Process Flow :-

EDA

EDA

Machine learning Algorithm

Machin Learning Algorithm

& Training Model

Testing Model & checking accuracy and Roc Auc Score

Balanced Data using SMOTE

Create an Ensemble for the better Result

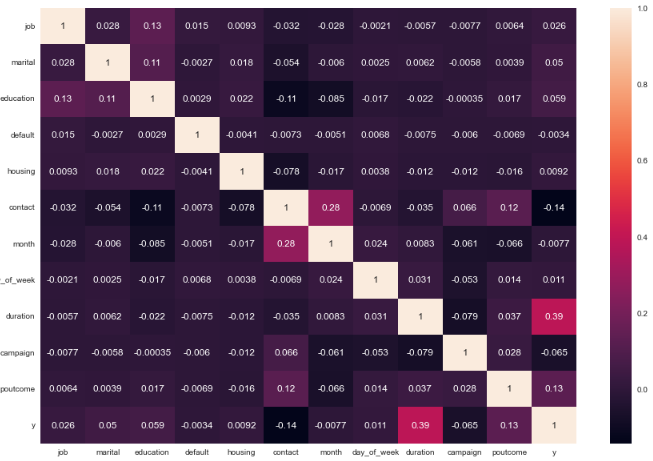
Testing Model and checking Roc Auc Score

Retrain All Machine learning model on new balanced Dataset

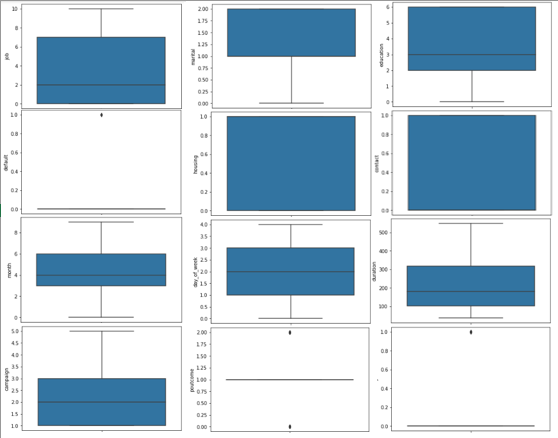
Balanced Dataset Using SMOTE

EDA :-

* There are a total of 13 independent features and one dependent features.
* There are a total of 32950 rows of data available.
* The datatypes of all data is Int64.
* There are a total of 55 duplicated rows. These rows will be removed as they may create bias in our mode.
* There are no null values present in our data, so there is no need to handle null values.
* To understand the data, we will find the statistical summary.
* We will find the correlation of the dataset and print a heatmap.
* From the heatmap, we can see that there is less correlation between the loan and y. Therefore, we can drop this column.



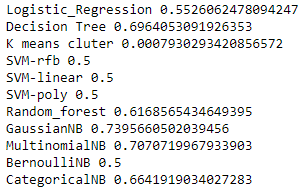
* Another insight is that age and marital status are highly correlated, which may create multicollinearity in our dataset. Additionally, age is less correlated with y. Therefore, we can drop age as well.
* We will find the outliers using box plots. However, there are no outliers, so we can move forward.



* We will divide data into X- independent features and y- dependent features.
* We will Split data into train and test and apply standard scaling.

Machine Learning Algorithm :-

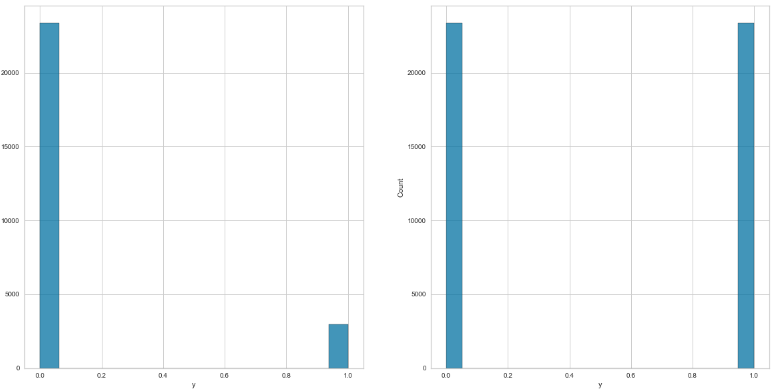
* Since this is a classification problem, some popular machine learning algorithms such as Logistic Regression, KMeans, Decision tree, Random Forest, Navie Bias, Support vector machine can be used. We apply each model and perform hyperparameter tuning to achieve maximum accuracy and roc auc score.
* We have calculated the roc auc score for all the models.



* Based on the roc auc score, I have selected the top 6 models for this dataset, which are GaussianNB, MultinomialNB, Decision Tree, CategoricalNB, Random Forest and Logistic Regression
* When I checked the classification report of all top 6 models, I found that the MultinomialNB model had a problem. Its accuracy for predicting 1 was only 28%, so I will not using this model further.

Balancing our Dataset :-

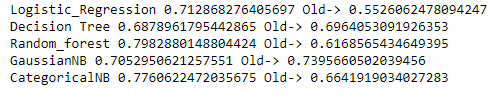
* Since our model’s accuracy is less than 75%, it is possible that this is due to an Imbalanced dataset. To address this, we need to balance the dataset using one of three popular methods.
* The first method is undersampling, which involves reducing the size of the majority class to match the size of the minority class. However, in our case, we have less data for 1, so this method is not useful.
* The second method is oversampling, which involves increasing the size of the minority class to match the size of the majority class. This can be done by duplicating the minority class data points.
* The third method is SMOTE(synthetic minority over-sampling technique) it’s also a oversampling method. It involves generating synthetic data points for the minority class by interpolating between existing minority class data points. We will use his method.



Distribution of data before and after SMOTE

Training & Testing Model again :-

* Thus, we retrained our models using the new balanced training dataset and the old hyperparameter. We then predicted y for thee testing dataset and obtained the roc auc score for each model, comparing it with the old roc auc score.



* After analyzing the new balanced training dataset, we found that some models perform better while others have slightly lower roc auc scores compared to the old dataset.
* In order to improve the results, we used an ensemble technique that combines the top 5 models’ results using the majority rule.
* We obtained the roc auc score for our ensemble model.

