CoE202 - Basics of Artificial Intelligence “Big data analysis and machine learning”

Mukanov Ayan 20170881

**The Titanic Project Report**

1. **Data analysis and Preprocessing**

During this part, we have two data sets: training and test. The target attribute is ‘Survival’ and it’s only available in training set, since we will use test set to predict if the person survived. The result should be a set where given some “PassengerId” from test set, our algorithm should tell whether that passenger survived. There are several attributes that constitute data set:

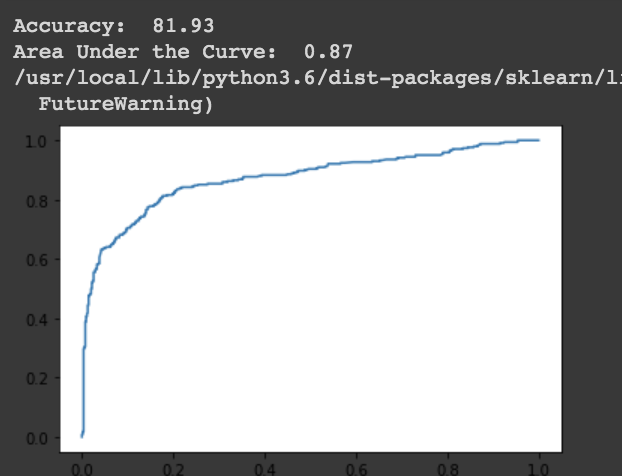
There are Name, Ticket, Cabin, Embarked, Sex- all categorical classes.

Also, there are Age, Pclass, SipSp, Parch, Fare – numerical classes/attributes.

All of the categorical attributes were subject to preprocessing, and some of numeric attributes too.

1. We create a new feature named FamilySize, by using and adding SibSp and Parch. It can be observed that the rate of “survival” of those having FamilySize of 3 or more is lower that of less than 3, so we assign values 0 to ones who have FamilySize 3 or less, and 1 to ones with greater FamilySize.
2. We restore the None values of Embarked with S, and assign numerical values to the names of destinations(vectorization)
3. We discard Ticket, since that data doesn’t have relevance to us since it can’t be related to Survived.
4. We set 0 for female in Sex, and 1 for male.
5. The Names are used to create Title, which in turn is vectorized. Since there are several Titles, and the frequency of them vary, we replace the rarest ones with 5, and more frequent ones, like ‘Mr’, ‘Ms’, etc with 1 to 4 (0- for unknown Title).
6. The we group data by Sex, Pclass and Title, and get the median age of this grouped features instead of average value of 30. We use this median age to fill in the missing data in the numeric Age attribute.
7. We then fill in the missing values in numeric Fare attribute with its median value, since it’s clearly a better choice that using the mean.
8. We then work on preprocessing the categorical Cabin attribute. Even though it may seem unnecessary or irrelevant to our task at a first glance, it is actually useful in a way that people from some Cabins, which basically signify sections of Titanic might have higher Survival chance, as a result of closer proximity to life boats or other factors. We first fill unknown Cabins with a letter U (stands for unknown), and extract the first letter of the Cabins and reassign it to them. We then vectorize this categorical data/attribute by using pd.get\_dummies() method for Pandas.
9. The last step was the “standardization” (removing the mean and scaling to unit variance) of Age and Fare attributes, for which I used sklearn.preprocessing.StandardScaler function.
10. **Machine learning**

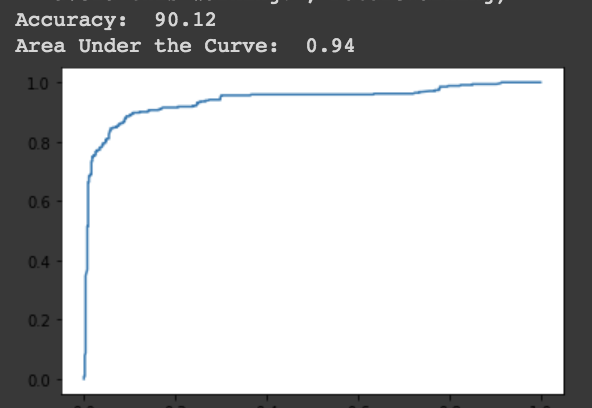
1.First we use **Logistic Regression**, a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. This model can be further improved by the utilization of SVM that actually can handle numerical and categorical attributes.

Accuracy : 81.93 Area Under Curve: 0.87

A receiver operating characteristic curve (ROC)

2.Next we use **Support Vector Machine,** and its classifier tries to predict the class of a given test observation by identifying the instances that are nearest to it. So given the importance of variable scales, some feature scaling is necessary.

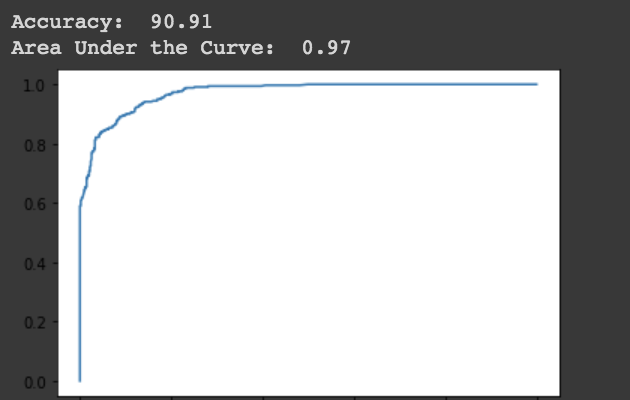
Accuracy: 90.12 Area Under Curve: 0.94



A receiver operating characteristic curve (ROC)

3.Lastly, **Random forest** is ML algorithm with high classification accuracy that improves the accuracy of predictions by generating a large number of bootstrapped trees (based on random samples of variables).

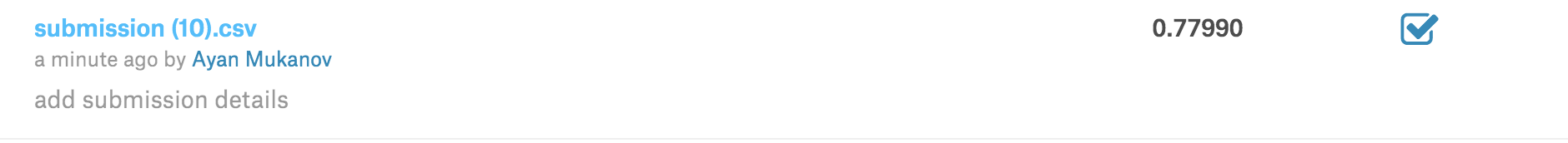
Accuracy: 90.91 Area Under Curve: 0.97



A receiver operating characteristic curve (ROC)

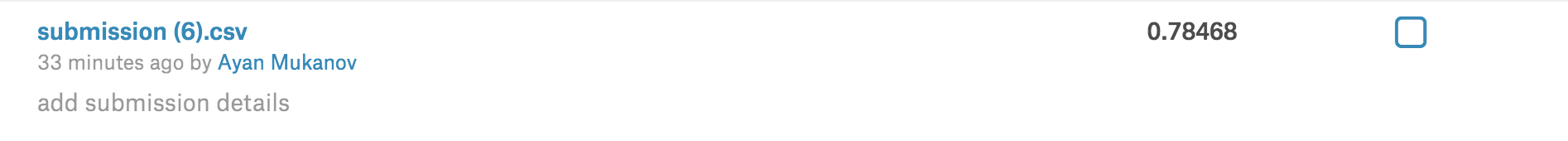
Overall, after numerous attempts to modify classifiers by adjusting hyperparameters and modifying and further elaborating the preprocessing process, I can see that the best-performing model is Random Forest classifier. We can notice its from the receiver operating characteristic curve (ROC curve) and by comparing the results of accuracy of these three models, of which Random forest, obviously, has the highest result.

1. **Kaggle competition**

****

The Last score – 0.77990

The Best Score – **0.78468**



The same preprocessing procedure and Random forest algorithm was used as mentioned in this report. For the best score run, I used Random forest algorithm, where I used GridSearchCV to pass in a range of parameters for obtaining the most optimal ones