Machine Learning Kaggle Competition

Name: Fareed Hassan Khan (ERP ID - 25367)

Before Model Implementation

Since the dataset contain category columns, that needs to be converted into string and then one-hot encoding is used to convert those into numeric.

Decision Tree

After applying on complete dataset with one-hot encoding and changing the max_depth parameter to different values, got the highest roc score value of 0.75760, but different approaches had to be done to check whether this value is increasable or not?

Logistic Regression

Logistic regression is the worst approach among all the models implemented on this dataset, as it gives the roc score of 0.72. Scaling the dataset, removing variables by p value does not make an impact on roc value. This may indicate that the target label has no strong linear correlation with the features.

KNeighborsClassifier

Applying KNeighborsClassifier on scaled dataset with n_neighbors up to 100 or 150 increases the roc value upto 0.82 which clearly shows a better approach to predict based on this dataset but by further increasing or decreasing the n_neighors does not return the same but decreases the roc value.

RandomForestClassifier

Applying random forest classifier without scaling and with different values of max_depth and n_estimators increase the roc value as compared to all the previous models approaches.

Scaling the dataset, removing variables by p value does not make an impact on roc value as compared to previous approach of random forest.

GradientBoostingClassifier

In Gradient Boosting, increasing the estimators value and decreasing the max depth value increases the roc score value not only as compared to random forest but also compared to all the previous models approaches. This may indicate that the data is unbalanced and has less noise.

Scaling the dataset, removing variables by p value or correlation does not make an impact on roc value as compared to previous approaches.

Log transformation of dataset

Log transformation is applied on dataset to normalize it and to test what impact will it make on area under the roc curve value?

Minor improvement did take place in random forest and gradient boosting, while the rest of the models have no impact because of this transformation.

Conclusion:

Based on area under the roc curve, Gradient Boosting did perform well as compared to other models and by increasing estimators to more than 500 the roc increases further but with slow rate. While for random forest, it produces results somewhat like gradient boosting in decimal places but not crossing the score of it. Rest of the model's performance are not as good as compared to these two. Normalizing, scaling and removing variables based on p value or correlation did not make a good impact on these models as compared to non-normalized dataset without such implementation.