Final Report

Problem Statement

In the financial industry, a term deposit is a type of savings account offered by banks and financial institutions. It involves depositing a fixed amount of money for a specific period, often ranging from a few months to several years, at a fixed interest rate.
A Portuguese bank is aiming to improve its customer acquisition marketing strategies for term deposits. By identifying potential subscribers, the bank will tailor its marketing efforts to increase subscription rates.
The insights gained from the model can be used to drive targeted campaigns for reaching out to clients most likely to subscribe, optimizing marketing expenses and improving overall customer satisfaction. This project will be working on predicting whether a customer will subscribe to a term deposit or not. This project also identifies the biggest predictors of term deposit subscription for the bank.

- *In summary, the classification task primarily aims to assist the financial institution in optimizing its marketing efforts by targeting potential customers who are more likely to subscribe to term deposits, thereby increasing the efficiency of their campaigns and ultimately improving the conversion rate.*

Data

The raw dataset contains 45,211 records and 17 attributes, organized in a long format. It contains the following:

- **Bank client data:** age (numeric), Job, marital status , education level, whether the client has credit in default, average yearly balance in euros, whether the client has housing loan, or other loans.
- **Data related to the last contact of the current campaign:** contact communication type , last contact day of the month, last contact month of year and last contact duration in seconds.
- **other attributes:** number of contacts performed during this campaign, number of days that passed by after the client was last contacted from a previous campaign, number of contacts performed for the client before this campaign and outcome of the previous marketing campaign
- **Output variable (desired target)**: subscription

Key Data source:

Moro,S., Rita,P., and Cortez,P.. (2012). Bank Marketing. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.
Link: (https://archive.ics.uci.edu/dataset/222/bank+marketing)
The dataset is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.

Data Wrangling

The raw data is more or less clean and doesn't have any duplicate values, however some features like 'previous outcome' have a high number of 'unknown' values which are essentially missing values.

By analyzing the numerical variables' statistics and frequency distribution, I've identified a few potential issues in the dataset such as:

- **out-of-range data or outliers**.'*previous'* representing past contacts, has an outlier with the value of '275' which seems like an error.
  '*balance'* representing customers' balance in their bank account has outliers on both ends.
  'campaign' i.e the number of contacts performed during this campaign - shows a few instances with values over 30, possibly due to a longer campaign duration.

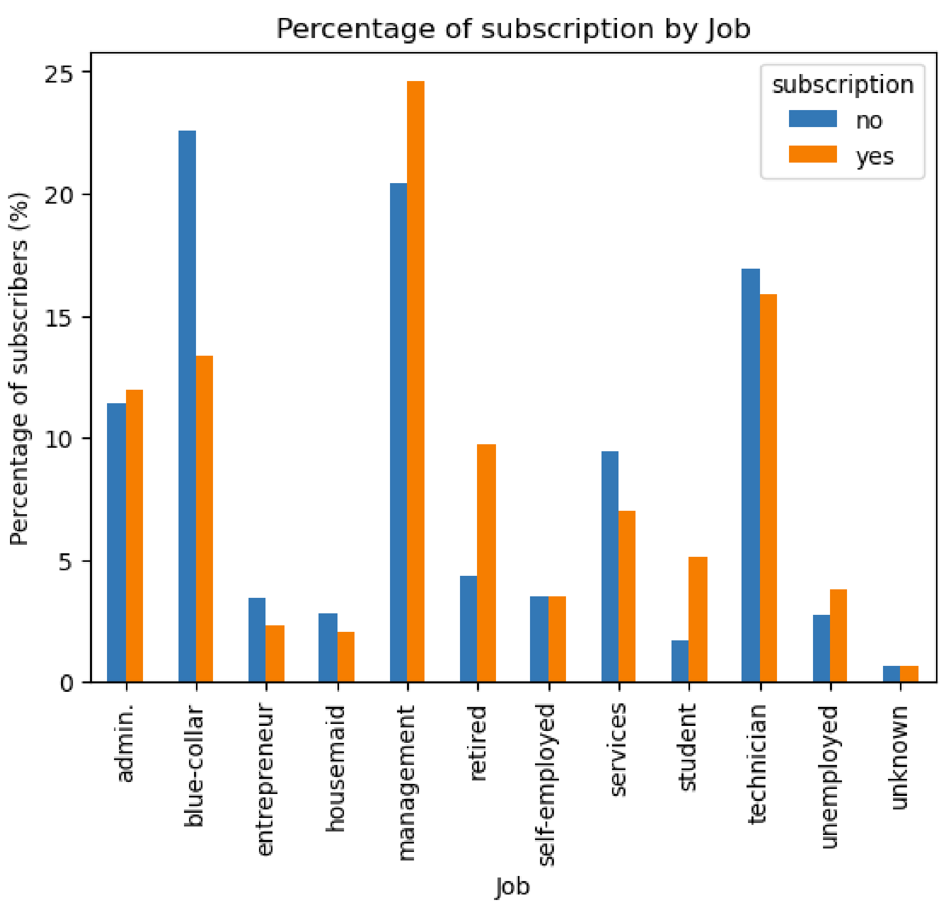
- **Category imbalance:** looking into the categorical variables, they appear to be mostly balanced, except *'defaults'*, which is highly imbalanced.
  '*job'* has many unique categories, which may need potential grouping before modeling.
  The target feature *'subscription'* is imbalanced, possibly requiring techniques to address this issue later on before modeling.

- **Skewness and low variability** - Variables representing the number of days that passed by after the client was last contacted from a previous campaign, the number of contacts performed during this campaign for each client, and duration of campaign calls are highly skewed and have clustered values towards the lower end.
  The variable representing past contacts, has limited variability, mostly having 0 values. Around 81.7% of clients were never contacted previously making the variable one with the least variability which was eventually removed before modeling.

Data Exploration

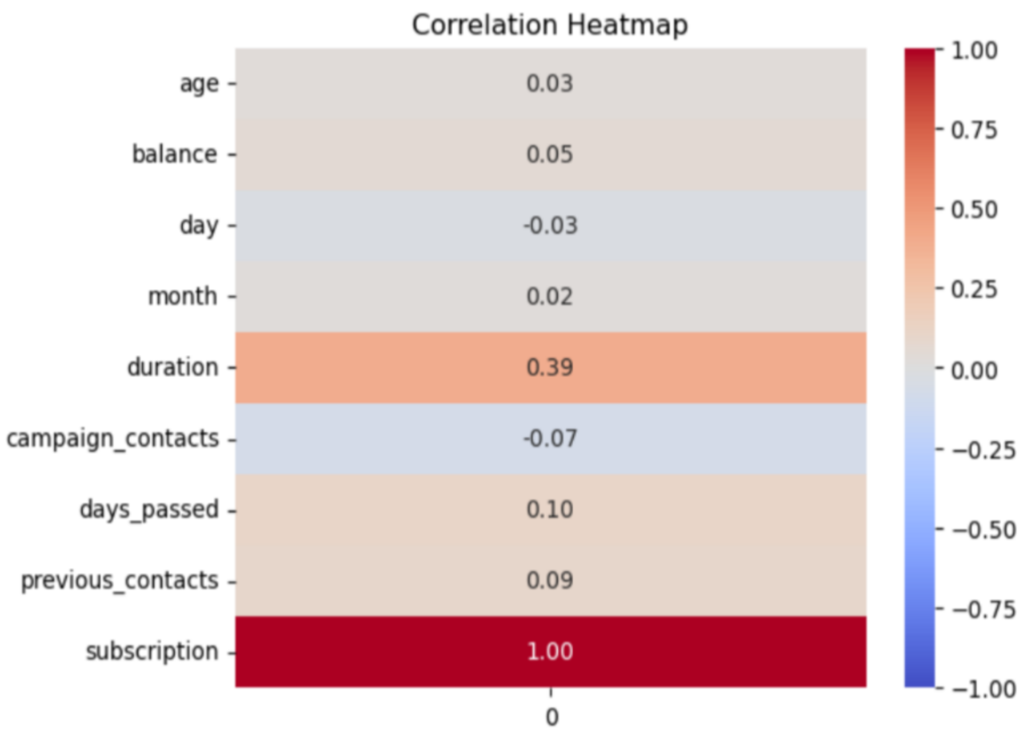
The initial data exploration revealed that there isn't a direct correlation among the independent variables. The following are the relationships that were found between the independent variables and the target.

- `'Age'` of the customer has a nonlinear relationship with subscription i.e. customers above the `age` of 60 and those below 30, are likely to be associated with a higher number of subscriptions.
- `'Job'` of the customer - Those in management as well as retirees and students are associated with a higher number of subscriptions. There appears to be a connection between age and job, where older individuals aged above 60 are likely to be retired and those under 30 students, aligning with the expected demographics.



- Customers without existing `loans` show a higher rate of subscription, aligning with intuitive expectations.

- We can conclude that, overall, longer call `duration` and a moderate number of contacts appear to be linked with a higher number of subscription rates. 'Duration' is the only numerical variable that shows a direct correlation with subscription.*(shown in the heatmap below)*
- Generally, moderately high `balances` (low positives) are associated with higher numbers of subscriptions. However, when plotting `age_group` against `balance` it's clear that only the age group spanning from the 20s to 50s with a higher balance demonstrate a higher mean subscription
- It also appears that higher number of `contacts` made to clients results in adverse effects on subscription.

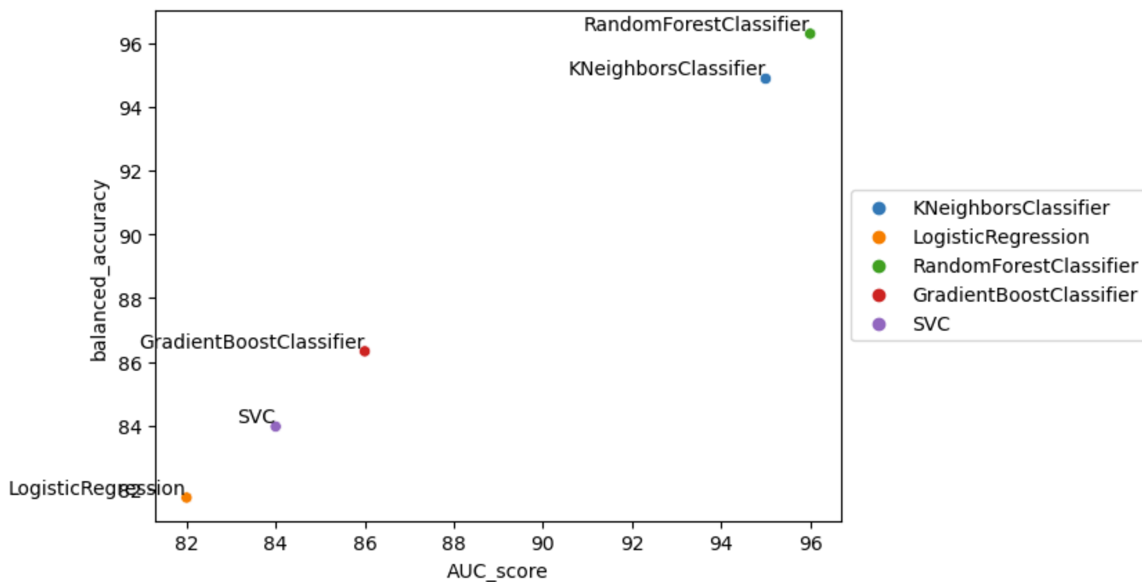


Model Selection

5 classifiers were considered for this classification task. The K nearest Neighbors classifier, Logistic regression model, Random Forest Classifier, the Gradient Boost Classifier and Support Vector Machine.

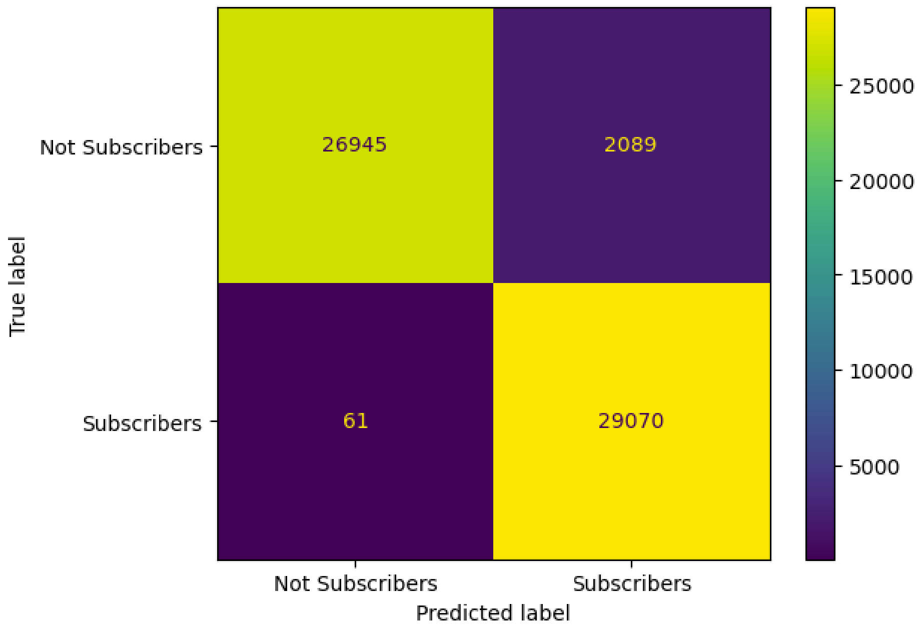
After fitting 5 different models with cross validation of 5 batches, initially, the accuracy metric indicated that KNeighbors Classifier with an accuracy of 94.88 and Random Forest models with an accuracy of 96.3 performed better than the rest. Other evaluation metrics like balanced accuracy, confusion matrix, F1 score, and AUC also identified these two as the best classifiers.

Ultimately, The Random Forest model was selected as the best model and hyperparameter tuning was conducted, which improved the model's performance to 97.67 accuracy level.



Conclusion

The resulting random forest model has an accuracy of 97.67% and a balanced accuracy of 95%. The model resulted in the following confusion matrix, which shows that it has a very high True positive rate.



Based on the results of the Random Forest model - key features impacting the model the most were identified as follows:

- The top features that the model showed to have the biggest impact are: **duration**, **contact_type_unknown**, **balance**, **month** and **age**.
- **Duration** appeared to be by far the biggest predictor of subscription.
- **Job type** and **credit default** showed to have the least impact on subscription.

While this model performs really well at minimizing false positives, The threshold for prediction can be adjusted (lowered) to allow the model to capture more positive instances, at the risk of raising the false positive rate but potentially reducing missed positive cases which could enable the bank to maximize efforts in targeting more customers leading to a higher potential subscription.
Similarly, the threshold can be increased to make the model more conservative in classifying instances as positive,risking increasing the false negative rate; however this would allow the bank to use less resources in the campaign and focus on a smaller number of customers that are highly likely to subscribe. The decision to adjust this threshold depends on the bank's Subscription goals for future campaigns.

Future Research

Exploring the variable 'contact_type_unknown' could be valuable as it was identified as one of the major indicators of subscription. Recalling that, a high portion of the 'contact_type' feature contains the label 'unknown', examining this substantial influence on subscriptions is crucial for a comprehensive understanding