N/LAB Platinum Deposit

Data Analysis Report - 20535493

Section A: Summarization

The dataset consists of characteristics and demographic features of individuals and their interactions with N/LAB bank. These features can be broadly classified into three categories: demographics, financial, and contact. **Demographic features** include the individual's age, job role, marital status, and education level. **Financial features** include credit default status, current balance, housing, and personal loan status. **Contact-related** features include communication type, day of the month last contacted, duration of the last contact, the number of contacts, and days since the client was last contacted in a previous campaign. The target output is binary, indicating whether the call resulted in a sale ('yes' or 'no'). The data types include numerical (age, balance, day, duration, campaign, pdays, previous) and categorical (job, marital, education, default, housing, loan, contact, poutcome, y).

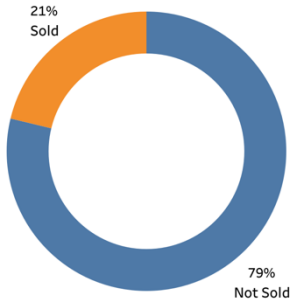


Figure 1: Distribution of target variable 'y'

Upon examining the data, the first noteworthy observation is the high imbalance in the dataset. Figure 1 shows the distribution of target variable in the available dataset, only **21%** of the calls made to individuals resulted in a sale. This imbalance could have implications for the development and evaluation of predictive models, as the class distribution is skewed towards one outcome (no sale) over the other. It's important to consider this imbalance when interpreting the model performance and making predictions related to the likelihood of a successful sale.

The dataset reveals that the average age of the individuals is around 41 years, and the median age is 38 years. Furthermore, it indicates that 75% of the individuals fall within the age of 48 years or younger.

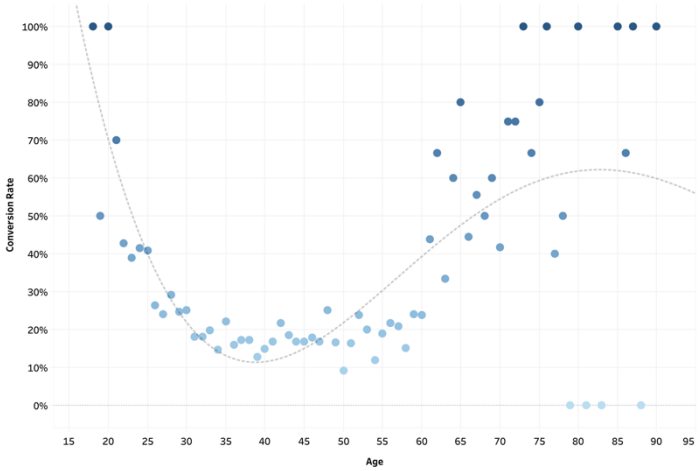


Figure 2: Sale conversion rate Vs Age

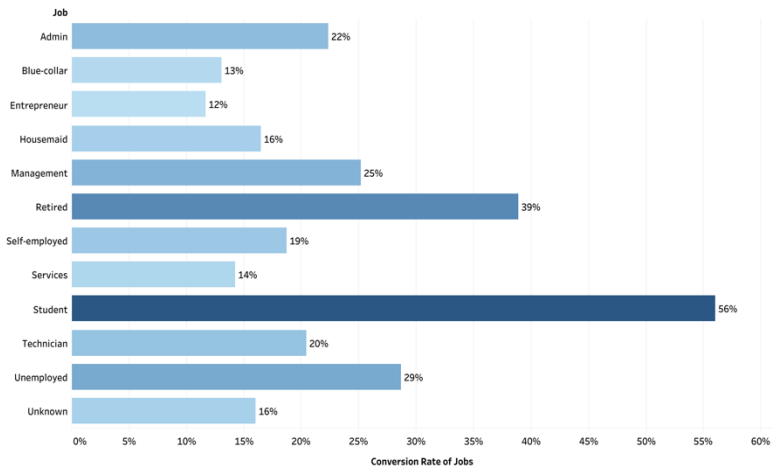


Figure 3: Sale conversion rates for various job categories

The above figures offer valuable insights into the effectiveness of calls made to customers across different age groups and job categories. Figure 2 indicates a notable trend where both the **younger age** demographic and those **above 60** exhibit a higher subscription rate to the deposit scheme compared to middle-aged customers. This trend extends to the job categories of the individuals, as depicted in Figure 3. Specifically, **students** and **retired** individuals show a higher inclination to subscribe to the Deposit scheme. This trend also extends to individual's marital status. Since many students would be bachelors, single customers are **37%** more likely to subscribe to the Deposit scheme than the married customers. It's also seen that there's a significant improvement in the subscription rate if the individual is educated at **Tertiary level** compared to Primary and Secondary levels.

Looking at the subscription rates of financial features, it's evident that the individuals that did not default on credit tend to subscribe to the scheme more than the ones who default. This trend extends to the customer's loan status as well. Looking at the right part of the Figure 4, its noteworthy that the individuals who did not take personal loan and housing loan tend to subscribe to the scheme the most. The subscription rate for such customers is almost **100%** more than the customers who had taken at least one kind of loan.

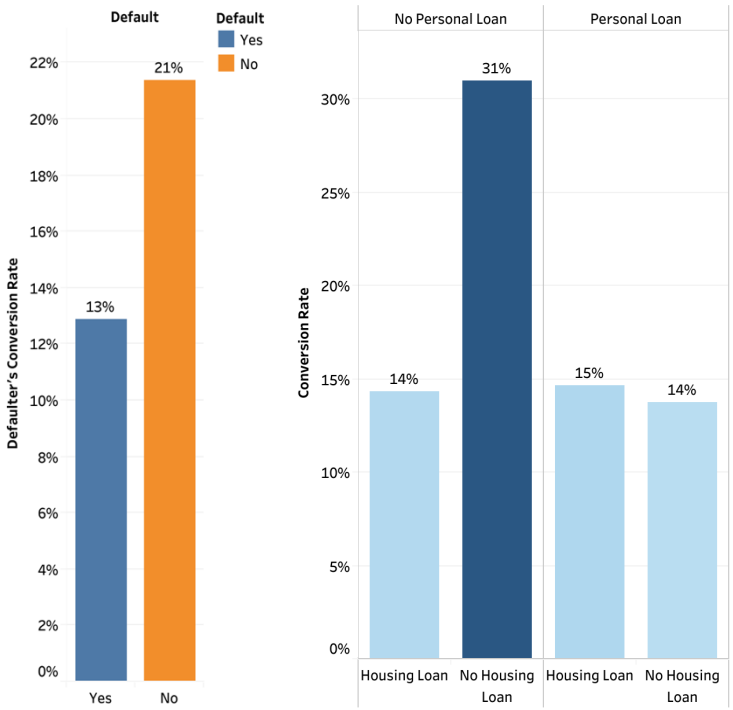


Figure 4: Subscription rates for financial features of customers

We can also see a linear trend in the subscription rate for bank balance of customers. As per the available data, lower the bank balance, less likely they are to subscribe to the new scheme. Only **8-10%** of individuals that subscribed had negative bank balance whereas approximately **25%** of the individuals who subscribed to the deposit scheme had bank balance greater than GBP 1000.

When it comes to the contact related features, the customers who were contacted fewer number of times tend to subscribe to the new scheme more (refer Figure 5). On an average, **50%** of the customers who subscribed to the scheme were contacted **10 times or fewer** before the current campaign. Customers who were frequently contacted showed no interest in buying the scheme. None of the customers who were contacted more than 15 times previously, subscribed to the scheme. This specific trend also extends to the 'campaign' feature where the customers contacted the least number of times during current campaign were more fruitful.

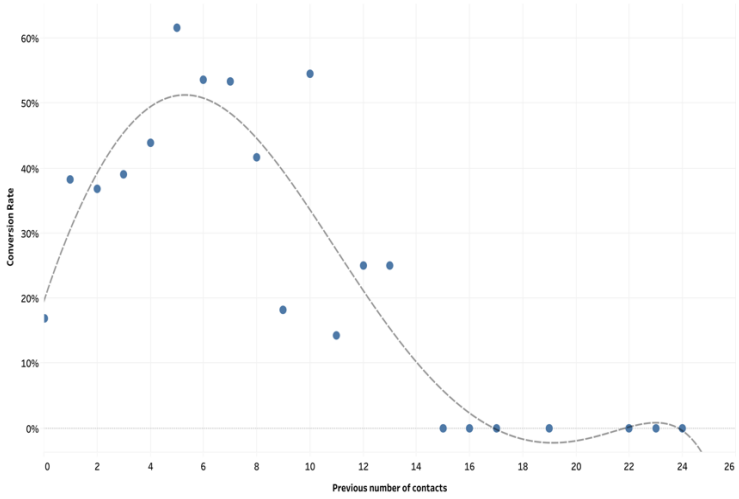


Figure 5: Subscription rate Vs Prior number of contacts

Figure 6 shows the subscription rate of individuals based on their marital status as well as the part of the month they were last contacted. Married and Single customers that subscribed were mostly last contacted in the beginning of the month or the end of the month. Whereas the subscribed customers who were divorced showed least conversion rate when contacted at the end of the month.

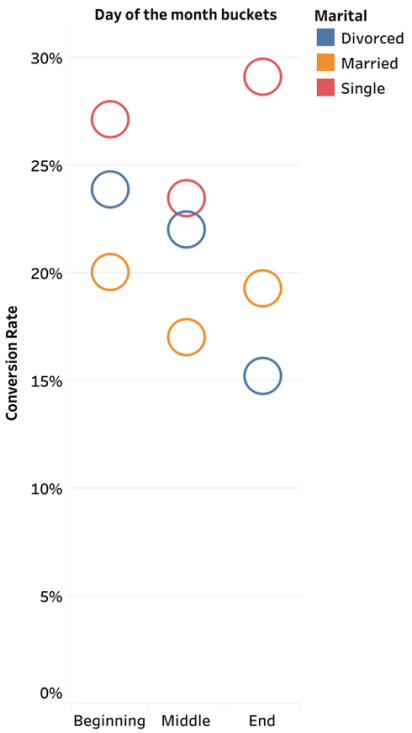


Figure 6: Subscription Rate Vs Day of the month buckets for various marital status

The correlation matrix, generated after label encoding all input features, provides valuable insights into the relationships among different features. Notably, the highest correlation, at **0.57**, is observed between the number of previous contacts and the days elapsed since the client was last contacted. Additionally, a moderate correlation of **0.45** is evident between the duration of the call and the target variable. Furthermore, there exists a moderate correlation of **0.42** between the individual's age and marital status. These correlation values offer a quantitative understanding of the interdependencies among various input features, shedding light on potential patterns and relationships within the dataset.

Section B: Exploration

A decision tree was applied to the given dataset to unravel the underlying factors that significantly influence the outcomes. The goal was to identify a preliminary set of influential features and assess their importance in predicting whether a call to an individual resulted in a sale.

Considerations: Some considerations were made with regards to the features and the data points. From the correlation matrix it was clear that the **duration** of the calls had a big impact on whether a call led to a sale. And since its evident, duration would always turn out to be the most influential feature in our analysis. And since we do not know the duration for future calls, we decided not to use this in our analysis. Another feature, '**poutcome**,' which tells us the outcome of the current campaign, had about **80%** of the data points marked as unknown. So, we decided to skip this feature too. For the other features, even though some data points were unknown, it wasn't a big deal. The unknown datapoints were replaced with the most common value (**Mode**) for that feature. This way, we simplified our analysis and focused on the features that matter most for predicting sales outcomes.

To meet the prerequisites of Scikit-Learn's decision tree model, the categorical features underwent label encoding before model application. This was necessary because Scikit-Learn's decision tree requires all features to be in a numeric format for accurate processing and analysis. Through this analysis, certain variables emerged as particularly important in shaping the decision-making process of the model. The examination of feature importance scores revealed that factors such as the **bank balance, age of the clients, day of the month last contacted, days passed since the client was last contacted, and job role** of the client played substantial roles in the decision tree's predictions. Notably, these variables were found to be informative in deciding patterns within the data, and their impact on the target variable. While not all variables were deemed equally influential, the decision tree allowed for the identification of key features that contribute significantly to predicting the desired outcomes. This initial exploratory analysis provides a foundation for understanding the relative importance of different variables in the dataset, laying the groundwork for further investigation and model refinement.

Surprisingly, personal loan and credit default status of an individual did not have a substantial influence when compared to the likes of the top 5 influential features looking at the Figure 7. As discussed in the summary section, **bank balance** almost had a linear relationship with the outcome of the call being fruitful. Therefore, its intuitive that this feature is the most influential in predicting the outcome precisely.

The trend seen in age and job roles of individuals and their subscription rates further fortifies in the feature importance graph. Both **age** and **job roles** are in the **Top 5** most influential features in predicting the outcome of the calls.

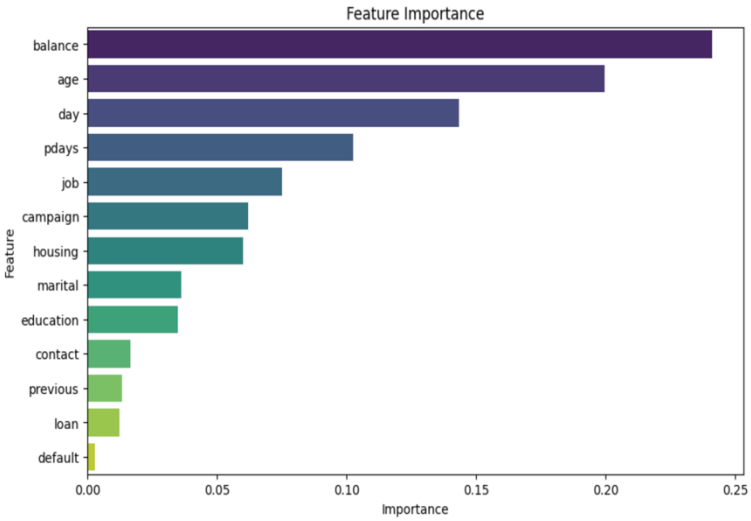


Figure 7: Feature Importance graph

Day of the month an individual was last contacted, and the **number of days passed** since last contact also stand out in the Top 5 influential features. As highlighted in the summarization section, it is noteworthy that the part of the month (beginning, middle and end) the client was last contacted has substantial effect on the outcome of the call.

The method used to contact an individual proved to be less influential in determining whether the call led to a sale. Although the number of previous contacts made to an individual (when selected exclusively, ignoring other features) showed good correlation with the subscription rate in Figure 5, it is not as influential when other features are also considered.

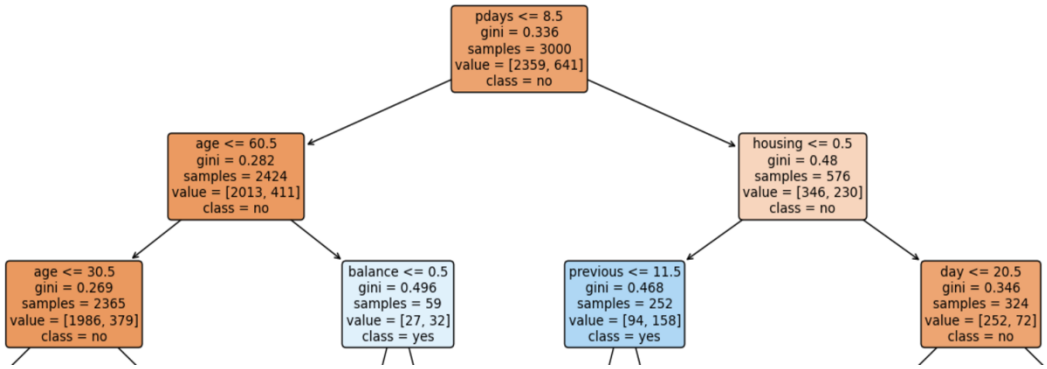


Figure 8: Decision Tree visualization (Top 3 rows)

By default, the Sklearn Decision Tree algorithm utilizes Gini impurity as the criterion for splitting nodes. When the criterion was changed to entropy-based, no significant difference in the tree structure was observed. As depicted in Figure 8, the first three rows showcase the decision tree with the default Gini impurity criterion.

```
Decision Tree Rules:
Rules
-----------------------------------------------------------
|--- pdays <= 8.50
|   |--- age <= 60.50
|   |   |--- age <= 30.50
|   |   |   |--- age <= 21.50
|   |   |   |   |--- day <= 14.50
|   |   |   |   |   |--- class: yes
|   |   |   |   |--- day >  14.50
|   |   |   |   |   |--- balance <= 385.00
|   |   |   |   |   |   |--- age <= 20.50
|   |   |   |   |   |   |   |--- class: yes
|   |   |   |   |   |   |--- age >  20.50
|   |   |   |   |   |   |   |--- class: no
|   |   |   |   |   |--- balance >  385.00
|   |   |   |   |   |   |--- class: yes
```

Figure 9: Decision Tree rules snippet from the root node

```
|   |--- age >  60.50
|   |   |--- balance <= 0.50
|   |   |   |--- class: no
|   |   |--- balance >  0.50
|   |   |   |--- housing <= 0.50
|   |   |   |   |--- education <= 0.50
|   |   |   |   |   |--- age <= 71.50
|   |   |   |   |   |   |--- age <= 65.50
|   |   |   |   |   |   |   |--- class: yes
|   |   |   |   |   |   |--- age >  65.50
|   |   |   |   |   |   |   |--- class: no
|   |   |   |   |   |--- age >  71.50
|   |   |   |   |   |   |--- day <= 26.50
|   |   |   |   |   |   |   |--- marital <= 0.50
|   |   |   |   |   |   |   |   |--- class: yes
|   |   |   |   |   |   |   |--- marital >  0.50
|   |   |   |   |   |   |   |   |--- balance <= 578.00
|   |   |   |   |   |   |   |   |   |--- class: yes
```

Figure 10: Decision Tree rules for a section of the top rows

The decision tree rules offer a detailed insight into the logic followed by the model in making predictions. Figures 9 and 10 are snapshots of rules from the top portion of the tree. Starting at the root node, the initial criterion considers the number of days since the client was last contacted (**pdays**) and whether it is less than or equal to **8.50**. This condition sets the groundwork for subsequent splits in the decision-making process.

Moving through the tree in Figure 9, the first split considers the client's age, examining whether it is less than or equal to 60.50. If this condition holds, the model proceeds to assess whether the age is further restricted. This hierarchical decision-making continues until a split occurs based on an age threshold of 21.50. The next critical decision point involves evaluating the day of the month (day) and determining whether it is less than or equal to **14.50** (First half of the month). If this condition is satisfied, the model predicts a successful outcome, potentially indicating that the client responded positively to the marketing call.

Otherwise, if the client was last contacted in the second half of the month with client's age being less than **20.5** and bank balance less than **GBP 385**, the model still predicts a positive outcome. This snippet conforms with the analysis and exploration made in previous sections, regarding age and balance features.

Figure 10 shows how the tree is structured when the condition for second split fails (i.e. when age > 60.50). The model predicts negatively for old, aged clients with either zero or negative bank balance. Further down the tree, the model checks if the client has any housing loan. As seen in the summary section and now in the decision tree, people having no loan tend to subscribe to the new scheme. Furthermore, the model predicts that if the age of the client is greater than **71.5** and is **divorced** (marital = 0) then it would turn out to be a positive outcome, provided the client was last contacted before 26th day of the month.

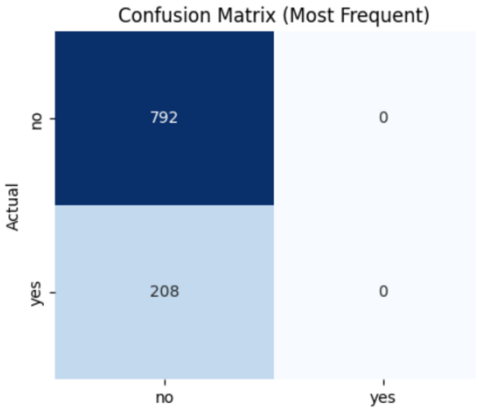
By focusing on key features such as the number of days since the last contact, age, bank balance and day of the month, we've identified patterns that significantly impact the likelihood of a positive outcome. The transparency of the decision tree makes it a powerful tool for uncovering actionable information and informing future strategies in the context of marketing calls and client interactions.

Section C: Model Evaluation

Three models underwent hyperparameter optimization using **GridSearchCV** and cross validation (to avoid overfitting) on the dataset, and their evaluation metrics were tabulated for comparison. Since Naïve Bayes model assumes that the features represent counts or frequencies, and they are **not inherently suitable for handling negative values**, we do not consider this algorithm in our analysis. To establish a baseline, a dummy classifier (Mode) was also applied. This comprehensive approach aids in selecting the most effective model for the given dataset. For all our models, the data is split into training and testing datasets in a **3:1** ratio respectively. As a **call resulting in sale is considered positive outcome**, below is our evaluation glossary:

- TP (True Positives): Count of observations correctly classified as positive.
- TN (True Negatives): Count of observations correctly classified as negative.
- FP (False Positives): Count of observations incorrectly classified as positive.
- FN (False Negatives): Count of observations incorrectly classified as negative.

1. **Dummy Classifier (or) Point Model:**
   A dummy classifier is a simple algorithm that makes predictions using straightforward rules, such as always predicting the **most frequent class** in the training data. It serves as a baseline comparison to assess the performance of other complex models, helping to determine whether a model provides substantial improvements over a basic, rule-based approach.



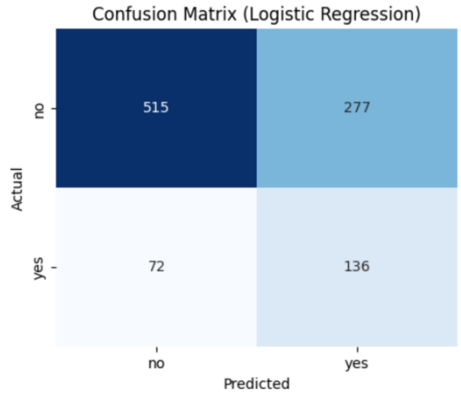
As it is evident in Figure 11, the model always predicted a negative outcome as it's the most frequent class.

Overall accuracy of the model is **0.79.** It is noteworthy that the recall and precision are **0** for positive class.

Classification Report:	precision	recall	f1-score	support
no	0.79	1.00	0.88	792
yes	0.00	0.00	0.00	208
accuracy			0.79	1000
macro avg	0.40	0.50	0.44	1000
weighted avg	0.63	0.79	0.70	1000

Figure 11: Confusion matric and classification report for dummy classifier

2. **Logistic Regression:**
   Logistic regression is well-suited for binary classification problems because it models the probability of an observation belonging to a specific class.



Classification Report:	precision	recall	f1-score	support
no	0.88	0.65	0.75	792
yes	0.33	0.65	0.44	208
accuracy			0.65	1000
macro avg	0.60	0.65	0.59	1000
weighted avg	0.76	0.65	0.68	1000

Figure 12: Confusion Matrix and Classification report for logistic regression model

The hyperparameters used for logistic regression training are:

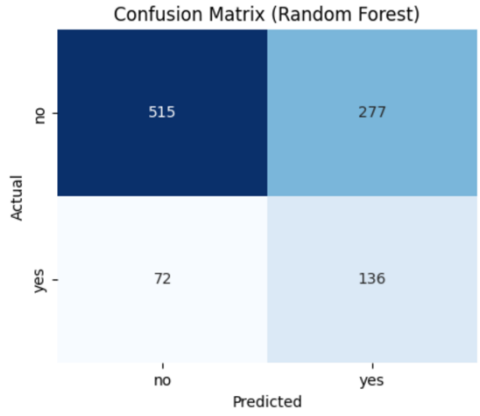
```
param_grid = {
    'penalty': ['l1', 'l2', 'elasticnet'],
    'fit_intercept': [True, False],
    'class_weight': ['balanced']
}
```

After running GridSearchCV, the best parameters turned out to be: {**class_weight**: balanced, **fit_intercept**: True, **penalty**: l2}

Best class weight is balanced as our dataset is highly imbalanced. l2 penalty method is used for loss function so it discourages overly complex models by penalizing large coefficients. And the decision boundary passes through the origin.

3. Random Forest:

Random Forest is an ensemble learning method that combines the predictions of multiple decision trees. By aggregating the results of multiple trees, it tends to provide more robust and accurate predictions compared to individual trees.



Classification Report:	precision	recall	f1-score	support
no	0.86	0.86	0.86	792
yes	0.47	0.48	0.47	208
accuracy			0.78	1000
macro avg	0.66	0.67	0.67	1000

Figure 13: Confusion matrix and classification report for Random Forest

The hyperparameters used for training are:

```
rf_param_grid = {
    'n_estimators': [30, 50, 100],
    'max_depth': [None, 4, 5, 6],
    'min_samples_split': [5, 10, 15],
    'min_samples_leaf': [10, 15, 20],
    'max_features': ['sqrt', 'log2'],
    'class_weight': ['balanced']
}
```

The best hyperparameters turned out to be:
{**class_weight**: balanced, **max_depth**: None, **max_features**: log2, **min_samples_leaf**: 10, **min_samples_split**: 15, **n_estimators**: 100}

This Random Forest consists of an ensemble of 100 decision trees. Increasing the number of trees generally improves model performance, but it also increases computational cost. And the algorithm considers the base-2 logarithm of the total number of features for each split.

4. kNN Classifier:

The k-Nearest Neighbors (kNN) classifier is often chosen for its ability to capture complex, non-linear relationships in the data, making it effective when decision boundaries are irregular or when there is no clear underlying mathematical model. We first find the optimum 'k' value for maximizing the F1-score.

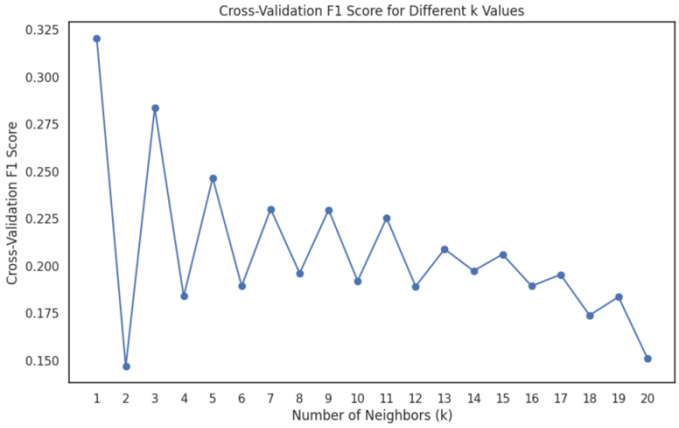
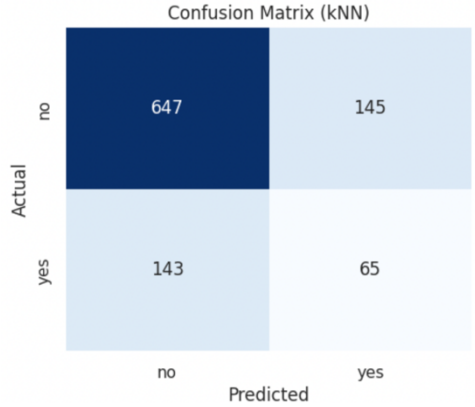


Figure 14: k Vs Cross-Validation F1 Scores in kNN Model

The reason for optimizing F1-score to choose best 'k' value will be illustrated in the next section. The model performs best when k=1, with an F1-score for classifying positive outcomes of **0.31** and an overall accuracy of **0.71.**

Figure 15 shows the confusion matrix and the classification report for kNN classifier.



Classification Report:	precision	recall	f1-score	support
no	0.82	0.82	0.82	792
yes	0.31	0.31	0.31	208
accuracy			0.71	1000
macro avg	0.56	0.56	0.56	1000
weighted avg	0.71	0.71	0.71	1000

Figure 15: Confusion Matrix and Classification Report for kNN model

Section D: Final Assessment

Below is the summary of evaluation metrics for all the models tested earlier:

Model	Accuracy	F1- Score	Precision	Recall
Dummy Classifier	0.7900	0.0000	0.0000	0.0000
Logistic Regression	0.6500	0.4400	0.3300	0.6500
Random Forest	0.7800	0.4700	0.4700	0.4800
k-Nearest Neighbors	0.7100	0.3100	0.3100	0.3100

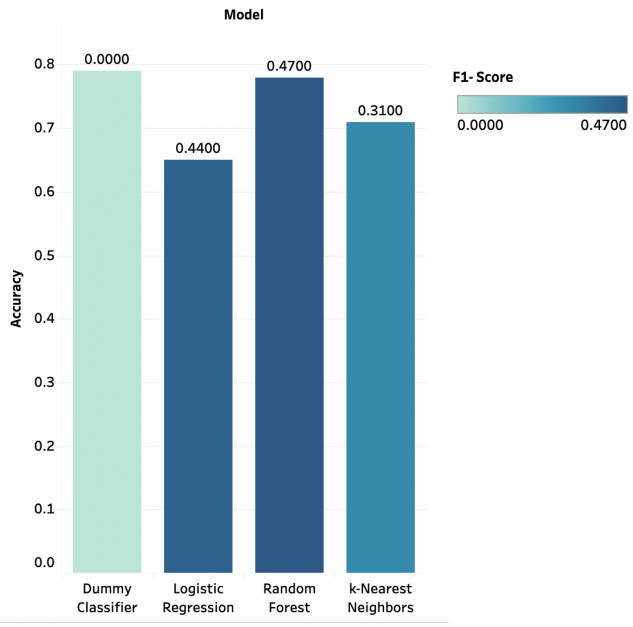


Figure 16: Model Vs Accuracy and F1 Score

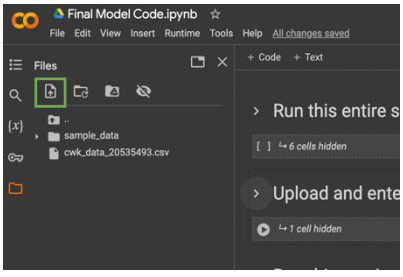
The Chief Data Officer (CDO) of N/LAB recognizes that some customers may find phone calls bothersome, and the real concern is in the expenses incurred from reaching out to clients who aren't interested, wasting valuable staff time. In this context, **the success measure for the classification problem is identified as the F1-Score**. This metric is chosen because it considers both **precision and recall**, offering a balanced evaluation that aligns with the business goal of efficiently identifying genuinely interested customers while minimizing resource wastage.

Going by the above reasoning, **Random Forest** algorithm offers best predictive model for determining if a call will result in a sale. While Logistic Regression also comes close to Random Forest in terms of F1 Score, it lacks in the overall accuracy of the model. Therefore, Random Forest will be chosen as the best model, and it will be trained again on full available dataset.

Section E: Model Implementation

Random Forest model is trained on the entire available dataset and below are the steps to run the model and test the hidden dataset:

1. Open the '20535493_Final_Model_Code.ipynb' file in Google Colab
2. Run the first section entirely. This section performs the following tasks:
    a. Downloads the provided individual dataset from google drive link
    b. Defines python functions to clean the data and train the model
    c. Cleans the dataset and trains the model
3. Upload the hidden dataset (test dataset) to the session storage as shown below:



4. Provide the name of the uploaded test file in the second section of the code
5. Run the third section which will print the confusion matrix and evaluation metrics for test dataset

Assumption: The uploaded test file has the output feature named 'y'.

Section F: Business Case Recommendations

After analyzing the previous campaign's data, demographics and financial status of individuals, below are the business recommendations:

- **Customer segmentation:** The analysis highlights distinct trends in subscription rates across different demographic and financial features. Further segmentation of customers based on these features, such as age groups, job categories, and financial status, could provide valuable insights into **targeted marketing strategies**. Understanding which segments are more likely to subscribe to the "N/LAB platinum" product allows for personalized and efficient campaign planning.

- **Duration and Contact Optimization**: The analysis indicates that the duration of the call and the number of contacts have a significant impact on the subscription outcome. Further investigation into **increasing the duration of calls** and determining an optimal contact frequency could enhance the efficiency of the sales process. It's important to balance effective communication with **avoiding customer annoyance**.

- **Loan Status Influence**: Further exploration of the impact of credit default, personal loans, and housing loans on subscription rates can help in deciding the target audience. For instance, marketing efforts could be tailored to individuals with a good credit score.

- **Education Level Impact**: The data indicates that individuals with a **tertiary education level** are more likely to subscribe. Understanding the specific preferences and behaviors of different education segments can inform targeted messaging and marketing strategies.

- **Collection of data**: The dataset reveals missing data on a crucial feature, '**poutcome**,' which signifies the outcome of the previous call. Since this information is essential for understanding the success or failure of past marketing efforts, it is imperative for N/LAB to prioritize improving data collection and maintenance practices.