Bank Marketing Prediction

OPTIMIZING TERM DEPOSIT CAMPAIGNS USING MACHINE LEARNING

AUTHORS: HALA ARAR, FAZEEIA MOHAMMED, RONG WAN

Importance of Bank Marketing Campaigns

- Promote products/services, especially term deposits
- Challenges: Identifying potential customers
- Consequences: Low response rates

Research Objective

Goal:

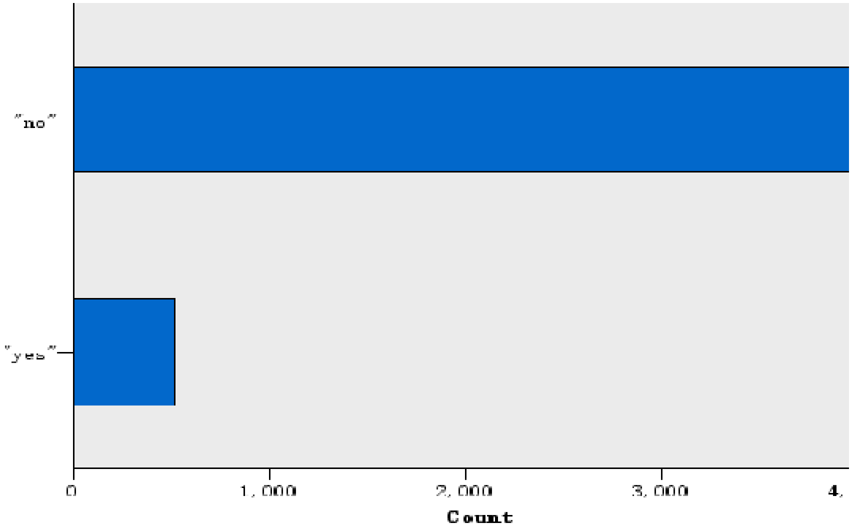
Predict customer subscription to term deposits using machine learning.

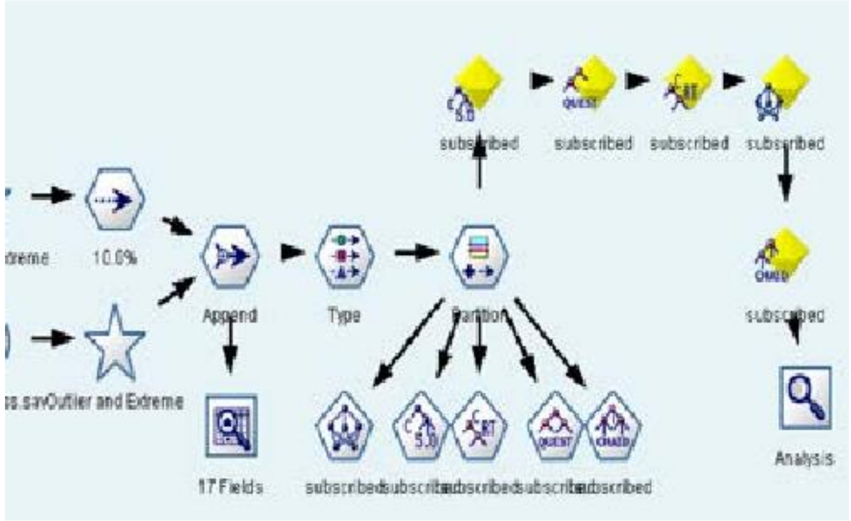
Purpose:

Improve campaign efficiency,

Minimize wasted resources,

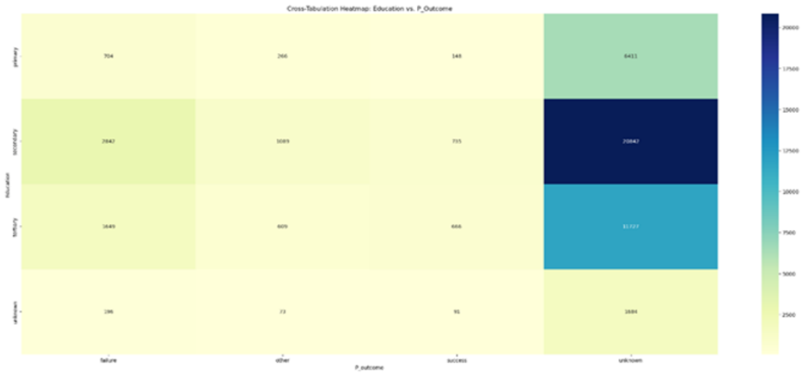
Increase customer satisfaction.

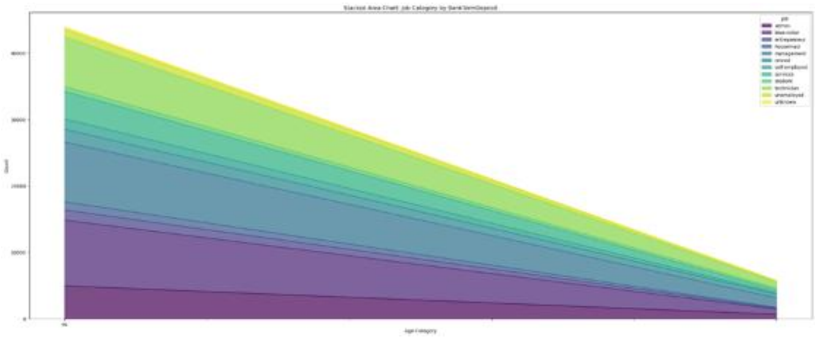




"Research on Bank Marketing Behavior Based on Machine Learning" **by Wang (2020)**

- Used the C5.0 algorithm to classify customers.
- Focused on segmentation for better-targeted campaigns.
- Highlighted importance of quality data





"Predictive Analytics and Machine Learning in Direct Marketing for Anticipating Bank Term Deposit Subscriptions" by Zaki et al. (2024)

- Explored machine learning models for predicting term deposit subscriptions.
- Outcome: Random Forest achieved 87.5% accuracy.
- Demonstrates machine learning's role in refining direct marketing strategies.

Dataset Overview

Dataset: Bank Marketing dataset from UCI Repository

Size: 45,211 records, 17 features.

Key Features:

- **Demographic**
- **Campaign-related**
- **Target Variable**: Subscription to a term deposit (yes/no).

Data Preparation

Tools Used:

- Python libraries: pandas, scikit-learn, matplotlib.

Validation Checks:

- Resolved duplicate rows.
- Passed outlier, category, and correlation checks.

Variables Table

Variable Name	Role	Type	Demographic	Description
age	Feature	Integer	Age	
job	Feature	Categorical	Occupation	type of job (categorical: 'admin.','blue-collar','entrepreneur','houser employed','services','student','technician','unemployed','unknown')
marital	Feature	Categorical	Marital Status	marital status (categorical: 'divorced','married','single','unknown'; n widowed)
education	Feature	Categorical	Education Level	(categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.c
default	Feature	Binary		has credit in default?
balance	Feature	Integer		average yearly balance
housing	Feature	Binary		has housing loan?
loan	Feature	Binary		has personal loan?
contact	Feature	Categorical		contact communication type (categorical: 'cellular','telephone')
day_of_week	Feature	Date		last contact day of the week
month	Feature	Date		last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', '
duration	Feature	Integer		last contact duration, in seconds (numeric). Important note: this at target (e.g., if duration=0 then y='no'). Yet, the duration is not know after the end of the call y is obviously known. Thus, this input shou purposes and should be discarded if the intention is to have a reali
campaign	Feature	Integer		number of contacts performed during this campaign and for this cl
pdays	Feature	Integer		number of days that passed by after the client was last contacted f -1 means client was not previously contacted)
previous	Feature	Integer		number of contacts performed before this campaign and for this cl
poutcome	Feature	Categorical		outcome of the previous marketing campaign (categorical: 'failure',
y	Target	Binary		has the client subscribed a term deposit?

Data Validation checks

Correct file format (CSV).

Valid column names and expected data types.

No duplicates or rows with all missing values.

No outliers or anomalous values, correct category levels.

No anomalous correlations between response and features or between features.

Analysis Process

Preprocessing Steps:

Categorical features: Encoded to numerical values.

Numerical variables: Scaled for consistency.

Data split: 80% training, 20% testing.

Model Training:

Logistic Regression:

- Pipeline with preprocessing.
- Hyperparameter tuning using GridSearchCV.

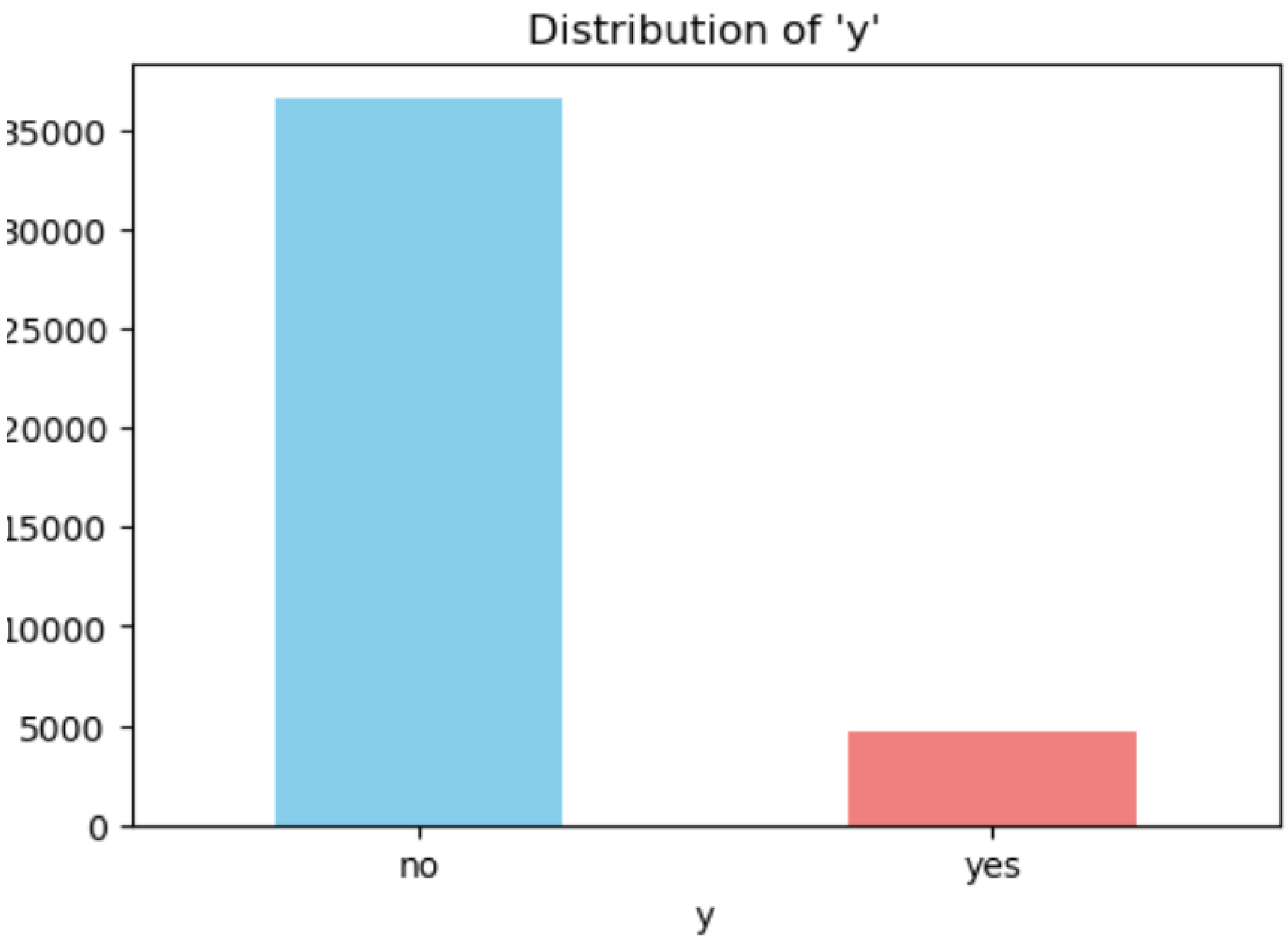
Decision Tree:

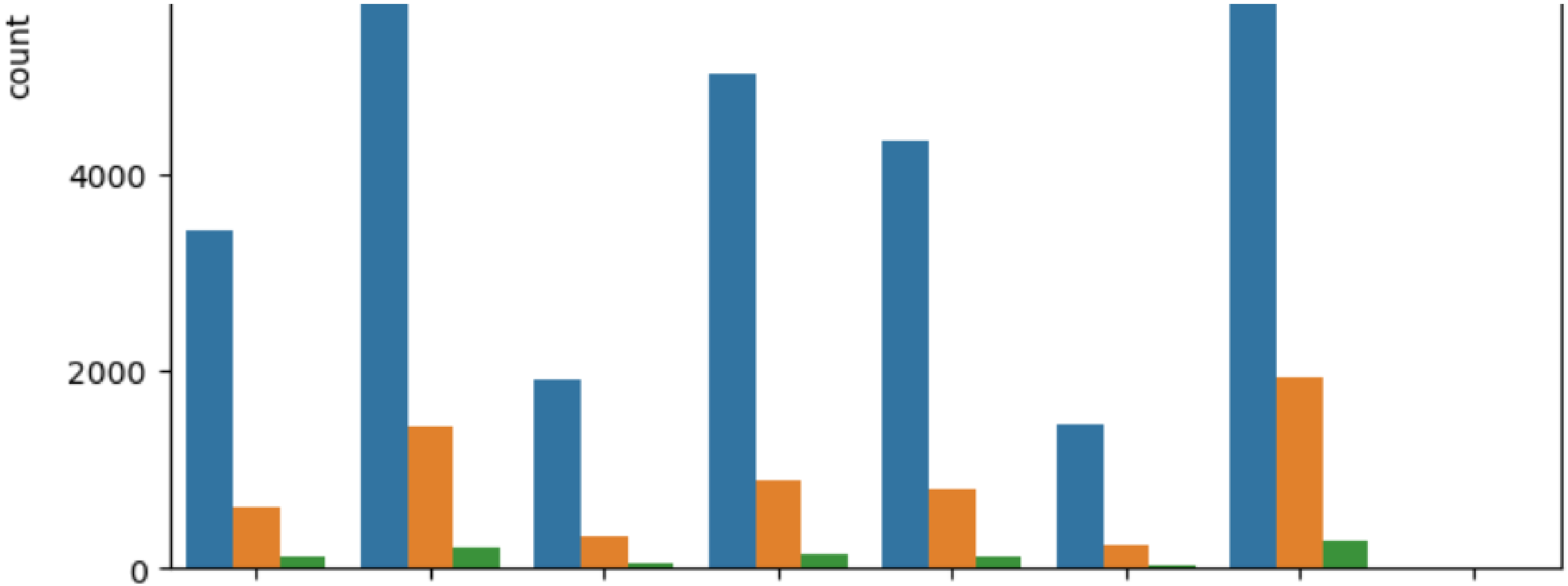
- Optimized depth, splits, and pruning techniques.

Target Variable Distribution:

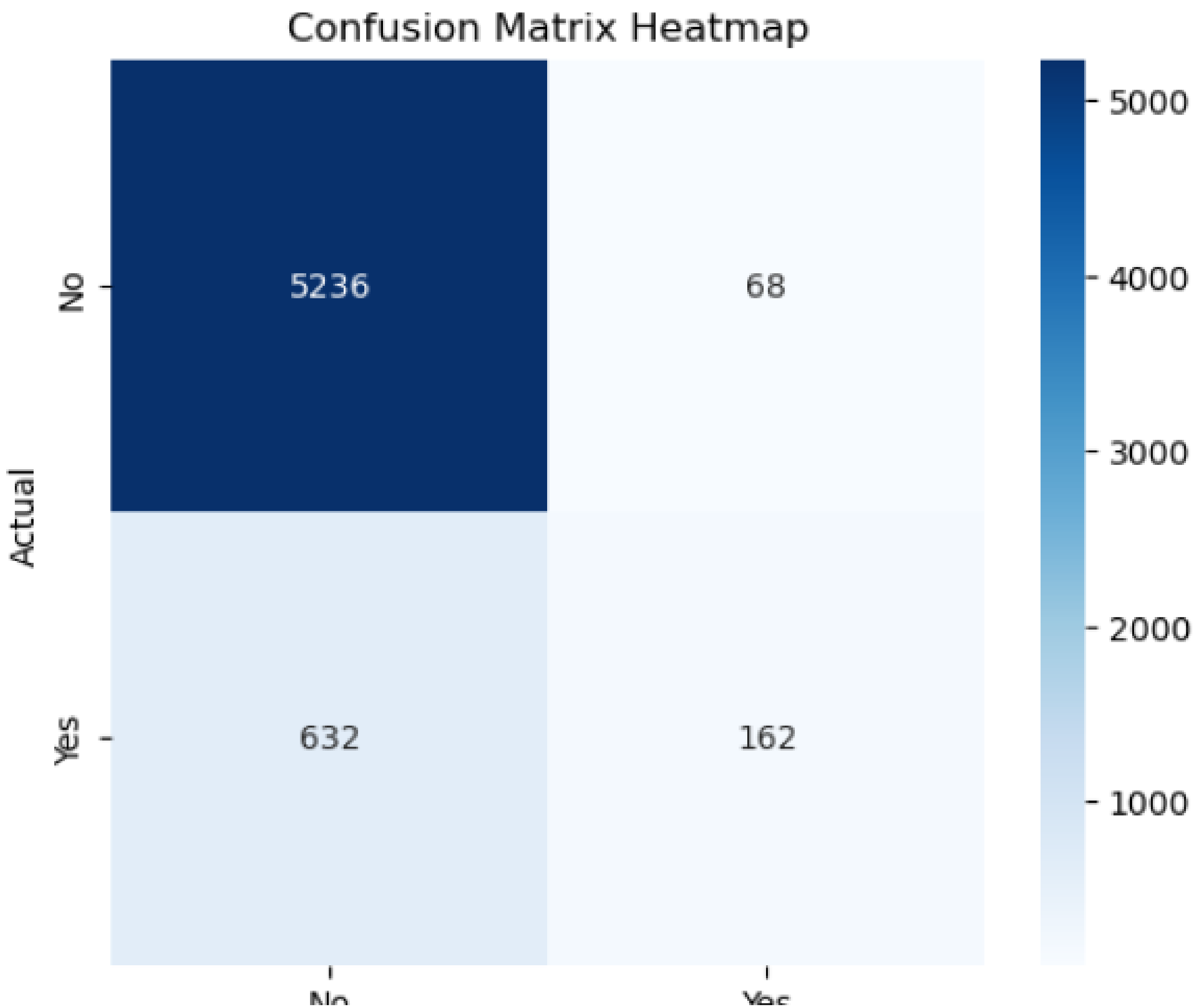
Yes: 10.3% (4,640 records).

No: 89.7% (36,548 records).





Education Level vs Loan Status



Logistic Regression

Strengths:

- Correctly classified **5236 negatives** with **only 68 false positives**.

Weaknesses:

- Identified **162 true positives** but misclassified **632 positives** as negatives.

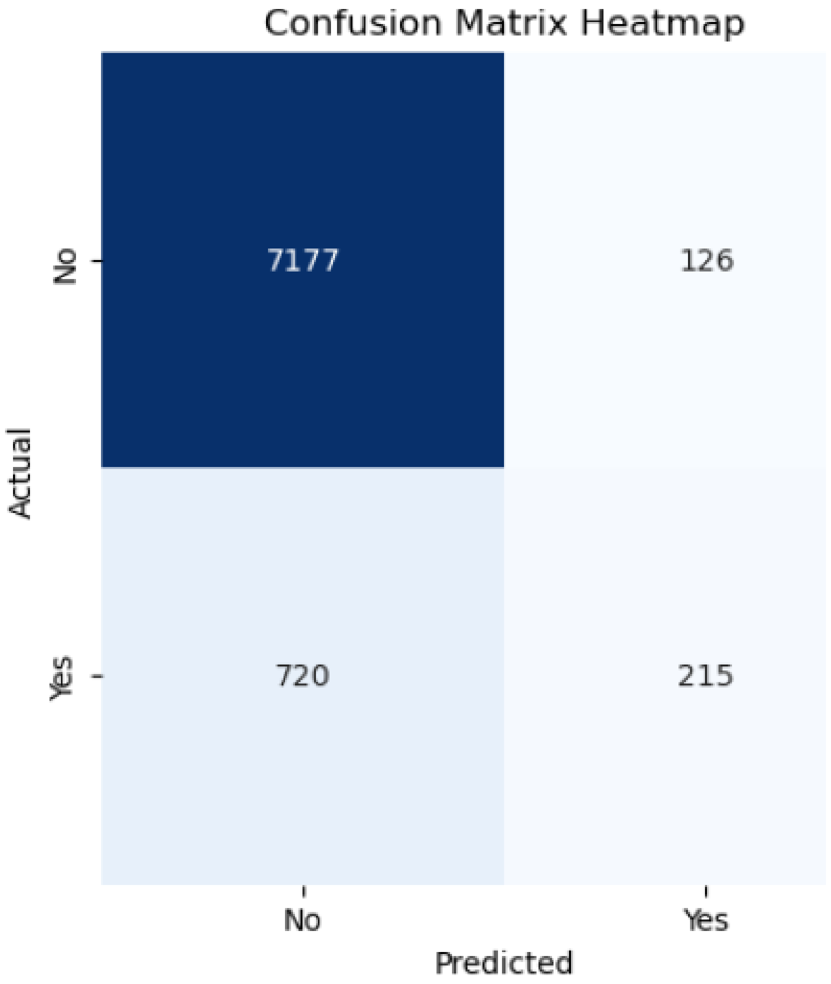


Figure 4: Confusion Matrix HeatMap - Decision '

Decision Tree

Decision Tree:

Strengths:

- Improved recall with **215 true positives**.
- Correctly classified **7177 negatives**, fewer false negatives.

Weaknesses:

- Increased false positives (**126 vs. 68 in Logistic Regression**).

Takeaway:

Logistic Regression prioritizes **precision**; Decision Tree enhances **recall**.

Business use depends on whether to minimize costs or capture maximum subscribers.

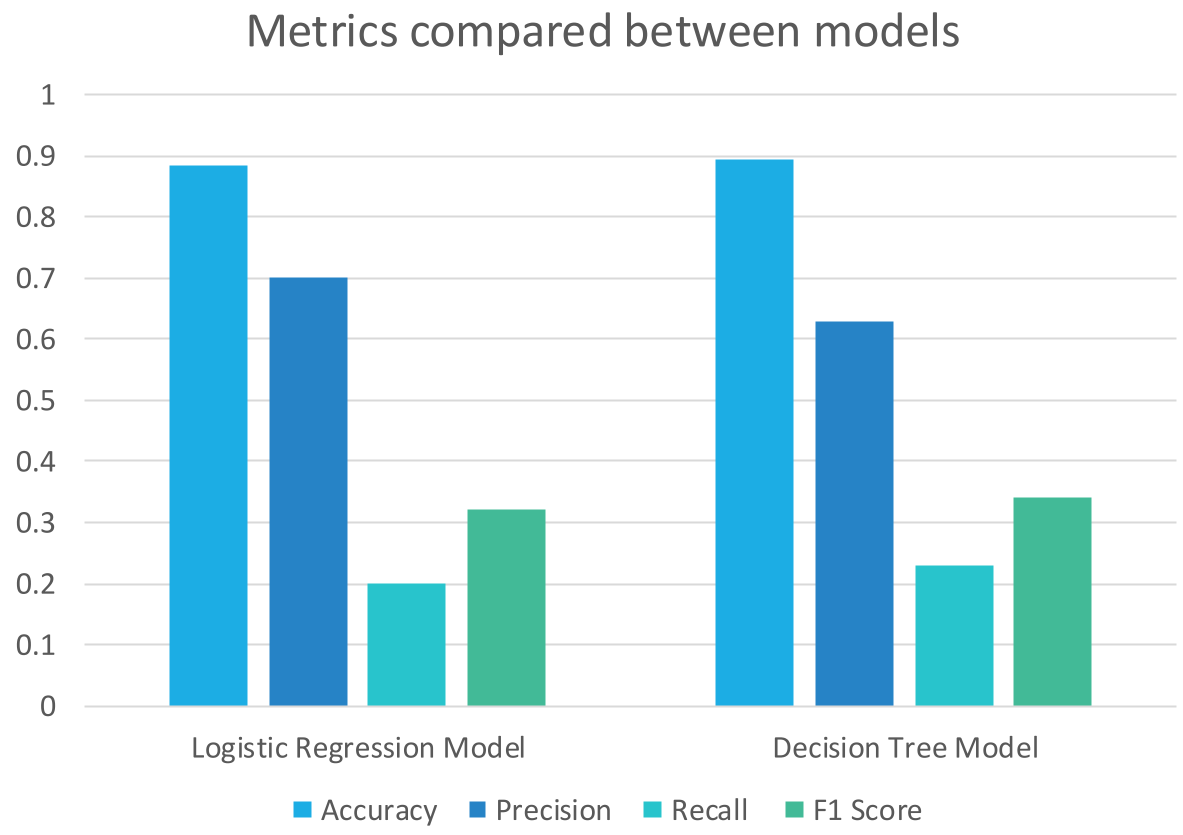
Results & Discussion

I. Logistic Regression Model

- Strengths: Minimizing false positives.
- Weaknesses: Low recall

II. Decision Tree Model

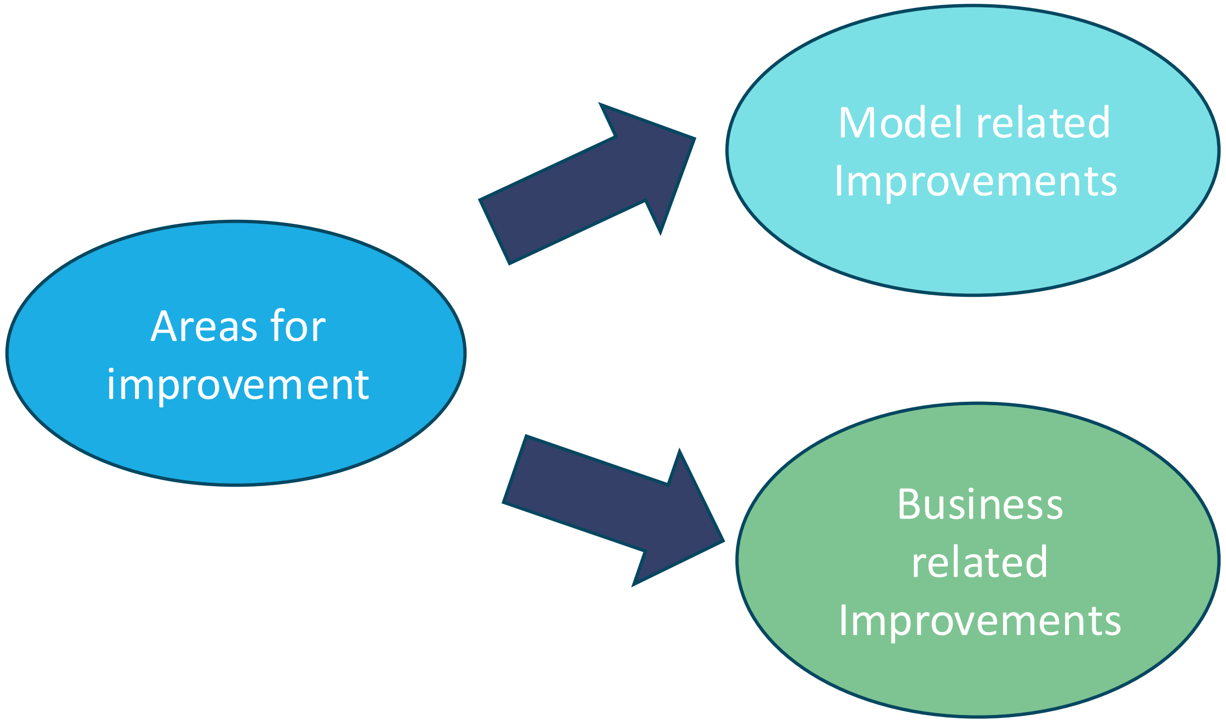
- Strengths: Better recall, identifies potential subscribers.
- Weaknesses: Higher false positives.

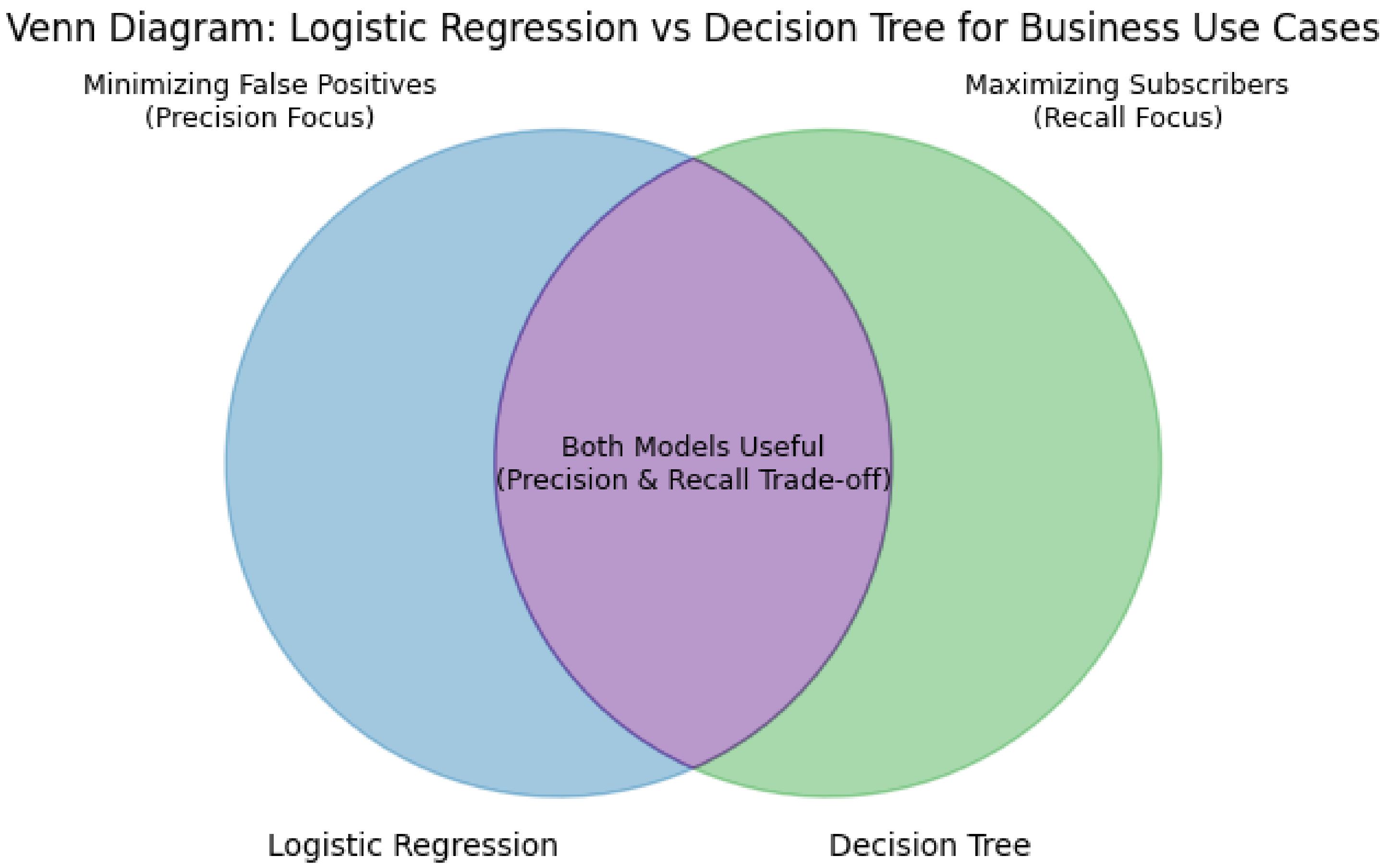


Implications

- Both models have their strengths and can be beneficial for marketing initiatives aimed at increasing customer subscriptions (term deposits).
- **Logistic Regression** is ideal when resource efficiency and minimizing unnecessary outreach are the priorities.
- **Decision Tree** is valuable when the goal is to maximize the number of identified potential subscribers, even if it means spending more resources on customers who may not convert.

Strategic Recommendations




Venn Diagram: Logistic Regression vs Decision Tree for Business Use Cases
Minimizing False Positives
(Precision Focus)
Maximizing Subscribers
(Recall Focus)
Both Models Useful
(Precision & Recall Trade-off)
Logistic Regression
Decision Tree

Conclusion

Questions