Amanda Moreland
9/8/15
Data Science Midterm

Questions:

1. I dealt with the missing values in a number of ways.

   There were two missing values in the ‘Embarked’ feature, for these values I replaced the missing rows with the most common port of departure, which was Southhampton (S). I think this was a valid method, because such a greater number of passengers embarked from this port, rather than Cherbourg or Queenstown, it is likely that the two passengers with missing information for port of embankment likely left from Southampton.

   For the missing values concerning Age, I filled in the values with the average of the age values. This method is not necessarily statistically sound, as the average of the ages may still cause bias in the data set. However, it is a quick fix for the missing values in this data set.

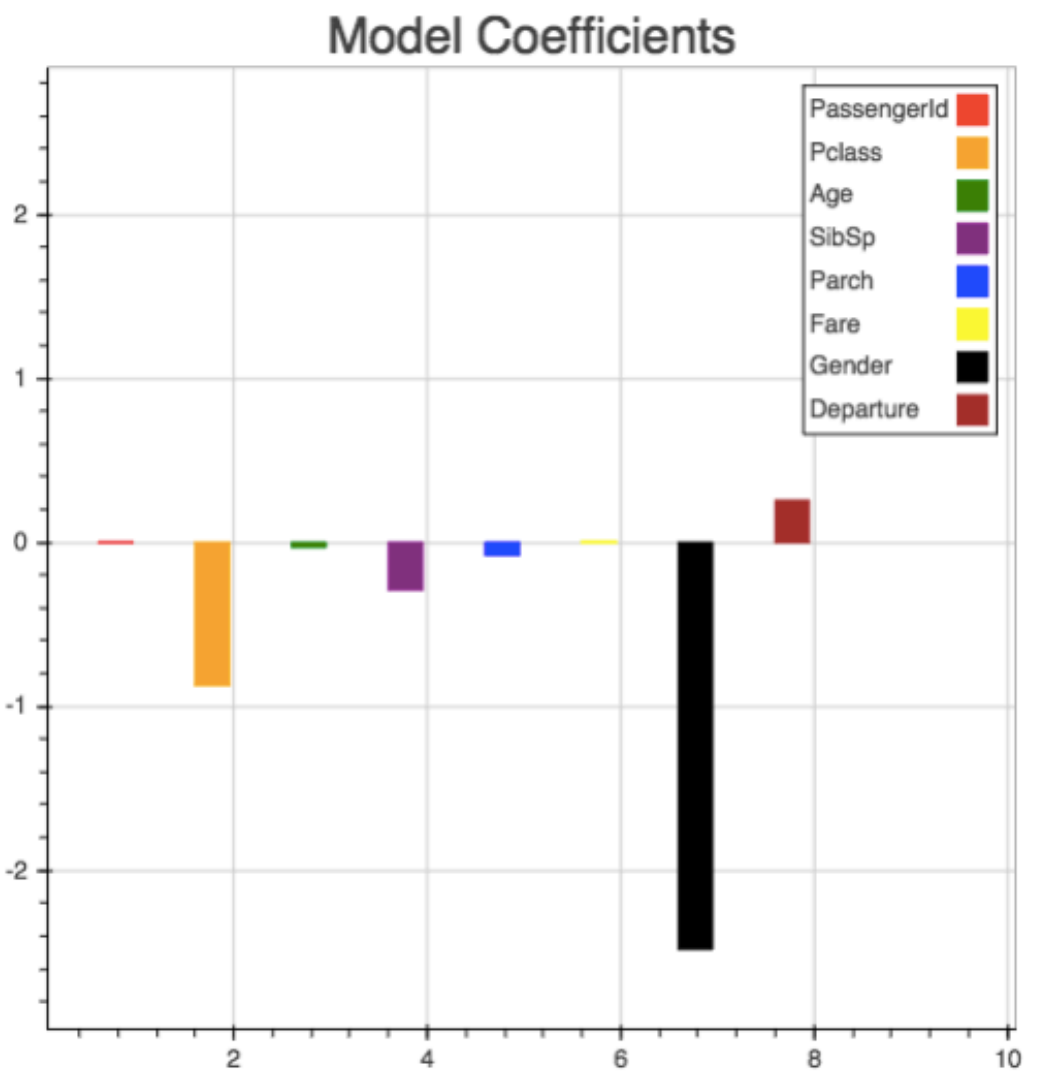
   The ‘Cabin’ feature had too many missing values to be relevant, so I decided to drop that feature from this analysis.

   I dropped the ‘Name’ feature because it would not be relevant in this analysis.

   I made ‘Sex’ a numerical binary variable and renamed it ‘Gender’, and I also made ‘Embarked’ a numerical variable (S=0, C=1, Q=2) and renamed it ‘Departure’.

2. Logistic Regression
   a. values :

```
[0 1 2 3 4 5 6 7] [0.78787726104700961, 0.79685746352413023, 0.79683572
806470793, 0.7957292518903023, 0.79794727613519567, 0.79909514658881931
, 0.80023582836082841, 0.7991021324354658, 0.79461411871524223, 0.79798
835235827104, 0.80025381545929486, 0.80024591776509935, 0.8012654260422
1196, 0.80028118709007079, 0.80022179027113238, 0.80022750916202645, 0.
80126984126984135, 0.7991382248405472, 0.80013175230566547, 0.798073281
46198586, 0.80242907499005056, 0.80147036668775795, 0.79913136821031561
, 0.80481605975723625, 0.80348883143000782, 0.80017288902092831, 0.8012
8800970534841, 0.80272077531868091]
```



b. The features that are highly predictive for this model are Pclass and Gender. The moderately predictive models are SibSp and Departure. The other features are hardly predictive.

   Gender would be highly predictive as more women survived this tragedy (ie save the women and children). I had also changed this variable to be binary where female was represented by 0 and male was represented by 1, indicating that with the negative representation on this graph, being female was highly predictive of survival on the Titanic.

   Pclass would also be highly predictive as the higher class had an overall greater survival in this tragedy, where the lower class had worse survival rates. Hence, it is logical that this feature would be highly predictive.

It is interesting that the number of siblings (SibSp) and Departure (Embark) are moderately predictive. The greater number of siblings could have had a greater representation due to the greater chance of survival of one child out of a large family. And Departure is represented by (S=0, C=1, Q=2), and the greatest number of people departed from Southampton, so the predictive strength of this feature is likely biased.

It is logical that passenger ID would not be highly predictive, but it is odd that Age, Parch and Fare are less predictive. I would have expected each of those variables to be more predictive. Age, as more children likely survived (greater number than adults), Parch, as SibSp had moderately great predictive strength and families would have likely boarded the ship together, and Fare, as it is another indicator of Pclass.

c. I think I passed values though the model correctly

3. I chose 7 folds due to a CV score scatter plot that I generated based on the variance. Seven has a relatively low variance, and a mediocre cv sore based on mean. I determined that this number of folds is the best compromise to create the most effective model.
4. ROC curve
   a. Plotted the curve
   b. AUC =0.821
   c. The model has achieved this level of accuracy due to relying on the predictive capability of the model that was generated from the features of this data set that I engineered.
   d. Improving the metrics of this curve could be carried out by improving the model, by such ways as eliminating the arbitrary features such as PassengerID, which has little predictive value as it is just an assigned identification number. Relying on the highly predictive features would improve the area under the ROC curve.
   e. I would likely use a threshold cut off value of 0.90, as I feel that it is a justifiable value to use to indicate that I can be confident in the predictive capability surrounding that threshold.