Joseph Bieseslin

Comparing LASSO Regression with Ridge Regression & Elastic Net - Titanic dataset

December 14, 2016

Table of Contents

Generic Project Information

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.

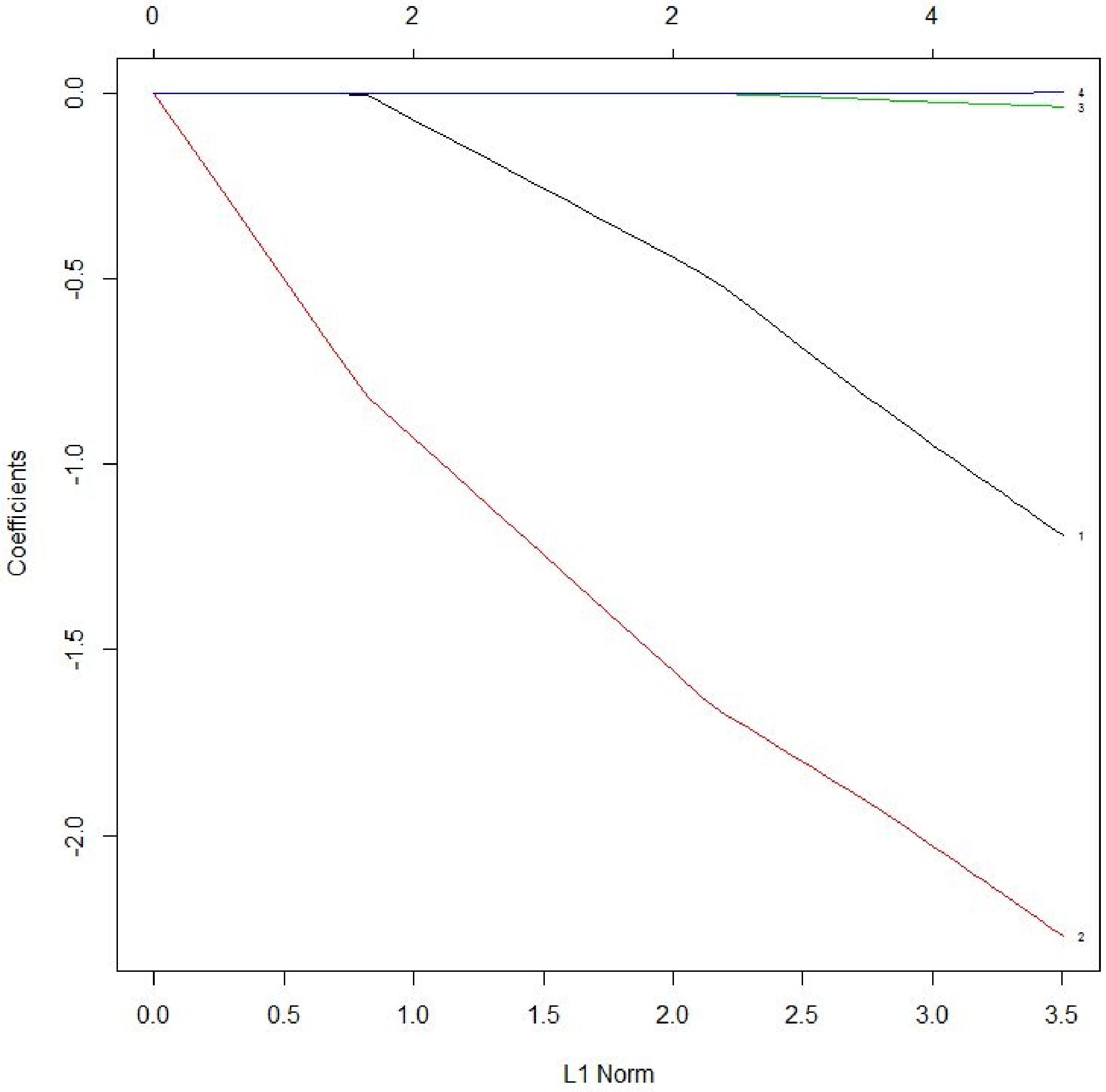
The focus of this project is to compare a model built using LASSO and compare it with Ridge regression and Elastic Net. Elastic Net supposedly always performs better than the other two because it combines the methods that both LASSO and Ridge perform into one algorithm. Therefore, simply changing to the Elastic Net algorithm should improve the model.

Lasso Regression

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 - basic model

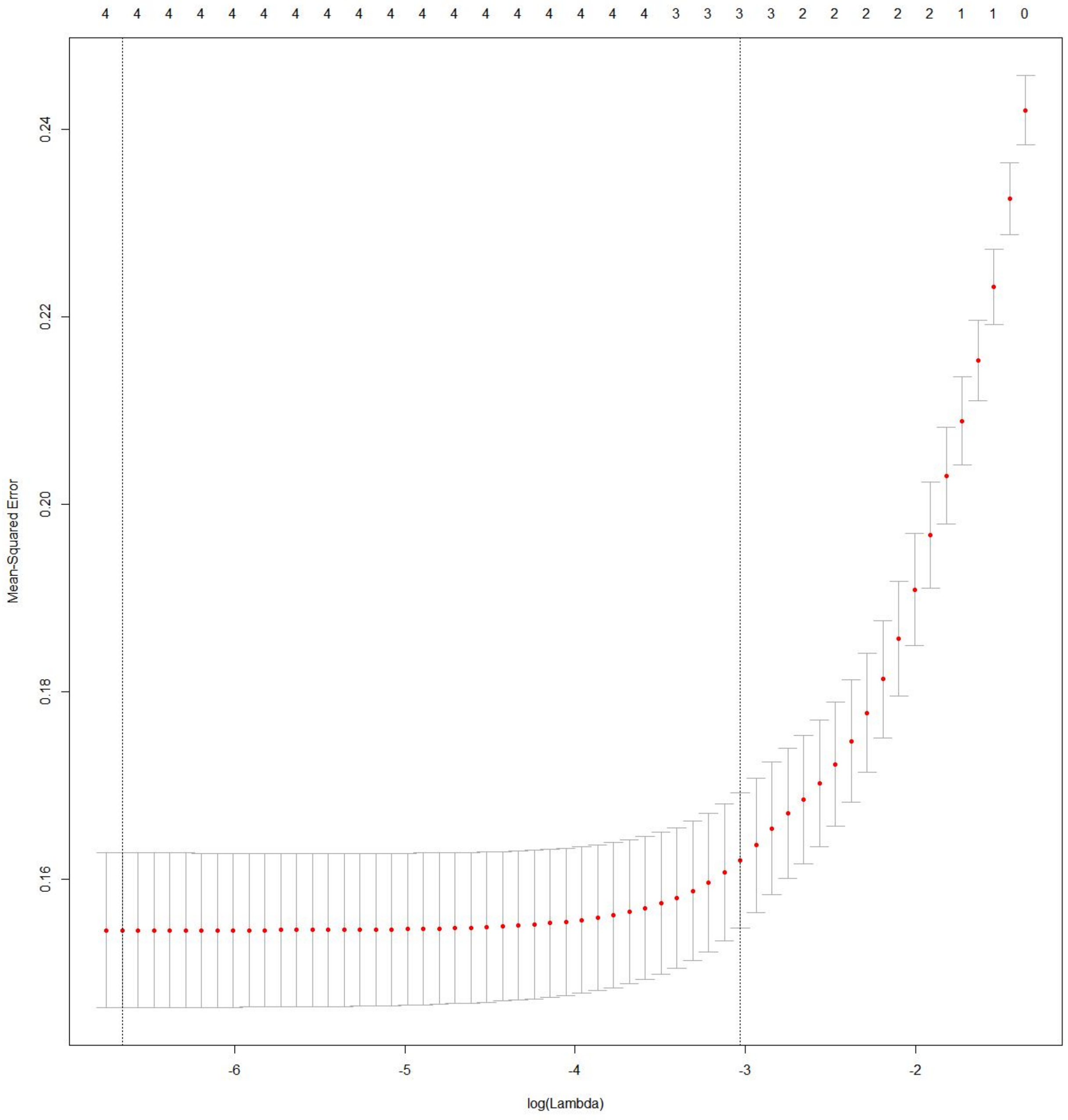
LASSO is useful for feature selection by setting variables coefficients to 0 to determine features with less importance. The following plot shows that features 3 & 4 which correspond to Age & Fare in the Titanic dataset do not have as much importance as Plcass & Sex (features 1 & 2).



Plot of cv.lasso.model: cross-validated model

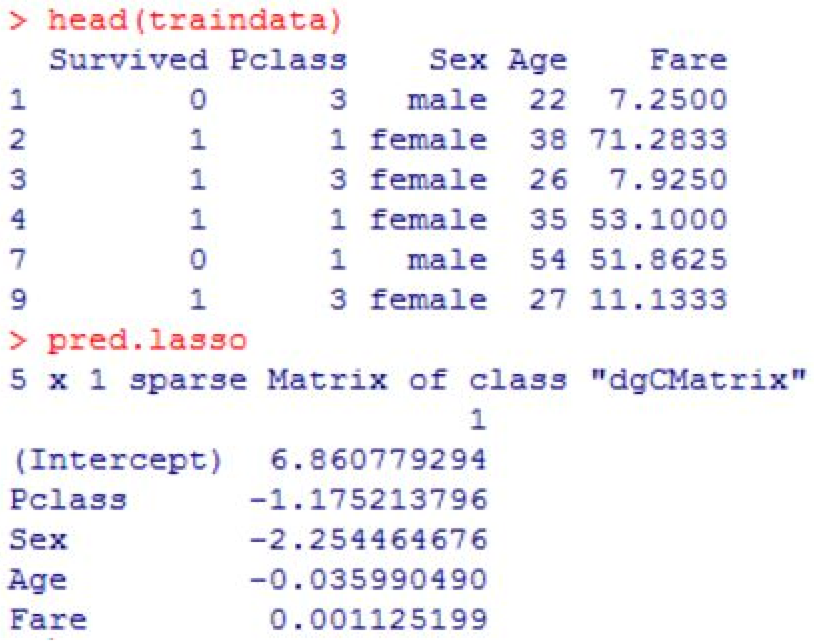
The cross-validated model uses n=10. The model is looking for an optimal lambda to size coefficients such that there is not too much complexity. As the size of a coefficient for a variable increases, the model is placing more emphasis on that variable. When a model is too complex and a coefficient becomes 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.



Prediction data

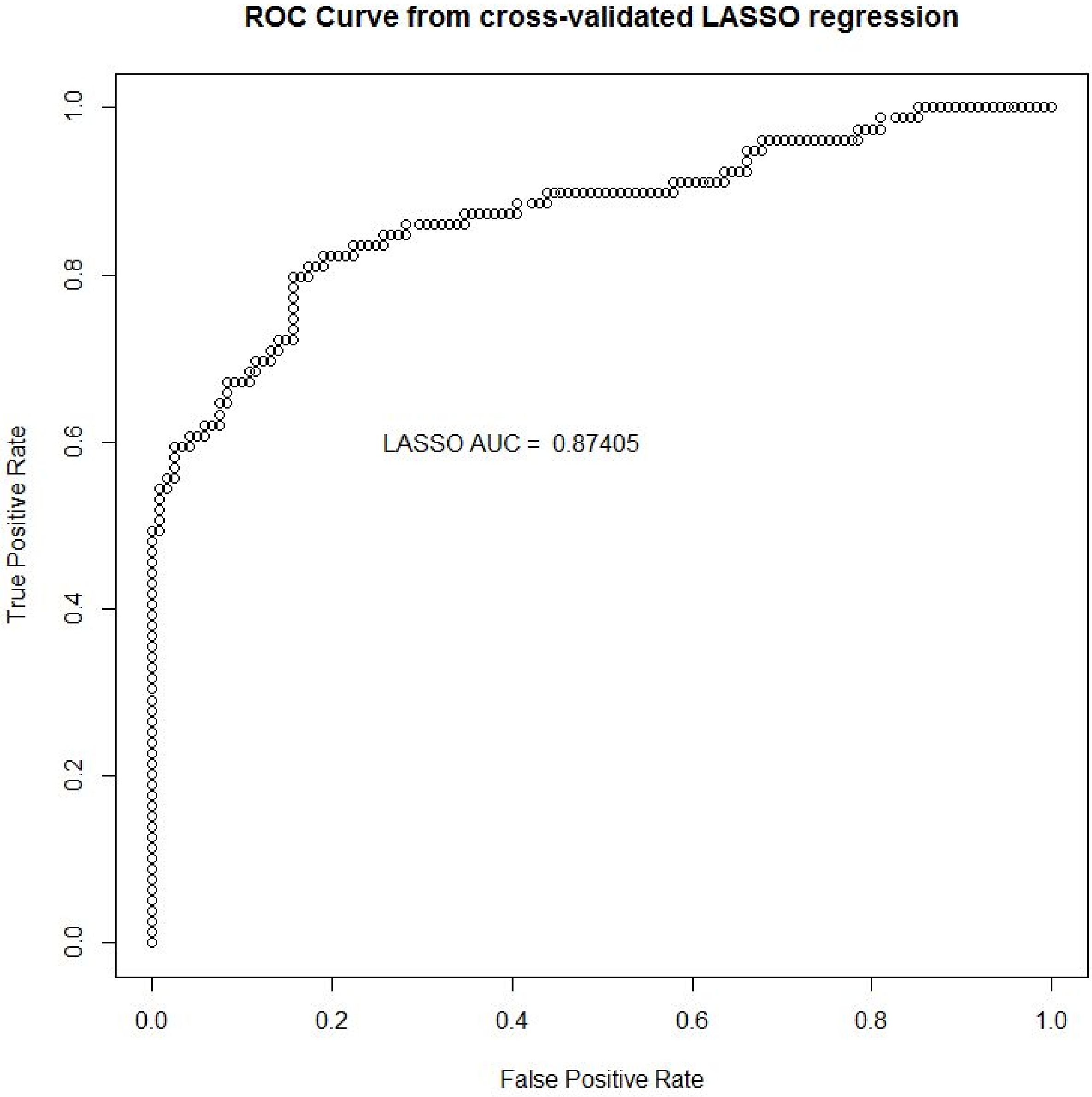
After running the predict function on the cross-validated LASSO model, we can see that Pclass and Sex are features with non-zero coefficients which affect the data most. This agrees with the lasso.model coefficient plot about the importance of these two features.



```
> 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
```

ROC-Curve with AUC

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.



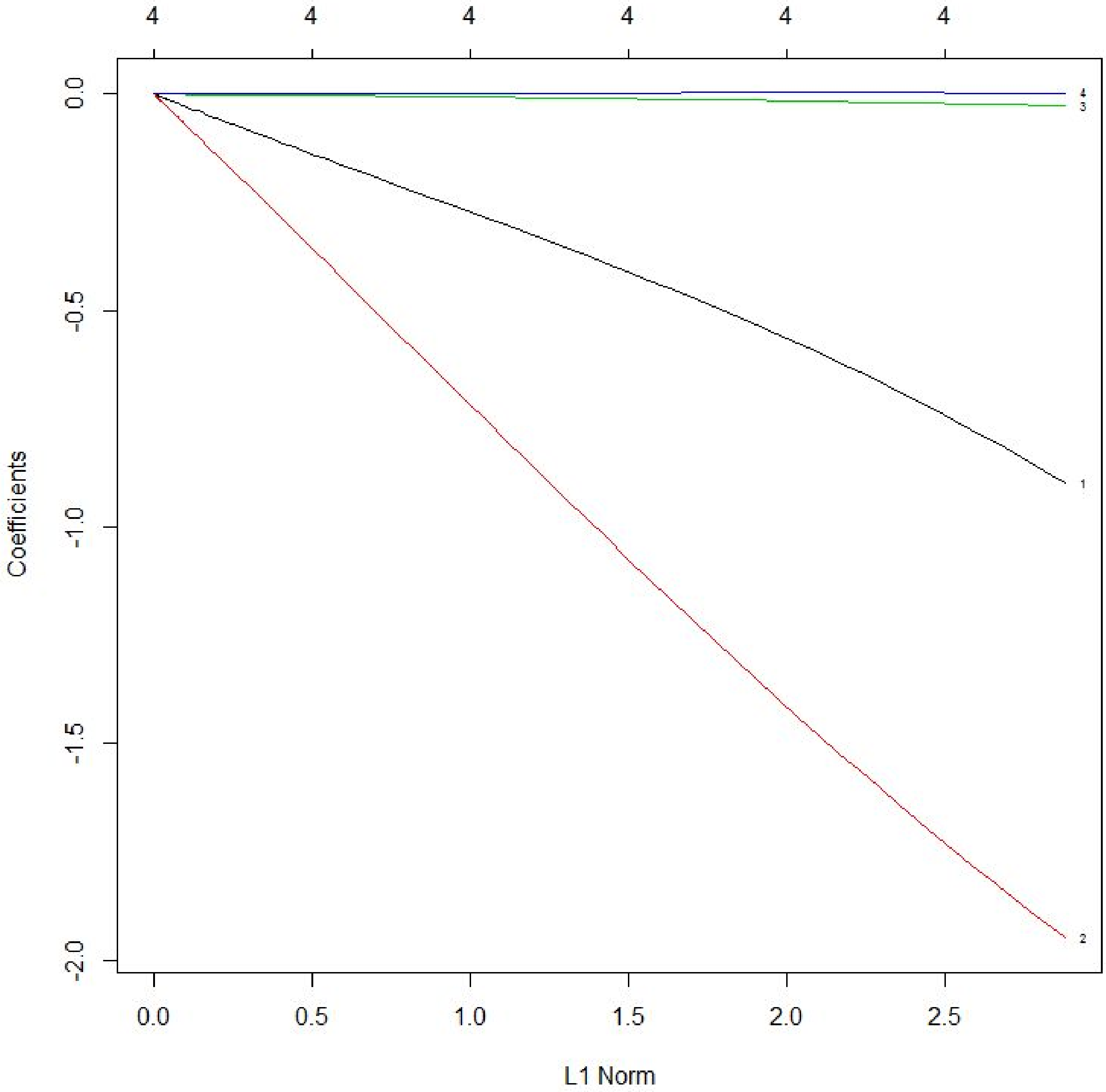
Ridge Regression

Ridge regression is used for similar reasons to that of LASSO, which is why I chose to compare the two algorithms. Ridge regression is also used to reduce a set of many fields to some key fields that have the most effect on the dataset. As was done with LASSO, a cross-validated model will be used for prediction. All steps from the LASSO data analysis will be replicated for Ridge and a plot of the ROC curve will be displayed.

Again analysis was done by trying to classify whether or not someone would have survived the Titanic shipwreck. The same training data was used for the Ridge model as the LASSO model

Plot of ridge.model - basic model

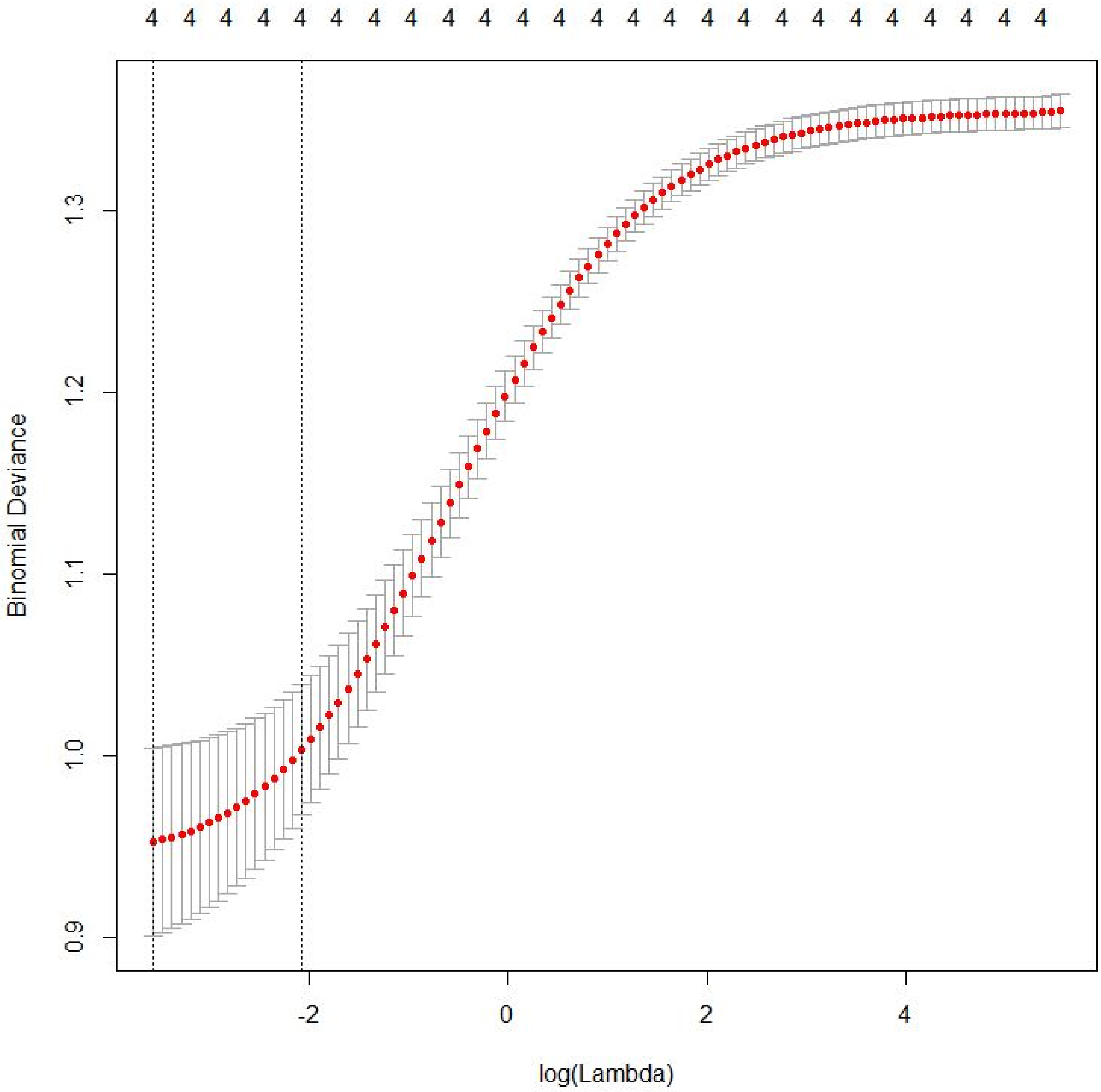
Ridge is useful for feature selection by setting variables coefficients to 0 to determine features with less importance. It does this by adding a penalty term to reduce larger coefficients which increases model complexity (and model complexity on training data leads to overfitting). The following plot 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.ridge.model: cross-validated model

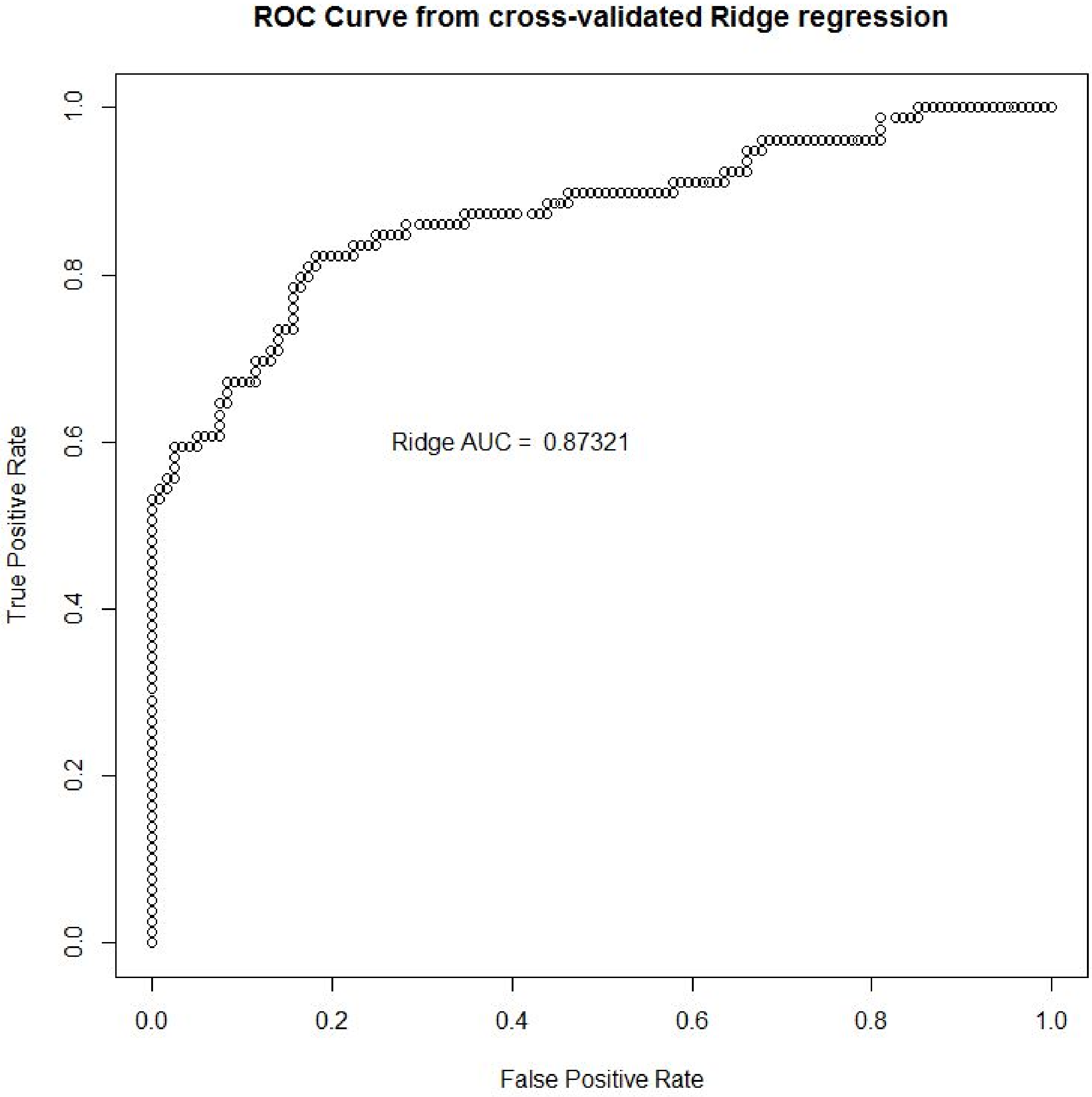
The cross-validated model uses n=10. The model is looking for an optimal lambda to size coefficients such that there is not too much complexity. As the size of a coefficient for a variable increases, the model is placing more emphasis on that variable. When a model is too complex and a coefficient becomes 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.



ROC-Curve with AUC

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.



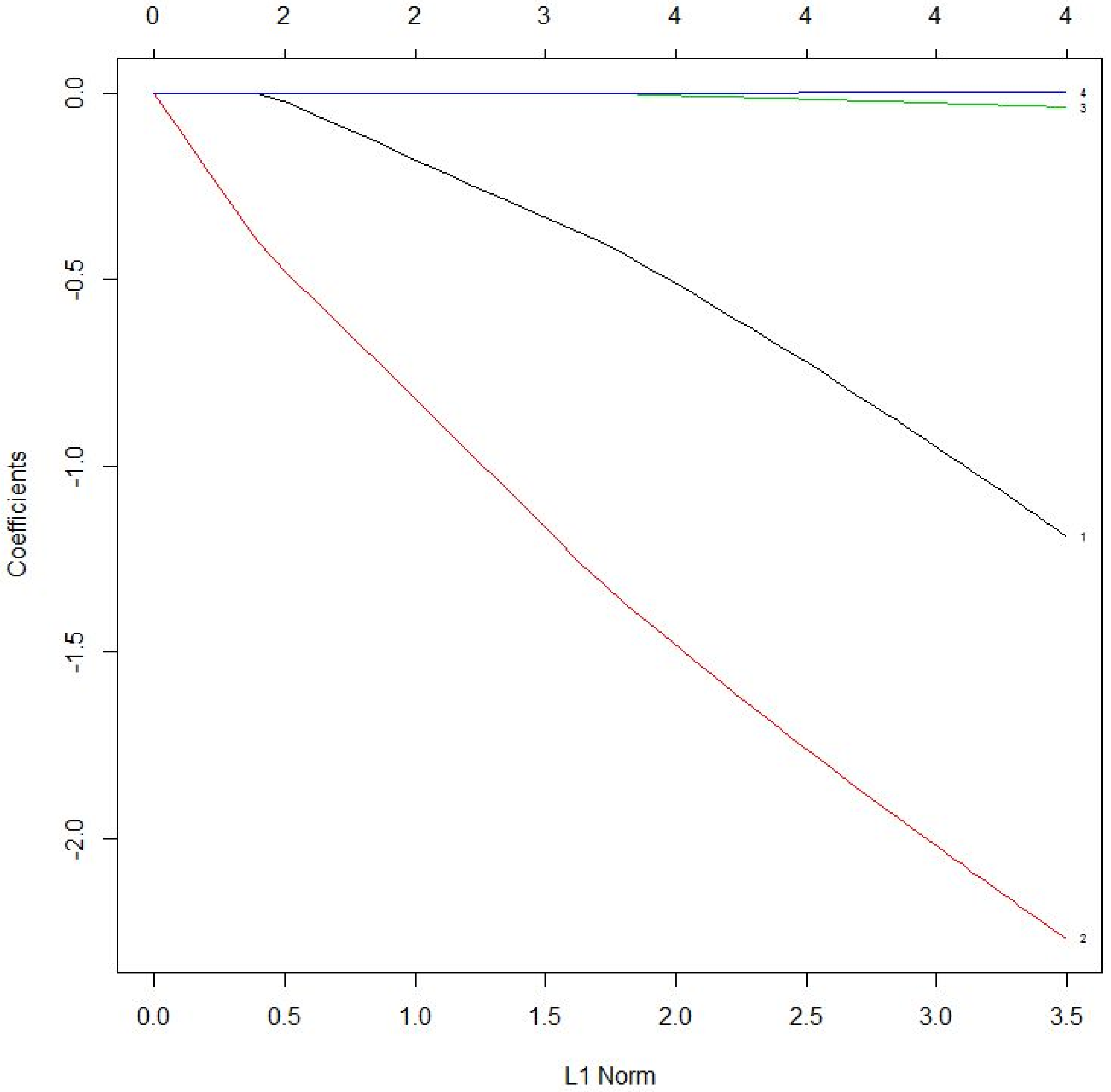
Elastic Net Regression

Elastic Net is used for similar reasons to that of LASSO and Ridge Regression, which is why I chose to compare the three algorithms. Elastic Net is supposedly the best of the three algorithms and always outperforms the other two. This is because Elastic Net combines the error correction analysis done by LASSO and Ridge and forms the best model from it. I expect the original model to improve because of this, and it is the main form of improvement for this project.

Again (as was the case with Ridge) analysis was done by trying to classify whether or not someone would have survived the Titanic shipwreck. The same training data was used for the Elastic Net, Ridge, and LASSO model.

Plot of elasticnet.model - basic model

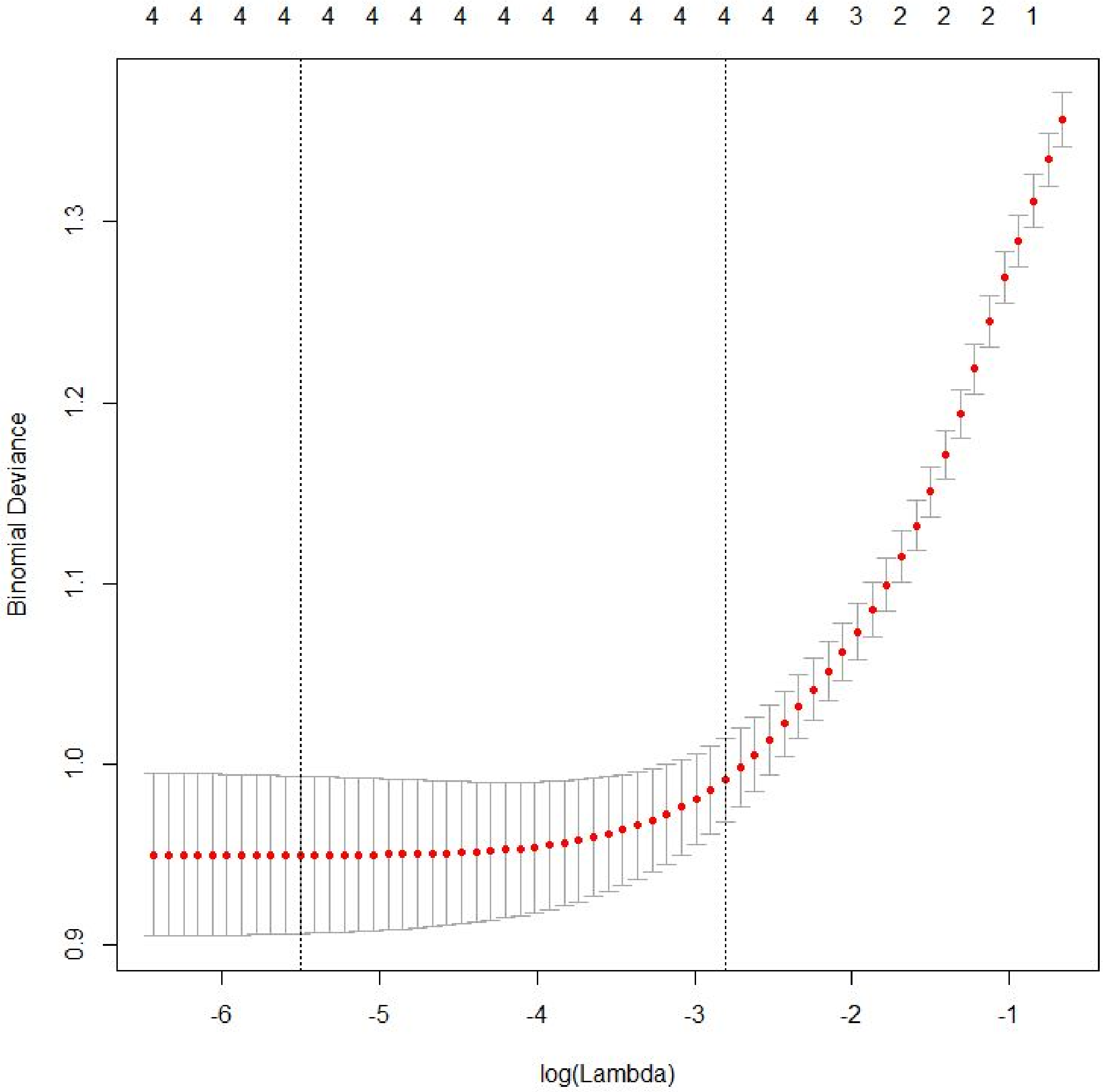
Elastic Net is useful for feature selection by setting variables coefficients to 0 to determine features with less importance. The following plot once again shows that features 3 & 4 which correspond to Age & Fare in the Titanic dataset do not have as much importance as Plcass & Sex (features 1 & 2). This is the same as LASSO and Ridge which is good since results should be similar.



Plot of cv.elasticnet.model: cross-validated model

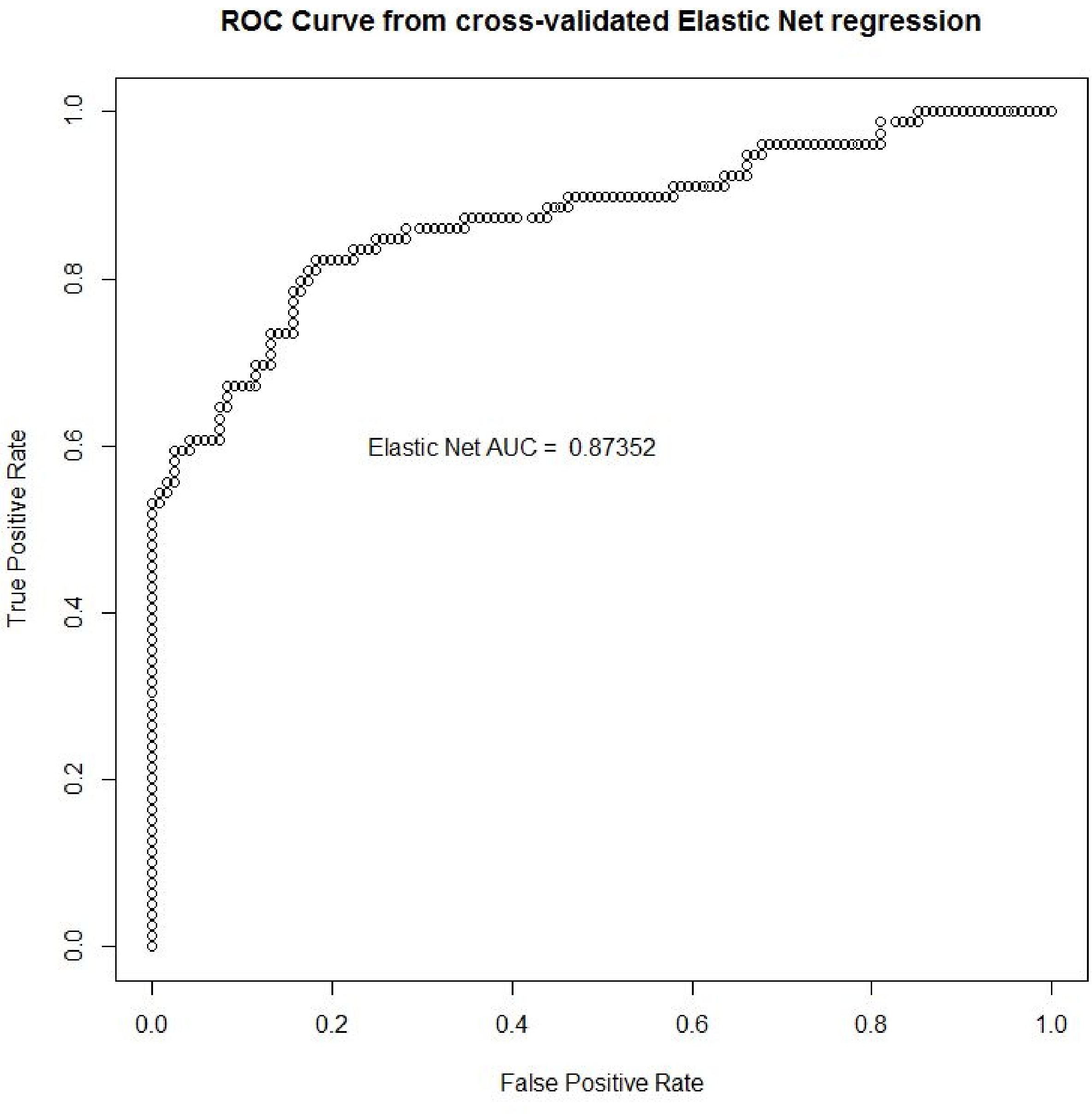
The cross-validated model uses n=10. The model is looking for an optimal lambda to size coefficients such that there is not too much complexity. As the size of a coefficient for a variable increases, the model is placing more emphasis on that variable. When a model is too complex and a coefficient becomes 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.



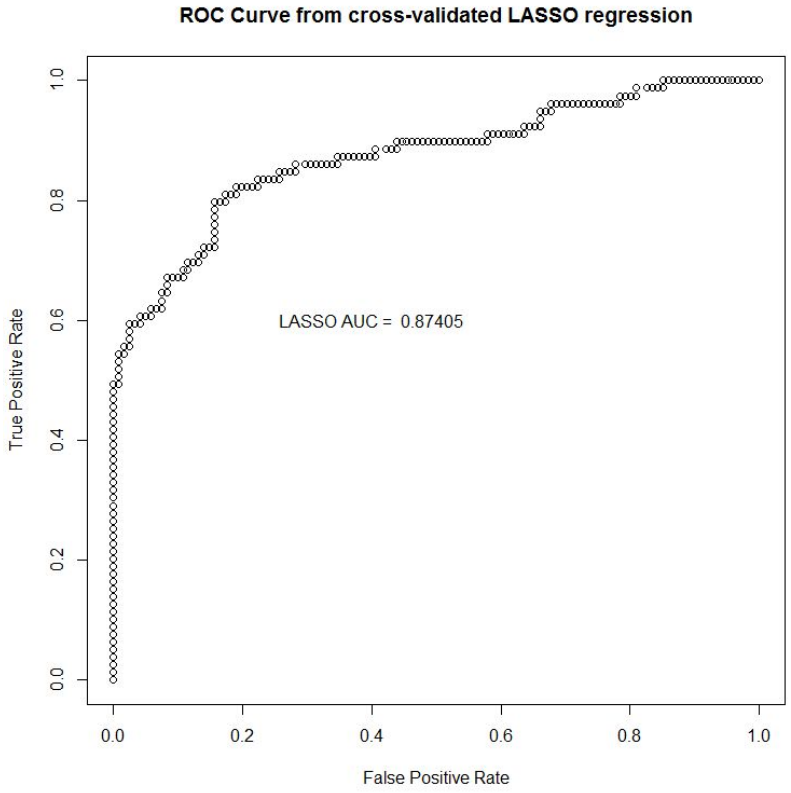
ROC-Curve with AUC

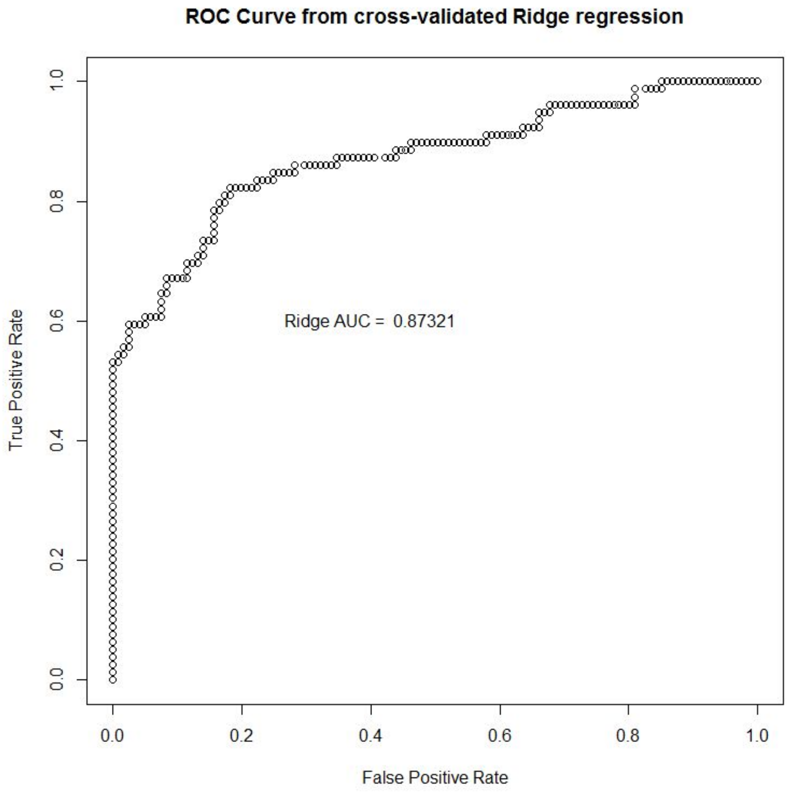
Plot of the prediction of testdata based on the cross-validated model. We can see that the ROC-curve performance is fairly good. The curve approaches great prediction results since the true positive rate is high and the false positive rate is low. This can be seen because the curve nears the top-left 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.



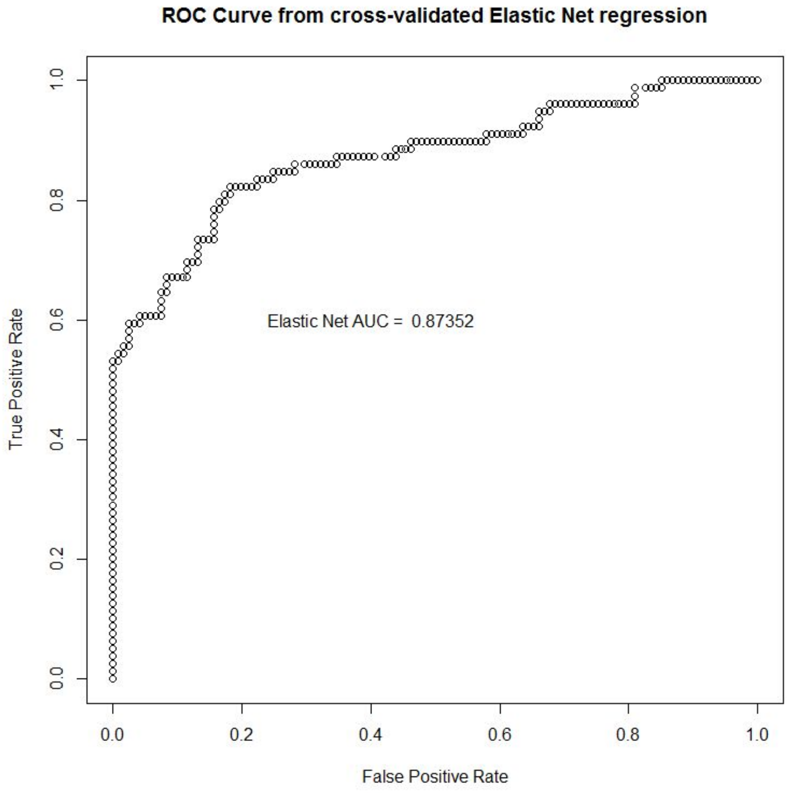
Comparison of Results

Graphical Results





ROC Curve from cross-validated Ridge regression



ROC Curve from cross-validated Elastic Net regression

Textual Analysis

From the plots of the ROC curves, the result is that LASSO performs the best out of the three algorithms, with Elastic Net performing second best. From this, the initial assumption that Elastic Net always performs better than LASSO and Ridge regression was not found to be true. However, all the results were within a thousandth of a percent of each other. LASSO's model may have performed better because of the lambda value that was selected from its cross-validated model.

This problem may also not have been perfectly suited for these regression algorithms. Each of these three algorithms is used to determine which features in a dataset introduce more error due to large coefficient values. This happens because of collinearity in predictor variables. In the dataset used, there are only 4 predictor variables for the 1 variable the model is trying to predict. With very few predictor variables and a relatively small sample size, it may be likely that extra model tuning for each of the algorithms would not bring that much change to the results.