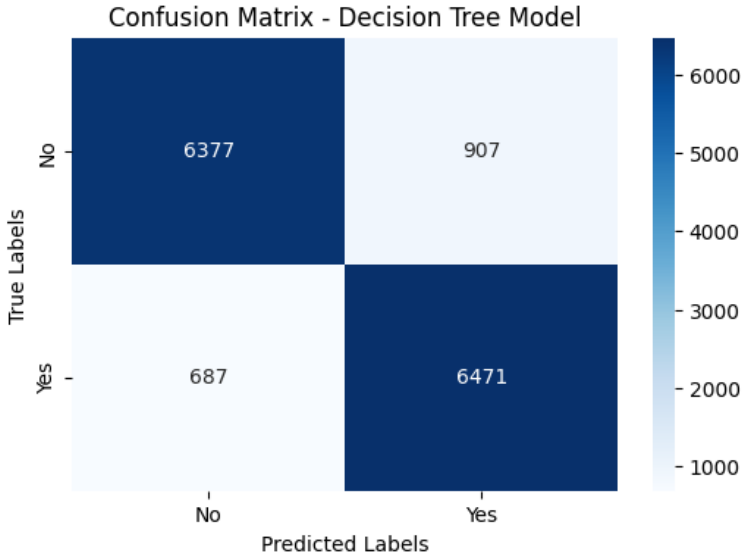

Figure: Confusion matrix- Decision Tree

The confusion report provides information into the performance of the decision tree model by showing the number of correct and incorrect predictions for each class, allowing for a comprehensive evaluation of its effectiveness in classification tasks. The model indicates the true negative predictions and positive predictions made by a model, where 6377 instances were correctly classified as negative, and 6471 instances were correctly classified as positive.

Model 2: Random Forest Classifier

Random forest classifier is an ensemble learning method that builds multiple decision tree and combines predictions to improve accuracy and reduce overfitting. It often yields more robust and accurate results compared to individual decision trees.

Classification report for Random Forest Classifier

	precision	recall	f1-score	support
no	0.93	0.93	0.93	7284
yes	0.93	0.92	0.93	7158
accuracy			0.93	14442
macro avg	0.93	0.93	0.93	14442
weighted avg	0.93	0.93	0.93	14442

Figure: Classification report for Random Forest Classifier

The accuracy of the Random Forest Classifier is approximately 0.93, indicating that around 93% of the model predictions are correct. F1-scores for both classes are approximately 0.93, indicating a good balance between precision and recall.

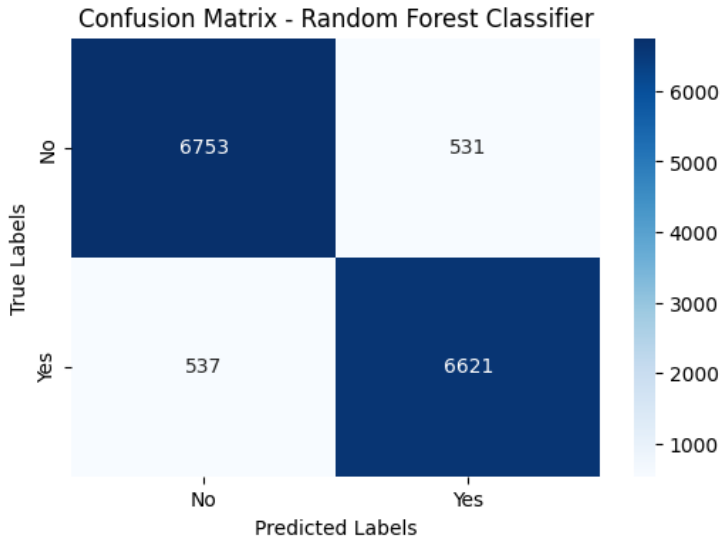

Figure: Confusion Matrix: Random Forest Classifier

According to the data given in the above figure, the confusion report of the random forest classifier indicates that the model correctly classified 6753 instances as negative and 6621 instances as positive.

Feature Importance

In the case of the random forest classifier, the figure below shows that the most influential features for this machine learning model are age, campaign, and euribor3m. It helps to understand valuable insights into prediction ability.

Similarly, for decision trees, the most influencing features are age, nr.employed, and euribor3m.

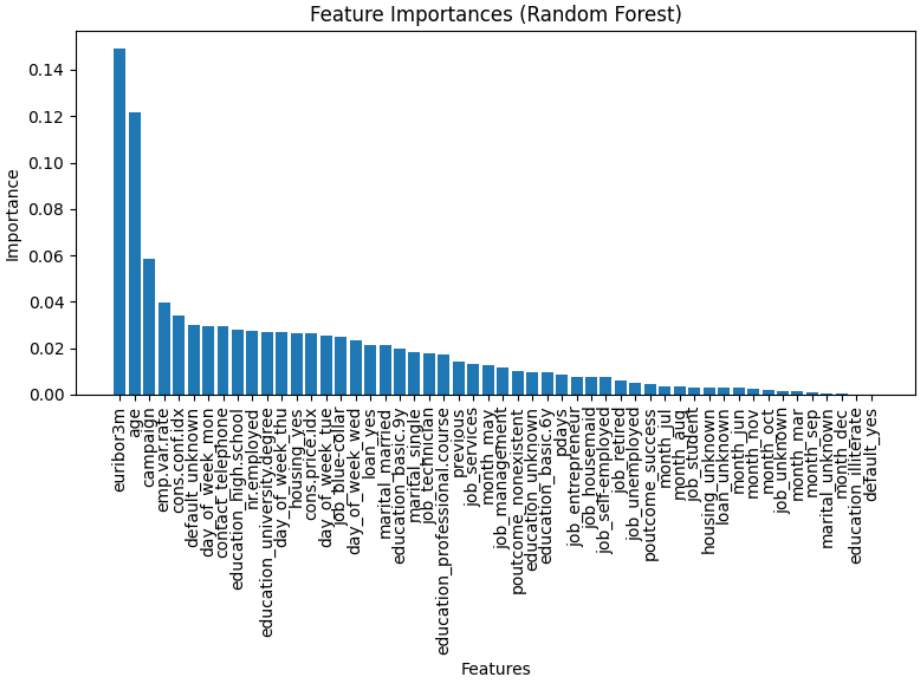

Figure: Feature Importance for Random Forest

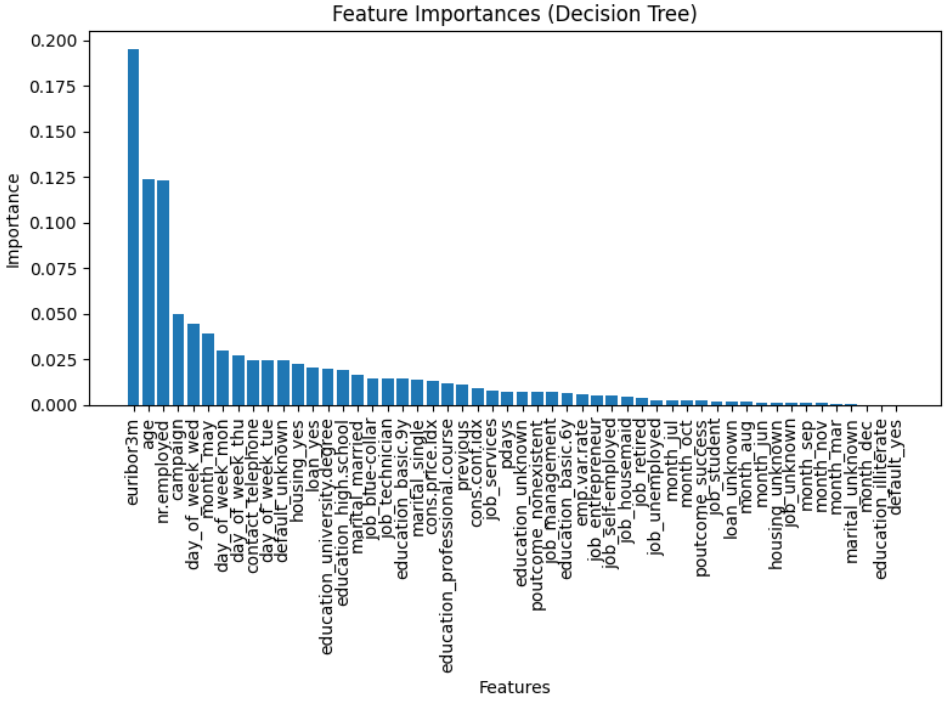

Figure: Feature Importance for Decision Tree

While comparing both models, I tried showing with visualization as:

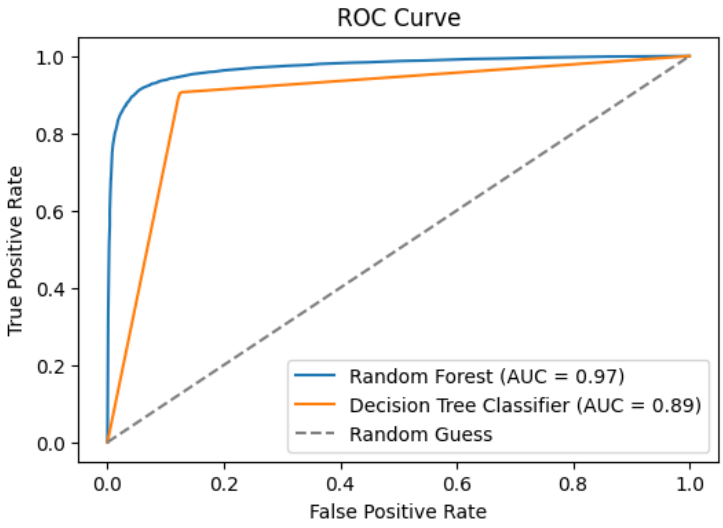

Figure: ROC Curve

The ROC (Receiver Operating Characteristics) curve is a graphical representation of the performance of a binary classification model across different threshold values.

As shown in the figure, the ROC curve for random forest is moving upwards towards the top-left corner signifying that the model is effectively distinguishing between positive and negative instances, resulting in better overall classification performance. It indicates that the model is making fewer errors and has higher predictive accuracy.

With an AUC of 0.97, the random forest model demonstrates excellent discrimination ability. The decision tree achieved an AUC of 0.89 which is still considered good but slightly lower than that of the random forest classifier.

Conclusion

In conclusion, the analysis demonstrates the effectiveness of machine learning techniques in predicting client responses to marketing campaigns in the banking sector. After checking the missing values and duplicates, data was visualized with a few of the variables from the dataset to understand the distribution features of the data. The provided data was imbalanced due to which while training with different other models, I was not getting the expected F1-value and precision value even though the accuracy rate was almost satisfying. Because of this, I tried oversampling the dataset using the Adaptive Synthetic Sampling Approach.

Both models, Decision Tree Classifier and Random Forest Classifier exhibit high precision, recall, F1-score, and accuracy, indicating strong predictive performance. The Random Forest Classifier slightly outperforms the Decision Tree Classifier in terms of accuracy, precision, recall, and F1-

score, suggesting that the ensemble approach of combining multiple decision trees leads to improved predictive performance. Overall, both models demonstrate good performance in predicting client responses to the banking campaign, with the Random Forest Classifier showing slightly better results.