BDM 3603: BIG DATA FRAMEWORK

Submitted to: VASIL KHACHIDZE

Submitted on: 2024-08-11

FINAL PROJECT REPORT

AJOMON JOSE c0900975

INTRODUCTION

This report details the process and findings of a predictive modeling project focused on direct marketing campaigns conducted by a Portuguese banking institution. The campaigns involved contacting clients through phone calls to determine whether they would subscribe to a term deposit. The primary objective of this project was to build a machine learning model capable of predicting whether a client would subscribe to a term deposit based on their demographic and socio-economic features.

DATASET OVERVIEW

The dataset used in this project contains information on direct marketing campaigns conducted by a Portuguese bank. The data includes a variety of features such as demographic details, past interactions with the bank, and other socio-economic indicators.

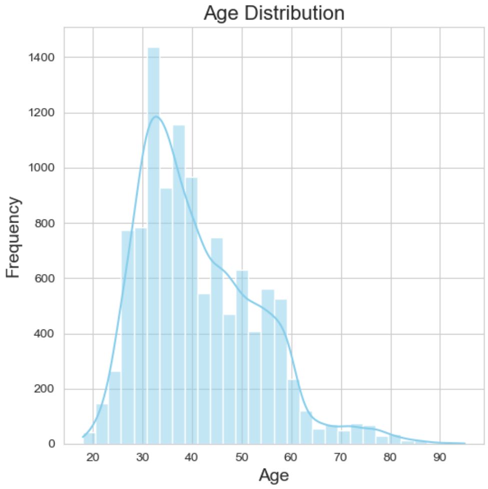
- **Target Variable**: The target variable (y) indicates whether the client subscribed to the term deposit (yes/no).
- **Features**:
    - **Bank Client Data**:
        - age: Age of the client.
        - job: Type of job.
        - marital: Marital status.
        - education: Level of education.
        - default: Has credit in default? (binary: 'yes','no')
        - balance: Average yearly balance.
        - housing: Has a housing loan? (binary: 'yes','no')
        - loan: Has a personal loan? (binary: 'yes','no')
    - **Related to the Last Contact of the Current Campaign**:
        - contact: Communication type (e.g., cellular, telephone).
        - day: Last contact day of the month.
        - month: Last contact month of the year.
        - duration: Duration of the last contact.
    - **Other Attributes**:
        - campaign: Number of contacts performed during this campaign.

- pdays: Number of days since the client was last contacted.
- previous: Number of contacts performed before this campaign.
- poutcome: Outcome of the previous marketing campaign.

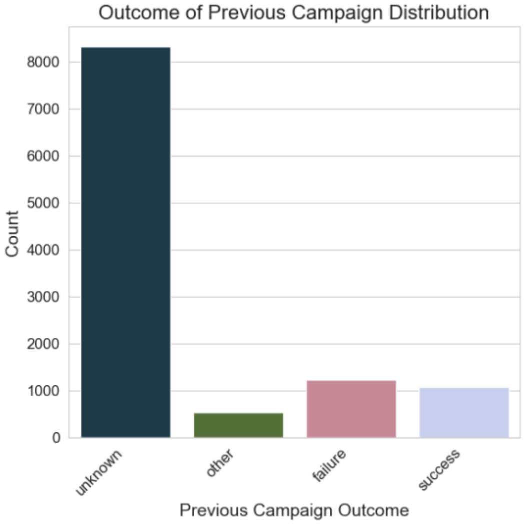
This dataset is well-suited for predictive modeling as it provides a mix of numerical and categorical variables that can be used to predict customer behavior.

DATA VISUALIZATION AND EXPLORATION

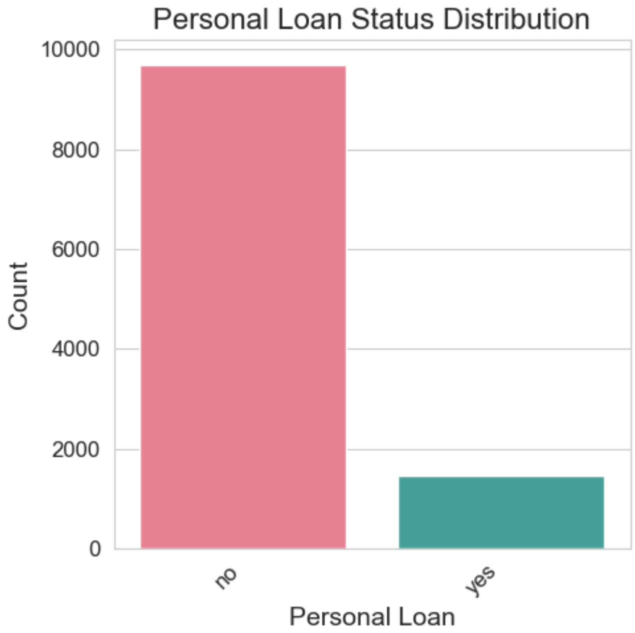
1. Age distribution



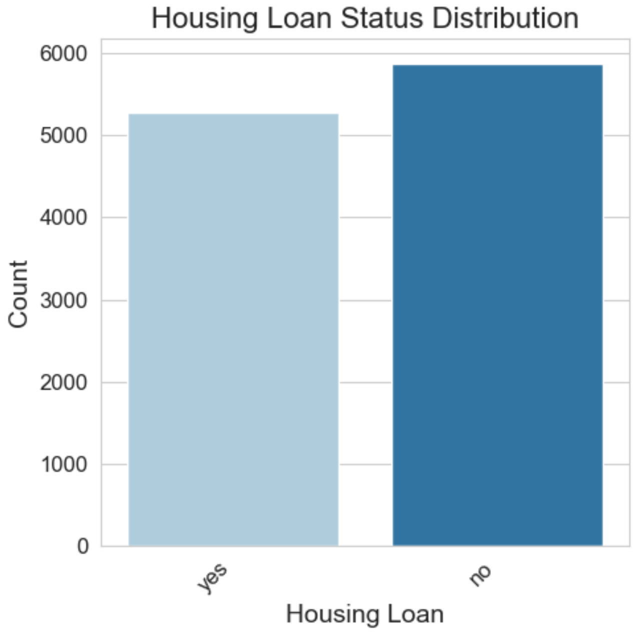
2. Outcome of previous campaign distribution



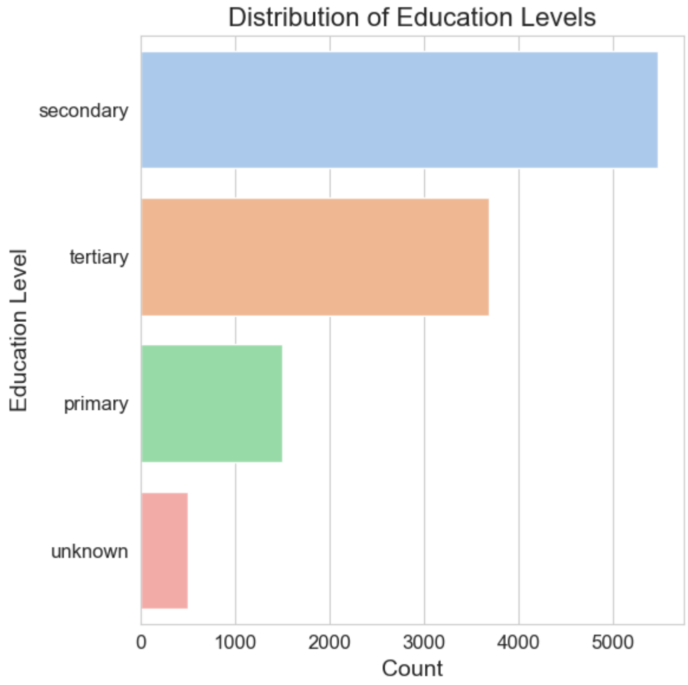
3. Personal loan status distribution



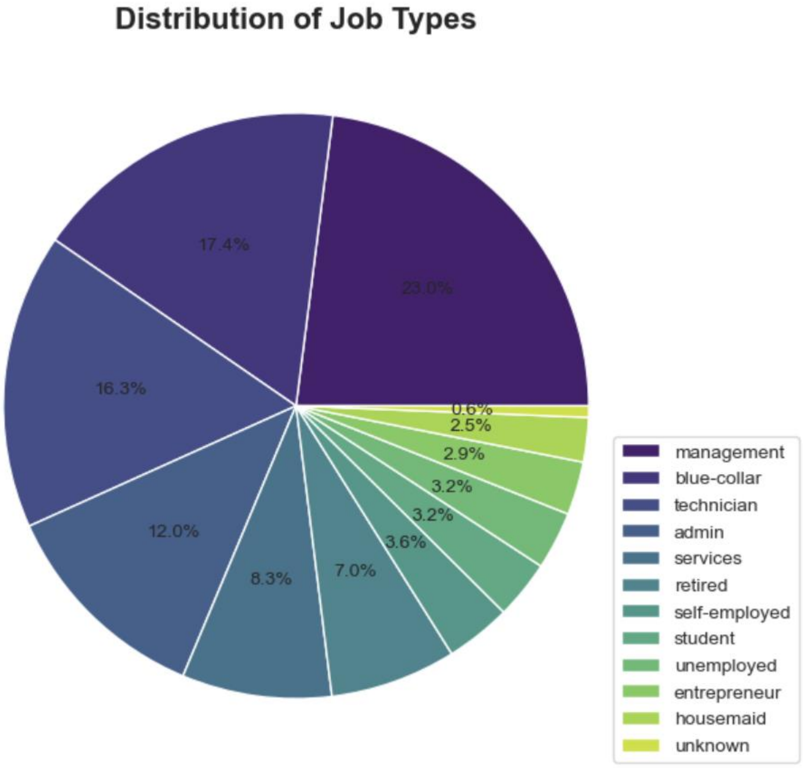
4. Housing loan status distribution



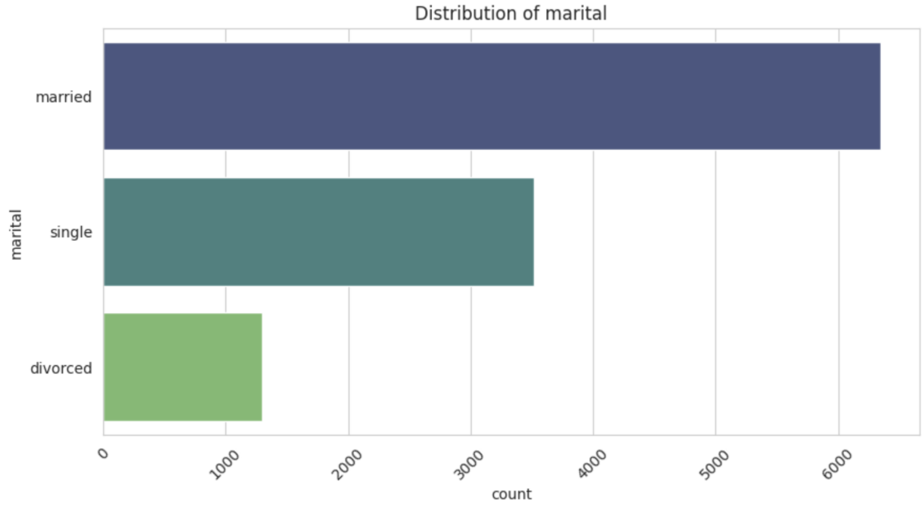
5. Distribution of education levels



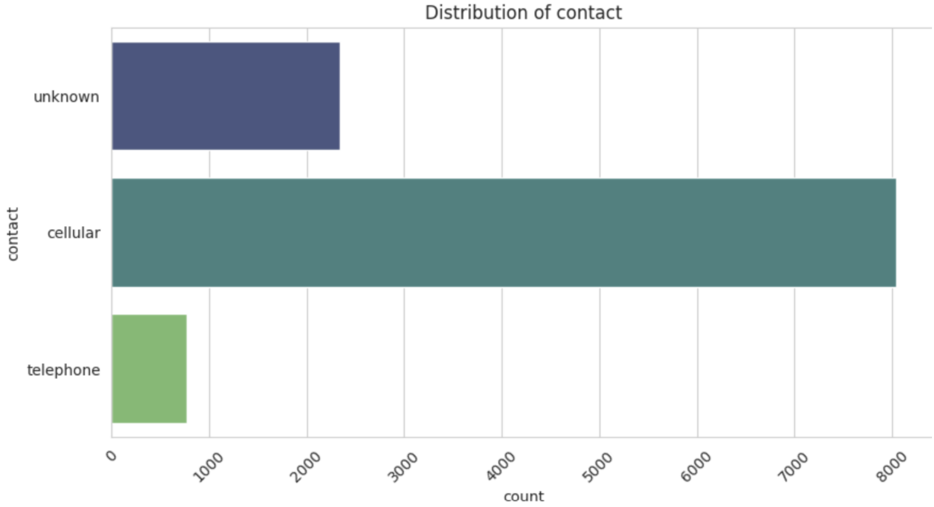
6. Distribution of job types



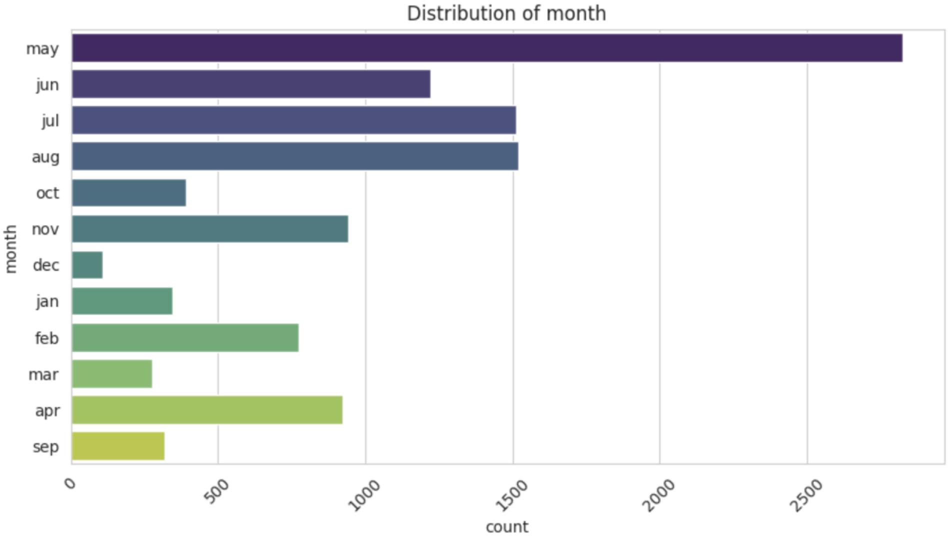
7. Distribution of marital status



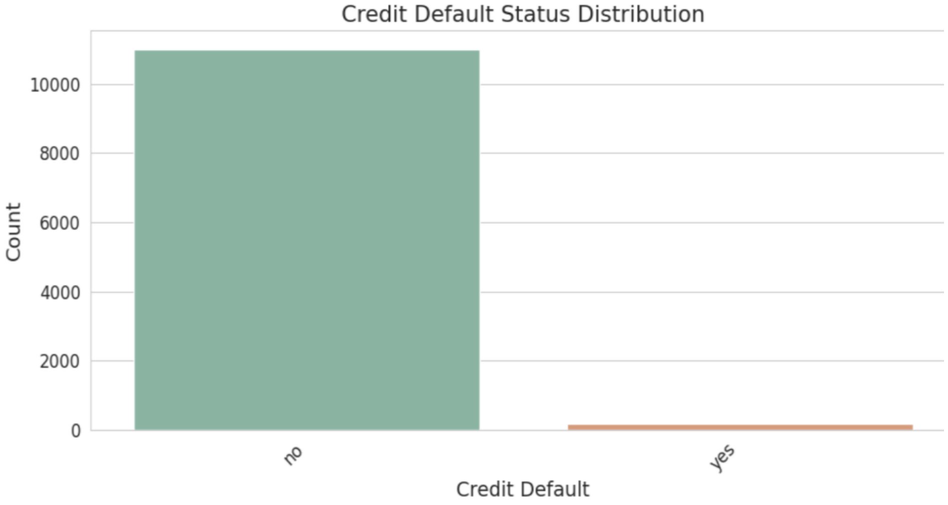
8. Distribution of contact method



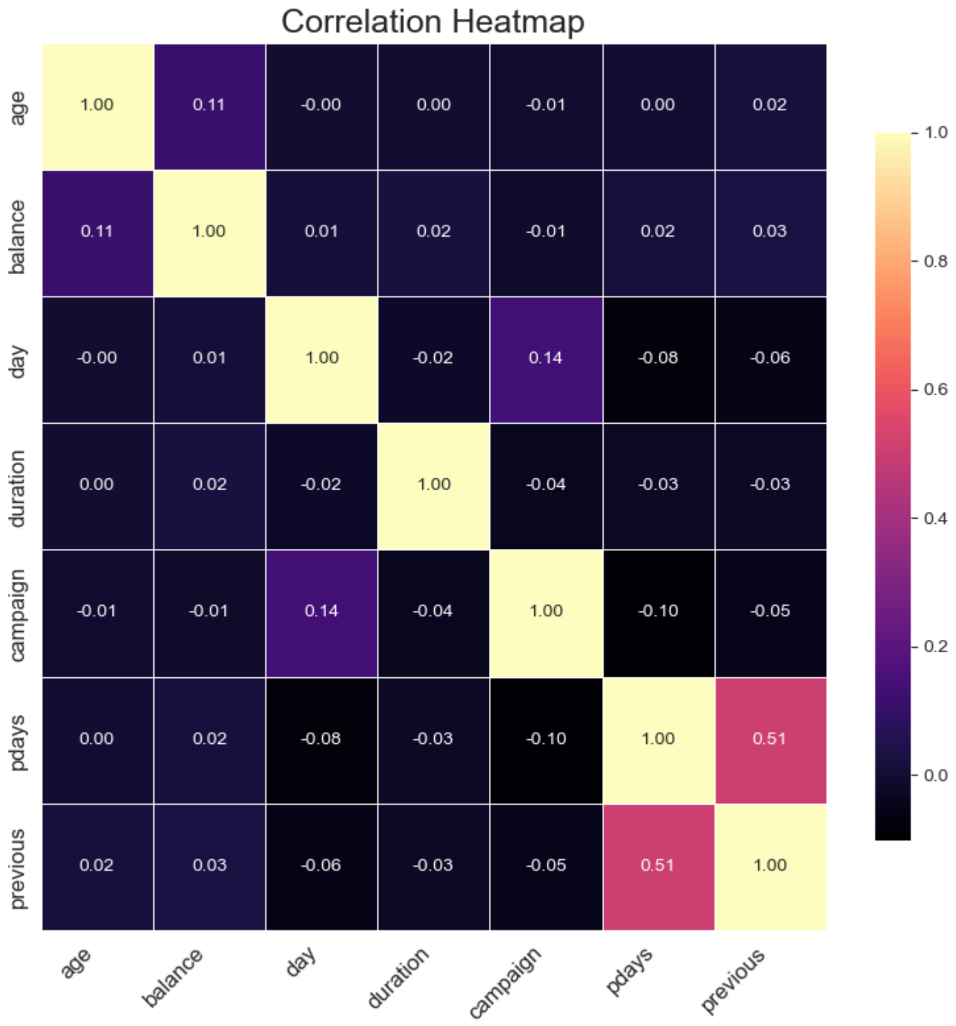
9. Distribution of month



10. Distribution of Credit default



11. Correlation heatmap



DATA PRE-PROCESSING

Preprocessing is a crucial step in preparing the data for model building. The following steps were applied to the dataset:

1. **Handling Missing Values**:
   - Any missing values in the dataset were handled appropriately. For categorical variables, missing values were imputed with the most frequent category or a placeholder category, while numerical variables were imputed with the mean or median value.
2. **Encoding Categorical Variables**:
   - Categorical variables such as job, marital, education, contact, and poutcome were converted into numerical form using one-hot encoding. This process is essential to make these variables usable in machine learning models.
3. **Feature Scaling**:
   - Numerical features like age, balance, and duration were standardized using z-score normalization to ensure that all features contribute equally to the model.
4. **Feature Engineering**:
   - New features were created based on domain knowledge. For example, campaign effectiveness could be derived from the campaign and poutcome columns. The Vector Assembler was then used to combine all features into a single vector, which is a necessary step before feeding the data into the machine learning models.

MODEL BUILDING AND EVALUATION

Four machine learning models were built and evaluated using the preprocessed dataset: **Logistic Regression**, **Decision Tree**, **Random Forest**, and **Gradient Boosting**. Each model was chosen for its specific strengths and was evaluated on a variety of performance metrics.

CLASSIFICATION MODELS: TECHNIQUES & EFFECTIVENESS

LOGISTIC REGRESSION:

- **Technique**: Logistic Regression is a linear model used for binary classification. It estimates the probability that a given instance belongs to a particular class by applying the logistic function to a linear combination of input features. The output is a value between 0 and 1, which is then thresholded to make a binary decision.
- **Effectiveness**: Logistic Regression is often used as a baseline model due to its simplicity and interpretability. It works well when the relationship between the input features and the target variable is approximately linear.

DECISION TREE:

- **Technique**: A Decision Tree is a non-linear model that splits the data into subsets based on feature values, forming a tree-like structure where each internal node represents a decision on a feature, each branch represents an outcome, and each leaf node represents a class label.
- **Effectiveness**: Decision Trees are highly interpretable and can capture complex relationships between features. However, they are prone to overfitting, especially with noisy data.

RANDOM FOREST:

- **Technique**: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (for classification) or mean prediction (for regression) of the individual trees. It reduces overfitting by averaging out the predictions from multiple trees, which are trained on different random subsets of the data.
- **Effectiveness**: Random Forest improves the generalization of Decision Trees by reducing variance, making it more robust to overfitting and effective on complex datasets.

GRADIENT BOOSTING:

- **Technique**: Gradient Boosting is an ensemble technique where models are built sequentially, and each new model attempts to correct the errors made by the previous models. The models are combined to form a stronger predictive model. This method is particularly effective in handling structured data.
- **Effectiveness**: Gradient Boosting is known for its high predictive accuracy and ability to capture complex patterns in data. It is often more effective than other models but may require careful tuning to avoid overfitting.

EVALUATION METRICS

Accuracy : Accuracy is the ratio of correctly predicted instances (both positive and negative) to the total number of instances. It gives a general sense of how often the model is correct.

Formula:

$$\text{Accuracy} = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})}$$

where:

TP = True Positives (correctly predicted positive instances)
TN = True Negatives (correctly predicted negative instances)

FP = False Positives (incorrectly predicted positive instances)
FN = False Negatives (incorrectly predicted negative instances)

Usefulness: Accuracy is useful for providing a quick overview of the model's performance, but it can be misleading if the data is imbalanced (i.e., one class is much more frequent than the other).

Precision : Precision is the ratio of correctly predicted positive instances to the total number of instances predicted as positive. It indicates the accuracy of positive predictions.

Formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Usefulness: Precision is particularly important when the cost of false positives is high, such as in scenarios where incorrect positive predictions can lead to significant consequences.

Recall : Recall (also known as Sensitivity or True Positive Rate) is the ratio of correctly predicted positive instances to the total number of actual positive instances. It measures the model's ability to capture all the positive instances.

Formula:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Usefulness: Recall is critical when the cost of false negatives is high, such as in medical diagnoses, where failing to identify a condition could be very harmful.

F1-Score : The F1-Score is the harmonic mean of precision and recall. It provides a single metric that balances the trade-off between precision and recall, making it particularly useful when you need to consider both false positives and false negatives.

Formula:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Usefulness: The F1-Score is a good metric to use when you want to find a balance between precision and recall, especially in situations where you have an uneven class distribution.

AUROC (Area Under the Receiver Operating Characteristic Curve) : The AUROC is a metric that evaluates the model's ability to distinguish between positive and negative classes across all possible classification thresholds. It plots the true positive rate (recall) against the false positive rate (1 - specificity) at various threshold settings.

Usefulness: The AUROC is a comprehensive metric that reflects the model's ability to discriminate between classes. A value closer to 1 indicates better performance, with 1 representing perfect classification and 0.5 representing random guessing.

EVALUATION RESULTS:

1. Logistic Regression

- **Performance Metrics**:
    - ➢ Accuracy: 0.7941
    - ➢ Precision: 0.8288
    - ➢ Recall: 0.7249
    - ➢ F1-Score: 0.7734
    - ➢ AUROC: 0.8845
- **Interpretation**: Logistic Regression provided a good balance between precision and recall. Its AUROC score indicates a strong ability to distinguish between customers who will subscribe and those who won't.

2. Decision Tree

- **Performance Metrics**:
    - ➢ Accuracy: 0.7714
    - ➢ Precision: 0.7211
    - ➢ Recall: 0.8615
    - ➢ F1-Score: 0.7851
    - ➢ AUROC: 0.8105
- **Interpretation**: The Decision Tree excelled in recall, making it effective at identifying customers likely to subscribe. However, it had a slightly lower precision, which means it predicted more false positives.

3. Random Forest

- **Performance Metrics**:
    - ➢ Accuracy: 0.7965
    - ➢ Precision: 0.8019
    - ➢ Recall: 0.7704
    - ➢ F1-Score: 0.7859
    - ➢ AUROC: 0.8785
- **Interpretation**: The Random Forest model offered a good balance of precision and recall, with slightly better recall compared to Logistic Regression but lower precision.

4. Gradient Boosting

- **Performance Metrics**:
    - ➢ Accuracy: 0.8210
    - ➢ Precision: 0.8150
    - ➢ Recall: 0.8160
    - ➢ F1-Score: 0.8155
    - ➢ AUROC: 0.8955
- **Interpretation**: Gradient Boosting emerged as the best-performing model with the highest overall accuracy, F1-Score, and AUROC. It provided a strong balance between precision and recall, making it highly reliable for predicting customer subscriptions.

MODEL SELECTION AND JUSTIFICATION

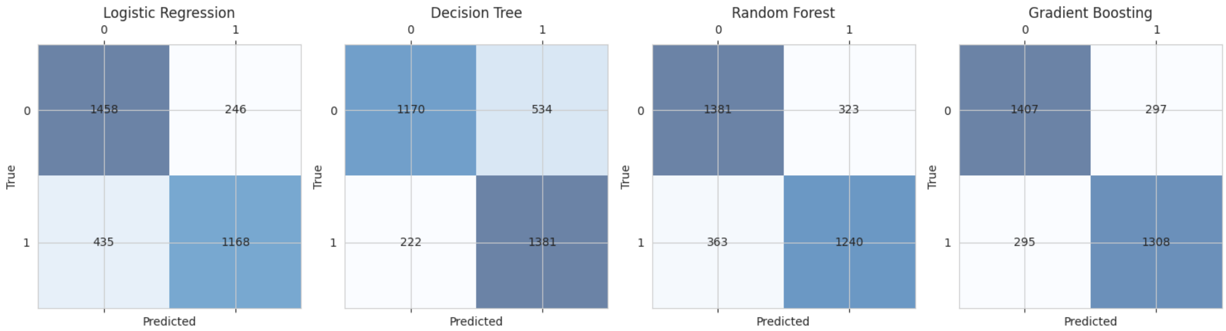
Given the business objective of predicting whether a customer will subscribe to a term deposit, the most appropriate model needs to balance precision and recall effectively. The **Gradient Boosting** model was selected for the following reasons:

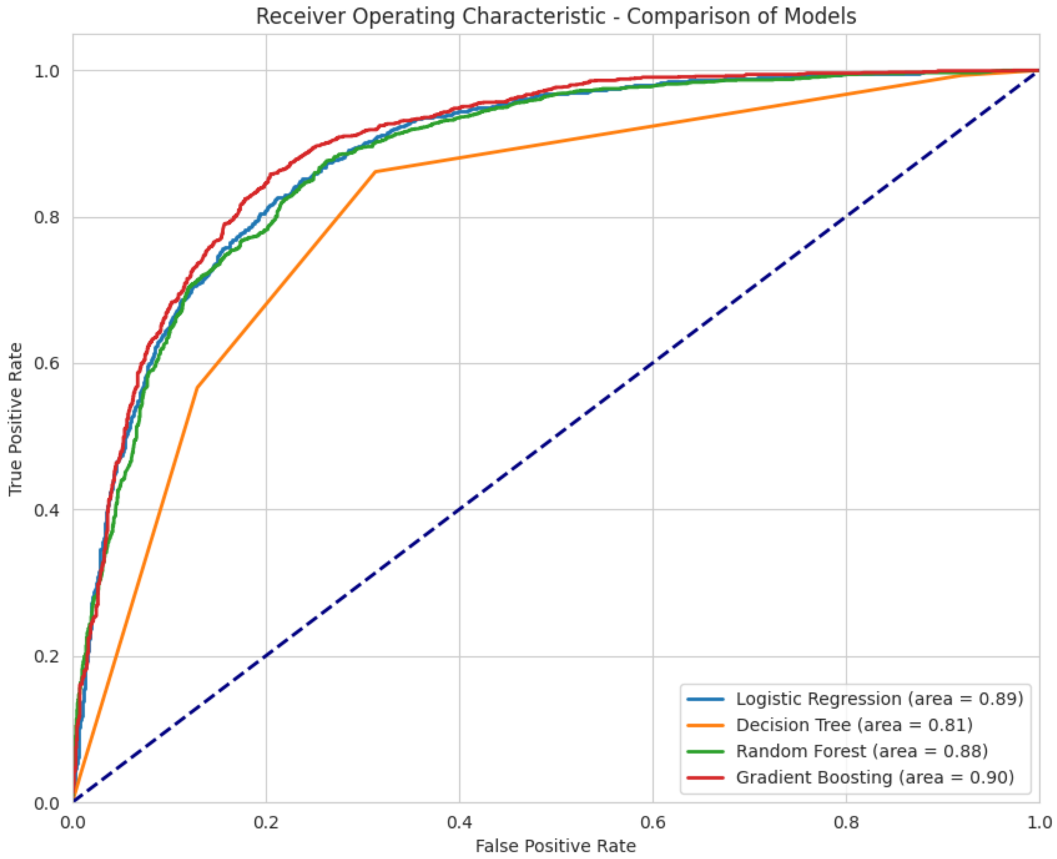
- **High Accuracy**: The Gradient Boosting model achieved the highest accuracy of 82.10%, meaning it correctly classified a higher percentage of the cases compared to the other models.
- **Balanced Precision and Recall**: It offered a balanced precision (81.50%) and recall (81.60%), making it effective at both identifying true positives and minimizing false positives.
- **Superior AUROC**: With an AUROC of 0.8955, Gradient Boosting demonstrated the best discriminative power, indicating it is excellent at distinguishing between customers who will subscribe and those who won't.

This balance is crucial in marketing campaigns where both identifying interested customers and avoiding unnecessary costs on uninterested ones are important.

For further enhancement, the following steps have been undertaken:

- **Hyperparameter Tuning**: Further fine-tuning the hyperparameters of the Gradient Boosting model yielded even better performance.
- **Cross-Validation**: Implemented cross-validation techniques to provide a more robust estimate of the model's performance and to reduce the risk of overfitting.


Logistic Regression
0
1
True
Predicted
1458
246
435
1168
Decision Tree
1170
534
222
1381
Random Forest
1381
323
363
1240
Gradient Boosting
1407
297
295
1308


Receiver Operating Characteristic - Comparison of Models
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (area = 0.89)
Decision Tree (area = 0.81)
Random Forest (area = 0.88)
Gradient Boosting (area = 0.90)

CONCLUSION

The project successfully built and evaluated multiple predictive models using machine learning techniques to predict customer subscription to a term deposit. Among the models tested, **Gradient Boosting** was identified as the most effective model for this task. It provides a strong balance of accuracy, precision, recall, and AUROC, making it the best choice for deployment in real-world banking scenarios.

FUTURE WORK

Feature Importance Analysis: By conducting feature importance analysis using methods like SHAP to identify key predictive features and exploring model ensembling by combining predictions from Gradient Boosting, Random Forest, and Logistic Regression models. Additionally, the model's predicted probabilities could be calibrated using techniques such as isotonic regression or Platt scaling, and advanced cross-validation techniques like nested or stratified K-fold cross-validation could be employed. Addressing class imbalance through methods like SMOTE or adjusting class weights, and optimizing hyperparameters with Bayesian optimization could further enhance the project. For deployment, using Apache Spark MLlib or exporting the model for REST API integration could be considered, along with implementing real-time data processing using Spark Streaming or Apache Kafka. Finally, the project could benefit from enhanced model interpretability with tools like LIME or SHAP and exploring alternative models like XGBoost or LightGBM for potentially better performance.

REFERENCES

- *Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Science & Business Media.
- *Zheng, A., & Casari, A. (2018). Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists.* O'Reilly Media.
- *Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks.* Information Processing & Management, 45(4), 427-437.