DATA MINING

Telemarketing of Banking institution

PRESENTED BY:

Giang Nguyen
Ky Anh Le
Minh Hoang Nguyen

SGH Szkoła Główna Handlowa w Warszawie

Topic justification

The problem of forecasting whether a customer will opt in for a term deposit is very important for monetary establishments. Marketing campaigns that use more direct measures such as telephone calls are expensive as well as quite tedious. Enormously, being able to narrow down on the appropriate clientele is likely to result in reduced advertising costs while enhancing the likelihood of effectiveness of the campaign.

This study utilizes the Bank Marketing dataset that contains detailed information about the prospects that were reached during previous marketing campaigns.

Research questions

Which **demographic** and **campaign-related factors** are **most influential** in determining whether a client will **subscribe** to a term deposit?

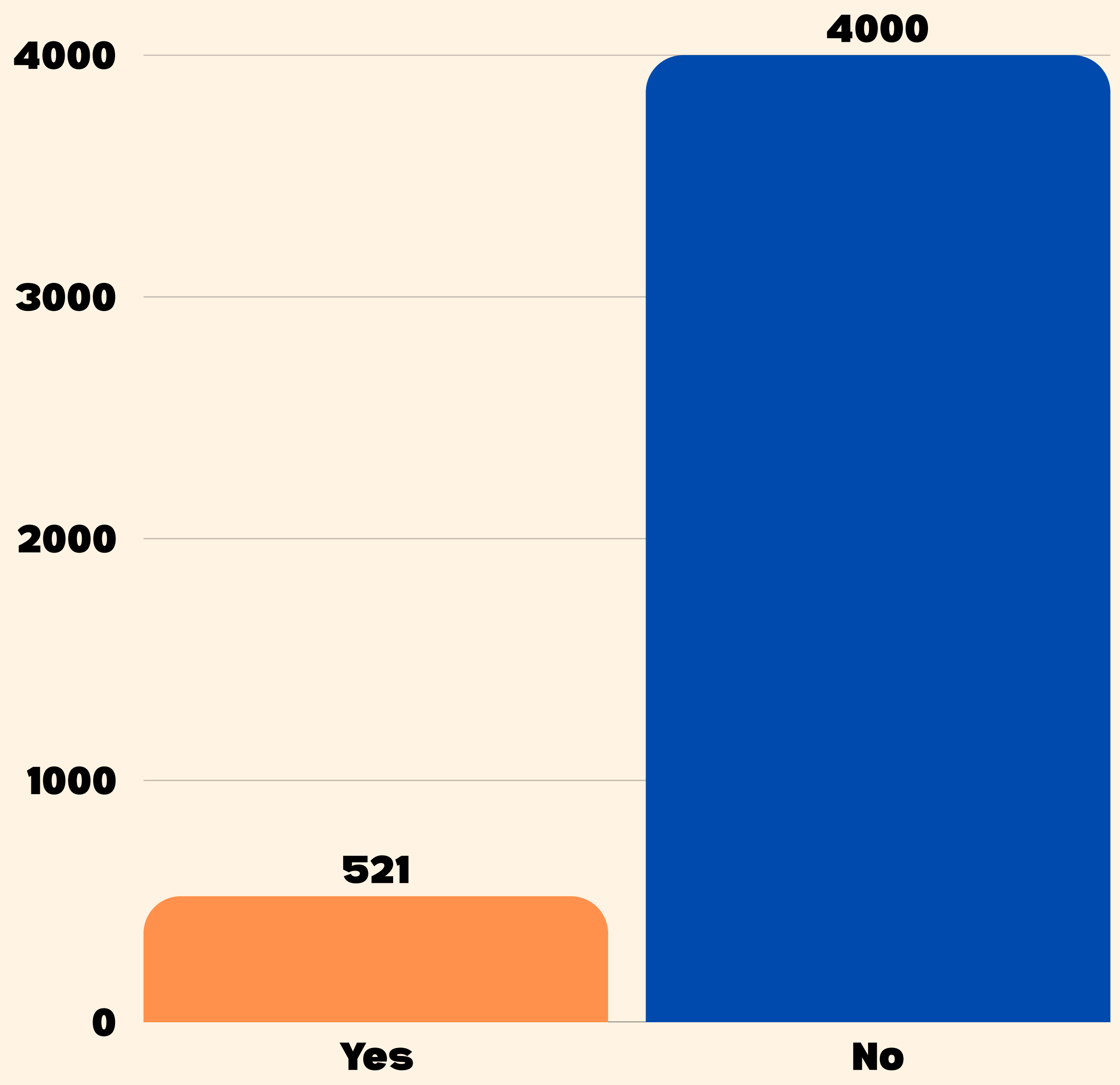
Dataset Overview

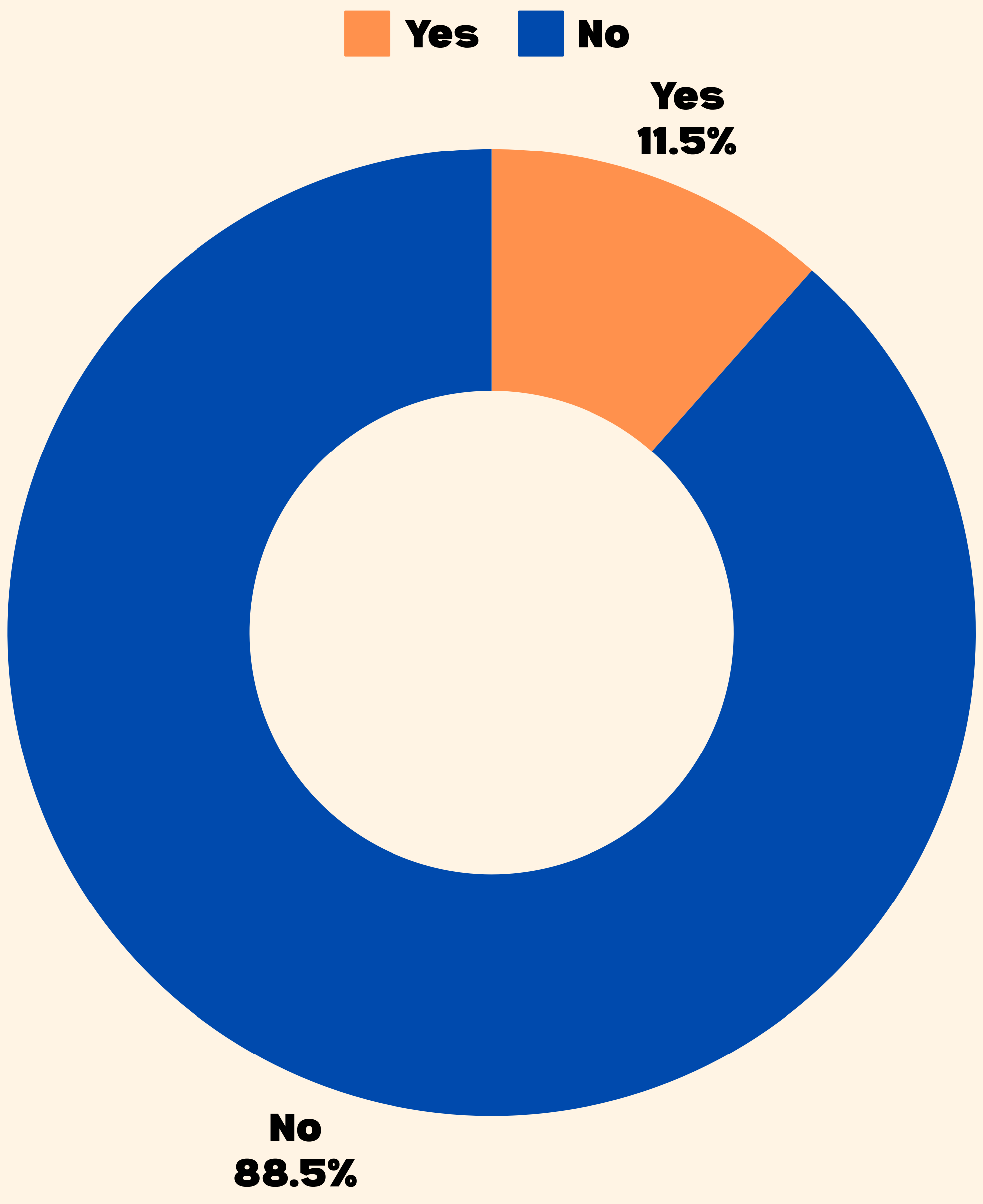
Target variable: y **(customer subscription)**

4521 observations

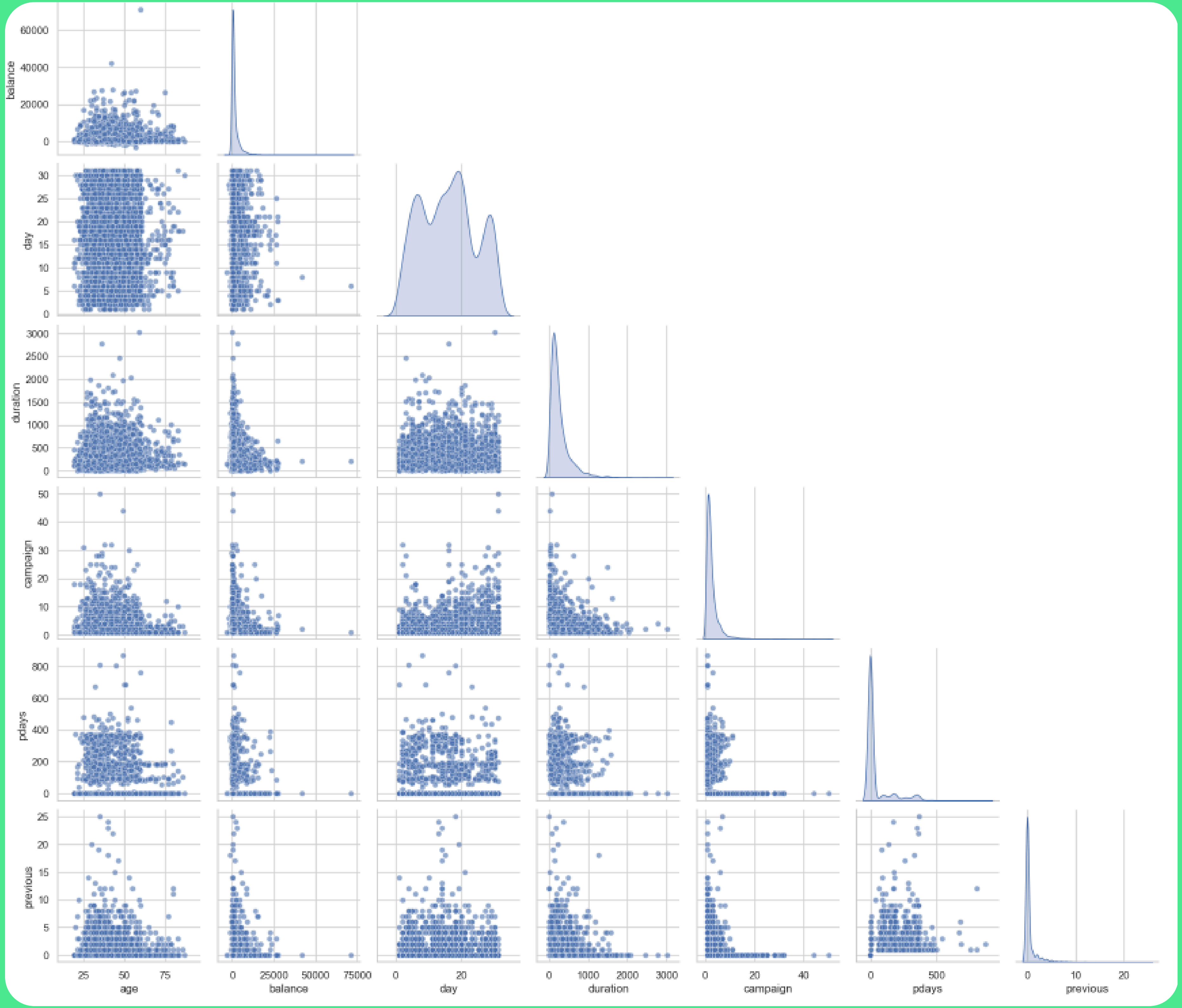
16 features
with 7 numerical and 9 categorical

Target variable





Numerical features



Observation:

No co-linearity

Categorical features

Feature	Job	Marital	Education	Default	Housing	Loan	Contact	Month	Poutcome
Unique value	12	3	2	2	2	2	3	12	4
Mode	management	married	secondary	no	yes	no	cellular	may	unknown
Mode frequency	969	2797	2306	4445	2559	3830	2896	1398	3705

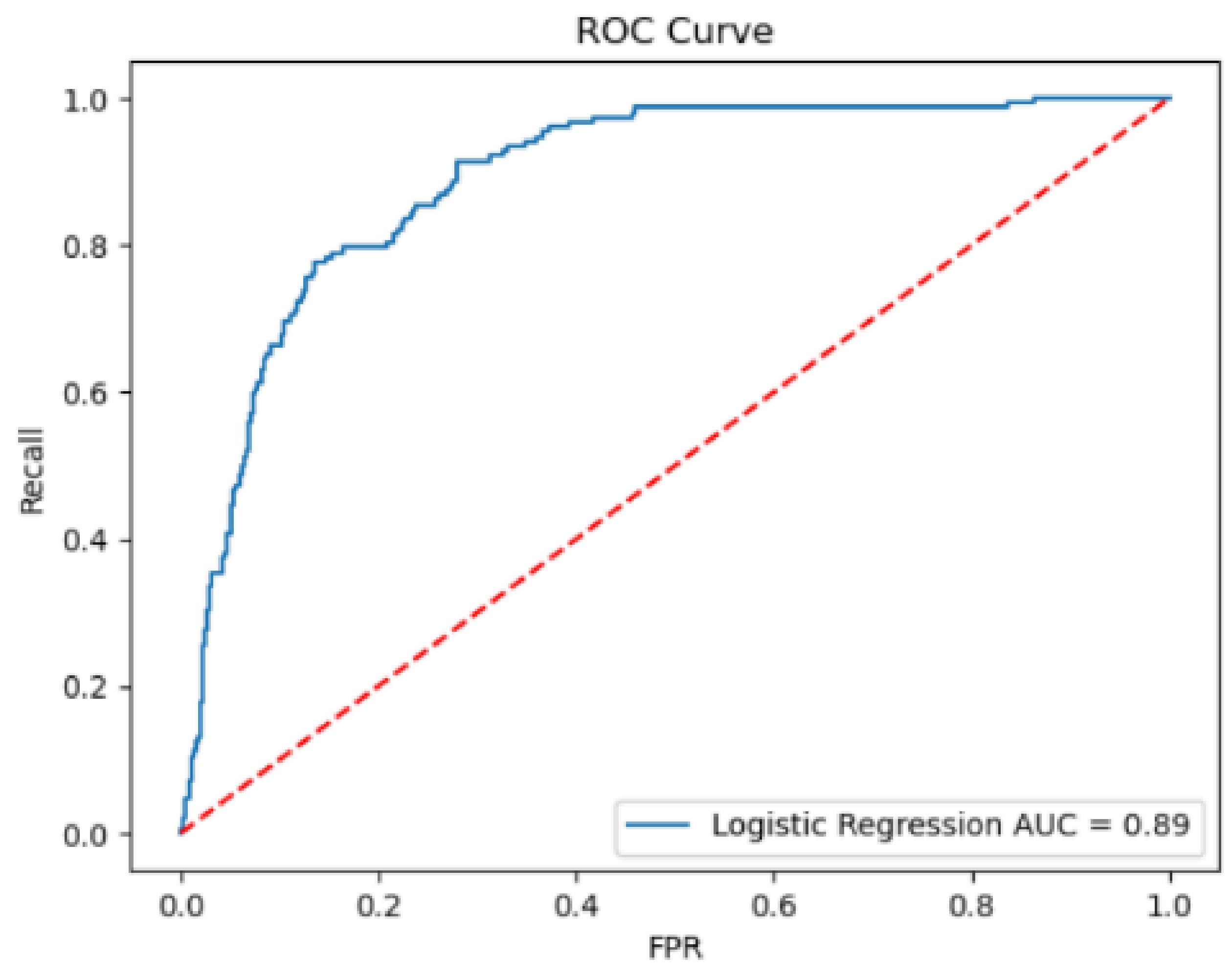
Model I: Logistic Regression

Theory:

Logistic Regression is a **data analysis technique** that **finds a relationship** between **two data elements**. It then uses the **discovered relationship** to **predict the value of those elements** based on the remaining elements.

The prediction usually produces a **finite number of results,** such as yes or no. Putting into reality, Logític Regression **calculates the probability that an event will occur based on the independent variables X**

AUC = 0.89 -> good model because it **discriminates well between the two classes.**



Model I: Logistic Regression

Precision (0.58): Precision above average means that out of all the customers that the model predicts as "Yes", only 58% are actually registered customers -> **High False Positive, the model is not really effective**

Recall (0.33) Recall is very low for the "Yes" class, meaning that the model misses 67% of potential customers. **This is a serious problem because missing registered customers means losing marketing opportunities and revenue.**

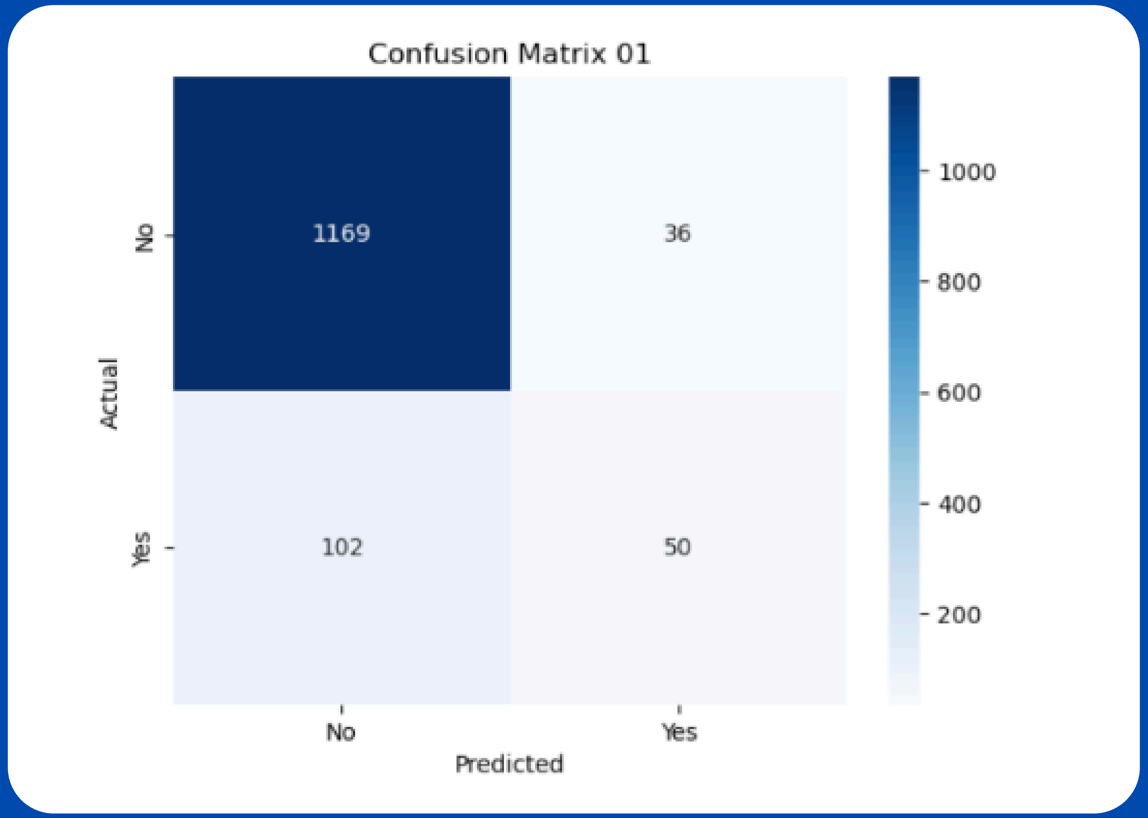
F1-score (0.42) Low F1-Score shows that the model is not well balanced between Precision and Recall.
Accuracy (0.9) Accuracy is high but is not a suitable index for the problem of data imbalance**. Because it only predicts all as "No" (the majority class).**

Macro Avg Recall (0.65) is low, showing that the model is **not good at detecting the minority class ("Yes").**

Weighted Avg Precision (0.88) and **Weighted Avg Recall (0.90)** are high because the "No" class is in the majority, affecting the overall results.

Classification Report 01:

	precision	recall	f1-score	support
no	0.92	0.97	0.94	1205
yes	0.58	0.33	0.42	152
accuracy			0.90	1357
macro avg	0.75	0.65	0.68	1357
weighted avg	0.88	0.90	0.89	1357



Model II: Decision Tree

Precision for "Yes" (0.41): This value indicates that out of all the customers predicted as "Yes," only 41% are actually registered customers. The model suffers from a high false positive rate, leading to inefficiency in targeting actual customers.

Recall for "Yes" (0.41): While recall for "Yes" is moderate, the model misses 41% of potential registered customers, which could result in lost marketing opportunities and revenue.

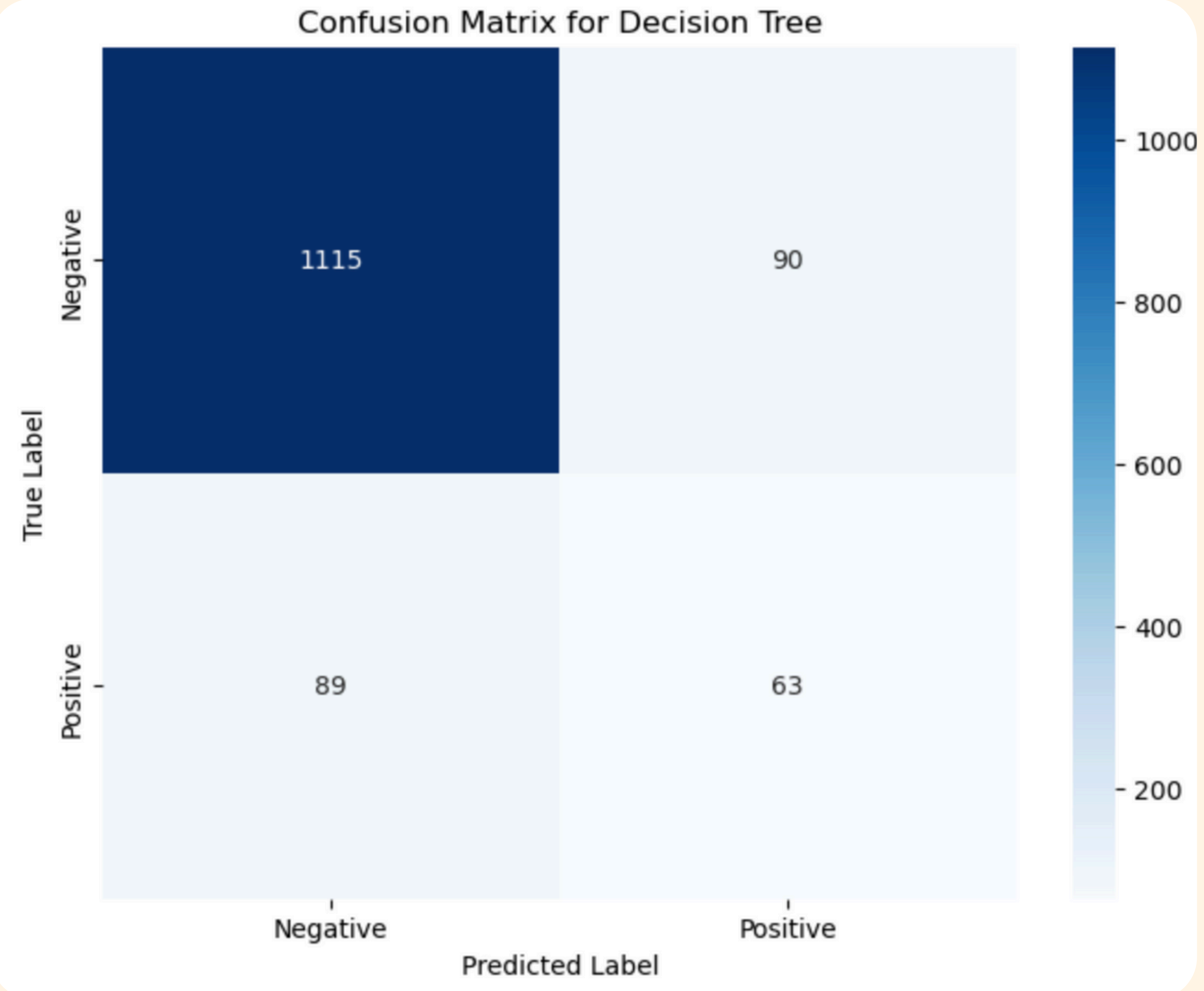
F1-Score for "Yes" (0.41): A low F1-Score demonstrates that the balance between precision and recall is poor for the minority class.

Accuracy (0.84): The accuracy is high but misleading due to class imbalance. Predicting most samples as "No" inflates accuracy without addressing the minority class effectively.

Macro Average Recall (0.67): This relatively low value highlights poor performance in identifying minority class instances.

Weighted Average Precision and Recall: High values for these metrics are skewed by the majority class performance and are not reliable indicators for minority class performance.

	precision	recall	f1-score	support
0	0.93	0.93	0.93	1205
1	0.41	0.41	0.41	152
accuracy			0.87	1357
macro avg	0.67	0.67	0.67	1357
weighted avg	0.87	0.87	0.87	1357



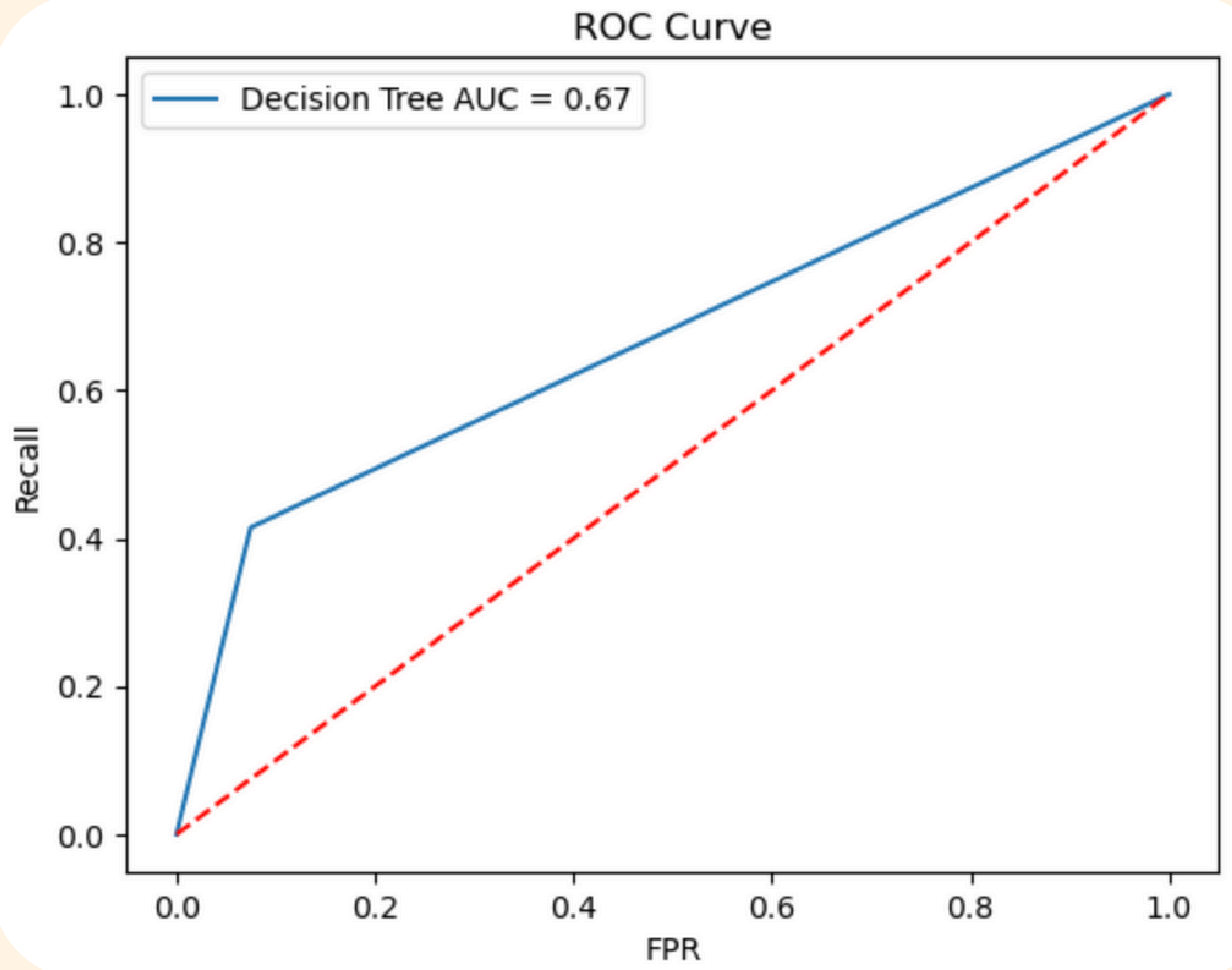
Model II: Decision Tree

The ROC curve shows an **AUC of 0.67**, indicating **limited** discriminative ability.

The curve's closeness to the diagonal suggests difficulty in separating classes, due to the dataset's **imbalance**.

To improve performance, we need to address class imbalance with techniques like **SMOTE** and **hyperparameter tuning**.

Evaluate metrics of outcome to analyze the impact of these technique on the result.



Optimize models

Feature engineering

SMOTE method

Optimize cut-off
(Model I)

Hyperparameter Tuning
(Model II)

Feature egineering

Feature Engineering is the **process of creating, transforming, or selecting features from raw data to improve the performance of a machine learning model**. It helps the **model better understand the relationships in the data,** thereby increasing accuracy and efficiency.

We create new features based on the dataset as follow:

- **Balance status**
- **Credit risk**
- **Contact times**
- **Contacted before**
- **Age group**
- **Previous outcome success**

Model I: Feature Engineering & Cut-off Optimized

When class 0 is the majority - 89%, (imbalanced dataset), keeping the default cut-off (0.5) will make the Logistic Regression model prioritize predicting class 0, so we need to find the optimal cut-off.

So we can reduce the cut-off point: It helps the model be easier to label class 1, increasing the ability to correctly detect class 1 samples (increasing Recall) and also reduces False Negative (FN), but may increase False Positive (FP).

Advantages:

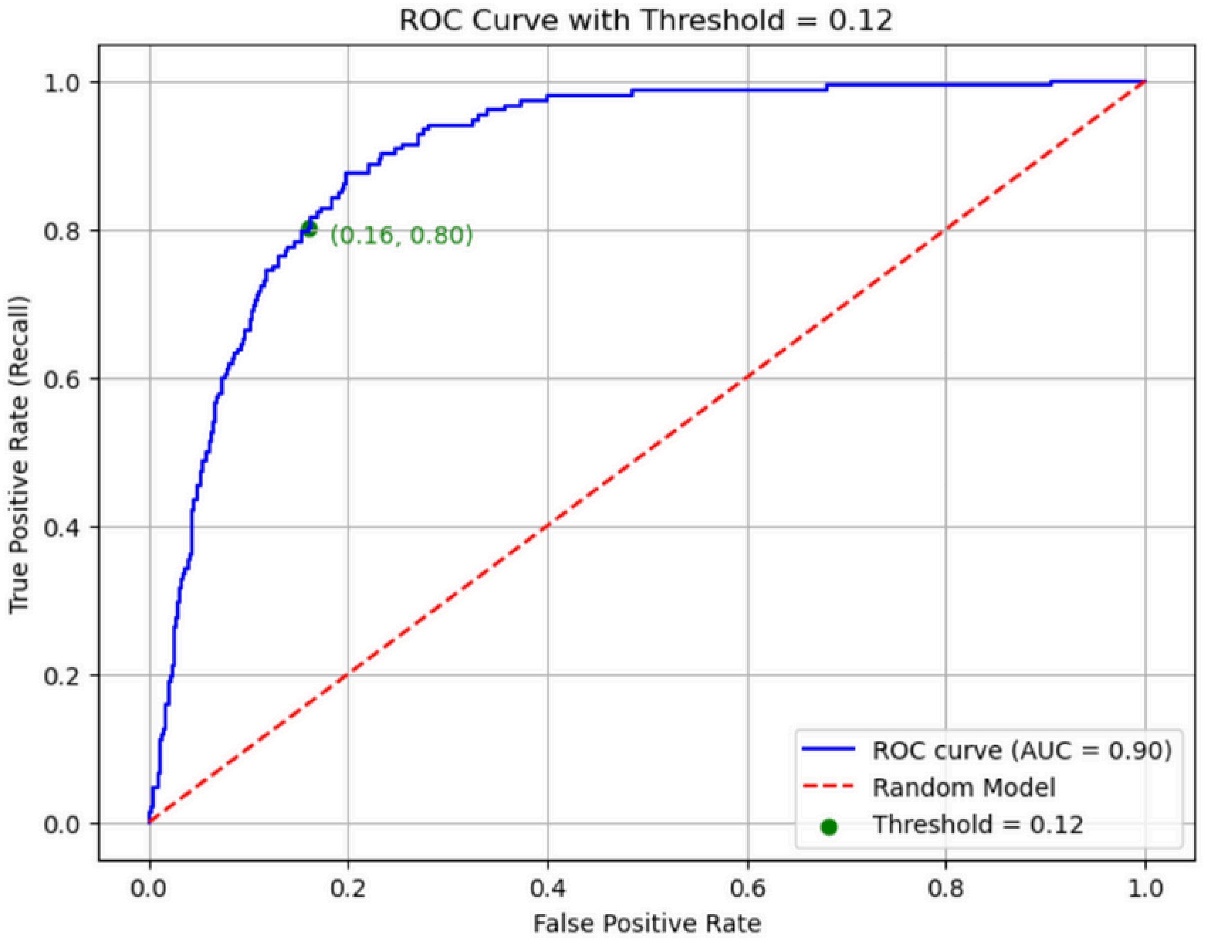
- Recall for class "Yes" increased sharply (0.33 → 0.80).
- F1-Score improved (0.42 → 0.52), meaning the overall model performs better in detecting registered customers.
- False Negatives decreased, helping the model miss fewer potential customers.

Disadvantages:

- Precision for class "Yes" decreased (0.58 → 0.39).
- This means the model reaches more non-potential customers, causing an increase in False Positives.

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.84	0.90	1205
1	0.39	0.80	0.52	152
accuracy			0.83	1357
macro avg	0.68	0.82	0.71	1357
weighted avg	0.91	0.83	0.86	1357



Model II: Feature Engineering

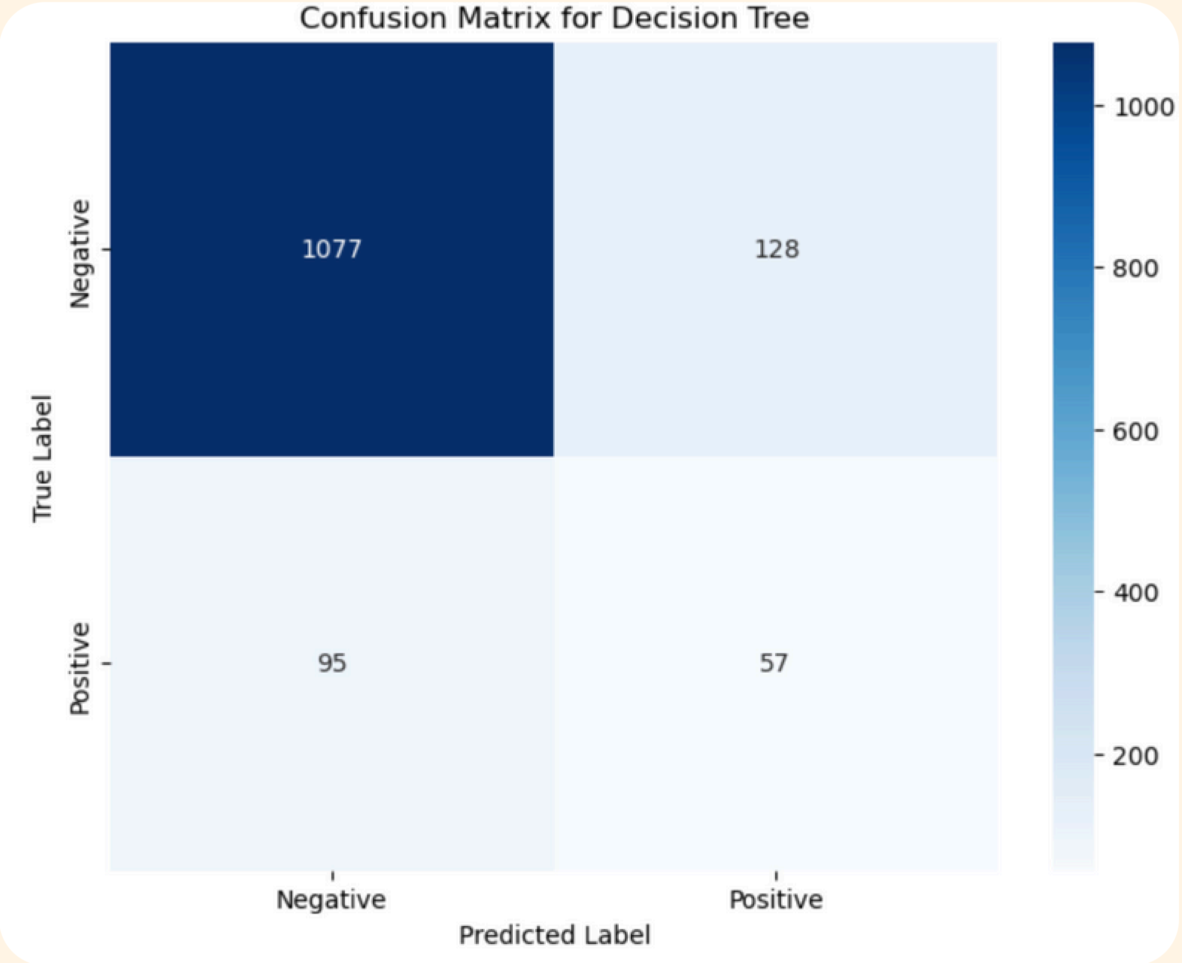
Disadvantage:

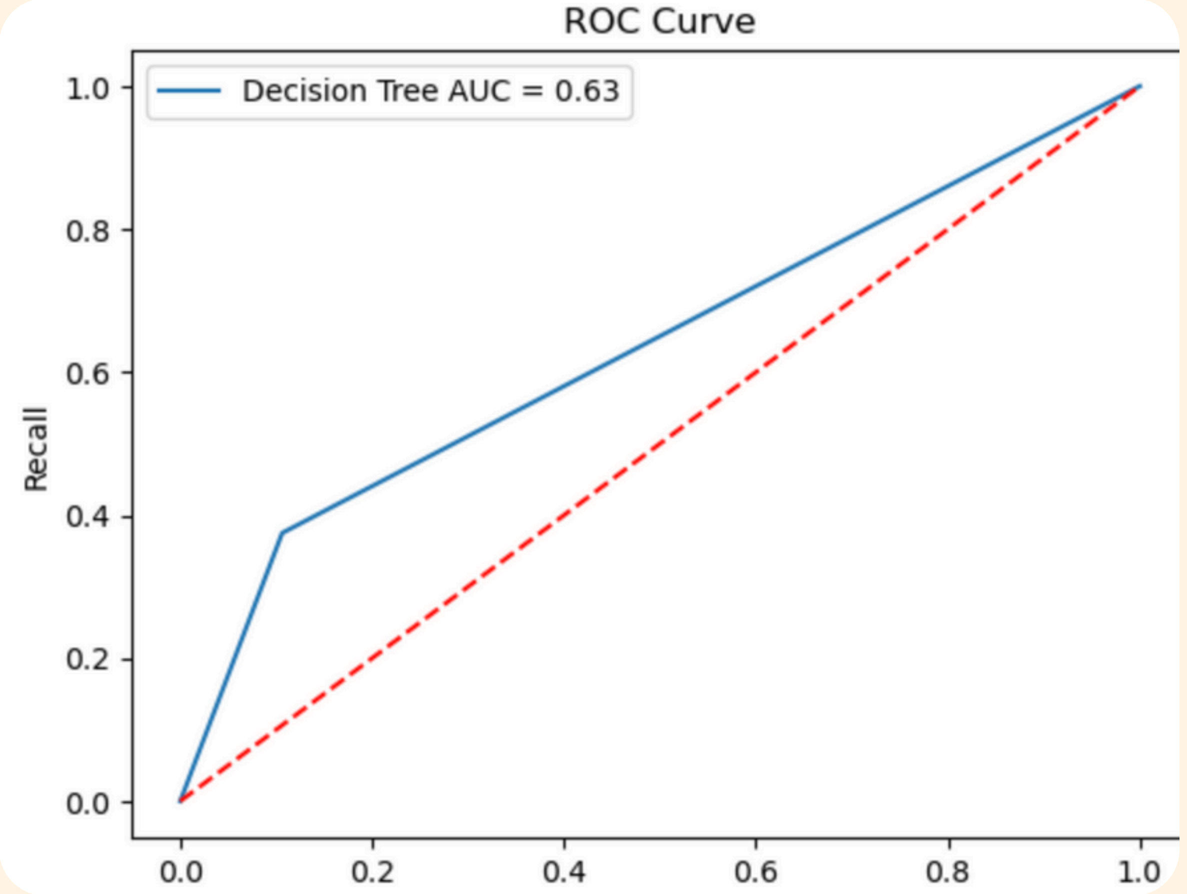
- Recall for class "Yes" decrease (0.41 → 0.31).
- Precision for class "Yes" decreased (0.41 → 0.31).
- F1-Score decresed (0.41 → 0.34), meaning the overall model performs worst in detecting registered customers.
- False Negatives increased, make the model miss more potential customers.

Advantage:

- Precision for class "No" increased slightly(0.92 → 0.93).

	precision	recall	f1-score	support
0	0.92	0.89	0.91	1205
1	0.31	0.38	0.34	152
accuracy			0.84	1357
macro avg	0.61	0.63	0.62	1357
weighted avg	0.85	0.84	0.84	1357





SMOTE method

SMOTE (Synthetic Minority Over-sampling Technique) is a technique to **deal with imbalanced data by generating additional simulated data for the minority class.** Instead of simply replicating the minority class samples like the random oversampling method, SMOTE generates new data samples based on interpolation between existing data points in the minority class.

The main goal of SMOTE is to **reduce the imbalance between the majority class and the minority class to improve the model performance,** especially metrics such as Recall and F1-Score.

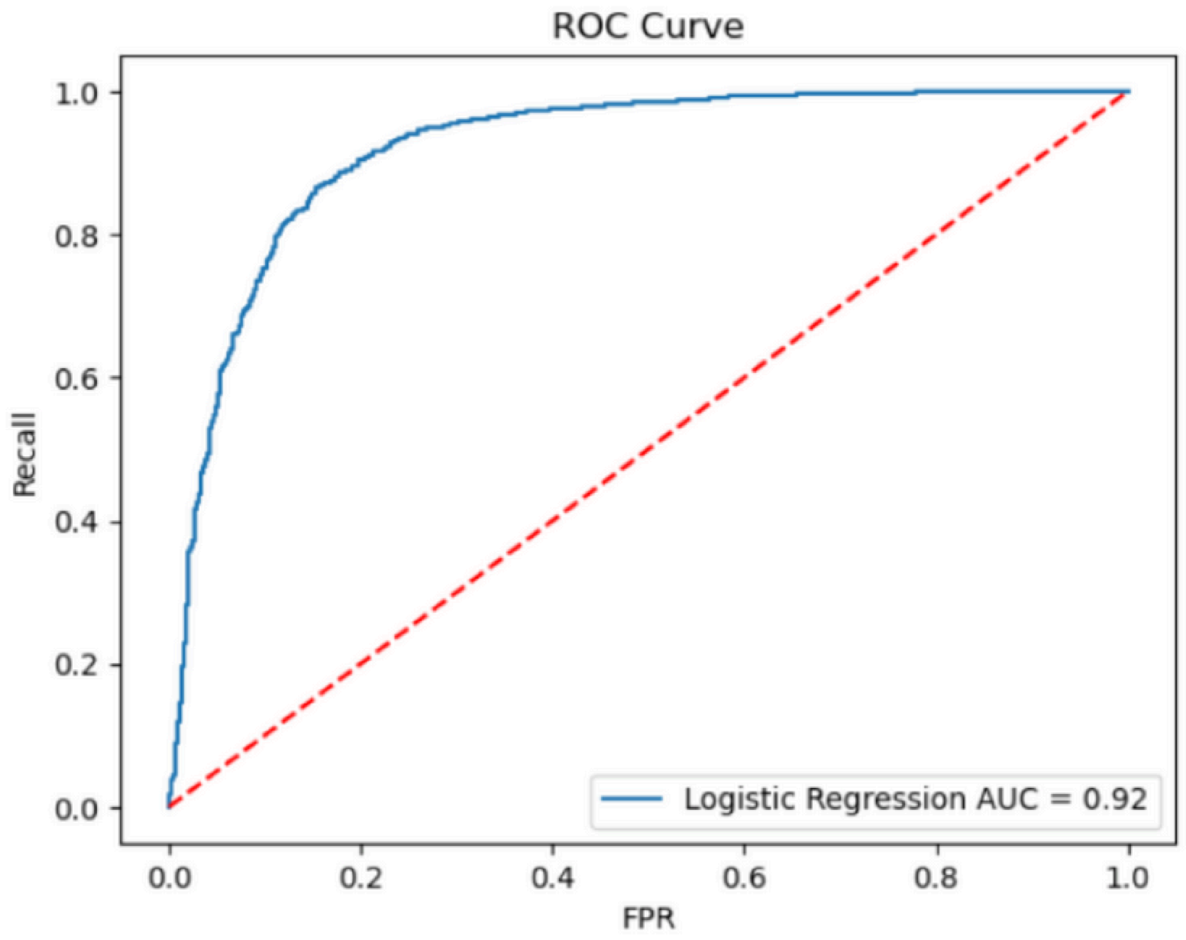
```python
from collections import Counter

counter = Counter(y)
print('Before', counter)

# oversampling the train dataset using SMOTE
smt = SMOTE()
X_sm, y_sm = smt.fit_resample(X, y)

counter = Counter(y_sm)
print('After', counter)
```

```
Before Counter({0: 4000, 1: 521})
After Counter({0: 4000, 1: 4000})
```

Model I: SMOTE

	precision	recall	f1-score	support
no	0.87	0.83	0.85	1206
yes	0.83	0.87	0.85	1194
accuracy			0.85	2400
macro avg	0.85	0.85	0.85	2400
weighted avg	0.85	0.85	0.85	2400

- **Precision (Yes) = 0.83: Significant increase,** meaning the model is more accurate in predicting customers who sign up.
- **Recall (Yes) = 0.87: Significant increase,** detecting 87% of customers who actually sign up.
- **F1-Score (Yes) = 0.85: Significant improvement,** showing the model has a good balance between Precision and Recall.
- **Accuracy = 0.85: Decreased** from baseline, but still high
- **AUC = 0.92: Increased** marginally



⟶ SMOTE **helps the model balance Precision and Recall**, significantly improving over both previous models. The model reduces False Negatives (FN) without increasing False Positives (FP) too much.

Comparison

Method	Advantages	Disadvantages
Baseline dataset	High Accuracy	Low Recall, losing potential customers
Feature Engineering + Optimal Threshold	Recall improved significantly with lower False Negative	Low Precision
SMOTE	Good balance between Precision and Recall, F1-score improved. Got the highest AUC	Slight decrease on Accuracy, get more False Positive

The SMOTE model gives the best results with both high Precision and Recall, especially the F1-Score for the "Yes" class reaching 0.85. As well as the highest AUC (0.92) compared to the other two datasets.

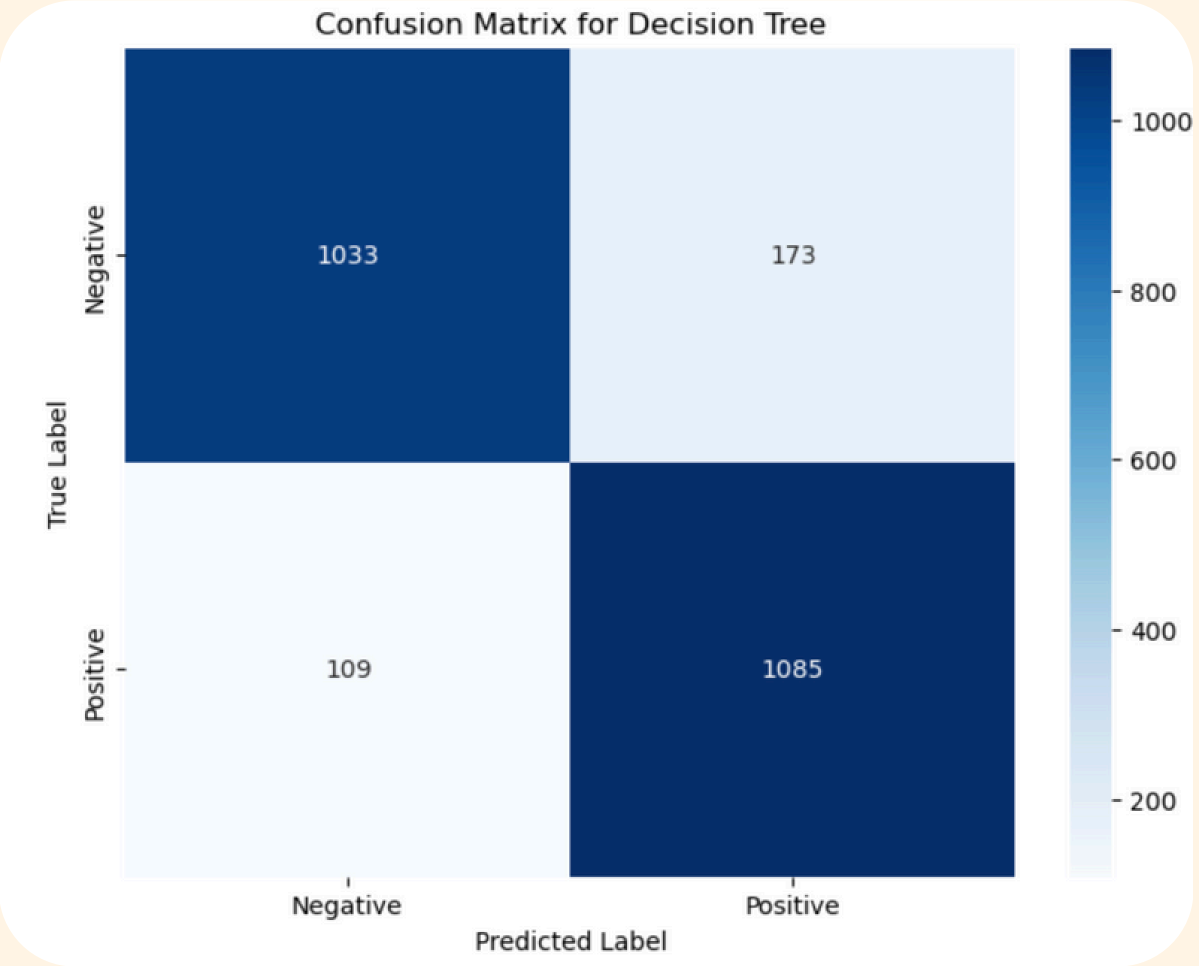
Model II: SMOTE

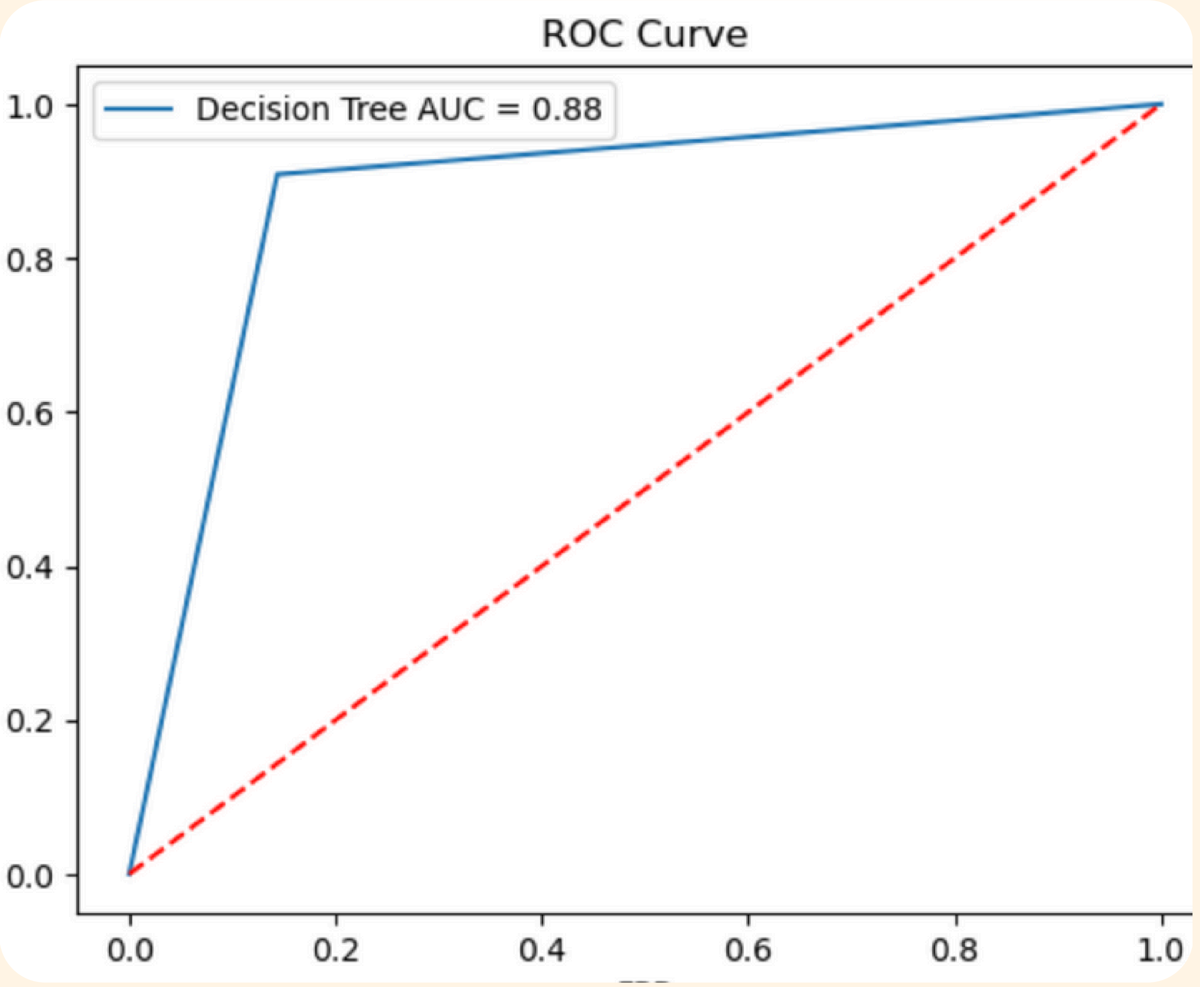
- **Precision (Yes) = 0.86: Significant increase,** meaning the model is more accurate in predicting customers who sign up.
- **Recall (Yes) = 0.91: Significant increase,** detecting 87% of customers who actually sign up.
- **F1-Score (Yes) = 0.88: Significant improvement,** showing the model has a good balance between Precision and Recall.
- **Accuracy = 0.88: Significant improvement** from baseline
- **AUC = 0.92: Significant improvement** from 0.63

SMOTE **helps the model balance Precision and Recall**, significantly improving over both previous models. The model reduces False Negatives (FN) significantly, overall improvement.

→

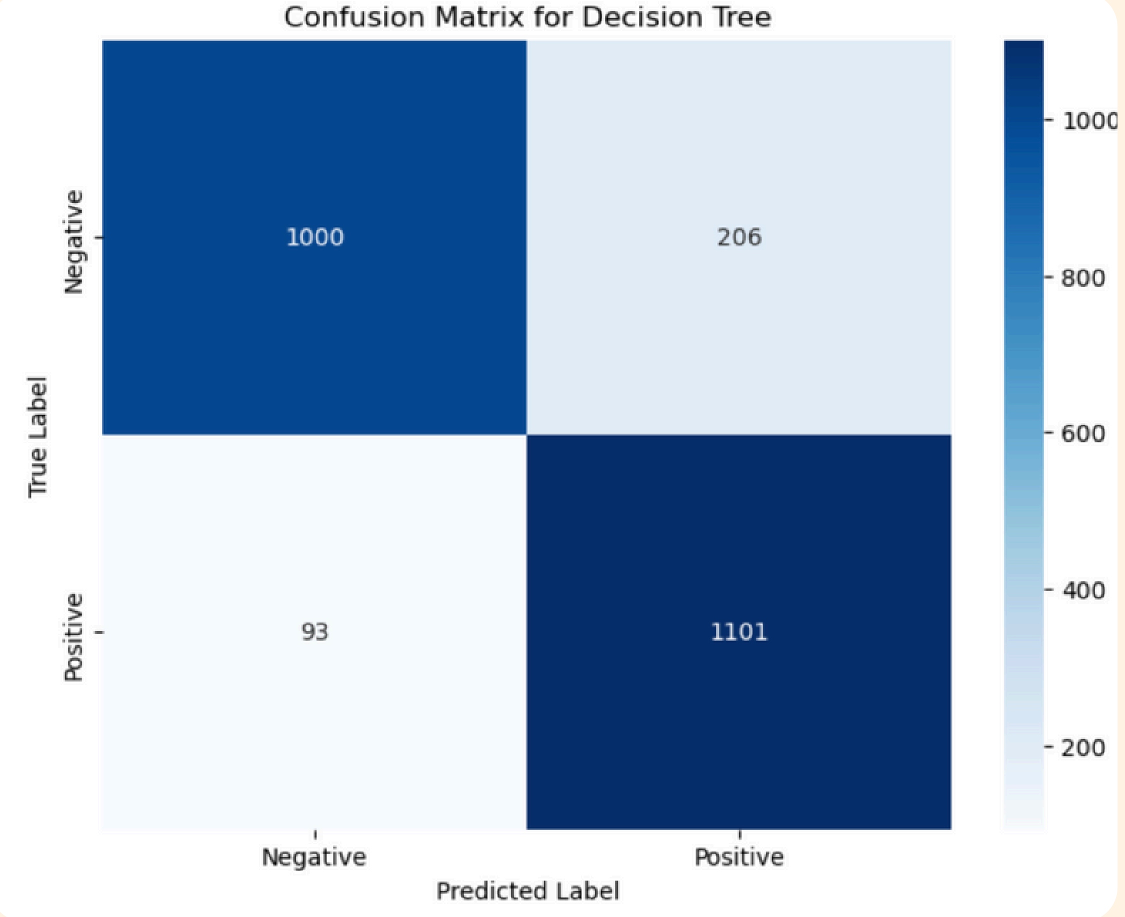
	precision	recall	f1-score	support
0	0.90	0.86	0.88	1206
1	0.86	0.91	0.88	1194
accuracy			0.88	2400
macro avg	0.88	0.88	0.88	2400
weighted avg	0.88	0.88	0.88	2400





Model II: SMOTE + Hyperparameters tuning

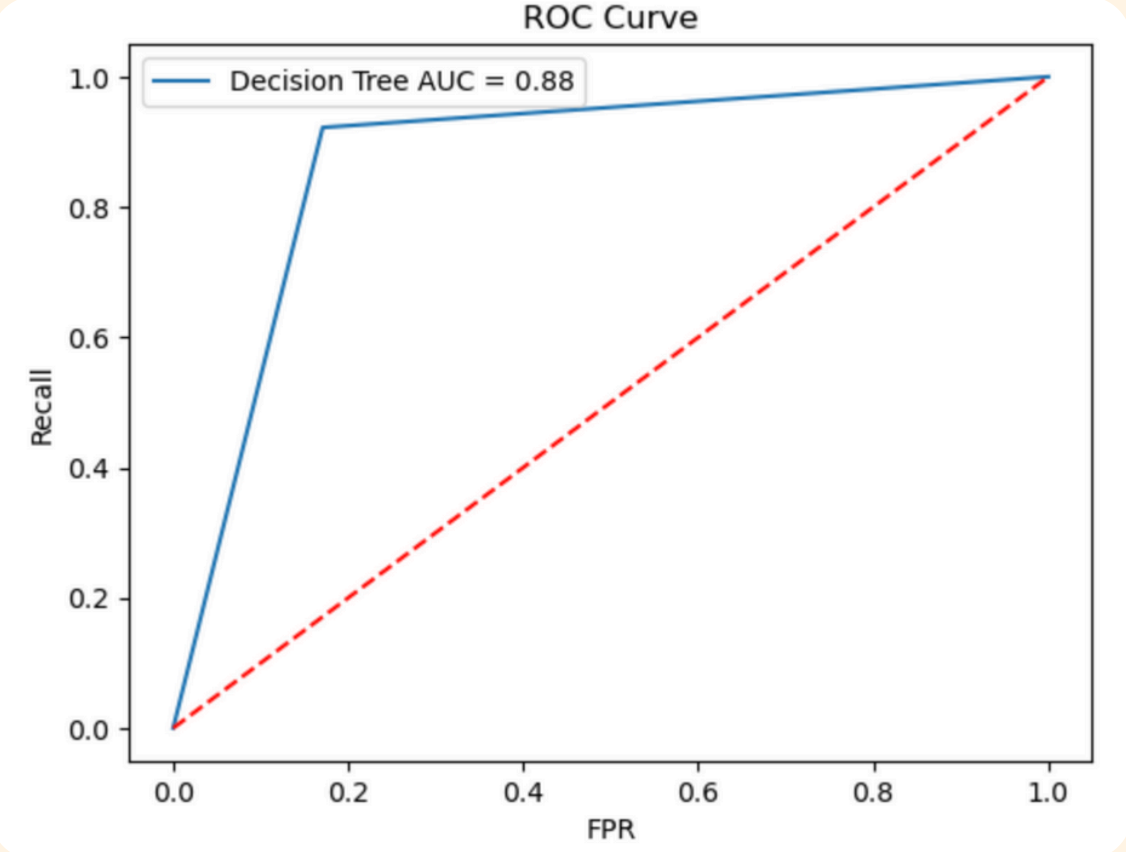
- **Precision (Yes) = 0.84: Slight decrease,** meaning the model is less accurate in predicting customers who sign up.
- **Recall (Yes) = 0.92: Significant increase,** detecting 92% of customers who actually sign up.
- **F1-Score (Yes) = 0.85: Decrease from 0.88,** showing the model has a less balance between Precision and Recall compared to before.
- **Accuracy = 0.88: Stay the same**
- **AUC = 0.88: Stay the same**



Hyperparameter tuninng with the target for **Recall**, significantly improving Recall as the goal is to detect customers who actually sign up.

→

	precision	recall	f1-score	support
0	0.91	0.83	0.87	1206
1	0.84	0.92	0.88	1194
accuracy			0.88	2400
macro avg	0.88	0.88	0.88	2400
weighted avg	0.88	0.88	0.88	2400

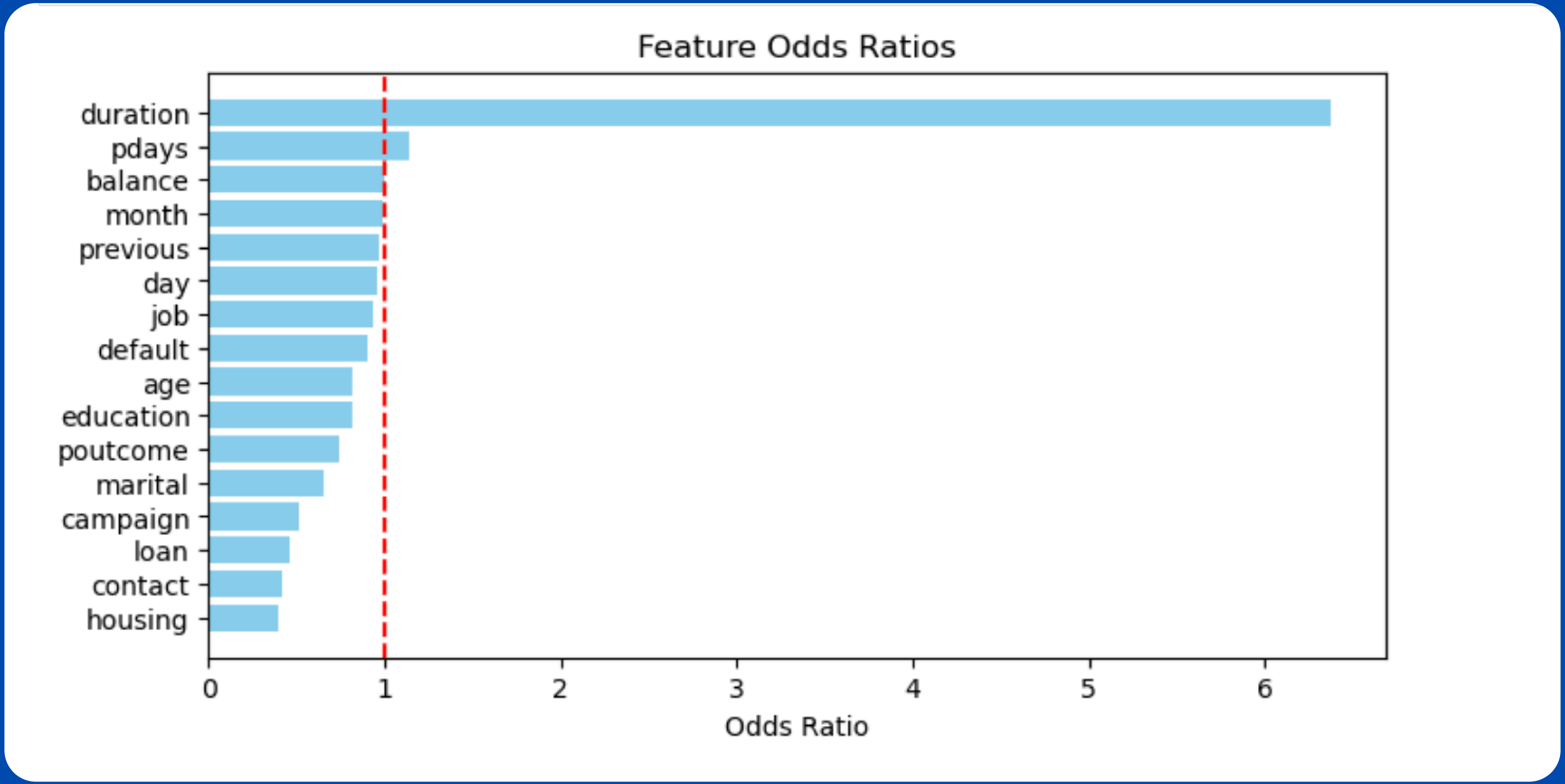


Comparison

Method	Advantages	Disadvantages
Baseline Model	- Simple implementation. - Decent precision for the "No" class (e.g., 0.93).	- Poor recall for the positive class (subscribed clients). - Many false negatives, missing subscribers. - Low overall F1-score and AUC.
Feature Engineering Only	Precision for class "No" increased slightly(0.92 → 0.93).	Worst overall on most metric
SMOTE + Tuning	- Significantly improved recall for the positive class. - Increased F1-score and AUC.	- Slight decrease in precision for the positive class compared to just SMOTE

The SMOTE model gives the best results with Precision, SMOTE and Tunning give the best result for Recall. So it depend on the bank strategy to employ which strategy.

Features Importants to the Dataset measured by Odds Ratio



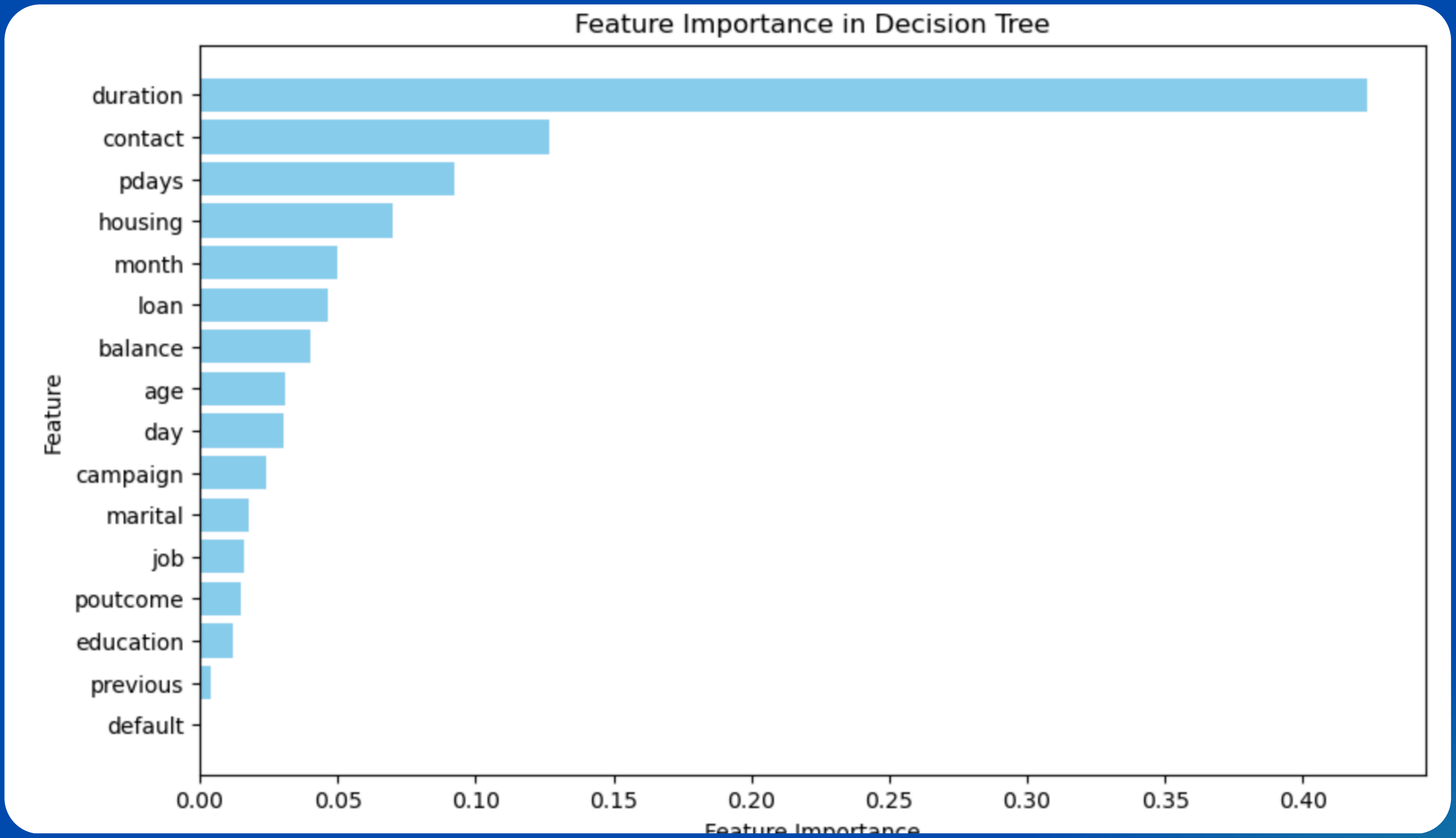
Most important feature:

- **duration** (6.37), showing that **call duration is the most important factor** in deciding whether customers will sign up for the term deposit package or not.
- **pdays is a positive feature,** with Odds Ratio = 1.13. This shows that if a customer has been contacted in advance, the likelihood of them signing up increases slightly.

Most influential campaign-related factors (Model II)

From the feature importances, **campaign-related factors clearly dominate the model II's decisions**. Specifically top 3:

1. **Duration of the last contact (duration_group)**: This is by far the most influential factor. Clients who stay on the call longer tend to be more receptive and ultimately more likely to subscribe.
2. **Contact method (contact**): How the client was reached (cellular vs. telephone, etc.) also plays a role.
3. **Outcome of previous marketing contacts (poutcome):** Whether a client's previous campaign resulted in success, failure, or no contact at all also heavily influences their likelihood of subscribing this time.



Models comparision

Metrics	Logistic Regression	Decision Tree
Precision	0.83	0.84
Recal	0.87	0.92
F1-Score	0.85	0.88

CONCLUSION

1. **Key Factor:**
   - Both models identify **duration** as the most critical factor in determining whether a client will subscribe to a term deposit, highlighting the importance of the length of the last contact.
2. **Logistic Regression Insights:**
   - **Pdays** (recency of prior contact) and **balance** (financial stability) are also influential, aligning with the decision tree findings.
   - Logistic regression emphasizes previous interactions and day of the month as additional important factors.
3. **Demographics:**
   - While **age** and **job** have moderate significance in logistic regression, demographic features generally play a smaller role compared to campaign-related factors.
4. **Campaign Features:**
   - Campaign-specific elements, such as **contact type** and **number of contacts** (campaign), show less importance in logistic regression compared to decision trees.

Conclusion:
Campaign-related features like **duration**, **pdays**, and **balance** are consistently influential across both models, making them key focus areas for improving subscription rates. Logistic regression provides additional insights into the importance of **previous interactions** and **timing** (day of the month). This highlights the need to prioritize both effective communication and strategic timing in marketing campaigns.

Thank you for listening!