SEP 6DA3 Final Project: Data Analytics and Predictive Modeling

INSTRUCTOR: SIQI ZHAO

A PROJECT REPORT

SUBMITTED BY

DARSHAN PARBADIYA

SHRUSTI DESAI

TABLE OF CONTENT

PROJECT BACKGROUND

Nowadays, marketing expenditures in the banking industry are massive, meaning that it is essential for banks to optimize marketing strategies and improve effectiveness. Understanding customers' need leads to more effective marketing plans, smarter product designs and greater customer satisfaction.

Main Objectives: predict customers' responses to future marketing campaigns & increase the effectiveness of the bank's telemarketing campaign

This project will enable the bank to develop a more granular understanding of its customer base, predict customers' response to its telemarketing campaign and establish a target customer profile for future marketing plans.

By analyzing customer features, such as demographics and transaction history, the bank will be able to predict customer saving behaviors and identify which type of customers is more likely to make term deposits. The bank can then focus its marketing efforts on those customers. This will not only allow the bank to secure deposits more effectively but also increase customer satisfaction by reducing undesirable advertisements for certain customers.

PART 2: ABOUT THE DATA

The dataset is about the direct phone call marketing campaigns, which aim to promote term deposits among existing customers, by a Portuguese banking institution from May 2008 to November 2010. It is publicly available at Kaggle.

There are 41,188 observations in the dataset, with no missing values. Each represents an existing customer that the bank reached via phone calls.

- For each observation, the dataset records **21 input variables** that stand for both qualitative and quantitative attributes of the customer, such as age, job, housing and personal loan status, account balance, and the number of contacts.
- There is **a single binary output variable** that denotes "yes" or "no" revealing the outcomes of the marketing phone calls. "Yes" means that a customer subscribed to term deposits.

OVERVIEW OF THE DATASET

The dataset comprises information collected from direct phone call marketing campaigns conducted by a Portuguese banking institution, focusing on customer responses to term deposit

offers. The dataset includes 41,188 observations, with no missing values, providing a robust foundation for analyzing customer behavior and predicting campaign outcomes.

Input Variables

The dataset contains 20 input variables, which are grouped into the following categories:

1. Bank Client Data

The **Bank Client Data** provides key demographic and financial details about each customer, offering a comprehensive understanding of customer profiles that can be used to predict the likelihood of term deposit subscriptions.

- **Age (numeric)**: Represents the age of the client.
- **Job (categorical)**: The occupation of the client, such as 'admin.', 'blue-collar', 'entrepreneur', etc.
- **Marital Status (categorical)**: The marital status of the client, with options like 'divorced', 'married', 'single', and 'unknown'.
- **Education (categorical)**: The educational background of the client, including levels like 'basic.4y', 'high.school', 'university.degree', etc.
- **Default (categorical)**: Indicates if the client has a credit in default ('yes', 'no', 'unknown').
- **Housing (categorical)**: Indicates if the client has a housing loan ('yes', 'no', 'unknown').
- **Loan (categorical)**: Shows whether the client has a personal loan ('yes', 'no', 'unknown').

2. Campaign-Related Attributes

These features provide insights into the client's interaction during the current marketing campaign.

- **Contact (categorical)**: The type of communication used for contact, such as 'cellular' or 'telephone'.
- **Month (categorical)**: The month during which the last contact occurred (e.g., 'jan', 'feb', 'nov').
- **Day of the Week (categorical)**: The day of the week on which the last contact was made (e.g., 'mon', 'tue', 'fri').
- **Duration (numeric)**: The duration of the last contact in seconds. Note: While this variable is highly correlated with the target, it should be excluded in predictive models as it is only known post-call.

3. Previous Campaign Data

These variables capture previous interactions and outcomes from past marketing campaigns.

- **Campaign (numeric)**: The number of contacts made during the current campaign.
- **Pdays (numeric)**: The number of days since the client was last contacted in a previous campaign (999 indicates no prior contact).
- **Previous (numeric)**: The number of contacts the client has had before the current campaign.
- **Poutcome (categorical)**: The outcome of the previous campaign ('failure', 'nonexistent', 'success').

4. Social and Economic Context

These features provide macroeconomic indicators that can influence customer behavior during the campaign period.

- **Emp.var.rate (numeric)**: Employment variation rate (quarterly indicator), reflecting changes in employment levels.
- **Cons.price.idx (numeric)**: Consumer price index (monthly indicator), tracking inflation and price changes.
- **Cons.conf.idx (numeric)**: Consumer confidence index (monthly indicator), showing the public's confidence in the economy.
- **Euribor3m (numeric)**: The 3-month Euribor rate (daily indicator), which represents the European banking interest rate.
- **Nr.employed (numeric)**: The total number of employees (quarterly indicator), reflecting employment trends in the economy.

Output Variable (Target)

- **y (binary)**: The target variable indicating whether the client subscribed to a term deposit ('yes' or 'no').

Insights from the Dataset

This dataset provides a mix of qualitative and quantitative attributes that capture customer profiles, campaign characteristics, and the surrounding economic environment. By leveraging these features, we aim to build a predictive model that can efficiently identify which customers are likely to subscribe to term deposits, enabling the bank to optimize its marketing efforts.

DATA PREPROCESSING AND EXPLORATION

To prepare the dataset for modeling and analysis, several preprocessing steps were implemented to clean, transform, and optimize the data for effective feature evaluation and modeling.

1. Handling Missing and Unknown Values

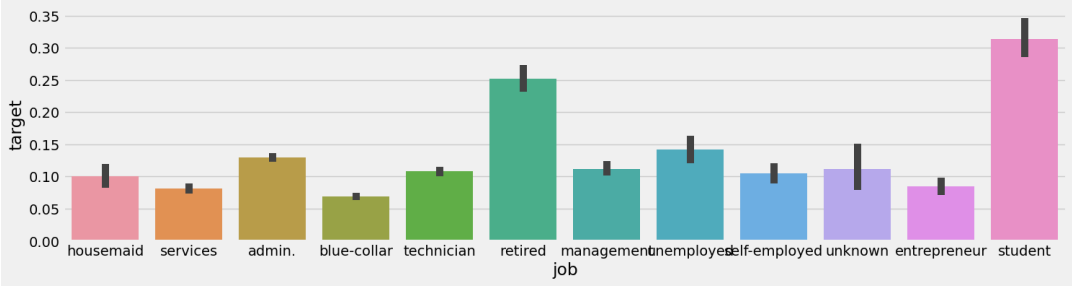
- **Categorical Variables:** Unknown categories (e.g., "unknown" in job, education, default, housing, loan) were either merged into relevant existing categories or handled separately based on their behavior relative to the target variable.
    - Example: unknown in **default** was grouped with "yes" due to its closer behavior.
- **Continuous Variables:** No missing values were observed in the dataset.

2. Feature Grouping and Simplification

To streamline the analysis and enhance interpretability, similar categories were merged where appropriate:

- **Education:**
    - Grouped "basic.4y," "basic.6y," and "basic.9y" into a single **"Basic"** category.
    - Merged "unknown" and "illiterate" into a combined group due to their small sizes.

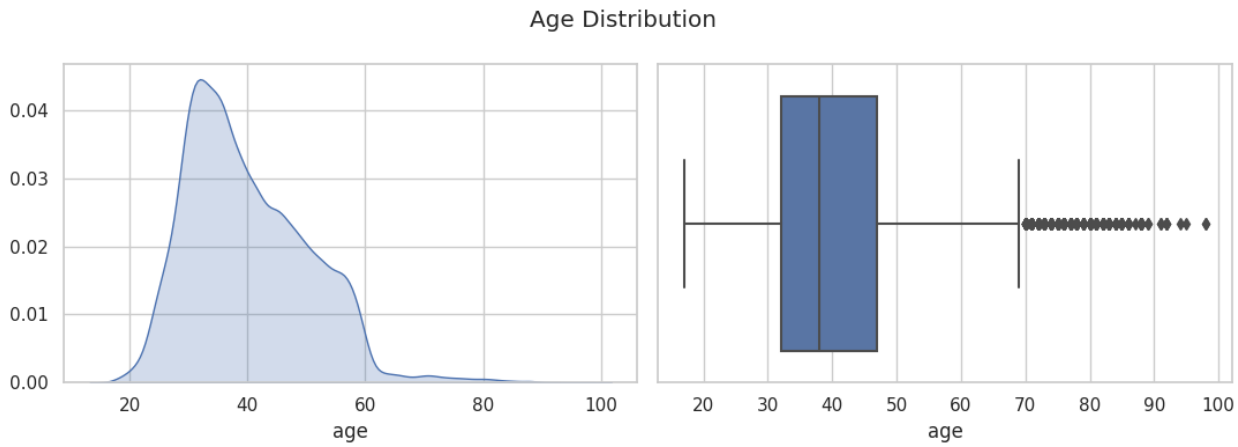
- **Job:**
    - Combined "unknown" with "unemployed" as they showed similar behavior.



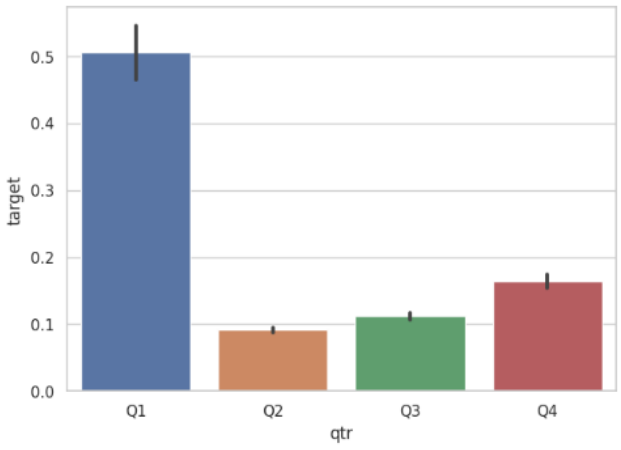
- **poutcome (Previous Outcome):**

- Merged "nonexistent" and "failure" into a single category for simplicity.

- **Age and Duration Binning:**
  - Divided continuous variables like age and duration into **5 bins (quantile ranks)** for easier pattern analysis.



- **Months and Quarters:**
  - Grouped month variable into **quarters (Q1, Q2, Q3, Q4)** to capture seasonal patterns.



- The analysis shows that in Q1, there's a 50% likelihood of customers subscribing to a term loan, indicating a strong response rate. This insight suggests that the marketing team should prioritize efforts early in the year for optimal results. However, in Q2, there's a noticeable decline in customer interest, marking it as a potentially "dry" period for campaigns. This insight can help in adjusting the marketing strategy to focus resources effectively across the year.

3. Encoding Categorical Variables

- Applied **one-hot encoding** for categorical features like contact and month to retain all categories.
- Implemented **label encoding** for ordinal-like features where grouping patterns emerged, such as poutcome and education.

4. Handling Continuous Variables

- **Binning Strategy:** Continuous features like age, duration, pdays, emp.var.rate, and euribor3m were divided into quantile-based ranks to identify trends more effectively.
- **Normalization and Scaling:**
    - Standardization techniques (z-score scaling) were applied to variables like campaign and previous to ensure uniformity in feature ranges.

5. Feature Importance and Evaluation

Post preprocessing, each variable was evaluated for its predictive power against the target variable:

- **Strong Predictors:**
    - duration (clear positive relationship; longer durations lead to higher success rates).
    - contact (cellular contact had significantly higher success rates).
    - poutcome (previous success indicates a much higher likelihood of subscription).
    - euribor3m and emp.var.rate (displayed clear patterns).
- **Moderate Predictors:**
    - job (students and retired individuals showed high success rates).
    - month and quarter (higher subscriptions observed in Q1).
    - age (U-shaped relationship; younger and older groups had slightly higher success rates).
- **Weak Predictors:**
    - day_of_week, marital_status, housing, loan, cons.conf.idx (showed little to no distinction in subscription rates).

Feature Engineering Summary

The following features were derived or optimized for model readiness:

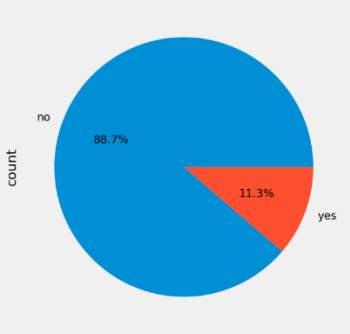
1. **Binned Features:**
    - age, duration, campaign, pdays divided into quantile-based bins for ranking behavior.
2. **Grouped Categories:**
    - Combined similar groups for education, job, poutcome, and month.
3. **Quarterly Analysis:**

- A new quarter feature was derived from the month variable to capture seasonal trends.

4. **Interaction Features:**
    - Explored interactions like age_rank with quarter and contact with quarter to identify combined effects.
5. **Normalized and Scaled Features:**
    - Continuous variables were standardized to ensure consistent input scales for models.
6. **Encoded Categorical Variables:**
    - One-hot encoding and label encoding were applied where appropriate.

BASELINE ANALYSIS: TERM DEPOSITORS AND SUCCESS RATES

The dataset contains information about **41,188 customers**, with the following distribution of term deposit subscription outcomes:



36,548 customers (88.7%) did not subscribe to a term deposit.

4,640 customers (11.3%) subscribed to a term deposit.

This baseline distribution highlights a significant imbalance in the data, with a majority of customers not opting for the product.

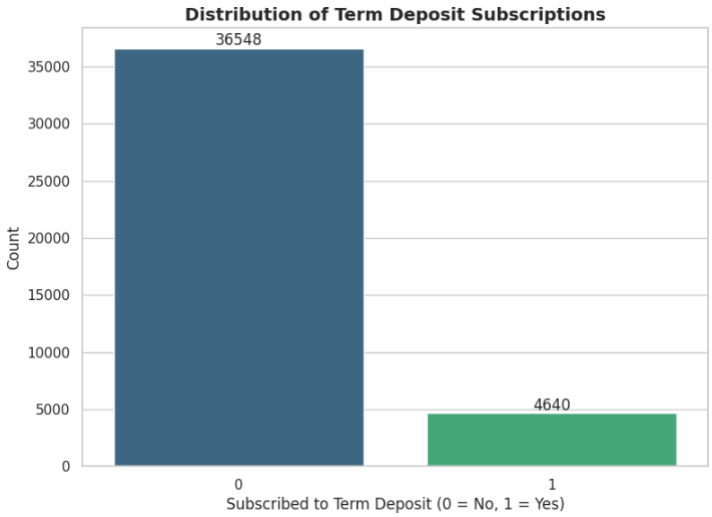
Baseline Likelihood of Term Deposits

To understand the baseline performance in the absence of a machine learning model, we calculated the success rate of term deposits based on historical data. A new binary variable, target, was created to represent the response variable (y) numerically:

- **1** for customers who subscribed to a term deposit (y = 'yes').
- **0** for customers who did not subscribe (y = 'no').

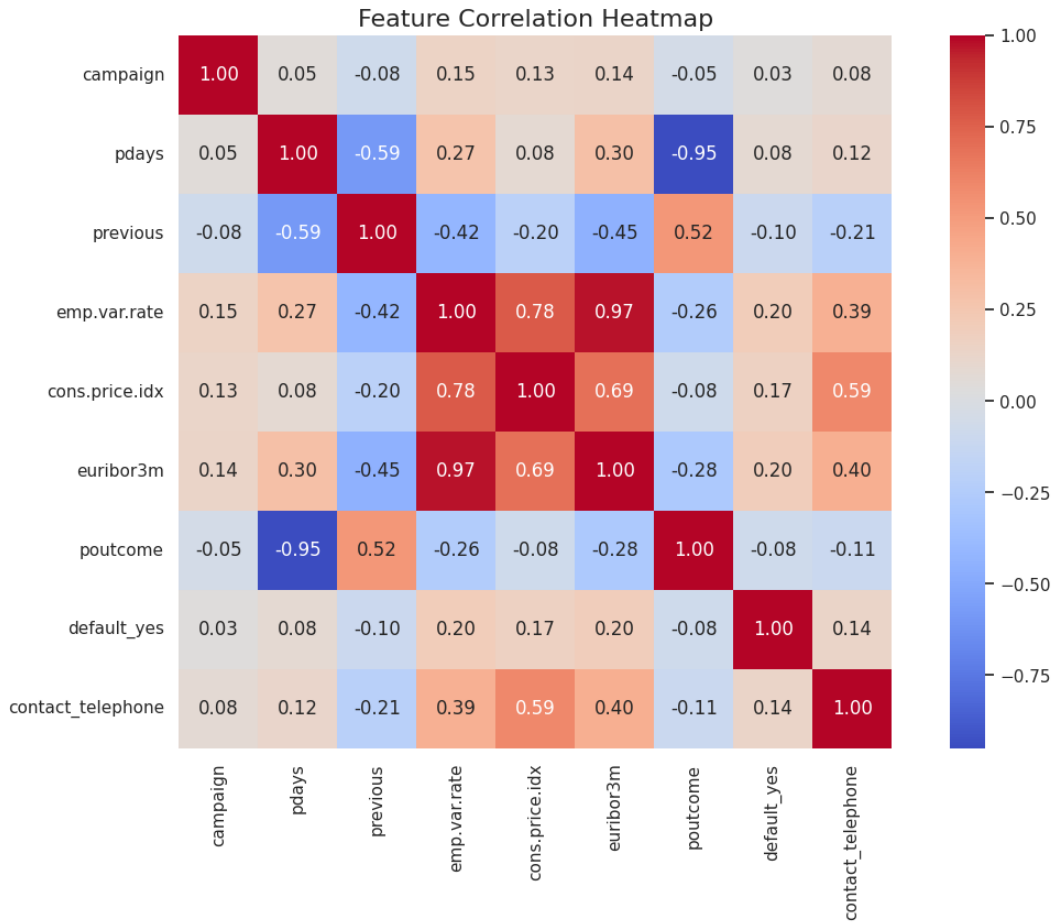
The mean value of the target variable provides the baseline likelihood of a customer subscribing to a term deposit. This calculation revealed a success rate of approximately **11.3%**:

Interpretation:



- Without a predictive model, **11 out of 100 phone calls** are expected to result in a customer subscribing to a term deposit.
- This low success rate highlights the inefficiency of randomly targeting customers and underscores the need for a data-driven approach to optimize telemarketing campaigns.

Heatmap



Key Observations:

1. **Strong Positive Correlations:**
    - **emp.var.rate** and **euribor3m** (0.97): A very high positive correlation suggests that changes in the employment variation rate are closely related to changes in the euribor3m index.
    - **emp.var.rate** and **cons.price.idx** (0.78): Indicates that fluctuations in employment variation rate are moderately positively correlated with the consumer price index.
2. **Strong Negative Correlation:**
    - **pdays** and **poutcome** (-0.95): A very strong negative correlation indicates that if the previous outcome was positive, the number of days since last contact (pdays) tends to be low. This suggests that successful outcomes are tied to more recent contacts.
3. **Moderate Correlations:**
    - **previous** and **poutcome** (0.52): A moderate positive correlation implies that the number of previous contacts is moderately related to the success of the previous outcome.
    - **cons.price.idx** and **euribor3m** (0.69): A moderate positive correlation suggesting that these two economic indicators often move together.
4. **Low or No Correlation:**
    - **default_yes** and other features generally show low correlations, indicating that the default status of the customer doesn't strongly relate to the other features.
    - **contact_telephone** and most other features have low correlations, showing that whether a customer was contacted via telephone doesn't have a strong relationship with other features.

Implications for Model Building:

- Features like **emp.var.rate** and **euribor3m** are likely to be influential in predicting the target variable, as they show a strong correlation with each other.
- The **pdays** and **poutcome** relationship suggests that previous contact outcome and timing are important for understanding customer behavior.
- The low correlation of **contact_telephone** may imply that this feature might not contribute significantly to model performance unless combined with other features.

DATA PREPROCESSING AND SCALING

To prepare the dataset for model training, the data was split into three sets: **training**, **validation**, and **test**. The function load_data_with_scaling was used to perform the following steps:

1. **Data Splitting**:
    - The dataset was divided into a **training set** (75% of the data), a **validation set** (12.5%), and a **test set** (12.5%). This split ensures that the model is trained on one portion of the data, validated on another to tune hyperparameters, and tested on a separate set to evaluate its final performance.
2. **Standardization of Numeric Features**:
    - The **numeric features** of the dataset were standardized to ensure that the models trained on this data perform optimally. This step helps avoid any single feature dominating others due to differing scales.
    - Standardization was applied to all numeric columns unless a specific list of columns was provided. The transformation was performed using StandardScaler from sklearn, which scales the data to have a mean of 0 and a standard deviation of 1.
3. **Handling of Categorical and Numerical Data**:
    - Categorical variables were encoded separately before applying this function, allowing for proper handling of different data types. The numerical features, including campaign, pdays, previous, and others, were subjected to scaling to ensure the model training process benefits from normalized data.

By applying these preprocessing steps, the dataset was properly prepared for training and validation, helping to improve the efficiency and effectiveness of the machine learning models.

Resampling with SMOTETomek

To address the class imbalance in the dataset, the **SMOTETomek** technique was applied. SMOTETomek combines two methods: **SMOTE (Synthetic Minority Over-sampling Technique)** and **Tomek Links**. SMOTE is used to create synthetic samples for the minority class, while Tomek Links remove instances from the majority class that are close to the decision boundary, helping to clean up the data and reduce noise.

Before SMOTETomek:

Prior to applying SMOTETomek, the dataset exhibited an imbalance in the target variable, with a significantly larger number of instances belonging to the majority class (target = 0). The distribution of the target classes was as follows:

- **Class 0 (No subscription):** 21,923 instances
- **Class 1 (Subscribed):** 2,789 instances

This imbalance could lead to model bias, where the model might predict the majority class more frequently, neglecting the minority class, and making it challenging to capture all positive cases.

After SMOTETomek:

After applying the **SMOTETomek** technique, the target class distribution became balanced, with both the majority and minority classes having an equal number of instances:

- **Class 0 (No subscription):** 21,767 instances
- **Class 1 (Subscribed):** 21,767 instances

By creating synthetic samples for the minority class and removing noisy majority class instances, the dataset now has an equal number of samples for each class, reducing the bias toward the majority class and improving the model's ability to correctly identify positive cases.

This balancing process through SMOTETomek can enhance the performance of models, especially in terms of recall, by ensuring that both classes are treated equally during the training phase. However, it is important to evaluate how the resampling affects other metrics, such as precision and accuracy, to ensure a well-rounded model performance.

EVALUATION AND METRICS

Supervised Learning:

In this project, we evaluated multiple supervised learning models to assess their ability to predict customer responses to telemarketing campaigns. The models considered were Logistic Regression, Decision Tree, Random Forest, AdaBoost, and Support Vector Machine (SVM). The following evaluation metrics were used to compare model performance:

- **Accuracy**: Measures the overall proportion of correct predictions.
- **F1-Score**: The harmonic mean of precision and recall, especially important when dealing with imbalanced datasets.
- **Precision**: The ratio of true positives to the sum of true and false positives, indicating the accuracy of positive predictions.
- **Recall**: The ratio of true positives to the sum of true positives and false negatives, indicating the model's ability to capture positive cases.
- **ROC-AUC**: The area under the Receiver Operating Characteristic curve, summarizing the model's ability to distinguish between classes.

Cross-Validation (5-Fold) Overview

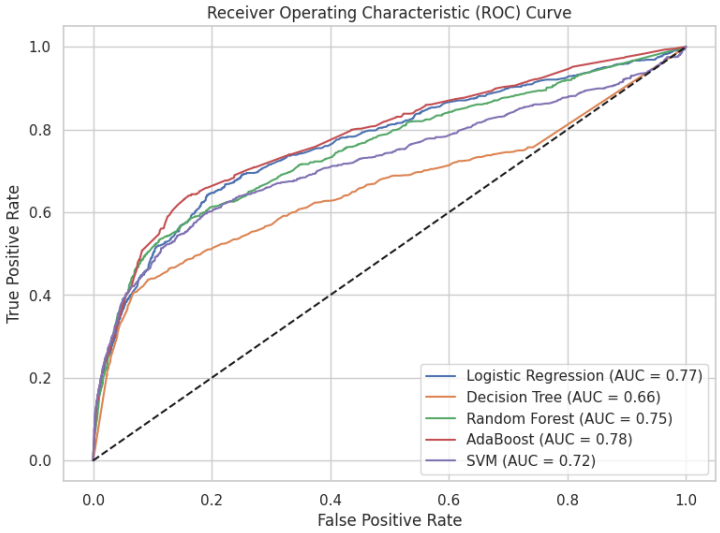
In order to ensure the robustness and generalizability of the models, 5-fold cross-validation was employed during the training and hyperparameter tuning process. This technique involves splitting the dataset into five equal subsets, using four for training and the remaining one for validation, repeated five times with each subset serving as the validation set once. The performance metrics were averaged over the five folds to provide a more reliable estimate of model performance. The cross-validation accuracy reported for each model reflects this averaged performance, helping to mitigate the risk of overfitting and ensuring that the models generalize well to unseen data. This approach was particularly important given the class imbalance in the dataset, as it provided a more comprehensive evaluation across different subsets of the data.

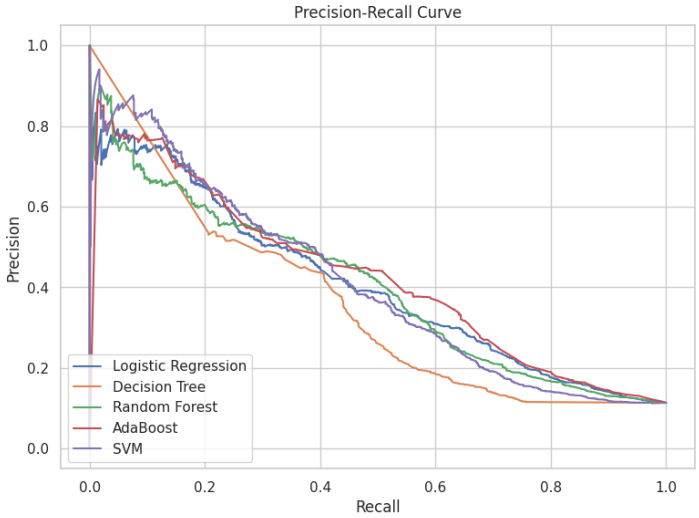
Model Evaluation Metrics: Without Hyperparameter Tuning

The following table presents the performance of multiple supervised learning models evaluated on the dataset without hyperparameter tuning. These models were assessed using key evaluation metrics: Accuracy, F1 Score, Precision, Recall, and ROC AUC.

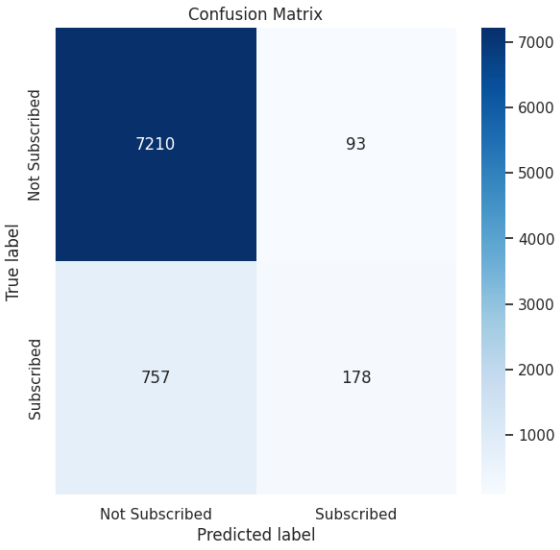
Model Evaluation Metrics:

	Accuracy	F1 Score	Precision	Recall	ROC AUC
Logistic Regression	0.896820	0.295191	0.656827	0.190374	0.767193
Decision Tree	0.888444	0.338373	0.517621	0.251337	0.659524
Random Forest	0.892207	0.368421	0.549894	0.277005	0.753084
AdaBoost	0.896334	0.317891	0.627760	0.212834	0.782388
SVM	0.897427	0.333070	0.635542	0.225668	0.723587

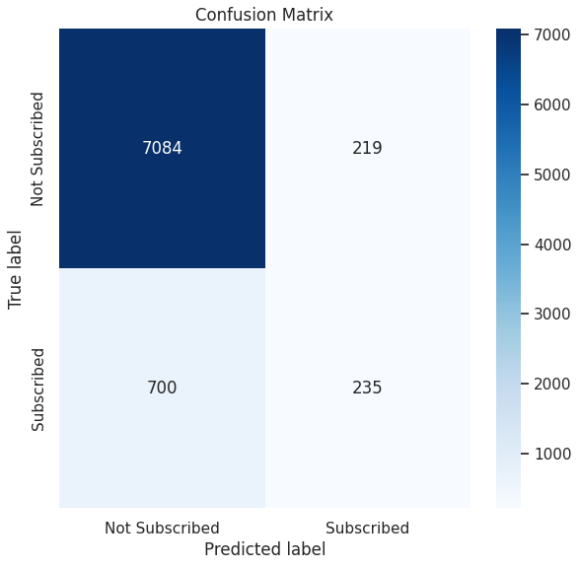




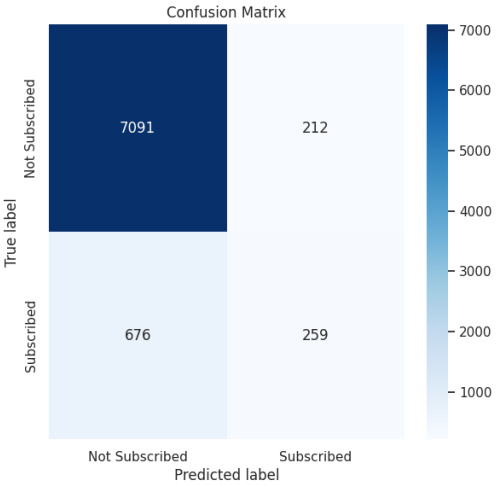
- **Logistic Regression** achieved the highest accuracy (89.68%) but demonstrated a relatively low F1 score (0.2952) and recall (0.1904), indicating that while it made accurate predictions, it struggled with identifying all positive instances.
- **Decision Tree** showed a decrease in performance compared to Logistic Regression, with a lower accuracy (88.84%) and F1 score (0.3384). It demonstrated a slightly better recall (0.2513) but had a relatively low precision (0.5176).
- **Random Forest**, an ensemble model, exhibited a good balance between accuracy (89.22%) and recall (0.2770), with an F1 score of 0.3684, suggesting its robustness in capturing positive cases while maintaining a higher precision (0.5499).
- **AdaBoost** performed similarly to Logistic Regression, with an accuracy of 89.63% but slightly lower precision (0.6278) and recall (0.2128). It achieved the highest ROC AUC score (0.7824), indicating its ability to differentiate between classes effectively.
- **Support Vector Machine (SVM)** displayed the lowest accuracy (89.74%) and F1 score (0.3331), but it performed well in recall (0.2257), showing it could identify some positive cases.



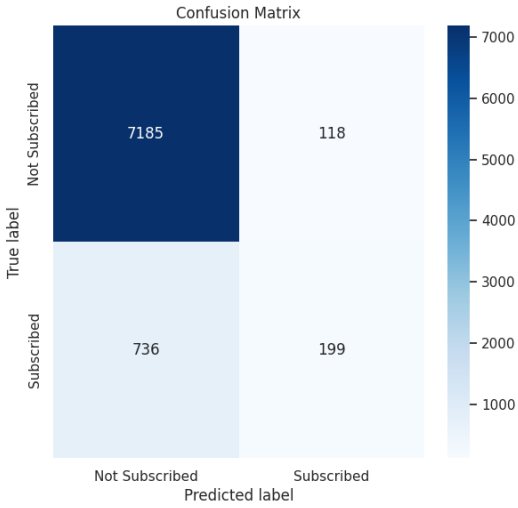
Logistic Regression



Decision Tree



Random Forest



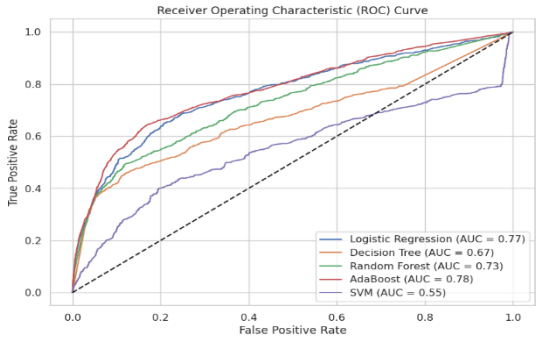
AdaBoost

Model Evaluation Metrics: After Resampling and Training

The table below summarizes the performance of the same models after resampling techniques were applied to handle any imbalances in the dataset. Resampling aims to improve the model's ability to detect the minority class, especially for imbalanced datasets.

Model Evaluation Metrics:

	Accuracy	F1 Score	Precision	Recall	ROC AUC
Logistic Regression	0.746905	0.379280	0.262789	0.681283	0.767360
Decision Tree	0.843166	0.378248	0.343832	0.420321	0.669633
Random Forest	0.844501	0.406120	0.358429	0.468449	0.729213
AdaBoost	0.823379	0.442315	0.344683	0.617112	0.779965
SVM	0.265113	0.182996	0.104710	0.725134	0.550228



- **Logistic Regression**, after resampling, showed a significant drop in accuracy (74.69%) but improved recall (0.6813), indicating that resampling helped the model capture more positive cases at the cost of precision (0.2628).
- **Decision Tree** exhibited an improvement in recall (0.4203) and precision (0.3438), with accuracy rising to 84.32%. However, the ROC AUC score decreased, suggesting it struggled to maintain class separation despite improved recall.
- **Random Forest** performed better in terms of recall (0.4684) and F1 score (0.4061), maintaining its accuracy (84.45%) and improving its ability to identify positive cases.
- **AdaBoost** saw a noticeable increase in recall (0.6171) and F1 score (0.4423) after resampling, but its accuracy (82.34%) remained lower than other models. The model continued to achieve the highest ROC AUC score (0.7800), suggesting it is the best at distinguishing between classes.
- **SVM** suffered a significant drop in performance, with an accuracy of just 26.51% and a drastically low F1 score (0.1830), indicating that resampling may have negatively impacted its ability to predict positive cases.

Conclusion

The models showed a clear difference in performance before and after applying resampling techniques. Resampling helped models like Logistic Regression, Random Forest, and AdaBoost improve their recall but at the cost of precision and overall accuracy. AdaBoost still outperformed others in ROC AUC, while SVM's performance deteriorated significantly after resampling. These

insights suggest that while resampling can enhance recall, careful consideration is needed to balance other evaluation metrics such as precision and accuracy for overall model performance.

MODEL TRAINING, HYPERPARAMETER TUNING, AND EVALUATION

To optimize telemarketing prediction accuracy, **Logistic Regression**, **Decision Tree**, and **Random Forest** classifiers were trained and evaluated. **Grid Search with Cross-Validation** was employed for hyperparameter tuning. Results were analyzed both **before and after resampling** using SMOTETomek, a technique for handling class imbalance.

1. Model Setup and Grid Search Parameters

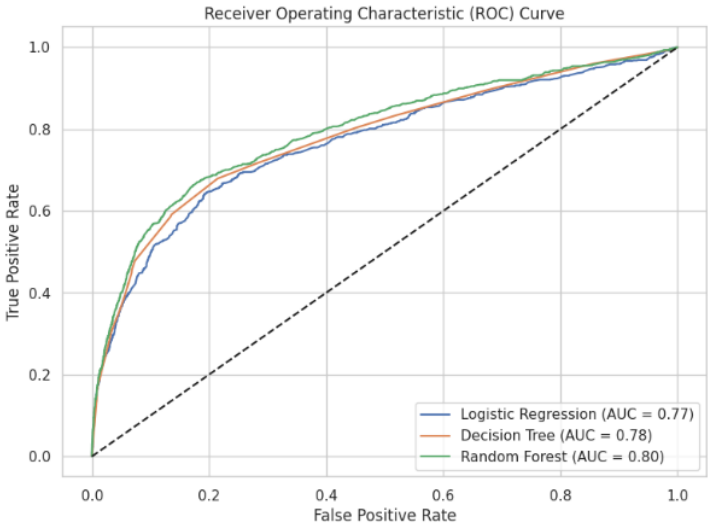
Model	Hyperparameters Tuned
Logistic Regression	- C: [0.01, 0.1, 1, 10] (Regularization strength) - max_iter: [100, 200, 300] - penalty: ['l1', 'l2']
Decision Tree	- max_depth: [5, 10, 20, None] - min_samples_split: [2, 5, 10] - min_samples_leaf: [1, 2, 4] - criterion: ['gini', 'entropy']
Random Forest	- n_estimators: [100, 200] (Number of trees) - max_depth: [5, 10, 20, None] - min_samples_split: [2, 5, 10] - min_samples_leaf: [1, 2, 4] - max_features: ['sqrt', 'log2']

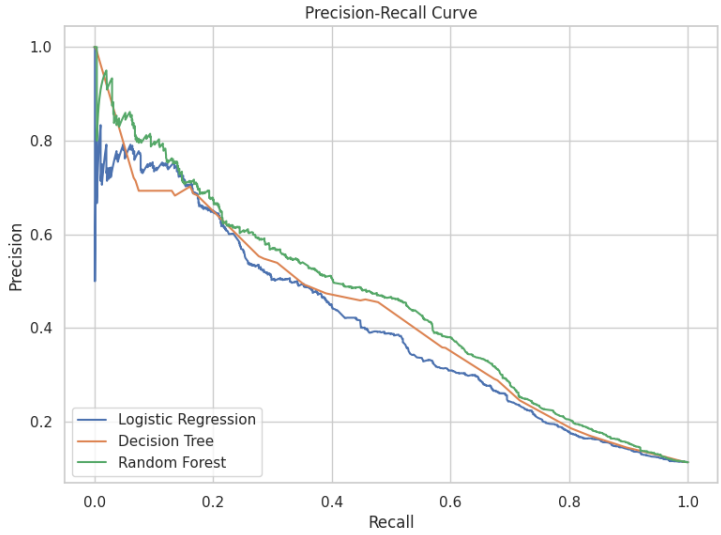
2. Model Performance Before Resampling

Evaluation Metrics

- Accuracy
- F1 Score
- Precision
- Recall
- ROC AUC

- Cross-Validation Accuracy





Results Before Resampling

Metric	Logistic Regression	Decision Tree	Random Forest
Accuracy	0.8968	0.8971	**0.8966**
F1 Score	0.2952	0.2765	**0.3436**
Precision	**0.6568**	0.6835	0.6143
Recall	0.1904	0.1733	**0.2385**
ROC AUC	0.7672	0.7802	**0.7953**
CV Accuracy	0.8990	0.8988	**0.9020**

Best Hyperparameters Before Resampling

- **Logistic Regression**: {'C': 1, 'max_iter': 100, 'penalty': 'l1'}
- **Decision Tree**: {'criterion': 'entropy', 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2}
- **Random Forest**: {'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 100}

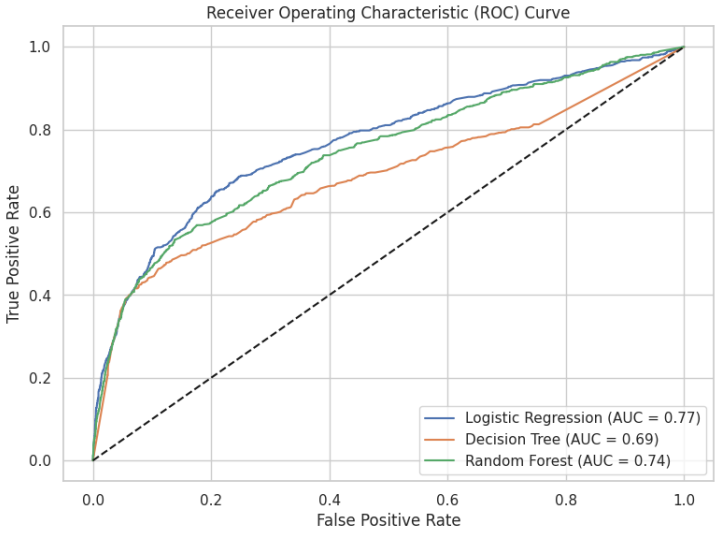
3. Model Performance After Resampling (SMOTETomek)

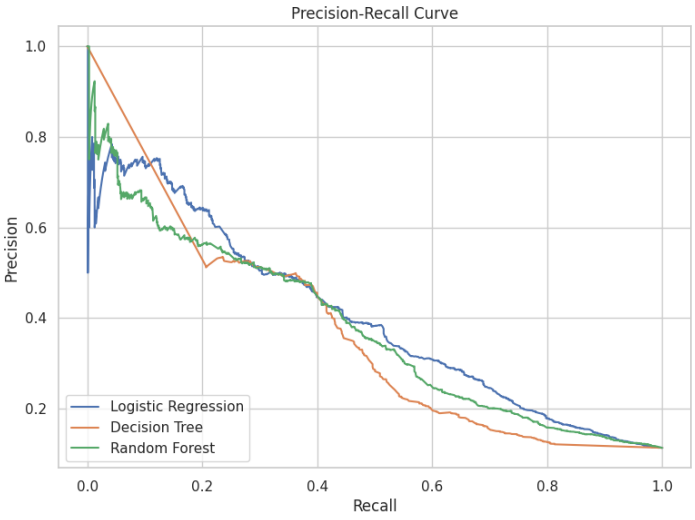
Impact of Resampling

SMOTETomek was applied to balance the dataset by:

- Generating synthetic samples for the minority class.
- Removing noisy samples from the majority class.

Results After Resampling





Metric	Logistic Regression	Decision Tree	Random Forest
Accuracy	0.7473	**0.8455**	0.8427
F1 Score	0.3800	0.3958	**0.4109**
Precision	0.2633	**0.3558**	0.3573
Recall	**0.6824**	0.4460	0.4834
ROC AUC	0.7673	0.6851	**0.7448**
CV Accuracy	0.7243	**0.8404**	0.8423

Best Hyperparameters After Resampling

- **Logistic Regression**: {'C': 10, 'max_iter': 100, 'penalty': 'l2'}
- **Decision Tree**: {'criterion': 'gini', 'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2}

- **Random Forest**: {'max_depth': 20, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200}

4. Model Comparison: Key Observations

Aspect	Before Resampling	After Resampling
Top Performer	Random Forest (Best F1, ROC AUC)	Decision Tree (Best Accuracy)
Class Imbalance	Models favored the majority class	Improved recall for minority class
Random Forest	Strong overall performance	Balanced F1 and Recall
Logistic Regression	High precision but low recall	Best recall, lower precision

Insights

1. **Random Forest** performed consistently well before and after resampling, showing its robustness.
2. **Decision Tree** excelled after resampling, demonstrating its ability to handle balanced datasets effectively.
3. **Logistic Regression** achieved the highest recall after resampling, making it ideal for applications prioritizing positive class detection.

5. Conclusion

- **Before Resampling**: Random Forest delivered the best overall performance.
- **After Resampling**: Resampling significantly improved minority class detection, with Decision Tree and Random Forest showing the highest gains in accuracy and F1 scores.
- **Recommendation**: Random Forest remains the most balanced and reliable model for this problem. However, when recall is critical, Logistic Regression should be considered.

Final Note: Hyperparameter tuning and resampling are essential steps to improve predictive performance, particularly in imbalanced datasets like telemarketing response prediction.

Bias-Variance Trade-Off in Model Evaluation

The **bias-variance trade-off** is crucial in understanding model performance, especially when evaluating supervised learning models for telemarketing prediction. In this analysis, models were assessed before and after resampling to handle dataset imbalance.

1. **Before Resampling**:
    - **Logistic Regression**: Achieved the highest accuracy (89.68%) but struggled with recall (0.1904), showing **high bias** (underfitting) and low variance. While accurate, it missed many positive cases.
    - **Decision Tree**: Showed lower accuracy (88.84%) and F1 score (0.3384), but slightly better recall (0.2513). The model exhibited **moderate bias** and **higher variance**, as it adjusted more to the data's noise.
    - **Random Forest**: With good accuracy (89.22%) and recall (0.2770), it balanced bias and variance well, showing robustness and **lower bias** with **moderate variance**, leading to better handling of both classes.
    - **AdaBoost**: Performed similarly to Logistic Regression in accuracy (89.63%) but had a lower recall (0.2128). The model had **higher bias** and **lower variance**, indicating it was less flexible in capturing the minority class.
    - **SVM**: Had the lowest accuracy and F1 score, reflecting **high bias** and **low variance**, showing a failure to capture the complexity of the data.
2. **After Resampling**:
    - **Logistic Regression**: Improved recall (0.6813) but saw a significant drop in accuracy (74.69%), indicating an increase in **bias** (less accurate predictions) and a reduction in **variance** (improved consistency in predicting the minority class).
    - **Decision Tree**: Improved recall (0.4203) and precision (0.3438), with a drop in ROC AUC. It showed **reduced variance**, balancing **bias** with better recall.
    - **Random Forest**: Achieved better recall (0.4684) and F1 score (0.4061), maintaining a balance of **bias** and **variance** while improving its ability to detect the minority class.
    - **AdaBoost**: Showed a significant increase in recall (0.6171), at the cost of precision (0.2628). This reflects **higher bias** with **moderate variance** but continued good class differentiation.
    - **SVM**: Suffered a drastic performance drop, with accuracy plummeting to 26.51%, indicating that **resampling worsened the bias-variance balance**, making the model less reliable.

In conclusion, resampling improved recall at the cost of precision and accuracy, especially in models like Logistic Regression and AdaBoost. While Random Forest and Decision Tree adapted well to the imbalance, Logistic Regression and AdaBoost showed improvements in detecting positive instances, with trade-offs in other evaluation metrics. The analysis emphasizes the importance of balancing **bias** and **variance** in model evaluation, particularly when dealing with imbalanced datasets like telemarketing response prediction.

BUSINESS INSIGHTS

The analysis of the telemarketing campaign highlights several critical features and strategies that can significantly improve campaign outcomes, resource allocation, and customer engagement. Below are the key insights derived from the data, which are actionable and impactful for enhancing marketing efforts:

1. **Targeted Campaigns Based on Key Features**:
    - **Focus on High-Value Customers**: Customers who have previously subscribed to term deposits and those with higher account balances show a higher likelihood of subscribing again. These segments should be prioritized in future campaigns to maximize returns.
        - **Actionable Insight**: Develop a strategy to target these customers with tailored offers or loyalty rewards, reinforcing the value of a term deposit and encouraging repeat subscriptions.
    - **Optimize Call Duration**: The duration of the call is positively correlated with successful outcomes. Longer, more meaningful conversations lead to higher conversion rates.
        - **Actionable Insight**: Train telemarketers to engage customers more effectively, ensuring that calls are not rushed and provide clear value to the customer, increasing the likelihood of subscription.
2. **Optimize Contact Frequency**:
    - **Limit Outreach Attempts**: Predictive analytics can help identify the optimal frequency of contact, ensuring that customers are not overwhelmed by excessive outreach.
        - **Actionable Insight**: Use data-driven insights to tailor contact strategies by customer segment, ensuring each group receives the right number of touchpoints without saturation. For high-potential groups, more frequent outreach may be appropriate, while lower-potential groups should be contacted less often.

3. **Resource Allocation Optimization**:
    - **Target Middle-Aged, Financially Stable Segments**: Customers aged 30-50 with higher account balances and stable employment have the highest likelihood of responding positively to telemarketing efforts.
        - **Actionable Insight**: Allocate more resources to outreach efforts targeting middle-aged individuals with financial stability. Tailor messaging to emphasize the long-term security and growth potential of term deposits.
    - **Tailored Solutions for Housing Loan Customers**: Customers with housing loans may have specific financial constraints, so targeted solutions should be developed to address their needs.
        - **Actionable Insight**: Create custom financial products or incentives for customers with housing loans, ensuring that their specific financial situation is taken into account when marketing term deposits.
4. **Dynamic Campaign Monitoring and Adjustment**:
    - **Real-Time Predictions with Random Forest**: Using machine learning models like Random Forest, predictions on customer behavior can be made in real-time, allowing for dynamic adjustments to marketing strategies.
        - **Actionable Insight**: Implement real-time tracking of key metrics like conversion rates and customer responses. Adjust the campaign strategy on-the-fly based on the model's predictions to continuously improve the campaign's effectiveness.
5. **Key Insights from Categorical and Continuous Variables**:
    - **Education**: Higher education correlates with a higher likelihood of subscription. University graduates and individuals with professional courses show a higher subscription rate.
        - **Actionable Insight**: Prioritize outreach to educated demographics, highlighting the financial benefits of term deposits as a tool for better financial planning.
    - **Job Type**: Target customers in "student" and "retired" categories, as they show higher success rates (31.4% and 25.2%, respectively).
        - **Actionable Insight**: Tailor campaigns to address the specific needs of students and retirees, emphasizing security, stability, and long-term benefits.

- **Contact Method**: Customers contacted via cellular phones show a higher likelihood of subscribing compared to those contacted via landline.
    - **Actionable Insight**: Prioritize mobile phone outreach, ensuring that calls are made to customers using their preferred contact method, which will enhance conversion rates.
- **Month and Quarter**: The first quarter (Q1) exhibits the highest likelihood of subscription, making it the optimal time for campaigns.
    - **Actionable Insight**: Schedule campaigns during Q1 to capitalize on the higher conversion rate, while adjusting strategies for lower-conversion months (e.g., Q2), focusing more on mobile outreach.
- **Previous Outcome (Poutcome)**: Customers who have previously subscribed to term deposits are more likely to subscribe again.
    - **Actionable Insight**: Retarget previous subscribers with follow-up campaigns offering loyalty rewards or incentives to drive repeat business.

LIMITATIONS AND IMPROVEMENTS

Limitations:

1. **Data Bias**:
    - The dataset only includes customers from a specific bank in Portugal, potentially limiting generalizability to other regions or industries.
    - The reliance on historical data may not account for changing economic or behavioral factors.
2. **Model Constraints**:
    - While Random Forest performed well, it is computationally expensive, which may hinder real-time deployment for large-scale campaigns.
    - Logistic Regression, while interpretable, struggled with precision, which could lead to false positives.
3. **Feature Limitations**:
    - Certain variables like “call duration” may not be practically usable for campaign planning, as duration is only known during the call. Pre-call prediction models must rely on other features.
4. **Class Imbalance**:

- Although SMOTETomek improved recall, synthetic sampling techniques may introduce noise into the data, reducing model reliability in real scenarios.

Improvements:

1. **Incorporate Additional Data**:
    - Integrate external features such as customer demographics, income, and spending patterns to enhance model predictions.
    - Include temporal data to analyze seasonal or monthly trends in customer behavior.
2. **Advanced Techniques**:
    - Explore ensemble methods like **XGBoost** or **LightGBM** to achieve further performance gains and reduce computation time.
    - Apply **stacked models** combining Random Forest and Logistic Regression to balance recall and precision.
3. **Dynamic Thresholding**:
    - Adjust model prediction thresholds to optimize trade-offs between precision and recall, depending on campaign goals (e.g., minimizing false positives vs. maximizing outreach).
4. **A/B Testing**:
    - Conduct real-world A/B testing of the model's predictions to validate effectiveness and refine strategies based on actual responses.
5. **Customer Feedback Loop**:
    - Collect post-campaign customer feedback to identify factors influencing decisions beyond the available dataset. Use this qualitative data to improve future models.

CONCLUSION

By leveraging machine learning, the analysis identifies key features influencing customer decisions and optimizes telemarketing strategies for higher success rates. Implementing these insights will not only improve response rates but also reduce operational costs by focusing on high-potential customers. Future improvements can enhance prediction accuracy and model generalizability, driving sustained business value.