Data Glacier

Your Deep Learning Partner

# Exploratory Data Analysis

## Bank Marketing (Campaign)- Group Scientists Project

**12th December 2021**

## Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommended Models

# Background

- ABC Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (telemarketing, SMS/email marketing etc) can focus only to those customers whose chances of buying the product is more.
- ABC bank launched a marketing campaign based on phone calls in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed, so ABC needs a ML model to recommend the more likely customers to subscribe in this service in order to save money and time and contact only the ML model recommended customers.
- **Problem Statement:**
- ABC Bank wants to sell it's 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).

# Approach

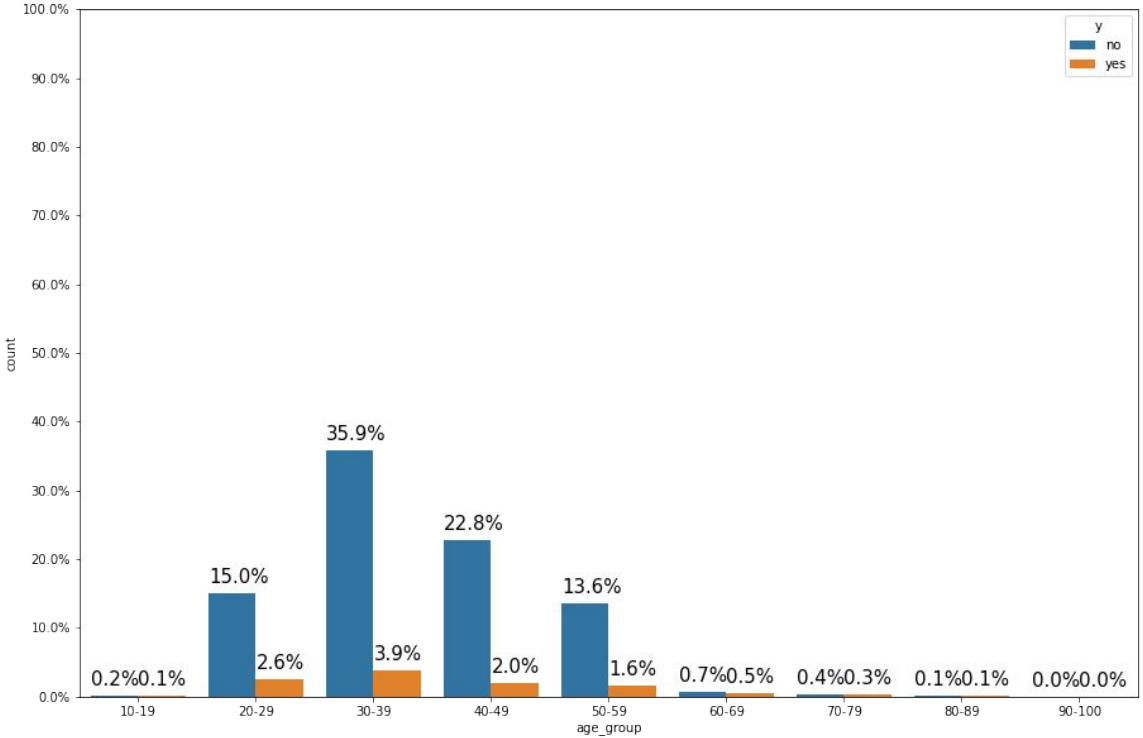
- The dataset had 2 sub datasets that contained information on the full dataset and a dataset that had additional information. We chose the full dataset that had all features. There were 21 and number of entries 411258.
- However the target variable is an imbalanced class with 88.7%- No and 11.3%- Yes.
- The solutions to the imbalanced class include a hybrid solution of using the oversampling and undersampling method together- SMOTE technique plus the RandomUnderSampling technique.

Data cleaning and transformation

- Mapping age to age groups(new feature)
- Mapping campaigns to campaigns groups(new feature)
- Dropping the column ‘pdays’
- Mapping all basic education to be basic only.

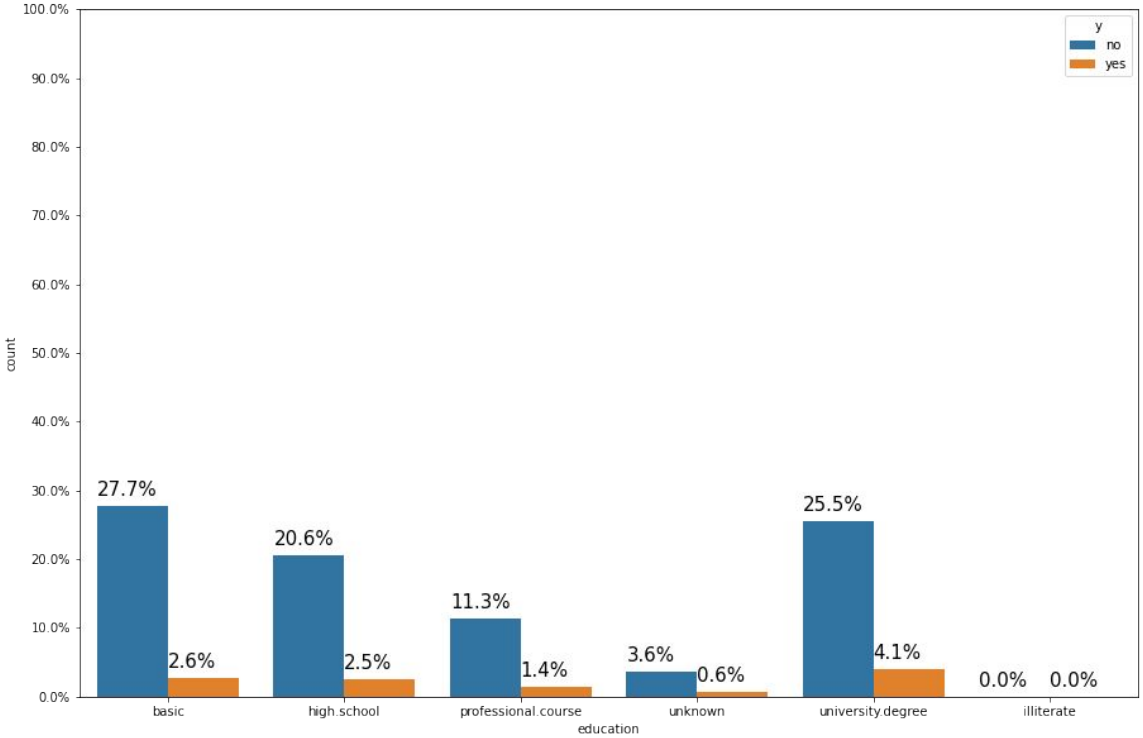
# EDA

# Age- groups and successful campaigns



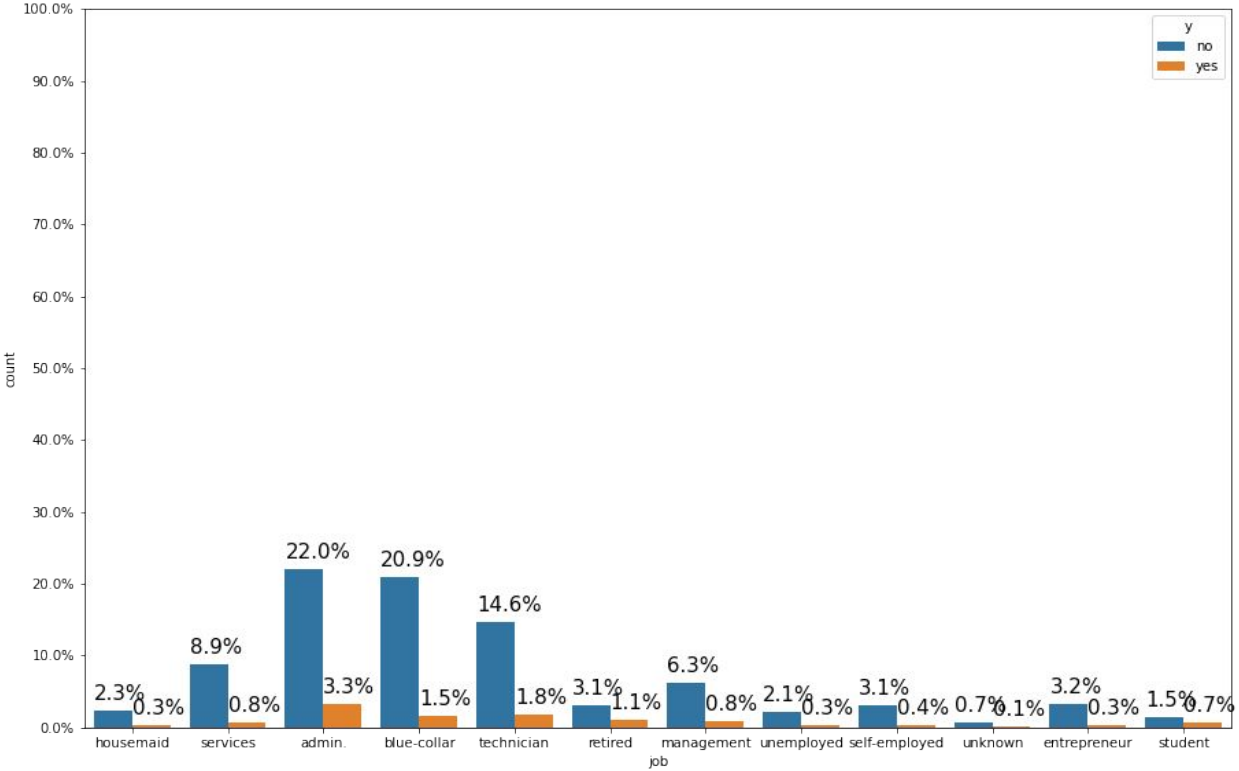

The age group 30-39 were the most contacted and also the ones who accepted the offer more. The number of customers accepting the offer lie between 20- 59 years of age.

# Education and successful campaigns



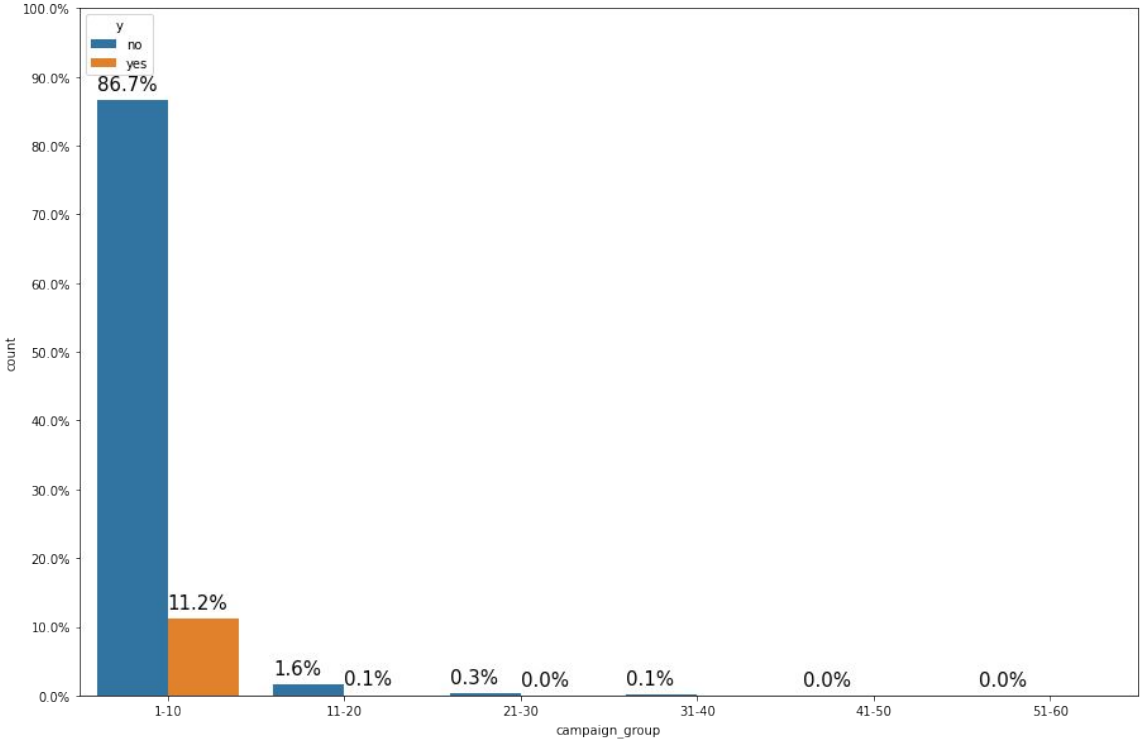

People who had higher education were more likely to accept a term deposit i.e people who have a university degree.

# Education and successful campaigns



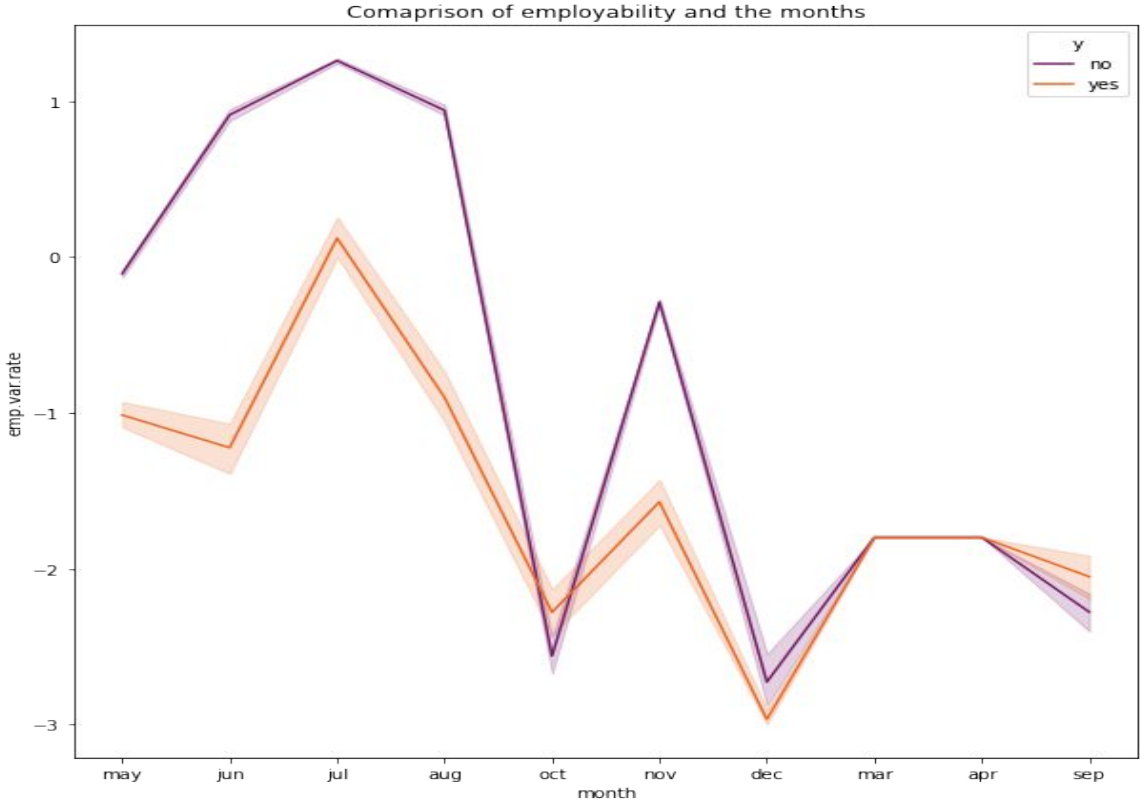

Customers who had a stable form of employment are more likely to subscribe for the term deposit. This accounts for students as well.

# Frequency of successful campaigns



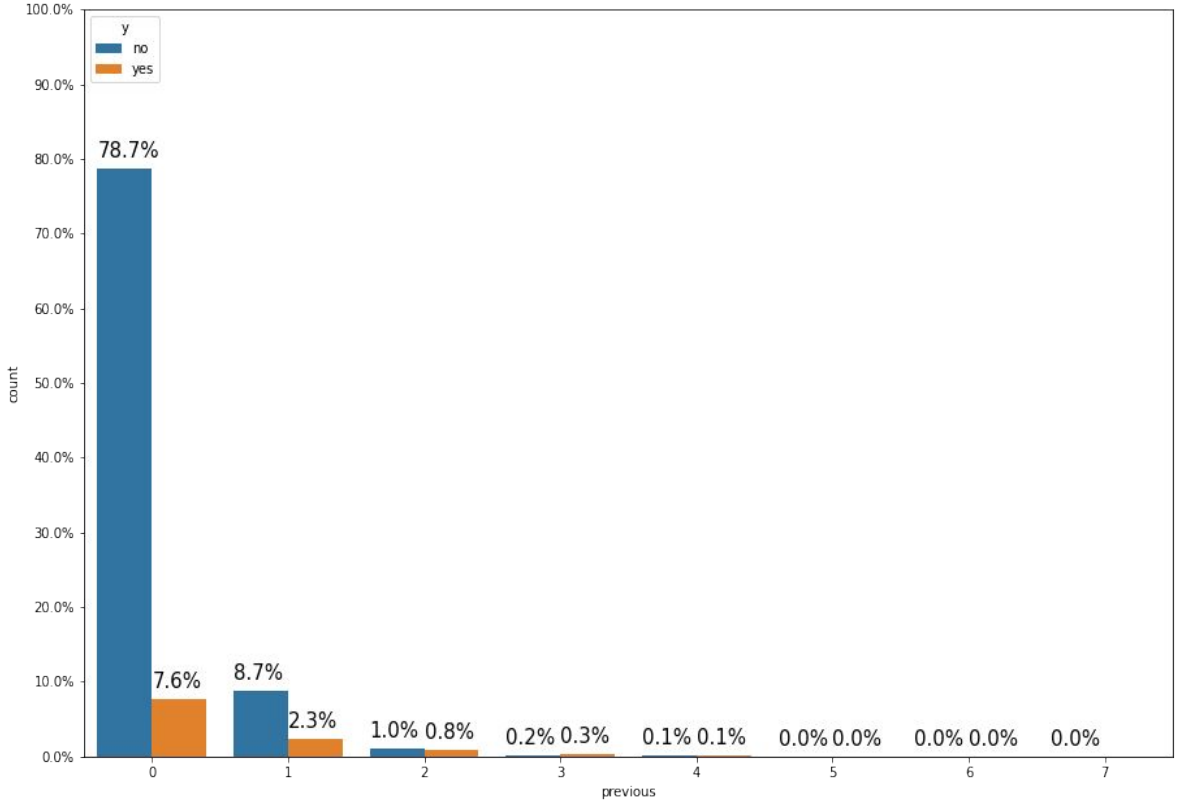

Most customers were contacted between 1 and 10 times and 11.2% subscribed during the campaigns.

# Monthly success of campaigns



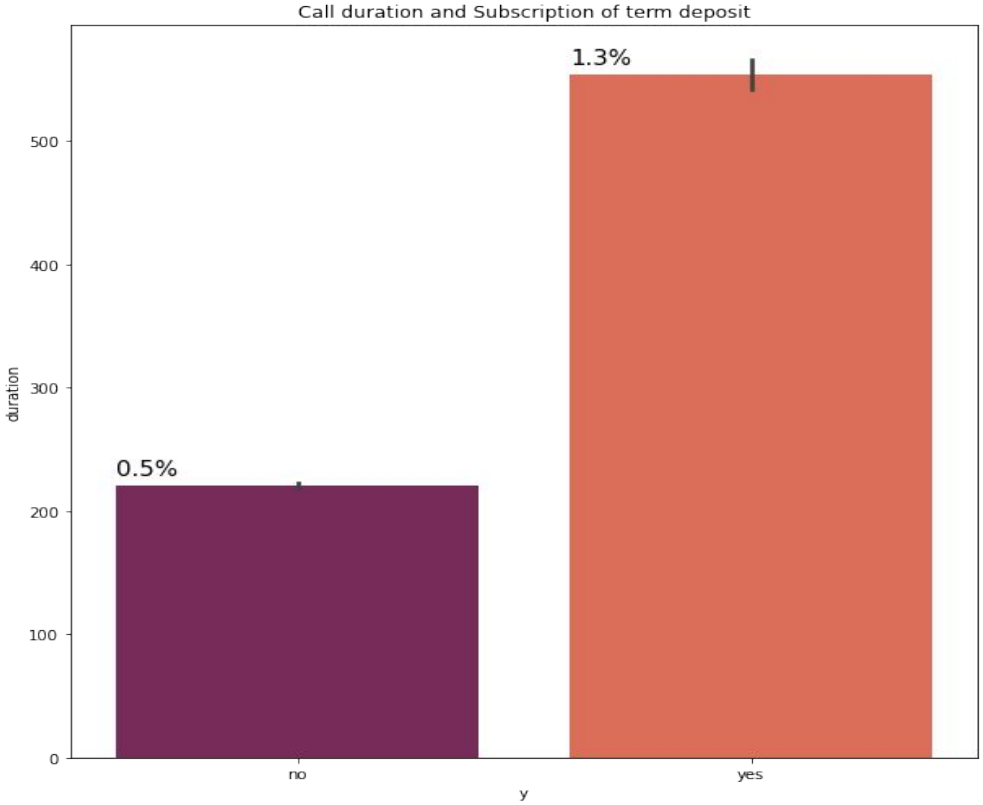

There was a stagnancy in March and April of those who took the term deposit. December had the lowest rate of employability and majority did not take the term deposit. However between May and July the numbers were quite high with July being the highest peak for the people who took the term deposit but it can still be noted that although the three months were good for the customers majority did not subscribe for a term deposit.

# Success of previous campaigns



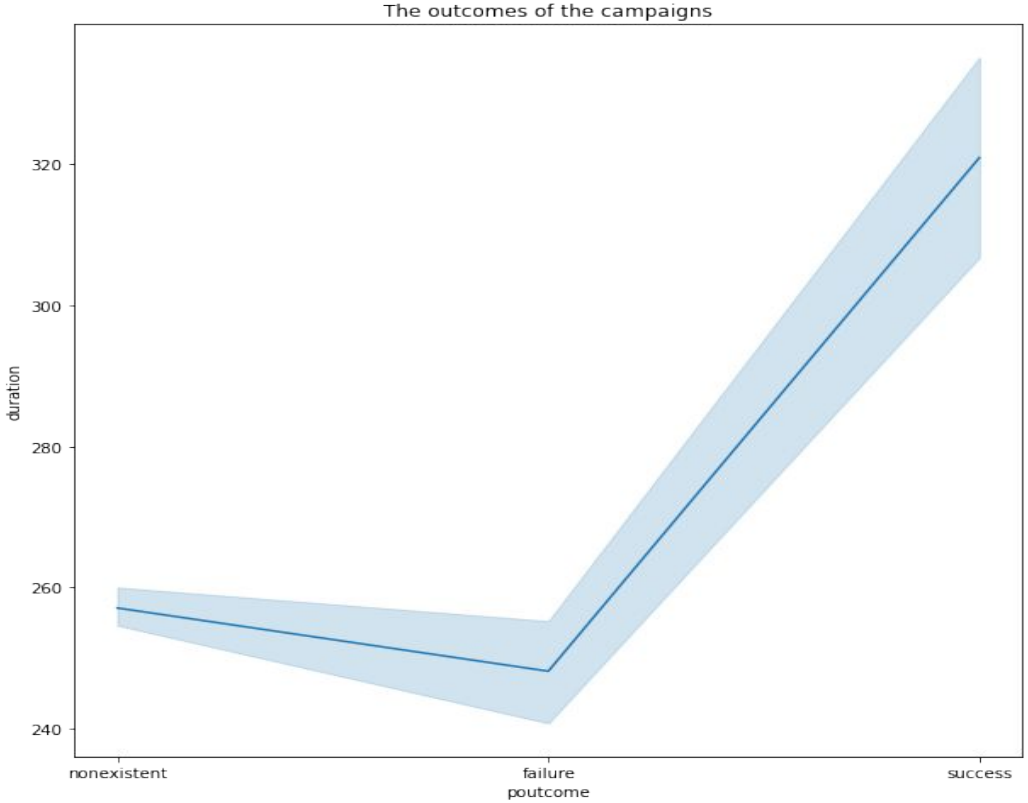

New customers were mostly targeted and had the highest rate of accepting the term deposit.Most people contacted more than two times were very few. Many customers who subscribed to the term deposit were never previously contacted.

# Call duration vs term deposit



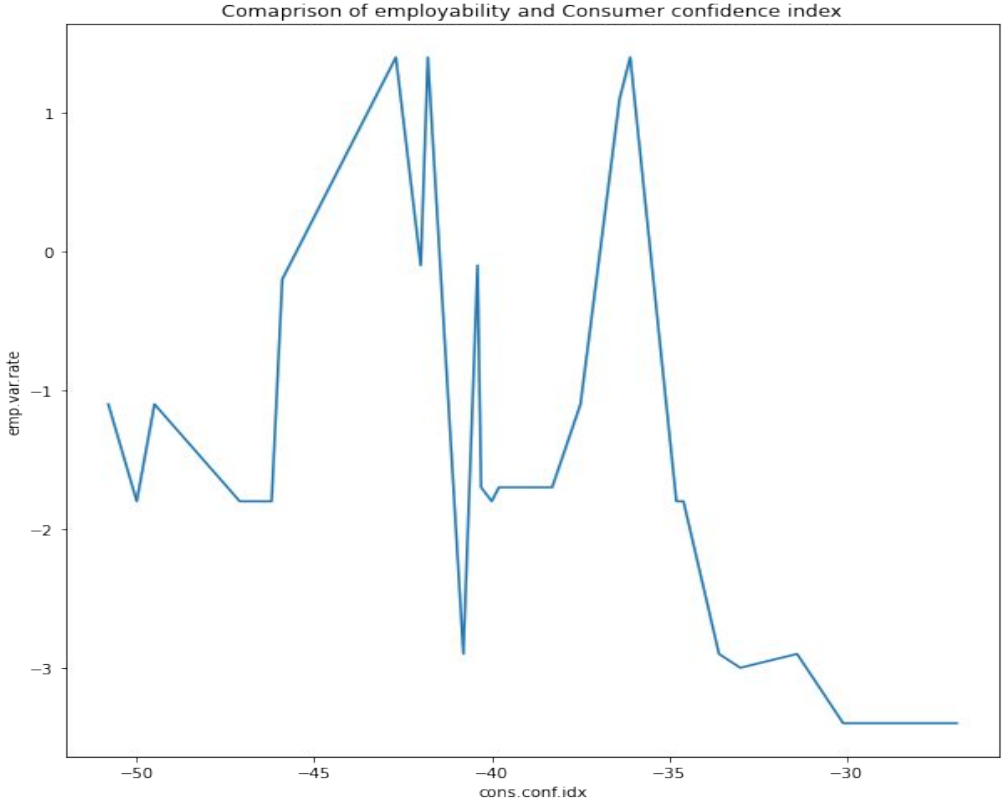

A long call duration has been a factor to influence those who had subscribed for a term deposit , this is accounted by 1.3%. The act of persuasion and persistence does pay off while marketing services and products.

# Call duration vs term deposit



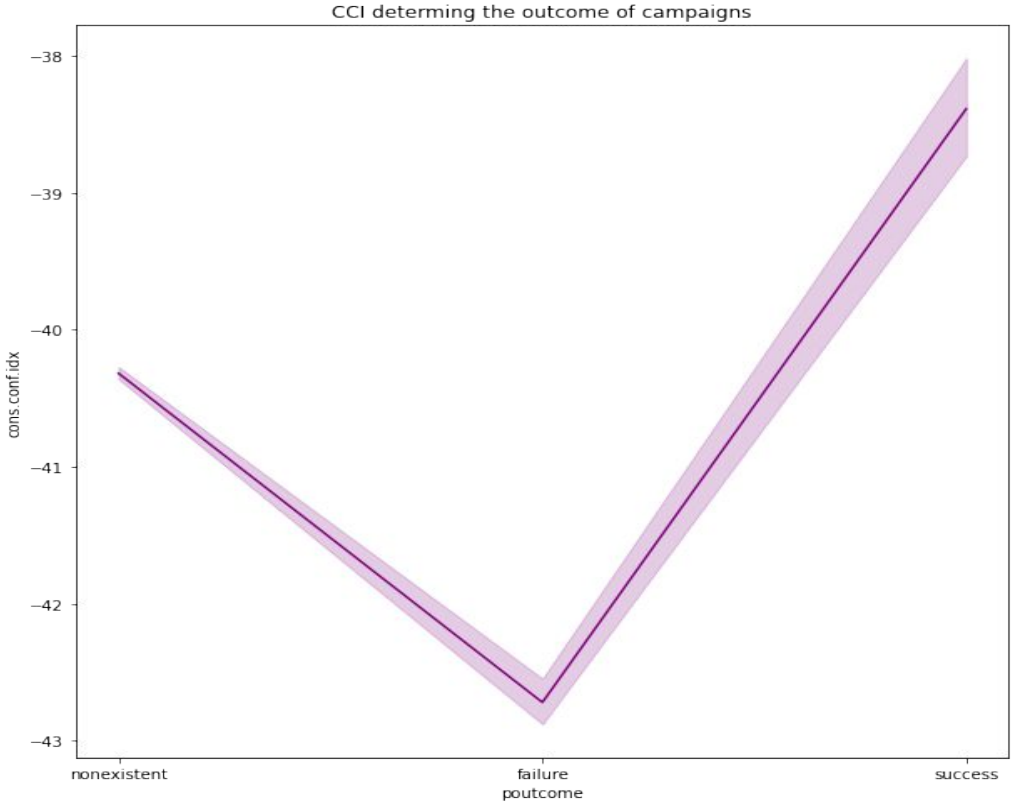

Successful campaigns in the past have been determined by the length of the calls done to customers.

# Economic indicators- CCI and emp.var.rate



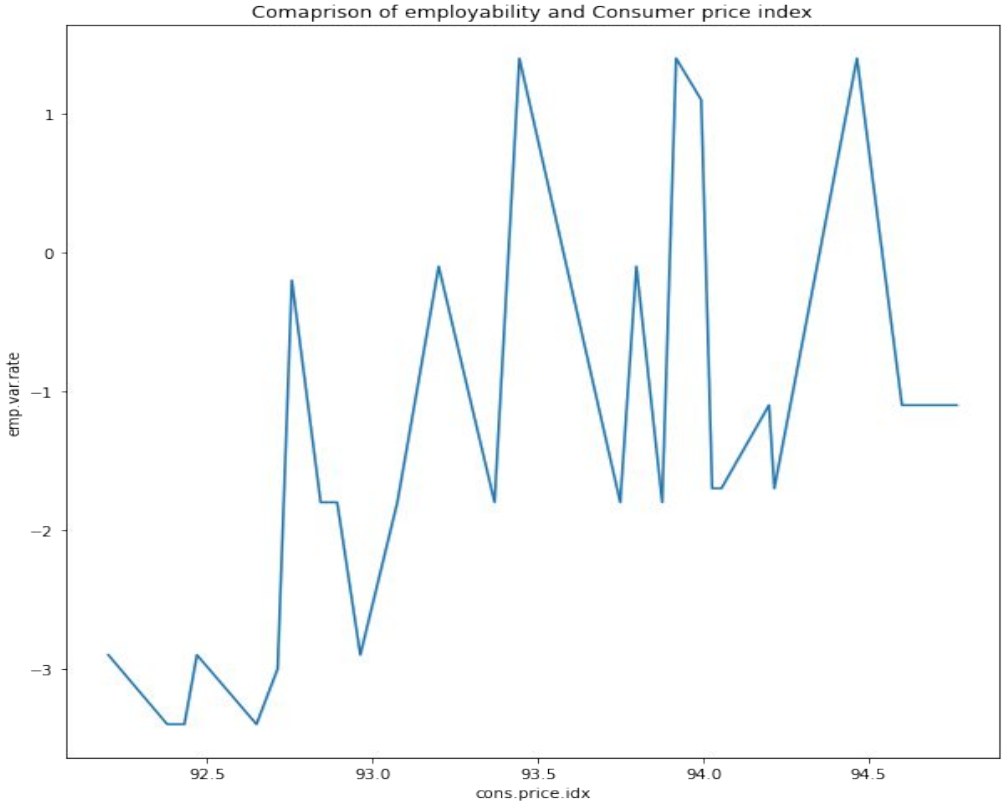

There is an immediate increase and decrease of how customers are likely to spend, that is particularly for the term deposit,those who do not have a stable source of income are likely not to take the offer unlike those who have a stable income.

# CCI and campaign outcomes



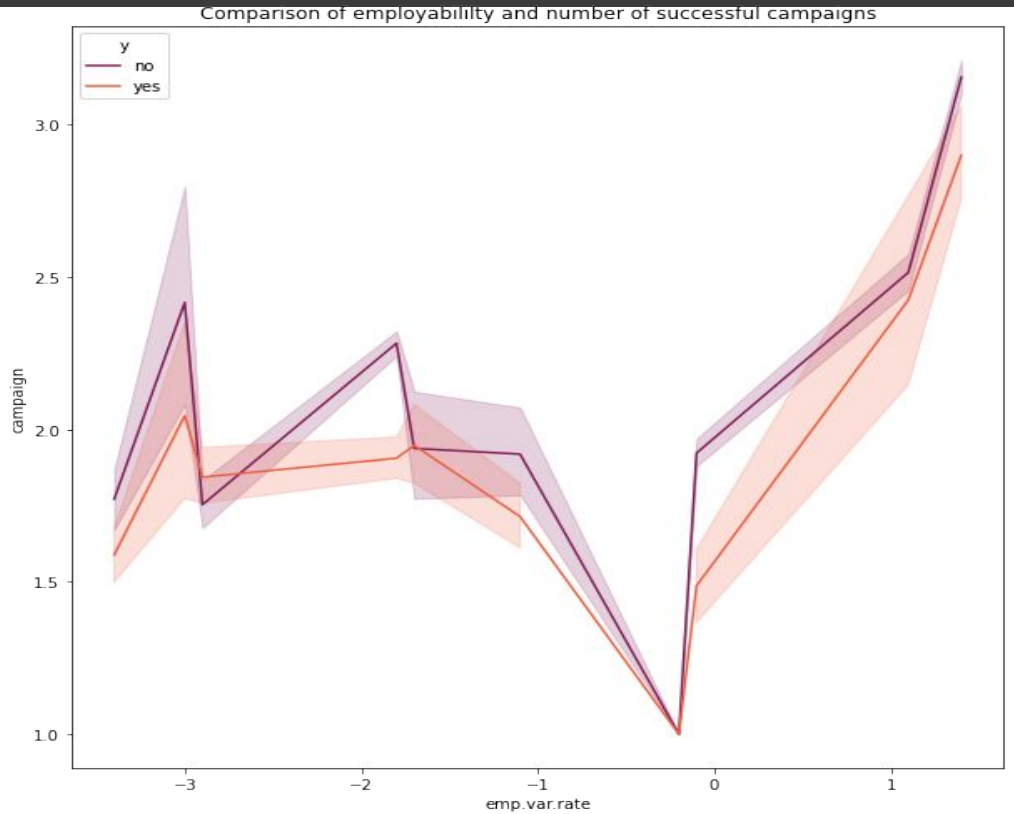

When customers are optimistic about the economy , they are more likely to spend more. A high CCI indicates positive economy and high chances of customers accepting a term deposit.

# CPI and emp.var.rate




An index of 110 means that there's been a 10% rise in the price of the market basket and 90% indicates a 10% decrease in the price of the market basket.There seems to be a positive correlation between the two variables

# Emp.var.rate and the rate of successful campaigns




We can clearly say that when the employability rate was at a zero(not many people having jobs) the number of successful campaigns were low. However when the rate changes to 1 which is a positive rate the customers subscribing to the term deposit are high.

# EDA Summary

# EDA summary

- Targeting 20-59 group age more than other groups.
- Admins, blue collar and technicians are the greatest potential to subscribe to the deposit.
- Married clients showed greater willingness to accept the offer.
- People without bank commitments like loans were contacted more.
- Choosing the write time where the employment rate was at highest significantly helped in increasing the probability of customers accepting the offer.
- Avoid contacting people at the last quarter of the year when holidays and rate of employment decreases, three months of summer are the best months having the highest subscriptions.

# Recommended Models

# Recommended models

- Since this is a classification problem so we need binary classifiers, here is a list of the most popular binary classifiers.

- Logistic Regression
- Random forest
- Gradient boost
- Xgboost
- Decision Tree

# Model Deployment:

- As we have an unbalanced dataset with the majority of responses in our target column being ‘no’,it was decided that we handle this by using a combination of oversampling and undersampling.

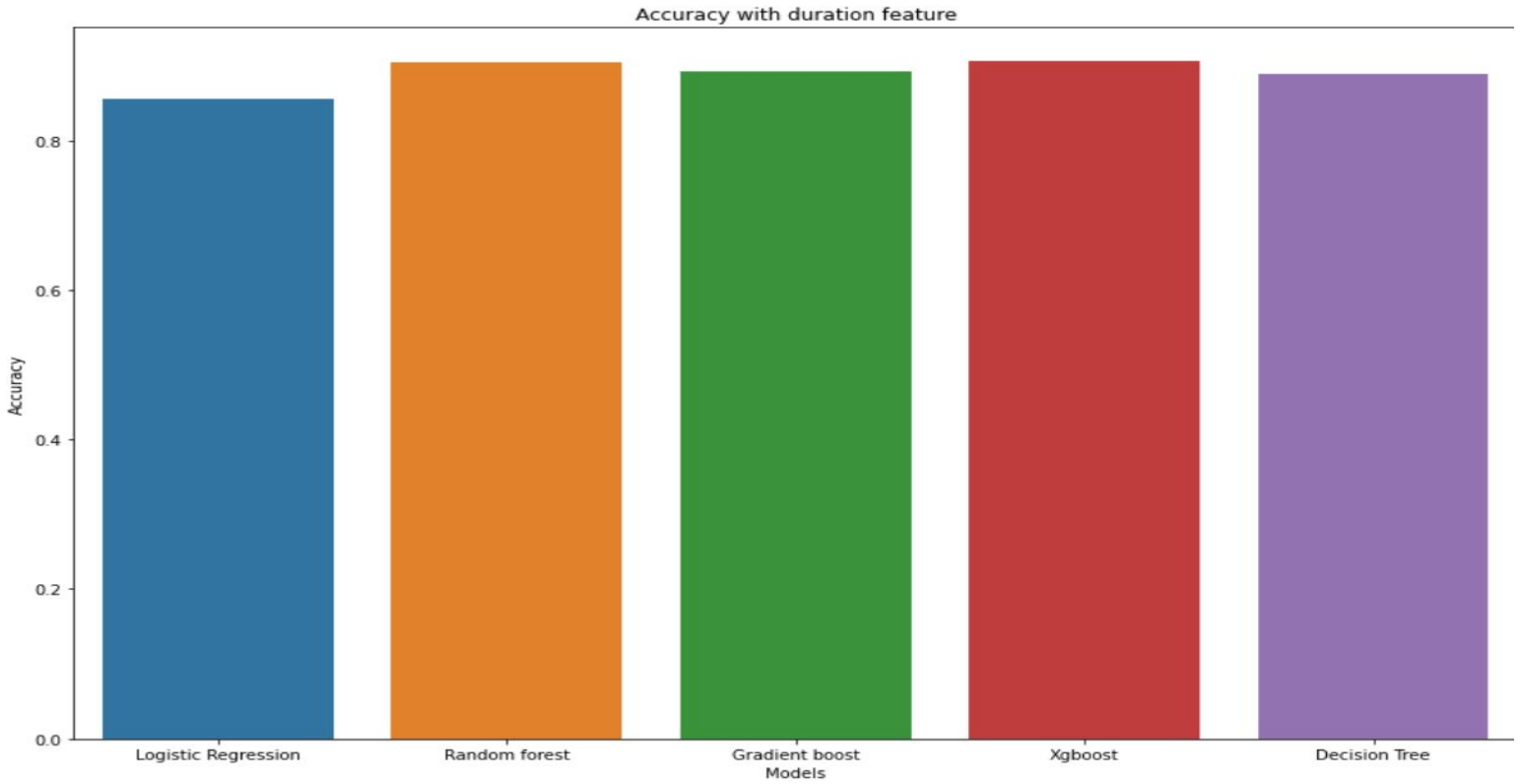
- Each of our 5 models was tested twice, once on the data set including the ‘Duration’ feature and once without it, and the accuracy results were compared.

# Model Deployment With ‘Duration’ Feature - 1:

After deploying and testing our models with the ‘Duration’ feature included in the data, we have the below accuracy for each model:

- Logistic Regression - Accuracy: 0.855668851663025
- Random forest - Accuracy: 0.9033746054867686
- Gradient boost - Accuracy: 0.8914785142024764
- Xgboost - Accuracy: 0.9058023792182569
- Decision Tree - Accuracy: 0.8889293517844137

# Model Deployment With ‘Duration’ Feature - 2:

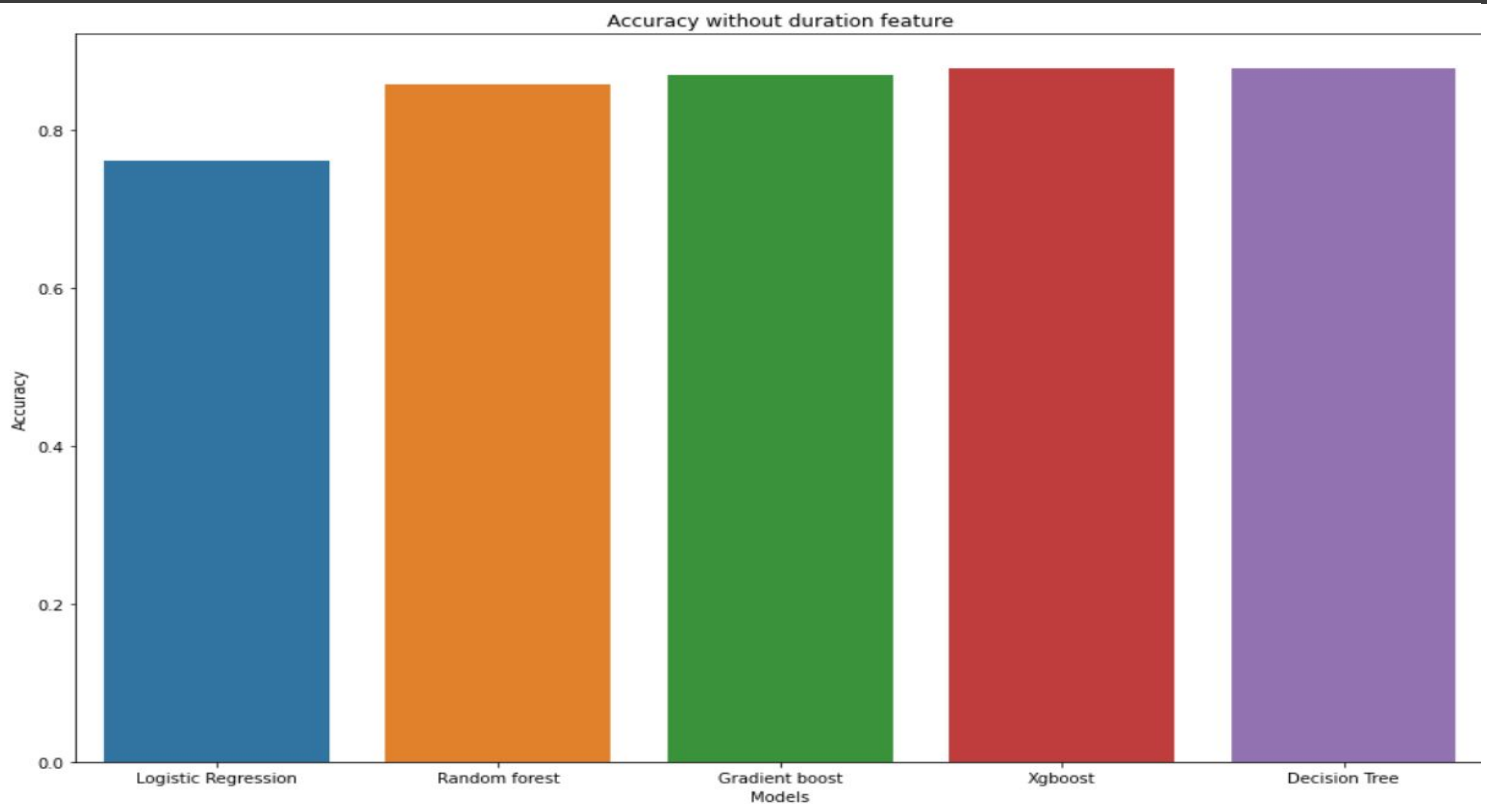


# Model Deployment Without ‘Duration’ Feature - 1:

After deploying and testing our models without the ‘Duration’ feature included in the data, we have the below accuracy for each model:

- Logistic Regression - Accuracy: 0.7609856761349842
- Random forest - Accuracy: 0.8573682932750668
- Gradient boost - Accuracy: 0.8692643845593591
- Xgboost - Accuracy: 0.8766690944403982
- Decision Tree - Accuracy: 0.8770332605001214

# Model Deployment Without ‘Duration’ Feature - 2:



# Model Deployment Conclusions - 1:

- If we check accuracy scores we see that the best 2 models are XGboost and then Decision Tree.
- While XGboost gave the highest accuracy on the dataset that includes ‘Duration’, its accuracy dropped noticeably when we tried it on the dataset without ‘Duration’.
- Decision tree accuracy was more consistent and was barely affected by the removal of the ‘Duration’ feature.

# Model Deployment Conclusions - 2:

| | Model Names | Acc with duration | Auc with duration | Acc without duration | Auc without duration |
|---|---|---|---|---|---|
| 0 | Logistic Regression | 0.855669 | 0.818311 | 0.760986 | 0.713781 |
| 1 | Random forest | 0.903375 | 0.818157 | 0.857368 | 0.669993 |
| 2 | Gradient boost | 0.891479 | 0.856458 | 0.869264 | 0.732966 |
| 3 | Xgboost | 0.905802 | 0.815853 | 0.876669 | 0.700720 |
| 4 | Decision Tree | 0.888929 | 0.832048 | 0.877033 | 0.712915 |

- Since accuracy metric is affected by the test set size so we added the auc metric which is not affected by the test set size.

- By referring to the above table we can conclude that **Gradient boost** classifier was the best model compared to it's AUC and accuracy metrics,so that is why we choose it to be our model.