

**Week 10: Report on Group Project**

**Topic: Bank Marketing (Campaign)**

**Group Name: Campaign Catalysts**

**Specialization:** Data Science

**Batch Code:** LISUM19

**Date:** 9th May 2023

**Submitted to:** Data Glacier

**Group Member Details**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No.** | **Name** | **Email** | **Country** | **College/ Company** |
| 1. | Yash Jayesh Doshi | yashjdoshi99@gmail.com | UAE | Orpheuss LLC |
| 2. | Anuj Singh | dsanuj21@gmail.com | India | Mumbai University |
| 3. | Yash Jadwani | yash.jadwani1998@gmail.com | United Kingdom | Kingston University |
| 4. | Harold Wilson | haroldwilson537@gmail.com | United Kingdom | University of Buckingham |

1. **Problem Description**

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).

Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc.) can focus only to those customers whose chances of buying the product is more.

# 

# 

# 

# 

# We will be following the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach to understand the data at hand. This approach involves the following steps:

# Collecting the data As described in the data intake report, we have imported four datasets for this project. The first two of these data sets are:

# bank-full.csv with all examples, ordered by date (from May 2008 to November 2010). The bank-full data set contains 45211 X 17 observations.

# bank.csv with 10% of the examples (4521), randomly selected from bank-full.csv. The bank data set contains 4521 X 17 observations and is 10% of the examples (4521), randomly selected from bank-full.csv to test more computationally demanding machine learning algorithms (e.g. SVM).

# More meta-data for these two datasets can be found below.

# Data set location:

# This dataset is publicly available for research. The details are described in [Moro et al., 2011] where it used for Data Mining for Bank Direct Marketing: relating to a direct marketing campaigns of a Portuguese banking institution. [1]

# The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be subscribed or not.

# The data is available and can be located at:

# [pdf] <http://hdl.handle.net/1822/14838>

# [bib] http://www3.dsi.uminho.pt/pcortez/bib/2011-esm-1.txt

# Data Attributes/Variables/Features:

# The main variables /attributes of the data are:

# 1 - age (numeric)

# 2 - job: type of job (categorical, “admin”,” unknown”,” unemployed”,” management”,” housemaid”

# entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services")

# 3 - marital: marital status (categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed)

# 4 - education (categorical: "unknown", "secondary", "primary", "tertiary")

# 5 - default: has credit in default? (binary: "yes", "no")

# 6 - balance: average yearly balance, in euros (numeric)

# 7 - housing: has housing loan? (binary: "yes", "no")

# 8 - loan: has personal loan? (binary: "yes", "no")

# 9 - contact: contact communication type (categorical: "unknown", "telephone", "cellular")

# 10 - day: last contact day of the month (numeric)

# 11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov”, “dec")

# 12 - duration: last contact duration, in seconds (numeric)

# 13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

# 14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

# 15 - previous: number of contacts performed before this campaign and for this client (numeric)

# 16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success")

# Output variable (desired target):

# 17 - y - has the client subscribed a term deposit? (binary: "yes", "no")

The other two datasets are:

1. **bank-additional-full.csv** with all examples, ordered by date (from May 2008 to November 2010). The bank-additional-full data set contains 41188 X 21 observations.
2. **bank-additional.csv** with 10% of the examples (4119), randomly selected from bank-additional-full. The bank-additional data set contains 4119 X 21 observations and is 10% of the examples (4119), randomly selected from bank-full.csv to test more computationally demanding machine learning algorithms (e.g. SVM).

# More meta-data for these two datasets can be found below.

# Data set location:

# This dataset is publicly available for research. The details are described in [Moro et al., 2014]

# S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems. [2]

# The data is available and can be located at:

* [pdf] http://dx.doi.org/10.1016/j.dss.2014.03.001
* [bib] <http://www3.dsi.uminho.pt/pcortez/bib/2014-dss.txt>

This dataset is based on "Bank Marketing" UCI, and is enriched by the addition of five new social and economic features/attributes collected from a national wide indicator from a 10M population country and published by the Banco de Portugal and publicly available at: https://www.bportugal.pt/estatisticasweb and was found to lead to a successful substantial improvement in the prediction process.

# Data Attributes/Variables/Features: 1 - age (numeric)

# 2 - job : type of job (categorical: "admin.","blue collar","entrepreneur","housemaid","management","retired","self-employed","services","student","technician","unemployed","unknown")

# 3 - marital: marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed)

# 4 - education (categorical: "basic.4y","basic.6y","basic.9y","high. school","illiterate","professional. course","university.degree","unknown")

# 5 - default: has credit in default? (categorical: "no","yes","unknown")

# 6 - housing: has housing loan? (categorical: "no","yes","unknown")

# 7 - loan: has personal loan? (categorical: "no","yes","unknown")

# 8 - contact: contact communication type (categorical: "cellular","telephone")

# 9 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

# 10 - day\_of\_week: last contact day of the week (categorical: "mon","tue","wed","thu","fri")

# 11 - duration: last contact duration, in seconds (numeric).

# 12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

# 13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

# 14 - previous: number of contacts performed before this campaign and for this client (numeric)

# 15 - poutcome: outcome of the previous marketing campaign (categorical: "failure","nonexistent","success")

# 16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

# 17 - cons.price.idx: consumer price index - monthly indicator (numeric)

# 18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

# 19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

# 20 - nr.employed: number of employees - quarterly indicator (numeric)

# Output variable (desired target):

# 21 - y - has the client subscribed a term deposit? (binary: "yes","no")

***Note:*** *The dataset chosen for further analysis and for ML model creation will be* ***bank-additional-full.csv*** *because it**seems to be ideal for the purpose of this project since it is more recent and has more variables****,*** *which helps us to build an efficient model.*

# 

# Describing the data

# Below images display the meta-data of the data that we will be using along with the meta-data of its attributes.

# 

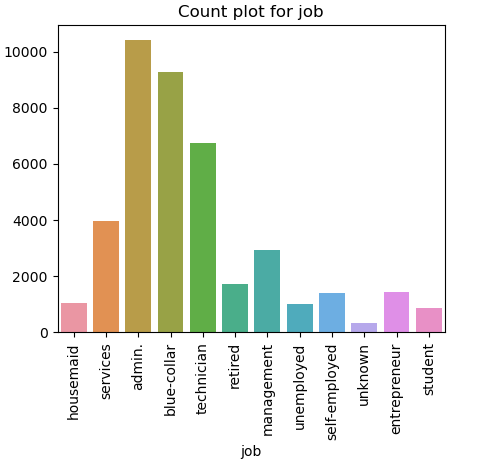
# Below images show the first five records, the last five records and, random five records of the data. This helps us better understand the data at hand. We used the datassist library for the following output.

# 

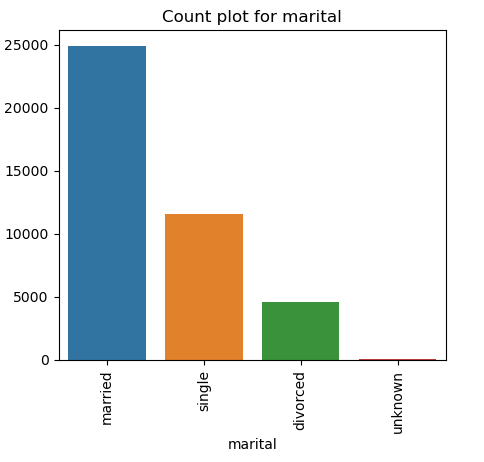
# Lastly, we generated a report about the type of data in each attribute with the help of pandas\_dq library. The following table shows the output of that report

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Feature Name** | **Data Type** | **Missing Values%** | **Unique Values%** | **Minimum Value** | **Maximum Value** |
| age | int64 | 0 | 0 | 17 | 98 |
| job | object | 0 | 0 | admin. | unknown |
| marital | object | 0 | 0 | divorced | unknown |
| education | object | 0 | 0 | basic.4y | unknown |
| default | object | 0 | 0 | no | yes |
| housing | object | 0 | 0 | no | yes |
| loan | object | 0 | 0 | no | yes |
| contact | object | 0 | 0 | cellular | telephone |
| month | object | 0 | 0 | apr | sep |
| day\_of\_week | object | 0 | 0 | fri | wed |
| duration | int64 | 0 | 3 | 0 | 4918 |
| campaign | int64 | 0 | 0 | 1 | 56 |
| pdays | int64 | 0 | 0 | 0 | 999 |
| previous | int64 | 0 | 0 | 0 | 7 |
| poutcome | object | 0 | 0 | failure | success |
| emp.var.rate | float64 | 0 | NA | -3.4 | 1.4 |
| cons.price.idx | float64 | 0 | NA | 92.201 | 94.767 |
| cons.conf.idx | float64 | 0 | NA | -50.8 | -26.9 |
| euribor3m | float64 | 0 | NA | 0.634 | 5.045 |
| nr.employed | float64 | 0 | NA | 4963.6 | 5228.1 |
| y | int64 | 0 | 0 | 0 | 1 |

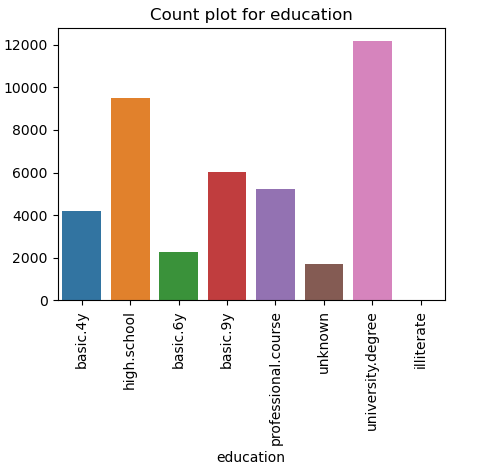
# Exploring the data



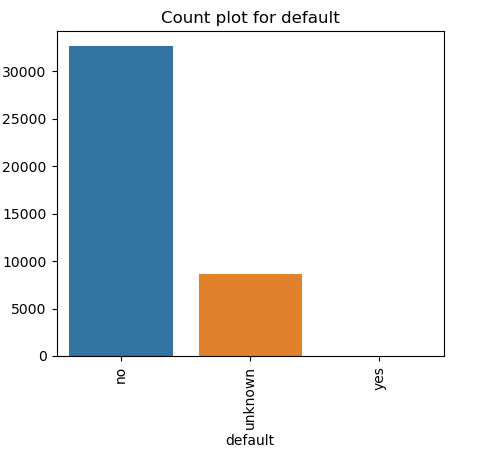
Observation: People with administrative job type were more involved in this project, followed by people with blue-collar jobs and technicians. The lease number of people involved in this survey for the project were students, housemaid and the unemployed.



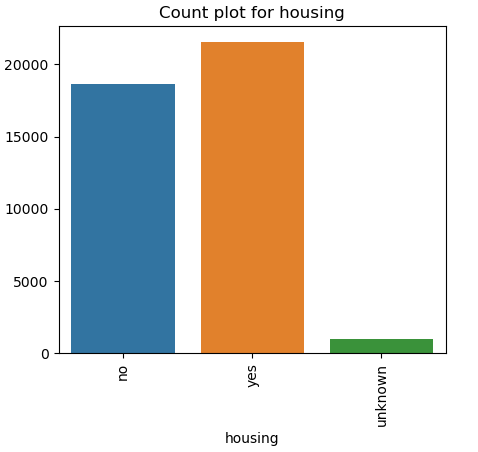
Observation: Married people were more involved in this project follwed by single and the divorced which recorded the least number of peole for this bank marketing campaign survey.



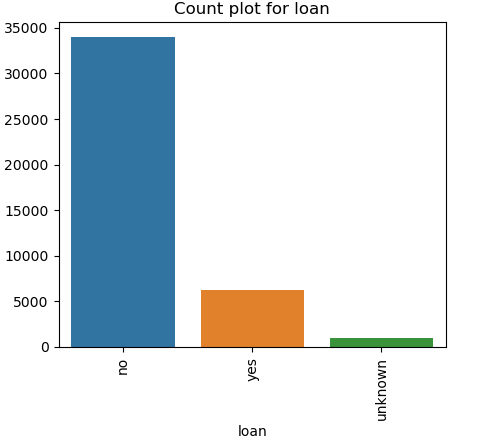
Observation: More people with a university degree took part in this bank marketing survey, followed by high school students. illiterate people recorded the least number considered in this bank marketing campaign survey.



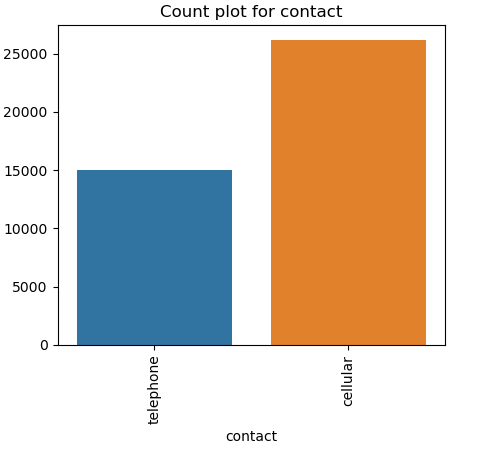
Observation: The number of people that defaulted on credit were much more than the people that did not creating a class imbalance which typically is always a fundamental issue when it comes to classification for machine learning process.



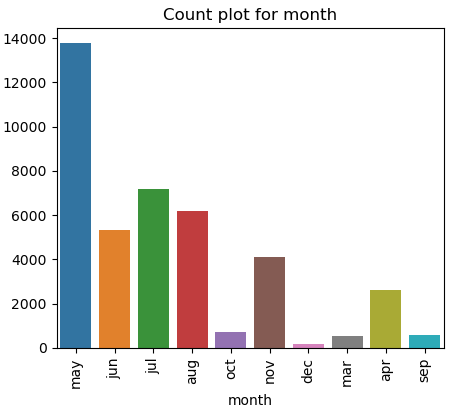
Observation: There were slightly more people with housing loan than people without.



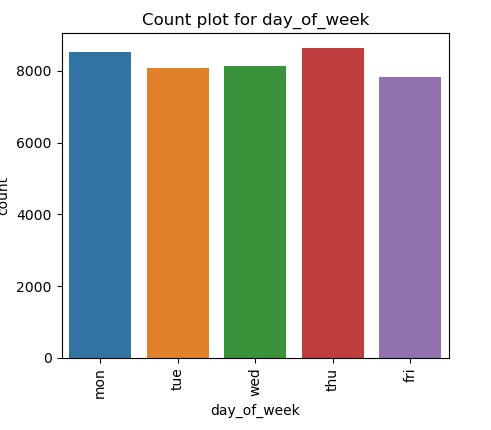
Observation: The number of people who had no loan were about ten times the people who had a loan, meaning more people without a loan took part in this survey than people with a loan.



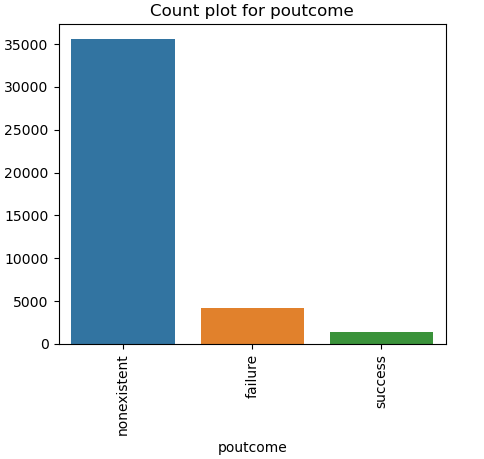
Observation: There was more contact made by cellular than telephone where for every three people contacted by telephone five more were contacted by cellular.



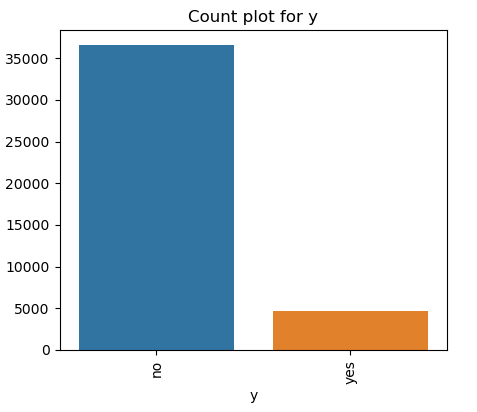
Observation: The month of May saw more last contacts been made than any month of the year according to the survey, followed by June and August with December recording the least number of contacts been made.



Observation: There were slightly equal and active number of contacts made for the days of the week.



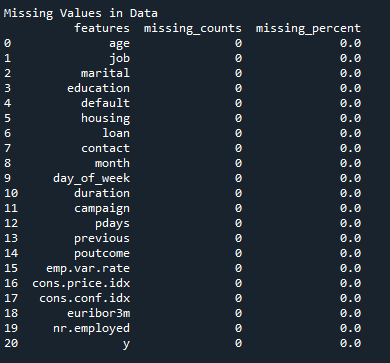
Observation: There were more failures than success for the outcome of the previous campaign recorded with a high number of non-existent records with no activity.



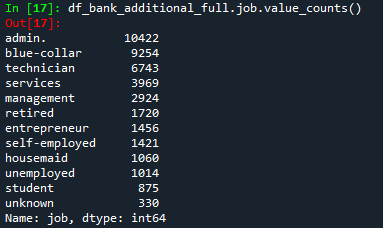
Observation: The number of clients that subscribed for a term deposit were ten times more than clients that did not subscribe for a term deposit creating yet again an imbalance class problem.

# Checking the quality of the data

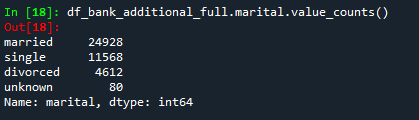
# Null/ Missing Values:



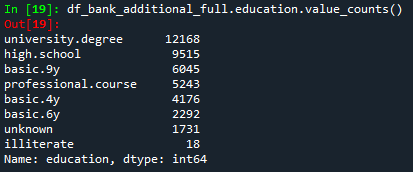
There are practically no missing values present in the data set, however there more unknown records present in the data set and this could be seen in the below for the different variables in the data set.



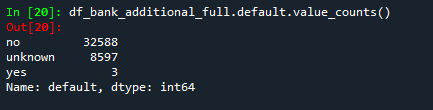
There are 330 unknown records for the job variable.



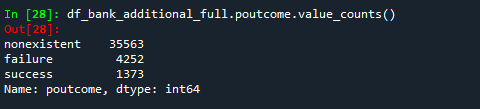
There are 80 unknown records in the marital variable.



There are 1731 unknown records for the education variable.



There are 8597 unknown records for the default variable.



There are 35563 non-existent unknown records for the poutcome variable.

1. **Duplicate Values:**



Twelve duplicate values were found in the dataset. But on further exploration, it was found that the duplicated records do have unique values in a few features, which means they are not entire duplicates. Hence, these records will not be removed.

# EDA of Categorical Features (Harold Wilson and Yash Jadwani)

# Data cleaning of categorical features:

# Introduction:

# Machine learning (ML) projects typically start with a comprehensive exploration of the provided datasets. It is critical that ML practitioners gain a deep understanding of:

# The properties of the data: schema, statistical properties

# The quality of the data: missing values, inconsistent data types

# The predictive power of the data: for example, the correlation of features with the target

# 2.1 Handling missing data for Bank marketing dataset

# Data cleaning is an important step in data pre-processing due to its ability to help improve the quality of the dataset for a more reliable output. The presence of impurities in real-world data application has brought about the development of several methods to eradicate this problem to help improve the accuracy and usability of existing data (Müller and Freytag, 2005). [3] The data cleaning process involves the detection or removal of outliers, smoothing noisy data, filling in missing values and resolving inconsistency within a dataset (Han, Pei and Kamber, 2011). [4]

# There is exactly no one way of dealing with missing data. There are different solutions for data imputation depending on the kind of problem and it always difficult to provide a general solution, and care should be taken when it comes to removing missing values in any given data set since doing so will introduce biasness in the model.

# Imputing or Deleting missing values of the Data:

# Before we decide to remove, replace or impute the data, we have to understand and establish the reason why data is missing.

# Missing at Random: This means that the tendency for a data point to be missing is not related to the missing data, but it is related to some of the observed data.

# Missing Completely at Random: The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.

# Missing not at Random: Possible reasons are that the missing value depends on the hypothetical value or missing value is dependent on some other variable’s value.

# 2.2 Data cleaning of categorical features in the data set:

# We have the following unknown values for some of the features in the data set:

# 

# Mode Imputation for Unknown/Missing values

# a) Deleting Duplicate values

# 

# b) Replacing Unknown/missing values with N/A and Checking for null values

# 

# c) Mode Imputation Steps

# We removed all the rows where job, default and housing are null.

# We removed all the rows where job and education are null.

# We encountered 75 missing values in the marital status variable and addressed this issue by creating an age marital map to impute these values.

# In the job variable, we found 197 instances of missing data, which we resolved by generating an age education job map and using it to impute the missing values.

# We found approximately 1600 instances of missing education records, which we addressed by generating a job education map and using it to impute the missing values.

# We removed all the rows where loan, default and housing are null.

# There was highly imbalance in default feature, only 3 individuals had defaulted. This suggests that the vast majority of individuals in the dataset have not defaulted on their payments. Therefore, we removed the default feature from the analysis as it may not be useful in making any meaningful conclusions

# We identified 763 missing values in both the loan and housing variables, which we imputed using information from the marital status, job, and education variables

# 2.3 Visualizing distributions for each category of target variable:

# 

# 

# 

# 

# 

# 

# 

# Observations 1:

* Age: Most of the calls were made to people aged 25-50. Percentage of subscriptions seems to be approximately constant across all ages.
* Duration: As expected, percentage of subscriptions increases with the increase in call duration.
* Campaign: There seems to be almost no subscriptions for more than 4 contacts in the current campaign
* Previous: Data is heavily skewed to number of contacts = 1. Percentage of conversion seems to be consistent with overall value.
* cons.price.idx, cons.conf.index and PC1: There seems to be some correlation with target variable. But the trend is not clear, this may be clearer when coupled with month and year values.

# 

# 

# 

# Observations 2:

# job: 'admin','blue-collar' and 'technician' jobs were contacted most. 'Retired' and 'Student' categories gave the highest percentage of subscriptions

# marital: most of the people contacted were married. The percentage of subscriptions didn't seem to change much with marital status

# education: most of the people contacted had either 'university. degree' or 'high. school' as their highest level of education. Though, 'illiterate' customers gave the highest percentage of subscriptions

# housing: There is no imbalance observed with respect to housing. The percentage of subscription also seems to be constant.

# loan: Most of the people contacted didn't have a personal loan. People without personal loan did seem to be more likely to subscribe but the difference between the two categories is small.

# contact: most of the people were contacted through a cellphone. This did result in a significantly higher percentage of subscriptions.

# month: Most of the contacts were made in the second quarter. Some months gave a significantly higher percentage of subscriptions than other months, but the trend is not very clear and there may be other factors at play here.

# day\_of\_week: Number of contacts and percentage of subscriptions doesn't seem to change much with day of the week.

# poutcome: The outcome of previous campaigns was "nonexistent" for most of the contacts. Although, the success of previous campaigns did seem to positively impact the subscriptions of current campaign.

# 3.0 EDA of Numerical Features (Yash Doshi and Anuj Singh)

# Data cleaning of numerical features:

# As discussed in the previous week’s report, going forward, we will be using the bank-additionalfull.csv for this project.

# Basic data exploration revealed that the dataset has 10 numerical features. In the data exploration performed in the last week, it was found that some of these features have outliers and some have high correlation amongst themselves. These issues need to be dealt with at this stage because these issues can significantly affect the accuracy and performance of the machine learning model that will be created later.

# Outlier detection:

# The following table shows the outliers for numerical features and rare categories for categorical features of the dataset. It also shows how we intend to deal with them.

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Outliers/ Rare Categories** | **How to deal with them?** |
| age | has 468 outliers greater than upper bound (69.5) or lower than lower bound(9.5). | Convert into categorical feature by binning the values. |
| job | 7 rare categories: ['retired', 'entrepreneur', 'self-employed', 'housemaid', 'unemployed', 'student', 'unknown']. | These can be grouped into a single category. |
| marital | 1 rare categories: ['unknown']. | This will not be changed/ transformed. |
| education | 2 rare categories: ['unknown', 'illiterate']. | This will not be changed/ transformed. |
| default | 1 rare categories: ['yes']. | This will not be changed/ transformed. |
| housing | 1 rare categories: ['unknown']. | This will not be changed/ transformed. |
| loan | 1 rare categories: ['unknown']. | This will not be changed/ transformed. |
| contact | No issue |  |
| month | 4 rare categories: ['oct', 'sep', 'mar', 'dec']. | This will not be changed/ transformed. |
| day\_of\_week | No issue |  |
| duration | has 2963 outliers greater than upper bound (644.5) or lower than lower bound(-223.5). | Convert into categorical feature by binning the values. |
| campaign | has 2406 outliers greater than upper bound (6.0) or lower than lower bound(-2.0). | Records with values greater than 15 will be removed as outliers. |
| pdays | has 1515 outliers greater than upper bound (999.0) or lower than lower bound(999.0) | Since 999 is just a placeholder, it will be replaced by -1. |
| previous | has 5625 outliers greater than upper bound (0.0) or lower than lower bound(0.0). | This will not be changed/ transformed. |
| poutcome | 1 rare categories: ['success']. | This will not be changed/ transformed. |
| emp.var.rate | No issue |  |
| cons.price.idx | No issue |  |
| cons.conf.idx | has 446 outliers greater than upper bound (-26.949999999999992) or lower than lower bound(-52.150000000000006). | This will not be changed/ transformed. |
| euribor3m | No issue |  |
| nr.employed | No issue |  |
| y | has 4639 outliers greater than upper bound (0.0) or lower than lower bound(0.0). Cap them or remove them. | Sampling methods will be used to deal with this imbalance |

# 3.1 Dealing with outliers (Anuj Singh)

# Checking the boxplots and the distribution plots of the numerical features revealed that three numerical features had a large number of outliers, namely, ‘duration’, ‘pdays’, and ‘campaign’. The outliers in the ‘duration’ and ‘campaign’ column were removed by only keeping the values less than the 95th percentile value for these features.

# 

# The boxplots below reveal that majority of the outliers were successfully removed

# 

# For the ‘pdays’ column, it was found that the outliers were due to 999 values, which were placeholders for missing data. Since more than 95% of the values are 999, we will drop the ‘pdays’ column entirely.

# 

# 3.2 Dealing with highly correlated features (Yash Doshi)

# The heatmap of the Pearson’s correlation matrix of the dataset shows that ‘euribor3m’, ‘emp.var.rate’, and ‘nr.employed’ are highly correlated. Note that the below figure also includes the target variable ‘y’, which was a categorical feature in the original dataset but was converted to a numerical feature using Label Encoding to study its correlation with the other numerical features.

# 

# We will be performing Principle Component Analysis (PCA) on these highly correlated features to reduce them into one or two transformed features. The below Scree plot shows that one principle component explains more that 90% of the variance in the data and hence, the three highly correlated features can be transformed into one feature without losing the information of the three features.

# 

# 

# After transforming the three highly correlated features into one feature using PCA, the below image shows the heatmap of the Pearson’s Correlation matrix of the updated dataframe. It is evident from the similar correlation value between the new feature PC1 and y that the correlation problem has been solved without losing the information contained in the features.

# 

# Target vector share before and after Imputation

# 

# 3.3 Individual feature analysis:

# Age:

# 

# 

# 

# Duration:

# 

# 

# Month:

# 

# March had the most conversions for the bank's term deposit product despite only around 500 calls. The success may have been influenced by factors like interest rates, promotions, and sales representatives.

# Campaign:

# 

# 

# 

# 3.4 Hypothesis testing:

# Testing correlation between month and economical indicators:

Relation between cons.conf.idx and month:

Chi-square statistic: 333819.0

P-value: 0.0

---------------------------------------------------------------------------------------

The p-value is below the threshold of 0.02. There is significant difference between month and cons.conf.idx.

===============================================================================

Relation between cons.price.idx and month:

Chi-square statistic: 333818.99999999994

P-value: 0.0

---------------------------------------------------------------------------------------

The p-value is below the threshold of 0.02. There is significant difference between month and cons.price.idx.

===============================================================================

Relation between PC1 and month:

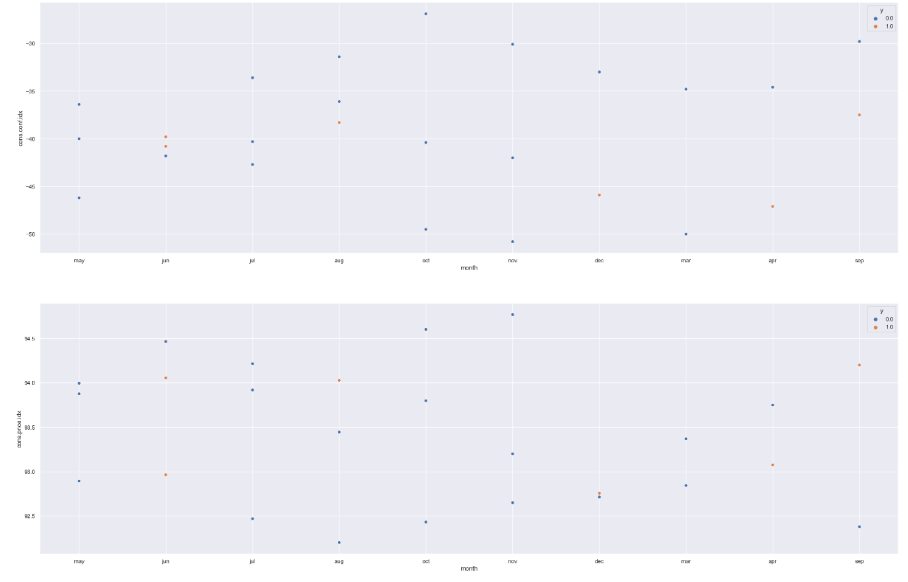
Chi-square statistic: 155791.85127762248

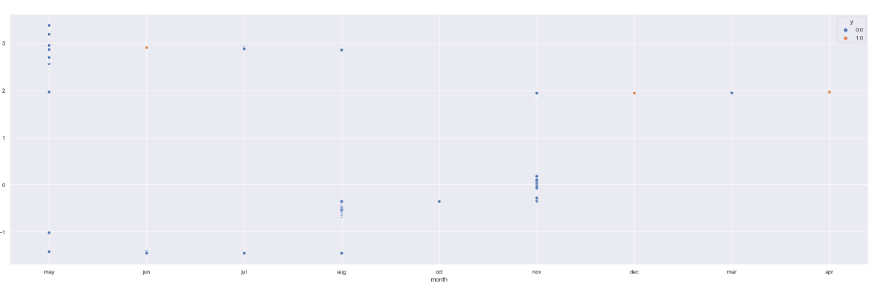
P-value: 0.0

---------------------------------------------------------------------------------------

The p-value is below the threshold of 0.02. There is significant difference between month and PC1.

===============================================================================





**References:**

1. [Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.
2. [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

# GitHub Repo Link:

# <https://github.com/singhanuj695/Data-glacier-Group-Project>