Report

(Coding Assignment-1)

EDA

I started with Exploratory data analysis (EDA). In training- dataset there are 891 examples and 11 features + the target variable (survived).

First I checked about missing data from dataset and learned which features are important. During this process I used seaborn and matplotlib to do the visualizations.

Data Preprocessing

During the data preprocessing part, First, I dropped ‘PassengerId’ from the train set, because it does not contribute to a persons survival probability. I did not drop it from the test set. For 'Cabin' converted the feature into a numeric variable and the missing values converted to zero. For 'Age ' computed based on the mean value in regards to the standard deviation and is_null. ‘Embarked’ feature has only 2 missing values, so just filled these with the most common one.

Convert Features

Converted 'Fare' from float to int64, using the “astype()” function pandas provides. Used the 'Name' feature to extract the Titles from the Name, so that we can build a new feature out of that. Convert ‘Sex’ feature into numeric. Since the Ticket attribute has 681 unique tickets, it will be a bit tricky to convert them into useful categories. So I just drop it from the dataset. Convert ‘Embarked’ feature into numeric. Thus I have computed missing values, converted features into numeric ones, dropped features, grouped values into categories and created a few new features. I added two new features to the dataset that I compute out of other features.

1. Age times Class

2. Fare per Person

Machine Learning Models

I trained several Machine Learning models and compare their results. Afterwards I started training different machine learning models- logistic regression, naive bayes, SVM, random forest, decision tree. Random forest and decision tree model gave

best result. Later on applied cross validation on random forest and trained the model again to get better result then before but result was same. Note that because the dataset does not provide labels for their testing-set, it's need to use the predictions on the training set to compare the algorithms with each other.

Accuracy of Logistic Regression: 81.82%

Accuracy of Gaussian Naïve Bayes: 78.34%

Accuracy of SVM: 81.83%

Accuracy of Random Forest: 92.56%

Accuracy of Decision Tree: 92.56%

Lastly, I looked at confusion matrix and computed the models precision, recall and ROC AUC score.

Precision and Recall:

Our model predicts 79% of the time, a passengers survival correctly (precision). The recall tells us that it predicted the survival of 70% of the people who actually survived.

I ploted the precision and recall with the threshold using matplotlib.

Confusion Matrix:

The first row is about the not-survived-predictions: 484 passengers were correctly classified as not survived (called true negatives) and 65 where wrongly classified as not survived (false positives).

The second row is about the survived-predictions: 102 passengers where wrongly classified as survived (false negatives) and 240 where correctly classified as survived (true positives).

ROC AUC score:

The ROC AUC Score is the corresponding score to the ROC AUC Curve. It is simply computed by measuring the area under the curve, which is called AUC.

A classifiers that is 100% correct, would have a ROC AUC Score of 1 and a completely random classiffier would have a score of 0.5.

The ROC AUC score of my model is 97% that is pretty satisfying.