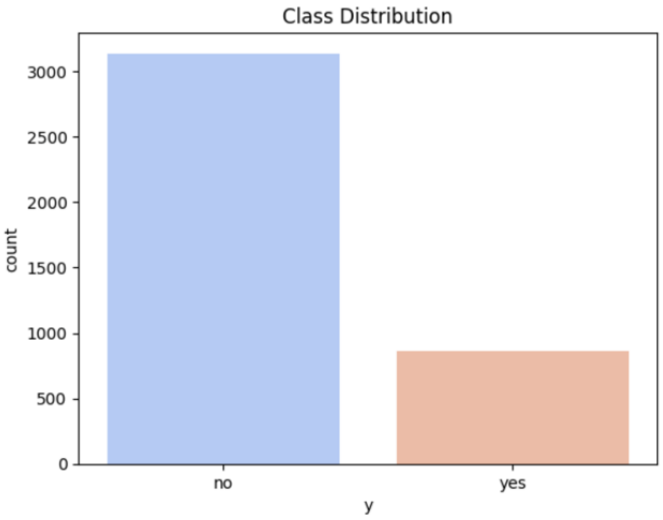
Section A: Summarization

The dataset has 4000 records with 16 features. The goal is to identify customers who are more likely to subscribe to the " N/LAB Platinum Deposit". The data showed some imbalances in class distribution, which led to the use of balanced class weights in the model and optimizing certain thesholds. Key features in the dataset include age (ranging from 18 to 95, with a median of 39), balance (ranging from -8019 to 66,653, with a median of 481.5), duration (maximum of 4918 seconds, median 197 seconds), and campaign (number of customers contacted in each campaign) (median of 2 and a maximum of 41). Moreover, variables like **previous** (median 0) and **pdays** (median -1, indicating that most customers were not contacted previously) this provide some insights into how N/LAB interact with customers across their campaigns.

Key Observations:

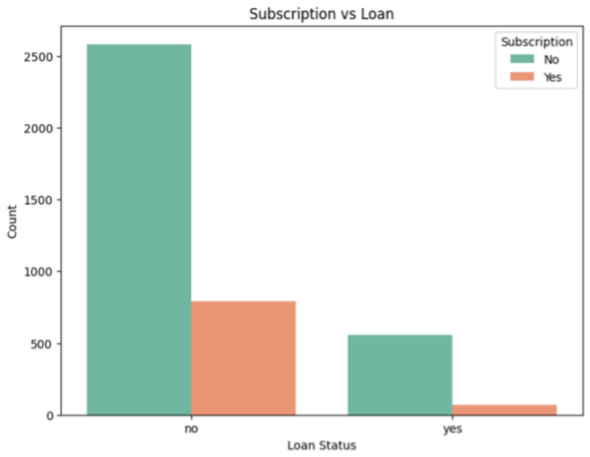
1. **Distribution of Target Variable (y):**
    - 85% of the dataset represents customers who did not subscribe and the other 15% represents customers who subscribed, meaning strategies are needed to address this during modeling.

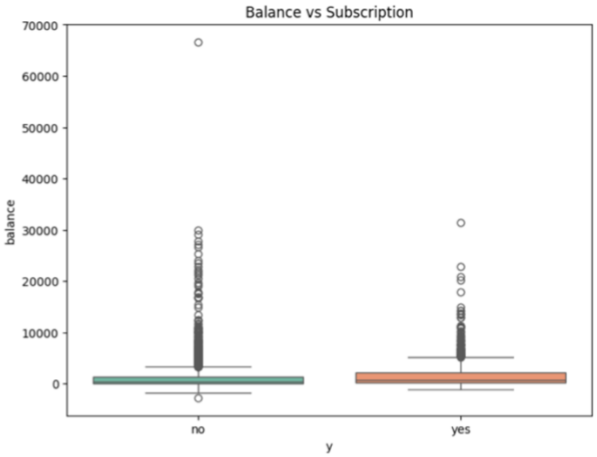


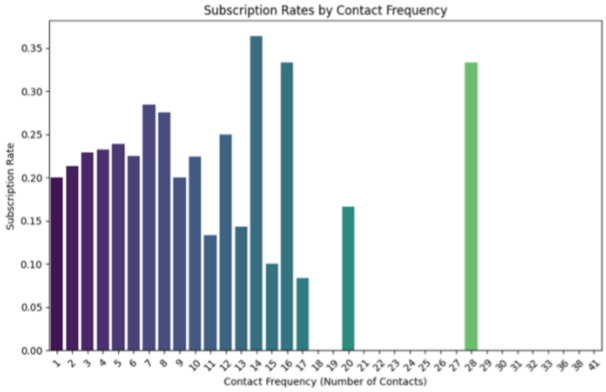
2. **Relationships Between Features and Target Variable:**

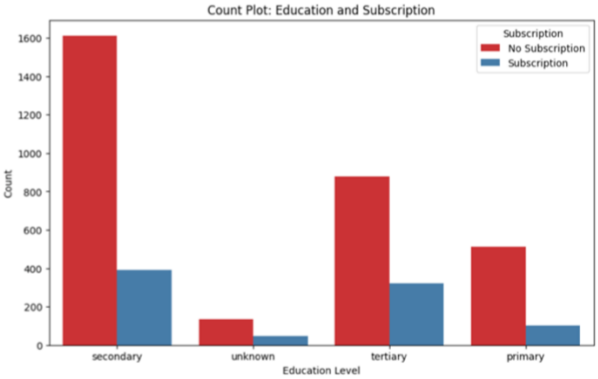
    - **Loan:** customers with no personal loans are more likely to subscribe. As shown in fig 2.
    - **Balance:** strong positive correlation between balance and subscribtion. Subscribers have higher median and interquartile range (IQR).
    - **Duration**: Strongest positive correlation with the target variable but due to its unavailability in the futur it is excluded from predictive modeling.
    - **Contact Frequency:** The higher the contact frequency from campaigns and poutcomes, the better the subscription rates.
    - **Education Level:** Customers with primary or tertiary education are less likely to subscribe than customers with secondary education.
    - **Job Type:** "Management" and "Technician" positions had higher subscription rates than others.

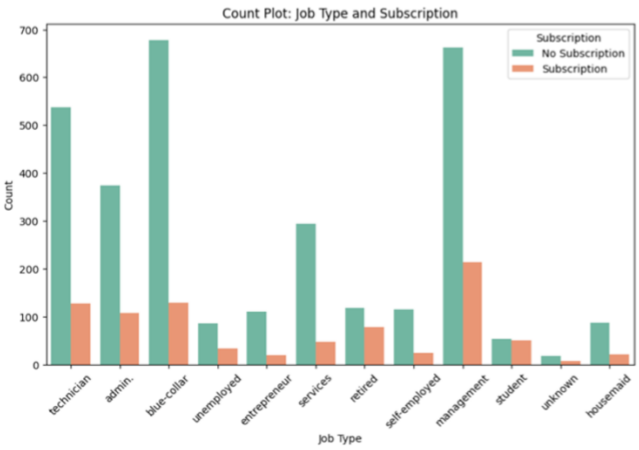
Visualizations:

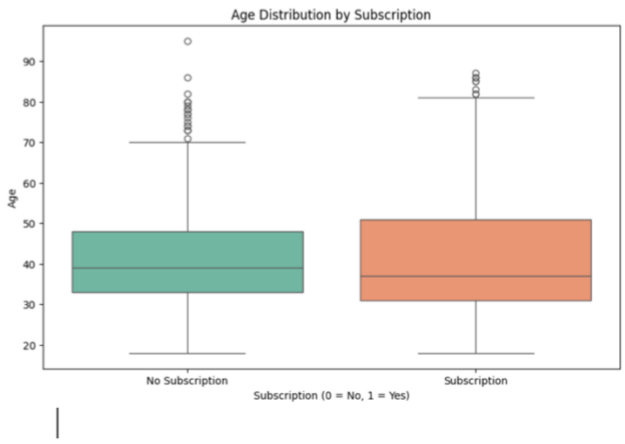








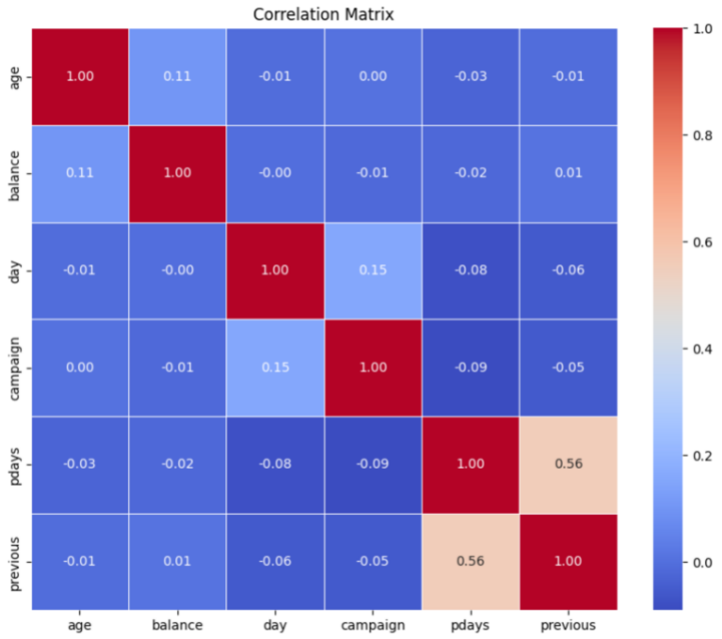




3. Feature Correlation

- Weak correlation between balance and other features (balance and age: 0.11). Indicates that balance is relatively independent.
- Moderately positive correlation between pdays and previous. Suggests potential redundancy.

- Duration shows the strongest positive correlation with the Variable (y). but cannot be used in future predictions.
- Customers with higher balances, frequent contacts, older age, and higher education levels are better prospects for subscription.
- Imbalance in the target variable requires careful handling during model training.

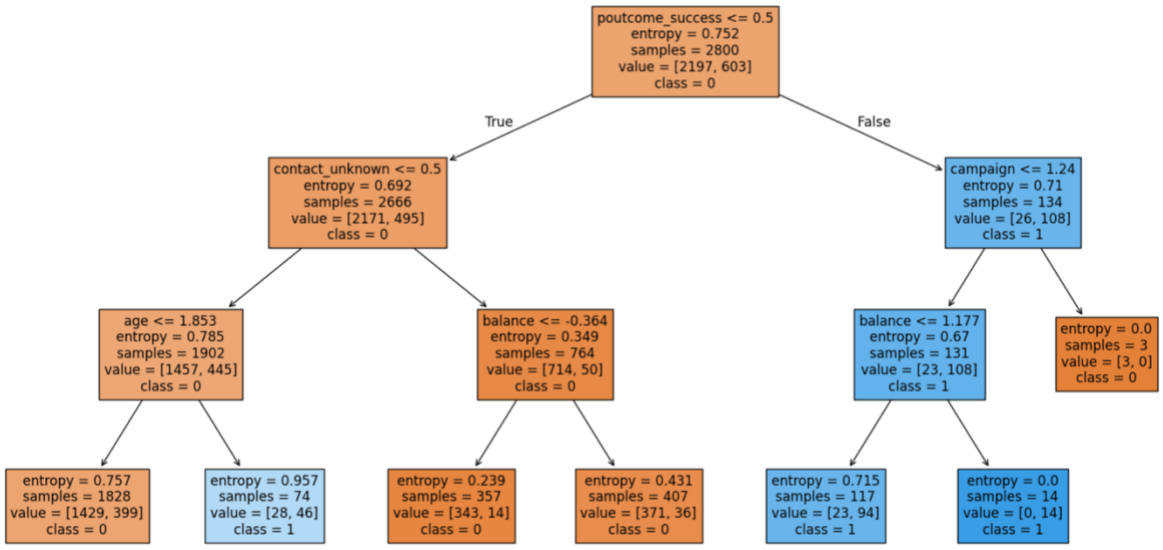


Section B: Exploration

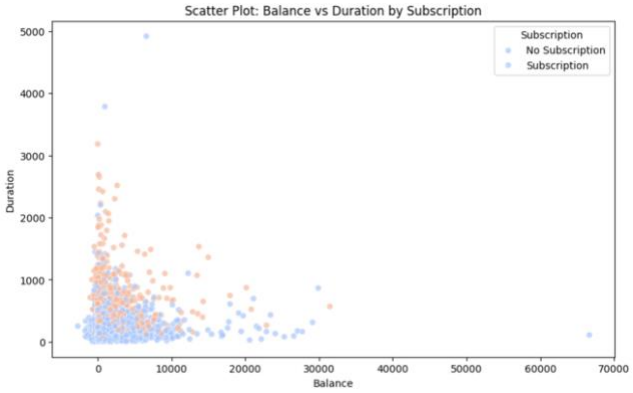
A Decision Tree was used to explore factors affecting the target variable (y). It showed the following primary splits and patterns.d

1. **Important Features:**
   - **poutcome_success:** This is the most important feature in the tree. Customers for whom the outcome in previous campaigns was successful had a much higher subscription rate.
   - **Contact_unknown:** In cases where contact type was unknown, this highly reduced the possibility of subscription. Therefore, this means that correct contact information through which customers are reached is crucial.
   - **Balance:** Higher balances had higher likelihoods of subscribing. balance_per_contact and recent_contact proved to be crucial derived features.
   - **Contact Frequency:** Customers contacted fewer times were more likely to subscribe. This means that over-contacting is less effective and might lose potential customers.
   - **Age:** Older customers, particularly those aged between 40 and 60, showed a higher likelihood of subscribing.

2. **Visualization:** The decision highlighted the following features as significant predictors of the target variable: poutcome_success, contact_unknown, balance, campaign, and age



3. **Variables**
    - The scatter plot of balance vs. duration showed positive relationships subscription likelyhood.
    - The histogram of age showed customer distributions and their correlation with the target variable. The distribution is slightly skewed to the right, with fewer older customers (above 60).



    - categorical features (job,education, and loan) also showed patterns in subscription behavior, as shown in the summarization section.
    - Derived features (balance_per_contact, recent_contact), were more useful for predictions than raw features.
    - categorical features helped uncover sub-populations with higher subscription rates.
4. **decision to use better performing models:**
    - While the Decision Tree provides valuable insights into the data, the simplicity and overfitting issues in a Decision Tree make it difficult to use for predictions. Thus, to overcome these limitations, other models have been considered so that there is better generalization and performance.

Section C: Model Evaluation

Three following models were evaluated to find the best classifier for predicting customer mosre likely to subscribe:

Models Selected and Parameterization:

1. **Logistic Regression:** For interpretability and baseline performance. Used class_weight = 'balanced', and changed the classification threshold to **0.58** (based on maximizing F1-score). These help in improving the balance between precision and recall and to handle class imbalance.
2. **Random Forest:** For reliable predictions. The best parameters according to the GridSearchCV are:`criterion`: 'gini', `max_depth`: 20, `min_samples_leaf`: 2, `min_samples_split`: 2, `n_estimators`: 50. The classification threshold was optimized to **0.28**, which improved recall while maintaining reasonable precision.
3. **Decision Tree:** For exploratory insights. The maximum depth is set to 3 and the minimum split of 2 to maintain interpretability and avoid overfitting.

Model	Accuracy	Precision	Recall	F1-Score	Confusion Matrix
Logistic Reg.	77%	47%	56%	52%	[[639 302] [84 175]]
Random Forest	75%	45%	58%	49%	[[901 40] [182 77]]
Decision Tree	42%	26%	90%	40%	[[276 665] [25 234]]

Evaluation Strategy:

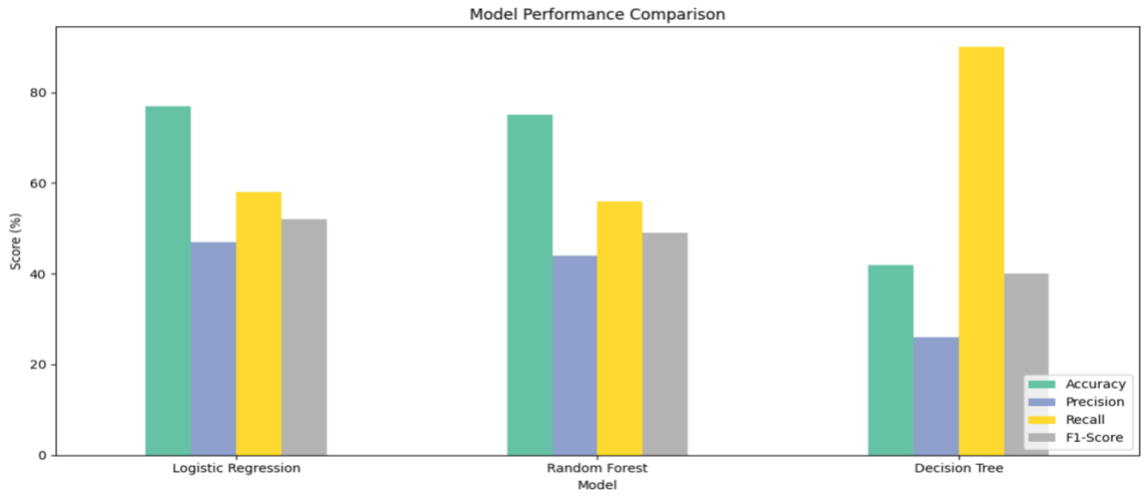
1. **Train-Test Split:** The dataset was split into 70% training and 30% testing.
2. **Cross-Validation:** To avoid overfitting and make sure the model is stable we used Stratified K-Fold cross-validation (k=5).

    - **Cross-Validation F1 Scores for random forest**: [0.31395349, 0.40677966, 0.45454545, 0.35365854, 0.34319527]
    - **Mean F1 Score**: 0.3744
    - Array ([0.314, 0.407, 0.455, 0.354, 0.343])

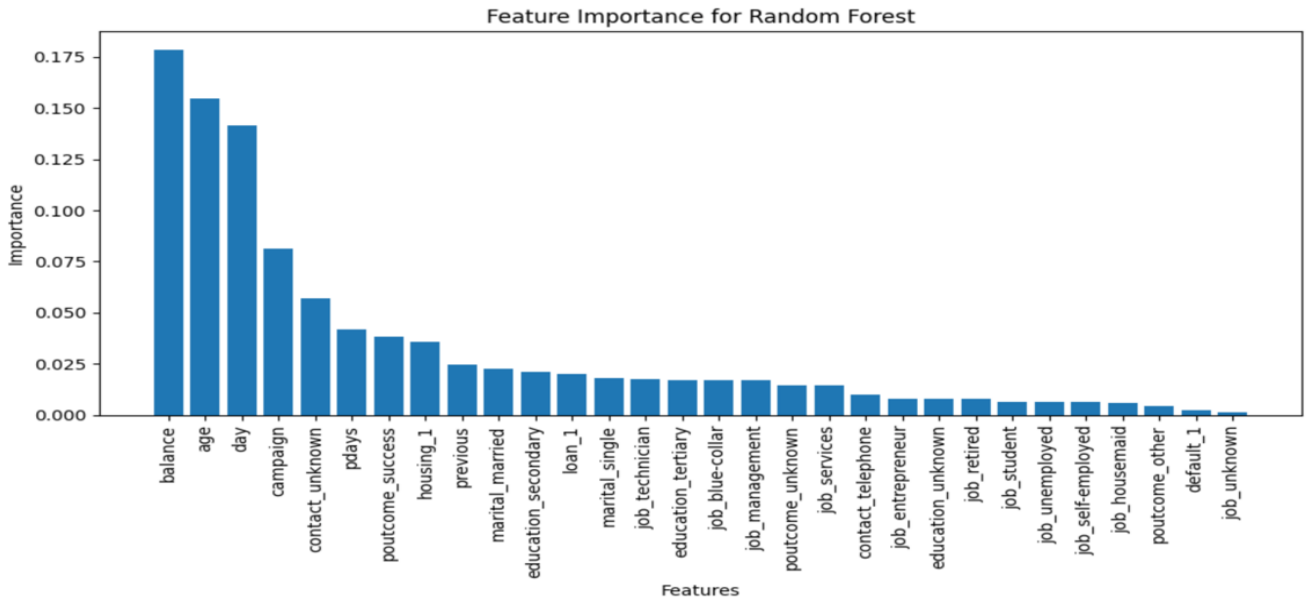
3. **Metrics:**
    - **Accuracy:** Ratio of correct predictions.
    - **Precision & recall:** Ratio of true positives (subscribers) and the true negatives.
    - **F1-Score:** The harmonic mean of precision and recall.

Key Findings:

- **Logistic Regression**: Better F1-Score after adjusting the threshold but lower recall compared to Random Forest. Having overall balanced performance.
- **Random Forest**: Best F1-Score, balancing precision and recall.
- **Decision Tree**: Simple and interpretable, provides strong recall but lacks in percision, leading to weaker overall performance in predictive models.

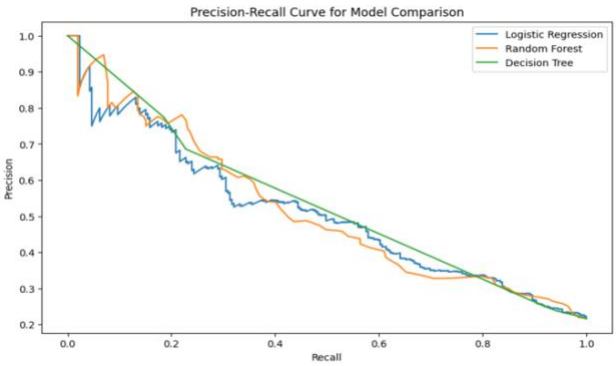


Feature Importance Analysis: The Random Forest feature importance analysis emphasized on the features discussed in Part A:



Analysis Of The Metrics:

1. **Accuracy:**
    - Accuracy varies across models, with Logistic Regession achieving the highest (77%) followed by random forest (75%), but it is not reliable because of the class imbalance.
2. **Precision and Recall:**
    - Precision for class 1 (subscribers) is highest in Logistic Regession (47%), showing control over false positives.
    - Recall is most effective in the Random Forest (56%) best for idintifying all potential subscribers.
    - Precision-Recall curves were used to optimize thresholds, to maintain a balance between metrics, even with imbalanced data.



3. **F1-Score:**
    - F1-Score is highest for Logestic Regression (52%) after it comes the Random Forest (49%) , even though LR is higher RF is more suitable.

Conclusion: Even if Logistic Regression does set a quite clean and interpretable baseline, some limitations about recall make the model useless insubscriber identification.On the other hand, Decision Treese shows the feature importances,but it cannot be used for prediction. Therefore, the balanced Random Forest keeps precision, recall, and overall predictive performance good for this task by processing class imbalance; its recall being larger makes this model the best.

Section D: Final Assessment

Winning Classifier is the **Random Forest** model due to its better overall performance:

- **High Performance:** Outperformed other models in recall.
- **Business Alignment:** It identifies high-probability subscribers and ensures that no potential subscribers are missed during targeting.
- **Feature Insights:** The feature importance analysis validates the focus on key customer attributes.
- **Class Imbalance Handling:** In training, it's class weighting did not allow the minority class to get mislaid.
- **Profit analysis:** The Random Forest model yields a net profit of $7,583, which is good for this approach and ensures campaign optimization.

The strength of Logistic Regression is its interpretability, giving clear coefficients for each feature; this simplicity makes it limited in capturing non-linear relationships. Therefore, Random Forest is better at finding complex and nonlinear patterns, making it better suited for this dataset.

Section E: Model Implementation

After selecting Random Forest based on the evaluation in Section C, we proceed by preparing the model for deployment and training it using the entire training dataset. To ensure reliable performance and optimize the model, we employed Grid Search for hyperparameter tuning and Cross-Validation to assess its generalizability.

The Random Forest model was finalized with the following steps:

1. **Preprocessing:**
    - Standardized and scaled numeric features using StandardScaler.

    - Categorical features were encoded using one-hot encoding.
2. **Model Deployment:**
    - Load the new dataset and apply the same preprocessing steps used.
    - Save the trained model using joblib.
    - Use the predict function to generate subscription probabilities.
3. **Threshold Setting:**
    - Adjust threshold to 0.28 to make sure it aligns with business objectives.

Instructions:

- Run the Python script, (Note: Make sure all dependencies are installed).
- Replace the dataset path in the script with the new path.
- run the script to get predictions.
- Use the predictions to target the right customers based on the business-defined threshold.

The provided code and instructions allow deployment for new dataset, to create new predictions for potential customers.

Section F: Business Case Recommendations

Based on the analysis, the following recommendations are provided:

1. **Target High-Probability Customers:** Focus on customers with higher balances, postive previous contacts, higher age, and secondary education.
2. **Timing Optimization:** Check customer behavior, such as seasonal or monthly patterns, schedule the campaigns at times when customers are most likely to subscribe.
3. **Incentivized Referral Programs:** Introduce Referral program with exclusive incentives. This will create better word of mouth marketing.
4. **Segment-Based Targeting:** Extract the feature importance analysis for segmenting customers, into high-balance customers and low-contact customers, and thereby communicate effectively with them. Additionally, Use insights from categorical variables like job types and education to design a group-specific offer.
5. **Monitoring and adaptation:** Ensure responsiveness and adptation to the constant market changes.
6. **Future Analysis**: Investigate other derived features such as income-to-loan ratio among others. And create depper analysis with other models which are more advanced such as Gradient Boosting.
7. **Model Improvement**: Although the Random Forest might be reallygood for immediate usage, the nature of the market-unpredictable-requires constant adaptations and refinements to be made.

Conclusion:

The recommendations mentioned above, supported by optimal uses of random forest, if effectively applied, would further help N/LAB enhance marketing efficiency and have better prospect targeting in general, which improves the overall business results. The analysis will state actionable strategies to enhance targeted, data-driven campaigns that will help N/LAB achieve fixed-term deposit subsriptions by balancing efficiency with customer acquisition goals.