Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

Bank Marketing (Campaign)

January 30, 2023

TEAM MEMBER

Name: Giovanna Vieira
E-mail: giihvieira1703@gmail.com
Country: Brazil
College/Company: Hashtag Treinamentos
Specialization: Data Science

Data Glacier
Your Deep Learning Partner

Agenda

Problem Statement
Data Analysis
Data Cleaning
EDA
Encoding
Models
Final Recommendations
GitHub Repo Link

Data Glacier
Your Deep Learning Partner

Problem Statement

Overview

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Solution

By analysing the dataset we will be able to convert this problem into a machine learning classification and build a model to predict whether a client will subscribe a term deposit or not.

Data Analysis

- 21 Features
- 41188 rows

Assumptions:

• The data seems to be cleaned and a little bit skewed, however you can see that the variables have outliers that need to be cleaned with data cleaning process.

• There are no null values, but there are some "unknown" values.

• We've got 41188 rows, some of the columns has 85% of the values repeated.

• The value types seem to be correct but as said there are some "unknown" values.

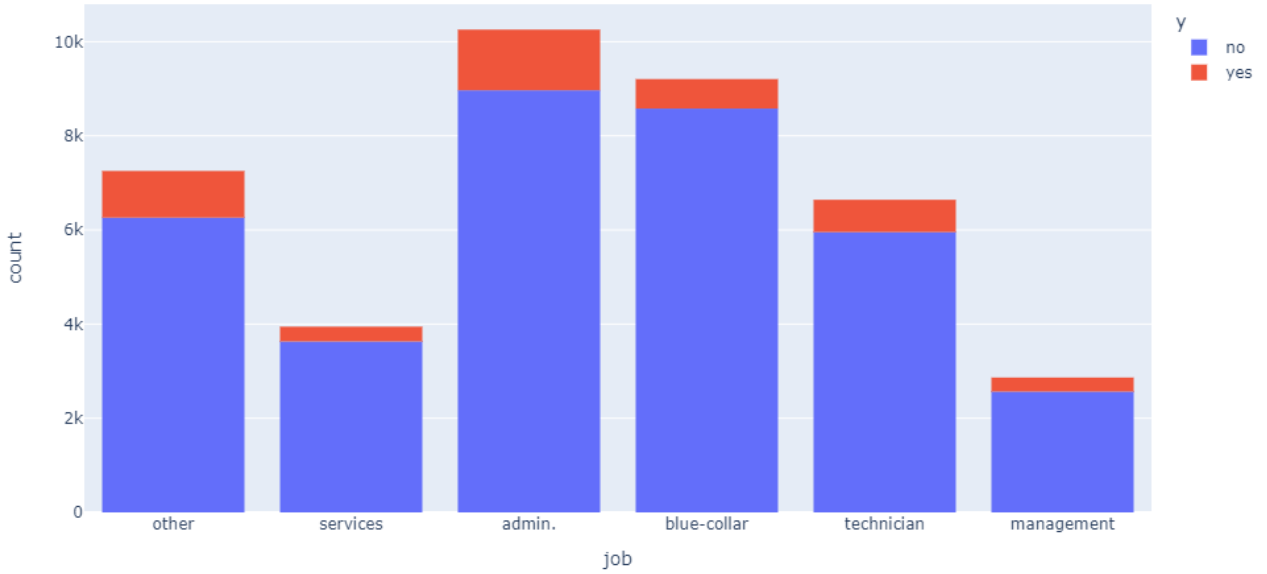
Data Cleaning

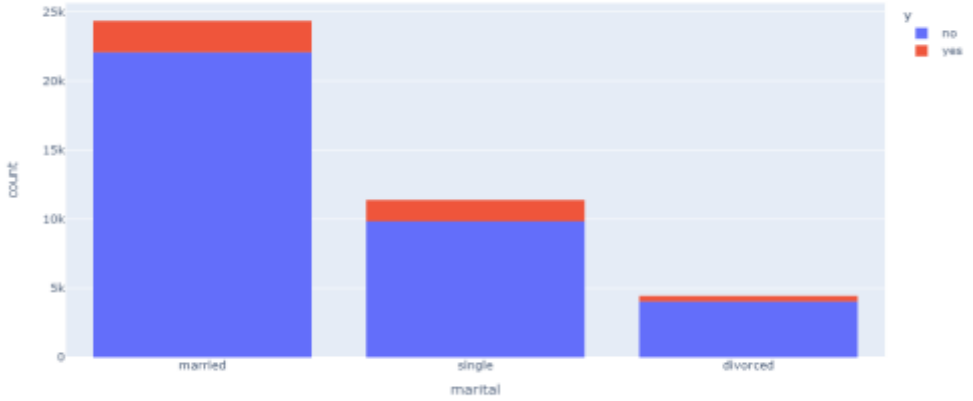
The most used techniques to treat the data frame were:

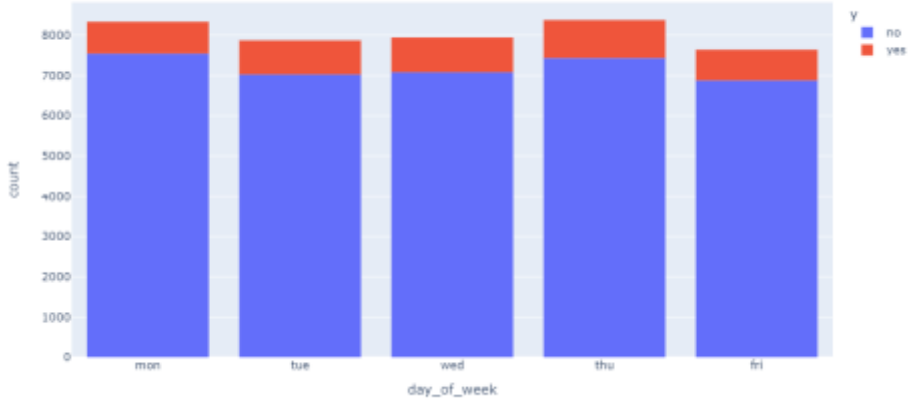
- Dropping columns and rows
- Replacing missing values with the mode
- Filling outliers with ffill/bfill based on interquantile range
- Grouping text features that doesn’t appear many time

Cleaned dataset reduced from (41188, 21) to (40165, 18)

EDA – Categorical Attributes

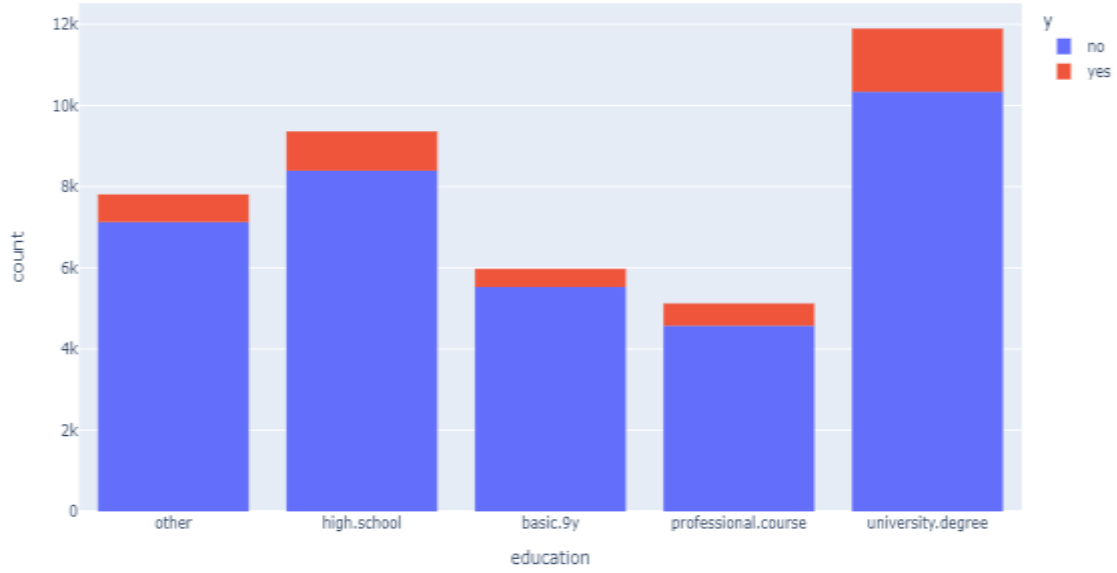


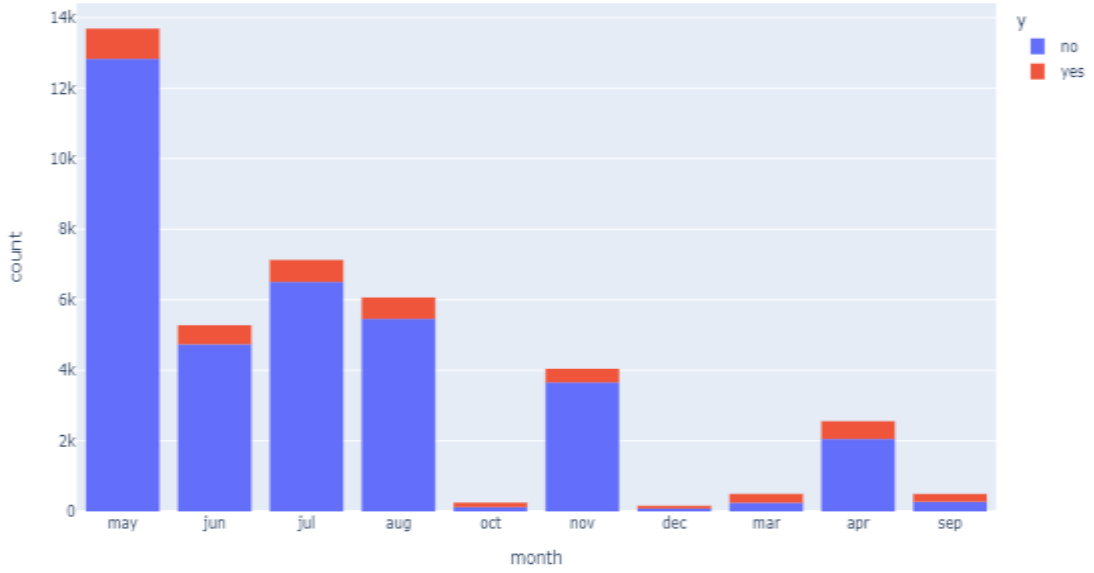




Job: Highest number of subscriptions to a term deposit work on admin.
Marital: Most of the clients approached were married.
Day of Week: There's no significant difference in the numbers of clients approached and people subscribed.

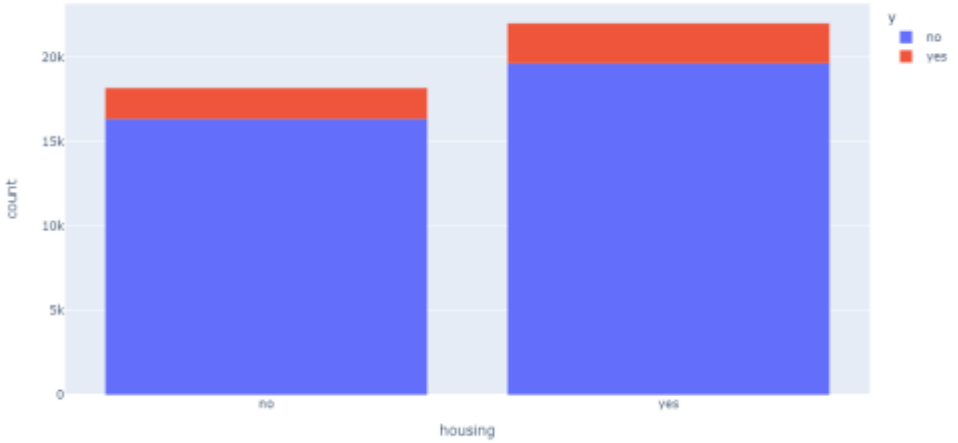
EDA – Categorical Attributes

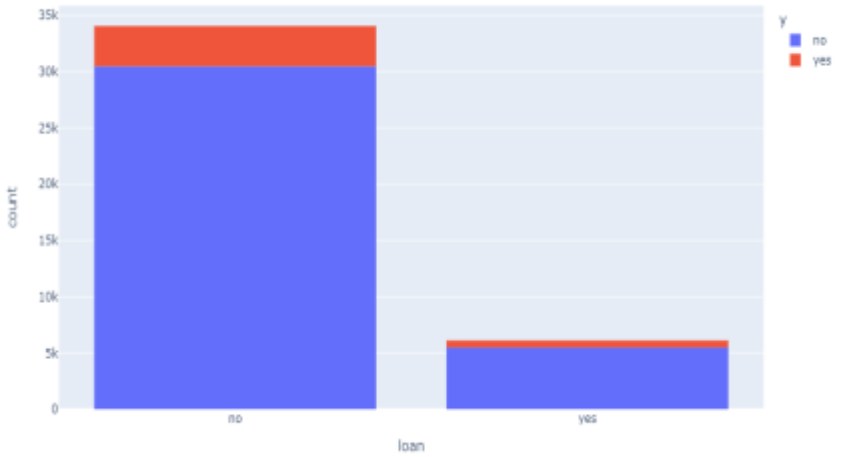


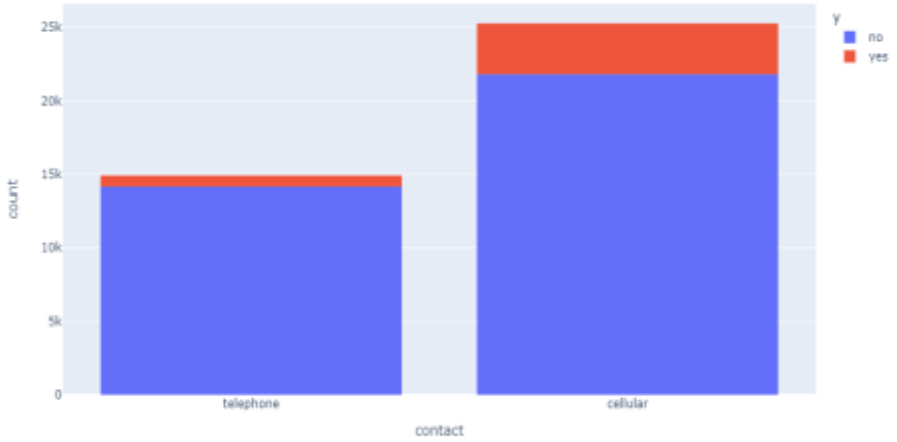


Education: In the education column we can see that most people who has subscribed have a university degree.
Month: The last contact month of year was way bigger in may.

EDA – Numerical and Other Attributes

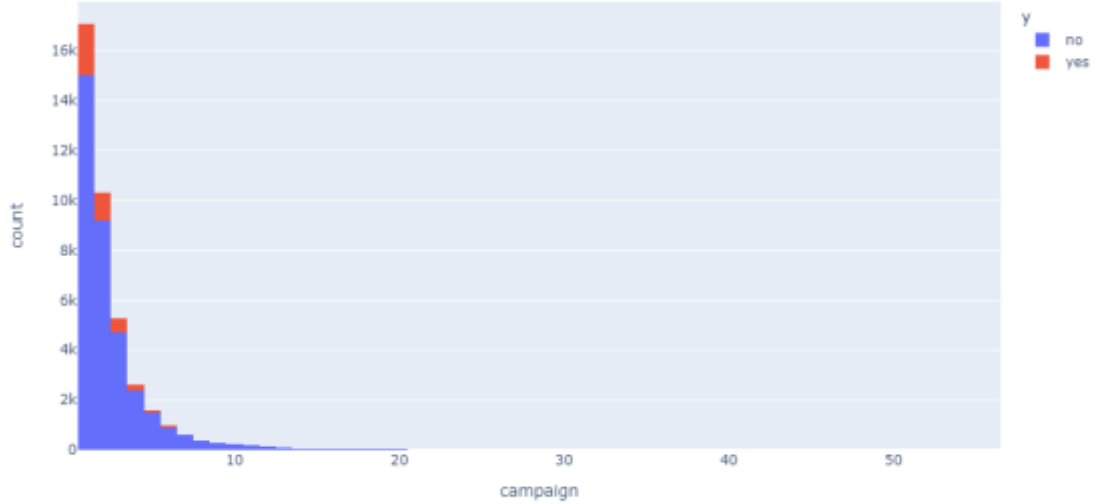


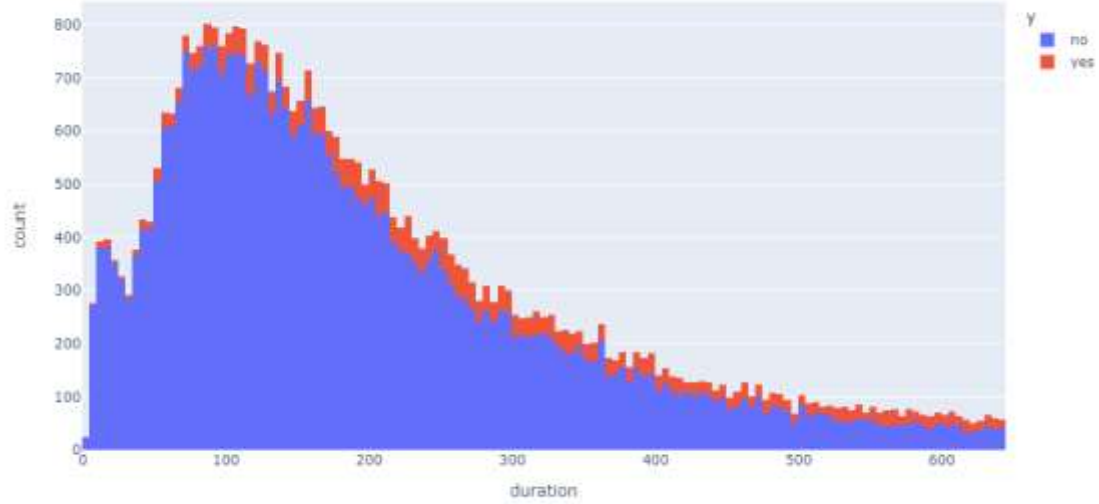




Housing: A housing loan does not have much of effect on the number of term deposits purchased.
Loan: Most of the clients approached did not have a personal loan.
Contact: The contact by phone seems to be more effective.

EDA – Numerical and Other Attributes

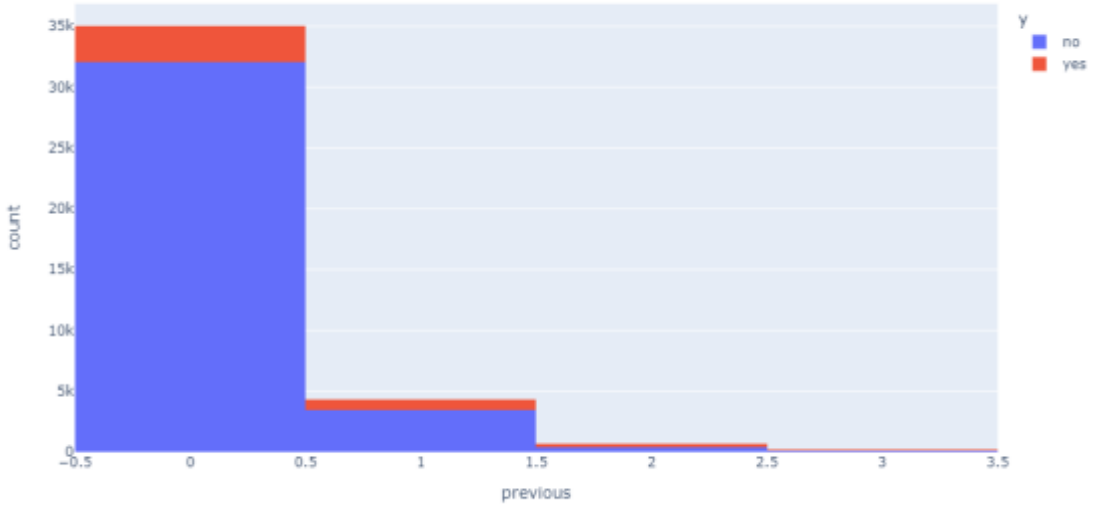


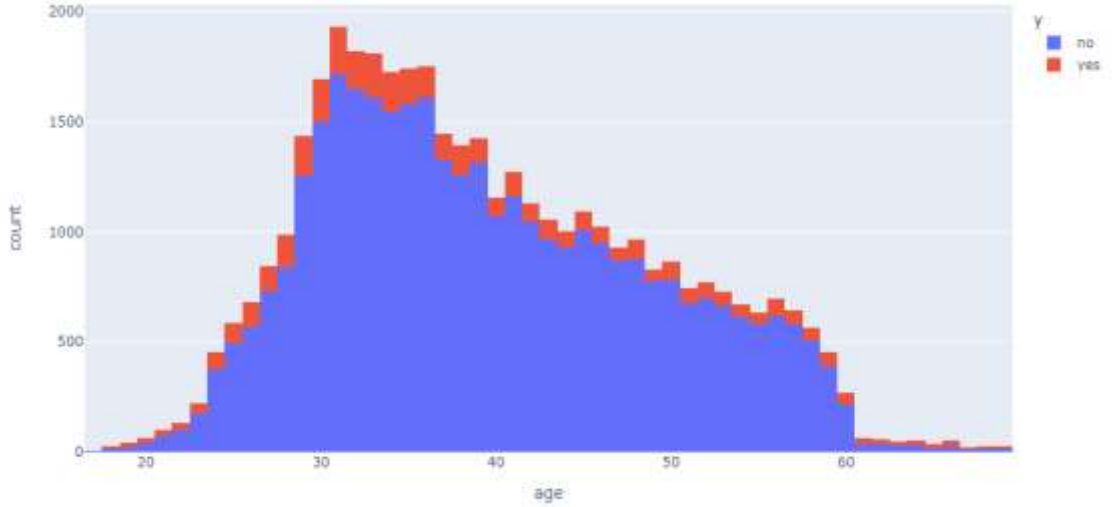


Campaign: Most of the numbers of contacts performed during the campaign and for the client were only once or twice.

Duration: The duration of the last contact is highly variated.

EDA – Numerical and Other Attributes

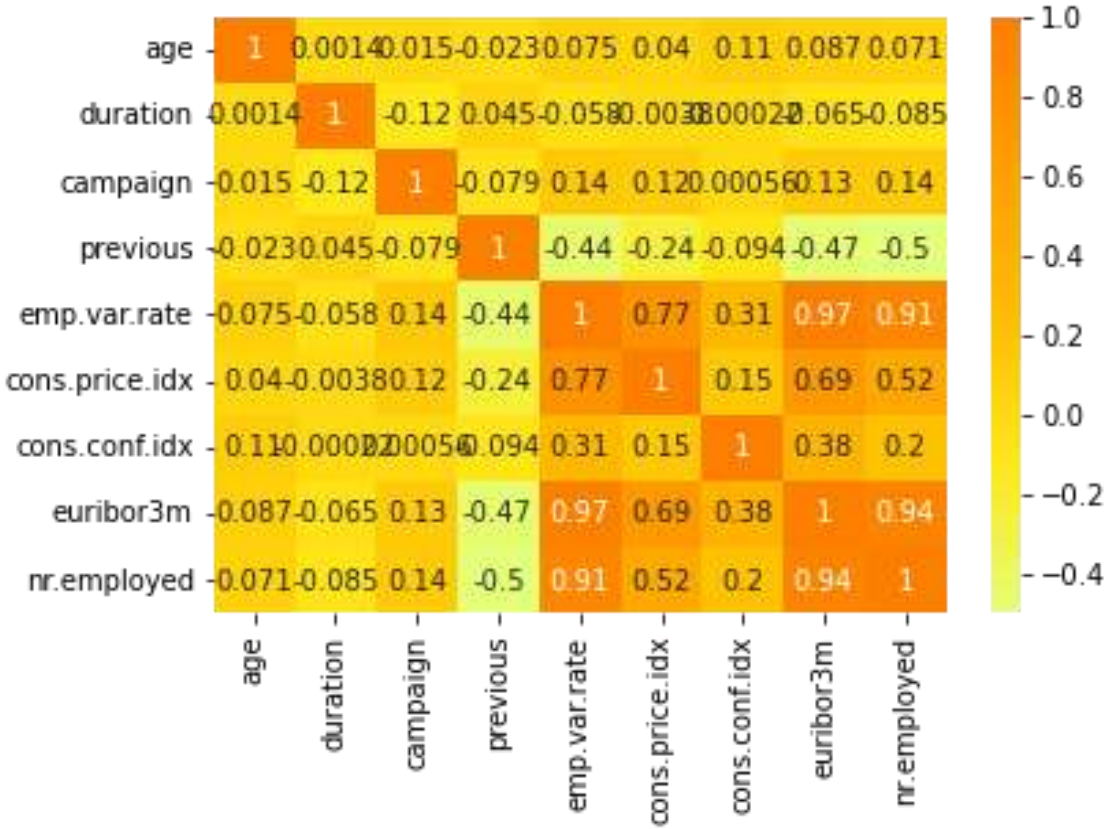




Previous: The number of contacts performed before this campaign and for a specific client was 0 in more than 80% of the cases

Age: Most of the people contacted was between 30 and 40 years old.

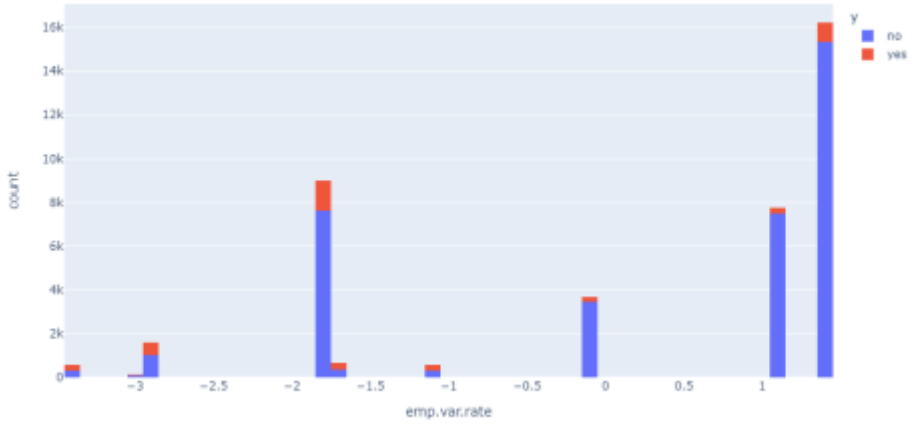
EDA – Numerical and Other Attributes

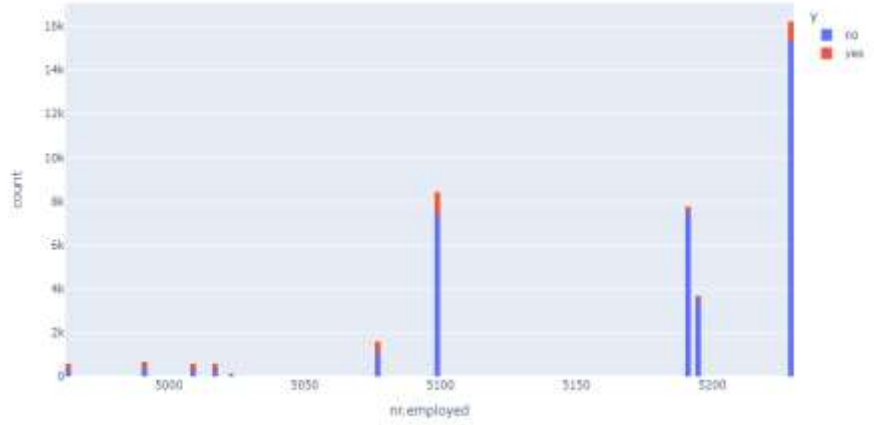


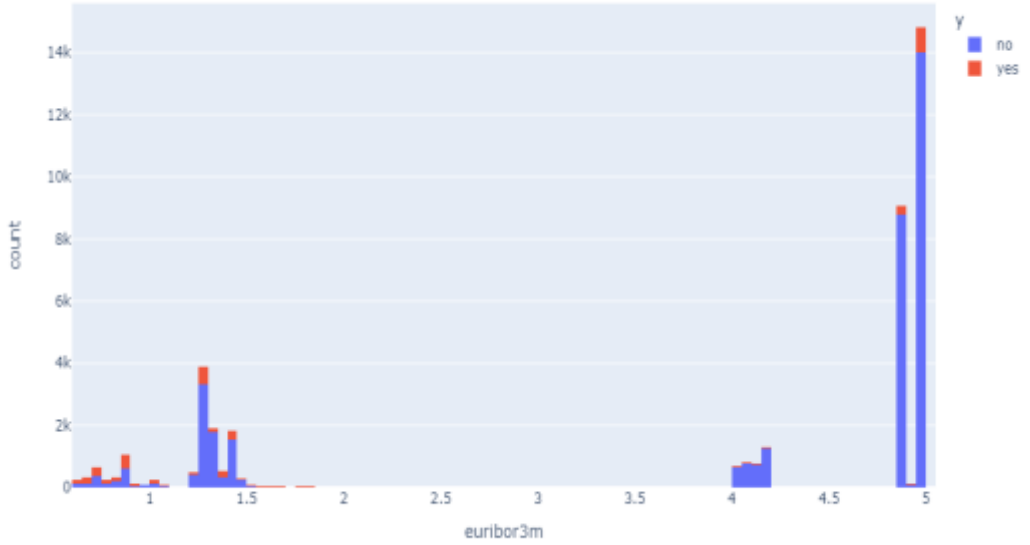
As we can see in this heatmap, the columns 'euribor3m', 'nr.employed' and 'emp.var.rate' are highly correlated.

You can see in the next page their graphics when comparing who did and who did not subscribe to a term deposit.

EDA – Numerical and Other Attributes







Encoding

How was the encoding done:

- Columns “month” and “day_of_week” were turned into numbers (jan = 1, feb = 2; mon = 1, tue = 2)
- Columns with values as “yes” and “no” and columns with 2 options (contact => telephone, cellular) were turned into 0s and 1s
- Column “marital” had its values turned into 0s (married), 1s (single) and 2s (divorced)
- Categorical features as “job” and “education” were turned into dummy variables
- Numeric features were left as they are

Models

Choosing Metric

7328	**1674**
385	655

False positive is the situation when the model marked a customer as potentially positive and a call has been made, but the customer refused to open deposit. This leads to increase campaign direct costs

False negative is the situation when the model marked a customer as negative => he would not be called. But this customer would agree to open deposit if he will be called. So here bank is loosing income

1. We are dealing with highly imbalanced dataset thus accuracy is not a good metric.

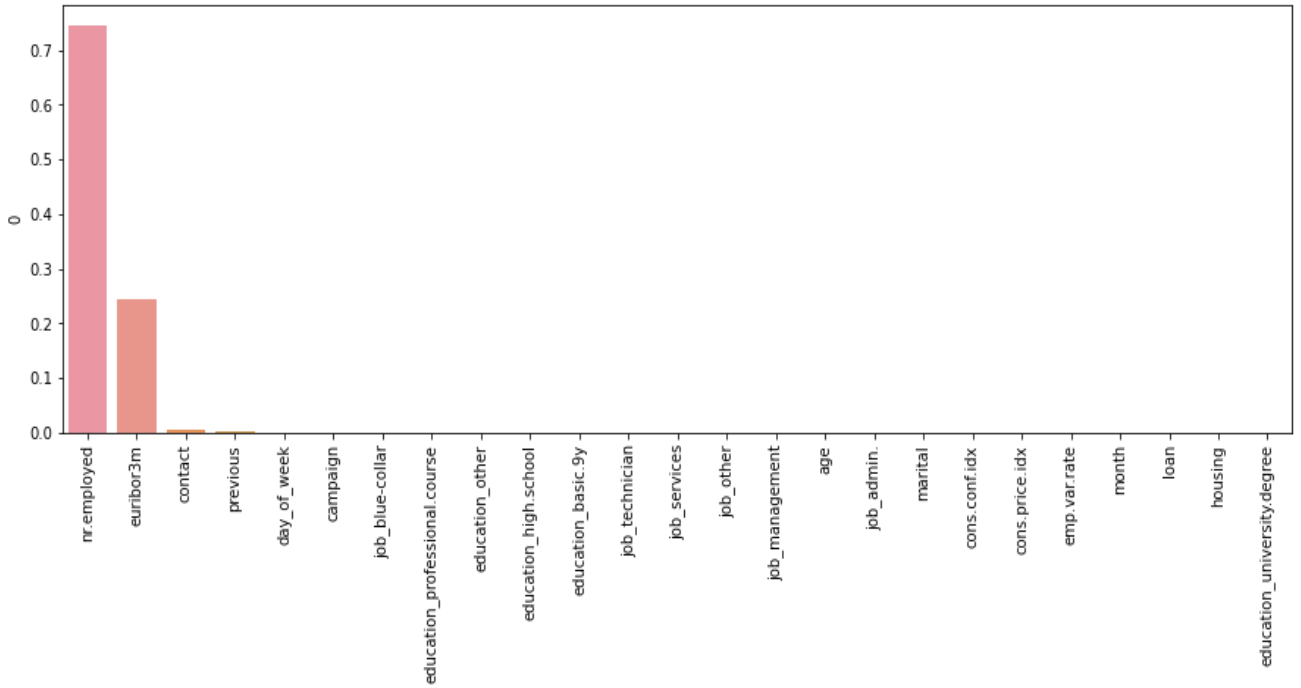
2. When choosing between recall and precision I thought about what is better for the bank: get less false positives or less false negatives

The target is to reduce False Negatives
So the metric chosen is **Recall**

Models

Experimented models:

- Logistic Regression
- Decision Tree
- Stochastic Gradient Descent
- Random Forest
- K Nearest Neighbors



Logistic Regression has the best results on almost all the datasets tested between the members of the group

Final Recommendations

Conclusion:

- Best metric to evaluate the model was recall
- The model that presented the best results is Logistic Regression

Final Recommendations:

- Use Logistic Regression to reduce costs and time and call those who are more likely to open a deposit
- The model can be tuned for better results
- Target relatively old people
- Try to engage customer and have long calls
- Prefer mobile calls over telephone

GitHub Repo Link

https://github.com/sharuhinda/bank_marketing_campaign/tree/review
https://github.com/vgiih/EvolveData

Thank You!