Exploratory Data Analysis

and proposed modelling techniques for business users

Bank Marketing Campaign

Agenda

1. Team member details
2. Problem Statement
3. EDA and EDA recommendation
4. Model Selection
5. Model Results
6. Final recommendation

Meet the team, ‘Sparagua’, specializing in Data Science:

- ***Daniela Alvarez*** (daniela.alvarez04@gmail.com)
  Country: Peru
  College/Company: Universidad de Piura (Peru), Datacamp, Kaggle Learn

- ***Akhil Nair*** (akhil.nair1908@gmail.com)
  Country: India
  College/Company: SIESGST Nerul, Mumbai University

Problem Statement

Experiencing a decrease in revenue, Portuguese bank now wants to predict which clients can subscribe to a term deposit.

Based on past activity, they want to develop a model to identify customers most likely to subscribe.

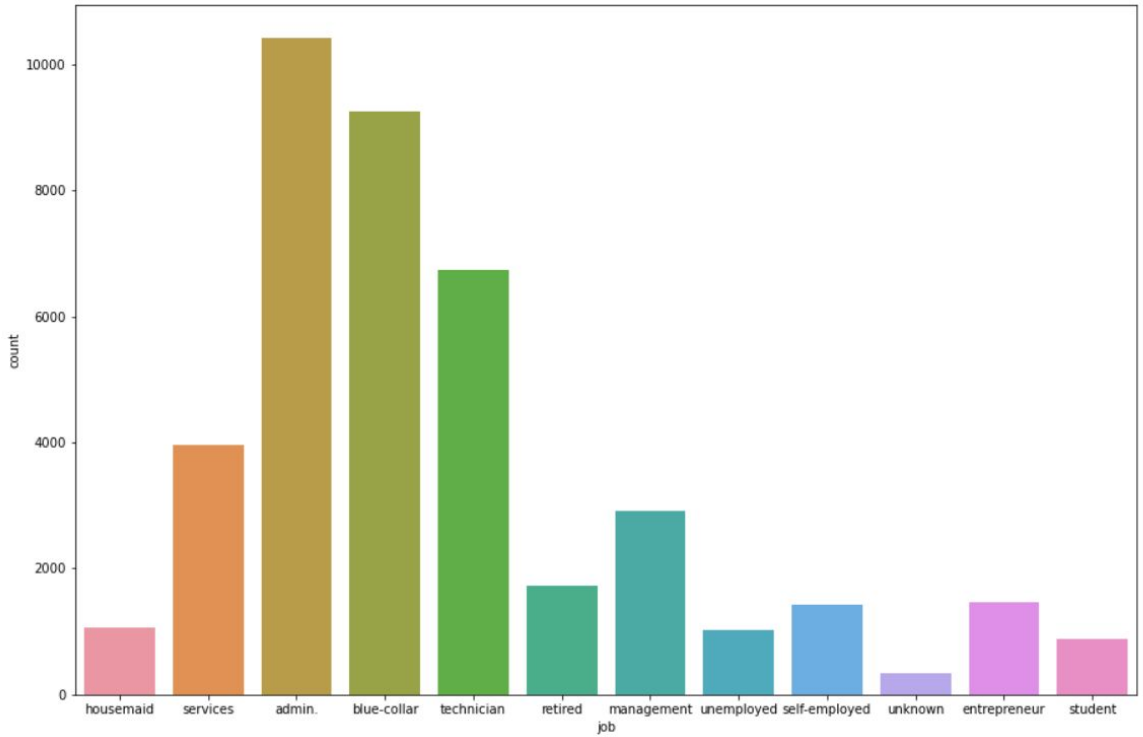
This would save their time, efforts and resources as they do not need to focus on clients that are unlikely to subscribe.

Exploratory Data Analysis (EDA)

Categorical Data

1. Jobs

```
countplot_features('job')
countplot_targetvsfeature('job', target)
```

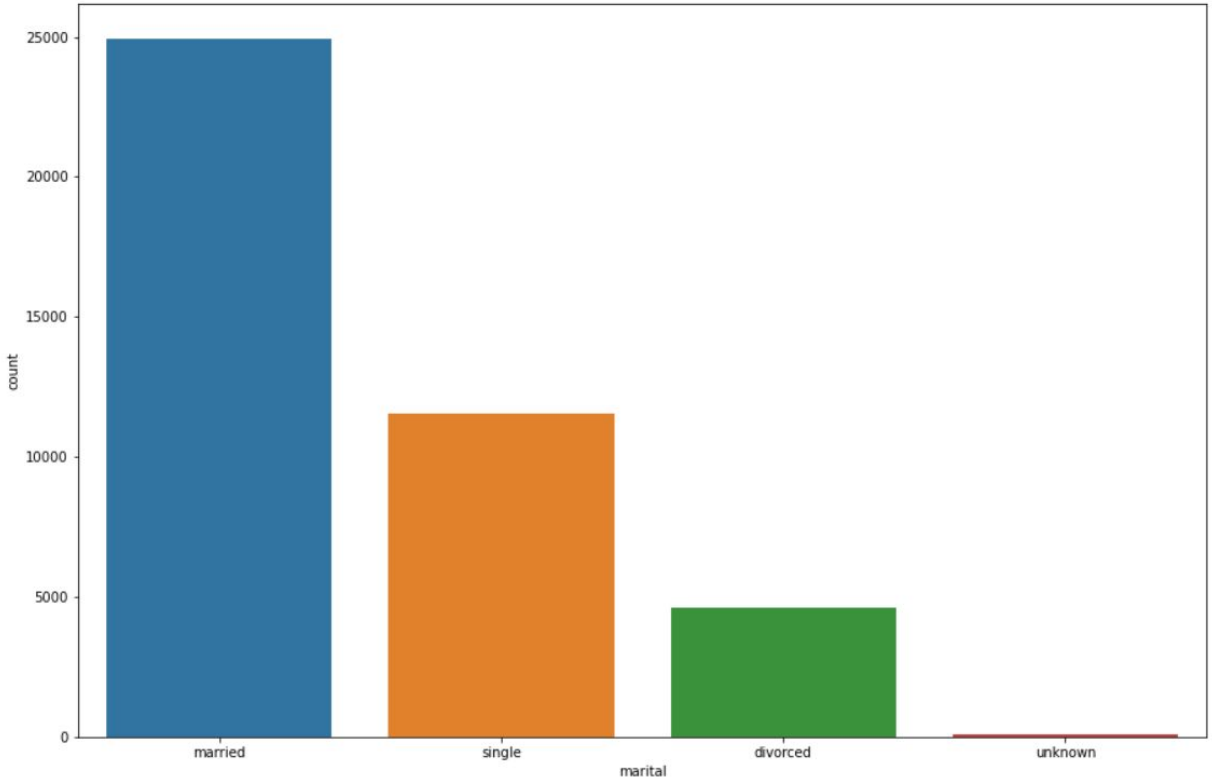


The most common jobs are administrative, blue collar and technical jobs,
whereas the least common ones are students, housemaids and unemployed individuals.

Categorical Data

2. Marital Information

```
countplot_features('marital')
countplot_targetvsfeature('marital', target)
```

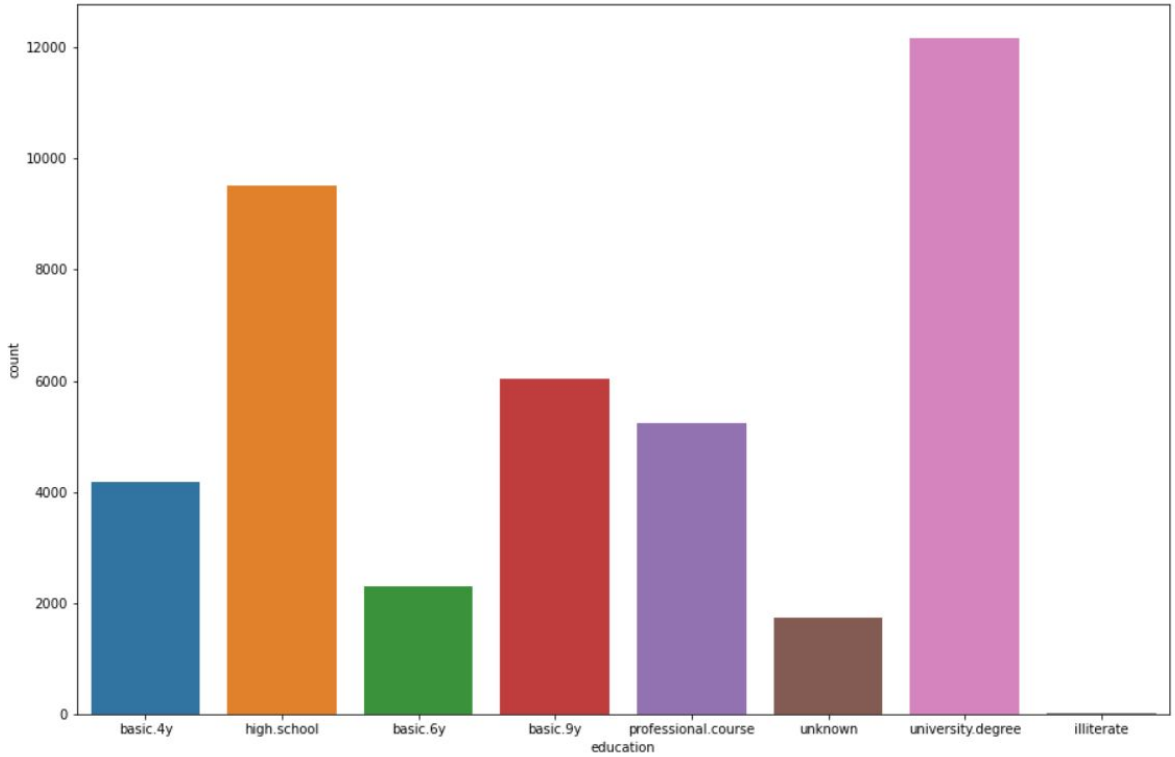


As we can see, most individuals are married.

Categorical Data

3. Education

```
countplot_features('education')
countplot_targetvsfeature('education', target)
```

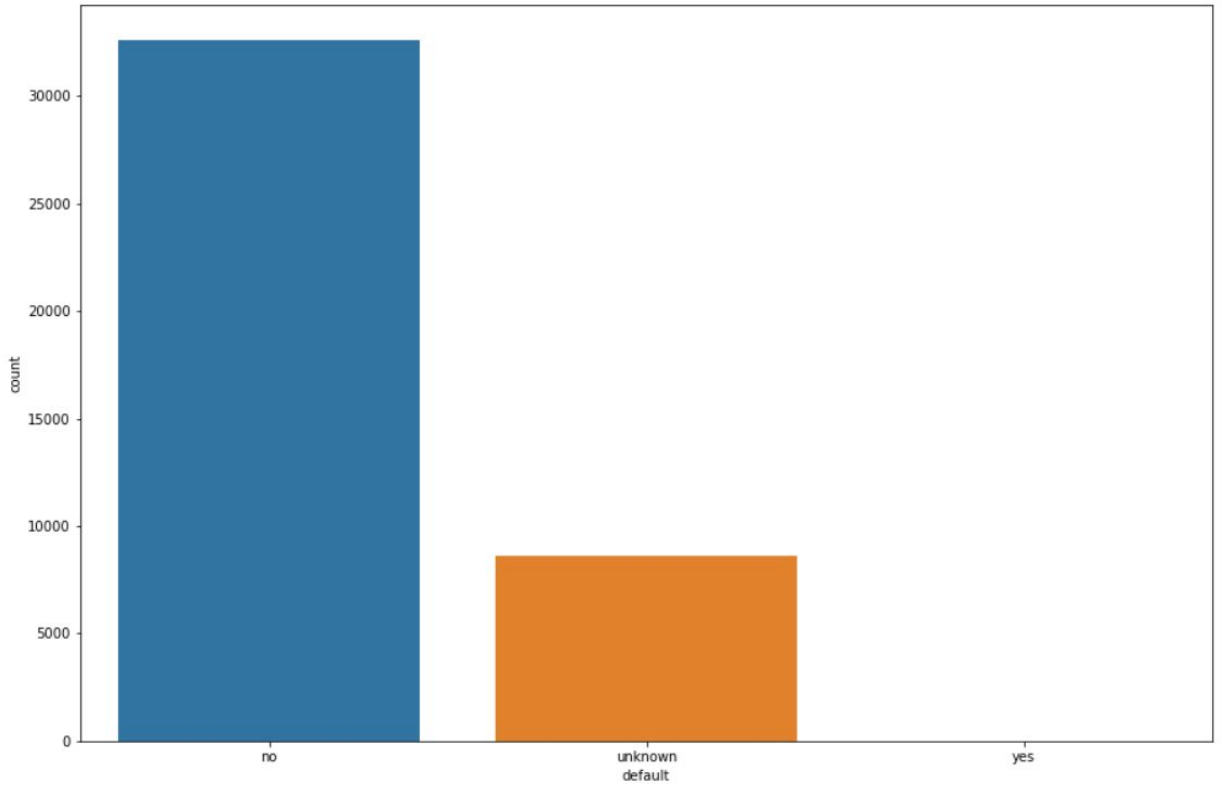


Most of the potential customers have a college degree, or a high
school degree. Very few are illiterate.

Categorical Data

4. Defaults

```
countplot_features('default')
countplot_targetvsfeature('default', target)
```

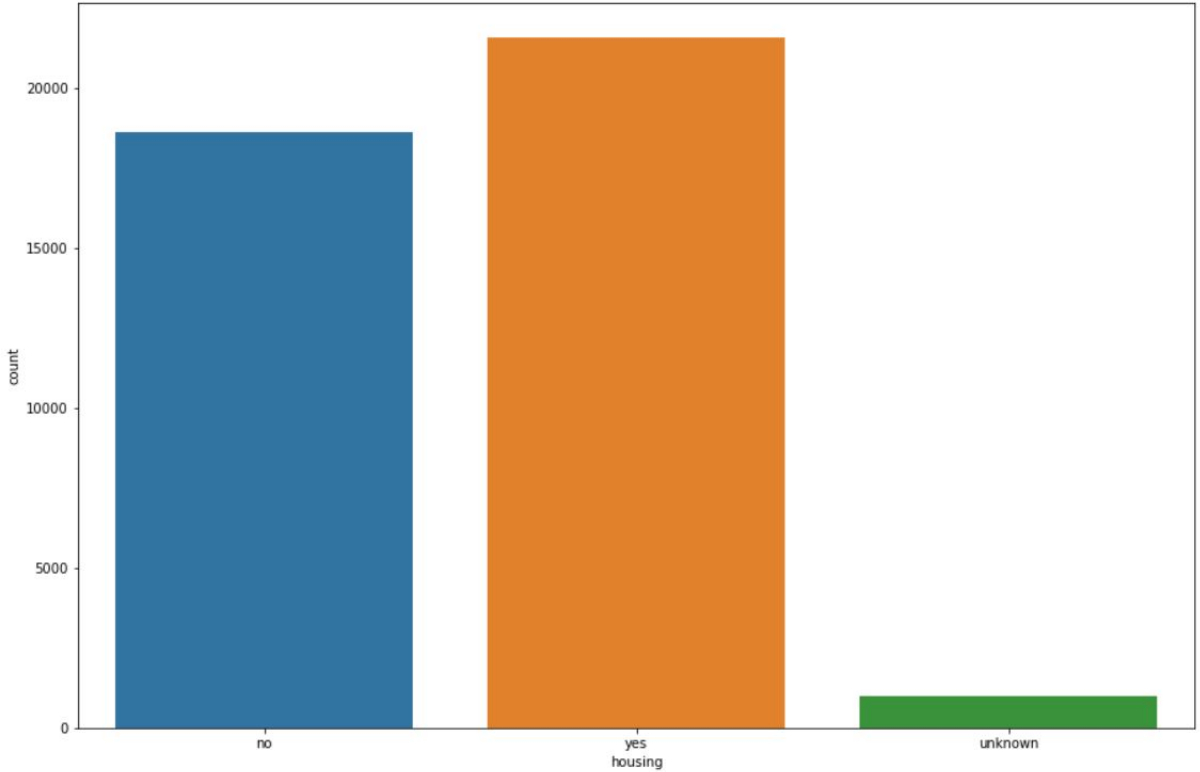


Most of the target clientele have no defaults.

Categorical Data

5. Housing Loans

```
countplot_features('housing')
countplot_targetvsfeature('housing', target)
```

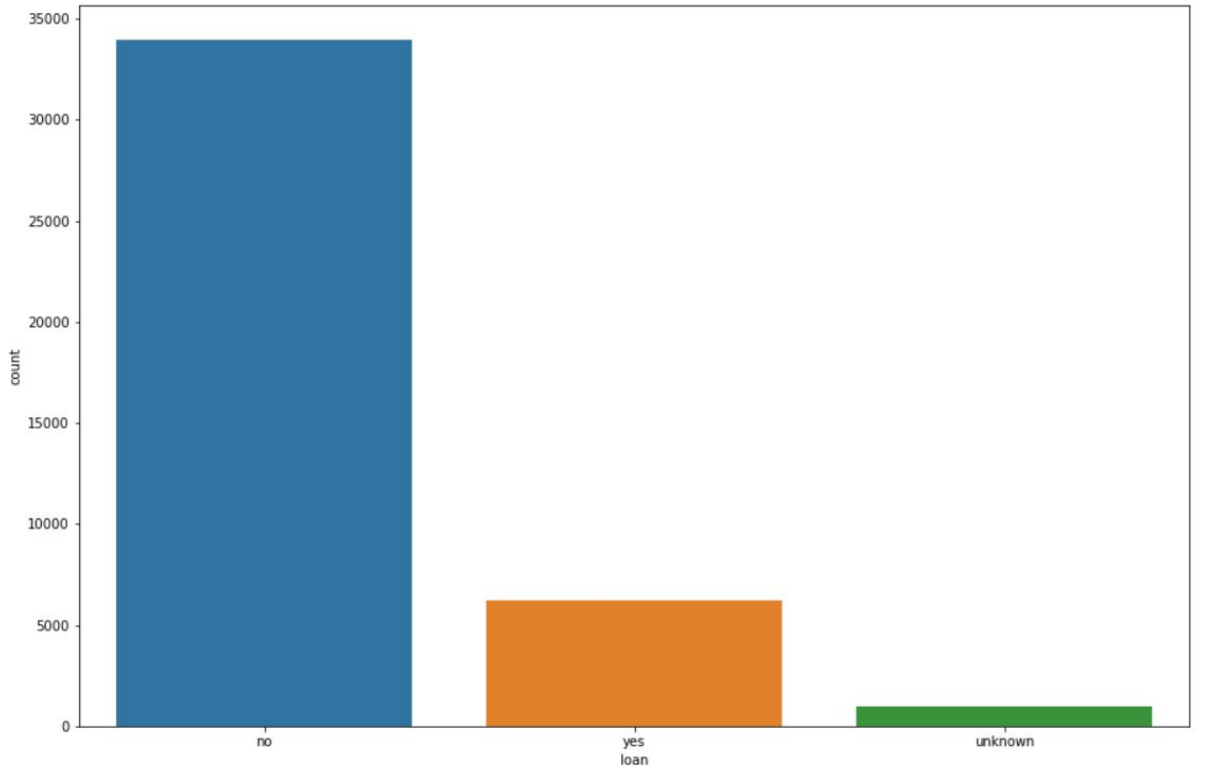


Most of the target clientele have a housing loan.

Categorical Data

6. Personal Loans

```
countplot_features('loan')
countplot_targetvsfeature('loan', target)
```

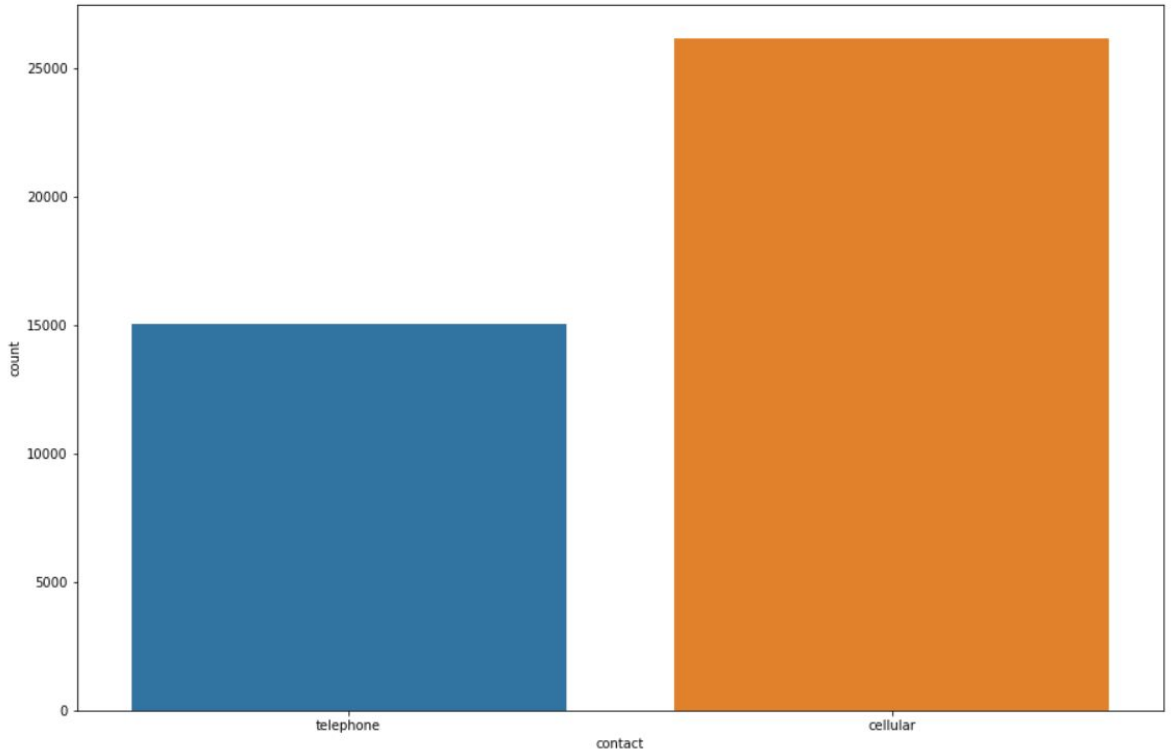


Most of the target clientele do not have any personal loans.

Categorical Data

7. Contact

```
countplot_features('contact')
countplot_targetvsfeature('contact', target)
```

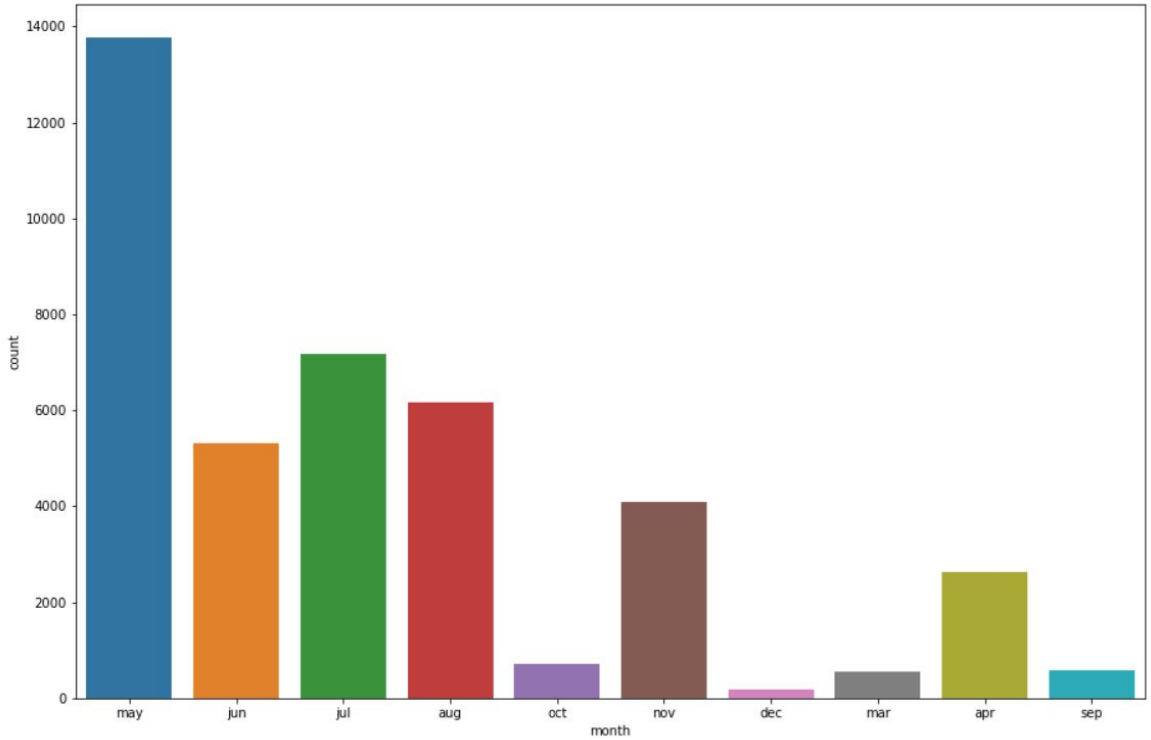


Most individuals have listed their preferred contact method as cellular phone over telephone.

Categorical Data

8. Month of contact

```
countplot_features('month')
countplot_targetvsfeature('month', target)
```

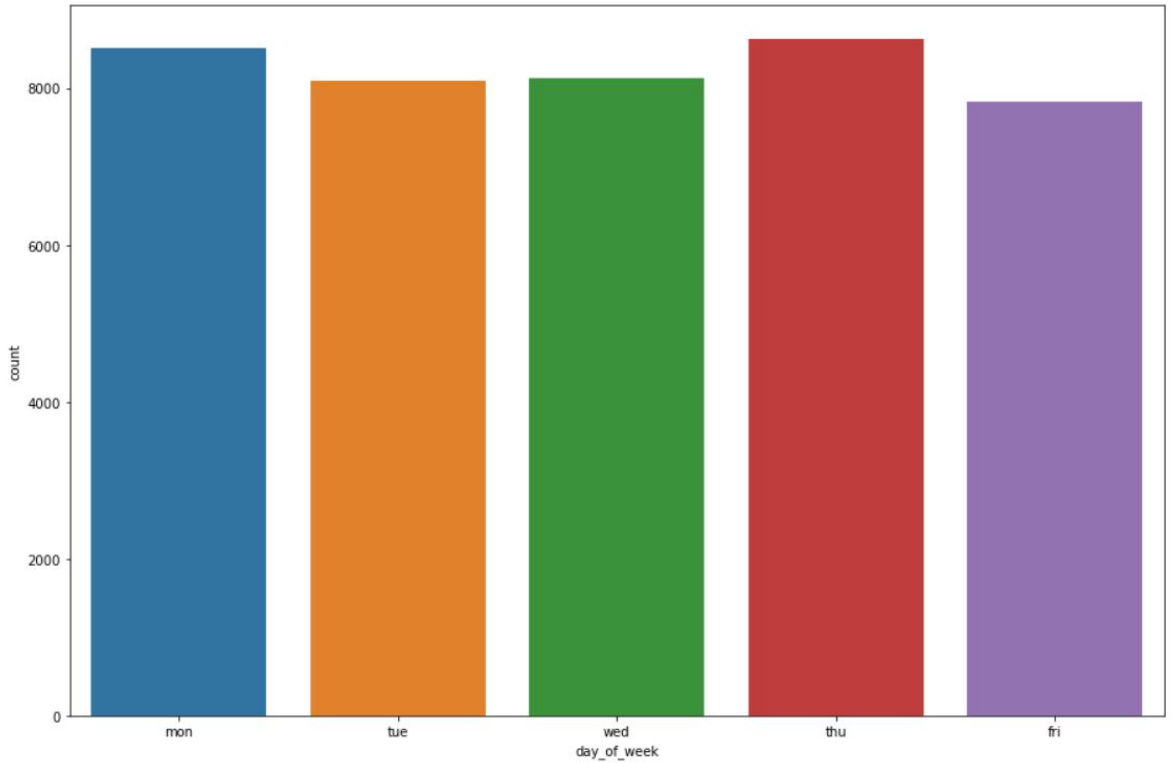


The last month of contact for most of them by far is May, followed by July, August and June.

Categorical Data

9. Day of week

```
countplot_features('day_of_week')
countplot_targetvsfeature('day_of_week', target)
```

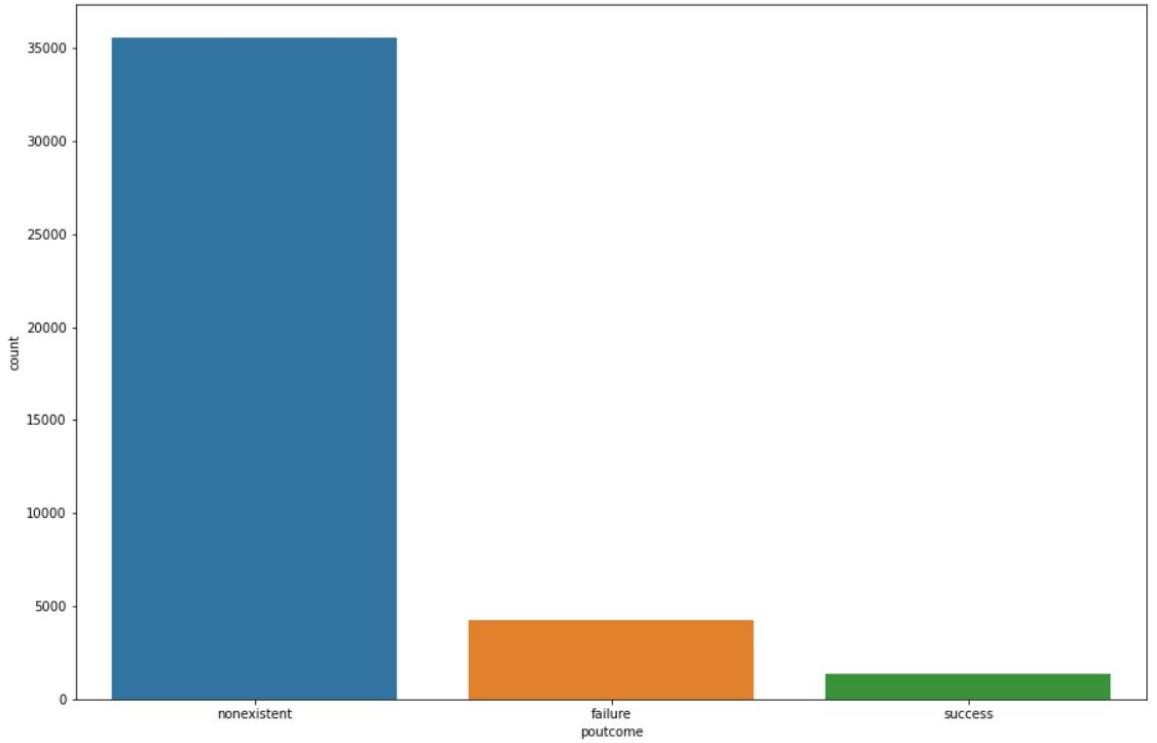


There is an even distribution in the last day of contact among the targets.

Categorical Data

10. Previous Outcome

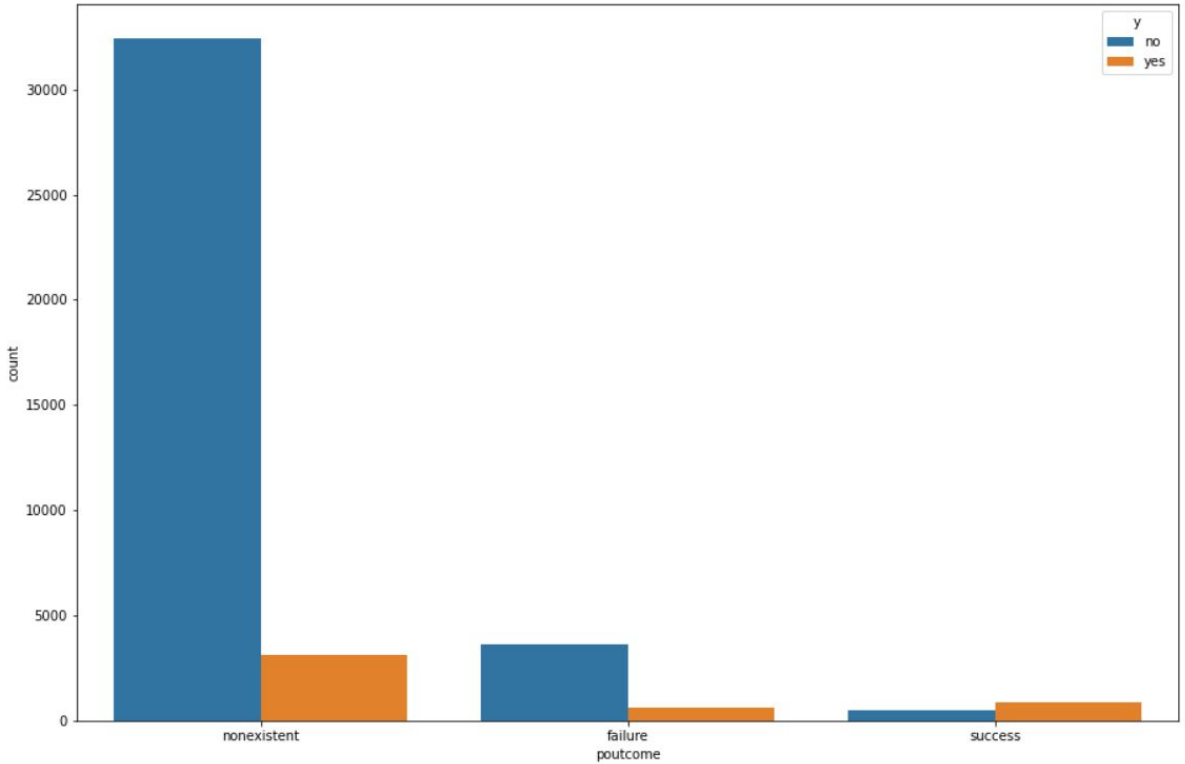
```
countplot_features('poutcome')
countplot_targetvsfeature('poutcome', target)
```



Categorical Data

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.



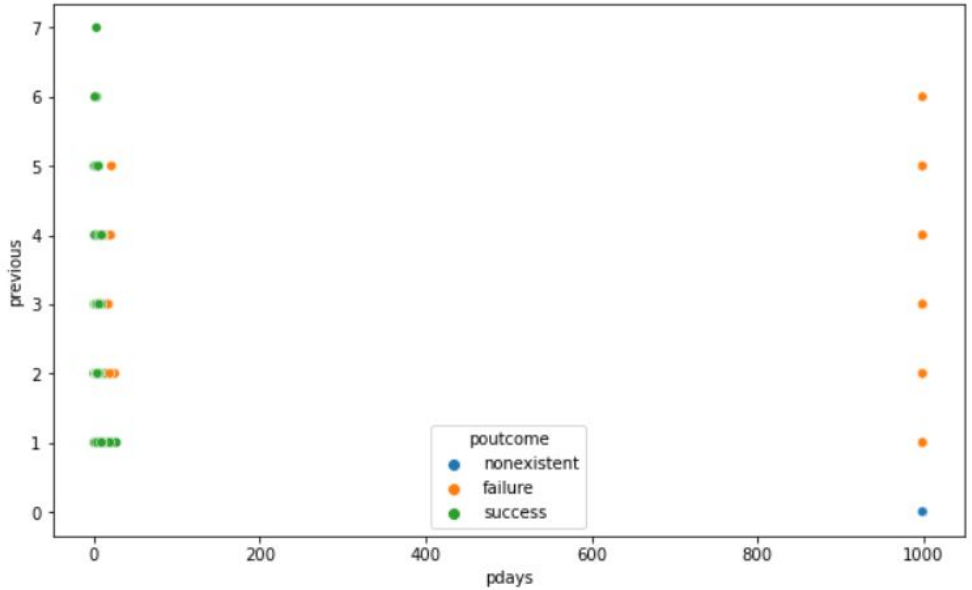
Numerical Data

1. Relation between ‘pdays’ and ‘previous’

```
#Let's verify that there is coherence between the pdays variable
#(#of days since last contacted----> if 999 then client was never contacted before)
#and previous variable (# of times contacted in last campaign).

plt.figure(figsize=(10,6))
sns.scatterplot(x=features['pdays'], y=features['previous'], hue = features['poutcome'])
```

```
<AxesSubplot:xlabel='pdays', ylabel='previous'>
```



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.

Numerical Data

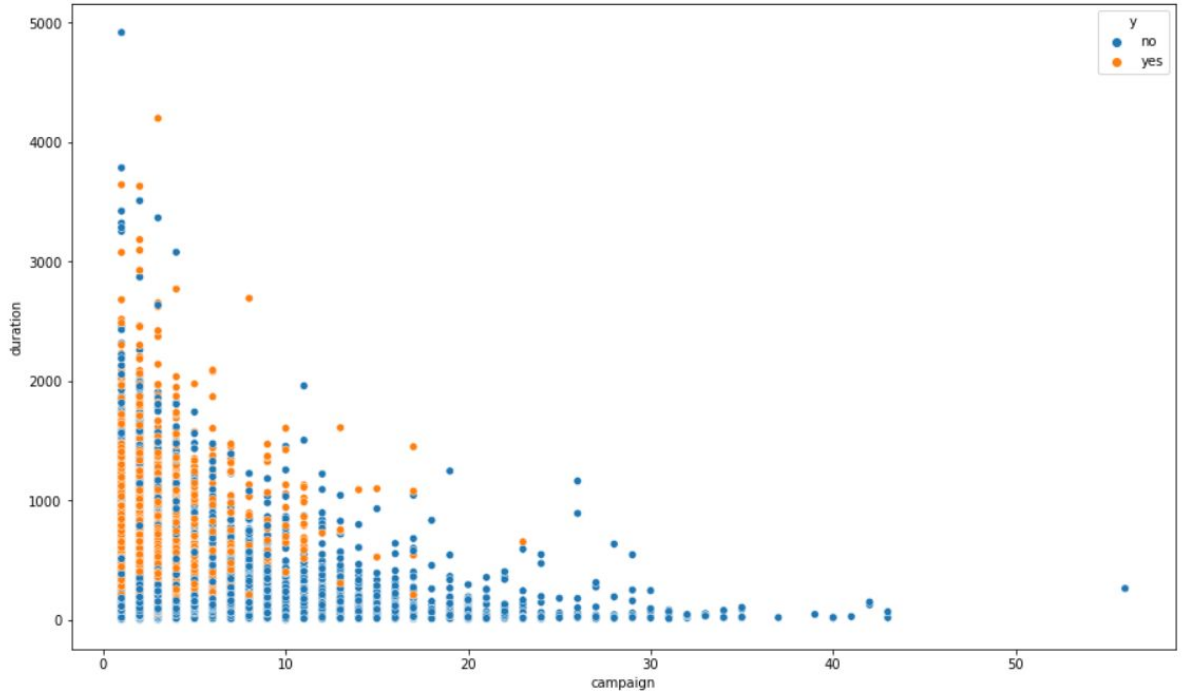
2. Relation between call duration and frequency

There is a negative relation between the number of calls made to a prospective client and the duration of these calls.

Clients that have been called over 12 times do not respond, or give a very brief response.

```
plt.figure(figsize=(15,9))
sns.scatterplot(x= features['campaign'], y= features['duration'], hue = target)
```

```
<AxesSubplot:xlabel='campaign', ylabel='duration'>
```



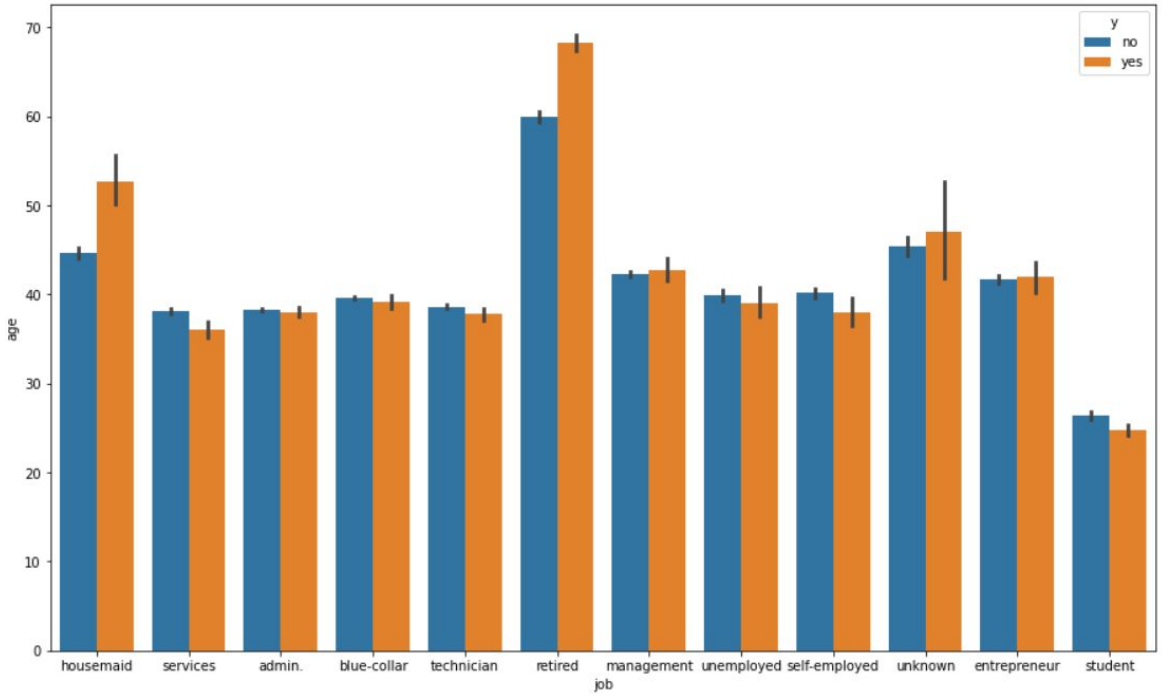
Numerical Data

3. Relation between job and age

```
#Can we visualize what job and age is a more common client for a term deposit?

plt.figure(figsize=(15,9))
sns.barplot(x= features['job'], y= features['age'], hue= target)
```

```
<AxesSubplot:xlabel='job', ylabel='age'>
```



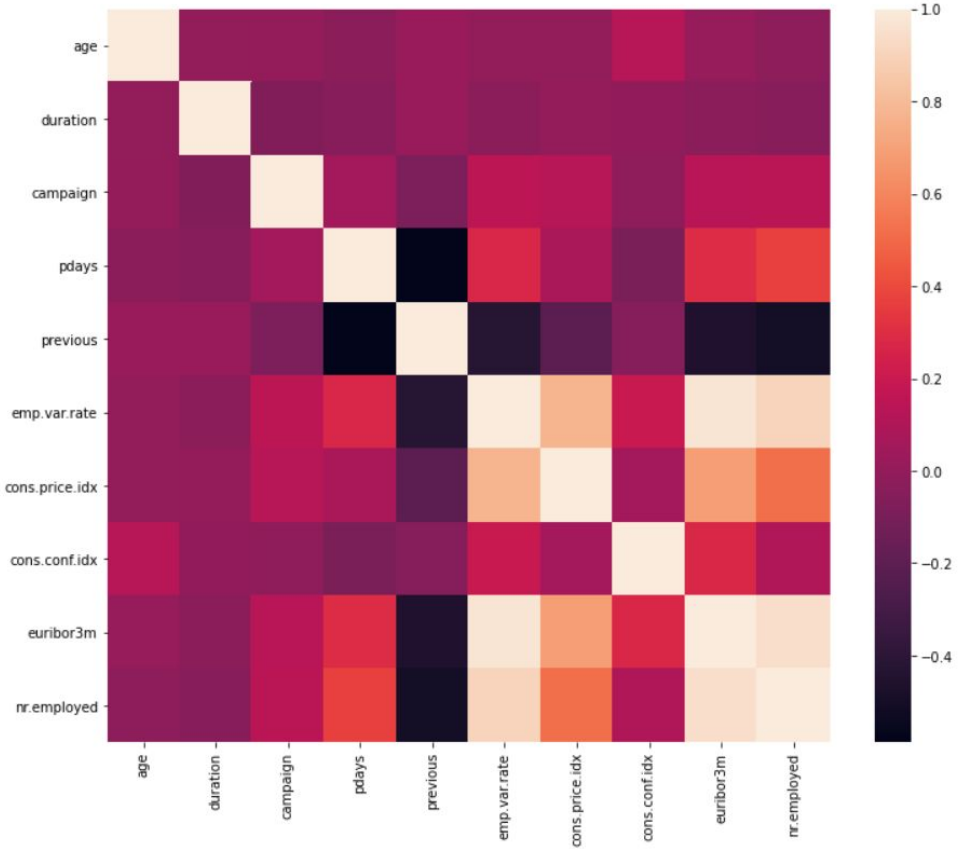
There seems to be a larger difference between the ‘yes’ and ‘no’ subscribers among the retired people aged 65-70 years old and the housemaids aged 50-55 years old than the other potential clients.

Numerical Data

4. Other insights

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.



EDA recommendation

along with proposed modelling techniques

EDA recommendation

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 for this dataset

- 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.

- Performing cross validation among classification models, we found these to be the best models for this case to be:

    - ➢ Logistic Regression
    - ➢ Support Vector Classifier (SVC)
    - ➢ Extreme Gradient Boosting (XGBoost)

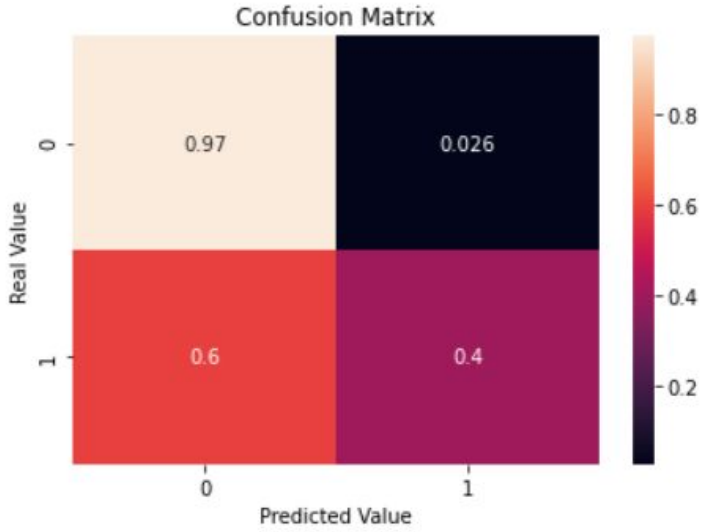
Note: We will be using grid search to optimize our hyperparameters.

Exploring each of these models:

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

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

	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

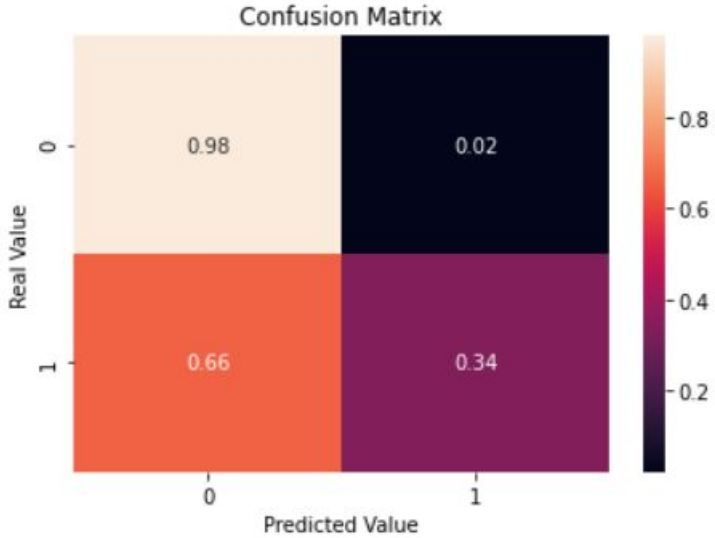


Exploring each of these models:

2. SVC: Is a supervised learning classifier that utilizes Support Vectors (coordinates) to segregate two classes using a hyperplane.

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

	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

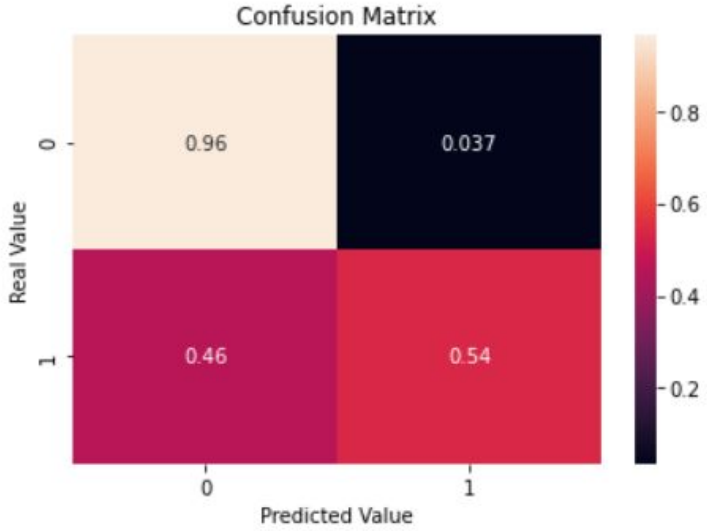


Exploring each of these models:

3. XGBoost: A supervised learning algorithm that uses gradient boosting (gradient descent algorithm) to minimize losses.

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

	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



Model Results

- We have chosen logistic regression as our Baseline model for classification purposes
- Other models we used were Support Vector Classifier and XGBoost Classifier.
- The F-1 score, recall and precision scores of all of them were quite satisfactory, however, the XGBoost model stood out compared to the other two (0.59 F-1 Score, 91% precision and 91% recall).

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 - Logistic Regression, SVC and XGBoost.
- Upon comparing these models using F-1 score, recall and precision, we found all three of them to perform really 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.

GitHub Links:

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

AND

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

Thank you.