CASE STUDY REPORT

Team : Data_degea

College : NIT Rourkela

Ashis Kumar Parida Alyal Samal

Code File That Describes Our Approach

UNDERSTANDING THE PROBLEM STATEMENT

'' A firm has launched a new investment product “Term Deposit” which has observed rapid growth since its inception. The increase in term deposit accounts has provided the firm with additional funds to create other offerings.
Hence, based on this performance, it further wants to increase its term deposit accounts amongst existing customers. To do so, the firm plans to run a campaign and wants EXL to identify a list of customers who could be contacted via telecommunication channels to open a term deposit account with ABC.
Since the campaign will be run for a short period and owing to budget constraints, the firm wants a list of only 1000 customers from EXL for the target list. ''

We are provided with three Datasets, i.e. **Data Dictionary**, **Historical data** and **New customer list data.**

We treated '**Historical Data**' as **Train Data** and '**New Customer List Data**' as **Test Data**. And For Information on columns, we used the '**Data Dictionary**' Data.

As the problem statement wants us to predict the top 1000 customers who would open a Term deposit account, we started with predicting probabilities for each customer and then arranging them by increasing probabilities. At last, we selected the top 1000 customers with the highest probabilities.

One more observation was that, this Dataset is **imbalanced** concerning the target column which is interpretable from **Fig-1**

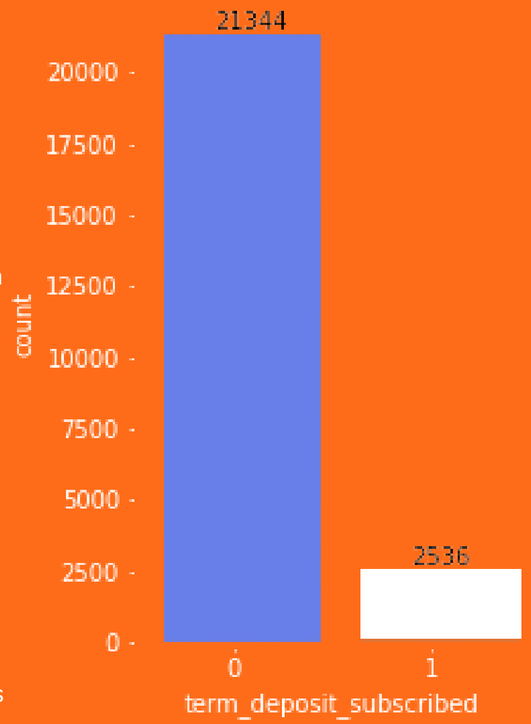


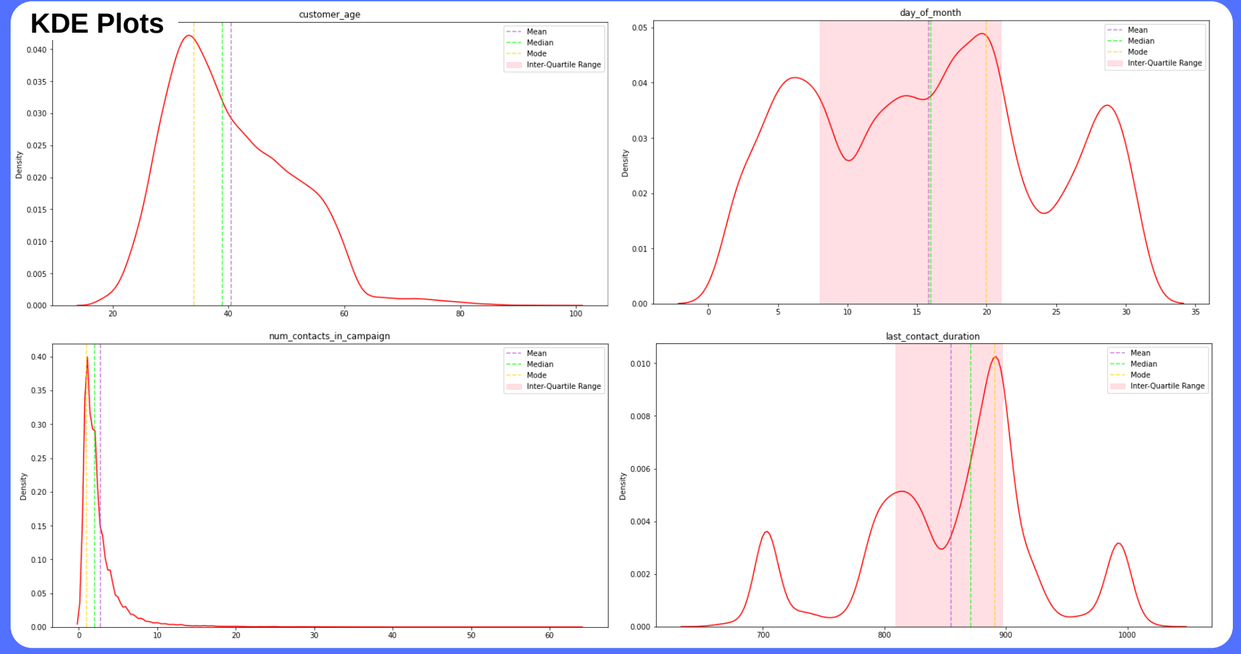
Fig-1

We then decided to **maximize the ROC-AUC** Score, so our focus was to increase the **area under the ROC curve**. Also, as we are predicting **probabilities** and not the class as it is more flexible to predict the probabilities for each class instead. The reason for this is to provide the capability to choose and even calibrate the threshold for how to interpret the predicted probabilities.

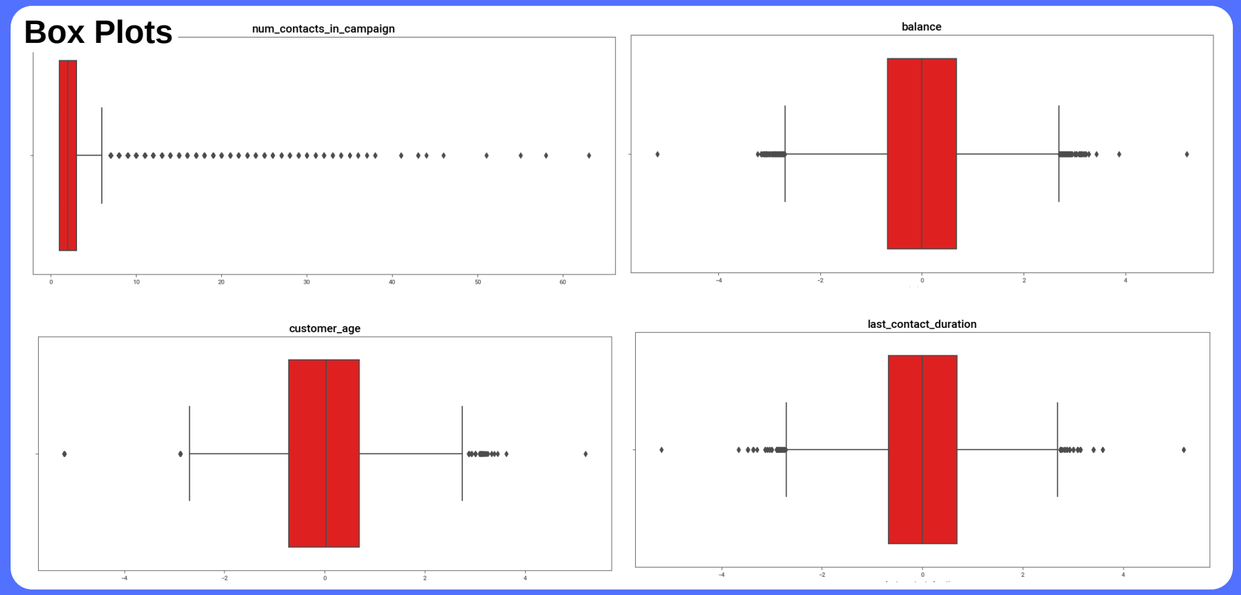
We also did many univariate and multivariate columns of the data, and then applied various models and finalised on blending of best performing models.

SOME COMMON GRAPHS OF DATA

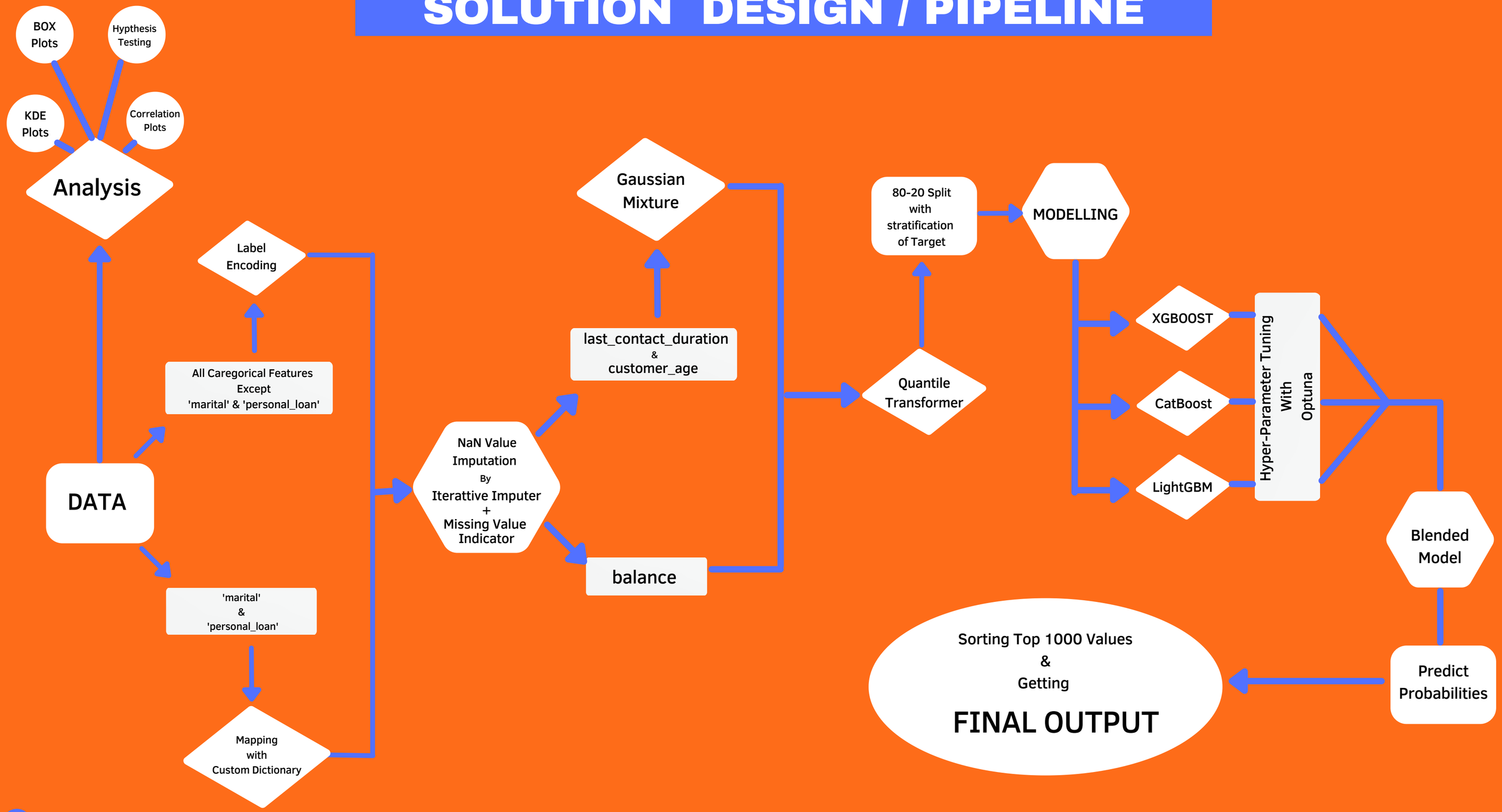
KDE Plots



Box Plots



SOLUTION DESIGN / PIPELINE



DATA PROCESSING AND TREATMENT

I. ENCODING THE CATEGORICAL FEATURES

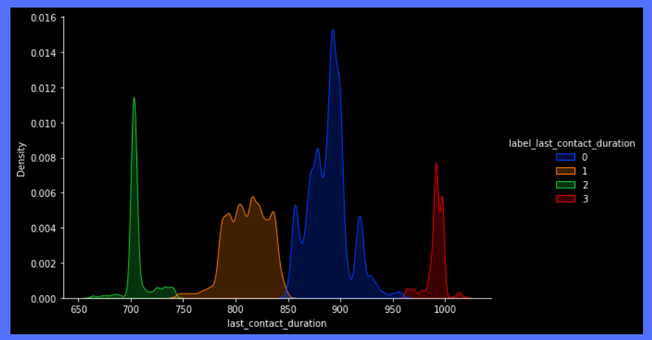
We have label encoded all the categorical features except 'marital' & 'personal_loan' columns. We have mapped the two columns to numerical because these two had NaN values. Then we dealt with them in the same way as we should treat numerical values.

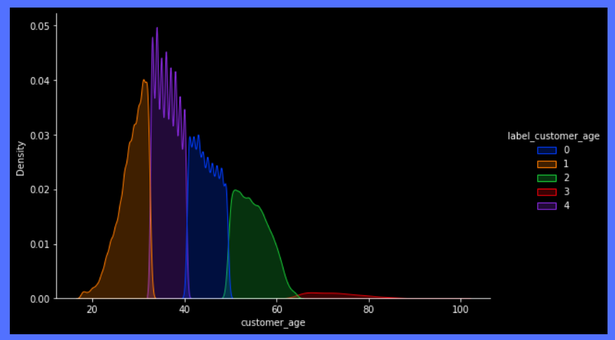
II. IMPUTING THE NAN VALUES

We noticed that few columns of the data had many instances of null values. So, we decided to impute them using iterative imputer with missing values indicator. The logic behind this is that **most of the time the missing values are not randomly distributed across observations but are distributed within one or more sub-samples. Therefore, missingness itself might be a good indicator to classify the labels.**

III. BINNING DISTRIBUTIONS WITH GAUSSIAN MIXTURE

We obseved that the columns named 'last_contact_duration' and 'customer_age' had multimodal disrtribution, so we decided to bin the values before transforming them with different kind of transformers. Here are the results we got:





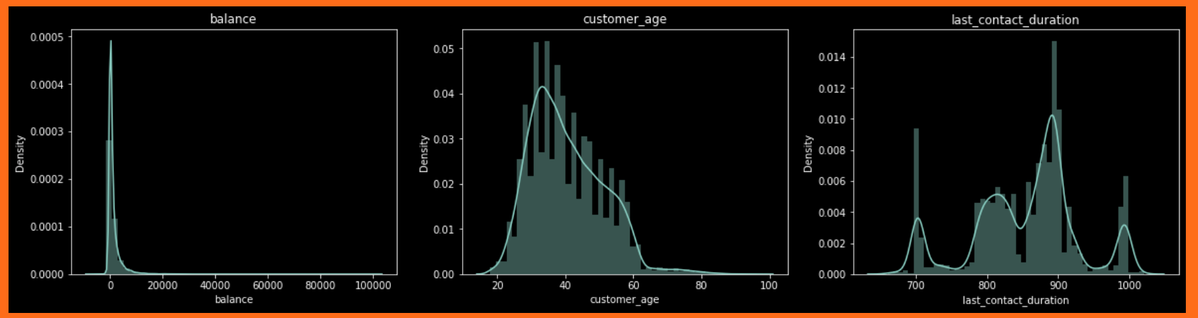
IV. SCALING THE VALUES

We have observed that the features of our data has values of different scale. So, in order to have a uniformity of values in the data, we decided to scale them using Robust Scaler. This will also help us yield better results with distance based algorithms. Robust Scaler also helps us deal with outlier values.

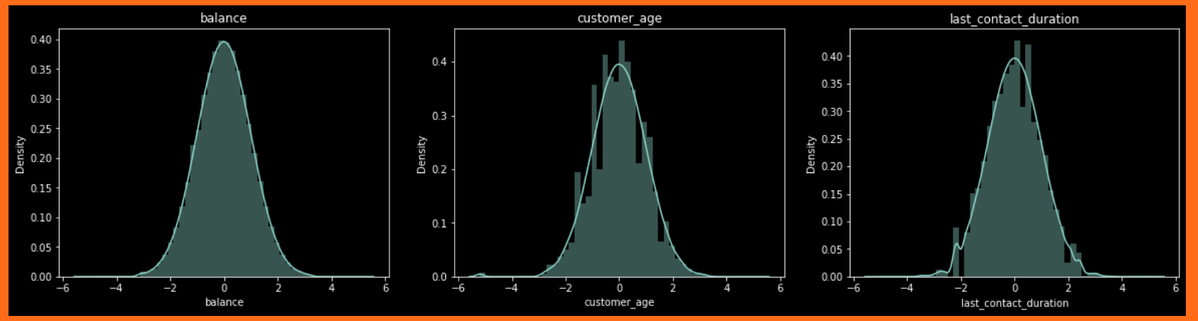
V. TRANSFORMATION OF COLUMNS

Finally we checked the skew-ness of columns and decided that 'last_contact_duration', 'customer_age' and 'balance' columns should be transformed by quantile transformer.

BEFORE TRANSFORMATION



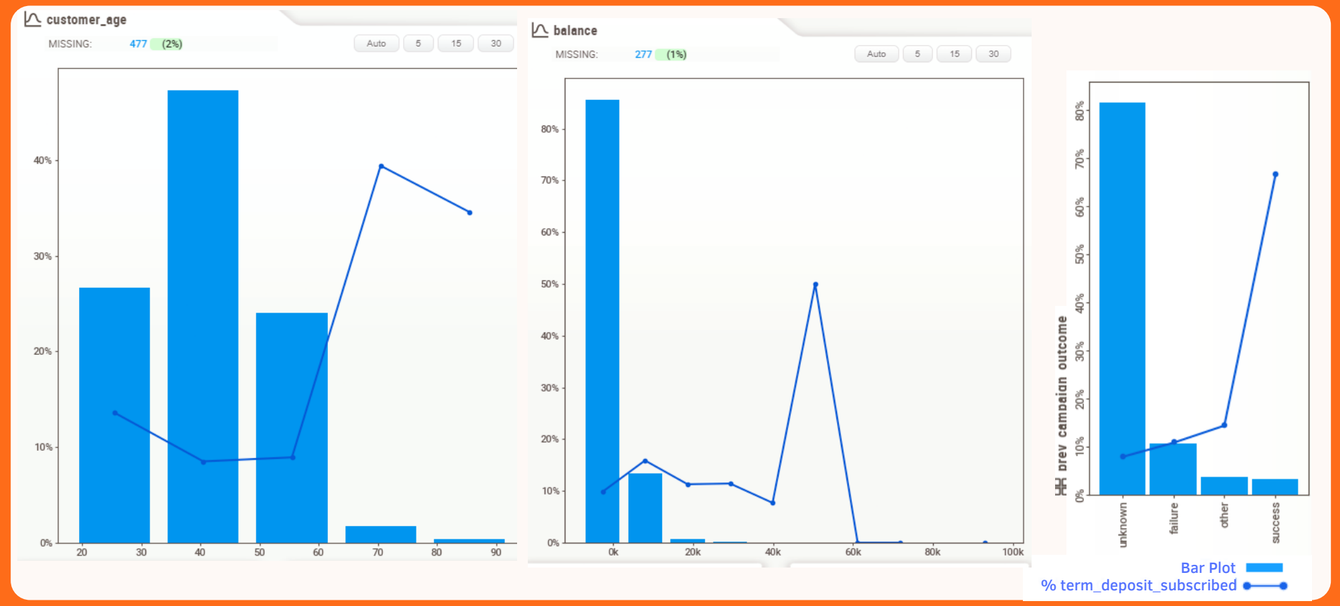
AFTER TRANSFORMATION

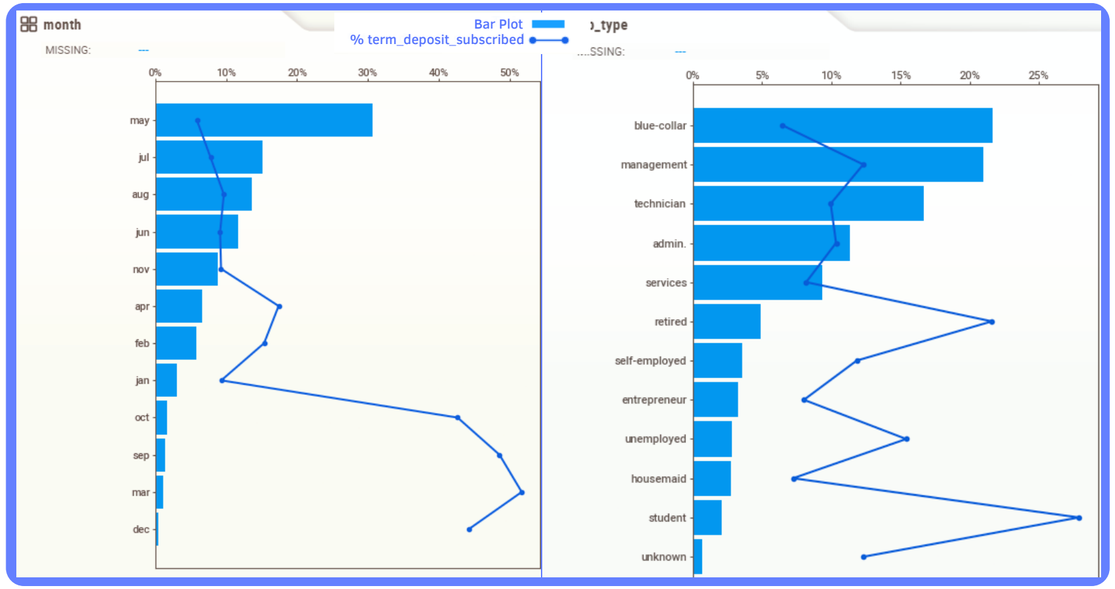


VI. DEALING WITH CLASS IMBALANCE

As we earlier seen that the target value is very much imbalanced, so we decided to use oversampling strategy by using algorithms like SMOTE and ADASYN for algorithms which are not tree based. For tree based algorithms, we used their built-in parameters to make the algorithm treat data as a balanced one.

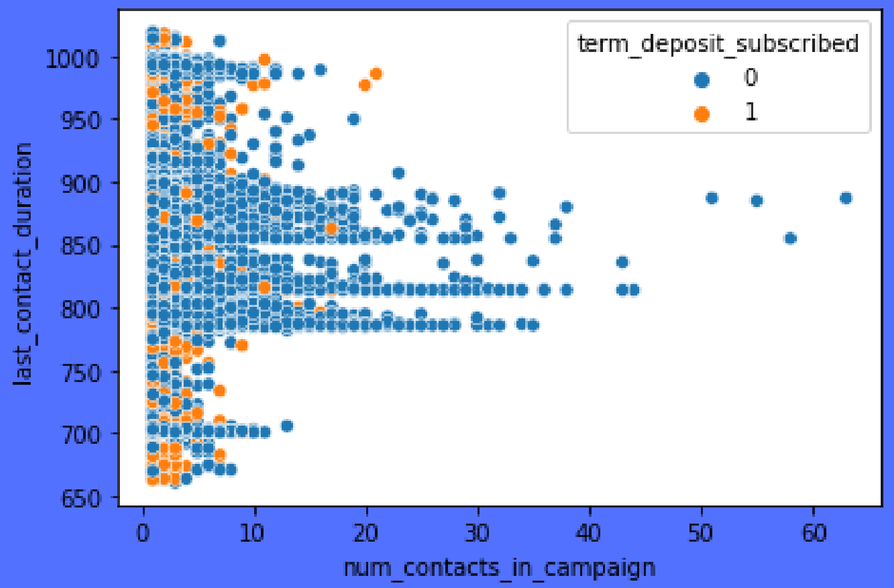
DATA VISUALIZATION





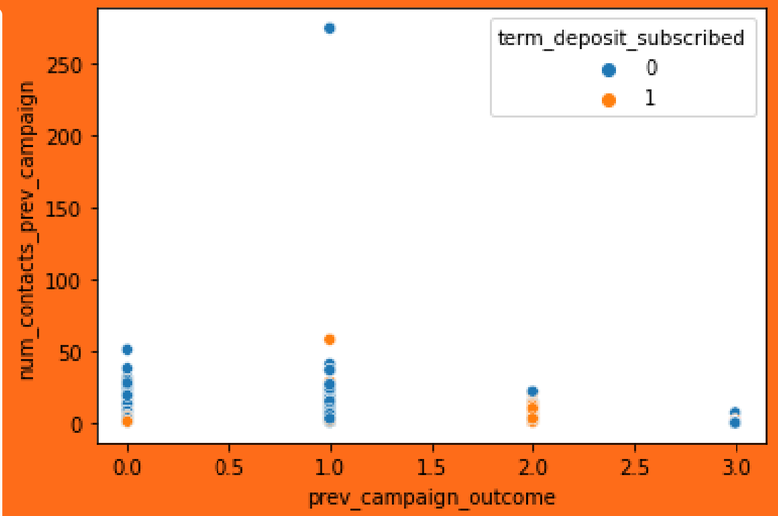
We have created a **scatter plot** drawn between **num_contacts_in_campaign** and **last_contact_duration** as we want to **analyse** how much **effect** does **last contact duration** have on **number of contacts in campaign** with respect to this year's **term deposit lead subscription.**

From this **graph**, we can observe that when the call duration is in **800-900** range, then only there is a **surge in number of contacts in campaign**. We can also observe that when the **last contact duration** is either **very high or very low**, then there are **higher percentage** of people who have **subscribed for term lead.**

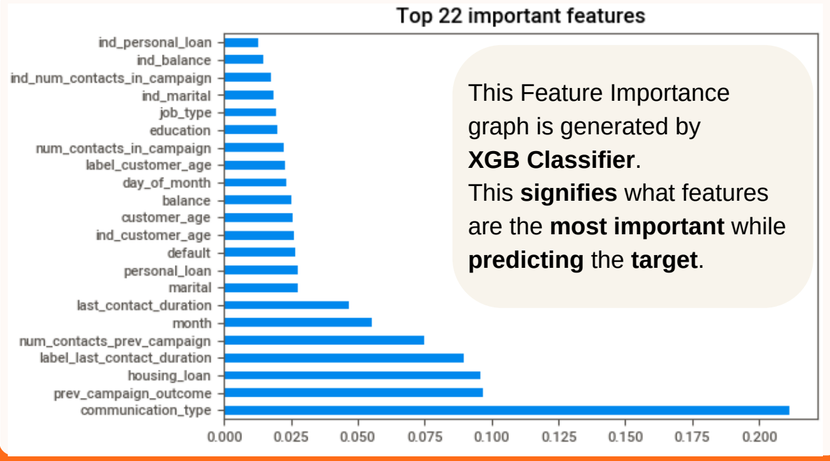


We have created a **scatter plot** drawn between **prev_campaign_outcome** and **num_contacts_prev_campaign** as we wanted to **analyse** how much **effect** did **number of contacts** done in **previous campaign** had on **successful outcomes**, with respect to this year's **term deposit lead subscription.**

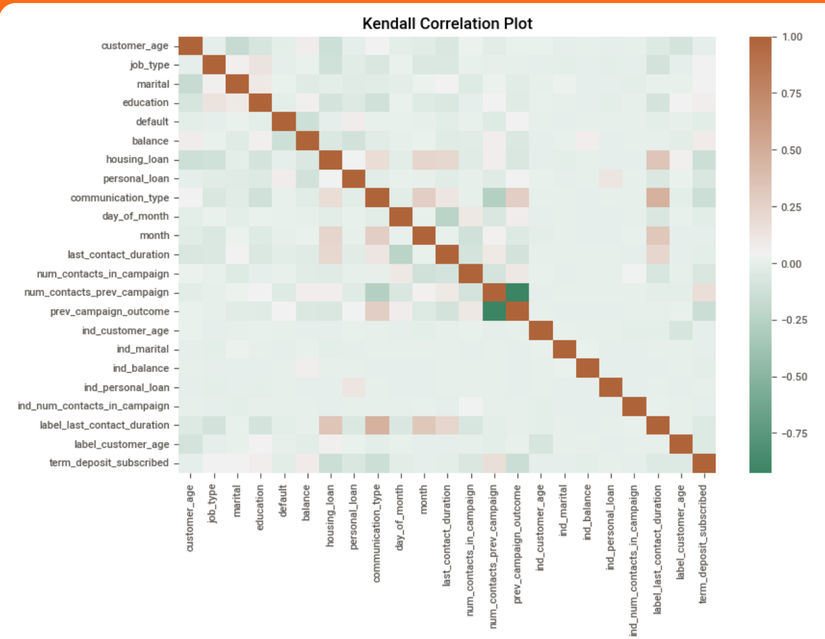
From this **graph**, we can observe that when previous campaign outcome was a **failure** , then also a **higher percentage** of that group has **subscribed** for **Term Deposit**, even if they were contacted **very less** no. of times in the **previous campaign.**



IDENTIFYING THE FACTORS THAT CONTRIBUTED THE MOST IN TERM DEPOSIT ACCOUNT OPENING



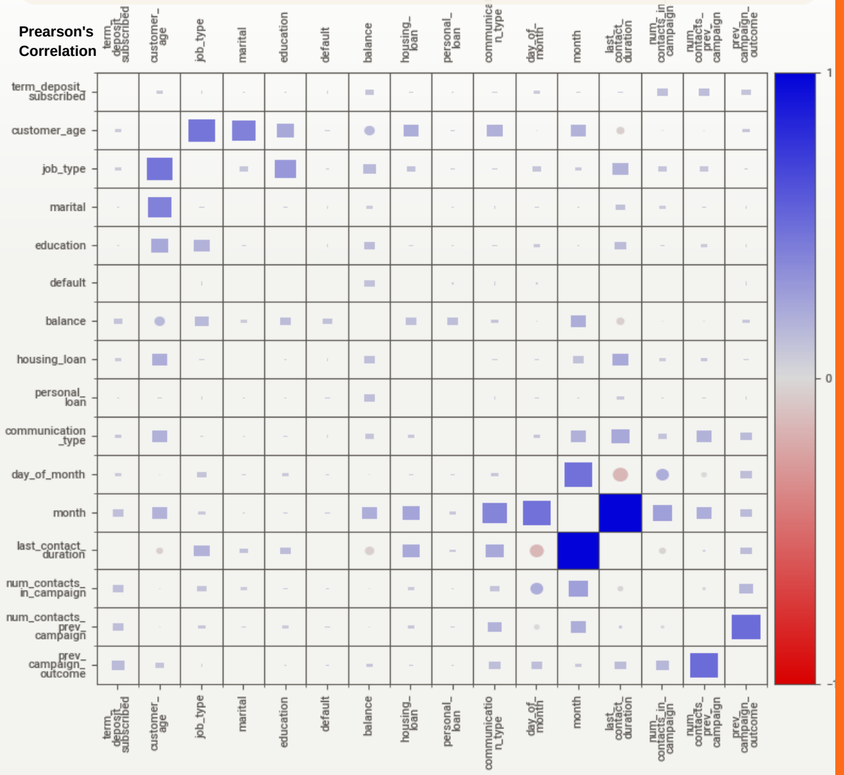
This Feature Importance graph is generated by **XGB Classifier**.
This **signifies** what features are the **most important** while **predicting** the **target**.



We have also checked **Kendall's tau** as **Pearson correlation** can **capture linear relations only** and **Spearman's rho** is **more sensitive to errors** and **discrepancies** in the data. When data is **normal**, Kendall's tau has **smaller gross error sensitivity** and **smaller asymptotic variance.**

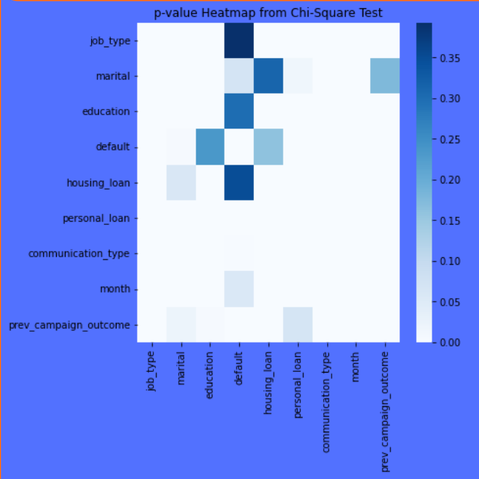
All of **these correlations** will eventually help us what columns are **more important for prediction** and what are their **correlations with the Target Column.**

■ Squares are **categorical associations** (uncertainty coefficient & correlation ratio) from 0 to 1. The uncertainty coefficient is asymmetrical, (i.e. ROW LABEL values indicate how much they PROVIDE INFORMATION to each LABEL at the TOP).
• Circles are the **symmetrical numerical correlations** (Pearson's) from -1 to 1. The trivial diagonal is intentionally left blank for clarity.



***Pearson's correlation coefficient** is very **sensitive** to **outliers**, which can have a **very large effect** on the **model fitting** and the Pearson correlation coefficient. This means including outliers in our analysis can lead to **misleading results.**

CHI-SQUARE TEST GRAPHS



Simply, the **more** these values **diverge** from each other, the **higher** the **chi square score**, the more likely it is to be significant, and the more likely it is we'll **reject** the **null hypothesis** and conclude the variables are associated with each other.

We have done this to get a rough idea of the dependencies.

we have observed that **communication_type** and **month** are **dependent on each other**.
This is something that **needs treatment** according to the model we decide to use.

If any of the column's **dependency** with the **target** column is **greater** than its dependency with another column, then we consider the latter one, because we don't want the **interference** of that variable **while determining** the **target** column value

THANK YOU

We thank **EXL Services** for conducting such an amazing competition where results are not just based on models, but also take into account data analysis, which is crucial for a data scientist. We have thoroughly enjoyed participating in the competition. Please hold more such competitions !!

CONTACT US
ashiskumarparida73@gmail.com
alyalsonu@gmail.com