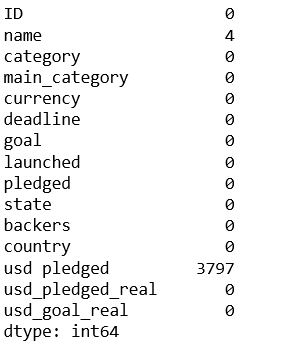
Report for SQL

**Machine Learning Models**

We implemented Machine Learning models to predict the various classes of attribute ‘State of the project’, based on provided independent variables. So, we have ‘state’ as target variable and 'ID', 'name', 'category', 'main\_category', 'currency', 'deadline', 'goal', 'launched', 'pledged', 'backers', 'country', 'usd pledged', 'usd\_pledged\_real', 'usd\_goal\_real' as predictors.

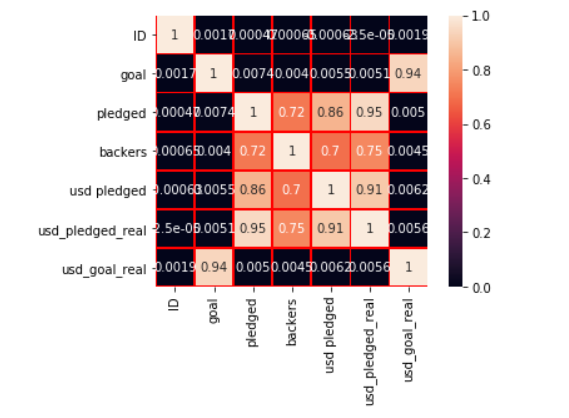
To begin with, we started with Data Preprocessing. First step was to check for the missing values, and we found 4 values missing for attribute ‘Name’ and 3797 values missing for attribute ‘usd pledged’.



There are a lot of methods to deal with missing values such as fill missing values with ‘mean’ or ‘median’ if the missing value is numerical and if the missing value is categorical then, replace the missing value with ‘mode’. In our case, since proportion of missing values were very low, hence we dropped the missing values, instead of plugging in the calculated values.

Our second step was to visualize the correlations between the variables in our data set. We used seaborn library and plotted ‘Heatmap’. If the correlation value is near to 1, means both the variables are highly correlated and if the correlation near to 0 then both variables are least correlated.

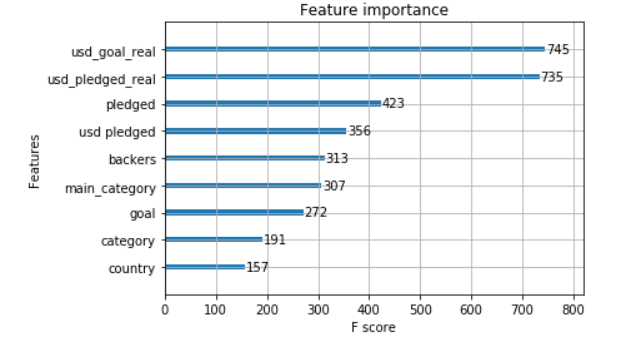
For example, usd\_pledged\_real and pledged are highly correlated.



In the next step, we dropped the columns which do not seem to add any value to prediction. Then, we proceeded with label encoding instead of one-hot encoding as there were a lot of categories in each categorical data.

For example: We converted different classes of ‘State’ : "Canceled Failed Suspended Successful Live" to 0 1 2 3 4 respectively.

Next, we selected the most important features for training our models and then testing our models on unseen data. This feature selection is one of the optimizing techniques which increased the accuracy and reduced the time duration of training and validation. Below is the plot of feature importance through XGBoost classifier. We can observe that, usd\_goal\_real and usd\_pledged\_real are most important features and on the other hand country is least important for predicting Success and failure.



**Models:**

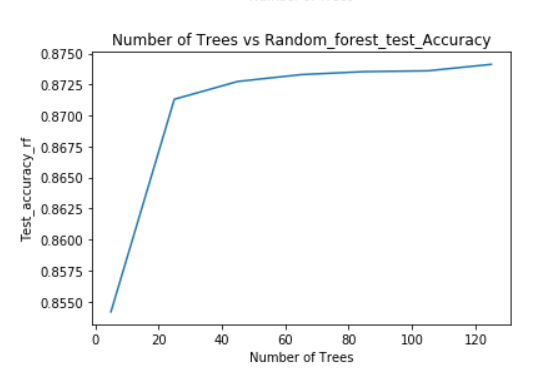
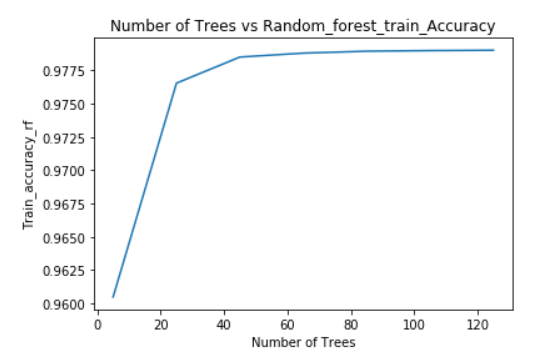
**Random Forests Classifier**

**Random forests are an ensemble learning method based on bagging technique. Random forests are widely used as it is less likely to overfit and being an ensemble learning method tends to provide high accuracy.**

**In our case, Random forest has performed well on predicting the target variable.**

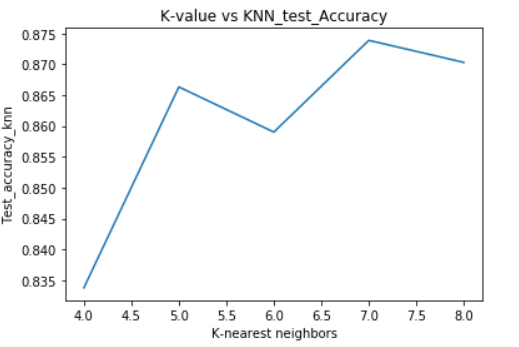
Hyperparameter tuning

We improved the performance of the model by tuning the hyperparameter as in this case ‘Number of trees’ was a hyper parameter and we can observe the relationship between the Number of trees and Accuracy. For 120 trees both train and test accuracy are reaching its highest value.



K-Nearest Neighbors

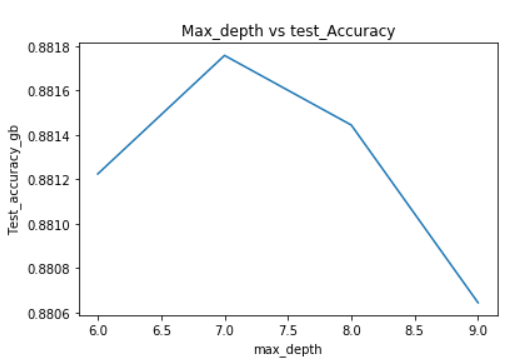
Similarly, for K-nearest neighbors, for 7 nearest neighbors, the model is performing best and for



**Boosting Algorithms**

Gradient Boosting and XGboost classifiers are the boosting methods, where we ensemble the weak learners in a sequence. If prediction is incorrect using the first learner, then it gives higher weight to observation which have been predicted incorrectly.

Gradient Boosting Classifier if the max\_depth parameter is set to 7, then it providing high test accuracy.

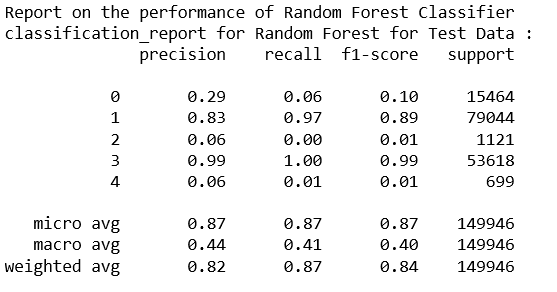


**Summary of Accuracies of different models**

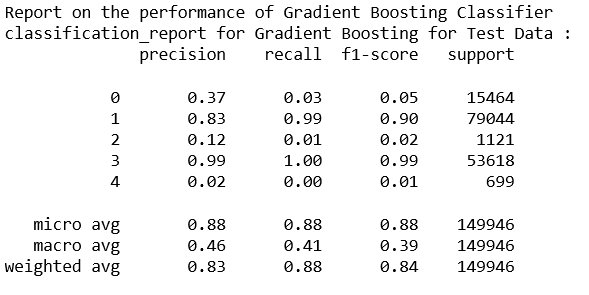
|  |  |  |
| --- | --- | --- |
| **Models** | **Train Accuracy(in %)** | **Test Accuracy(in %)** |
| **Random Forest Classifier** | **97.9** | **87.412** |
| **Gradient Boosting Classifier** | **90.25** | **88.157** |
| **K-Nearest Neighbours** | **88.77** | **88.39** |
| **XGBoost Classifier** | **90.24** | **88.10** |

**Model performance based on metrics such as Precision, Recall, f1-Score**

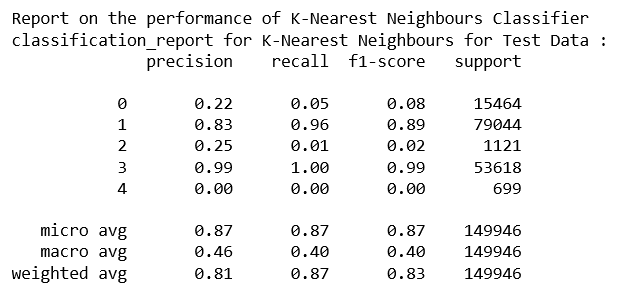
**Random Forest Classifier**



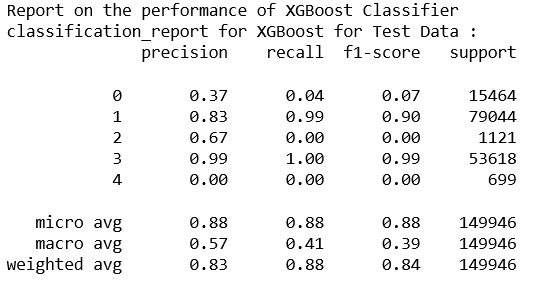
**Gradient Boosting Classifier**



**K-Nearest Neighbors Classifier**



**XGBoost Classifier**



Conclusion:

Using the above models, one will now be above to predict the success, failure, suspension, cancellation and Live with an accuracy of 88 percent.

Scope of Improvement

One aspect that we think will improve the prediction is inclusion of brief description of project and technology being used