Unveiling the Factors Influencing Term Deposit Subscriptions:Data and EDA

November 5 2024

Abhayraj Rana
ar34814n@pace.edu
Practical Data Science
MS in Data Science
Seidenberg School of CS and IS
Pace university

Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis

Executive summary

Problem: The bank is facing challenges in accurately predicting which customers are likely to subscribe to a term deposit.

Solution: This project aims to build a machine learning model to accurately predict customer behavior. By analyzing customer data, we will create a model that can identify potential subscribers, optimize marketing efforts, and increase conversion rates.

Project plan recap

Deliverable	**Due Date**	**Status**
Data & EDA	11/05/24	Complete
Methods, Findings, and Recommendations	11/12/24	In Progress
Final presentation	11/19/24	Not Started

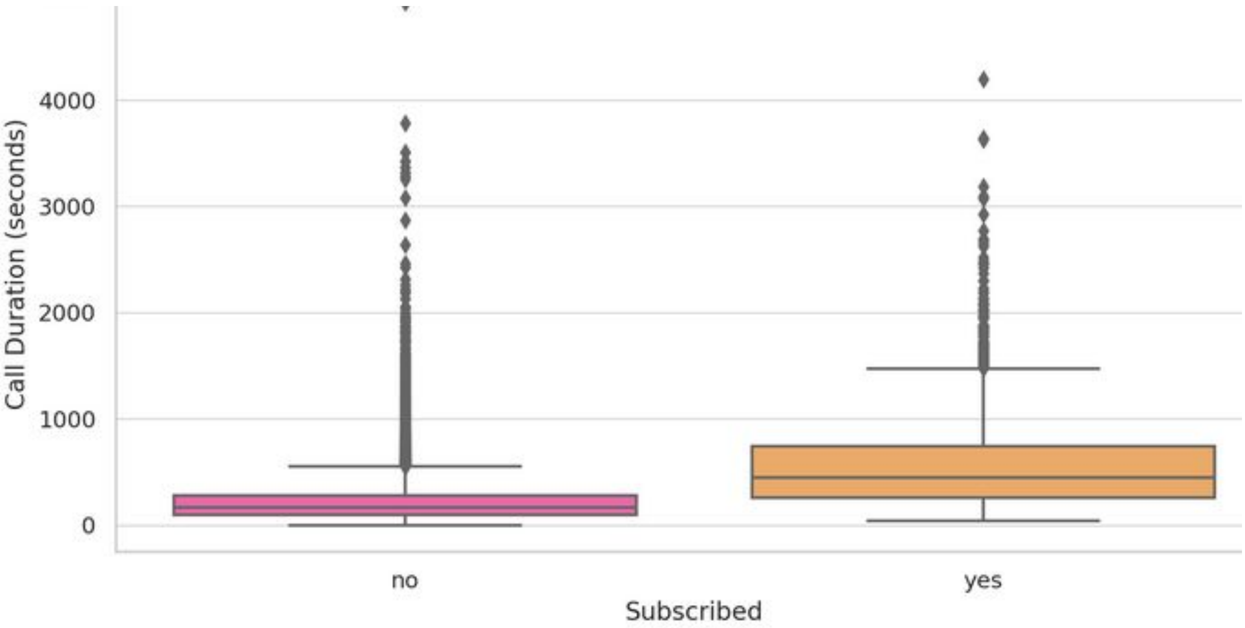
Data

Data

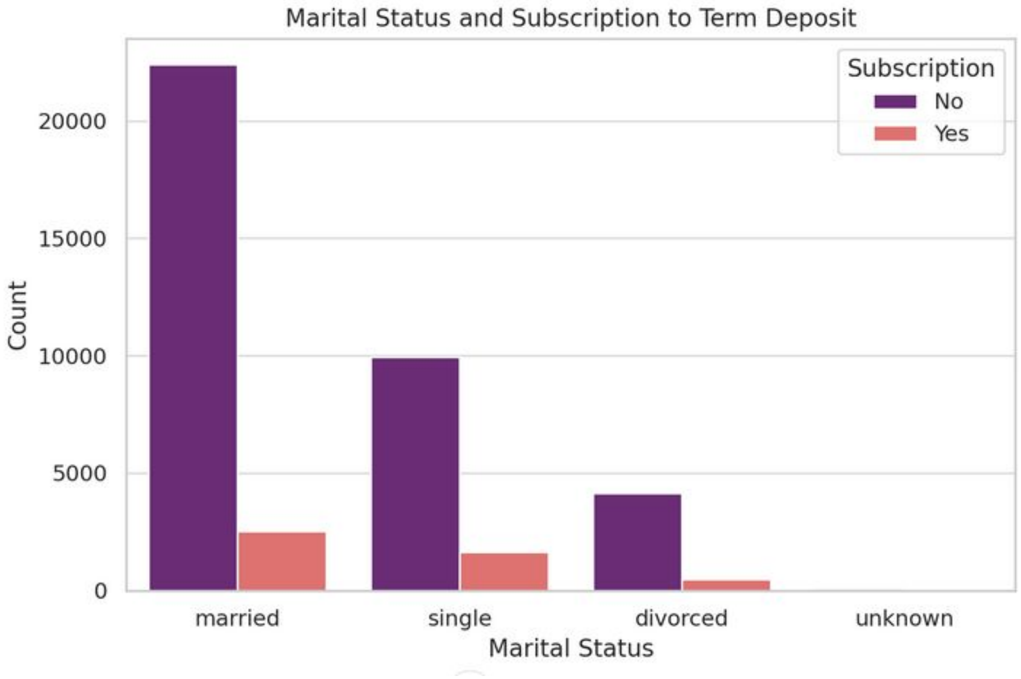
- Details:
  - Data source: Kaggle
  - Sample size: 41188 Records
  - Time period: Does not specify exact dates, as it covers multiple campaigns
  - Data that was purposefully excluded**:** Specific customer names or other personally identifiable information to ensure privacy.
- Notes:
  - The target variable, "y," indicates whether a customer subscribed to a term deposit
  - The data has categorical variables, such as job, marital status, and loan, which were analyzed to discern patterns but may require encoding for advanced modeling.
- Assumptions:
  - We assume that longer call durations suggest higher engagement, which may influence subscription likelihood.

Exploratory Data Analysis

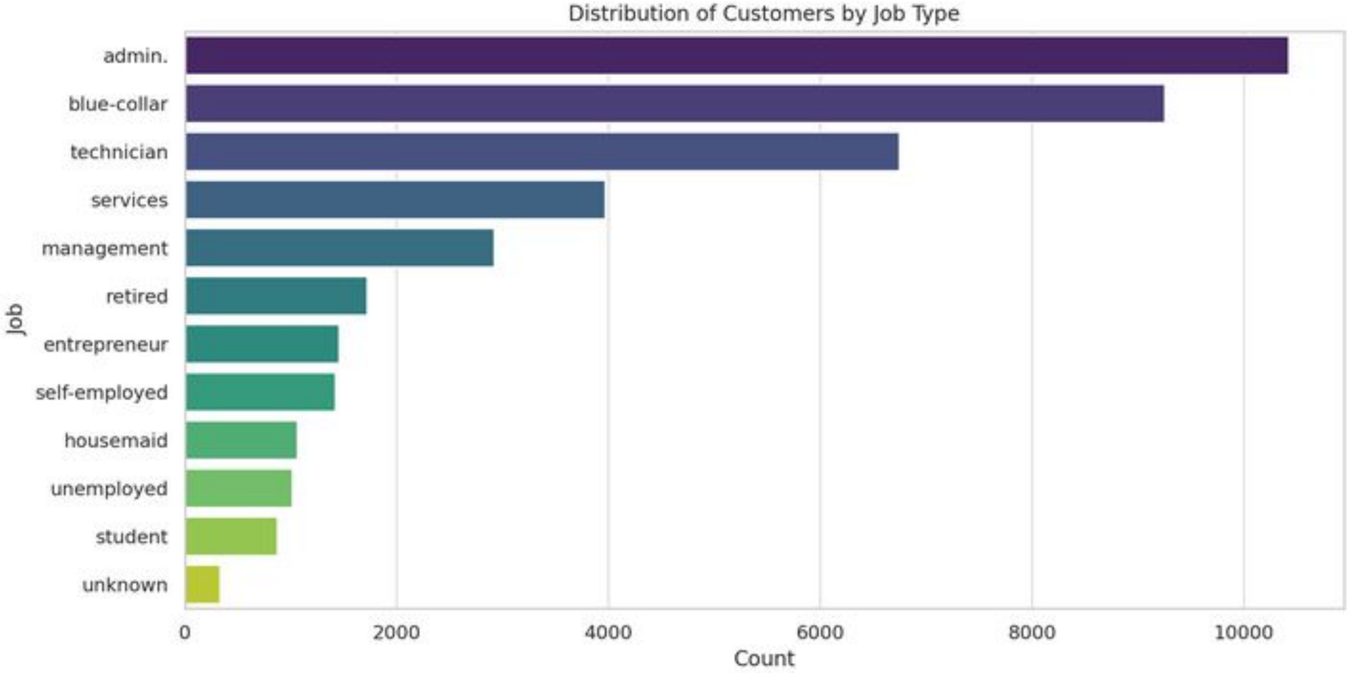
Effect of Call Duration on Subscription Status



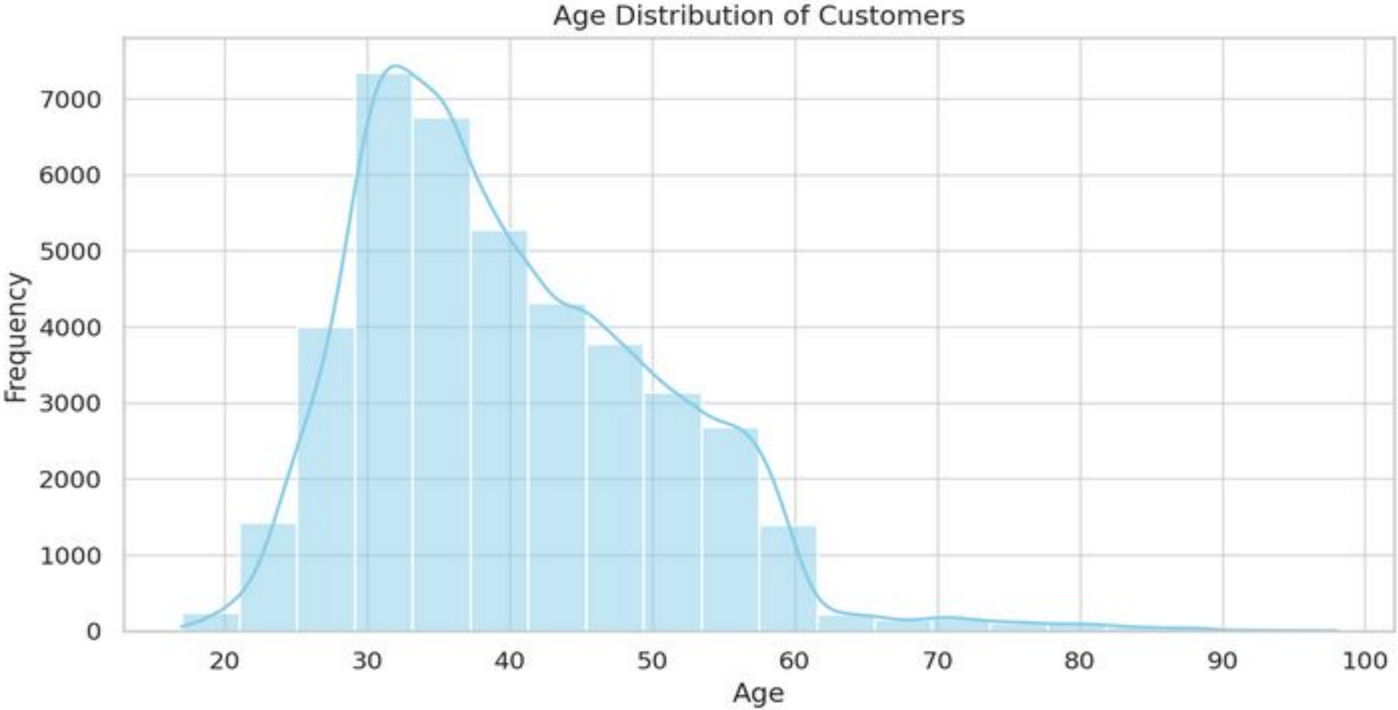
Count of Subscriptions of customers with Marital Status



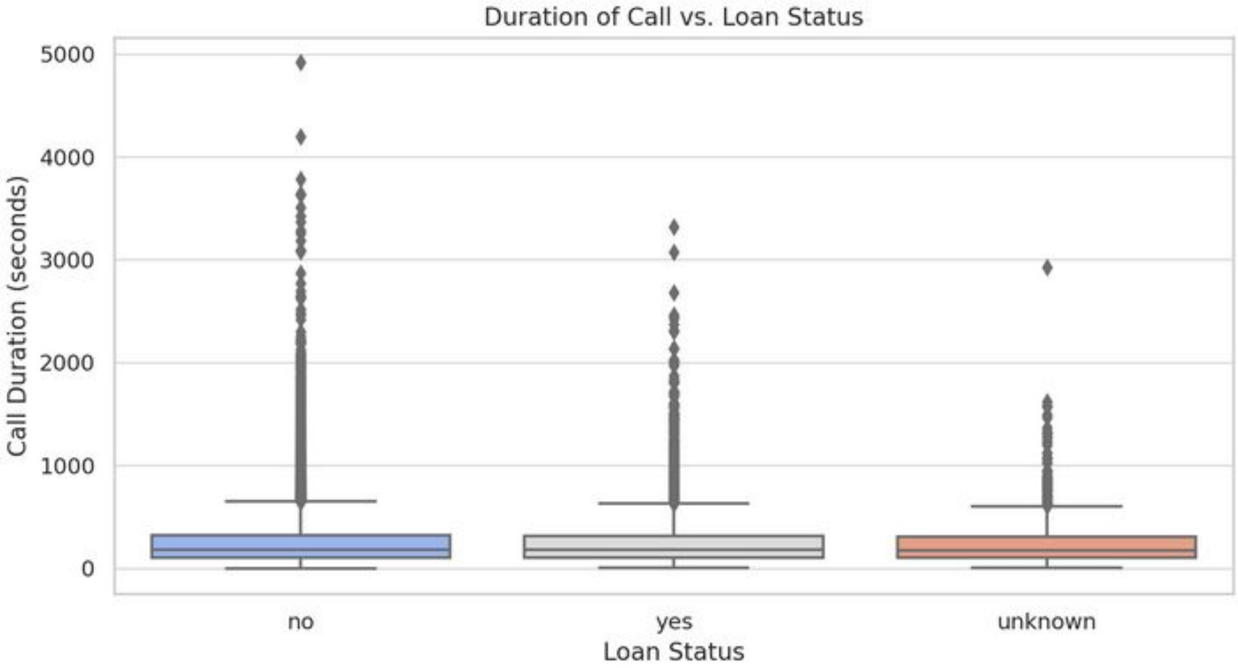
Distribution of Customers by Job Type



Distribution of Customers by Age



Duration of Call vs Loan Status



Agenda

- Modeling methods
- Findings
- Recommendations and technical next steps

Project plan recap

Deliverable	**Due Date**	**Status**
Data & EDA	11/05/24	Complete
Methods, Findings, and Recommendations	11/12/24	Complete
Final presentation	11/19/24	In Progress

Modeling methods

Modeling methods

- **Outcome variable -** This is a binary variable indicating whether a client subscribed to the term deposit. This will be the primary focus for model predictions, as it directly ties to the business goal of increasing subscriptions.

- **Features -** The dataset includes a mix of demographic information, campaign details, and economic indicators. These feature groups help the model learn patterns from different perspectives: demographic data provides insight into client profiles, campaign details show the context and timing of past interactions, and economic indicators reflect broader market conditions. Together, these features offer a well-rounded basis for predicting the likelihood of a client subscribing to a term deposit.

  Grouping Strategy:

❖ Client Demographics: age, job, marital status, education, housing, loan

❖ Campaign Information: contact method, month, day of the week, campaign duration (duration), previous contact details (campaign, pdays, previous, poutcome)

❖ Economic Indicators: emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed

Modeling methods

Model Choice: Logistic Regression

We chose Logistic Regression because it's effective for predicting yes-or-no outcomes, like whether a client will subscribe to a term deposit. Think of this model as estimating the likelihood of a "yes" or "no" based on various factors about each client and the campaign.

For example, imagine you're trying to guess if someone will go to a concert. You'd consider things like their interest in the band, if friends are going, and the ticket price. Logistic Regression works similarly, For example:

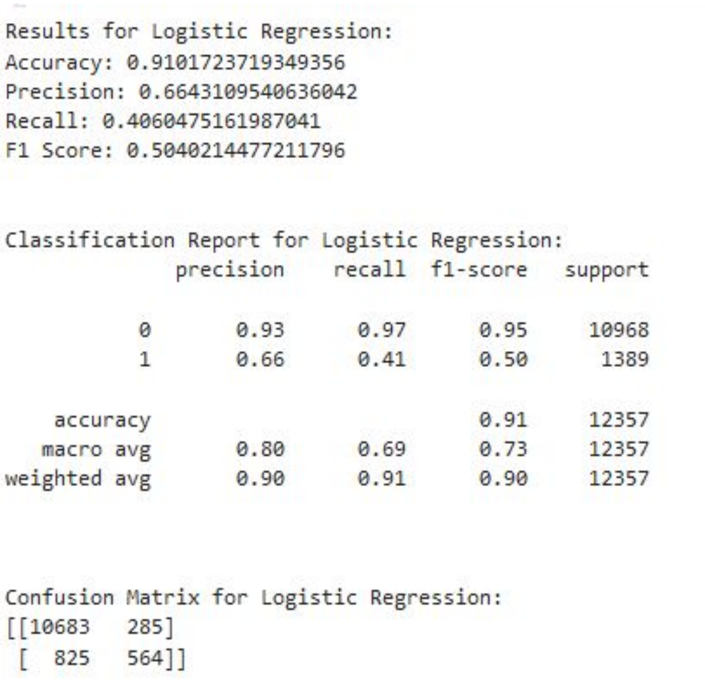
- If a client has been contacted before and already has a loan, the model might use these factors to suggest a lower chance of subscribing this time around.
- On the other hand, if the client is younger, works in a high-paying job, and the interest rates are favorable, the model might nudge the prediction more toward a "yes."

This way, the model isn't guessing but instead looking at all these indicators together to make a more informed prediction.This model is ideal because it provides a clear probability for each client, making it easier to focus marketing efforts on those most likely to subscribe.

For detailed explanation in technical terms, please refer to this index slide.

Findings

Model Performance



```
Results for Logistic Regression:
Accuracy: 0.9101723719349356
Precision: 0.6643109540636042
Recall: 0.4060475161987041
F1 Score: 0.5040214477211796


Classification Report for Logistic Regression:
              precision    recall  f1-score   support

           0       0.93      0.97      0.95     10968
           1       0.66      0.41      0.50      1389

    accuracy                           0.91     12357
   macro avg       0.80      0.69      0.73     12357
weighted avg       0.90      0.91      0.90     12357


Confusion Matrix for Logistic Regression:
[[10683   285]
 [  825   564]]
```

Summary of Findings: (For definition of terms: Index)

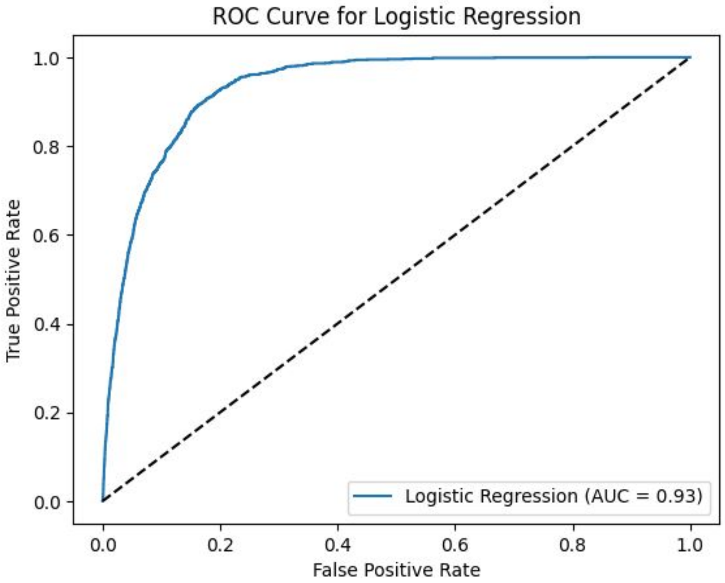
1. **Accuracy**: The model correctly predicts a client's subscription decision 91% of the time, making it generally reliable.
2. **Precision and Recall**:
   - For non-subscribers(Class 0), the model is 93% precise and has a 97% recall, meaning it's very good at identifying clients who are unlikely to subscribe.
   - For potential subscribers(Class 1), the model is only 66% precise and has a 41% recall, meaning it sometimes misses clients who might subscribe.
3. **Confusion Matrix**:
   - The model accurately identified 10,683 non-subscribers and 564 subscribers.
   - It missed 825 actual subscribers, meaning these clients were overlooked as potential subscribers, indicating room for improvement.

Model Performance

Connection to the Business Problem

- **Focus Resources**: With high accuracy for non-subscribers, the bank can avoid spending resources on clients unlikely to subscribe.
- **Missed Opportunities**: The model sometimes misses potential subscribers. Targeted adjustments could help capture these clients more effectively.
- **Overall Value**: Despite some gaps, the model provides a strong foundation for prioritizing high-potential clients, enhancing campaign effectiveness.

Model Performance



This graph shows how well our model can distinguish between customers who will subscribe to a term deposit and those who won't. The closer the curve is to the top-left corner, the better the model's ability to make accurate predictions.

Breakdown:

- True Positive Rate (TPR): Measures the model's ability to correctly identify positive cases.
- False Positive Rate (FPR): Measures the model's tendency to incorrectly identify negative cases as positive.
- AUC (Area Under the Curve): Represents the overall performance of the model. A higher AUC indicates better predictive power.

Recommendations & Data Science Next Steps

Recommendations and Data Science next steps

Recommendations:

1. Target High-Potential Demographics

- **Finding**: The model indicates that certain client demographics—such as younger individuals in stable employment—are more likely to subscribe to a term deposit.
- **Connection to Business Problem**: By identifying which types of clients are most likely to respond positively, the bank can make its marketing efforts more effective and cost-efficient.
- **Actionable Recommendation**: Focus marketing outreach on these high-potential groups, tailoring messages to appeal to their financial needs and goals. For example, a campaign specifically targeting young professionals could highlight long-term savings benefits, appealing directly to this audience.

2. Align Campaigns with Favorable Economic Conditions

- **Finding**: Economic conditions like low interest rates increase the likelihood of clients subscribing.
- **Connection to Business Problem**: Understanding how external factors influence client interest allows the bank to time campaigns for maximum effectiveness.
- **Actionable Recommendation**: Align marketing efforts with favorable economic periods, such as periods of low interest rates, to increase client engagement. For example, launching campaigns during these times could yield higher response rates, as clients are more open to investing in term deposits when rates are appealing.
    - ○

Recommendations and Data Science next steps

Technical Next Steps for the Data Science Team:

1. Enhance the Model for Greater Predictive Power

- **Rationale**: Our current model provides useful insights, but a more sophisticated approach could increase accuracy and help answer additional questions about client behavior.
- **Next Step**: Explore more advanced modeling techniques that may capture complex relationships in the data, such as client behavior patterns or seasonality in responses. This could improve predictions and uncover deeper insights.

2. Expand Data Collection for Broader Insights

- **Rationale**: Additional data points could give the team a more comprehensive understanding of client preferences and financial habits, allowing the bank to refine its outreach strategies.
- **Next Step**: Gather new data on factors like client lifestyle, online engagement with the bank, or interaction frequency with other services. This expanded dataset could allow the team to explore additional business opportunities and fine-tune client targeting further.

Appendix

Additional Information: Model Choice

Main Slide

Model Choice: Logistic Regression

We selected Logistic Regression for this project because it’s a reliable and interpretable model for binary classification tasks, especially useful when we want a straightforward "yes" or "no" outcome. Logistic Regression estimates the probability of a client subscribing to a term deposit by applying a logistic function to a weighted sum of the input features. The model outputs a probability between 0 and 1, which we can interpret as the likelihood of a subscription.

Each feature in our dataset (e.g., client age, job type, economic indicators) is assigned a coefficient, representing its impact on the probability of a subscription. A positive coefficient increases the likelihood of a "yes" outcome, while a negative one decreases it. For example:

- **Previous Contact Frequency**: If the client has been contacted frequently, this feature’s coefficient might push the probability down, as frequent contacts could indicate previous disinterest.
- **Euribor 3-Month Rate**: Economic factors like interest rates can play a major role, with higher rates possibly leading to lower subscription probability, represented by a negative coefficient.

The primary reason for choosing Logistic Regression is its interpretability. We can analyze each feature's coefficient to directly understand how it influences the outcome, making it easier to communicate actionable insights to stakeholders. Additionally, its simplicity helps avoid overfitting, particularly valuable in marketing contexts with potentially limited or noisy data.

Additional Information: Model performance

Main Slide

1. **Accuracy:** This is the overall correctness of your model's predictions. It's like your overall score on a test. A higher accuracy means your model is making more correct predictions.
2. **Precision:** This measures how often your positive predictions are actually correct. It's like your accuracy in guessing who will pass an exam. A higher precision means fewer false positives.
3. **Recall:** This measures how often your model correctly identifies positive cases. It's like remembering all the correct answers on a test. A higher recall means fewer false negatives.
4. **F1-Score:** This is a balance between precision and recall. It's a good metric to use when you want to consider both types of errors.
5. **Support:** This refers to the number of actual occurrences of a class in the dataset. For example, the number of customers who actually bought the product.
6. **Confusion Matrix:** This is a table that shows how many predictions were correct and incorrect. It helps visualize the performance of your model in more detail.

By understanding these metrics, you can assess the effectiveness of your model and make informed decisions about how to improve it.

Additional Information

- Link to the Git Repo with the code