**Smart banking customer targeting using ML Ensemble for improved business operational efficiency through reduced cost per call (CPC)  
EDA Report**

By Sai Sharan

University of Memphis

# Introduction

Traditionally businesses reach out to prospect customer for encashing potential opportunities (can be in terms of cross sell or upsell). Though this process of targeting customer for potential cross sell or upsell in observed in many industries, it is very wide and frequent seen in banking industry were a customer holding a with a bank will be targeted for cross selling opportunities like loans, fixed deposits and term deposits etc., Banks traditionally uses various channels to reach their customer and one of the major such channel is Telemarketing. Though telemarketing seems like a very easy way to reach a customer at the same time it is very costly. As per industry standard typical Cost per Call (CPC) is around $2.7 to $5.6 [1] (it might difference from business to business), based this statistic we can estimate the possible impact of targeting a wrong prospect and importance of accurate targeting strategy. In the current project we plan to address this problem of high operating cost due to inaccuracy customer targeting using machine learning. As part of the analysis, we will be using banking telemarketing call data for predicting the propensity of a customer opting for cross selling, which can later be used by banking businesses for making better call plan as well customer target list.

Keywords: Business optimization, Machine Learning for CPC reduction, Call Center Optimization

Contents

[Introduction 2](#_Toc127260439)

[Data Overview 4](#_Toc127260440)

[Source 4](#_Toc127260441)

[Data Description 4](#_Toc127260442)

[Collection Process 5](#_Toc127260443)

[Data Summary Stats 5](#_Toc127260444)

[Metadata, Counts and Datatypes 5](#_Toc127260445)

[Quick Summary Stats 6](#_Toc127260446)

[Univariate Analysis 7](#_Toc127260447)

[Insights and Key Distributions from Univariate Analysis 7](#_Toc127260448)

[Age 7](#_Toc127260449)

[Balance, Duration & Campaign 8](#_Toc127260450)

[Pdays and Previous 8](#_Toc127260451)

[Categorical Variable Distribution 9](#_Toc127260452)

[Bi-Variate Analysis 10](#_Toc127260453)

[Insights from Bi-Variate Analysis 10](#_Toc127260454)

[Numerical Variable 10](#_Toc127260455)

[Categorical Variable 12](#_Toc127260456)

[Correlation Analysis 13](#_Toc127260457)

[One to One Correlation Analysis using Pearson’s Correlation Coefficient 13](#_Toc127260458)

[Multicollinearity Analysis using Variance Inflation Factor (VIF) 14](#_Toc127260459)

[Target Variable Class Imbalance 14](#_Toc127260460)

[Conclusion and Next steps 14](#_Toc127260461)

[References 15](#_Toc127260462)

# Data Overview

## Source

Data for current project has been procured from UCI Machine learning repository. Following is the description of repository (as per the website) “*The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine. Since that time, it has been widely used by students, educators, and researchers all over the world as a primary source of machine learning datasets.*”

## Data Description

We are using “Bank Marketing Dataset” hosted in UCI repository for the current analysis. Data actually belongs to a Portuguese bank where existing bank customers are targeted for term deposit subscription over phone calls. The data provided has information related to customers and past behavior when targeted with marketing campaign. Following are key fields and their description as per UCI repository

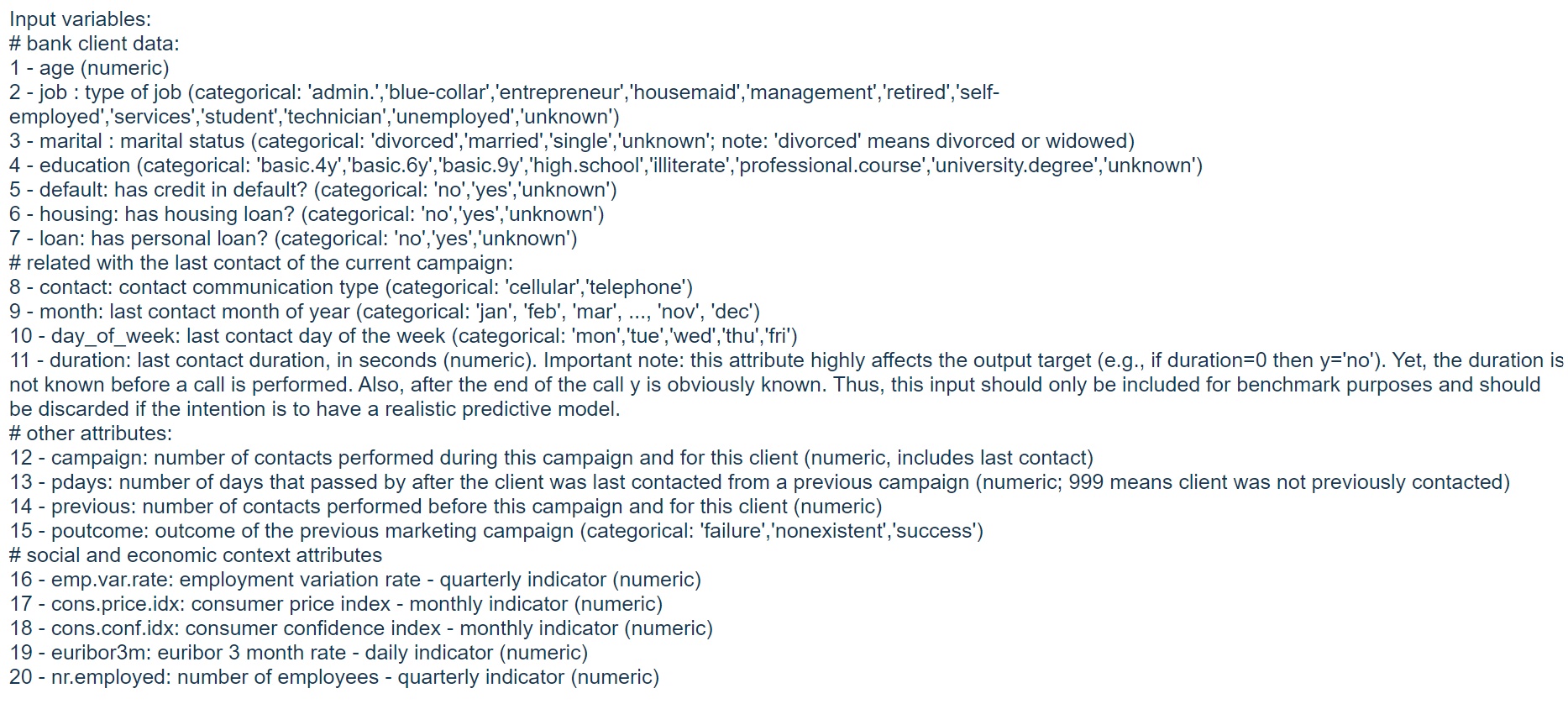


Image showing description of key fields in the dataset

## Collection Process

Data is directly downloaded from UCI website. As required information is available as single file there is no need for data merges. Data is available as CSV file which can be directly loaded into python for model development.

# Data Summary Stats

Understanding metadata and other key attributes like data size, data types only helps in making data preparation process smoother but also makes it faster as proactive data treatments can be made.

## Metadata, Counts and Datatypes

Dataset has close to 45K records with 17 columns. A good split of both categorical (10 fields) and continuous (7 fields) variables are available. Table 1 below show the datatype of each field.

|  |  |
| --- | --- |
| **Variable** | **Datatype** |
| job | object |
| marital | object |
| education | object |
| default | object |
| housing | object |
| loan | object |
| contact | object |
| month | object |
| poutcome | object |
| y | object |
| age | int64 |
| balance | int64 |
| day | int64 |
| duration | int64 |
| campaign | int64 |
| pdays | int64 |
| previous | int64 |

Table 1: *Datatypes of key fields*

## Quick Summary Stats

A high-level glance at the quick summary stats below (Table 2) shows that

1. All categorical fields have limited set of factors (max factor = 12 for job) hence no need for further bucketing
2. All the numerical feature ranges are in expected limits except balance and pdays which has negative values



Table 2: *High-level summary key summary states of data fields*

# Univariate Analysis

To understand the distribution of individual features univariate analysis has been performed on each feature of the dataset. Insights from the univariate analysis will be used for data treatment (if required for missing values, outliers and feature engineering) as well as model approach selection. To understand the univariate distribution Histogram have been used for numerical fields (post binning) and frequency plots are using for categorical fields.

Though univariate analysis has been performed on all the features only insightful distribution are reported in the current document. For detailed univariate EDA charts refer to EDA html file.

## Insights and Key Distributions from Univariate Analysis

Age**:** Age of the customer is nearly normally distributed with few outliers having age above 65 years which can be treated during data preprocessing stage (after careful observation)

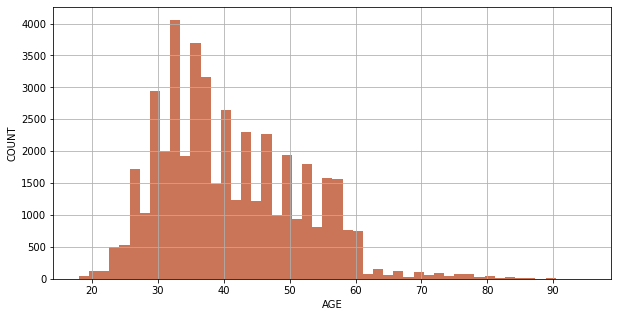


Chart 1: *Histogram of Customer Age*

Balance, Duration & Campaign**:** Balance, Duration and & Campaign variables are exhibiting right skewed distribution which kind of makes sense from intuition perspective. But these variables might need further treatment if models like linear regression are using for prediction**.** Additionally, balance has negative numbers which can be loans etc.,

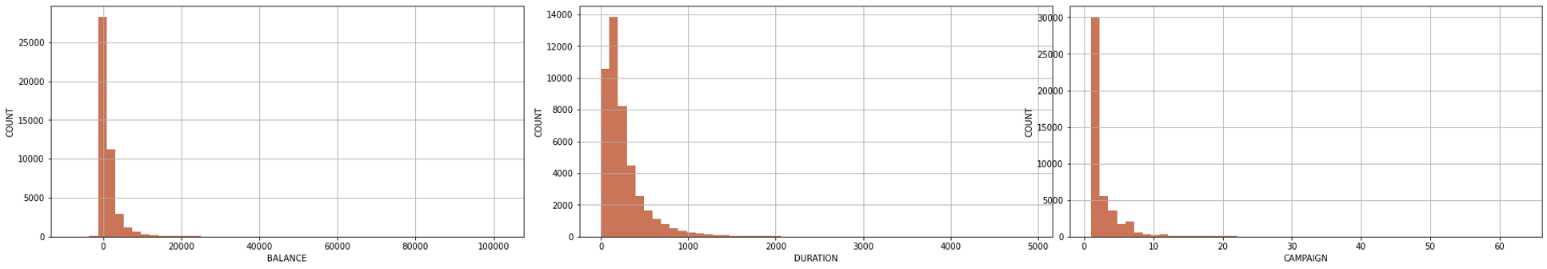
****

Chart 2: *Right Skewed Distribution of Balance, Duration & Campaign*

Pdays and Previous**:** Most of the values in Pdays and Previous are mostly zero. Non-zero values in these are of importance as not every customer is retarget during any campaign.

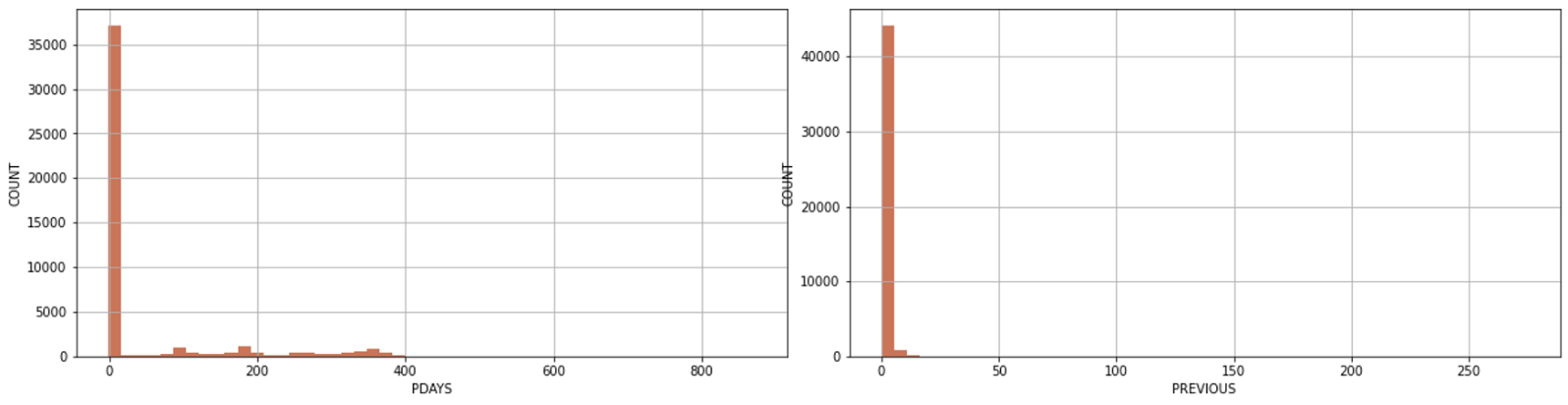
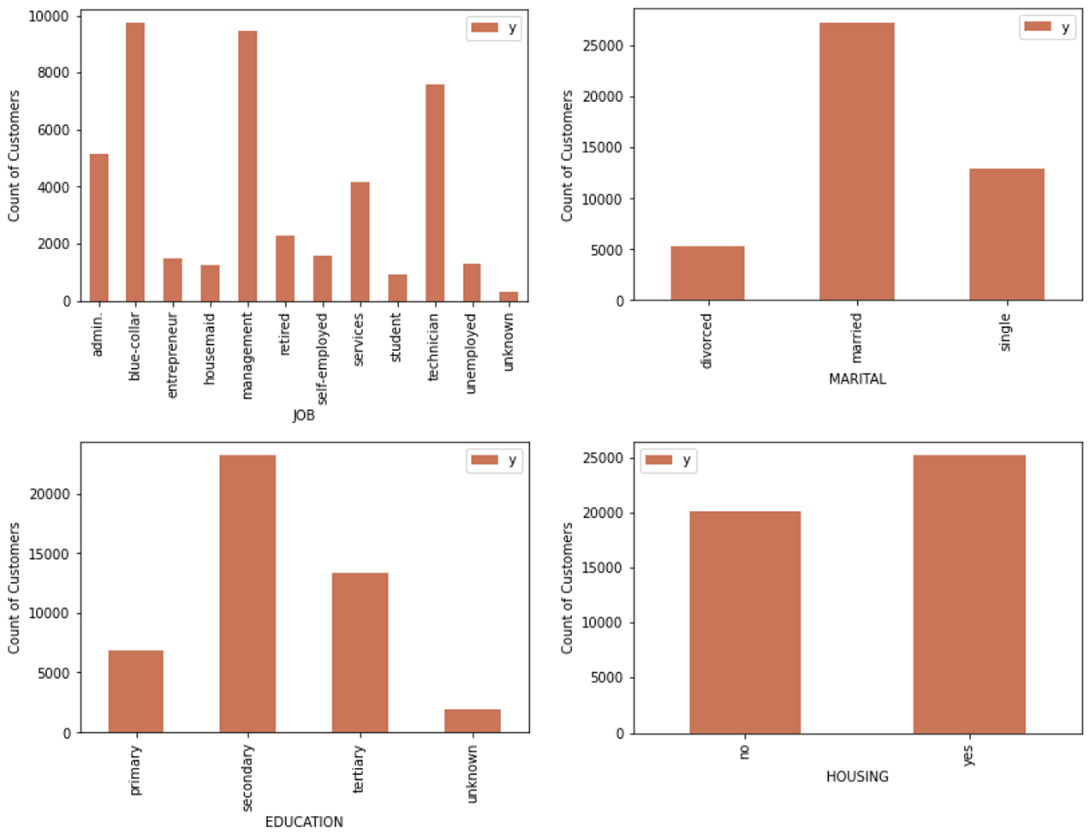


Chart 3: *Distribution of Pdays and Previous*

Categorical Variable Distribution**:**

1. All categorical variables have limited set of factors hence no need for further bucketing of categories before dummy coding of variables.
2. Very few customers are defaulters (default = ‘Yes’). Based on business intuition this can an important feature for model development
3. Job among all categorical variable has relatively higher number of factors, some of the factors can be excluded during preprocessing/model development stage based on their importance
4. Job and Education have factor values as ‘Unknow’, we need to be cautious while using this factor in the model as any insight on this factor is not actionable



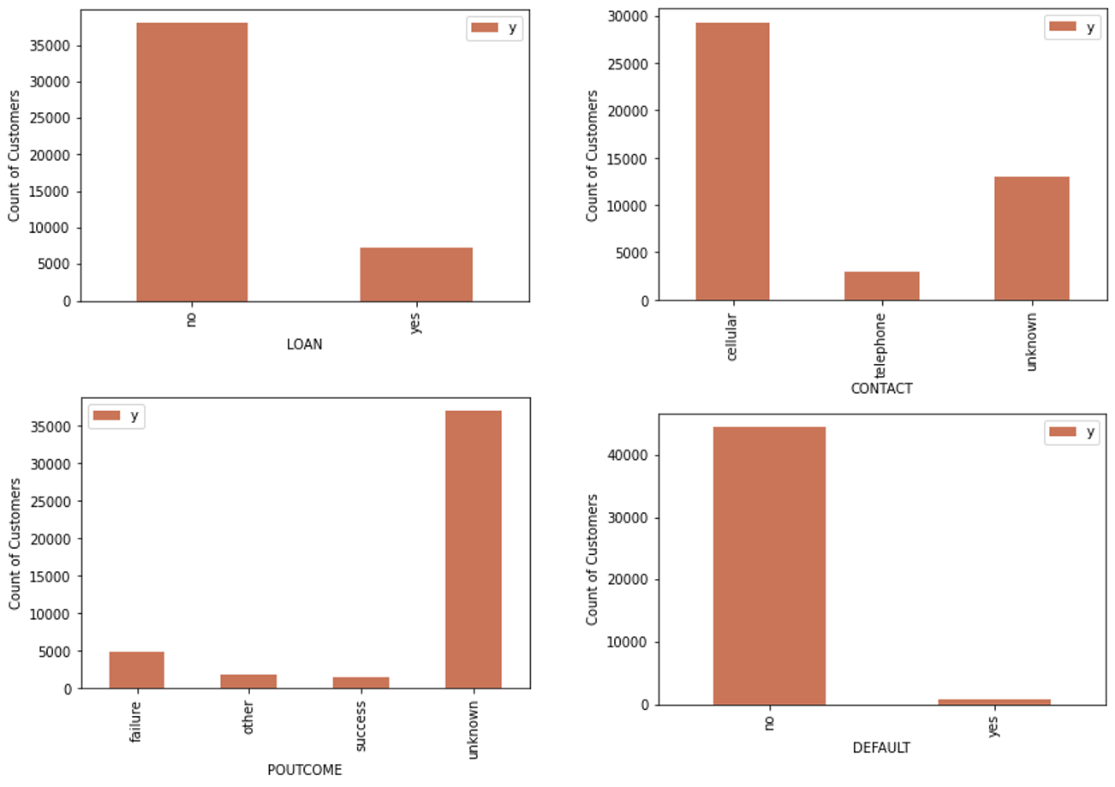


Chart 3: *Frequency plot of categorical variables across various factors*

# Bi-Variate Analysis

## Insights from Bi-Variate Analysis

Bi-Variate analysis on each feature (both categorical and continuous) has been performed for identifying possible the key features/factors impacting term deposit subscription. For categorical variables percentage term deposit distribution and for continuous variable average value feature value for with and without term deposit subscription has been used.

### Numerical Variable

* Average of age and day doesn’t show any significant difference for positive vs negative class [Refer to Chart 4]
* Balance, Duration, Campaign and Pdays variance shows very significant difference for positive vs negative classes [Refer to Chart 5]
  + People with high balance are showing higher tendency to opt for term deposit
  + Customers spending more time when an agent called for the first time are having higher tendency to opt for term deposits
  + Reaching a customer who has been targeted in previous campaigns are showing higher receptiveness in comparison with other customers

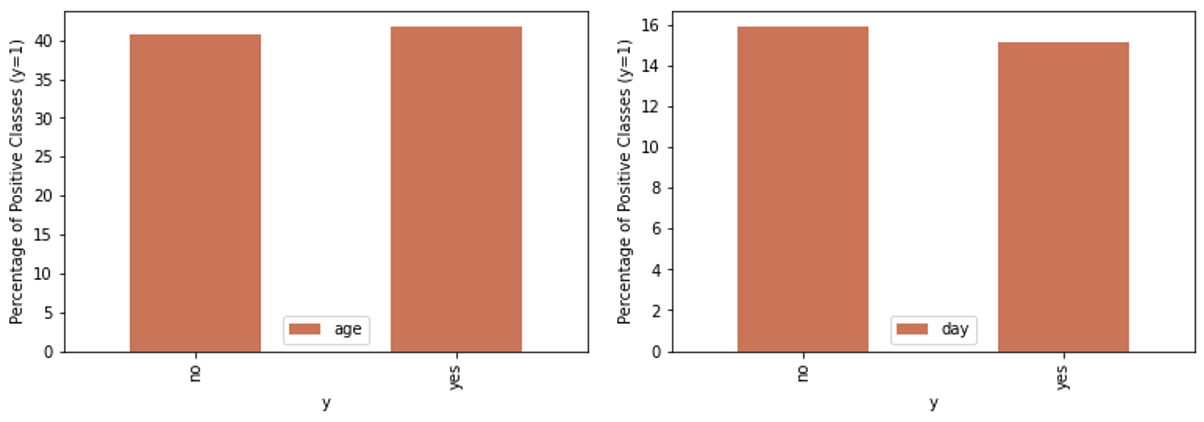


Chart 4: *Average of age, day vs positive & negative classes*

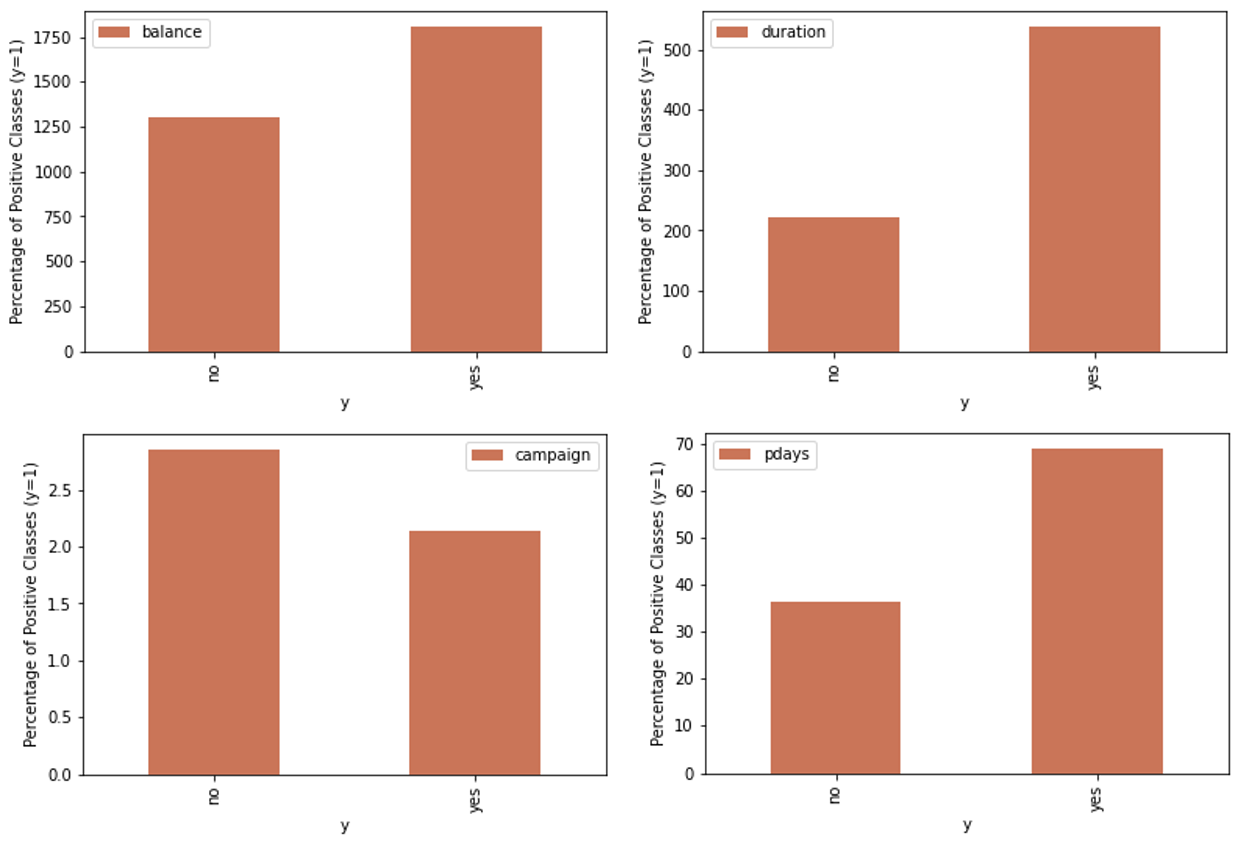


Chart 5: *Average of Balance, Duration, Campaign, Pdays vs positive & negative classes*

### Categorical Variable

* Customers who successfully opted in previous marketing campaign (poutcome) have higher changes of opting for term deposit if targeted
* Students and retired employees have higher percentage of term deposit opter than that of other job groups
* Customer who doesn’t own a house are preferring term deposit than that of customer with own house
* Lower the education level lower are the chances for taking term deposit
* Loan defaulting customers are showing lower receptance to term deposit
* Customers who don’t how an existing loan are opting more for term deposit than that of other customers

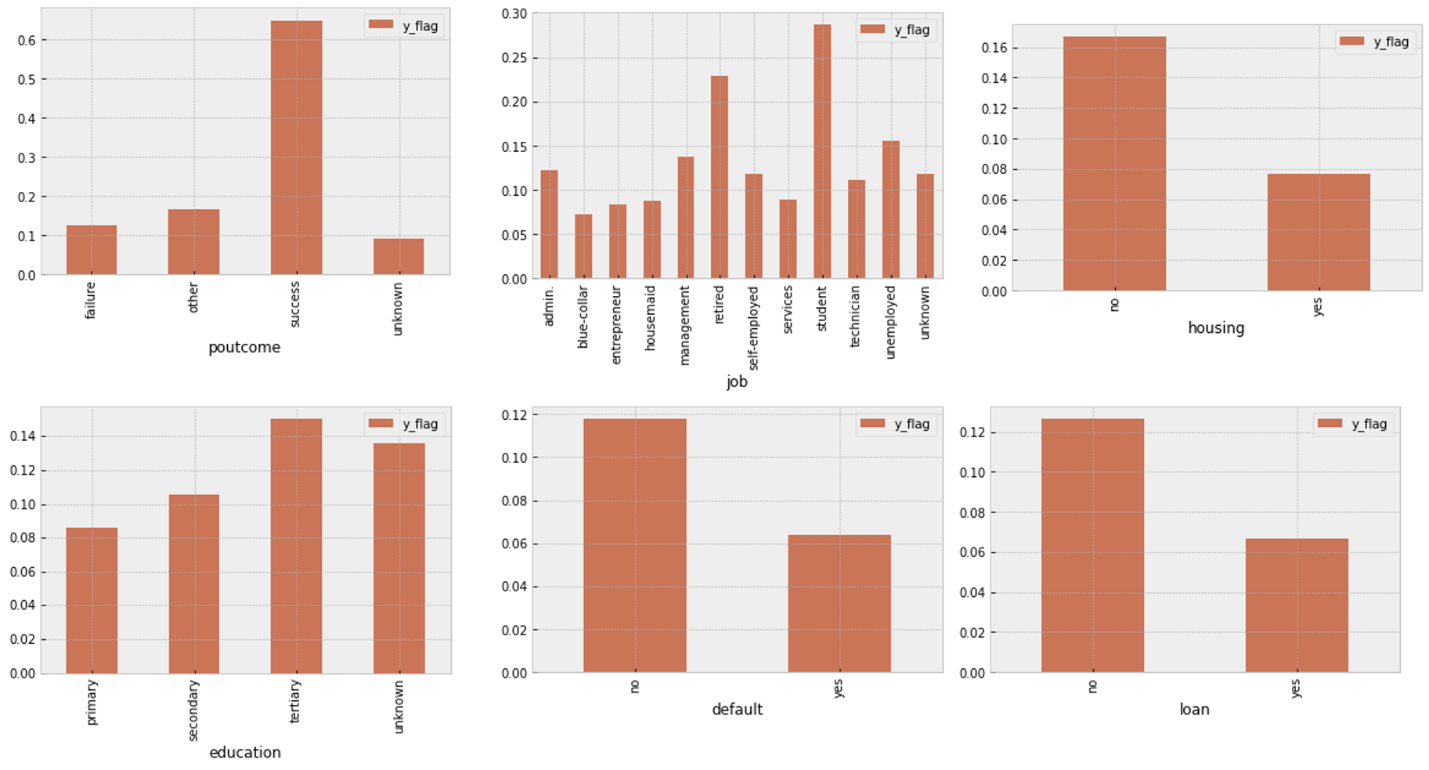


Chart 6: *Percentage of positive classes across different factors of categorical variable*

# Correlation Analysis

Correlation analysis has been performed on the given dataset post dummy coding categorical variables. We have used Pearson’s correlation coefficient for understanding variable one on one correlation and VIF has been used to check the multi collinearity of the variables. Following are key observations from correlation analysis

* poutcome – pdays, Education secondary - Education tertiary and Married\_single - marital\_married are very highly correlated with each other hence they to be appropriately treated while modeling [Refer to Chart 7]
* poutcome\_unknown, month\_may, day and marital\_married feature have > 5 VIF suggesting a strong presence of multicollinearity [Refer to Table 3]

## One to One Correlation Analysis using Pearson’s Correlation Coefficient



Chart 7: *Variable correlation heatmap*

## Multicollinearity Analysis using Variance Inflation Factor (VIF)



Table 3: *Variables with High VIF*

# Target Variable Class Imbalance

Only ~12% of the total customers have subscribed to term deposit (Refer to Chart 8) which means for every 100 records we have only 12 positive classes. This suggests a slight presence of class imbalance issues. During model development incase we observe large amount false negative class balancing might have to be performed.

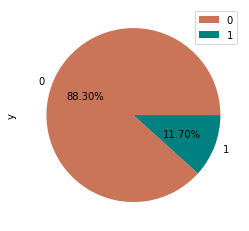


Chart 8: *Distribution customers subscribing for term deposit*

# Conclusion and Next steps

1. Data has no missing values hence no need for missing value treatment
2. Certain outliers have been observed for age variable which can be treated during preprocessing stage
3. Majority for the numeric variables are showing skewed distribution, it is suggested to avoid methods like linear regression as data is violating modeling assumptions
4. High Balance, High Duration, customer targeted in previous campaign, customer last targeted, customers without house, customer without existing loans and customers without any defaults are characters showing high tendency to opt for term deposits – (Observation made from initial EDA) Hence these can be import variables for model development
5. poutcome – pdays, Education secondary - Education tertiary and Married\_single - marital\_married are very highly correlated with each other hence they to be appropriately treated while modeling
6. poutcome\_unknown, month\_may, day and marital\_married feature have > 5 VIF suggesting a strong presence of multicollinearity
7. Data shows the presence of slight class imbalance issue which has to be addressed using appropriate class balancing techniques like over sampling

# References

[1] <https://www.liveagent.com/customer-support-glossary/cost-per-call/>

[2] <https://archive.ics.uci.edu/ml/datasets/bank+marketing>

[3] Stefan Lessmann a, Johannes Haupt a, Kristof Coussement b, Koen W. De Bock c *Targeting customers for profit: An ensemble learning framework to support marketing decision-making* [*https://doi.org/10.1016/j.ins.2019.05.027*](https://doi.org/10.1016/j.ins.2019.05.027)

[4] Justin M. Johnson & Taghi M. Khoshgoftaar *Survey on deep learning with class imbalance,* Journal of Big Data volume 6, Article number: 27 (2019)