Bank Marketing Analysis

Francisco Valle

1. Summary

The main objective of this project is to predict if clients will subscribe to a bank deposit term through a direct marketing campaign. The analysis was performed using Jupyter Notebooks to obtain an exhaustive visualization exploratory analysis and run the Machine Learning models. In terms of the data preprocessing, a resampling method was used given the imbalanced characteristics of the output variable. Only one in ten customers who were contacted, subscribed. The following models were applied: logistic regression, lasso logistic regression, k-NN and random forest. The best model for predictions was Random Forest. The most important factors for campaign success were related to economic conditions, which included the variables emp.var.rate, euribor3m, cons.conf.idx and cons.price.idx.

2. Models Assessment:

To assess performance of the models and determine its metrics, it is important to understand the business problem. Phone/Mobile marketing campaigns are known to be operational intensive; they need a considerable workforce, and their overall success rate is low. A study from Baylor University at the Keller Research Center study showed a 0.3% success rate during a cold call experiment. These results are reflective of the current marketing trend of more competition and lower conversion rates. Therefore, predicting a successful contact is highly important. In the case presented, the Bank's campaign had a 10% success rate per customer contacted. However, from a business perspective, this implies spending resources on the other 90% of unsuccessful calls.

There are two main objectives of this project: predict the highest number of clients that will subscribe and be efficient on these predictions. In this way we reach out to a more targeted audience which will help both the business growth and save resources.

The two metrics utilized were Recall* and Precision**. Recall served to determine how effective the model is in predicting the highest number of subscriptions. And Precision determined how efficient were those predictions.

Additionally, as a sanity check, we used the AUC which is a statistical value that provides a general assessment of how correct our models performed.

3. Data Source

The dataset is a list of 41,188 clients that were contacted during the marketing campaign to their telephone or mobile. Clients were contacted one or more times. The dataset has 20 features (10 categorical and 10 numeric) and one output variable, which is binary a yes/no if client subscribed. Assumption was made the bank is located in Europe given that the dataset includes variable called eurobor3m (the average interbank interest rate at which European banks are prepared to lend to one another) .

A summary of the data can be found in the Jupyter Notebook (section 1.1). More information on each variable can be reviewed on the challenge description.

---

* Recall: True Positive / (True Positive + False Negative)

** Precision: True Positive / (True Positive + False Positive)

4. Data Preprocessing and Feature Engineering

4.1. Imbalanced data

As stated before, the data is imbalanced and there is an overall 10% success rate. This might affect the predictions of our models. We managed these issues in 2 ways:

- Recall and Precision – which allowed to look for imbalances in our predictions. These metrics usually have a trade-off. For instance, if we predict all clients will subscribe, this means a 100% Recall but a very low Precision value (near 10%).
- Resampling of the data – we randomly selected samples that were successful and used them to balance the number of unsuccessful outcomes. We ran the models using both the original and the new balanced data to confirm. As expected, the balanced data had overall better results, which is what we will discuss in this summary report.

4.2. Data with Missing Values

There were 6 variables with missing values, all of them were categorical variables. The missing values % are shown in the following table:

	Values Missing	% of Total Values
default	8597	20.9
education	1731	4.2
housing	990	2.4
loan	990	2.4
job	330	0.8
marital	80	0.2

After reviewing each variable's characteristic and their distribution, the missing (null) values were handled in three ways:

- Generated a new category:
  - For the default variable a new category called "unknown" was created to replace the null values. This is because null values had very different success rates (5%) compared to other categories (13%).
- Used the most frequent category:
  - The variables education, housing, and loan were imputed using the most frequent category. The null values had similar behavior to the most frequent categories.
- Deleted rows:
  - This was used on the variables marital and job which have less than 1% of null values. Because of their lower impact, those rows were deleted.

4.3. Feature Engineering

We performed two modifications to the features:

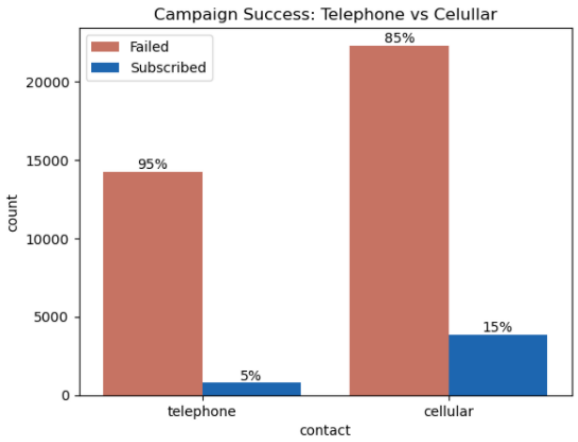
- Removed variables:
  - day_of_week. This variable did not have considerable impact on the success rate nor had variability between the number of calls between days. For that reason, it was not used in the models.
  - duration was removed because it is a variable unknown before contacting clients. It is however a proxy for the success or failure.
- One-hot encoding: Categorical values were transformed using one-hot encoding. For each category, a binary column was generated to run the models (one column is removed due to multicollinearity).

We attempted to keep unmodified original variables as much as possible so they can be recognized and used appropriately by the business teams.

5. Exploratory Data Analysis

From a business perspective, the dataset had three main categories:

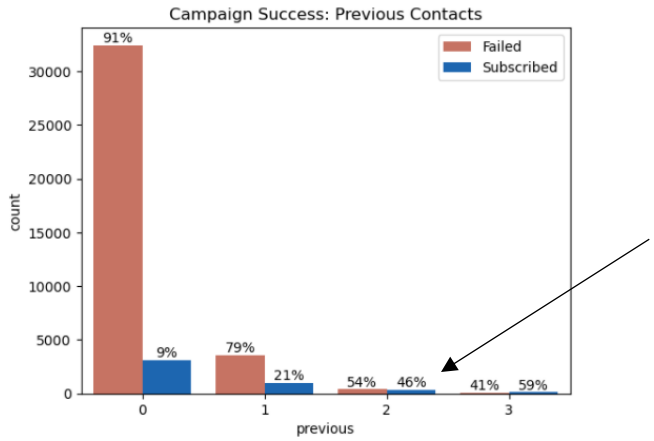
- **Campaign operations variables**: these variables describe the marketing campaigns and if modified, can have a positive impact in the success of the campaign. Variables: contact, month, day_of_week, duration, campaign, pdays, previous, poutcome were the most impactful:
  - **contact:** marketing team should focus on obtaining clients mobile number instead of home telephone. The following plot clearly shows that mobile has three times more successful contacts compared to using telephone (15% vs 5%)



Of note, bar size represents the quantity of clients.

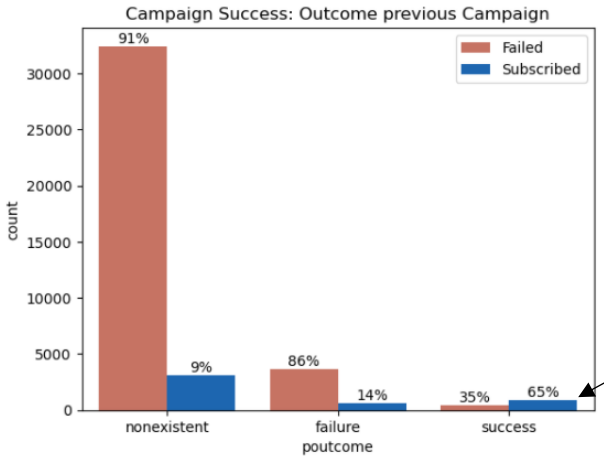
Percentages noted reflect fail vs subscribed proportion of clients for the same category

- **previous**: variable that provides how many times a client was contacted before the campaign by the bank. This had a positive effect, a customer that was contacted two times prior had a 46% subscription rate vs 9% that weren't contacted. This might be related to customer service and keeping the clients more engaged with the company.



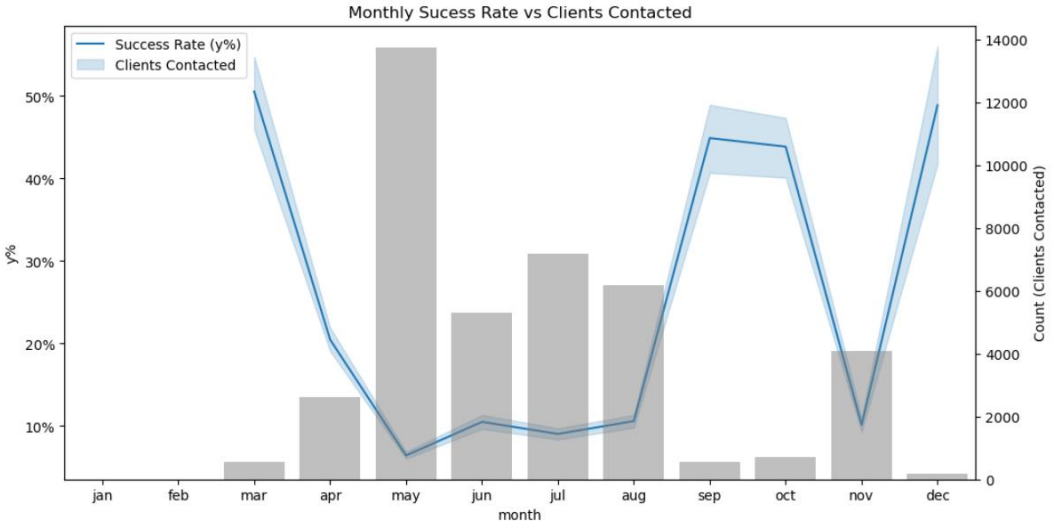
Clients contacted 2 times before have a **46%** subscription rate, but very low quantity of clients

- Variables related to previous campaign (**pdays, poutcome): pdays** is the number of days that passed by after the client was last contacted from a previous campaign and **poutcome** is the outcome or failure of the previous campaign. Both variables impact the success of the current campaign. The plot of poutcome shows it clearly, however we need a bigger sample of successful previous campaigns to have a bigger impact on our model.

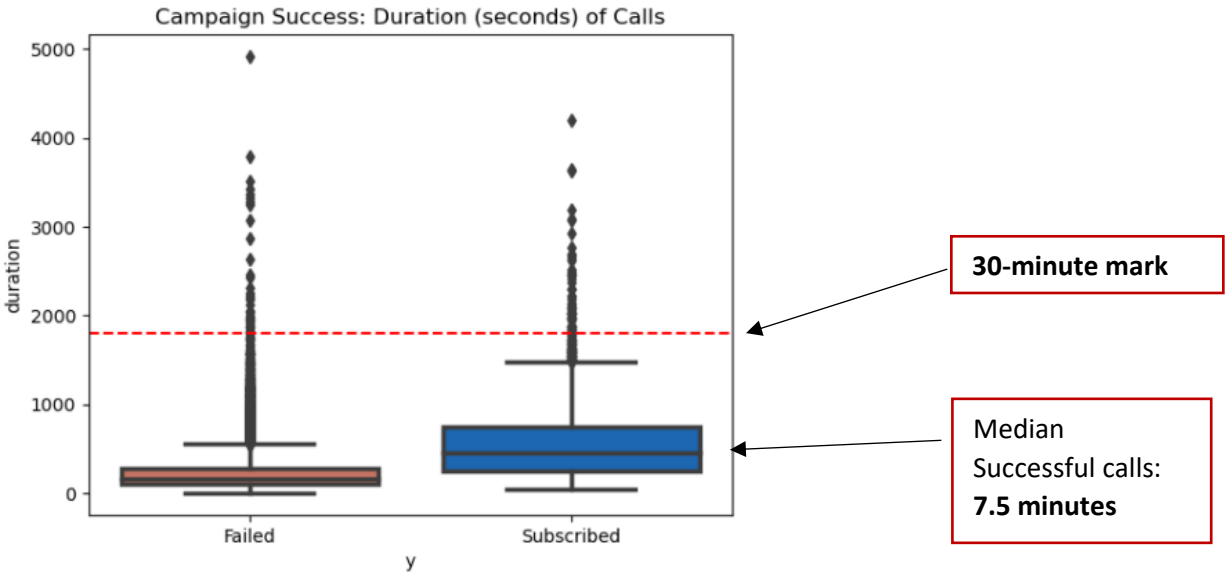


Impressive **65%** subscription rate, but very low quantity of clients

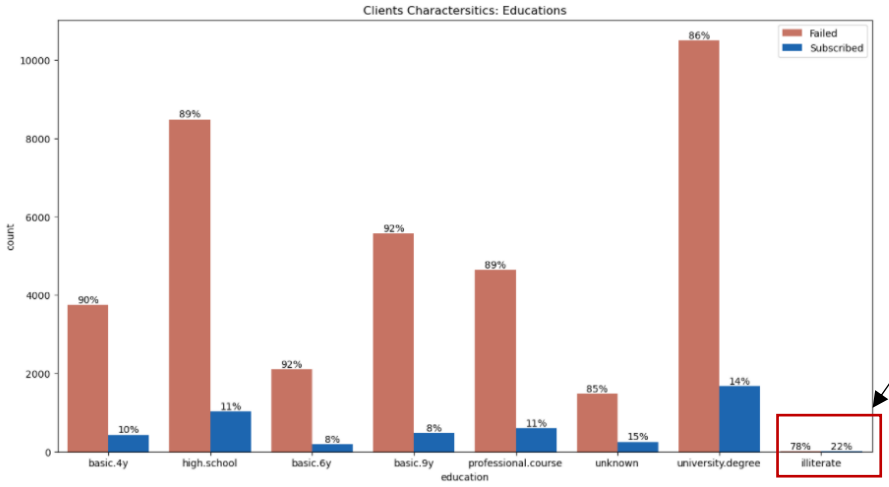
- **month:** this variable provides the last contact month of year. Certain months have impressive success rates such as September and October which both have above 40% (blue line). However, those same months have fewer people being contacted (grey bars). The opposite happens in May, June, and July, with low success rates and higher number of clients contacted. Marketing team might call more but the total results stay stagnant. Therefore, we need to increase the number of potential good leads to maintain a high success rate.



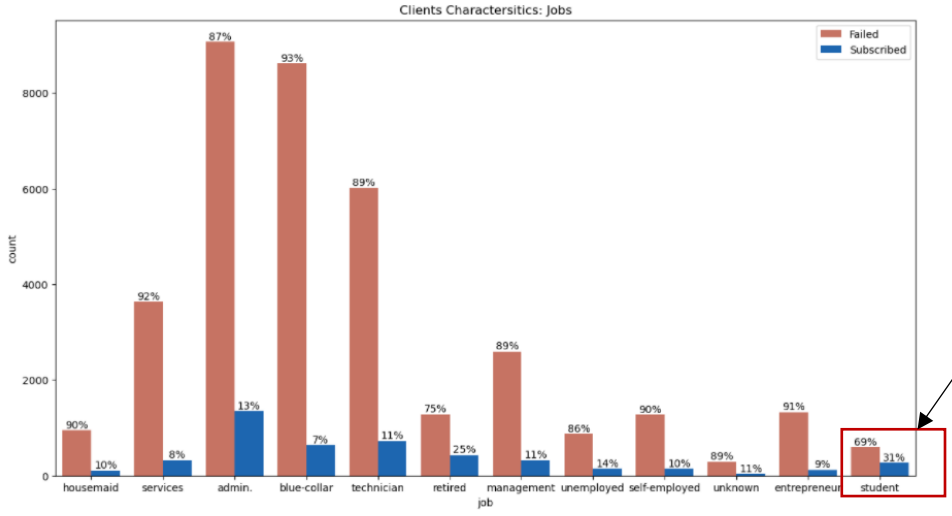
- **duration:** last contact duration, in seconds (numeric). This variable contains outliers calls above 30 minutes duration. The company should investigate these cases which can reflect an operational problem or could represent people talking about things unrelated to the campaign. The next boxplot shows the considerable quantity of outliers above the red line which is set at a 30 minute threshold.



- **<u>Client variables</u>** (age, job, marital, education, default, housing, loan): These variables help provide background and characterize the customers contacted, improving segmentation. In our case, we had a majority of customers who were young adults and middle-aged people with a median of 38 years old. Most of them work in administration roles, blue collar or as a technician.
    - Variables **job**, **education** and **default** do not show a major impact on the success rate. However, there are some exceptions, such as clients with a credit default which do not subscribe and illiterate and students clients which tend to subscribe more. With a 20% and 31% subscription rate respectively. These groups are relatively small as shown in the plot.

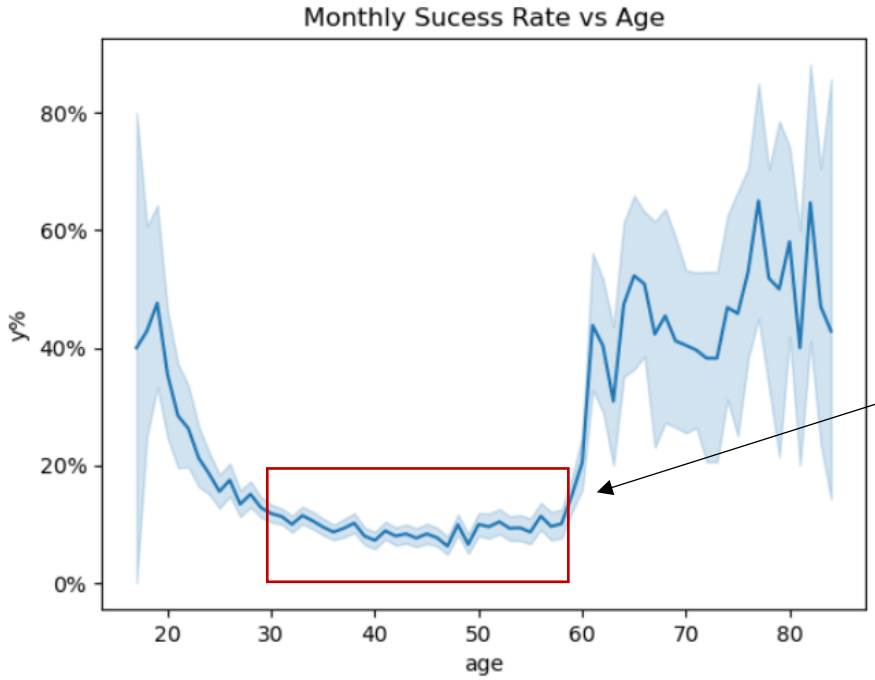


Illiterate: Highest subscription rate **22%,** but very low quantity of clients



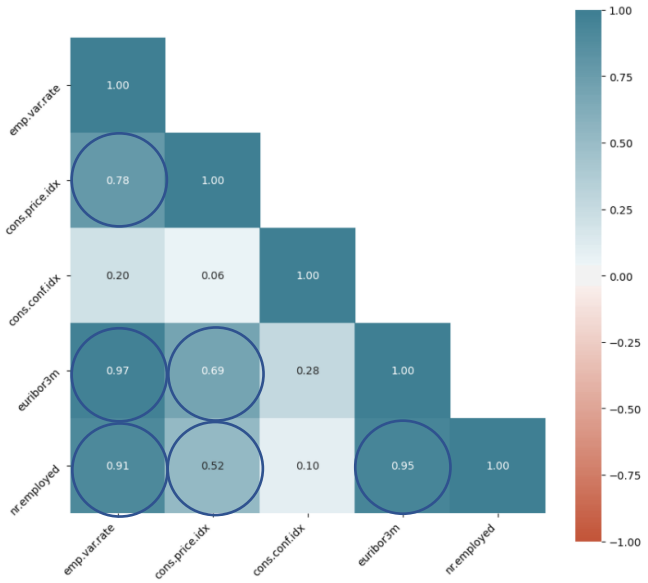
Students: Highest success rate **31%,** but very low quantity of clients

- In terms of age, there is a higher success rate in the younger and older population. However, most of the clients are between the ages of 30 and 60. This is both a challenge and opportunity to focus in the younger and older segments.



Low success rate corresponding to over 80% the clients

- **<u>Social and economic variables</u>**: These variables provide a general context of the potential success of the campaign. They are useful to decide when to allocate resources or modify the value proposition. These variables had a big impact on the success of the marketing campaign.
  - According to the correlation matrix, euribor3m, nr.employed, emp.var.rate, and cons.price.idx had the biggest impact. They are highly correlated to each other which is useful to simplify decision making. By analyzing one or two of these variables we can have an overall assessment of the economic conditions.



Very high correlations among variables, some of them very close to 1, almost perfectly colinear. **nr.employed** or **euribor3m** are good proxy value for this category

Next, we ran the models to further analyze the impact of the variables on the success rate of the campaign.

6. Statistical Modeling

We used a total of 4 models. We randomly split the data in 80% training and 20% testing, the detailed methods and code can be found on the Jupyter Notebook. The characteristics of each of them are the following:

Logistic Regression:

Logistic Regression:

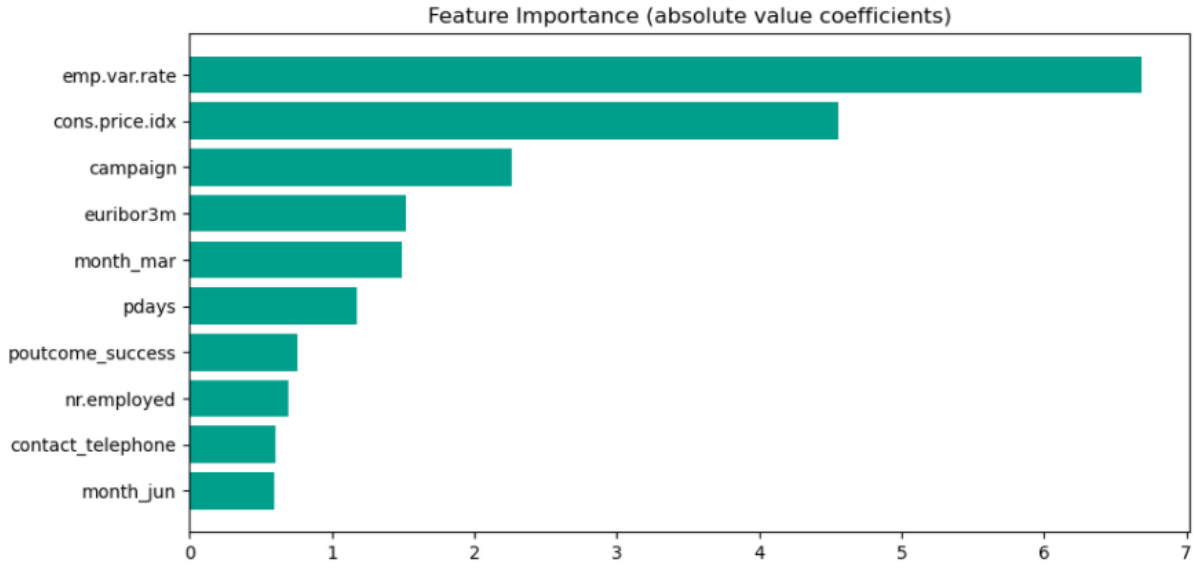
It is a classification model, which uses the logistic sigmoid function to obtain the probability of obtaining the result, which is later used to predict a binary output. The sigmoid function is constrained between a 0 and 1 probability. One of the useful aspects of this model is that it provides a coefficient per variable. This can be used to assess the importance of such variable or even calculate the monetary impact of business decisions (we will evaluate this with the regularized model).

Regularized Lasso Logistic Regression:

It is a modified version of logistic regression model with the difference being that it includes a penalty for variables that do not have a major impact on the model predictions. This can be used for feature selection. Ineffective variables will have low coefficients or even zero. In our case, it will be used to drop variables that are highly correlated to each other.

The most important variables according to the coefficient (absolute value) of the logistic regression are the following:

- Social economic: emp.var.rate, cons.price.idx, euribor3m, nr.employed
- Customer: job (not included in the top ten variables)
- Operation: month, pdays, poutcome, contact, campaign



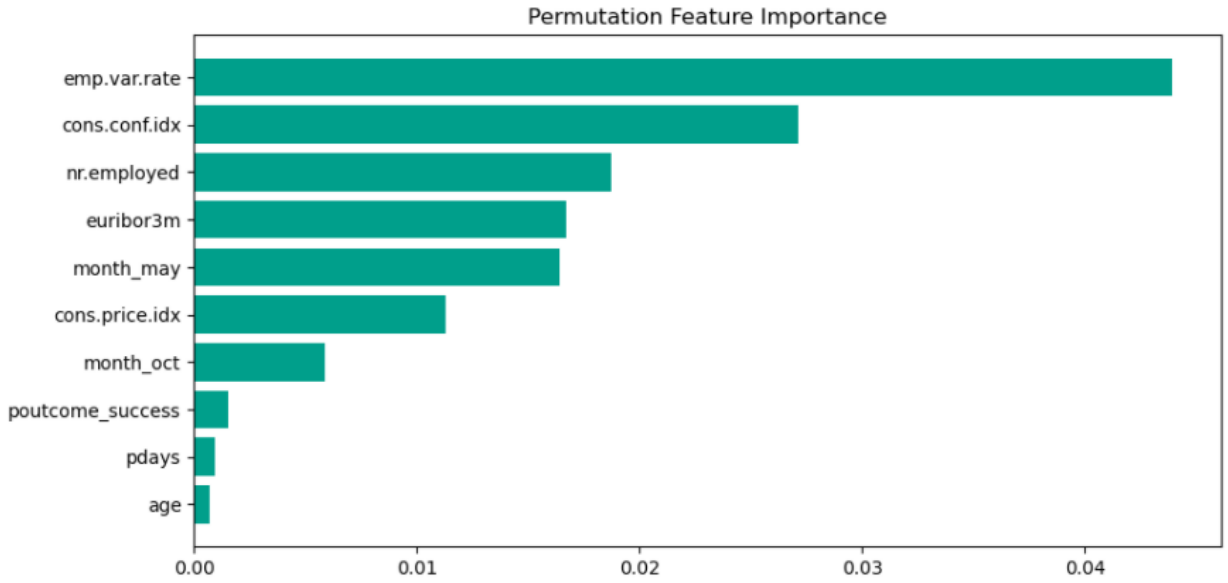
K-Nearest Neighbor (KNN):

This algorithm is more flexible than the logistic regression because it has fewer assumptions and lower biases. The basic logic is that similar things exist in proximity. The algorithm calculates the Euclidian distance to the closest k-number of neighbors and with this, obtain a majority vote from those neighbors to predict one of the two classes. However, it provides less interpretation to use for business decision making.

Random Forest:

Random forest is a classification algorithm that consists in generating a group of decision trees from which a majority vote is computed to select the final prediction. This model is very robust to outliers and works well with non-linear data. It does not provide coefficients as logistic regression does; however, it provides interpretability by providing variable importance. The most impactful being:

- Social economic: emp.var.rate, euribor3m, nr.employed, cons.price.idx, cons.conf.idx
- Customer: age
- Operation: campaign, previous



7. Results and Model Selection:

Each of the models output using the balanced dataset are described as follows:

	precision	recall	AUC
Logistic Regression - Balanced	0.360202	0.622416	0.741032
Logistic Lasso Regression - Balanced	0.360656	0.622416	0.741171
KNN - Balanced	0.553571	0.202394	0.590836
Random Forest - Balanced	0.404971	0.602829	0.745191

As stated, our main metrics were recall and precision. Logistic regression, lasso logistic regression, and random forest models all achieved very similar levels of recall (around 62%). However, Random forest had the highest precision, meaning, its predictions were more correct in finding customers that subscribe. Therefore, we selected this model as the main model. Additionally, this model had the best overall AUC value.

The Random Forest model also provided actionable information to use for the campaign with its feature importance. The feature coefficients of the Lasso Logistic Regression were used to better understand the impact of such variables.

8. Business Recommendation:

The **socio-economic variables** had the most impact according to the models. The emp.var.rate (employment variation rate) had a negative impact on subscription, while the euribor3m, cons.price.index, and cons.conf.idx had a positive impact on subscriptions. This clearly shows that clients subscribe more with better and stable economic conditions. Banks should focus their campaigns efforts when the job market is stable, the interest rates are higher, and the economy is growing.

Additionally, nr.employed (the bank's workforce growth) affects positively the campaign results. Hence as the bank hires more peoples it gets more subscriptions. This could be explained by brand recognition, better service, and more people working on the campaign.

Regarding **client variables,** our visualization analysis showed they are not a major contributing factor for subscriptions. However, the models confirm that age presented a nonlinear impact on subscriptions and a potential new segment to target would be students and the retired.

Month was the most important **campaign operations variables**, which impacted positively or negatively depending on the month. More information is needed to understand the reason for such difference.

The bank should also focus on obtaining more cell phone numbers from clients instead of telephone numbers given the positive impact on subscriptions. Additionally, the bank should target clients that were contacted in previous campaigns and, even better, if they had subscribed in the past given the negative impact of pdays and the positive impact of "success" from the poutcome variable.