BUSINESS UNDERSTANDING

Problem Statement: Improve the effectiveness of the bank's telemarketing campaign by understanding its client base.

Problem Motivation: By analysing clients features, the bank will be able to predict customer saving behaviours and identify which type of customers is more likely to make term deposits.

This dataset was collected from each client about their basic information, financial status , the frequency of contact with last campaign and social and economic context. This dataset is based on "Bank Marketing" UCI dataset (please check the description at: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing).The data is enriched by the addition of five new social and economic features/attributes (national wide indicators from a ~10M population country), published by the Banco de Portugal and publicly available at: https://www.bportugal.pt/estatisticasweb.

DATA EXPLORATION

All coding is done in Python 3. pandas , numpy , matplotlib , seaborn, sweetviz and sklearn packages were extensively used for data pre-processing, analysing, cleaning, visualizing and building models. This dataset contains 21 different features on 41188 observation. There have categorical and numerical features. Target variables was binary (‘Yes’or ‘No’). Numerical variables were visualized by matplotlib,seaborn and sweetviz package to analyse its distribution and its relation with the clients’ decision. Sweetviz package also used to check kurtosis and skewness of each features.

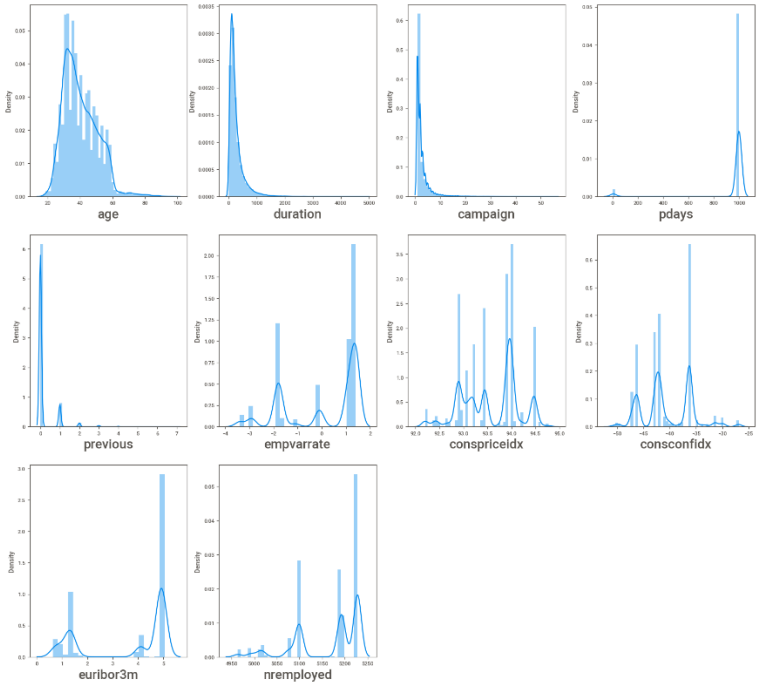


Figure 1: Univariate Analysis of numeric features

Categorical variables were visualized by matplotlib and seaborn package to analyse the distribution

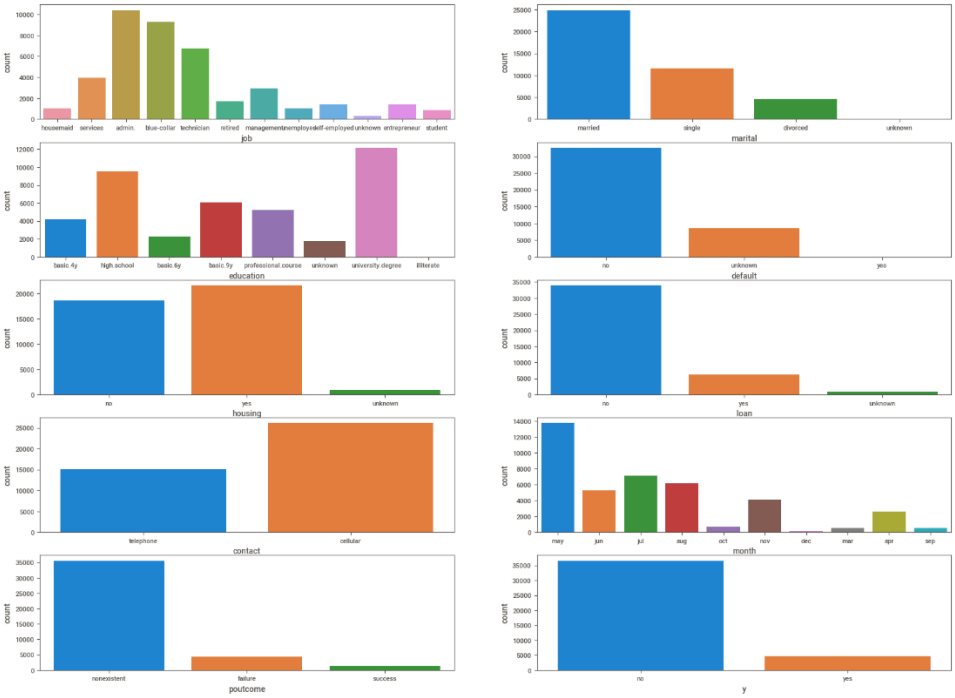
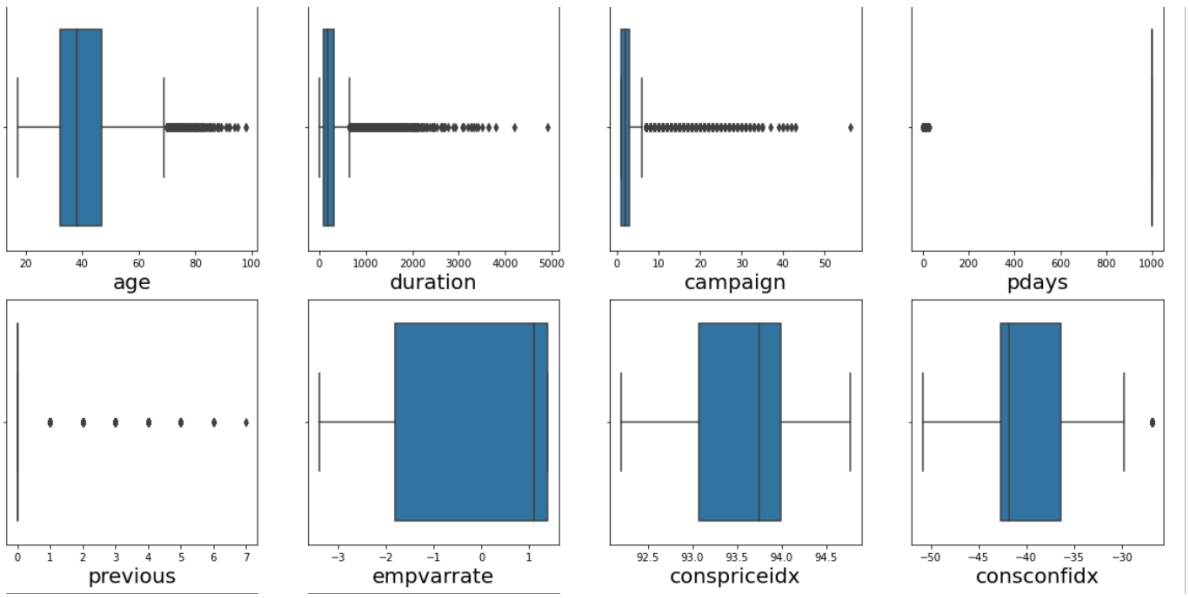


Figure 2: Univariate Analysis for Categorical Data

There are several missing values in some categorical attributes, all coded with the "unknown" label. These missing values are treated as imputation techniques by replace with the label which has highest density. For instances , 'job' has 'unknown ' (missing values). Use pandas replace methods to replace 'unknown' in the column 'job' with the mode of the attribute. As machine learning algorithms only take numerical values,all 11 categorial variables(job, marital, education, default, housing, loan, contact, month, poutcome , y , day_of_week) were converted into numerical variables , which simply represent different types by using map function. Next we need to handle the outliers, so we use boxplot for that.



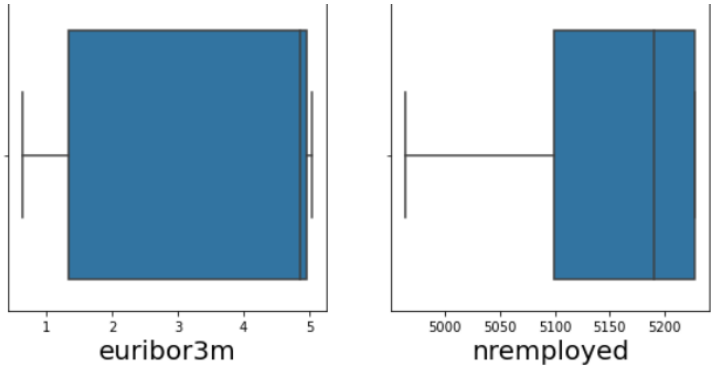


Figure 3: BoxPlot

From the above plot we can see that most of the features having outliers more than 5%, so we need not to apply any outlier treatment here. If we go to treat these outliers we may loose information.

FEATURE SELECTION

For this pipeline at first we check the correlation among features, if any features are highly correlated to each other, then we need to drop one of them(threshold value is 0.9). We use heatmap to find correlated features.

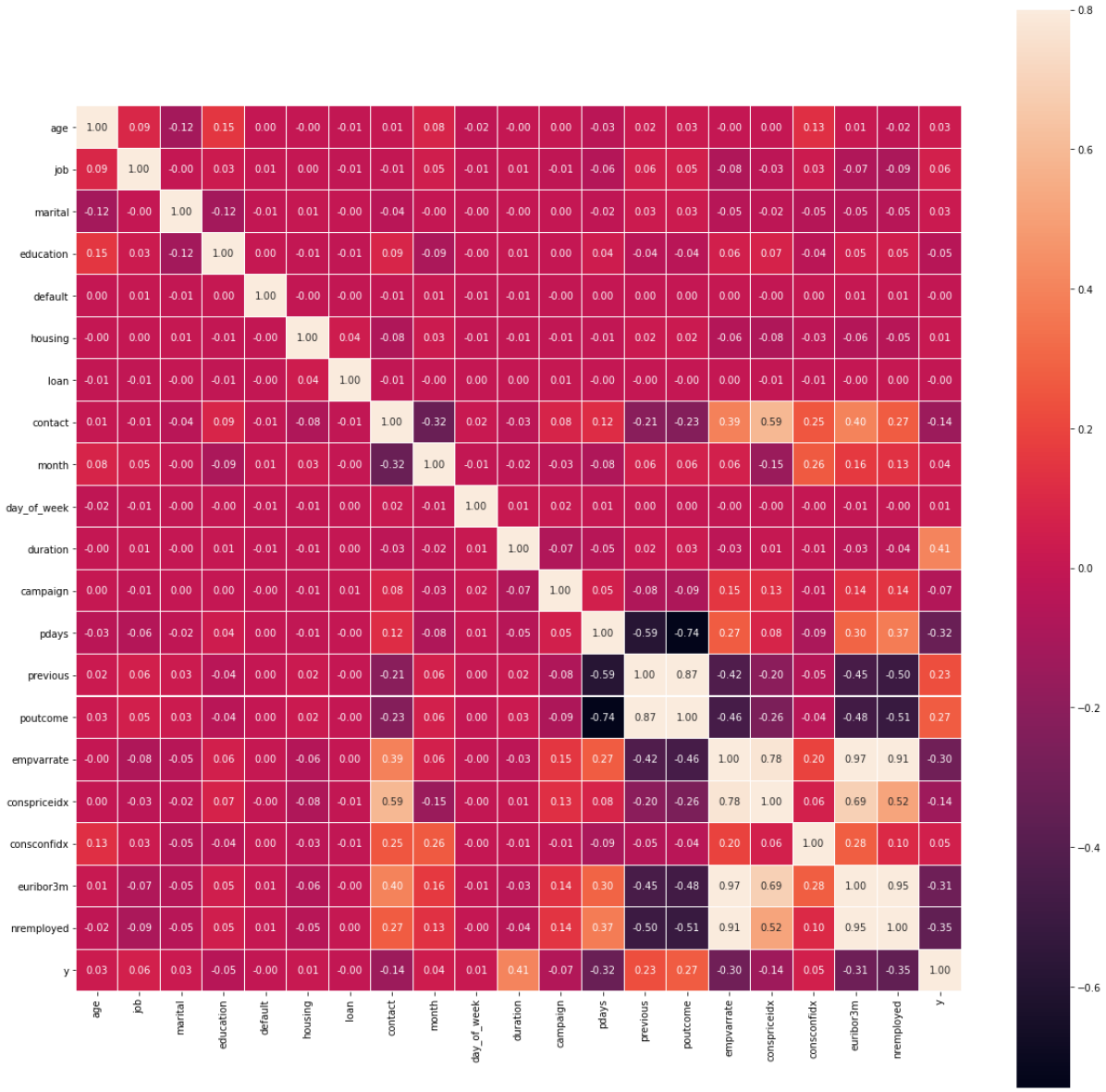


Figure 4: Heatmap

There is highly correlative columns like. emp_var_rate, euribor3m are highly correlated to each other, so we will drop one of them. emp_var_rate, nr_employed are highly correlated to each other, so we will drop one of them. euribor3m, nr_employed are highly correlated to each other, so we will drop one of them. Then we need to check if there are any duplicated rows exist or not. In this dataset there are 12 duplicated rows, we have drop those rows to increase model performances. We have checked if there was any constant features(Standard deviation is 0 of any feature to check) or unique features are there, if there are then we needed to drop them, in this dataset we do not have any type of them.

MODEL BUILDING

The Portuguese Bank marketing problem involves a classification challenge that whether a customer will subscribe to a fixed deposit scheme in response to a telemarketing campaign. It is a binary classification problem with YES or NO. YES means the customer will subscribe to the bank deposit scheme and NO means the customer will not subscribe it.

The data is observed to be very much unbalanced as shown here Counter({0: 36535, 1: 4639}). Most of the customers belongs to the 0 class that is they will not subscribe to the deposit scheme, this can reduces the performance of the model so the data is balanced using SMOTE technique. The data is standardised to remove the skewness, kurtosis and to improve the performance of the model.

The classification models used here are the following.

1)Logistic Regression

2)KNN

3)Bagging Classifier

4)Random Forest Classifier

5)SVM

6)Decision Tree

7)Gradient Boosting

8)XGB

The major parameters used for evaluating the model used are precision, recall and f1 score. Precision is the ratio between the True Positives and all the Positives. For our problem statement, that would be the measure of customer that were correctly identify to subscribe for the deposit scheme out of all the customers actually having it. The recall is the measure of our model correctly identifying True Positives. Thus, for all the patients who actually have subscribed, recall tells us how many we correctly identified as subscribed. There are also a lot of situations where both precision and recall are equally important. In such cases, we use something called F1-score. F1-score is the Harmonic mean of the Precision and Recall.

1) **Logistic Regression:** Logistic regression is the simplest classification algorithm we had used. It is fastest algorithm. The performance indicators are tabulated below

	precision	recall	f1-score	support
0	0.86	0.85	0.86	7332
1	0.85	0.86	0.86	7282
accuracy			0.86	14614
macro avg	0.86	0.86	0.86	14614
weighted avg	0.86	0.86	0.86	14614

2) **KNN:** The results for KNN classifier is tabulated below. The performance of KNN is ide ntical to that of Logistic Regression

	precision	recall	f1-score	support
0	0.86	0.85	0.86	7332
1	0.85	0.86	0.86	7282
accuracy			0.86	14614
macro avg	0.86	0.86	0.86	14614
weighted avg	0.86	0.86	0.86	14614

3) **Bagging Classifier:** A bagging classifier model is built with base estimator as KNN. It gi ves a good f1 score of 92%. The performance of bagging Classifier is tabulated below

	precision	recall	f1-score	support
0	0.99	0.86	0.92	7332
1	0.88	0.99	0.93	7282
accuracy			0.93	14614
macro avg	0.94	0.93	0.93	14614
weighted avg	0.94	0.93	0.93	14614

4) **Random Classifier:** Random forest classifier with n_estimators 100 gives the best res ult with an f1 score above 95%. The performance of Random Forest Classifier is tabul ated below.

	precision	recall	f1-score	support
0	0.97	0.93	0.95	7332
1	0.93	0.97	0.95	7282
accuracy			0.95	14614
macro avg	0.95	0.95	0.95	14614
weighted avg	0.95	0.95	0.95	14614

5) **Naive Bayes:** Naïve Bayes classifier gives only 73 % accuracy. The performance of Nai ve Bayes Classifier is tabulated below.

	precision	recall	f1-score	support
0	0.71	0.81	0.75	7332
1	0.77	0.66	0.71	7282
accuracy			0.73	14614
macro avg	0.74	0.73	0.73	14614
weighted avg	0.74	0.73	0.73	14614

6) **Gradient Boosting Classifier:** It gives a very good performance of 95%. The performa nce of Gradient Boost Classifier is tabulated below.

	precision	recall	f1-score	support
0	0.98	0.92	0.95	7332
1	0.92	0.98	0.95	7282

accuracy			0.95	14614
macro avg	0.95	0.95	0.95	14614
weighted avg	0.95	0.95	0.95	14614

7) **<u>XGB Classifier:</u>** It gives a very good performance of 95%. The performance of XGB Classifier is tabulated below.

	precision	recall	f1-score	support
0	0.95	0.95	0.95	7332
1	0.95	0.95	0.95	7282
accuracy			0.95	14614
macro avg	0.95	0.95	0.95	14614
weighted avg	0.95	0.95	0.95	14614

8) **<u>SVM Classifier:</u>** SVM classifier gives a good performance of 90%. The performance of SVM Classifier is tabulated below.

	precision	recall	f1-score	support
0	0.95	0.85	0.89	7332
1	0.86	0.95	0.90	7282
accuracy			0.90	14614
macro avg	0.90	0.90	0.90	14614
weighted avg	0.90	0.90	0.90	14614

RESULT & RECOMMENDATIONS

The final model chosen is Random Forest Classifier as it shows the best performance of f1 sc ore above 95%. Hyper parameter tuning will be performed on this model. However due to li mited computation capabilities the hyper parameter tuning takes marginally higher times in cluding hours and was unable to process it.