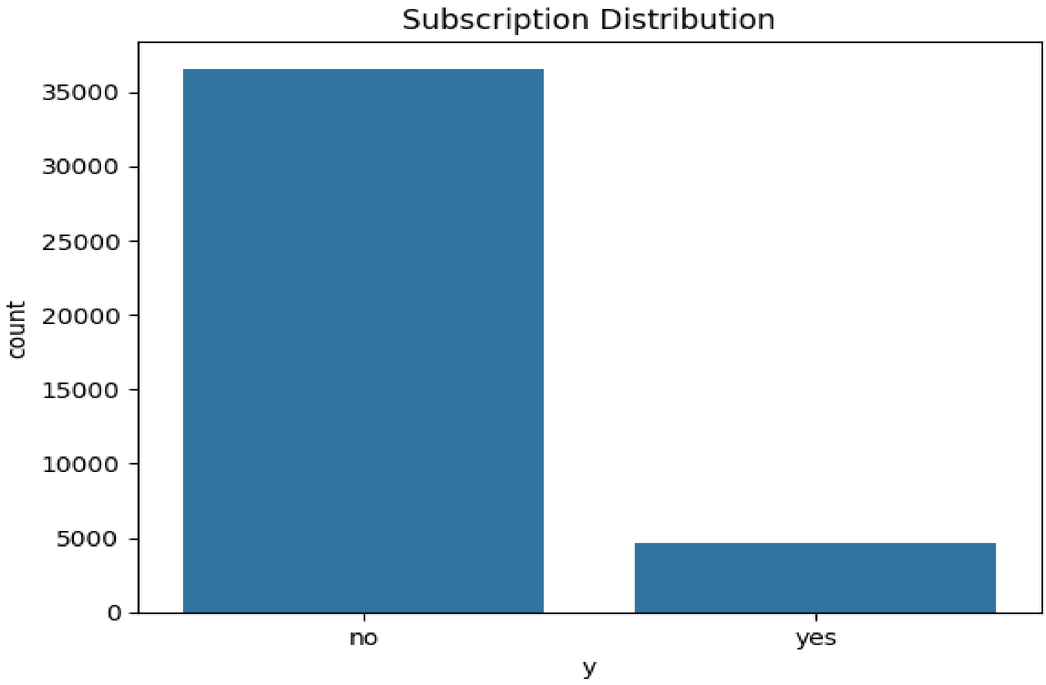
This is a detailed report on every step that was taken to make this assessment project successful.
The aim of this project is to predict whether a client will subscribe to a term deposit. I decided to perform every task online; from uploading the dataset to deploying my model.

Tools Used

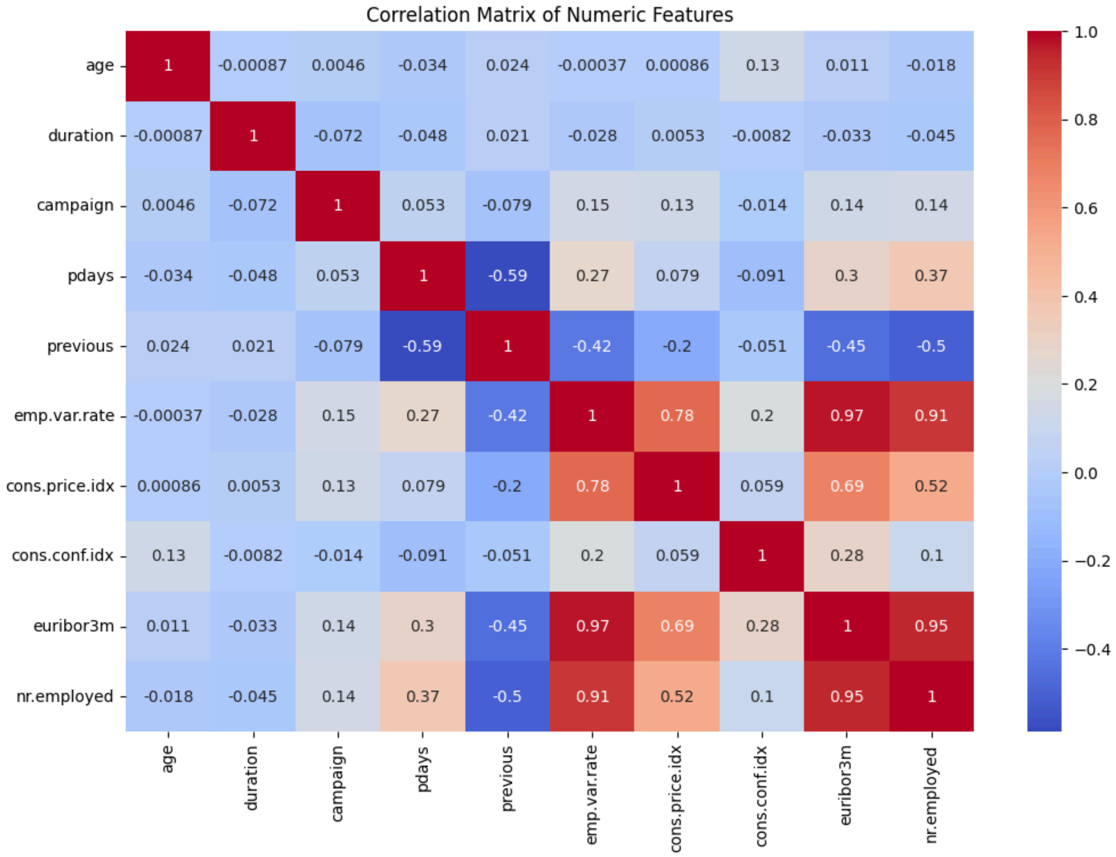
Process	Platform
Data cleaning and EDA	Google Collab
Modeling and Coding	Google Collab
Report	Google Docs
Code and Hosting	GitHub
Model Hosting	Streamlit Cloud

EDA Findings

1. **Target Variable** - Using the bank-additional-full.csv dataset, there was a significant data imbalance as there were more people that did not subscribe (**36548**) than people that subscribed(**4640**)



2. **Numerical feature Correlation**



Correlation Matrix of Numeric Features

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
age	1	-0.00087	0.0046	-0.034	0.024	-0.00037	0.00086	0.13	0.011	-0.018
duration	-0.00087	1	-0.072	-0.048	0.021	-0.028	0.0053	-0.0082	-0.033	-0.045
campaign	0.0046	-0.072	1	0.053	-0.079	0.15	0.13	-0.014	0.14	0.14
pdays	-0.034	-0.048	0.053	1	-0.59	0.27	0.079	-0.091	0.3	0.37
previous	0.024	0.021	-0.079	-0.59	1	-0.42	-0.2	-0.051	-0.45	-0.5
emp.var.rate	-0.00037	-0.028	0.15	0.27	-0.42	1	0.78	0.2	0.97	0.91
cons.price.idx	0.00086	0.0053	0.13	0.079	-0.2	0.78	1	0.059	0.69	0.52
cons.conf.idx	0.13	-0.0082	-0.014	-0.091	-0.051	0.2	0.059	1	0.28	0.1
euribor3m	0.011	-0.033	0.14	0.3	-0.45	0.97	0.69	0.28	1	0.95
nr.employed	-0.018	-0.045	0.14	0.37	-0.5	0.91	0.52	0.1	0.95	1

- ‘emp.var.rate’, ‘euribor3’`, and ‘nr.employed' are **extremely tightly linked (0.91 -0.97 correlation)**. When one moves, the others follow.
- This makes sense as employment rates, interest rates, and economic stability often shift together.
- Since they’re so similar, we might only need one (like `euribor3m`) or a combined version to avoid redundancy.
- ‘cons.conf.idx’ has a **moderate negative correlation (~ -0.40 to -0.45)** with those economic indicators.
- When employment and interest rates improve, consumer confidence tends to go down or vice versa. People get nervous when the economy changes.

- ‘pdays’ (days since last contact) and ‘previous’ (number of past contacts) are **moderately linked (-0.59)**.
- This makes sense because if someone was contacted many times (`previous` is high), they were probably reached more recently (`pdays` is low).
- Why this matters: It helps us understand client touchpoints and engagement patterns.

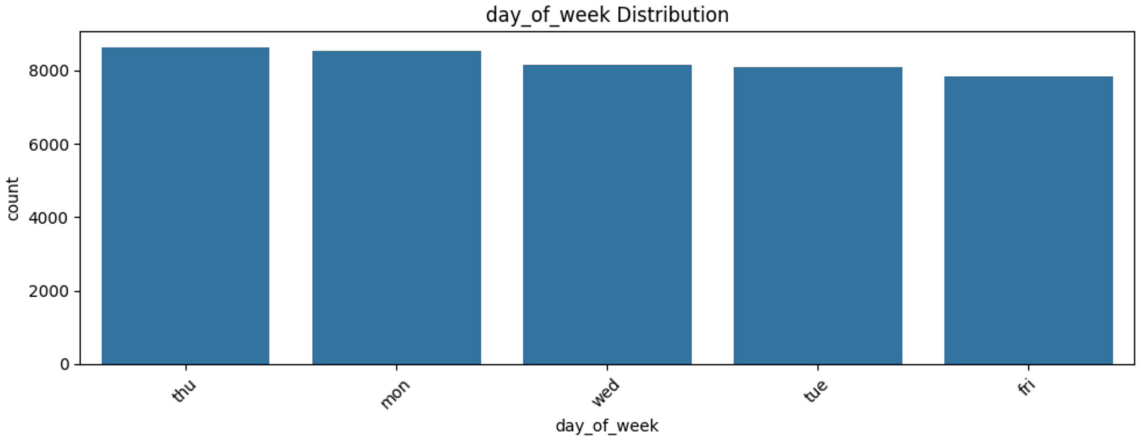
- Call Duration - ‘duration’ (length of last call) has almost no correlation with anything else. It’s a powerful signal for whether someone subscribes, but since we only know it after the call, it’s great for analysis, not for pre-call targeting.

- Age- ‘age’ doesn’t really correlate with other features

3. Categorical Distribution

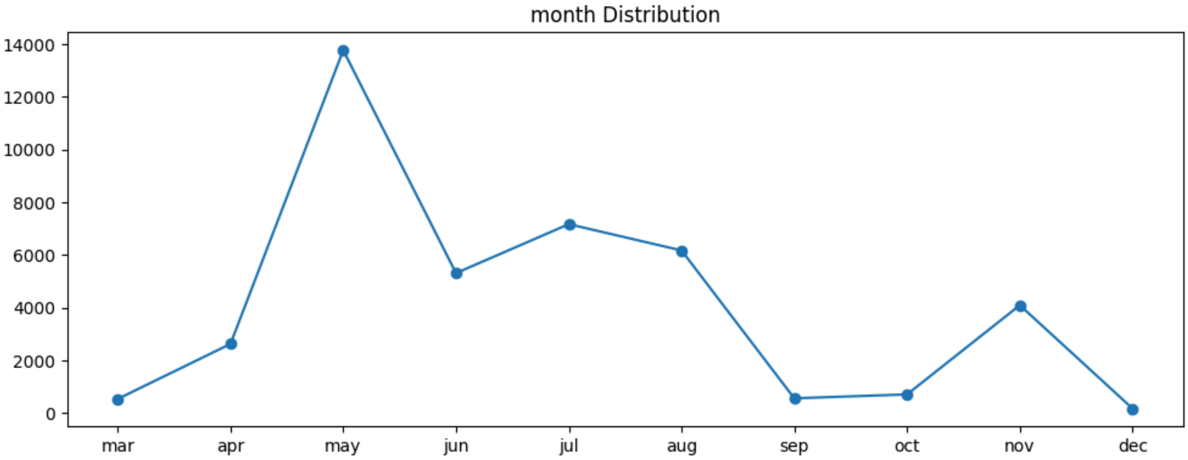
Daily Contact Trends

- **Observation:** Call volume is fairly consistent across weekdays, with Monday and Thursday being the busiest (8,500 calls each).
- **Possible Implication:** While daily call volume is similar, it's crucial to investigate if certain days are more effective. For example, conversion rates might be higher on Thursdays than on Mondays



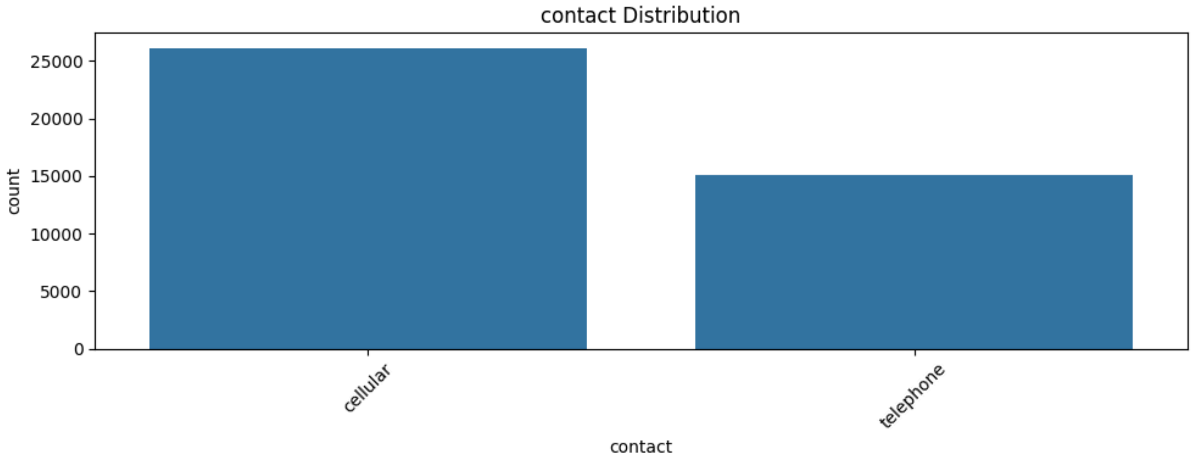
Monthly Contact Seasonality

- **Observation:** The data heavily concentrates calls in May (14K calls), with July and August showing secondary peaks. Other months, like October and March, have significantly fewer calls (1K).
- **Possible Implication:** This likely reflects either a strategic focus during historically successful periods or resource constraints (staff/budget) that vary by season.



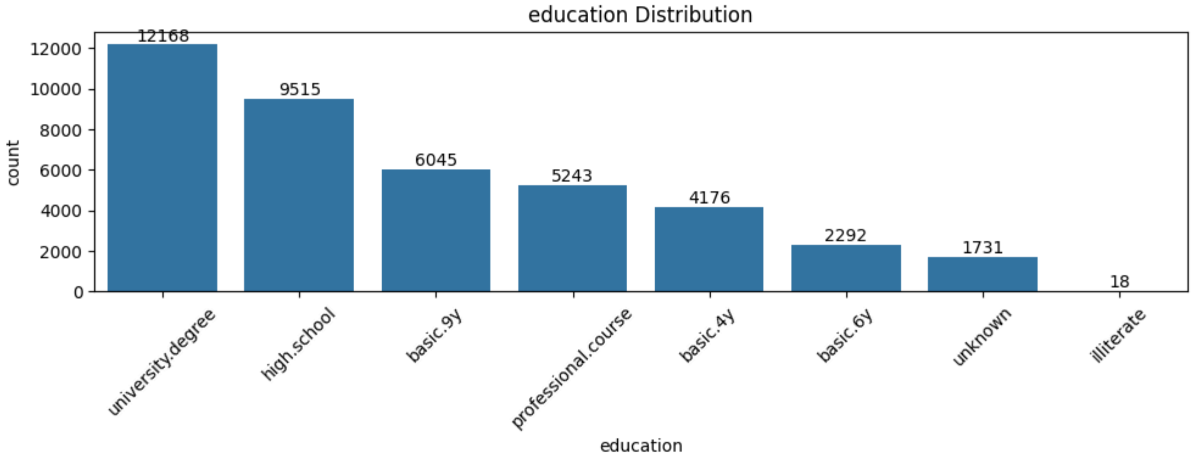
Contact Method

- **Observation:** Cellular calls (26K) significantly outnumber landline calls (15K).
- **Possible Implication:** Mobile outreach is clearly the primary method, likely due to customer accessibility. However, it's important to assess if mobile calls also deliver better results.



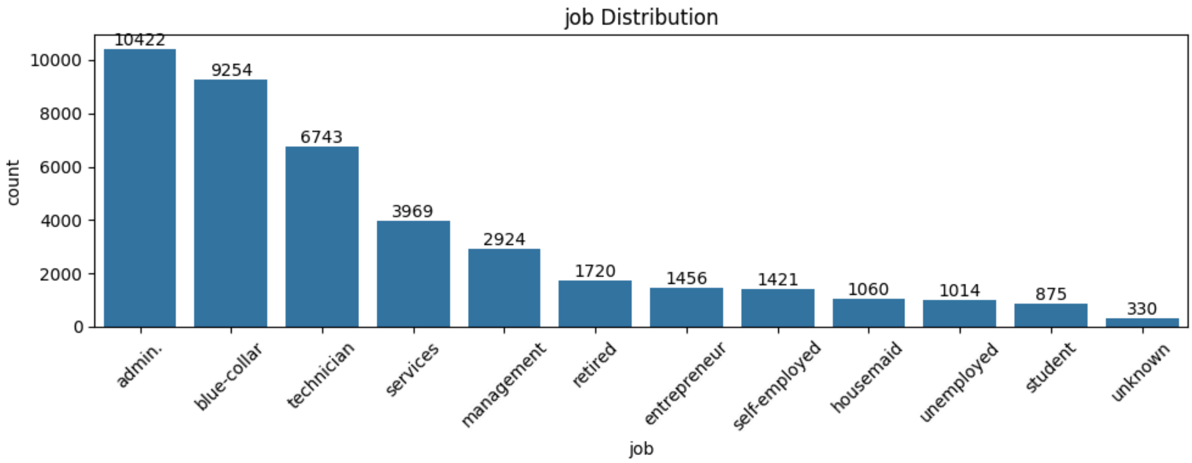
Education Level of Clients

- **Observation:** The majority of clients hold a university degree (12K) or a high school diploma (9.5K). There are very few "unknown" or "illiterate" entries.
- **Possible Implication:** An educated audience may respond differently to marketing messages, perhaps favoring analytical over emotional appeals.



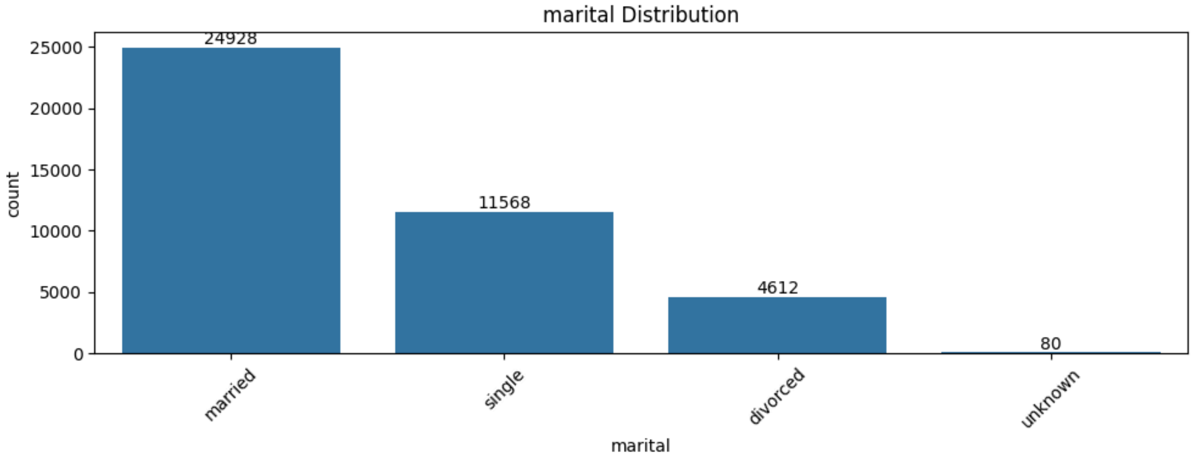
Client Job Types

- **Observation:** "Admin" and "blue-collar" workers are the most frequently contacted groups (9K each), with "technicians/services" close behind. Students and unemployed individuals represent a small proportion.
- **Implication:** Job types offer insights into income and financial needs. For example, blue-collar workers might prioritize liquidity differently than a student, which could influence product pitches.



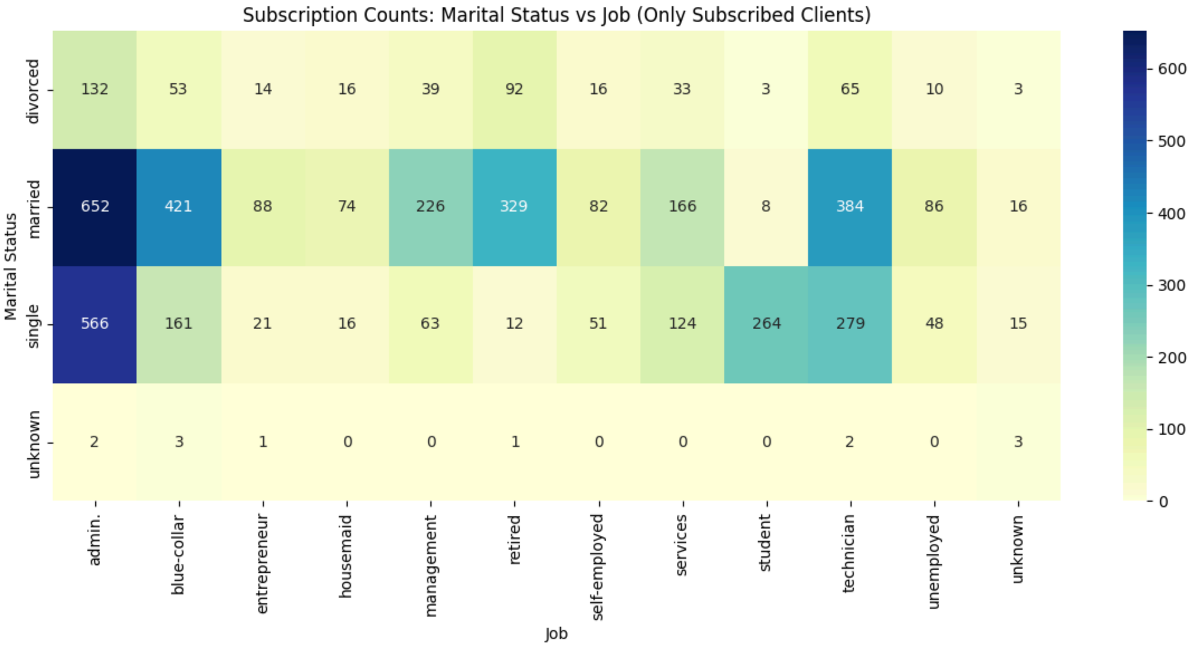
Marital Status Distribution

- **Observation:** Married clients constitute the largest group (25K), followed by single (12K) and divorced (4.5K) clients.
- **Possible Implication:** Marital status often correlates with financial priorities (mortgages, family savings, etc). If single clients show higher conversion rates, tailoring messaging to this segment could be effective.



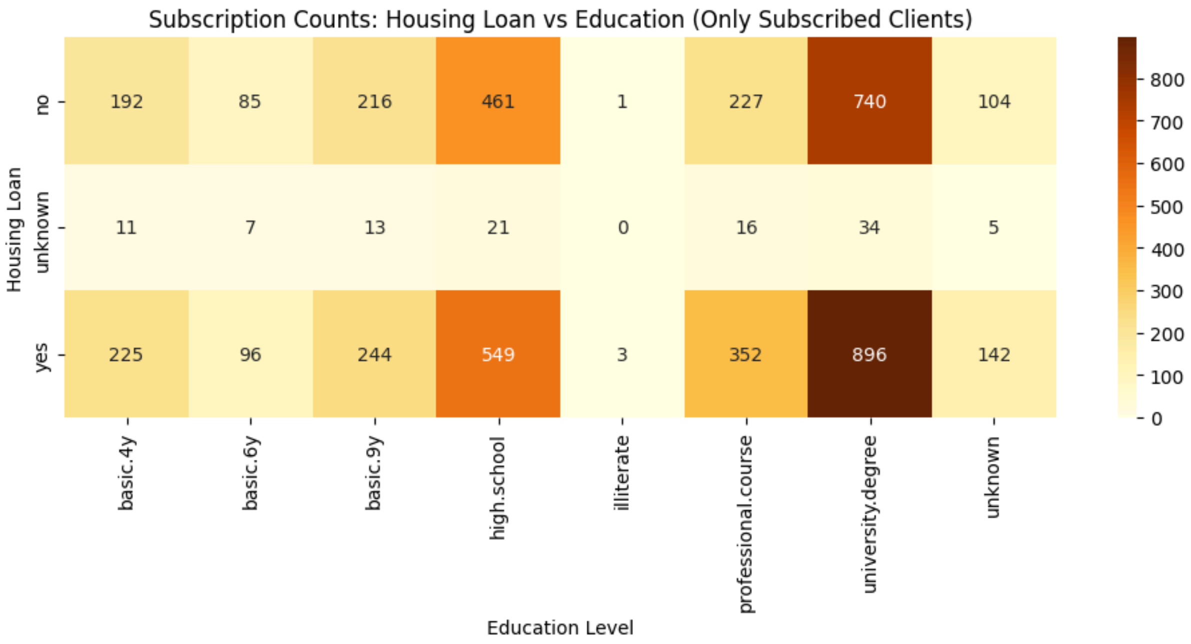
Heatmaps: Marital Status vs. Job (Subscribers Only)

This heatmap compares the marital status and job types of clients who subscribed to the term deposit. The highest subscription counts are for admin ,blue-collar and technician jobs.Some job categories like housemaid and student have very few subscribers. Marital status also plays a role, with certain groups showing higher subscriptions.

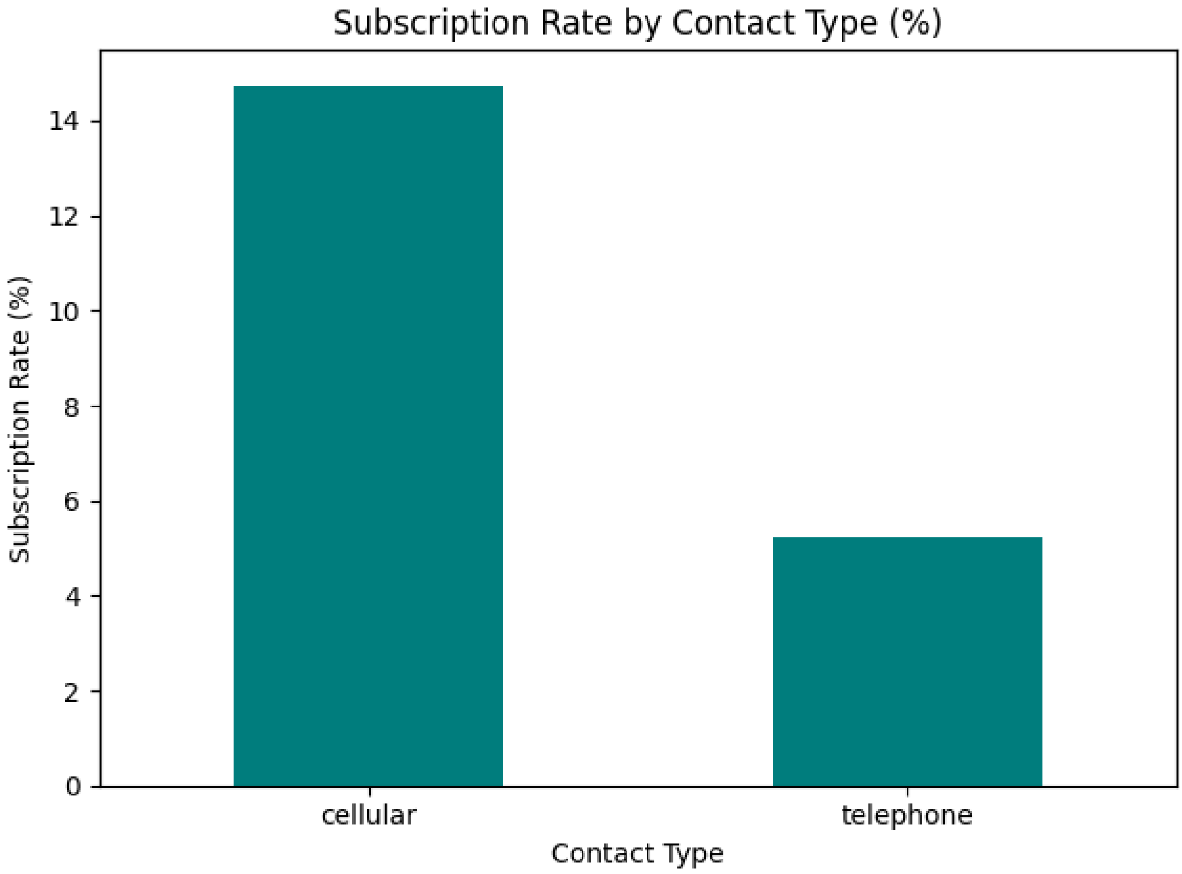


Heatmap: Housing Vs Education (Subscribers Only)

This heatmap compares the housing and education of clients who subscribed to the term deposit. Clients with a higher education and had housing loans were the most subscribers.Those with unknown housing loan status had fewer subscriptions.

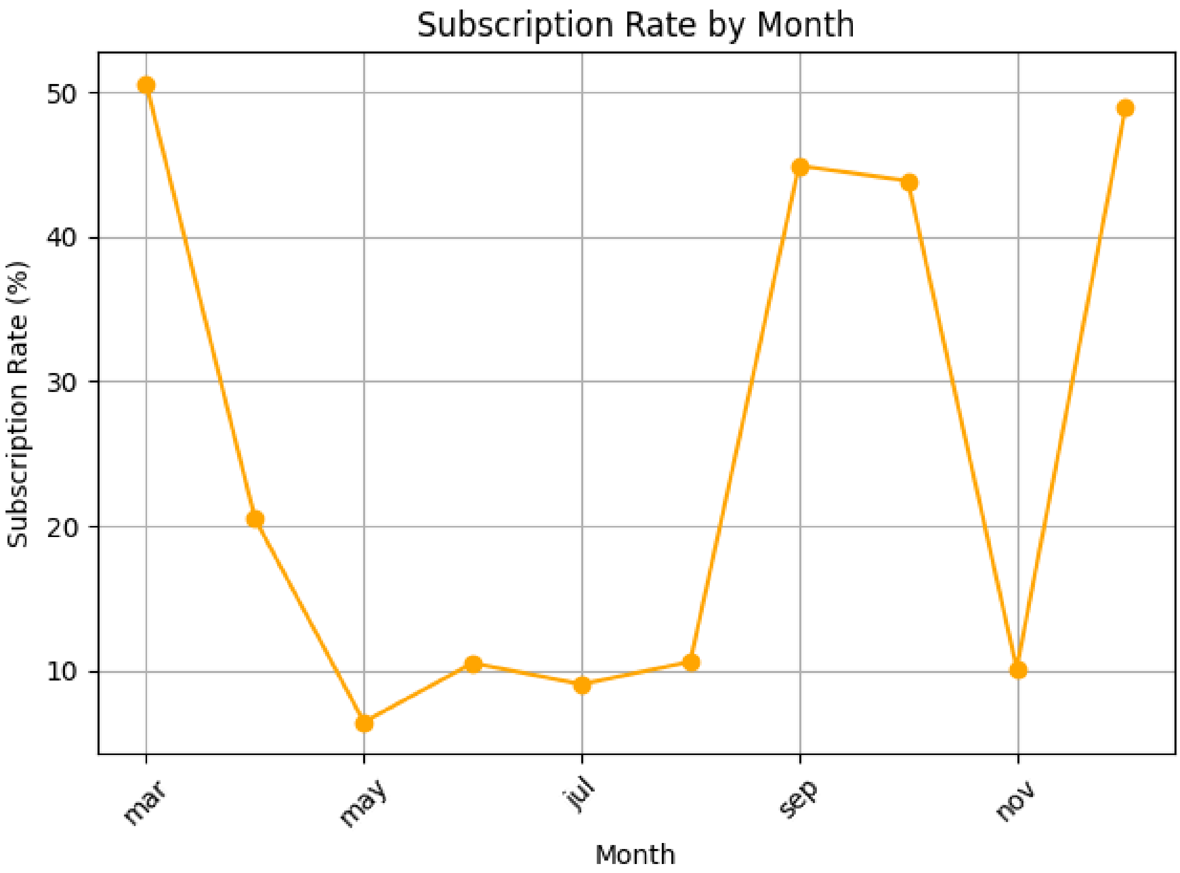


Subscription Rate by Contact Type :Campaigns conducted via cellular phones are nearly three times more effective in securing subscriptions than those conducted via traditional landline telephones.



Subscription Rate by Month : March and December show exceptionally high subscription rates, both around 50%. September and October also have high rates, around 45% and 43% respectively.

May has the lowest subscription rate, dipping to around 6%. July and November also show relatively low rates (around 9% and 10% respectively).

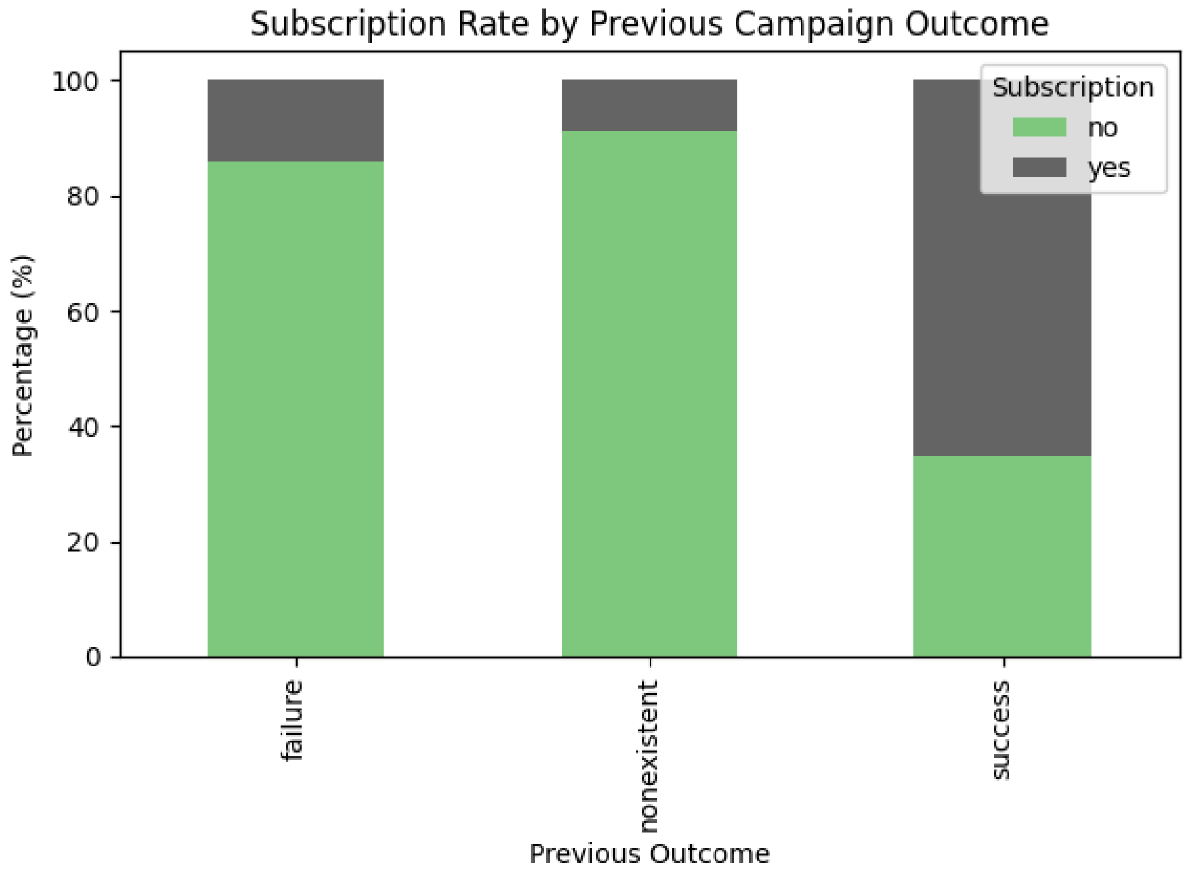


Subscription Rate by Previous Campaign Outcome: The stacked bar chart shows the percentage of 'no' (green) and 'yes' (gray) subscriptions based on the outcome of previous marketing campaigns for a client.

- **'success':** Clients with a previous campaign 'success' have an overwhelming subscription rate, with approximately 65% subscribing again (gray portion).
- **'nonexistent':** Clients for whom there was no previous campaign outcome show a

very low subscription rate, with around 7% subscribing.

- **'failure':** Clients who failed to subscribe in a previous campaign also have a low subscription rate, with about 15% subscribing.



DATA PRE-PROCESSING

To prepare the dataset for modeling, a series of preprocessing steps were carried out to ensure that the data was well-structured and machine learning ready. This includes encoding categorical variables, feature scaling, and addressing class imbalance.

- **Encoding categorical Variables**

The dataset contains multiple categorical features (job, marital, education, default, housing, loan, contact, month, day_of_week, and poutcome.)

To convert these into numeric format: Binary features like y (target), default, housing, and loan were label encoded (e.g., yes = 1, no = 0).

Multi-class categorical variables were transformed using one-hot encoding, with drop_first=True to prevent multicollinearity (dummy variable trap).

- **Feature Scaling**

  To ensure that numerical features contributed equally to the model, StandardScaler was applied to normalize continuous variables (age, duration, campaign, pdays, previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed).

  This scaling brings all values to a standard normal distribution with mean 0 and standard deviation 1.

- **Train-Test Split**

  To evaluate model performance reliably, the dataset was split into 80% training and 20% test sets.The split was stratified on the target variable (y) to maintain the same proportion of subscribers (yes) and non-subscribers (no) in both sets.

- **Class Imbalance Handling**

  EDA revealed that the target variable was imbalanced, with significantly fewer clients subscribing to a term deposit. To mitigate bias, SMOTE (Synthetic Minority Oversampling Technique) was applied to the training data to achieve a balanced class distribution, improving on model fairness and generalization.

Data Modeling

A Logistic Regression model was trained to predict client subscriptions, with key configurations:

- Class imbalance handling: **class_weight= ‘balanced’** to adjust for minority class (subscribers).
- Reproducibility: **random_state=42** and **max_iter=1000** for consistent convergence.
- Evaluation metrics: Accuracy, precision, recall, F1-score, ROC AUC, and confusion matrix.

Key Insights

1. **Model Performance**:
   - **Accuracy**: 86% (general correctness).
   - **Recall (Class 1)**: 91% (captures 91% of actual subscribers).
   - **Precision (Class 1)**: 45% (45% of predicted subscribers are correct).

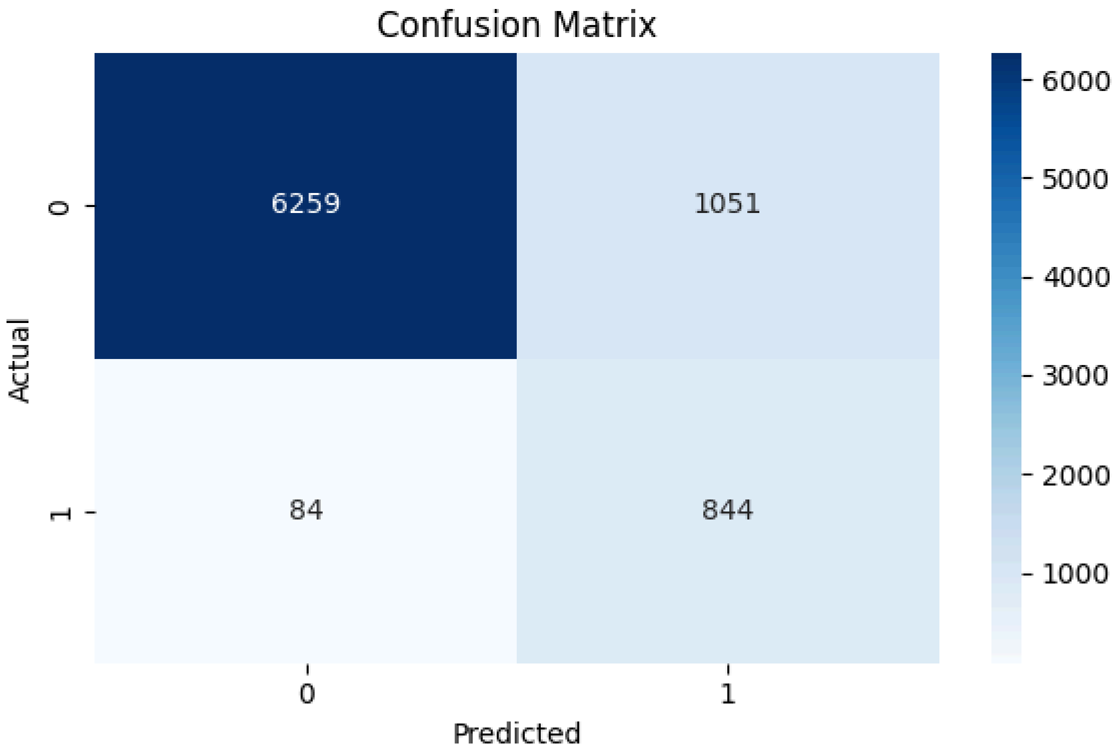
    - **ROC AUC**: 0.94 (excellent class separation).
2. **Confusion Matrix**:
    - **True Positives**: 844 (correctly predicted subscribers).
    - **False Positives**: 1051 (non-subscribers flagged incorrectly).
    - **High recall** prioritizes minimizing missed subscribers, ideal for marketing campaigns.
3. **Class Imbalance Handling**:
    - **class_weight= ‘balanced’** effectively improved minority class predictions without overfitting.

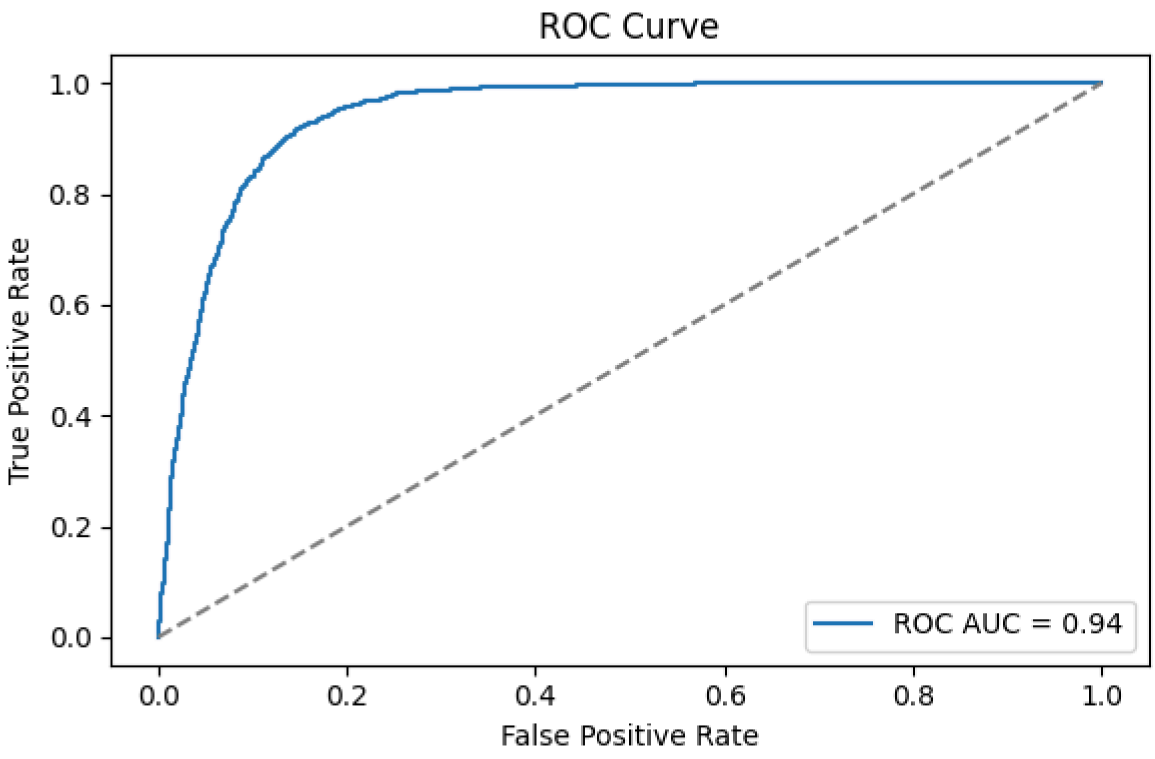
Key outcomes

- High recall (91%) for subscribers, ensuring menial missed opportunities
- Moderate precision (45%), indication some false positives but acceptable for marketing outreach
- Strong overall performance (ROC AUC = 0.94, accuracy = 86%))

Conclusion

The model effectively identifies potential subscribers, making it a valuable tool for targeted marketing campaigns. While precision could be improved, the high recall ensures broad coverage of potential clients. Future enhancements such as hyperparameter tuning, alternative algorithms, and threshold optimization, can further refine performance.





Decision Tree Model

Performance Overview

The Decision Tree model delivers a decent performance with:

- **Accuracy**: 88.31%
- **Precision (Subscription)**: 48.37%
- **Recall (Subscription)**: 56.03%
- **F1 Score (Subscription)**: 51.92%
- **ROC AUC**: 0.7422

Non-Subscription Class Report

- **Precision**: 0.94
- **Recall**: 0.92

- **F1 Score**: 0.93

Confusion Matrix Breakdown

- **True Negatives**: 6755
- **False Positives**: 555
- **False Negatives**: 408
- **True Positives**: 520

ROC Curve Insight

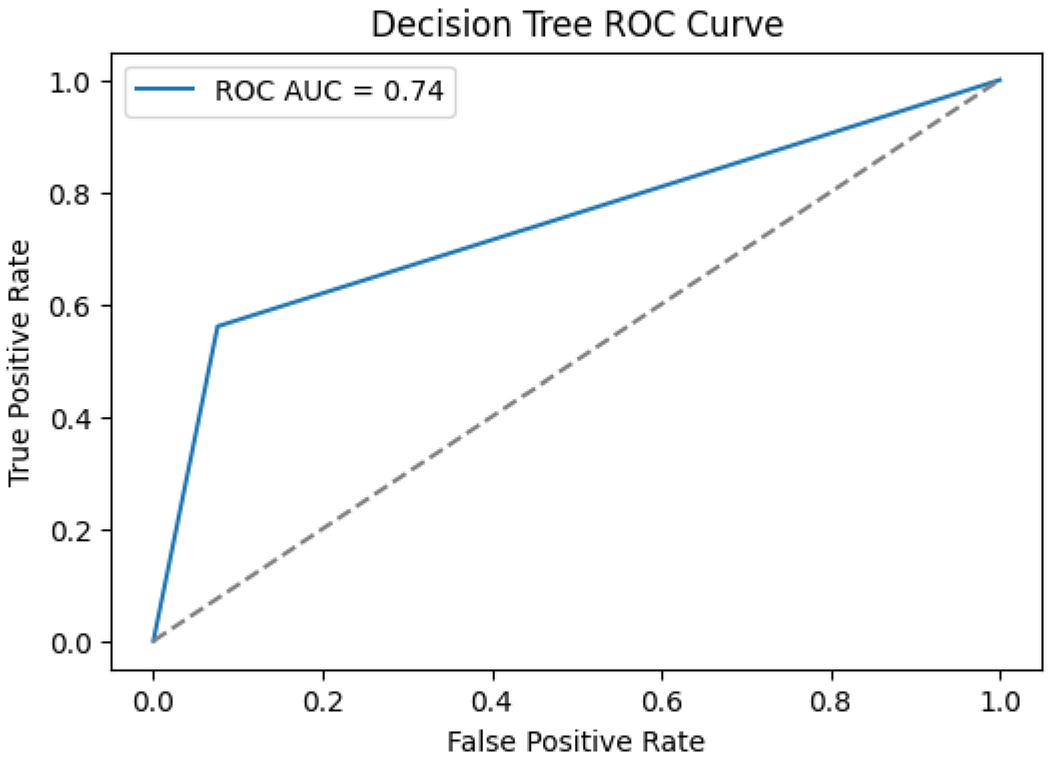
The ROC curve stays above the diagonal, showing reasonable class separation. However, it takes on a jagged, step-like shape, which is expected from a single tree's hard splits. An AUC of 0.74 is acceptable but leaves room for improvement.

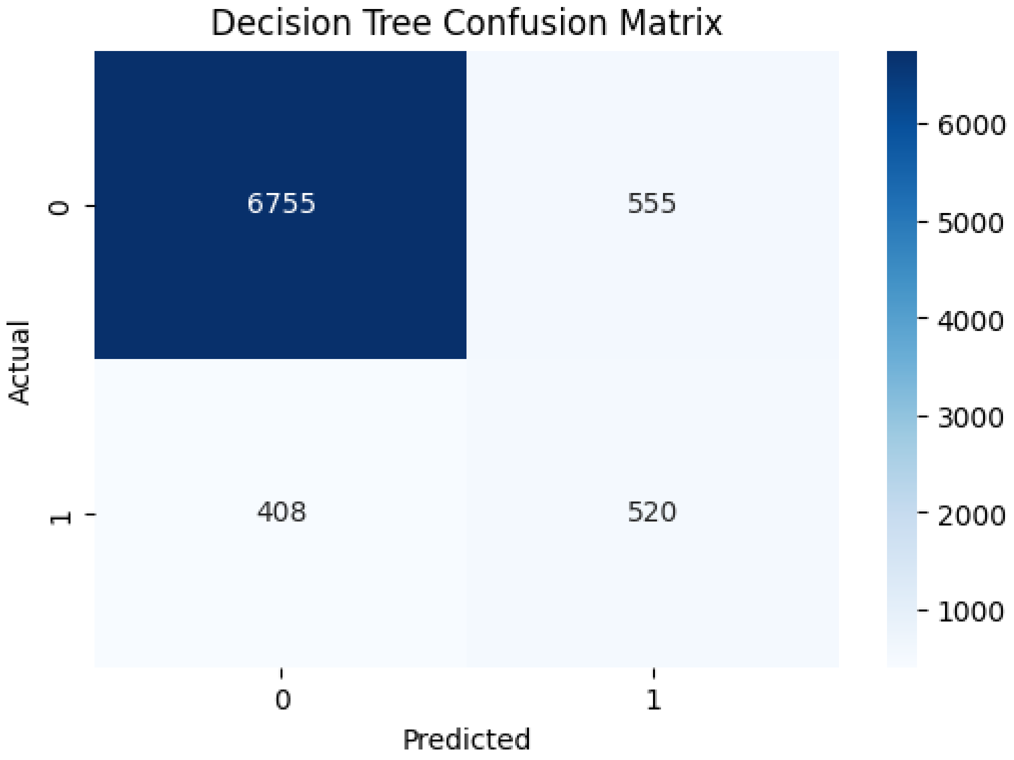
Feature Importance (Top 10)

- `duration`: **45.15%**
- `nr.employed`: **15.61%**
- `cons.conf.idx`: **7.20%**
- `euribor3m`: **4.77%**
- `campaign`: **3.78%**
- `age`: **3.40%**
- `cons.price.idx`: **1.72%**
- `housing_yes`: **1.66%**
- `emp.var.rate`: **1.30%**
- `default_unknown`: **1.12%**

General Observations

The Decision Tree improves significantly over the baseline Logistic Regression model. It captures non-linear relationships and is easy to interpret, which is helpful for feature insight. That said, it can struggle with generalizing to more complex patterns and is more sensitive to overfitting when working with high-dimensional or noisy data.


Decision Tree ROC Curve
ROC AUC = 0.74
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


Decision Tree Confusion Matrix
6755
555
408
520
Actual
Predicted
0
1
0
1
6000
5000
4000
3000
2000
1000

Random Forest Model

Performance Overview

The Random Forest model delivers strong, well-rounded performance:

- **Accuracy**: 91.53%
- **Precision (Subscription)**: 61.64%
- **Recall (Subscription)**: 65.62%
- **F1 Score (Subscription)**: 63.57%
- **ROC AUC**: 0.9489

Non-Subscription Class Report

- **Precision**: 0.96
- **Recall**: 0.95
- **F1 Score**: 0.95

Confusion Matrix Breakdown

- **True Negatives**: 6931
- **False Positives**: 379
- **False Negatives**: 319
- **True Positives**: 609

ROC Curve Insight

The ROC curve is smooth and tightly wrapped around the top-left corner, clear evidence of a highly capable model. An AUC of 0.9489 confirms excellent class separability and strong probability estimation.

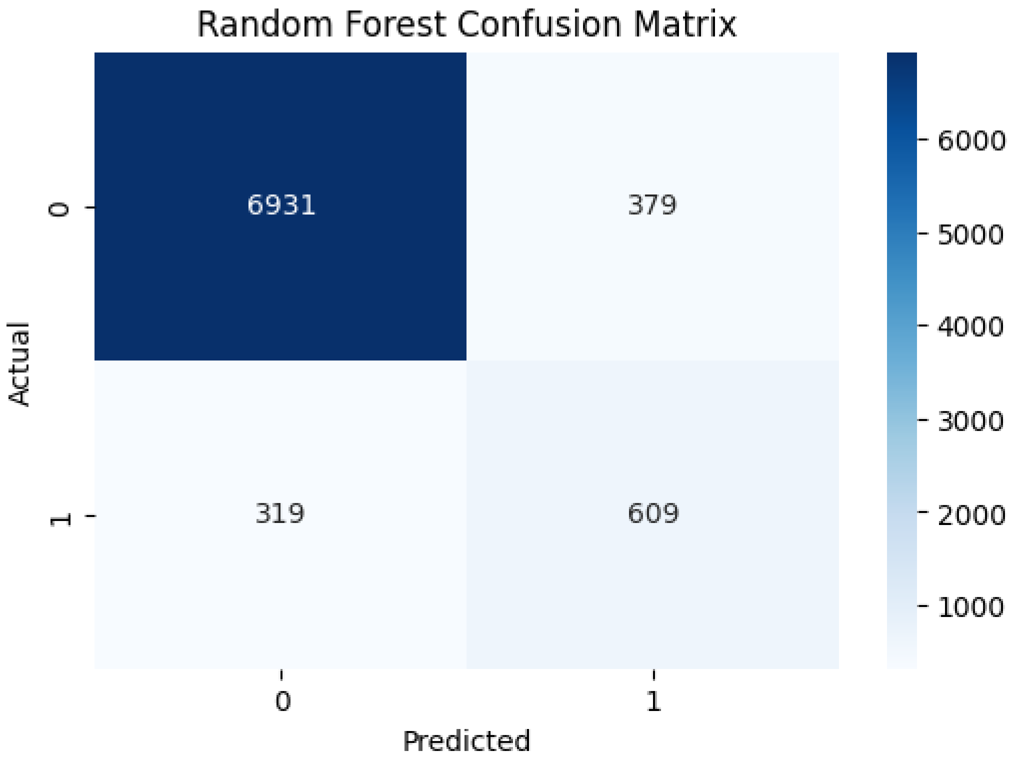
Feature Importance (Top 10)

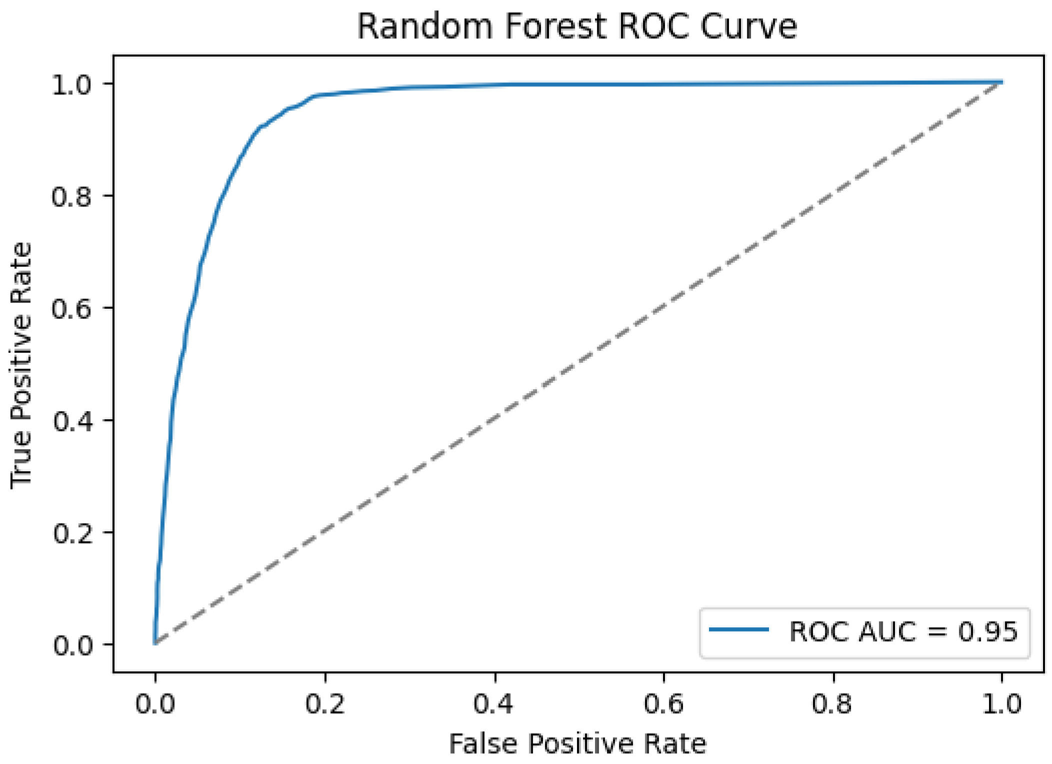
- `duration`: **33.81%**
- `euribor3m`: **8.21%**

- `nr.employed`: **6.77%**
- `emp.var.rate`: **4.95%**
- `campaign`: **4.69%**
- `cons.conf.idx`: **3.78%**
- `age`: **3.68%**
- `housing_yes`: **3.25%**
- `cons.price.idx`: **2.55%**
- `contact_telephone`: **2.49%**

General Observations

Random Forest stands out for its consistency and robustness. The model handles class imbalance well and balances precision and recall for the subscription class effectively. As an ensemble method, it reduces the risk of overfitting and generalizes better to unseen data, making it a solid candidate for deployment in production environments like a web application.





Streamlit Web App

After evaluating both models individually, Random Forest clearly stood out as the better option for deployment in the web app. Here's the reasoning behind the choice

Higher Overall Accuracy

With an accuracy of **91.53%**, Random Forest makes the most correct predictions across the dataset. This level of performance provides confidence in how the model will behave in a real-time environment.

Better at Identifying Potential Subscribers

For the minority class, **those who actually subscribe**, Random Forest offers a strong balance between **precision (61.64%)** and **recall (65.62%)**. The F1 Score of **63.57%** reflects this consistency.

Strong Discriminative Power

The **ROC AUC of 0.9489** shows that Random Forest is excellent at distinguishing between subscribers and non-subscribers. This matters a lot when the app needs to rank users by likelihood and show clear probability scores.

Robust and Production-Ready

Ensemble models like Random Forest tend to generalize better than single estimators. It's more stable, handles noise better, and doesn’t overreact to outliers or weird edge cases, important traits when moving from development to a live, user-facing product.

Actionable Insights

The model doesn't just perform well, it also surfaces **clear feature importance rankings**, with variables like duration, euribor3m, and campaign standing out. These insights are useful for shaping marketing strategies and customer engagement campaigns.