Bank Marketing Analysis

Rong Wan, Hala Arar & Fazeeia Mohammed

2024-11-21

Contents

Summary

This lab report analyzes bank marketing campaigns with the goal of using machine learning to predict whether a customer will subscribe to a term deposit. The dataset, sourced from the UCI Machine Learning Repository, contains demographic and campaign-related information on customers who were contacted via phone for a Portuguese bank's direct marketing campaign (Moro and Cortez 2014). The target variable is whether or not the customer subscribed to a term deposit. This study evaluates the performance of Logistic Regression and Decision Tree models in predicting customer subscription to term deposits, using metrics such as accuracy, precision, recall, and F1 score. The Logistic Regression model achieved 88.5% accuracy with high precision (0.70) but low recall (0.20), making it suitable for minimizing false positives. Conversely, the Decision Tree model achieved 89.7% accuracy with improved recall (0.23) but lower precision (0.63), better identifying potential subscribers at the cost of higher false positives. Both models emphasize the majority class (non-subscribers) and highlight challenges in detecting true positives. Strategic recommendations include targeted marketing, personalized offers, and continuous monitoring and adjustment of the models to improve performance. By leveraging these models, banks can enhance marketing strategies, optimize resource allocation, and increase conversion rates.

Introduction

Bank marketing campaigns are a critical tool for financial institutions to promote their products and services, particularly time deposit subscriptions (Meshref 2020). However, identifying potential customers who are likely to respond positively to these campaigns can be challenging (Meshref 2020). Despite advances in targeted marketing strategies, response rates for bank marketing campaigns remain low, and ineffective campaigns can lead to wasted resources and decreased customer satisfaction (Xie et al. 2023).

One notable study in this area is "Predictive Analytics and Machine Learning in Direct Marketing for Anticipating Bank Term Deposit Subscriptions" by (Zaki et al. 2024). The authors explore how machine learning models, including the SGD Classifier, k-nearest neighbor Classifier, and Random Forest Classifier, can be used to predict bank term deposit subscriptions. The study employs various data exploration and feature engineering techniques to build and evaluate the models, ultimately identifying the Random Forest Classifier as the most effective, achieving an impressive accuracy of 87.5%. This study underscores the potential of machine learning to enhance marketing strategies in the banking sector, providing valuable insights that can help institutions refine their direct marketing approaches and improve customer acquisition.

In recent years, the use of machine learning and data mining techniques in the banking sector has gained significant traction, particularly for customer targeting and marketing optimization. A study by (Wang 2020) examines the application of machine learning algorithms, specifically

the C5.0 algorithm, to classify bank customers in order to improve marketing strategies. Using the Bank Marketing dataset from the UCI Machine Learning Repository, the study demonstrates how data mining can help identify customer segments, allowing banks to tailor their marketing campaigns more effectively. The classification model results can enhance decision-making processes for banks, ultimately improving marketing efficiency and customer satisfaction. The study highlights the importance of selecting relevant features, handling outliers, and balancing the dataset to ensure more accurate predictions.

This research raises the question of whether a machine learning algorithm can predict whether a customer will subscribe to a term deposit based on customer demographics and campaign-related data. This is an important inquiry because traditional marketing methods often rely on manual segmentation or generalized strategies, which may not capture the nuances of customer behavior. Additionally, by excluding customers who are unlikely to subscribe, banks can reduce campaign costs and improve customer experience. Conversely, accurately identifying potential subscribers allows banks to concentrate efforts on the right audience, improving both efficiency and outcomes. Therefore, if a machine learning algorithm can accurately predict customer subscriptions based on the bank marketing dataset, it could enable more effective, scalable, and data-driven marketing strategies, leading to better resource allocation and enhanced campaign performance.

Methods

Data

The dataset used in this project is the Bank Marketing dataset, sourced from the UCI Machine Learning Repository (Moro and Cortez 2014). It contains information related to direct marketing campaigns (via phone calls) conducted by a Portuguese banking institution to predict if a client will subscribe to a term deposit. The dataset contains 45,211 rows and 17 columns and it includes features such as age, job type, marital status, education, balance, and details about previous marketing campaigns. The target variable in this study is "y," which indicates whether a customer subscribed to a term deposit (binary: "yes" or "no"). We processed and analyzed this data using Python with libraries such as pandas, scikit-learn, and matplotlib to implement data cleaning, exploratory data analysis, and machine learning models. The data has been pre-processed and contains no missing values.

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
age	Feature	Integer	nan	Age	years	no
job	Feature	Categorical	Occupation	Type of job (categorical: ‘admin.’, ‘blue-collar’,	nan	no
marital	Feature	Categorical	Marital Status	Marital status (categorical: ‘divorced’, ‘married’	nan	no
education	Feature	Categorical	Education Level	Education level (categorical: ‘basic.4y’, ’basic.6	nan	no
default	Feature	Binary	nan	Has credit in default? (binary: ‘yes’, ‘no’)	nan	no
balance	Feature	Integer	nan	Average yearly balance (numeric)	euros	no
housing	Feature	Binary	nan	Has housing loan? (binary: ‘yes’, ‘no’)	nan	no
loan	Feature	Binary	nan	Has personal loan? (binary: ‘yes’, ‘no’)	nan	no
contact	Feature	Categorical	nan	Contact communication type (categorical: ’cellular	nan	yes
day_of_week	Feature	Date	nan	Last contact day of the week (categorical: ‘mon’, …	nan	no
month	Feature	Date	nan	Last contact month of the year (categorical: ‘jan’	nan	no
duration	Feature	Integer	nan	Last contact duration, in seconds (numeric). Impor	seconds	no
campaign	Feature	Integer	nan	Number of contacts performed during this campaign …	nan	no
pdays	Feature	Integer	nan	Number of days that passed by after the client was	days	yes
previous	Feature	Integer	nan	Number of contacts performed before this campaign …	nan	no
poutcome	Feature	Categorical	nan	Outcome of the previous marketing campaign (catego	nan	yes
y	Target	Binary	nan	Has the client subscribed to a term deposit? (bina	nan	no

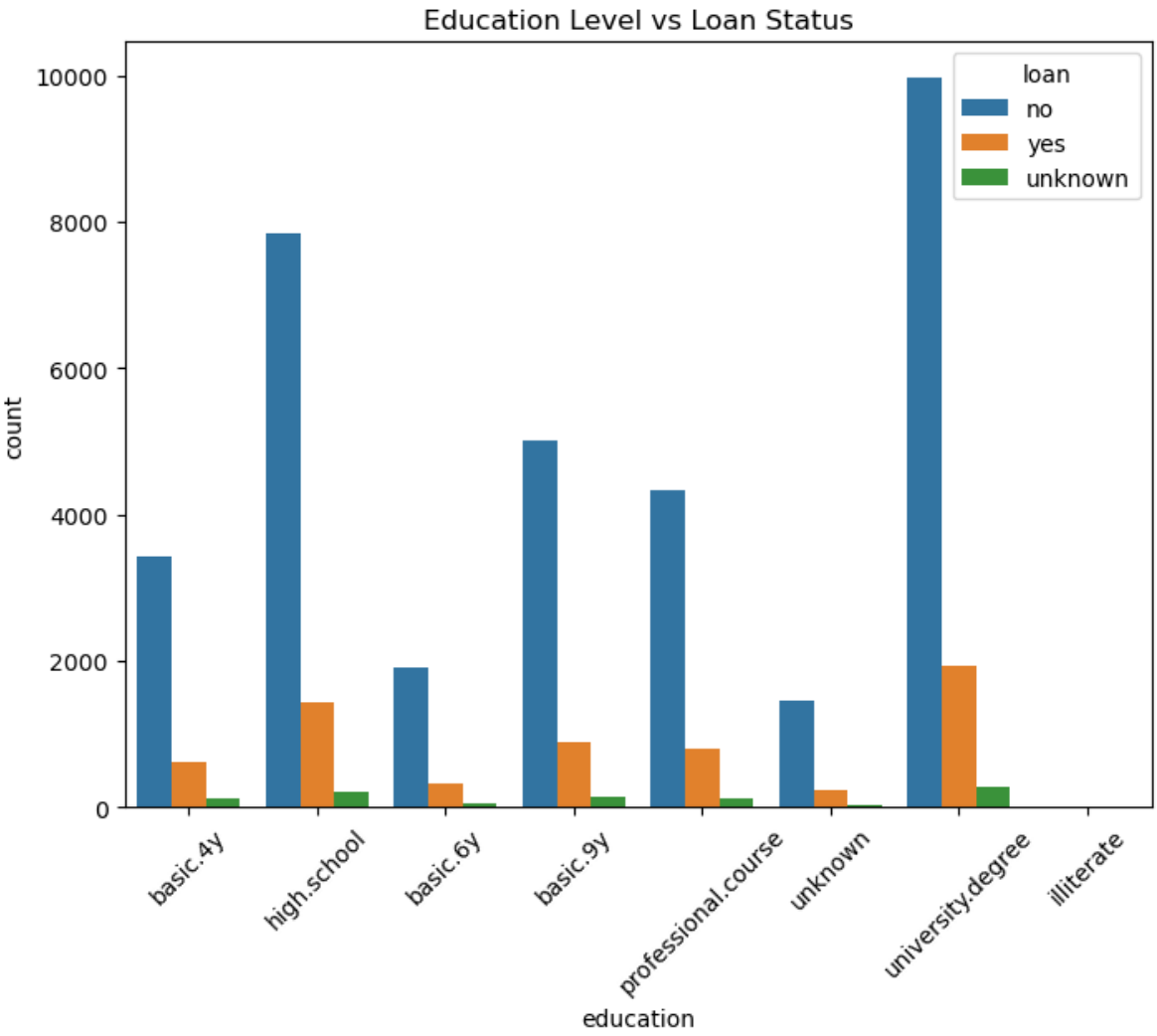

Figure 1: Loan Status distribution across education levels

Data Validation Check

```
Data validation failed with the following errors:
- Dataset contains duplicate rows.

Data passed outlier validation checks.

Data passed category level validation checks.

Target validation passed.

No anomalous correlations detected between target and features.

Warning: Anomalous correlations detected between features: ['age', 'job', 'marital', 'educati
```

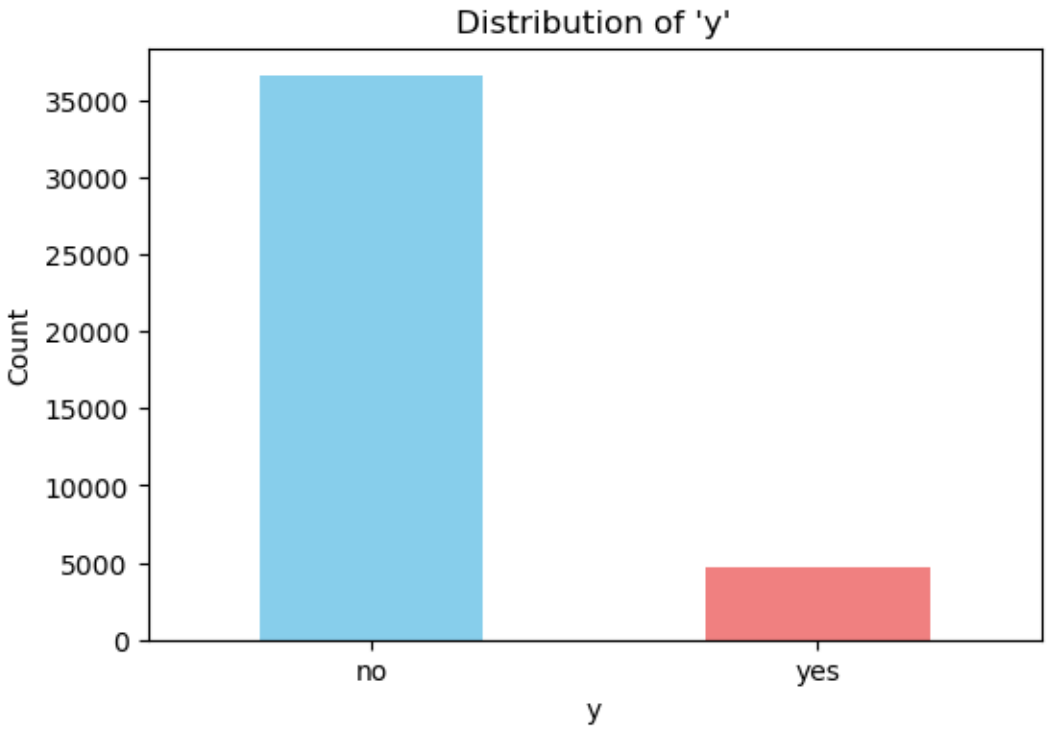

Figure 2: Distribution of Y

Analysis

The analysis began with loading and preprocessing the dataset, addressing missing values, encoding categorical features, and scaling numeric variables to ensure consistency across features. The dataset was then split into training and testing sets, with 20% allocated for testing to evaluate model performance. A logistic regression model was chosen for binary classification, implemented through a Pipeline to streamline preprocessing, encoding, and model fitting. To optimize model accuracy, GridSearchCV was used for hyperparameter tuning, and cross-validation was employed to assess the model's robustness. After training the model, its performance was evaluated using various metrics such as accuracy, precision, recall, and F1-score, with confusion matrices and heatmaps created using Seaborn for better visualization. These tools provided insights into the model's ability to differentiate between classes.

Results

To evaluate the utility of each predictor in predicting the response variable (y) for the bank marketing dataset, we visualized the distributions of each predictor in the training dataset, coloring them by the class (yes: orange and no: blue). These visualizations include univariate distributions, pairwise correlations, and scatterplots, as seen in the attached figures. In analyzing these plots, we observe significant differences in the distribution centers and spreads of predictors like duration and campaign between the two classes. However, some variables, such as age and balance, show overlapping distributions with less apparent class separation. Furthermore, categorical predictors, such as job and month, exhibit class imbalance but may still hold valuable predictive information. Based on these insights, predictors demonstrating clear separability and meaningful patterns are prioritized for inclusion in the predictive model, while those showing little to no differentiation may be considered for exclusion.

Model Creation

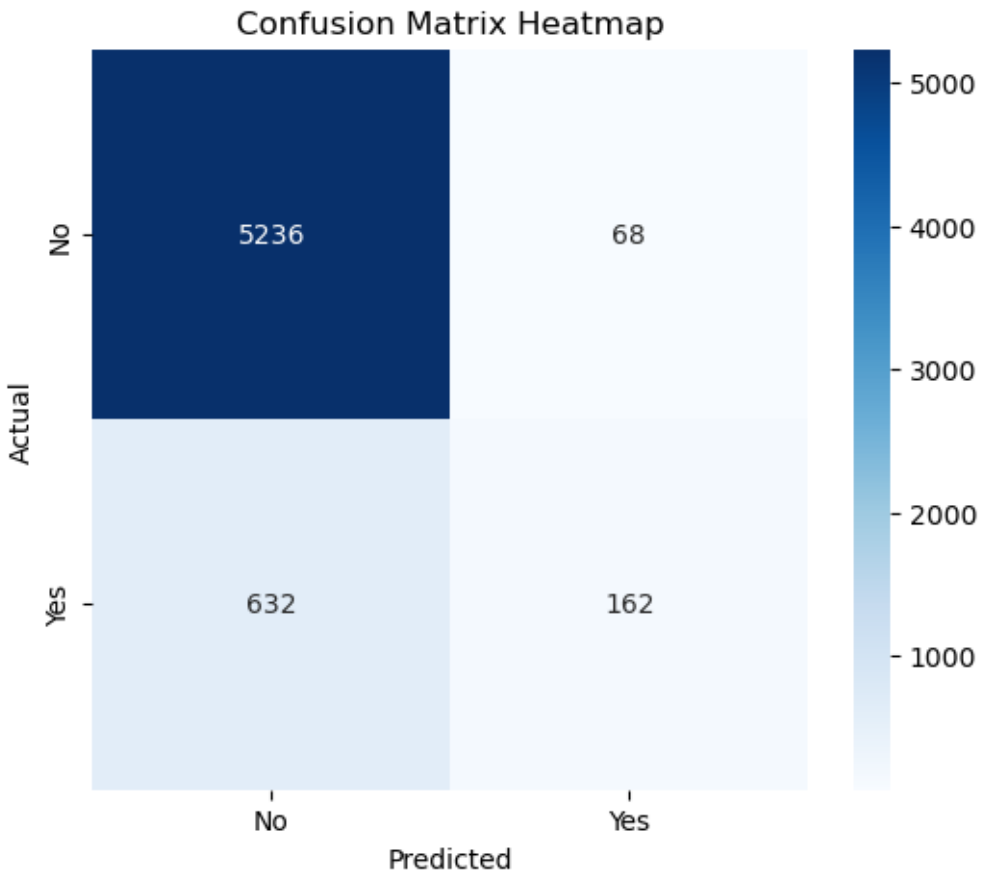

Figure 3: Confusion Matrix Heatmap - Liner Regression Model

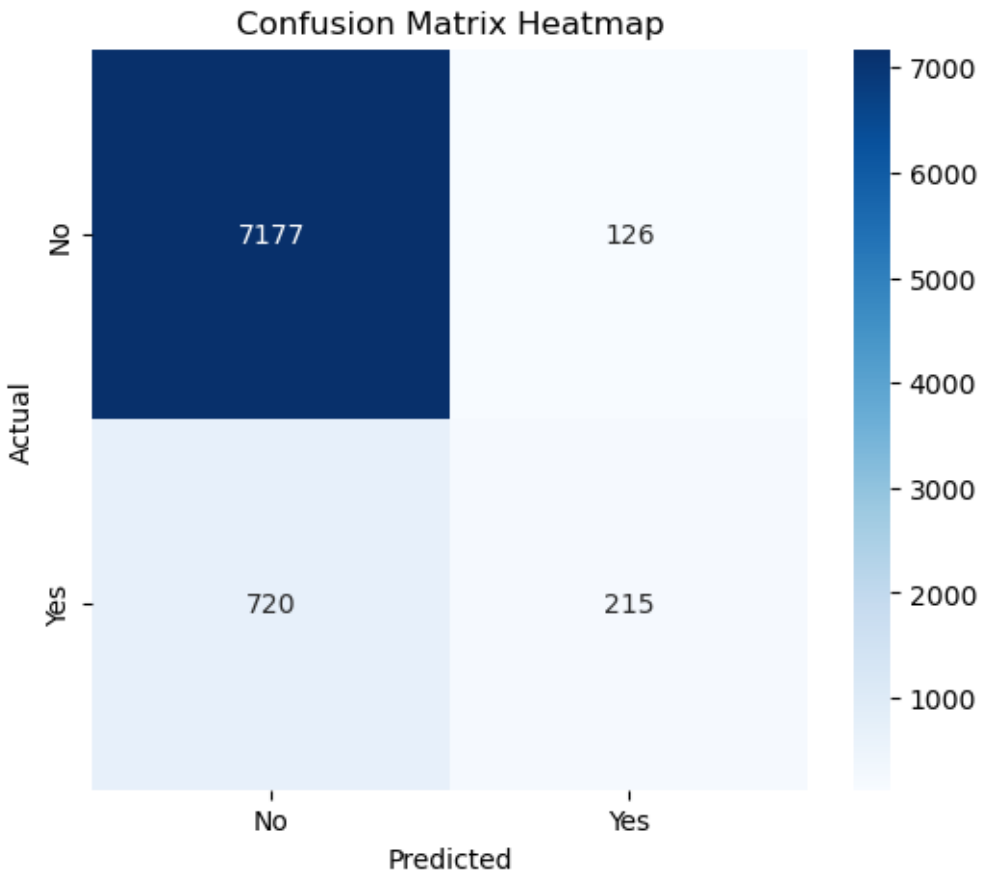

Figure 4: Confusion Matrix HeatMap - Decision Tree Model

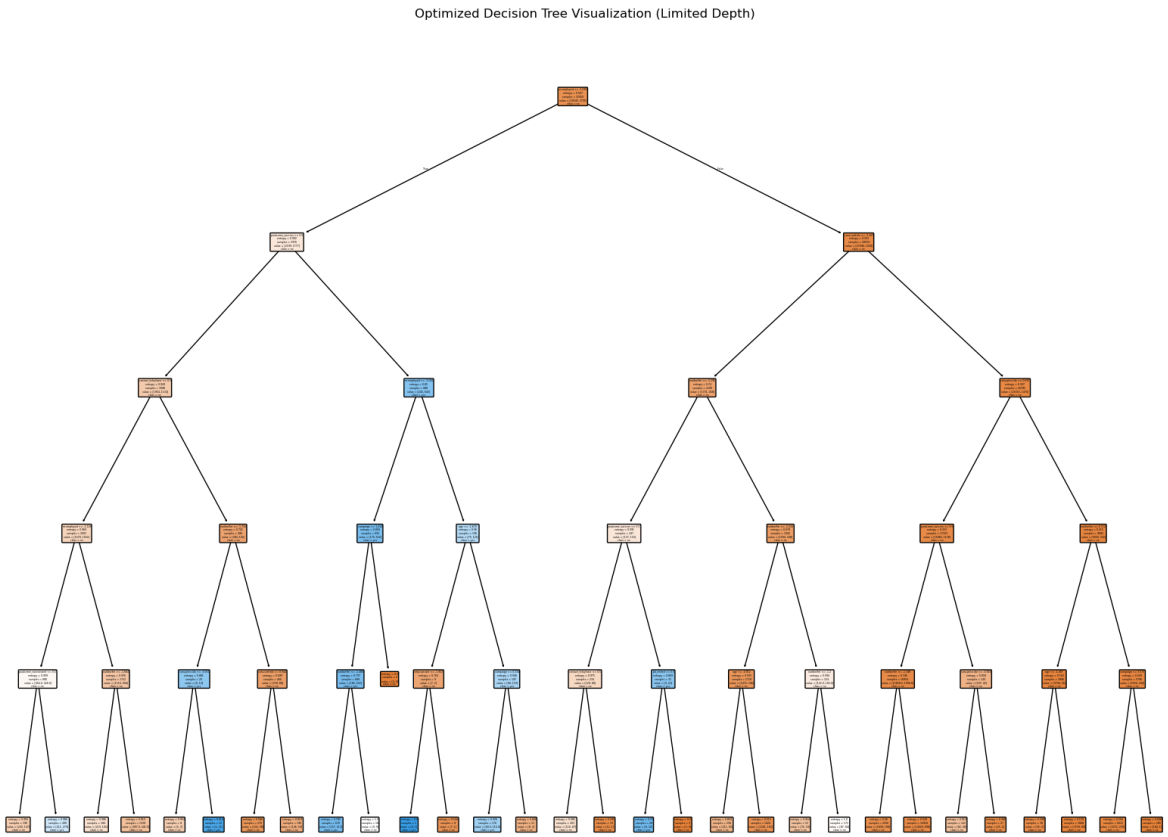

Figure 5: Optimized Decision Tree Visualization

Discussion

Logistic Regression Model:

The Logistic Regression model has achieved an accuracy of approximately 88.5%, with the best hyperparameters found as: as follows `{'classifier__C': 0.1, 'classifier__max_iter': 100, 'classifier__penalty': 'l1', 'classifier__solver': 'liblinear'}`.

The confusion matrix for this model as shown in Figure 3:

- **True Negatives (5236)**: The model correctly identified 5236 non-subscribers, which indicates its strong performance in predicting the majority class (non-subscribers).
- **False Positives (68)**: There are 68 instances where the model incorrectly predicted that non-subscribers would subscribe. This is a relatively low number, indicating that the model is relatively efficient at avoiding unnecessary targeting.
- **False Negatives (632)**: The model missed 632 actual subscribers, which is a significant number and highlights the low recall.
- **True Positives (162)**: The model correctly predicted 162 subscribers, but this number is still quite low, reflecting the model's struggle to identify potential subscribers.

The **Precision** is 0.70, meaning that 70% of the customers predicted as subscribers are actually subscribers. However, the **Recall** is only 0.20, meaning the model captures just 20% of the actual subscribers, which is quite low. This results in an **F1 Score** of 0.32, reflecting a poor balance between precision and recall. Despite the good precision, the low recall suggests that the model is not effectively identifying many actual subscribers, pointing to a significant trade-off between false positives and false negatives. This version of Logistic Regression is more suited to scenarios where **precision** (minimizing false positives) is prioritized over **recall** (capturing all potential subscribers).

Decision Tree Model:

After performing a grid search for hyperparameter optimization, the best hyperparameters found are: `{'classifier__criterion': 'entropy', 'classifier__max_depth': 5, 'classifier__min_samples_leaf': 1, 'classifier__min_samples_split': 2}`.

The model achieved an accuracy of approximately 89.7%, with the confusion matrix as shown in Figure 4

- **True Negatives (7177)**: The Decision Tree correctly predicted 7177 non-subscribers, showing solid performance in predicting the majority class (non-subscribers).
- **False Positives (126)**: There are 126 instances where the model incorrectly predicted non-subscribers as subscribers, which is a moderate number compared to the Logistic Regression model, indicating a higher sensitivity to identifying potential subscribers.

- **False Negatives (720)**: The model failed to predict 720 actual subscribers, a somewhat higher number, reflecting a lower recall than might be ideal.
- **True Positives (215)**: The Decision Tree correctly predicted 215 subscribers, which is an improvement over the Logistic Regression model, suggesting it is better at identifying potential subscribers.

The **Precision** is 0.63, meaning 63% of the customers predicted as subscribers are indeed subscribers. The **Recall** is 0.23, meaning the model captures only 23% of actual subscribers, indicating it still misses a significant portion. This results in an **F1 Score** of 0.34, which is slightly higher than the Logistic Regression model but still reflects an imbalance between precision and recall. The Decision Tree model performs better than Logistic Regression in terms of recall but still struggles to capture a large proportion of the potential subscribers. It might benefit from further adjustments, such as pruning, to reduce the number of false positives and improve its recall.

Although the Decision Tree has a slightly lower accuracy, its **higher recall** (more true positives) suggests it is better at identifying potential subscribers. However, its higher **false positives** indicate that the model might be overfitting, capturing noise in the data. This suggests that the Decision Tree is more sensitive to patterns in the data but might benefit from **regularization** or **pruning** to reduce overfitting.

Comparison and Implications:

Both models indicate that the most common outcome in the dataset is non-subscription, as reflected in the confusion matrices, where the number of true negatives vastly outweighs the number of true positives. This confirms that "no" is the statistically likely outcome for customer subscription.

- **Logistic Regression Model**: The Logistic Regression model is better suited for situations where minimizing false positives is critical, as its **precision** (0.70) is higher than that of the Decision Tree model. However, its **recall** (0.20) is lower, meaning it misses a significant portion of actual subscribers. This makes the Logistic Regression model more effective in contexts where avoiding unnecessary targeting of non-subscribers is more important than capturing every potential subscriber.
- **Decision Tree Model**: The Decision Tree model, while slightly less accurate overall (**accuracy = 89.7%**), has a better **recall** (0.23), meaning it identifies more true positives compared to Logistic Regression. However, this comes at the cost of an increased number of **false positives** (126). As such, the Decision Tree is better at capturing potential subscribers but may lead to more resources being spent on non-converting customers.

Implications:

Both models show reasonable accuracy and can be useful for the business's marketing initiatives to increase term deposits (subscriptions). The Logistic Regression model would be advantageous in scenarios where reducing false positives and minimizing resource expenditure is a priority, while the Decision Tree model could be valuable in situations where capturing more potential subscribers (even at the cost of more false positives).

Future iterations of these models should focus on improving both **precision** and **recall**, possibly through regularization, pruning, or incorporating more diverse data to better identify customers likely to subscribe. By fine-tuning the models, the business can maximize the effectiveness of its marketing campaigns and increase its return on investment.

Strategic Recommendations:

Given the insights from the evaluation of both models, here are some actionable strategies to enhance the bank's marketing efforts and improve conversion rates:

Targeted Marketing:

- Use these models to segment customers into two groups: those with a high likelihood of subscribing (identified by the model as potential positives) and those with a low likelihood (predicted as negatives). Focus marketing efforts on the high-probability segment to optimize resource allocation.

Campaign Timing:

- Refine marketing strategies by focusing efforts on customers during certain times when they are more likely to respond. The model can be expanded to include temporal features (e.g., day of the week or month) to optimize campaign timing.

Personalized Offers:

- Tailor offers to individual customers based on characteristics like age, occupation, or previous interactions with the bank (e.g., loan status). The models' predictions can guide personalized messaging, increasing engagement with customers and improving the chances of subscription.

Improve Conversion Rates:

- Implement **follow-up campaigns** targeting customers predicted as high-likelihood subscribers but who still did not convert. For those predicted as low-likelihood, consider creating new or improved offers to address specific concerns or barriers to subscription.

Monitor and Adjust:

- Continuously track the performance of both models over time, paying close attention to precision and recall. As more data becomes available, adjust the models and marketing strategies to ensure increasing accuracy and the development of more effective campaigns.

By applying these insights and strategies, the bank can improve its targeting for **long-term deposit** products, increasing conversion rates while making sure the marketing efforts are cost-effective and personalized.

–For the markdown rendering Chat-gpt was used to correct code

References

Meshref, H. 2020. "Predicting Loan Approval of Bank Direct Marketing Data Using Ensemble Machine Learning Algorithms." *International Journal of Circuits, Systems and Signal Processing* 14: 1–9. https://doi.org/https://doi.org/10.46300/9106.2020.14.117.

Moro, Rita, S., and P. Cortez. 2014. "Bank Marketing." UCI Machine Learning Repository.

Wang, D. 2020. "Research on Bank Marketing Behavior Based on Machine Learning." In *AIAM2020: Proceedings of the 2nd International Conference on Artificial Intelligence and Advanced Manufacture*, 150–54. https://doi.org/https://doi.org/10.1145/3421766.3421800.

Xie, C., J.-L. Zhang, Y. Zhu, B. Xiong, and G.-J. Wang. 2023. "How to Improve the Success of Bank Telemarketing? Prediction and Interpretability Analysis Based on Machine Learning." *Computers & Industrial Engineering* 175: 108874. https://doi.org/https://doi.org/10.1016/j.cie.2022.108874.

Zaki, A. M., N. Khodadadi, W. H. Lim, and S. K. Towfek. 2024. "Predictive Analytics and Machine Learning in Direct Marketing for Anticipating Bank Term Deposit Subscriptions." *American Journal of Business and Operations Research* 11 (1): 79–88. https://doi.org/https://doi.org/10.54216/AJBOR.110110.