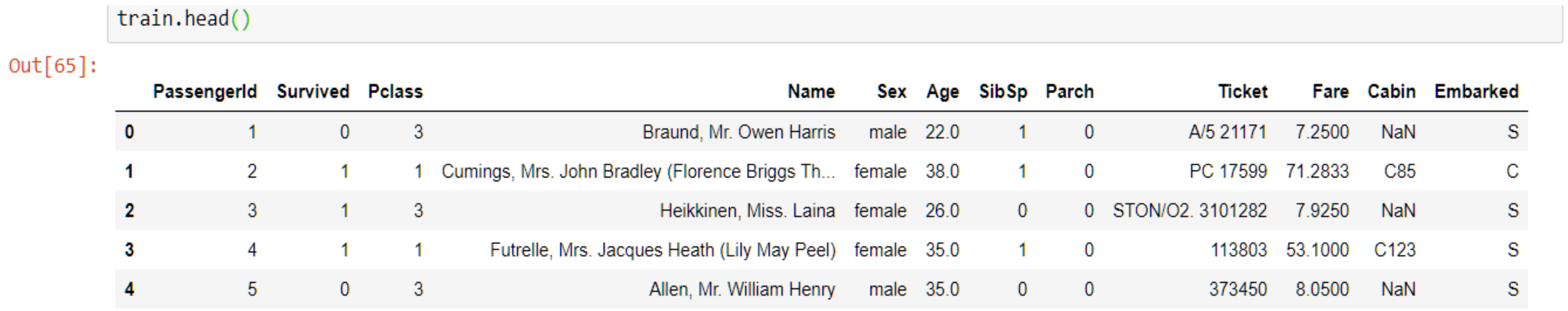
CSCI 334 Project

Titanic: Machine Learning from Disaster

About the Dataset

The Titanic dataset is of the shape: (891, 12). This means that we have 891 rows and 12 columns. The column labeled “Survived” represents our target values. The class, Survived = 0 means that a passenger did not survived and Survived = 1 means that a passenger survived. Here is a snapshot of the first few rows of the dataset:



```
train.head()
```

Out[65]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

I chose to disregard the columns “PassengerID”, “Name”, “Ticket”, and “Cabin” while training my classification models.

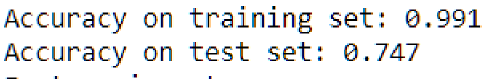
The feature “Sex” and “Embarked” had to be converted into numeric values from the provided string values in order to make it easier for calculations.

The feature “Age” had many missing values. We dealt with this by replacing the missing values with the median value of this feature.

Training and test data

I chose to use 75% of the dataset provided in training.csv as the training data for my prediction models and the other 25% as my test data.

I chose to go with this split because as I increased the percentage of test data, the models showed more overfitting. For example, the Decision tree classifier showed the following accuracies when the data was split 50/50



```
Accuracy on training set: 0.991
Accuracy on test set: 0.747
```

However, when the data was split 75/25 the accuracy on test data jumped up by 4-5%

Models used for prediction

1. Decision Tree and Random Forest.

   The Decision Tree performed very well on the training set(98% accuracy), but only reached a max accuracy of 80% on the test data. This can be a result of overfitting.

   Decision Tree parameters and their effects on accuracy

   a. Random state: lower random states resulted in lower test data accuracy, but the training accuracy was the same for all random states

      Random state = 12 resulted in the best accuracy. For the accuracy dropped as it got higher than 12.

   b. Max Depth: While using all features to train the model, a max depth of 25 was the optimal value. Lower max depths resulted in higher inaccuracies. Depths larger than 25 resulted in the same accuracy as a depth of 25.

   Feature Importance based on the decision tree

   The feature “Embarked” had the least information gain and when the data was trained without this feature, the acuuracy stayed the same for both test and training data.

   The feature “parch” had the second lowest information gain, but omitting this feature while training resulted in drops in both training and test data.

Random Forest Classifier

Because the random forest classifier uses a voting result from n decision tree outputs, it can minimize overfitting.

As we increase the number of estimators the accuracies on both test and training data increases.

If we use 200 estimators, it results in the best accuracy (training accuracy = 98.1%, test accuracy = 81.2)

2. **Logistic Regression**

The Logistic regressor should be a good model for prediction on binary class classification. However, from the accuracy results we find that it did not have better accuracy than the decision tree classifier. The parameter of choice was solver = “lbfgs”. The training accuracy was far below that of the decision tree classifiers.

Training accuracy = 81.1%

Test accuracy = 77.1%

3. **Naïve Bayes classifier**

This classifier does not fit well with our data because the features are not independent. The accuracy we get are: Training accuracy = 80.5% Test accuracy = 74.9%

4. **MLP classifier**

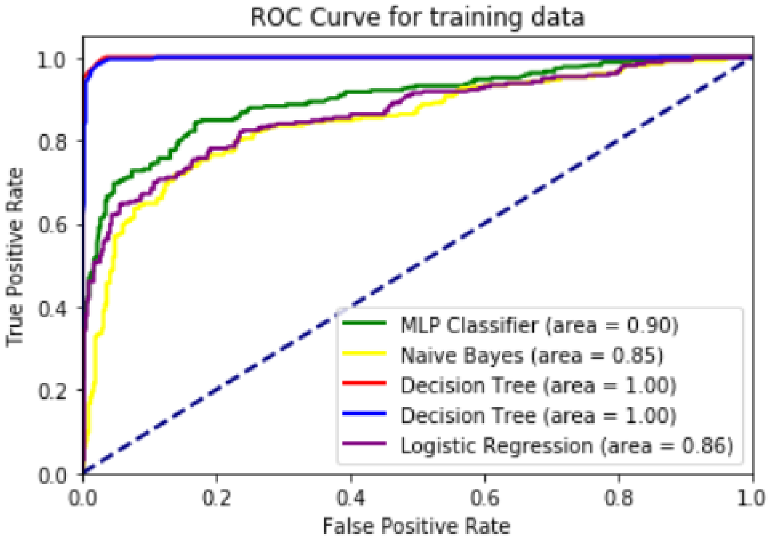
This classifier did not vary in results when the parameters activation function and number of hidden layers were changed. It was highly inaccurate on the training data.

Training accuracy = 78.4%

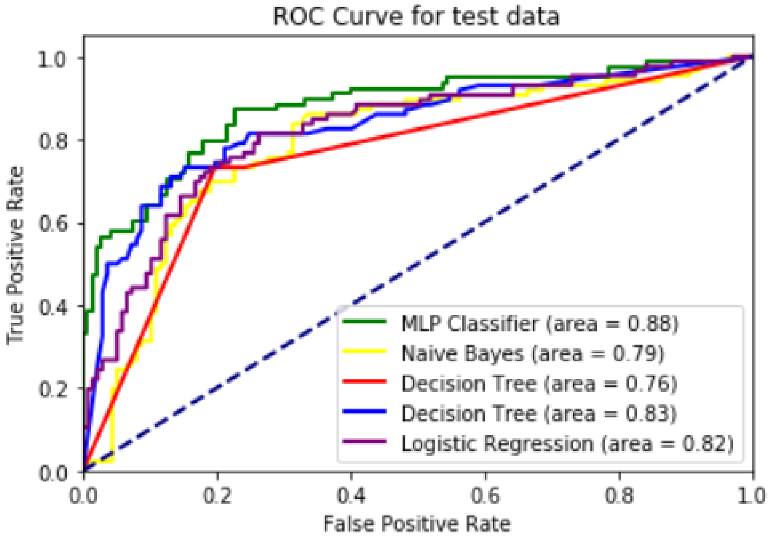
Test accuracy = 81.2%

Comparing the prediction models using ROC curve and AUC values

Training data:



Test data:



The MLP classifier performs the best on the test data based on ROC curve

The Decision Tree and Random Forest Classifier performs the best based on the ROC curve.