Data Glacier

Your Deep Learning Partner

# Exploratory Data Analysis

## Exploratory Data Analysis

## Team Muadh

Muadh Faizan
muadhfaizan@gmail.com
Data Science

**10 August 2021**

# Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

# Executive Summary

This presentation will illustrate the different methods used to analyse a Portuguese bank dataset, in order to understand the patterns, trends and gather some insights into the dataset. The programming language used to perform this exploratory data analysis (EDA) was Python, using Jupyter notebook.

The business problem will first be described, followed by the methods and tools used in the EDA. Then different graphs and visuals where meaningful information can be inferred will be shown. Following that, the key findings and recommendations will be listed out, which the bank may use. Finally, a few machine learning models will be suggested, which could be used to develop a classification model to solve the business problem.

# Problem Description

The project chosen to be done is the Bank marketing project, which is regarding a term deposit product of a Portuguese bank. The bank wants to sell a term deposit product to its customers but want to know which of their customers will be more likely to buy it, based on the past interactions with the bank. A classification problem is at hand, where the dataset contains information of several customers who were already informed about the term deposit, such as their age, gender, and other information pertaining to their bank accounts and loans. Whether the customer had bought the product or not, is also given for each customer.

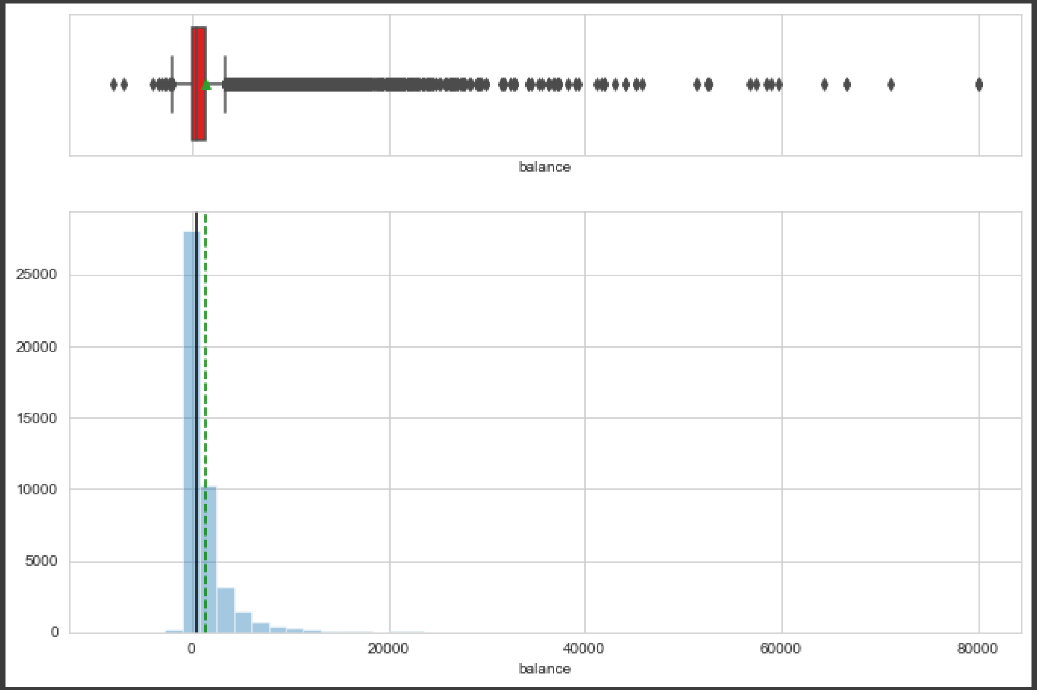
The aim of the project is to analyse the dataset and come up with a classification model which would be able to predict if a customer would buy the product or not.

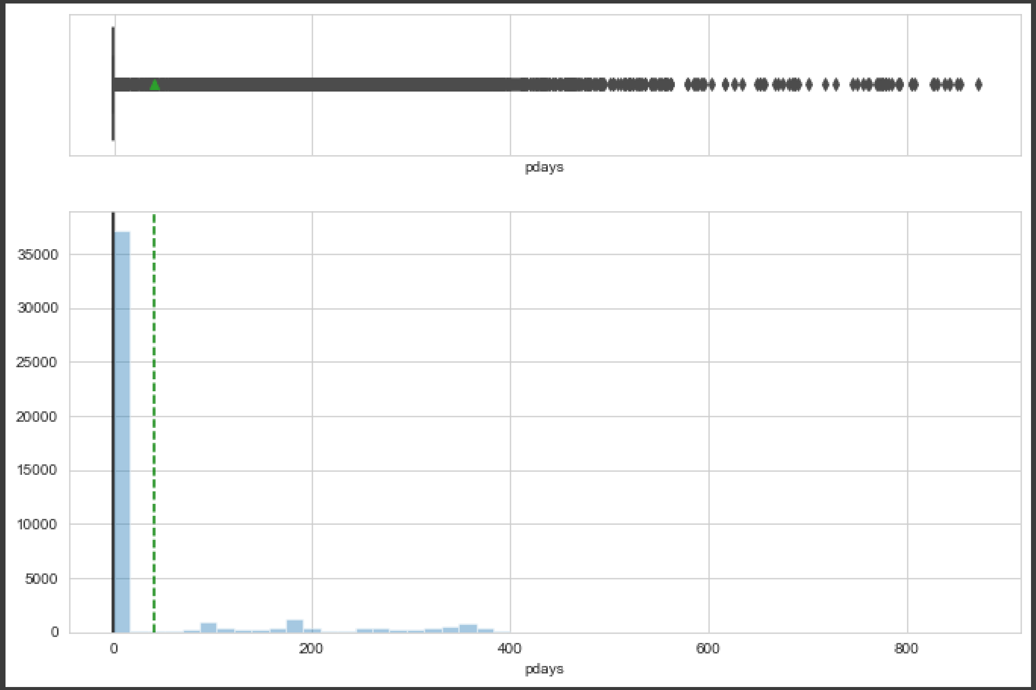
# Approach

The approach followed in the EDA is as follows:

- Histogram with boxplot performed on the numeric variables
- Count plots performed on the categorical variables
- A correlation heatmap was generated to check any correlations
- Stacked plots were generated with the dependent variable as the hue
- Boxplots were done with the dependent variable as the hue
- A few scatterplots were done.

# EDA - Histogram

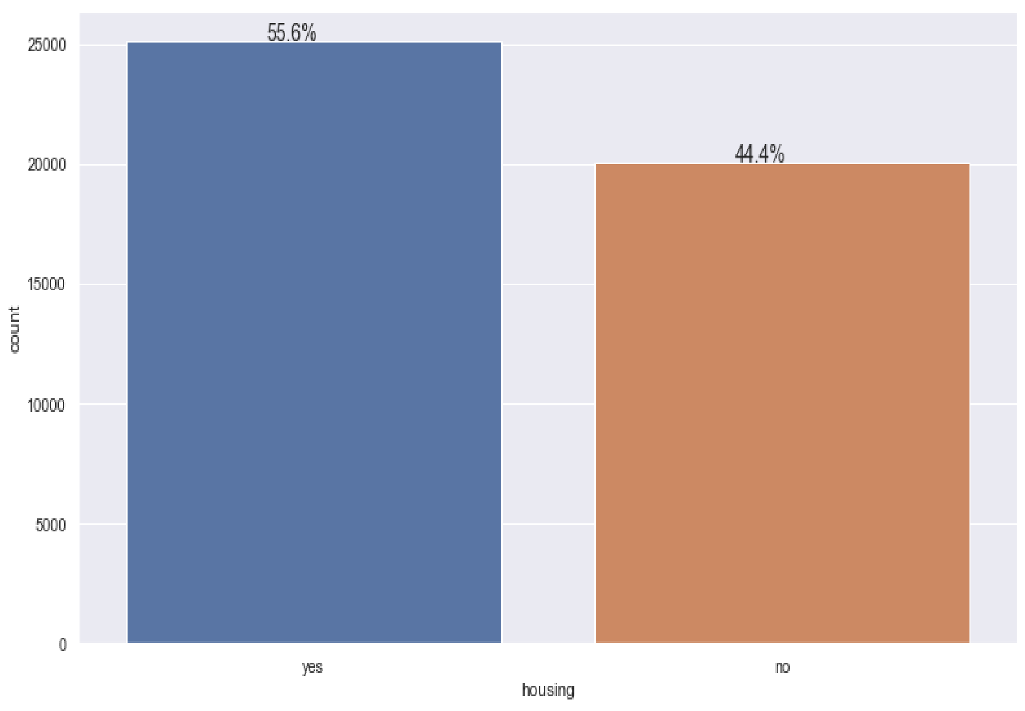


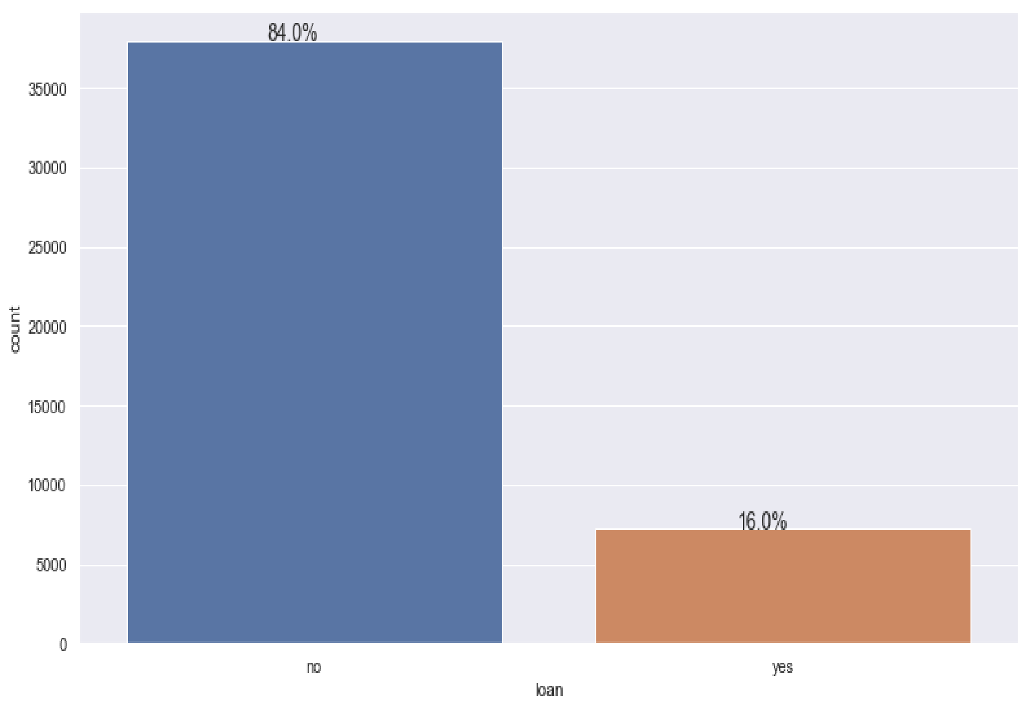


The bank balance of majority of customers is quite low or even negative. There are a lot of customers with more than 5,000 in the bank, but not as much when taken overall.

This shows that more than 35,000 of the 45,000 customers were not contacted for a previous campaign. This fact causes many outliers in the features.

# EDA – Count plots

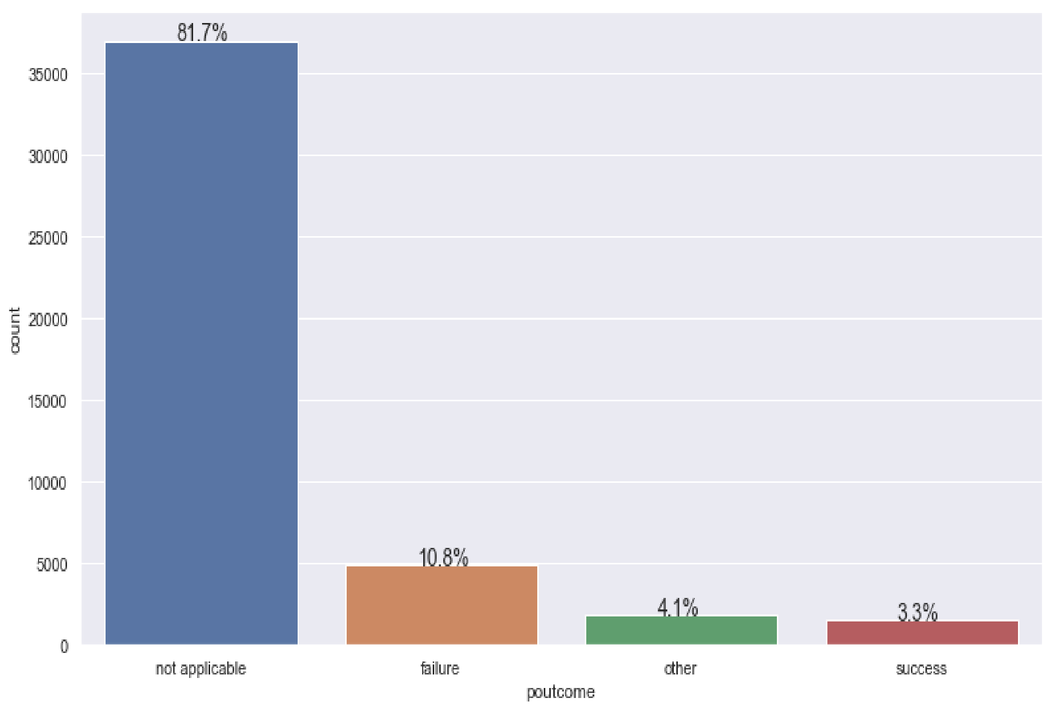


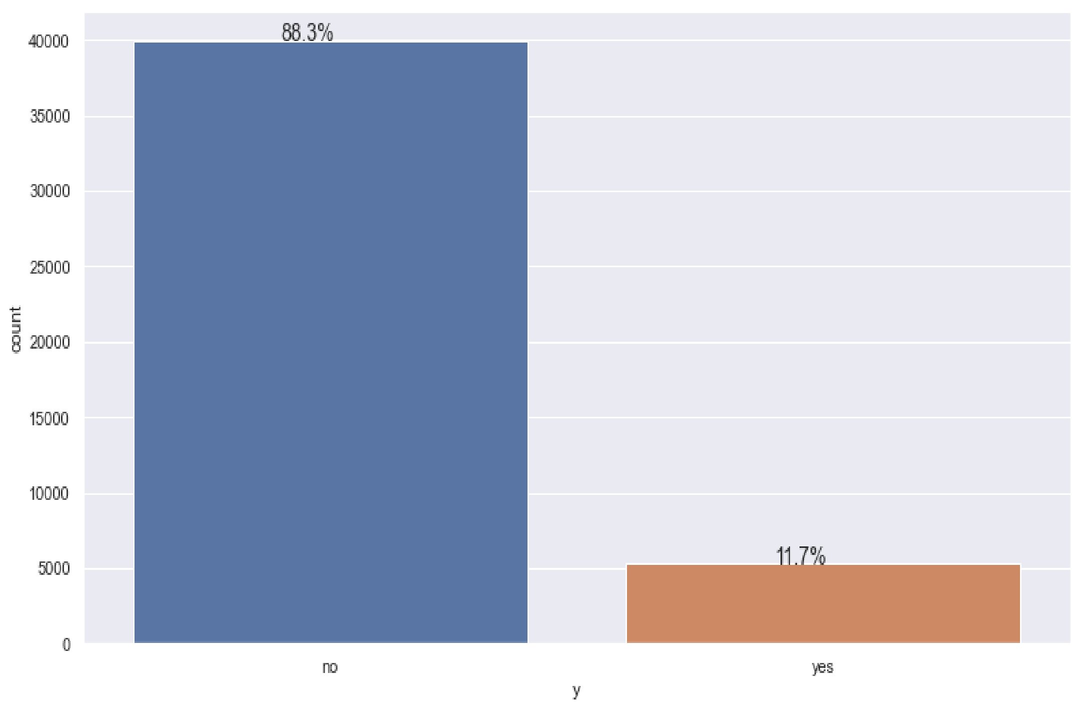


Around 55% of customers have housing loans, but only 16% of them have personal loans.

Data Glacier
Your Deep Learning Partner

# EDA – Count plots

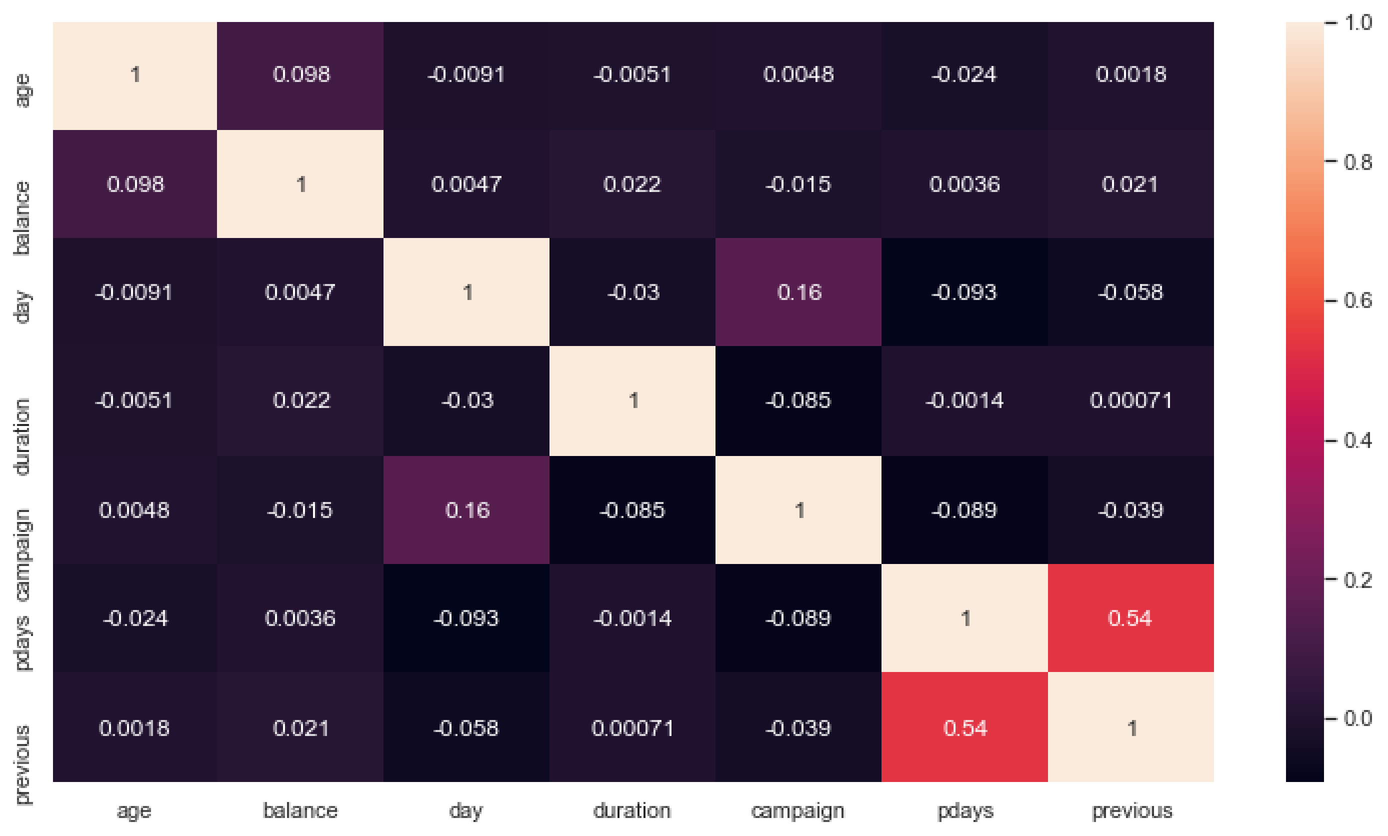




In the previous campaign, 81% were not contacted for it, so no outcome is given. Only 3% of the 18% that were contacted, ended up buying the product.

In the current campaign for the term deposits, only 11.7% of customers have purchased the product. But compared to the previous campaign, this is an improvement.

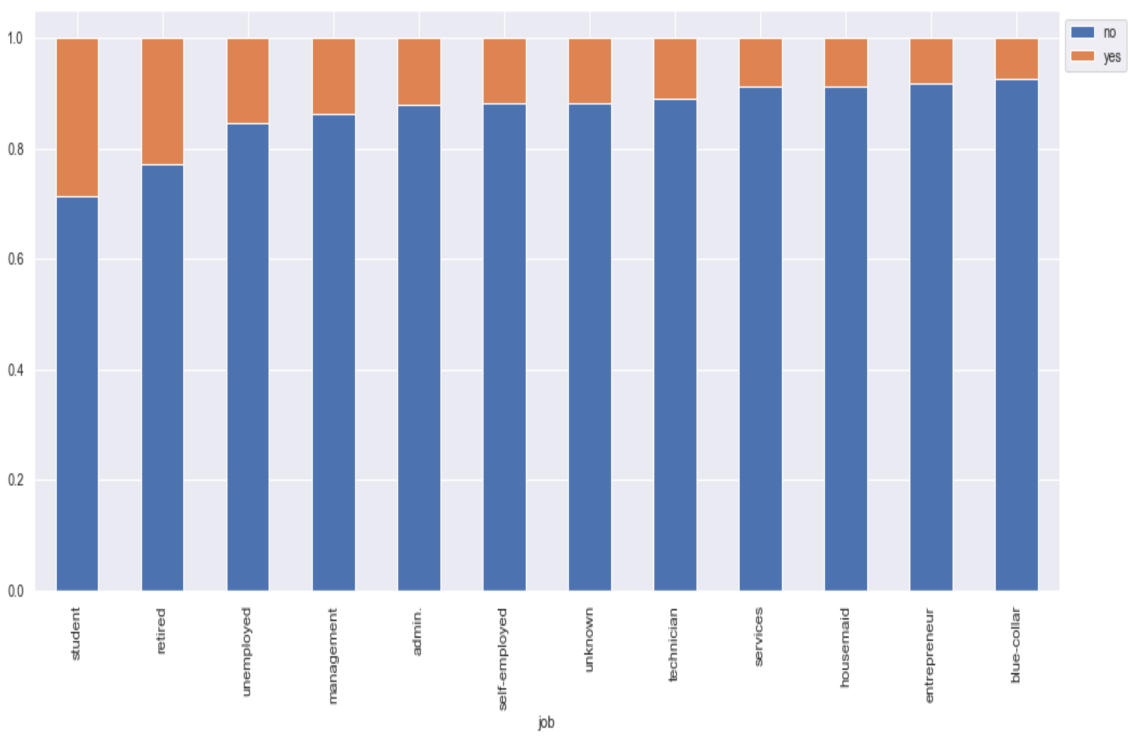
# EDA – Correlation Heatmap

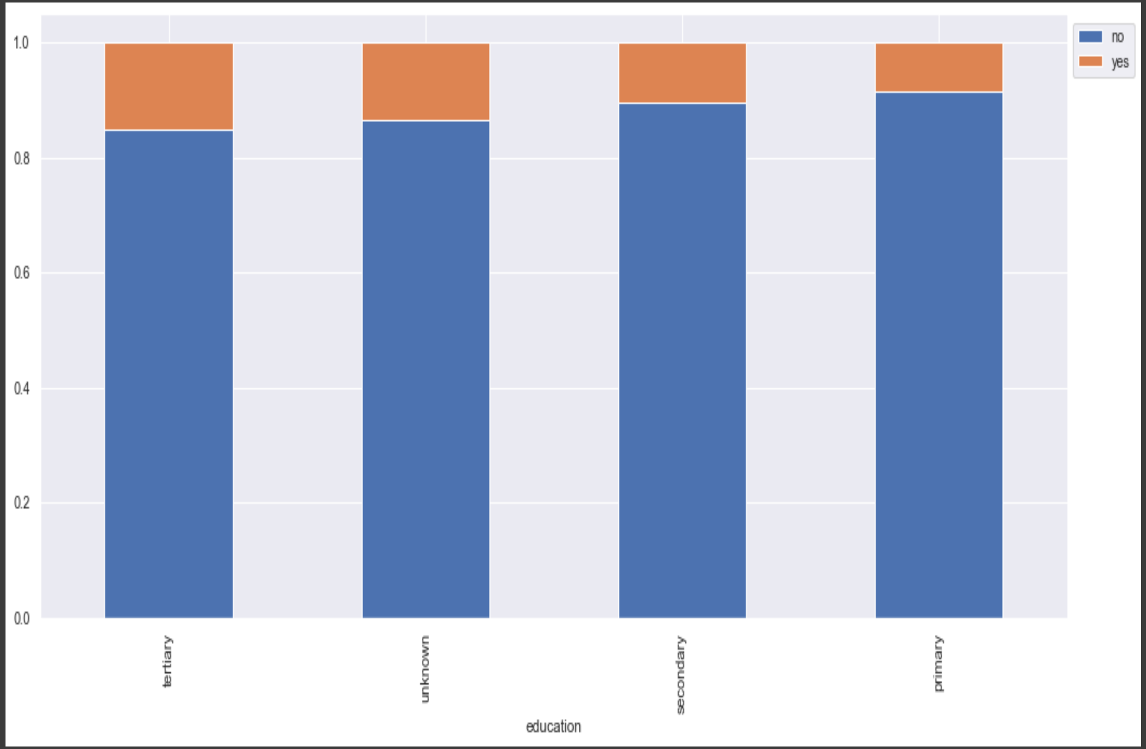


There seems to be no correlation between any of the features. This is good for from a modelling point of view, as correlation would cause multicollinearity and affect the reliability of the model.

Data Glacier
Your Deep Learning Partner

# EDA – Stacked plots

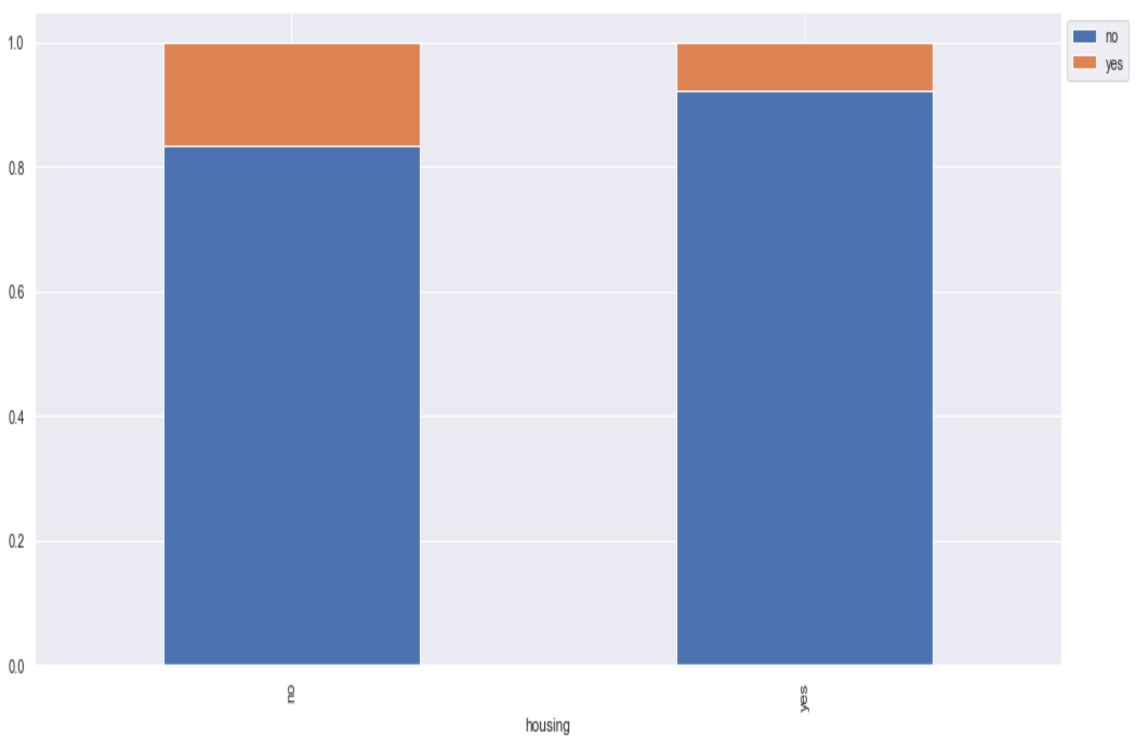


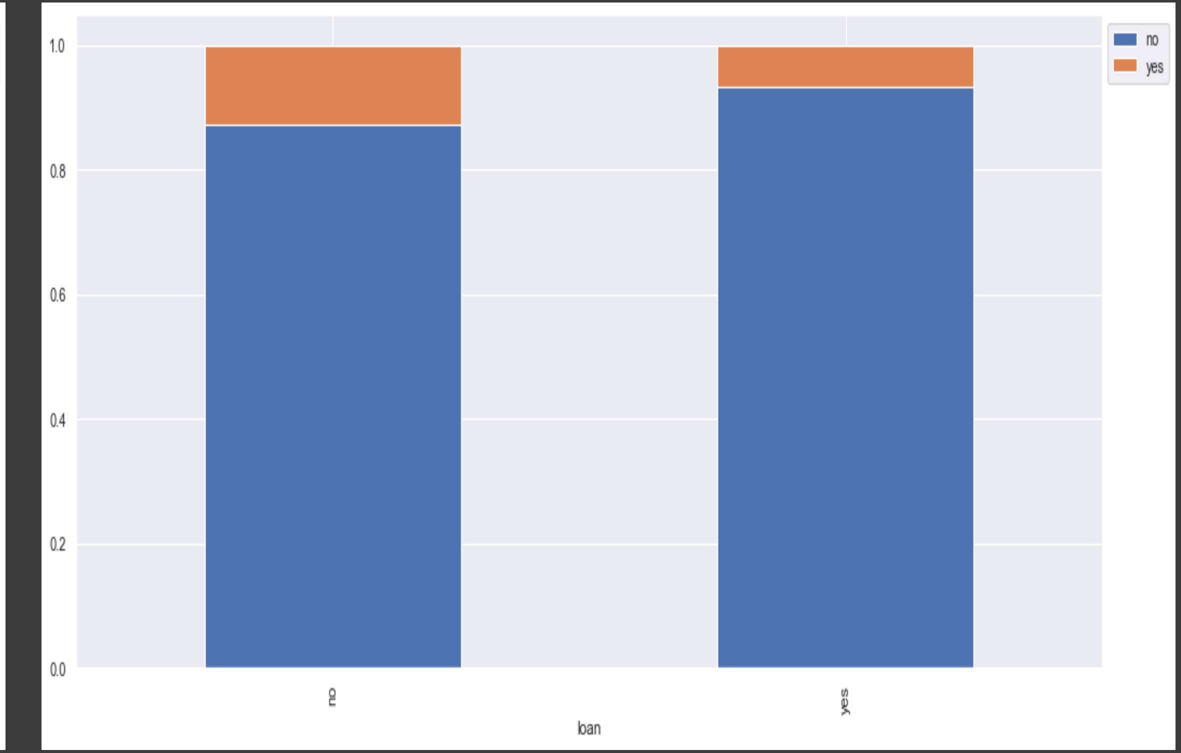


Students seem to have higher conversion rate than others, with around 30%. Retirees and unemployed customers follow behind.

Customers with tertiary education are almost twice more likely to buy the product than those with just secondary or primary education.

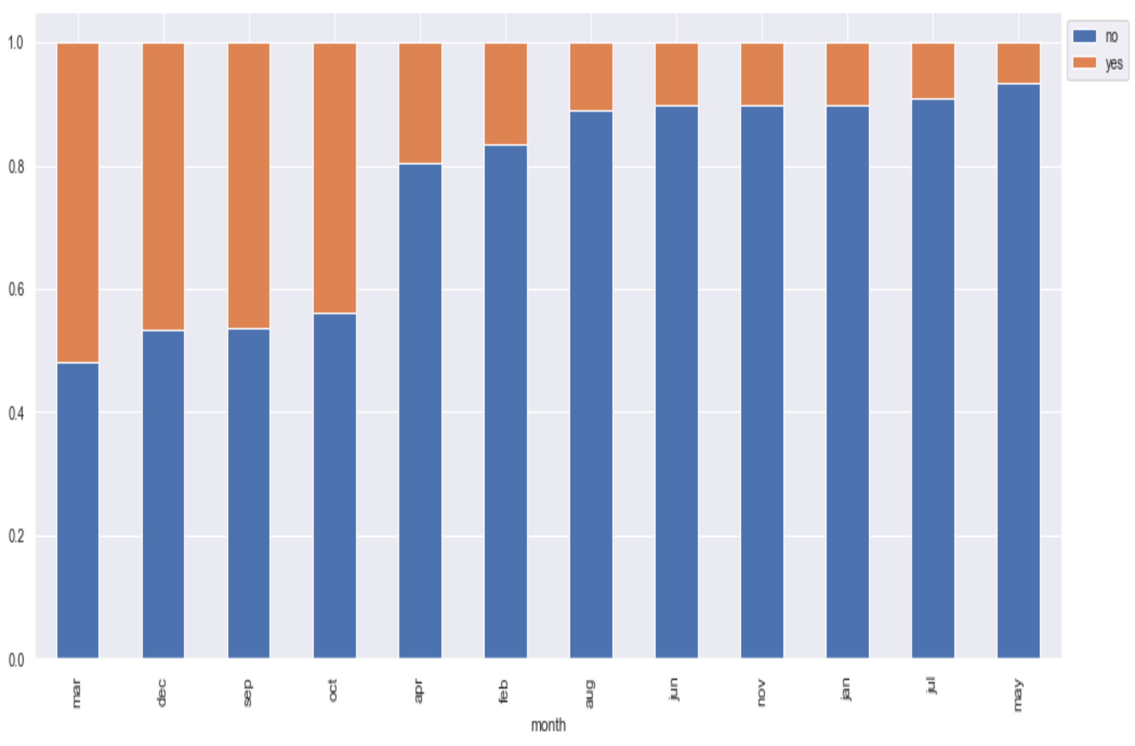
# EDA – Stacked plots

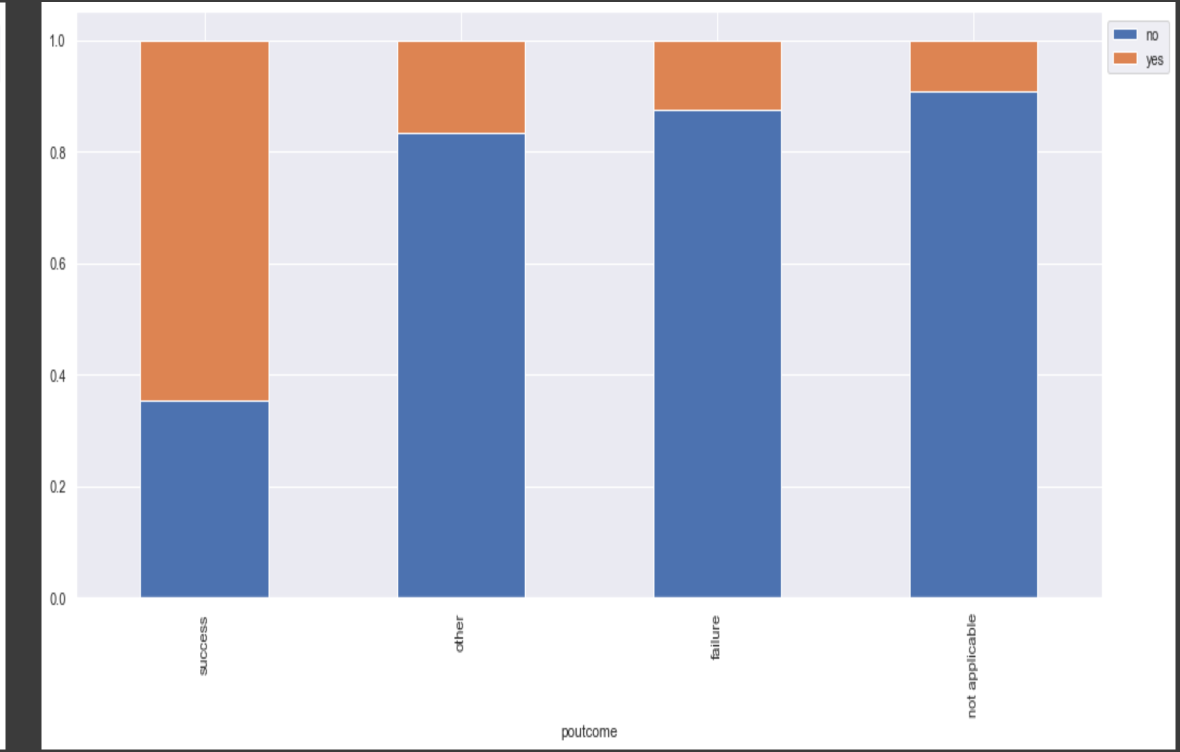




Customers that had no personal loans or no housing loans were twice more likely to buy the product than those who did have those loans.

# EDA – Stacked plots



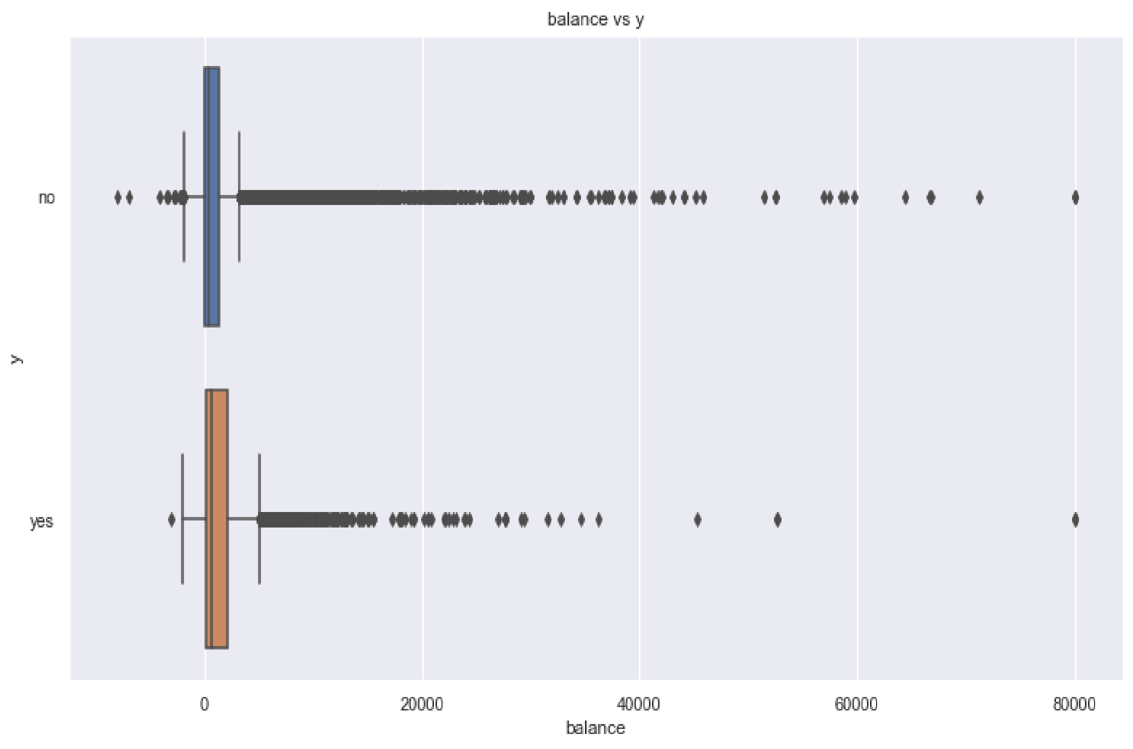


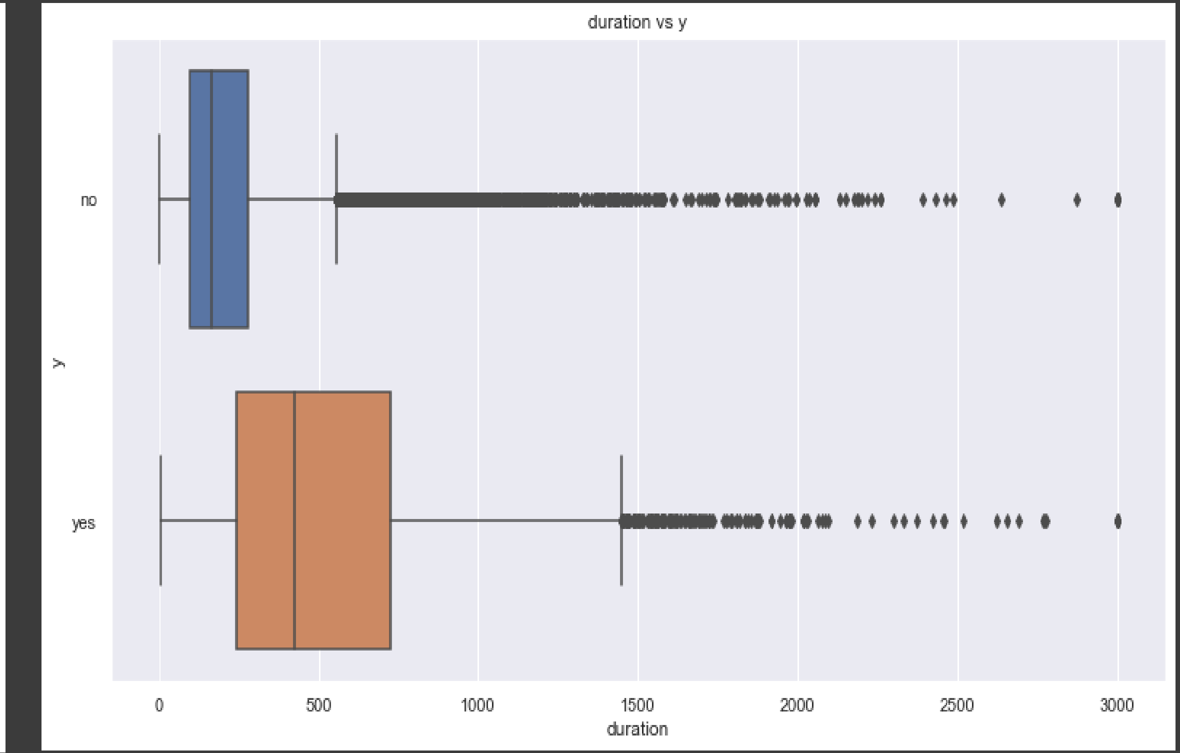
The months of March, September, October and December had almost 50% conversion rate. The reason for this is unclear, but an investigation into this would be recommended, so that any trends can be leveraged.

More than 50% of customers that bought the previous campaign’s product also bought the term deposit. This is a clear indication of the power customer loyalty.

Data Glacier
Your Deep Learning Partner

# EDA – Boxplots

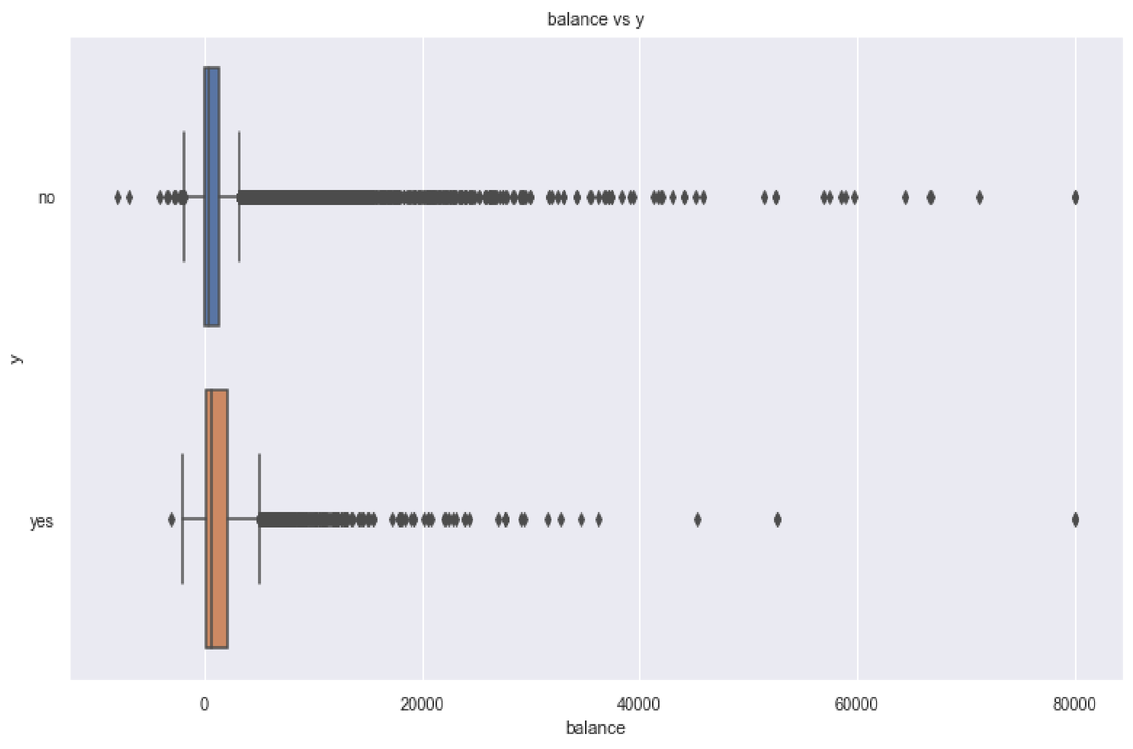


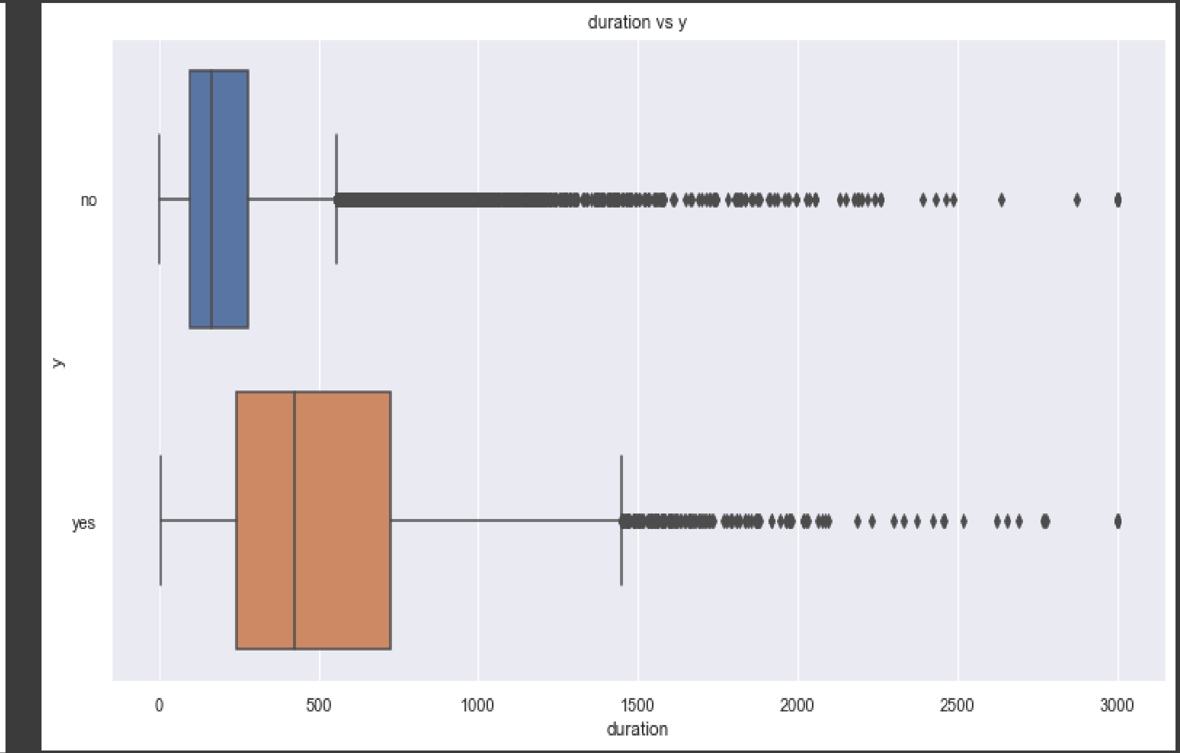


The bank balance of those who bought the product tended to be slightly higher on average.

The duration of the last call of customers who accepted the product seemed to be much larger than those who didn’t buy. This could be due to the purchase, so more details needed to be discussed.

# EDA – Boxplots

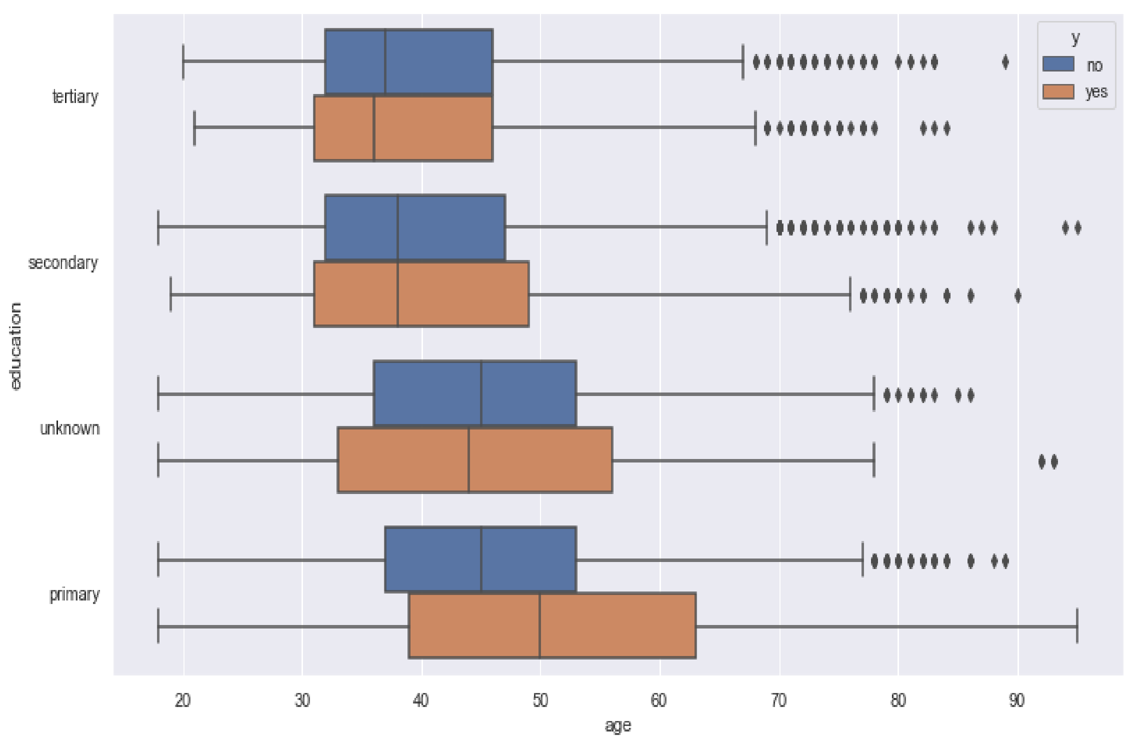


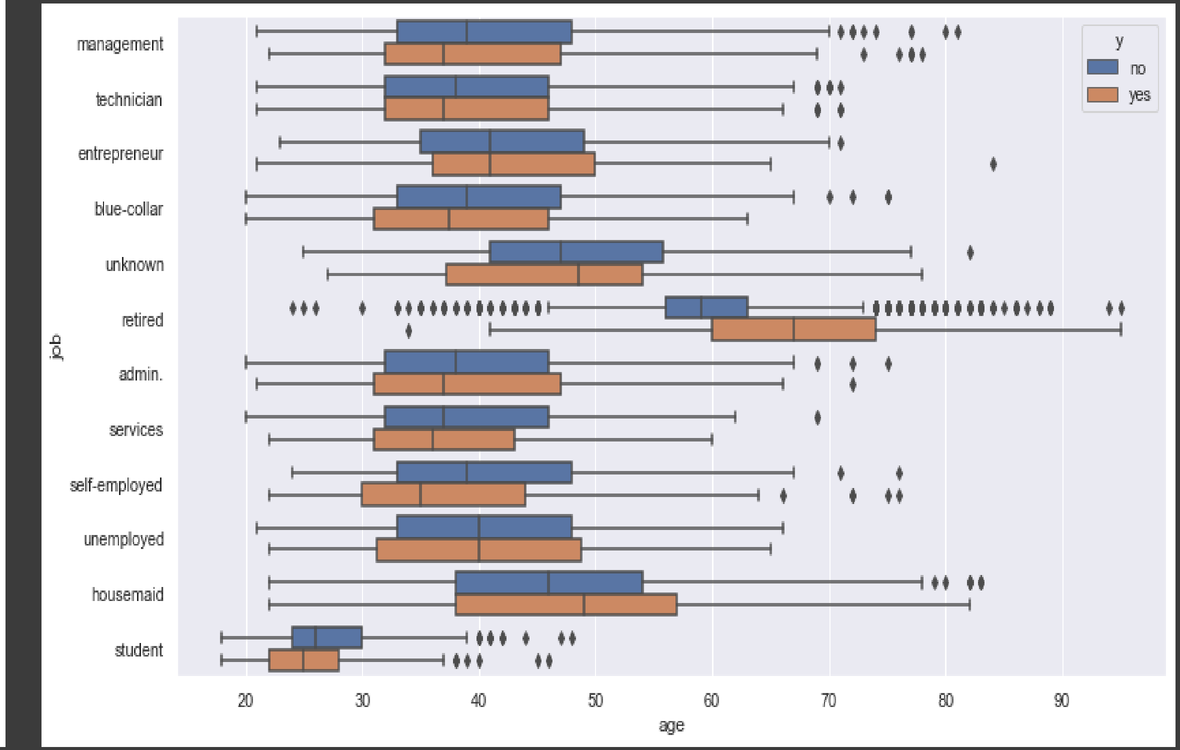


The bank balance of those who bought the product tended to be slightly higher on average.

The duration of the last call of customers who accepted the product seemed to be much larger than those who didn’t buy. This could be due to the purchase, so more details needed to be discussed.

# EDA – Boxplots





Among customers with primary education, older customers tended to convert more often than younger ones.

Among retirees, older customers ( ~ 65 years) tended to buy the product rather than younger (~55 to 60 years) customers.

Data Glacier

Your Deep Learning Partner

# EDA Summary

The key findings from the EDA (including graphs not included here) are:

- Around 4,000 customers have negative bank balances.
- Majority of customers have not been contacted for a previous campaign.
- Majority of customers are either blue collar workers, technicians, or in management.
- Most customers are married. Most customers also have at least a secondary education.
- Only 1.8% of customers have credit default.
- Around 55% of customers have a housing loan, but only 16% have a personal loan.
- More customers were contacted through cellular phone rather than telephone.
- Summer season had the greatest number of customer contacts, in the months of May to August.
- Out of the 18% that were contacted for a previous campaign, only 3.3% were successful in buying the product.
- 11.7% of customers have bought the term deposit. This is an improvement from the previous campaign, it seems.
- The correlation heatmap showed no signs of correlation between any of the features.

# EDA Summary

- Students are the most likely to buy the product, nearly 30% of student customers had bought the term deposit. They are followed by retired and unemployed customers.
- Single customers are slightly more likely to buy than divorced or married customers, but there is not much difference in conversion rate.
- Customers with tertiary education are more likely to buy the product.
- Customers with no credit default, or no personal loan or no housing loan are generally twice more likely to purchase the product.
- Method of communication doesn't seem to have much of an impact on whether the customer buys or not.
- Customers who bought the product in the previous campaign, are far more likely to buy the term deposit. Around 50% of them made the purchase in the current campaign.
- Customers who bought the product tend to have higher bank balances.
- Among retirees, older customers (~ 65 years) tended to buy the product rather than younger (~ 55 to 60 years) customers.

# Recommendations – Customer Profile

The customer profile that can be extracted from the data is as follows:

- Age between 30 to 50 years
- Positive bank balance slightly on the higher side
- Have tertiary education
- Is a student or retiree
- No credit defaults
- No personal loan
- No housing loans
- Purchased product on previous campaign

# Recommendations – Other

Other recommendations include:

- The bank should try to improve their marketing and try to get the customer to confirm the purchase as early as possible, because as the number of contacts increase for the campaign, very few conversions took place.
- The months of March, September, October, and December had the highest conversion rates. The bank can investigate the reason, and if it is not a random occurrence, then the bank can focus most of their marketing in these months. It could, however, be related to the product itself.
- The bank should segment their customers, based on their loyalty and interactions and tailor their services to them.

# Recommendations – Classification Models

For the creation of the classification models, the following techniques can be tried:

- Linear Regression Model
- Bagging Classifier
- Gradient Boost
- XG Boost
- Stacking Classifier
- For all except linear regression models, GridSearchCV can be used to tune the hyperparameters and find a better model.

# Thank You

Data Glacier

Your Deep Learning Partner