Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

Bank Marketing Campaign

15th Aug 2022

Agenda

Background

Problem Statement

Approach

Data Exploration

Descriptive Analysis

EDA

Feature selection

Quantitative Analysis

Data Glacier

Your Deep Learning Partner

Background : Bank Management Campaign

- ABC Bank wants to sell its term deposit product to customers and before launching the product they want to know that based on the customer's past interaction with the bank or other Financial Institution, how successful their product will be.

- Objective: By converting this problem into a machine learning classification problem we will build a model to predict whether a client will subscribe a term deposit or not so that the banks can arrange a better management of available resources by focusing on the potential customers “predicted” by the classifier .
  Technique to be used: Classification

- Analysis has been divided into 4 parts:
  - Data Understanding
  - Forecasting profit and customer engagement
  - Finding whether the product is profitable or not
  - Recommendations and conclusion of the analysis.

Data Exploration

- Dataset has been checked and reviewed by all the peers and concluded that there were no missing values found in the dataset however we found some “unknown” and “other” values which needs to be converted to numerical values or needs to be removed in order to clean the dataset. No “skewness” in dataset was found, data seems symmetrical dataset. Statistics summery such as slandered deviation, distribution and skewness has been checked.

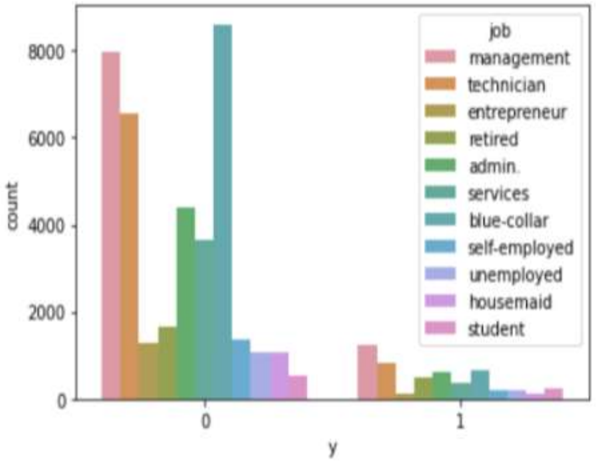
- Exploratory Data Analysis refers to the bank dataset provides critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

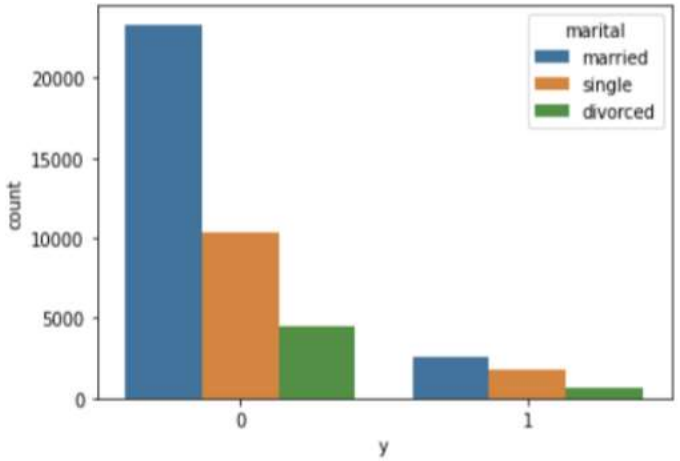
Descriptive Analysis(Univariate Analysis)

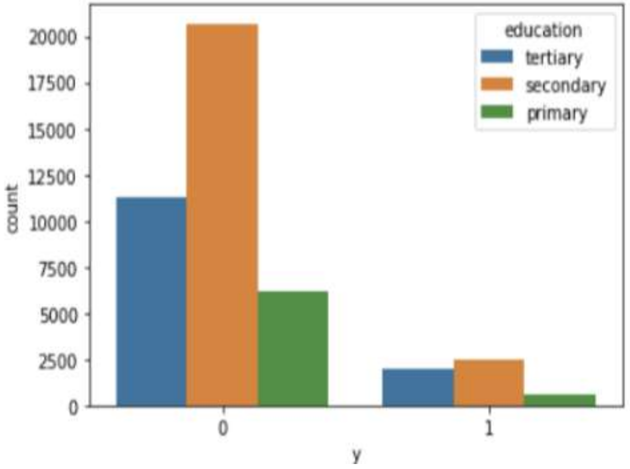
- While performing this analysis we provide an understanding of the characteristics of each attribute of the dataset which is important evidence for feature selection.
- There were no missing values but found some “unknown” values, we decided to drop the outliers and ambiguous values, such as “others” and “unknown”.
- The columns which have two values(‘yes’ and ‘no’) and are slightly imbalanced such as default, loan, and y, have been converted to (1,0) numerical values. The rest are continuous variables that were binned so that outliers values are converted into count values.
- Skewness doesn’t provide many insights into data, as the values of columns are nearly zero apart from ‘previous’. Data seems symmetrical.
- Flooring and clapping using interquartile range(IQR) Outliers are removed by dropping values that are below 25% and 75% percentile.
- We classified the dataset into numerical and categorical attributes.

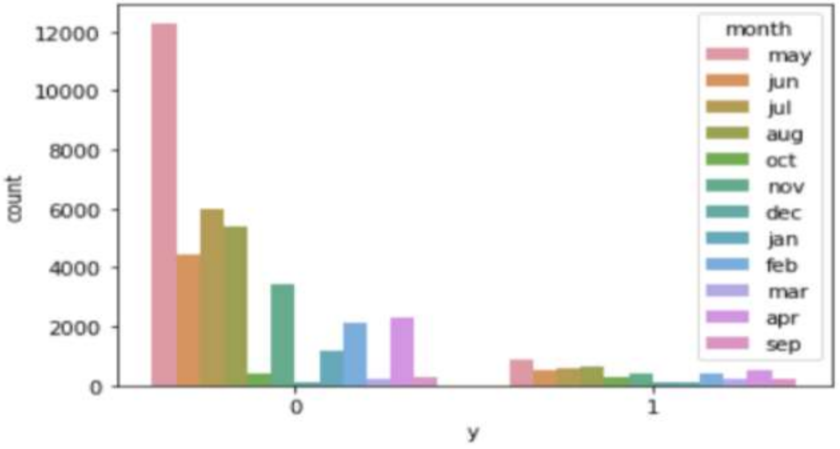
- Numerical Attributes: The following table provides statistical information in descriptive analysis.

Categorical Attribute Analysis





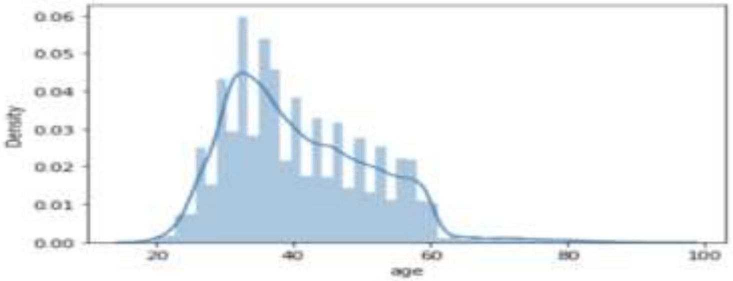


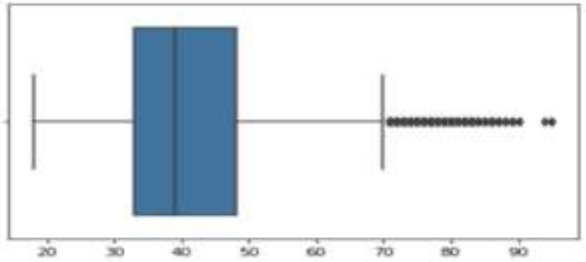


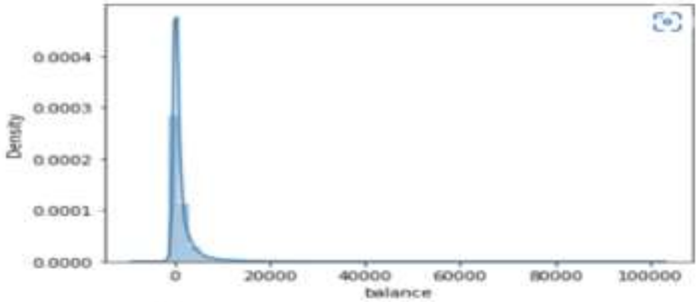
Job: Highest Number of applications are from blue collar type
Of job.
Marital: most of the clients approached were married.
Education: Clients with a university degree and high school
Were approached more as compared to others and they have a higher success rate compared to others. Default: it doesn't
Show much impact.
Month: Around 33% were approached in may and in January, and February we don't have data or no one was approached. The success rate was almost the same in June, July, and August.

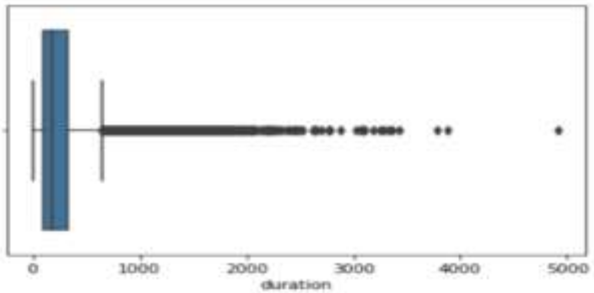
Data Glacier

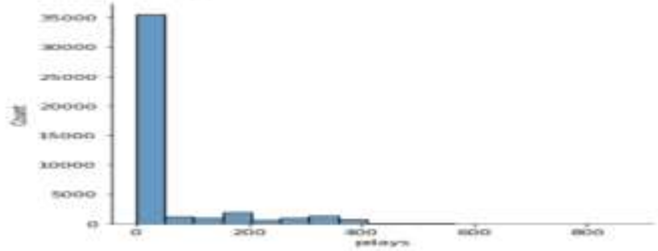
Numerical Attribute Analysis

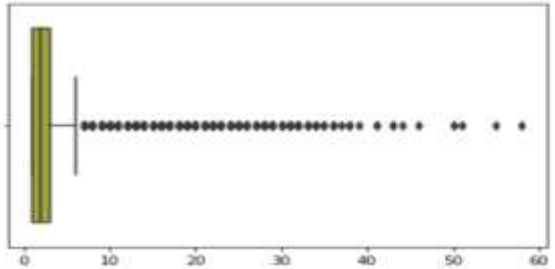






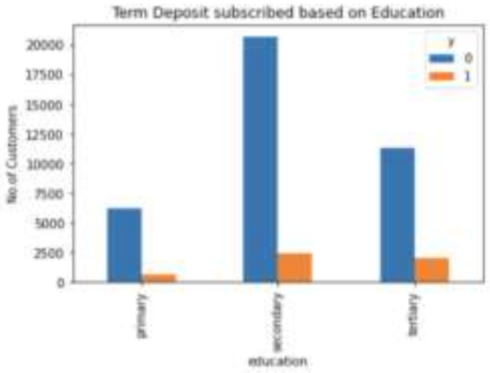


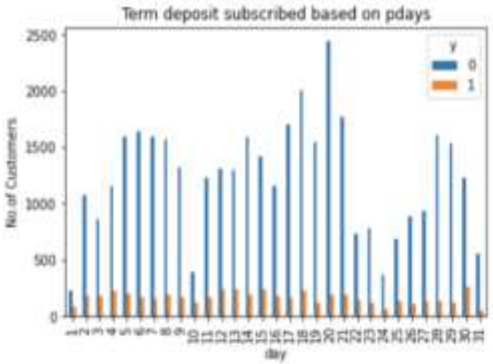


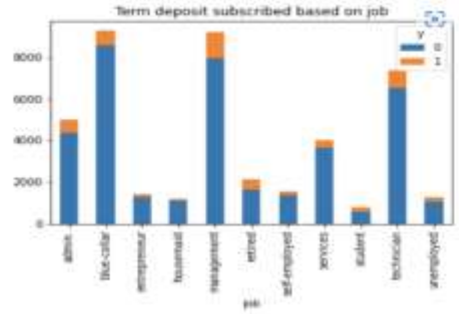


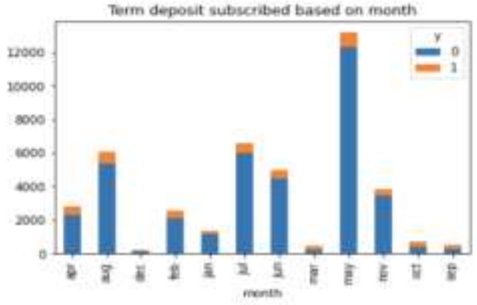
Housing: A housing loan does not have much effect on the number of term deposits purchased.
Loan: most clients with not have personal loans were approached most.
Contact: Around 64% of calls are from cellular.
Day_of_week: We have 5 days of collected values. There is no significant difference in the number of clients approached and the number of people subscribed. So we will drop this feature.
Poutcome: If a client took the term deposit last time then
There is a higher chance of that client subscribing to it again.

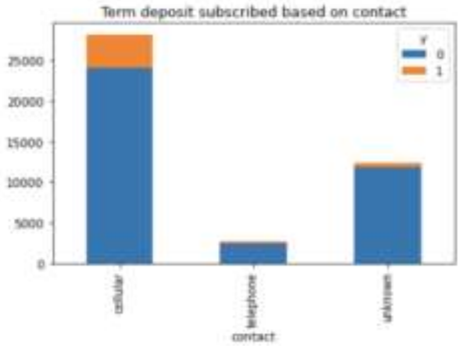
Term deposit Analysis

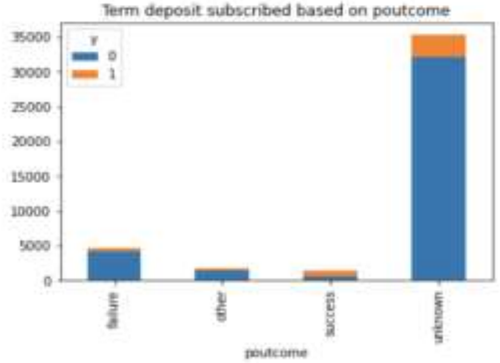


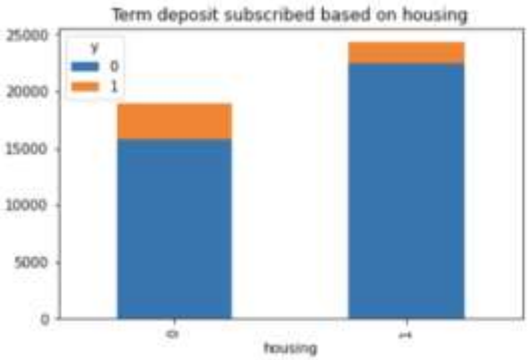


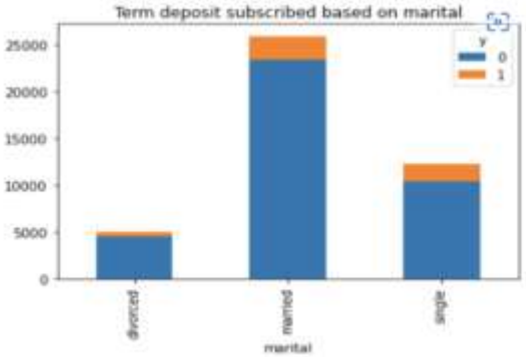


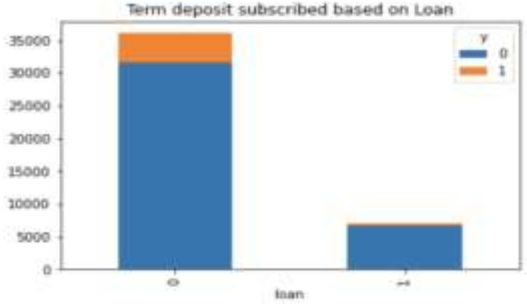


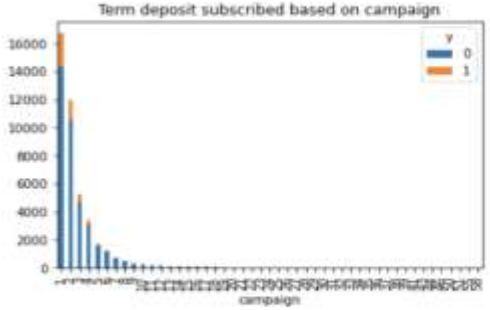




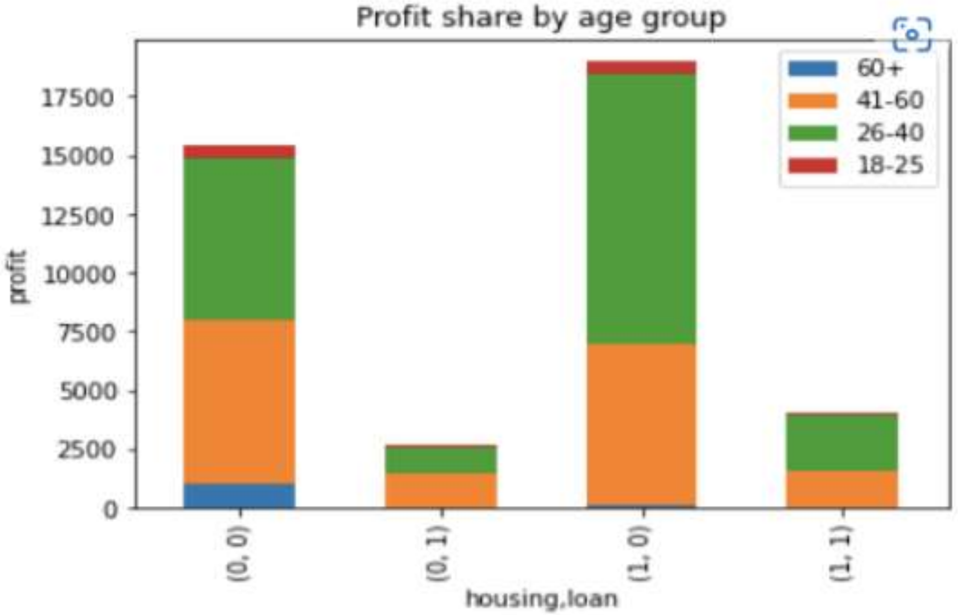


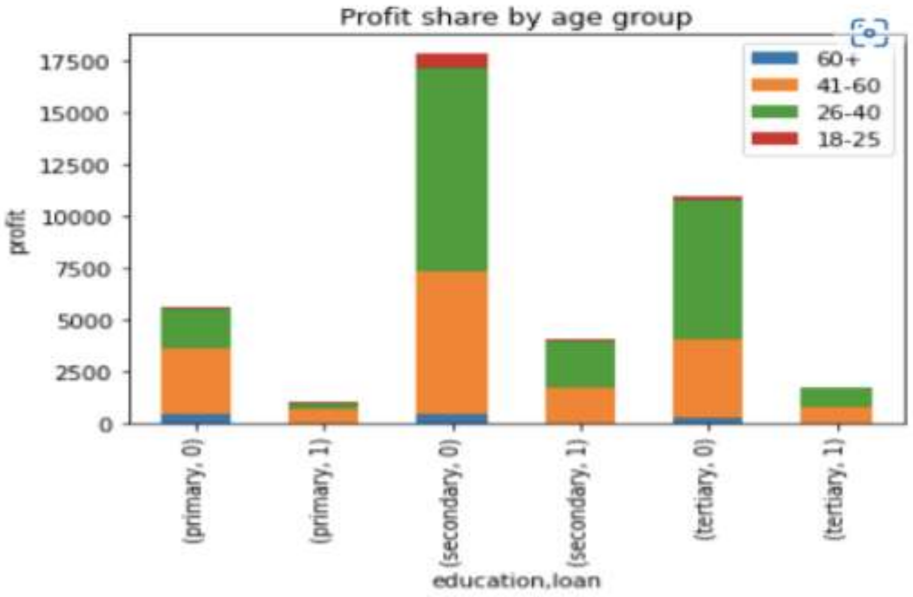






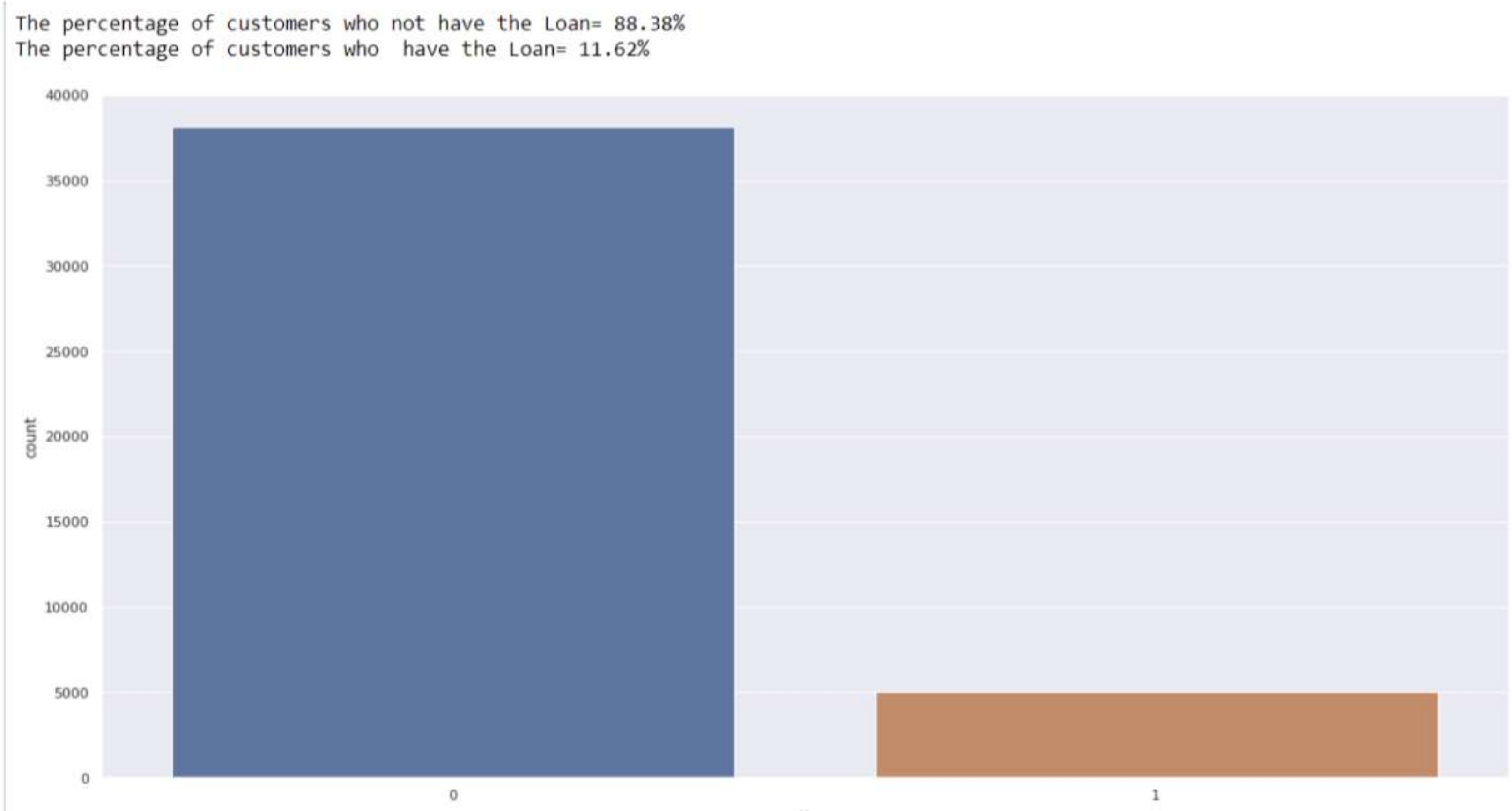
Profit Share Analysis





From the bar charts we can conclude that the age range 26-40 has taken more loan than other age groups.

Feature selection based on Descriptive Analysis

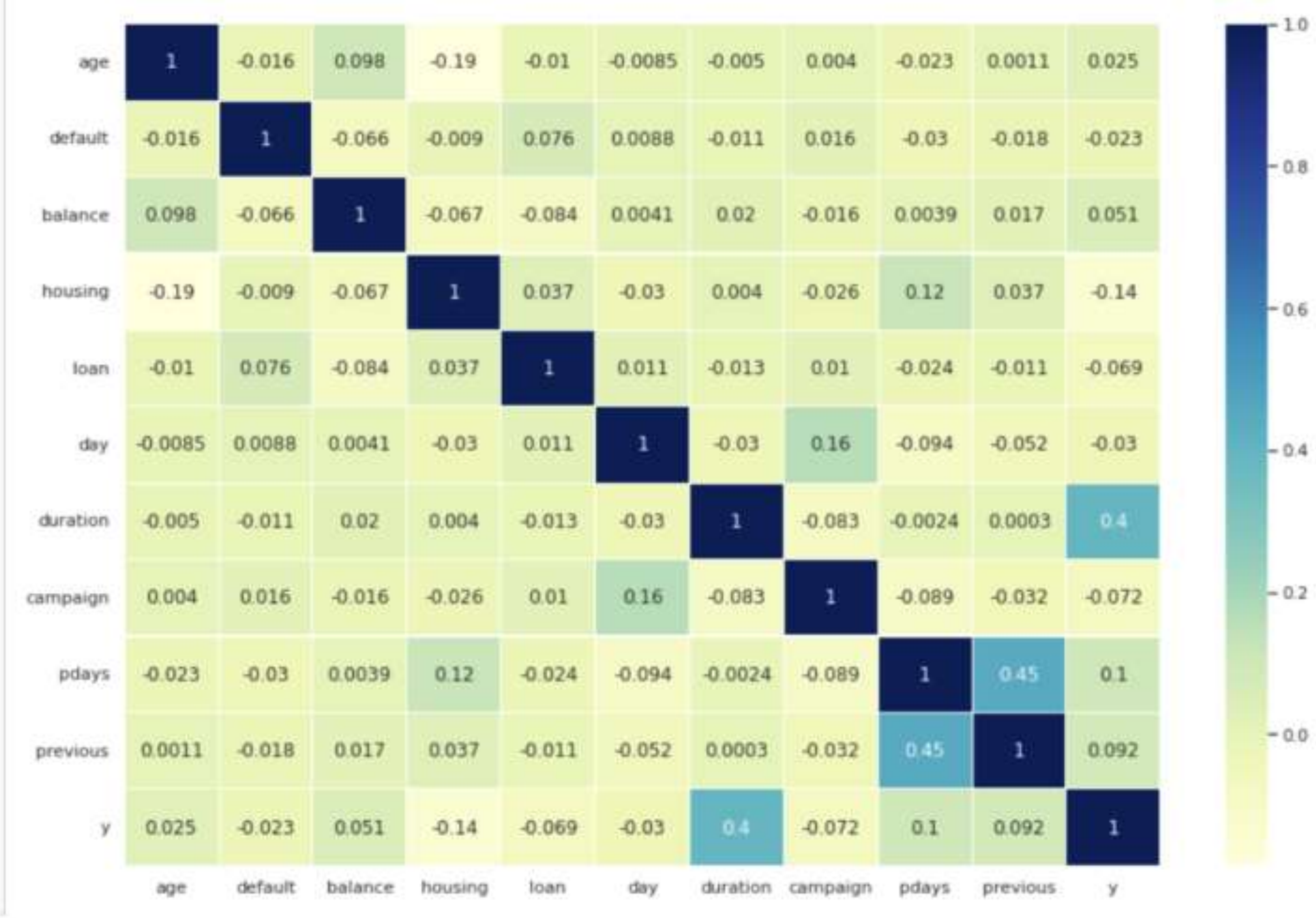


```
The percentage of customers who not have the Loan= 88.38%
The percentage of customers who  have the Loan= 11.62%
```

We have provided basic statistics of each attribute in the dataset, based on this some of the problems we have identified such as imbalance of categorical target. In this dataset, there are a higher number of instances of the major class and fewer number of instances of the minor class as shown in the bar chart below.

So we decided to use the under-sampling method to delete the samples in the majority class as there is a huge difference between y=1 and y=0, So the ML algorithm will omit the smaller value, which may affect the performance of the algorithm.

Feature selection based on Correlation Analysis (Bivariate Analysis)



After performing the under-sampling method and removal of highly positively correlated features we got the features useful in terms of predicting the desired target.

Quantitative Analysis

We also tried to perform the chi-squared test, The comparison is deemed statistically significant if the relationship between the categorical attribute “marital” where we got to know that P- value of 1 depicts that there is no difference in value of the groups other than due to chance.

Whereas in Z-test, we found Z-stat is less than Z-critical we accept the null hypothesis test.

Dedicated to technical user :

- Imbalanced dataset is a common problem in data science; however some approaches have been used on the dataset such as over and undersampling methods as well as boosting algorithm (for traditional machine learning approach) like adaboost, so we would like to use Adaboost in the model.
- For deep learning and imbalanced datasets, the two main topics that comes to mind is the loss function and the use of dropout layers. The loss function plays a crucial role in deep learning models, and selecting the right one for imbalanced datasets is important, therefore experimenting with different loss functions like weighted cross entropy or the Generalized Dice overlap loss functions. The dropout layer helps to stop overfitting which can help with reducing bias especially if used injunction with oversampling minority classes.
- We can also use booting trees like adaboost, xgboost and random forests that are more robust to imbalanced datasets.
- Also a popular method for dealing with imbalanced datasets for machine learning models and deep learning models within the preprocessing phase is data augmentation.
- We would also recommend logistic Regression model as it can provide better accuracy after providing model pipeline like min_max normalization or dimensionality reduction using PCA.
- Heterogeneous ensembling is also good option which combines several base model to produce final optimum solution.

Thank You