Joseph Bieselin

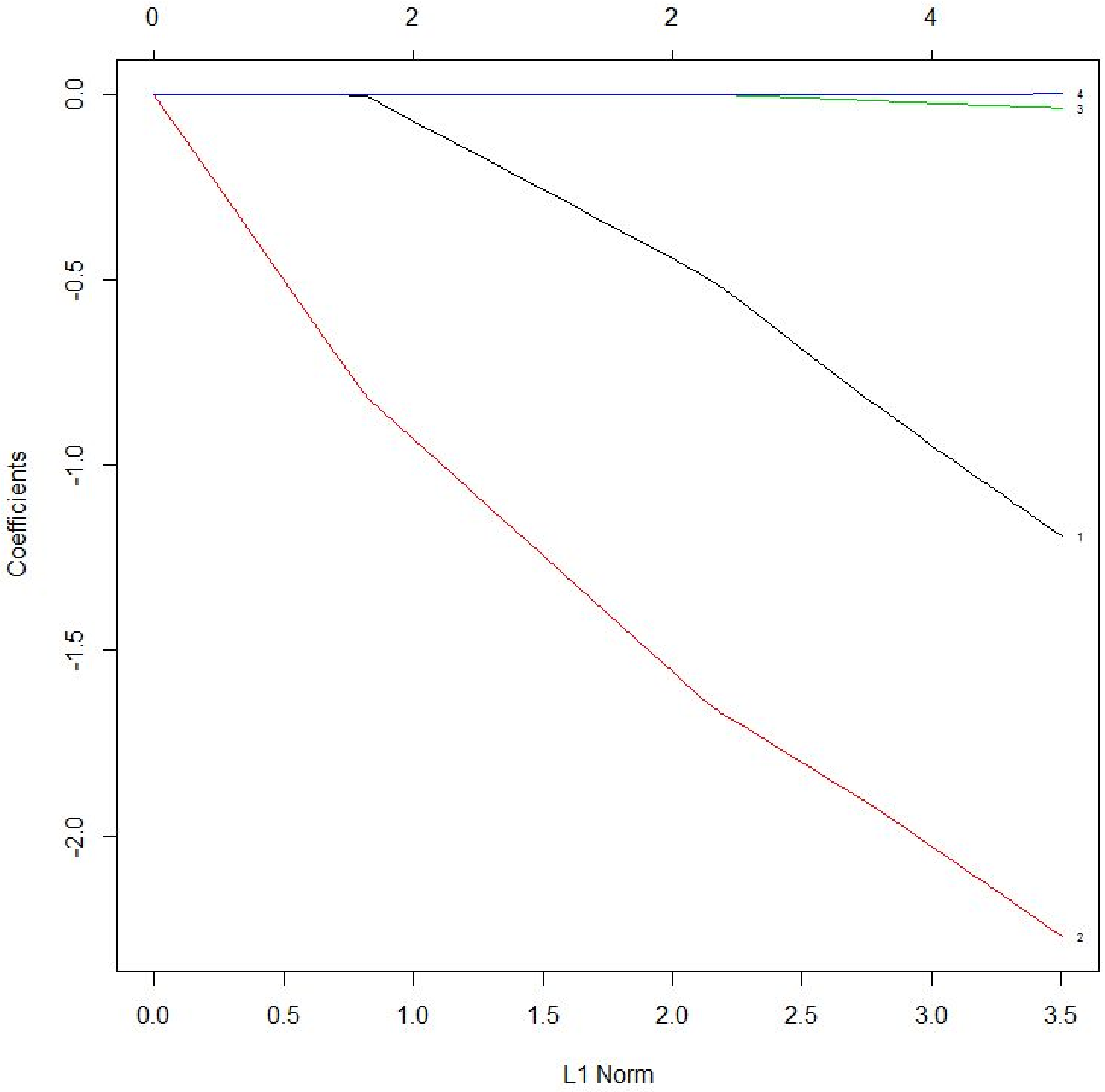
LASSO Regression - Titanic dataset

December 12, 2016

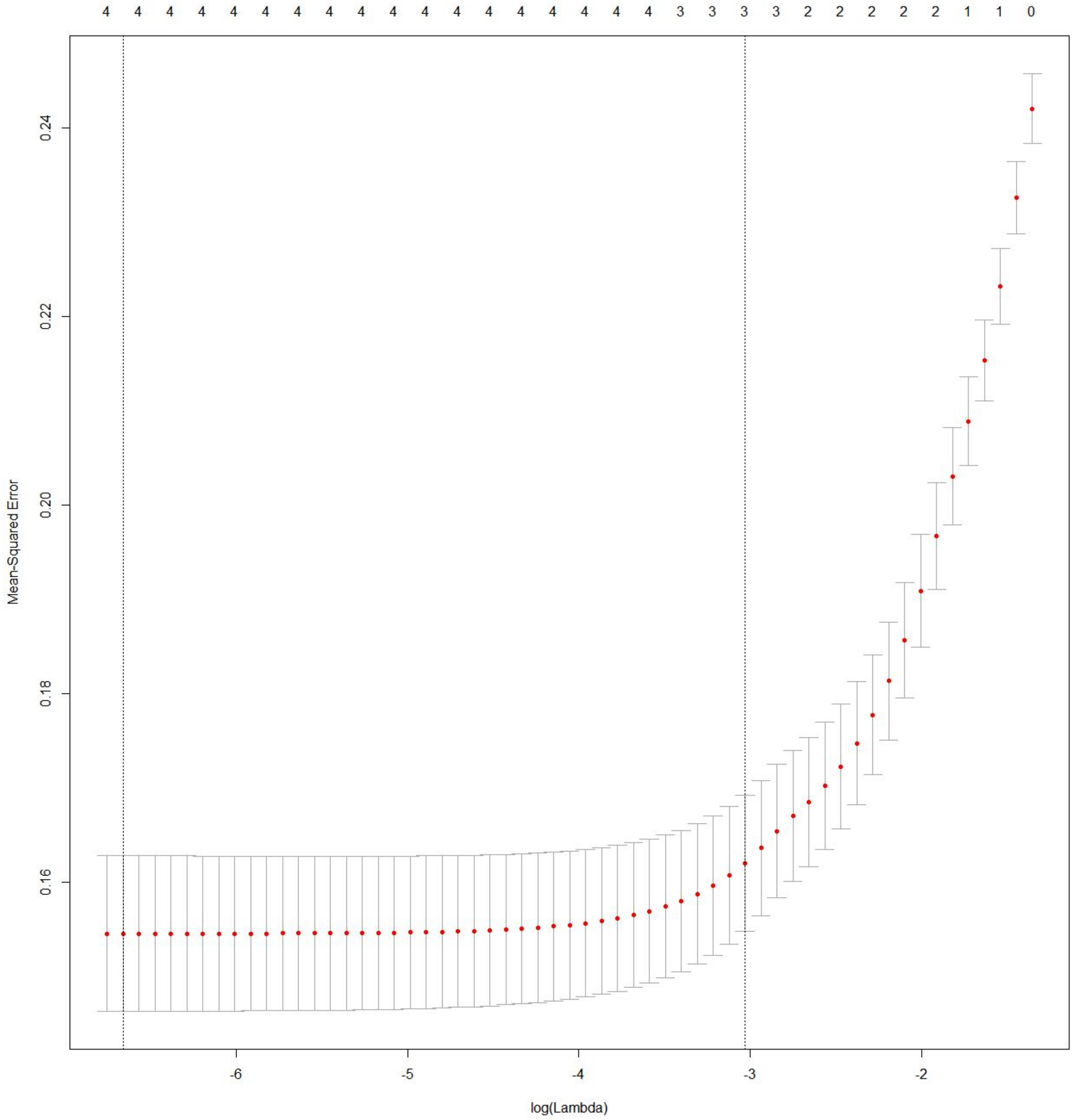
Dataset: https://www.kaggle.com/c/titanic/data
Data Preparation & Analysis Scripts are included in the zip file.
How-To-Help.txt is included in the zip file.
References.txt is included in the zip file.
Analysis & Summary-of-Findings is done below.
Presentation (PowerPoint) is included in the zip final.

Analysis was done by trying to classify whether or not someone would have survived the Titanic shipwreck. The fields for PassengerId, Name, SibSp, Parch, Ticket, Cabin, and Embarked were removed for this project. There was currently no need to use these features to determine whether or not LASSO regression would be useful in selecting a more limited amount of features for prediction.

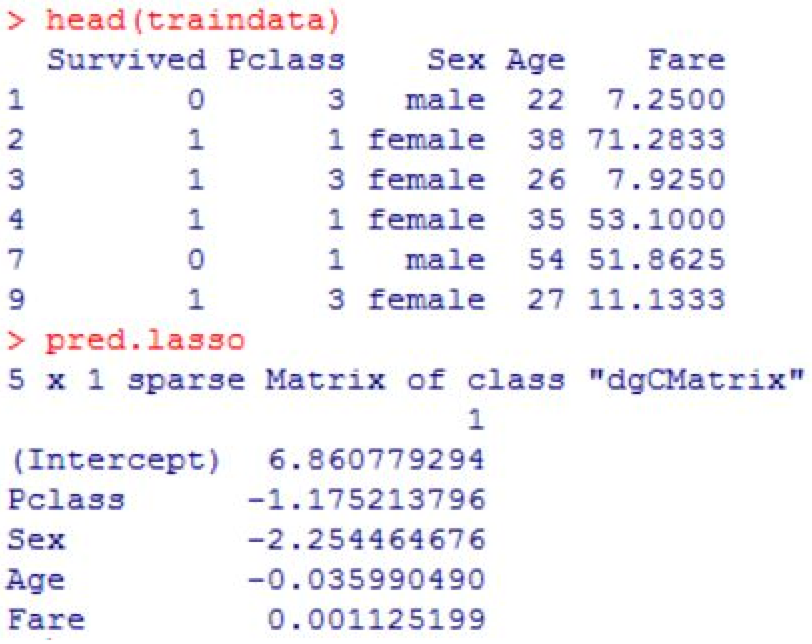
Plot of lasso.model. LASSO is useful for feature selection by setting variables coefficients to 0 to determine features with less importance. The following plot shows that features 3 & 4 which correspond to Age & Fare in the Titanic dataset do not have as much importance as Plcass & Sex (features 1 & 2).



Plot of cv.lasso.model which is the cross-validated model using n=10. The model is looking for an optimal lambda to size coefficients such that there is not too much complexity. As the size of a coefficient for a variable increases, the model is placing more emphasis on that variable. When a model is too complex and a coefficient becomes too high, this can result in overfitting. This model's minimum lambda value will be used for prediction because the minimized lambda is given as parameter in the cv.lasso.model glmnet object. However, this lambda.min would be different if the cv.glmnet algorithm was run again (even if on the same data). If this model needed to be improved, the algorithm could be run multiple times to get the average lambda.min.

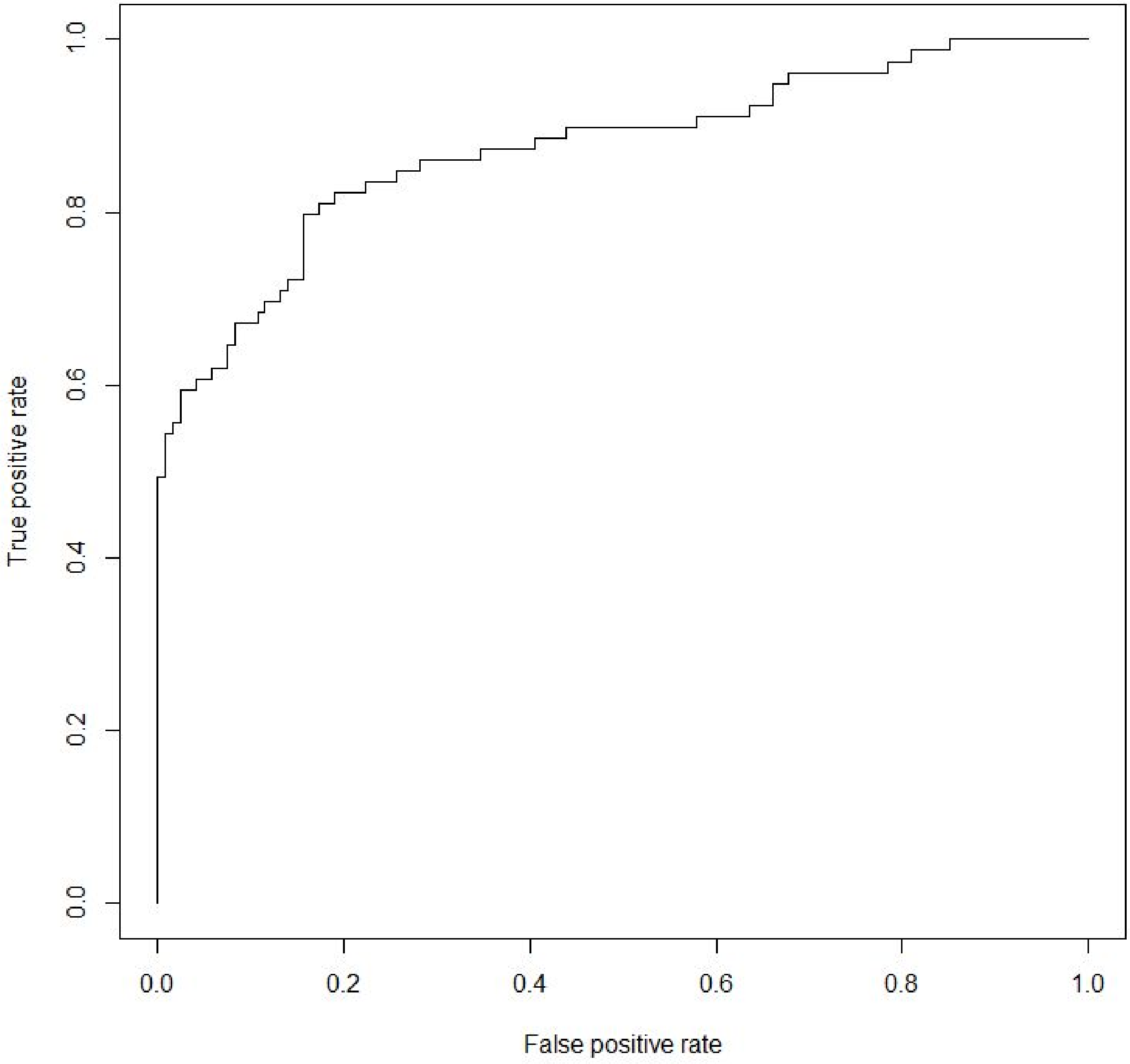


After running the predict function on the cross-validated LASSO model, we can see that Pclass and Sex are features with non-zero coefficients which affect the data most. This agrees with the lasso.model coefficient plot about the importance of these two features.



```
> head(traindata)
  Survived Pclass    Sex Age    Fare
1        0      3   male  22  7.2500
2        1      1 female  38 71.2833
3        1      3 female  26  7.9250
4        1      1 female  35 53.1000
7        0      1   male  54 51.8625
9        1      3 female  27 11.1333
> pred.lasso
5 x 1 sparse Matrix of class "dgCMatrix"
                       1
(Intercept)  6.860779294
Pclass      -1.175213796
Sex         -2.254464676
Age         -0.035990490
Fare         0.001125199
```

Plot of the prediction of testdata based on the cross-validated model. We can see that the ROC-curve performance is fairly good. The curve approaches great prediction results since the true positive rate is high and the false positive rate is low. This can be seen because the curve nears the top-left corner of the plot. If the model is improved more, the curve would approach the top-left corner indicating a better way to predict observations.


True positive rate
False positive rate