BANK MARKETING CAMPAIGN

Group Name:

Sparagua

Name, Email:

- Daniela Alvarez
  daniela.alvarez04@gmail.com
- Akhil Nair
  akhil.nair1908@gmail.com

Country:

- Peru
- India

College/Company:

- Daniela: Datacamp, Kaggle Learn, Universidad de Piura (Peru)
- Akhil: SIESGST Nerul, Navi Mumbai

Specialization:

Data Science

Problem description:

Portuguese bank is having a decrease in its revenue. The bank wants to be able to predict which clients are most likely to subscribe a term deposit. Before selling its term deposit product, it wants to develop a model based on the interactions of these clients with other banks in the past. This will allow them to focus their marketing efforts and resources on these select few customers with better chances of subscribing. This will help significantly reduce the expenditure of time and energy on clients that will probably not subscribe.

Exploratory Data Analysis:

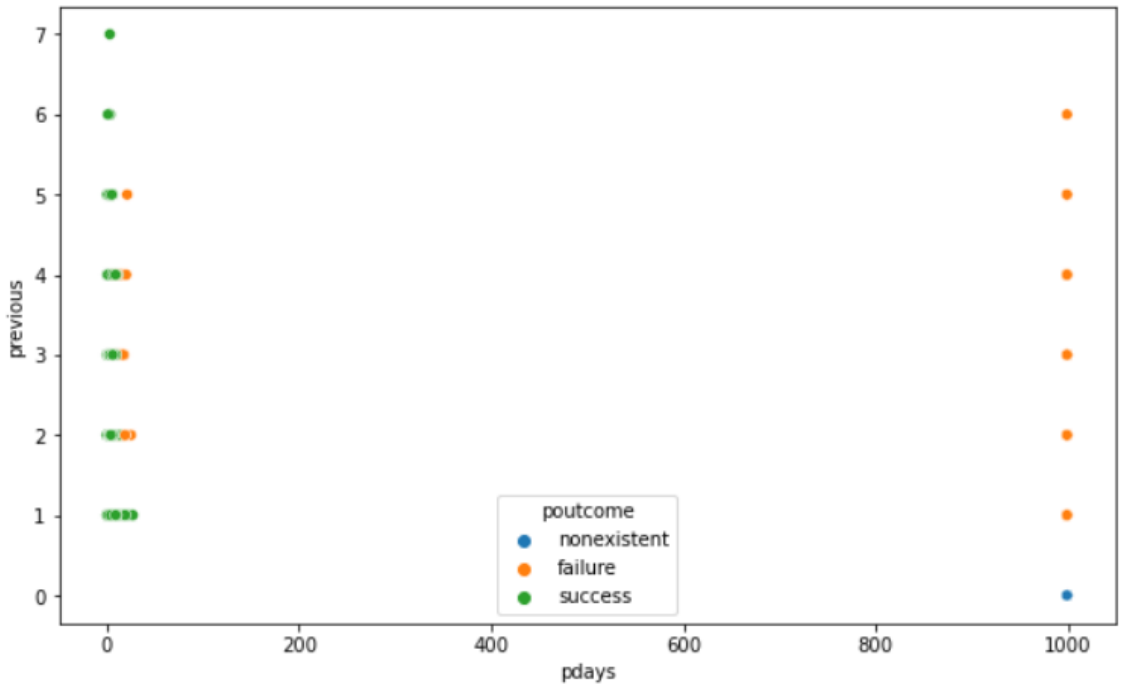
For exploratory Data Analysis, we have separated our prediction target from our features. A summary of our findings of this data has been listed below.

Categorical Data:

1. Jobs: The most common jobs are administrative, blue collar and technical jobs, whereas the least common ones are students, housemaids and unemployed individuals.
2. Marital status: Most individuals are married.
3. Education: Most of the potential customers have a college degree, or a high school degree. Very few are illiterate.
4. Defaults: Most of them have no defaults.
5. Housing Loans: Most of them have a housing loan.
6. Personal Loans: Most individuals do not have a personal loan.
7. Contact: Most individuals have listed their preferred contact method as cellular phone over telephone.
8. Month of contact: The last month of contact for most of them by far is May, followed by July, August and June.
9. Day of contact: There is an even distribution in the last day of contact among the targets.
10. Previous Outcome: Most of the past data shows us a nonexistent outcome, but this time a good portion of the individuals answered even made a term deposit.

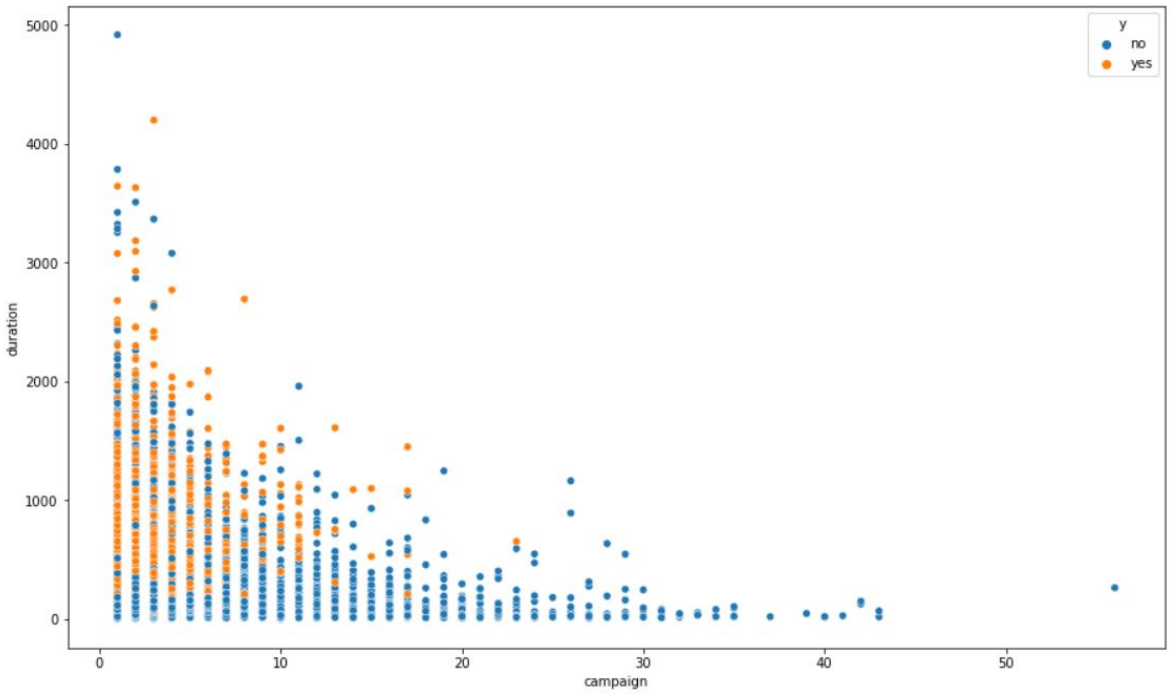
Numerical Data:

1. Relation between Pdays and Previous:
   Here ‘Pdays’ is the number of days that passed after the client was previously contacted, and ‘Previous’ is the number of contacts made earlier for the client.



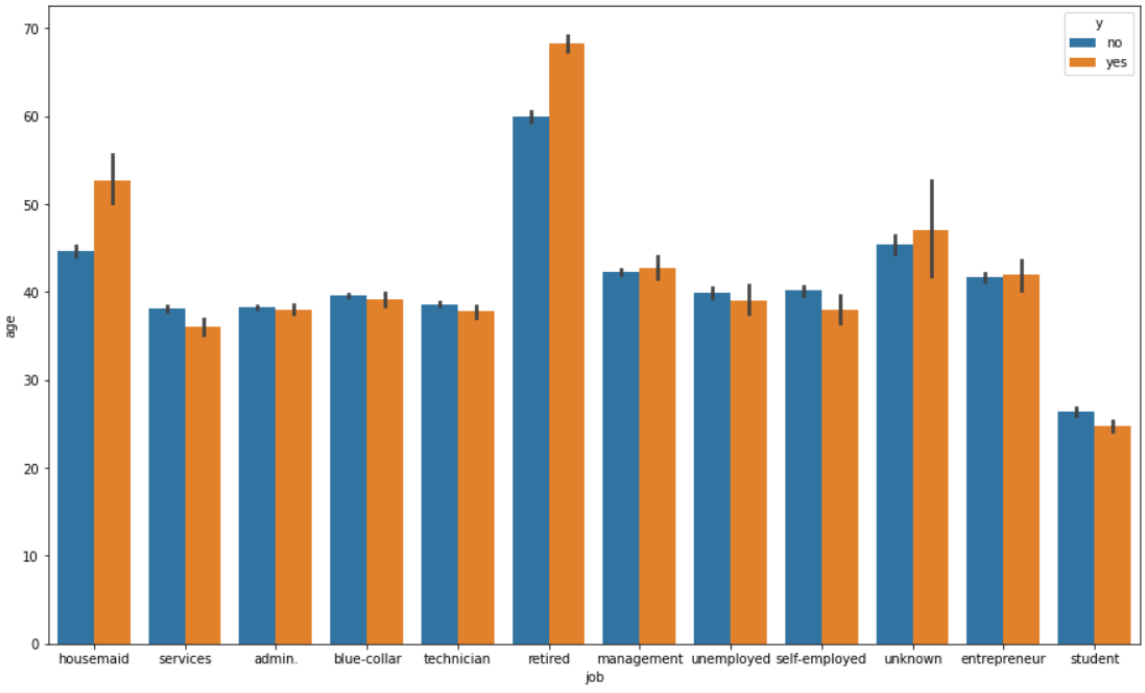
The 999 valued blue dot confirms that these people were never contacted before, hence there was no previous outcome. The orange dots on the left side represent negative responses from people contacted 2-5 times, and the green dots represent positive responses from people contacted 1-7 times. There have however, been instances where people have never been contacted but gave a negative response (orange dots on the right), so to avoid any confusions in the future, we shall eliminate the ‘Pdays’ feature and use the ‘Previous’ feature to know whether someone had previously been contacted.

2. Relations between call duration, call frequency and the corresponding outcome:



The obtained graph shows that there exists an inverse relation between the number of calls made to a prospective client and the duration of these calls. The lesser the frequency of calls, the longer the calls go on. Clients that have been called over 12 times do not respond, or give a very brief response.

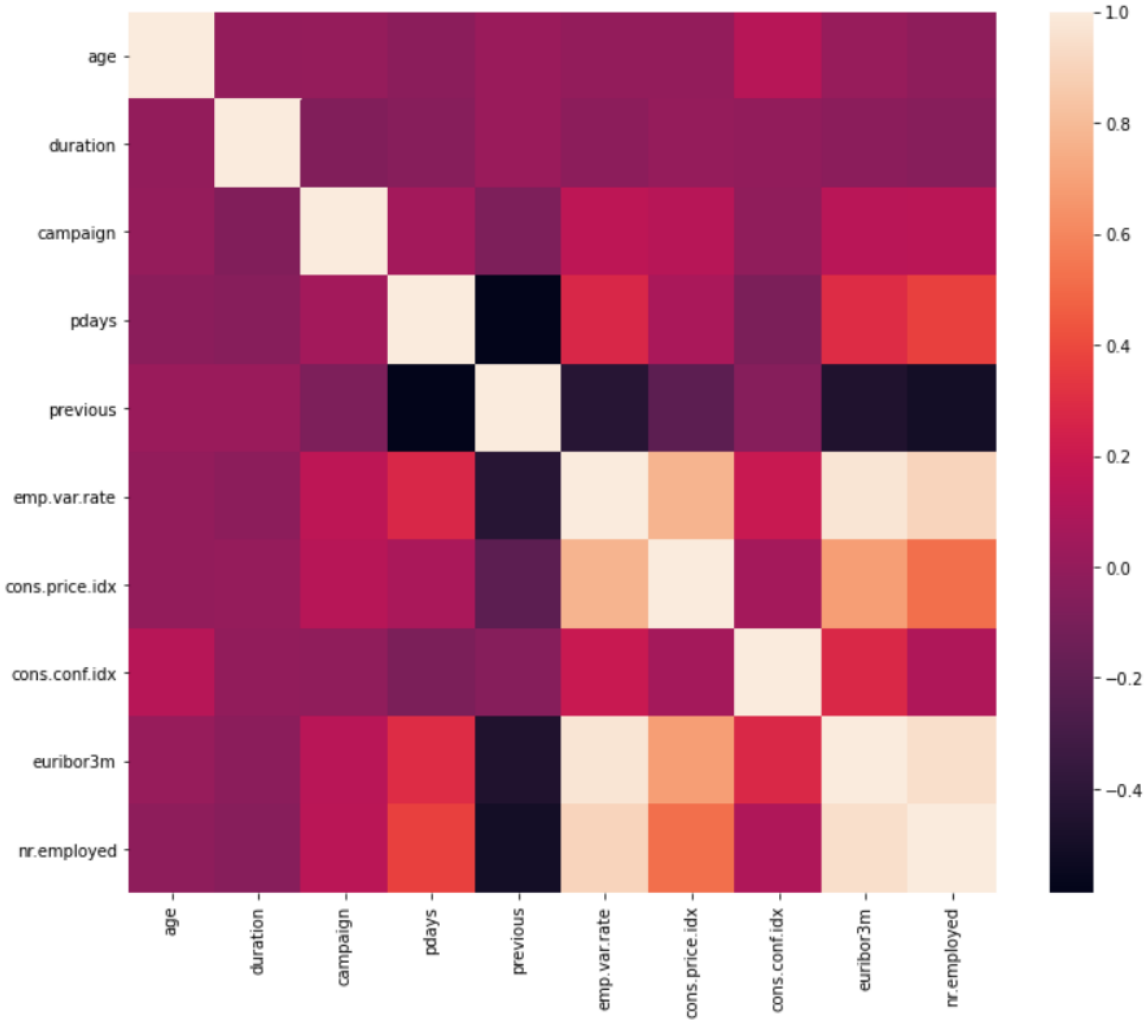
3. Relations between job and age:



There seems to be a larger difference between the ‘yes’ and ‘no’ subscribers among the retired people aged 65-70 years old and the housemaids aged 50-55 years old than the other potential clients.

EDA Recommendation:

A heatmap of our features will help gain even more insight:



As we can clearly see, the consumer price index is strongly correlated with the bank’s interest rates and employee variation rate, i.e., the higher the price index, the greater the interest rate. The employee number also has a strong correlation with the employee variation rate and bank interest rates.

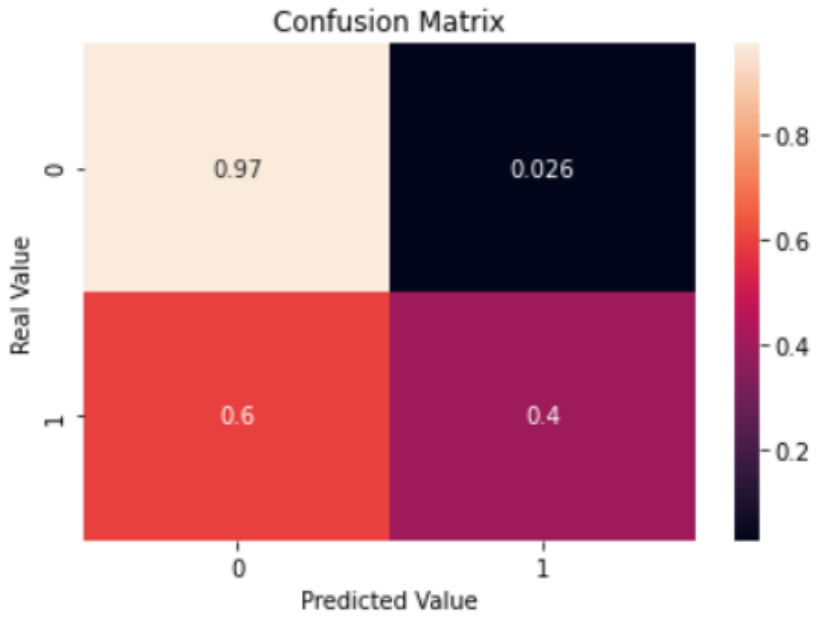
We can say with confidence that the frequency of contacts made with the prospective clients has a very strong negative correlation with the bank’s interest rates and employee variation rates, i.e, the greater the rates of interest, the lesser the number of contacts that had been performed before this campaign. A lower interest rate could therefore increase the number of contacts made this campaign.

Model Selection and building:

As our primary goal is to predict if a deposit will be made or not, the output would be binary. Classification models would therefore be our best bet.
We have used Grid Search to optimize our hyperparameters. Performing cross validation among classification models, we found these to be the best models for this case to be:

1. Logistic Regression: A supervised learnin algorithm that helps us predict a dependent categorical variable using independent variables.

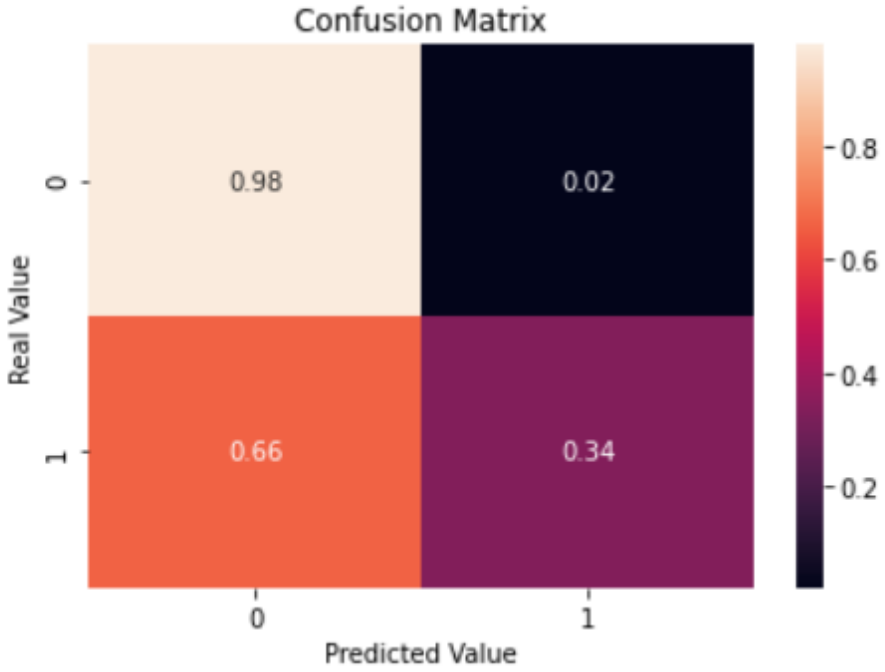
	precision	recall	f1-score	support
0	0.93	0.97	0.95	10931
1	0.67	0.40	0.50	1413
accuracy			0.91	12344
macro avg	0.80	0.69	0.73	12344
weighted avg	0.90	0.91	0.90	12344



As we can see, this model did really well for being the base model, with an F-1 score of 0.5 and Precision of 90% and Recall of 91%.

2. Support Vector Classifier (SVC): Is a supervised learning method that utilizes support vectors (coordinates) and hyperplanes to separate the two classes.

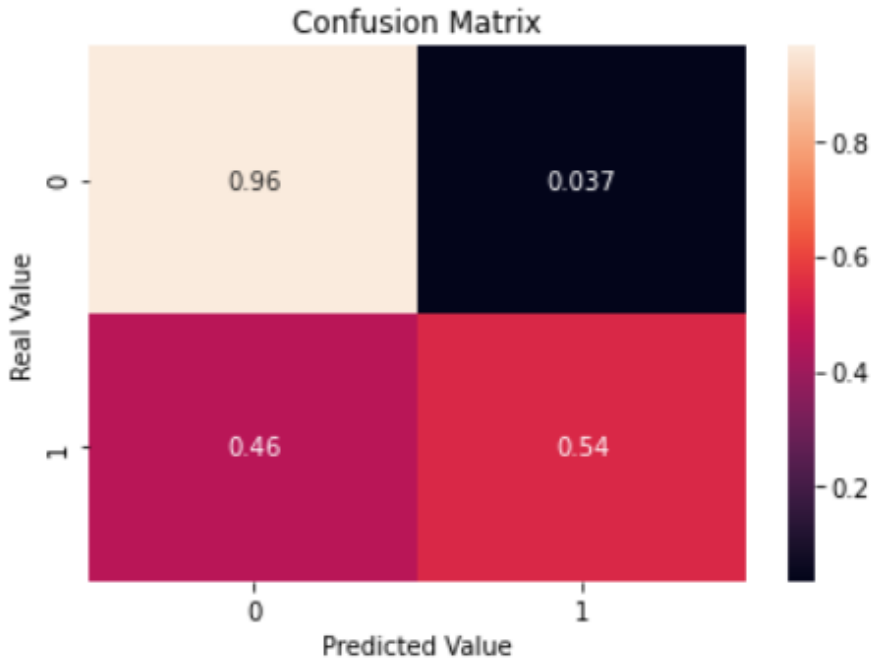
	precision	recall	f1-score	support
0	0.92	0.98	0.95	10931
1	0.68	0.34	0.45	1413
accuracy			0.91	12344
macro avg	0.80	0.66	0.70	12344
weighted avg	0.89	0.91	0.89	12344



As we can see, this model also did well, with an F-1 score of 0.45 and Precision of 89% and Recall of 91%.

3. XGBoost Classifier: Extreme Gradient Boosting Classifier uses gradient boosting (gradient descent algorithm) to minimize losses. It is a very widely used model that is designed for speed and high performance, due to which it finds usage in competitive machine learning.

	precision	recall	f1-score	support
0	0.94	0.96	0.95	10931
1	0.66	0.54	0.59	1413
accuracy			0.91	12344
macro avg	0.80	0.75	0.77	12344
weighted avg	0.91	0.91	0.91	12344



As we can see, this model has performed the best, with an F-1 score of 0.59 and Precision of 91% and Recall of 91%.

All three of the models that we have explored have performed better than our expectations, with very slight differences in their precision, recall and F-1 score. However, the XGBoost model has edged the other two out by a small margin, and this is the model that we would select for our purpose.

Final Recommendation:

As the Portuguese Bank required a model to check which customers will make a deposit or purchase a subscription, we conducted EDA on the data to understand hidden trends. We realized that a classifier model is most suitable for our task, so we compared three of the best models for our case, which are - Logistic Regression, SVC and XGBoost. Upon comparing these models using F-1 score, recall and precision, we found all three of them to perform well.

However, XGBoost Classifier stood out by a little, and would be the model that we would use to shortlist customers with the best chances of success. This is the Machine Learning model that would be most efficient in selecting clients that are most likely to subscribe and deposit money, and this is our final recommendation to Portuguese Bank after conducting our analysis and research.

Github Repo link:

https://github.com/danielaaz04/Bank-Marketing-Campaign

https://github.com/oak-hill/Bank-Marketing-Campaign