STK-IN4300 Project 2

Steinn Hauser Magnusson

November 16, 2019

Exercise 1

1.

Section one is mostly about implementing some code. This has been done in the first function of the 'Exercise1' class. The scaling is done typically, where the mean and standard deviations of the non-categorical variables should be zero and one, respectively. The main concern of the scaling is whether or not to one-hot encode the categorical variables. All the categorical variables presented in the data are binary, such that the one-hot encoding is not computationally efficient. If these columns were to be one hot encoded, then we would be left with two columns which are highly correlated, as one column could be derived from the other. One-hot encoding of the binary categorical data is therefore not done.

2.

Following are figures illustrating the Covariance and Correlation matrices of the data. A specific figure is also generated of the 'FFVC' column. Figure 1 illustrates the covariance matrix of the data:

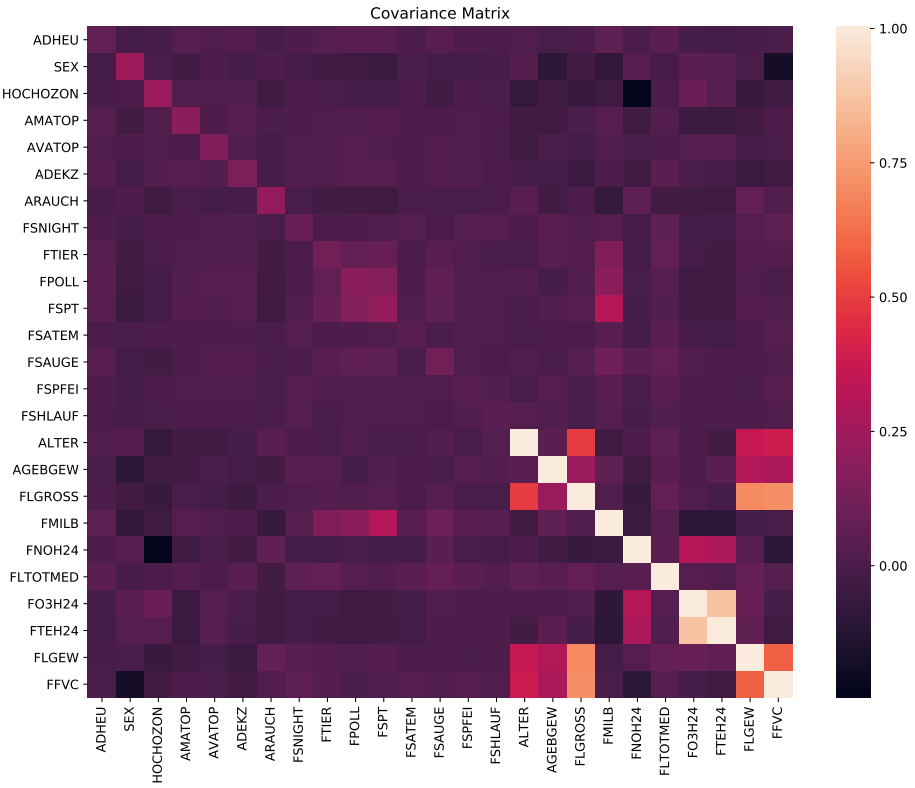


Figure 1: The covariance matrix of the data

Figure 2 illustrates the covariance vector of the FFVC feature:

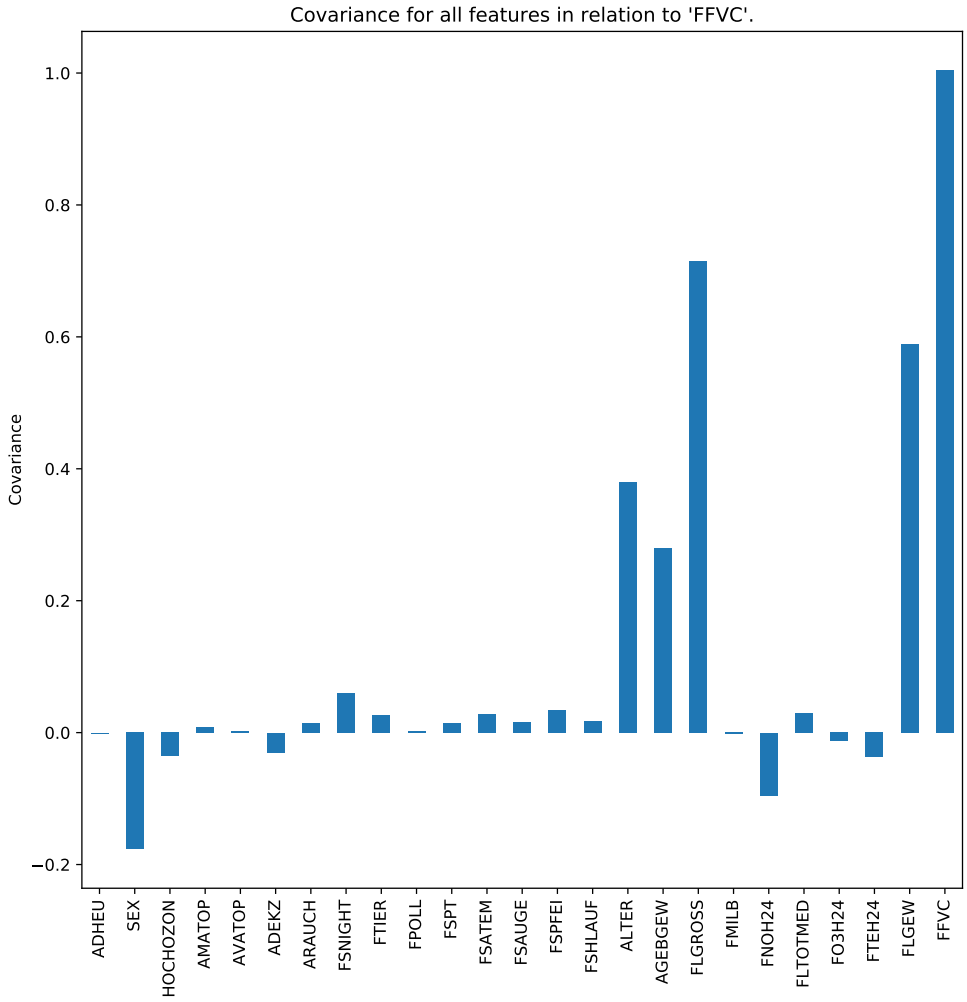


Figure 2: The specific covariance of the FFVC features.

The covariate which has the strongest association with the forced vital capacity (FFVC) is the FLGROSS feature. Following is a report on the coefficient estimates, their standard error and the associated p-values:

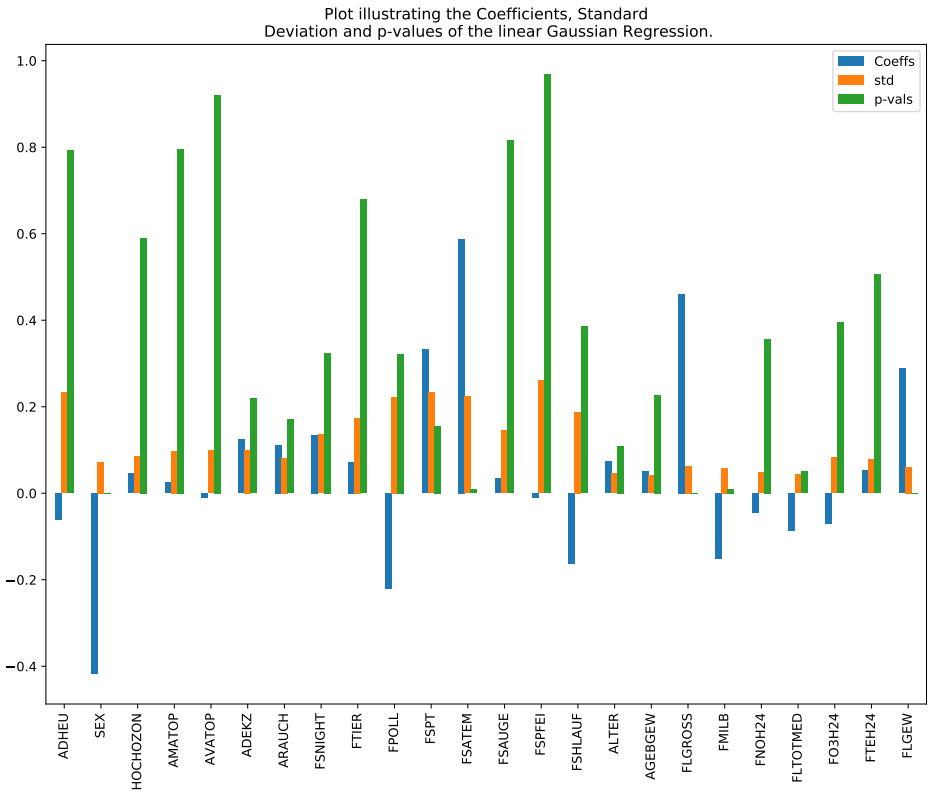


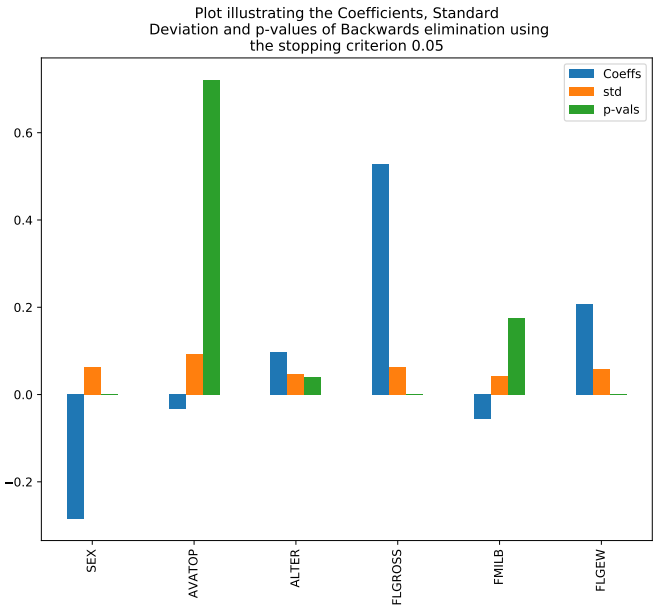
Figure 3: Figure illustrating the coefficients, standard deviations and p-values of all the features.

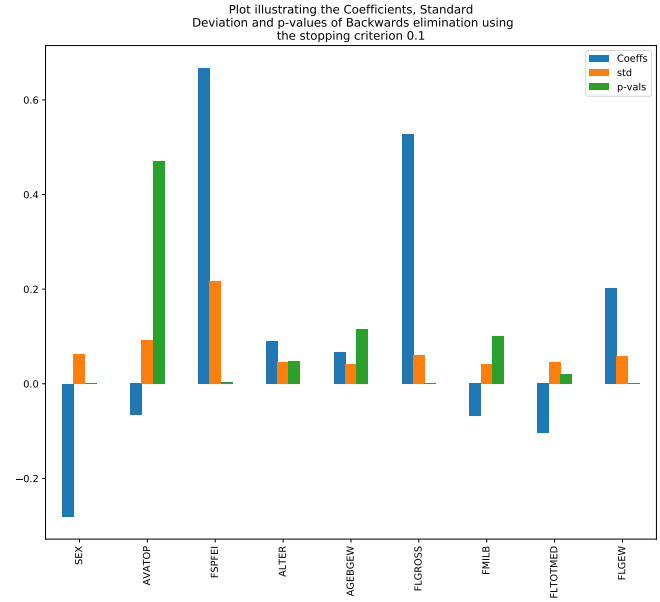
It is clear that several of the p-values are quite large in relation to the typical p-value standard of ~ 0.05. It is also quite interesting to see in one figure the relation between the coefficient size and the p-value. It is clear to see for the 'SEX' and 'FSATEM' columns that the bigger the regression coefficient, the smaller the p-value of the feature. This can also be seen in the 'FLGROSS' column, which we remember from the covariance matrix as having the strongest association with the 'FFVC'. The opposite is also noticeable; the largest p-values, namely the 'FSPEI' and 'AVATOP' columns have coefficient values close to zero.

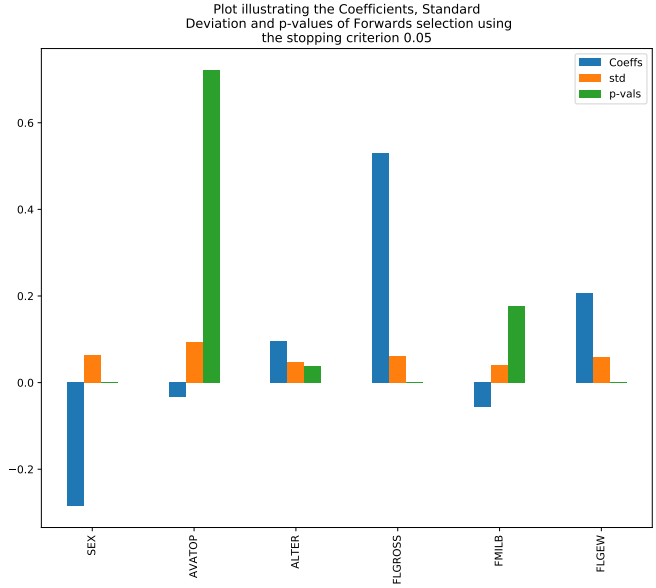
Some features do not exhibit this trend, such as the 'FSPT' and 'FPOLL' features exist somewhere in between, where they are not quite relevant to the prediction, but not quite irrelevant either.

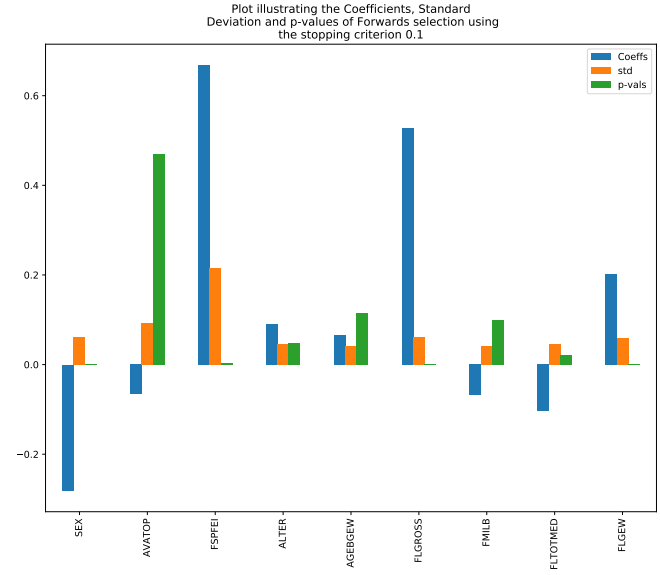
3.

The backwards and forwards selection algorithms performed using two different stopping criterion: One was 0.05 and the other was 0.1. Following are four figures of all these cases, where the coefficients, standard deviations, and all the p-values are listed in the same way as before.









These are all quite similar predictions, as the backwards elimination and forwards selection essentially perform the same operation. They simply set a limit to the p-values, and filter out accordingly.

These methods are implemented in order to build a high quality regression model with little to no unnecessary features. This is done in a way that hopefully does not compromise the predictive abilities of the model, though some line must be drawn on this subject. Even features such as the 'FTIER' feature (see figure 3) helped in predicting the model somewhat. Although it was a feature that seemed to be quite linearly dependent, it still had some use. Removing this feature is not severely jeopardizing to the model's prediction, though some accuracy is lost. I therefore hypothesize that the mean-squared error will overall increase after the backwards elimination and forwards selection techniques.

Note, that when reproducing the following figures, there is a large amount of stochasticity in the data shuffling method.

This means that the outcomes sometimes differ quite largely from one and other, so the figures illustrated in this report are simply one of many different cases.

4.

Use both a bootstrap and k-fold CV method to find the best (in terms of deviance minimization) complexity parameter of a lasso regression. Following is the bootstrap method results. The data was split up into 50/50, and the training data was further split up into 25% testing data for the bootstrap samples. The R^2 scores were calculated using the average of a lot of bootstrap samples, for multiple hyperparameter α values. Figure 4 illustrates the results of this analysis:

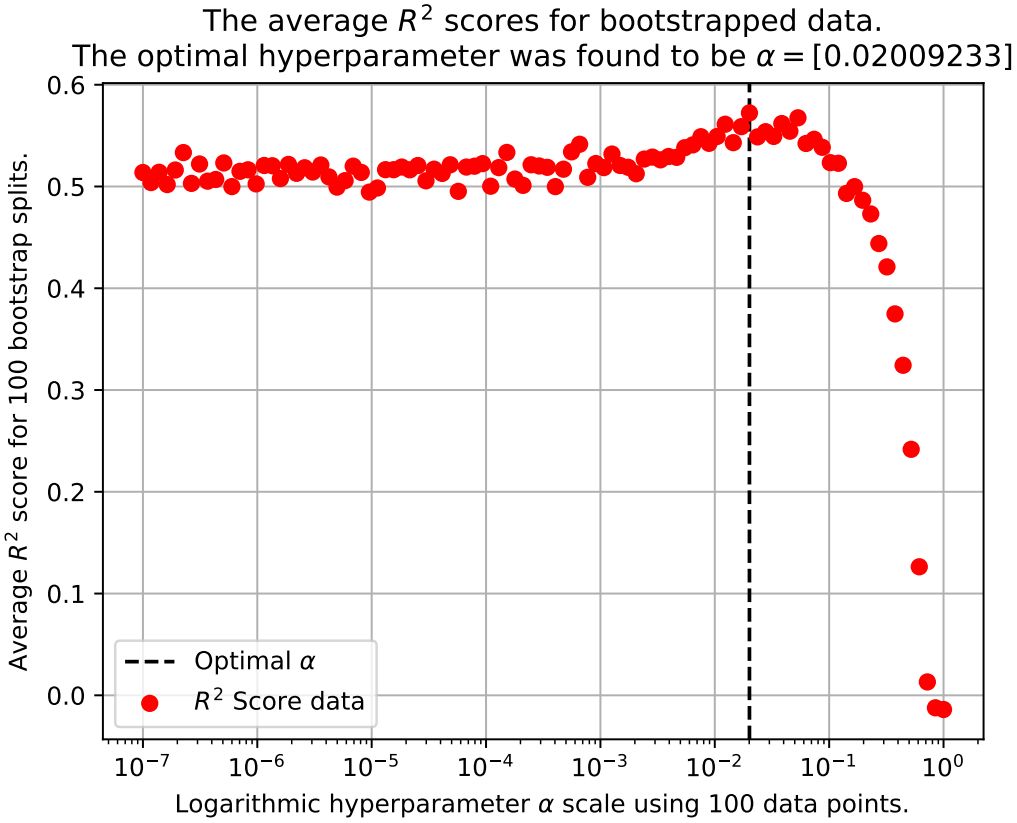


Figure 4: Figure illustrating the averages of the R^2 scores of multiple bootstrap samples. The optimal α parameter is found to be $\alpha = 0.02$.

Doing the same for 5-fold CV, the two optimal R^2 scores are generated:

Table 1: Maximum R^2 scores and minimum MSE scores of the two methods.

Scheme	Bootstrap	5-fold CV
R^2	0.64222	0.56684

5.

Following are the results from the GAM simulations:

Only linear terms:

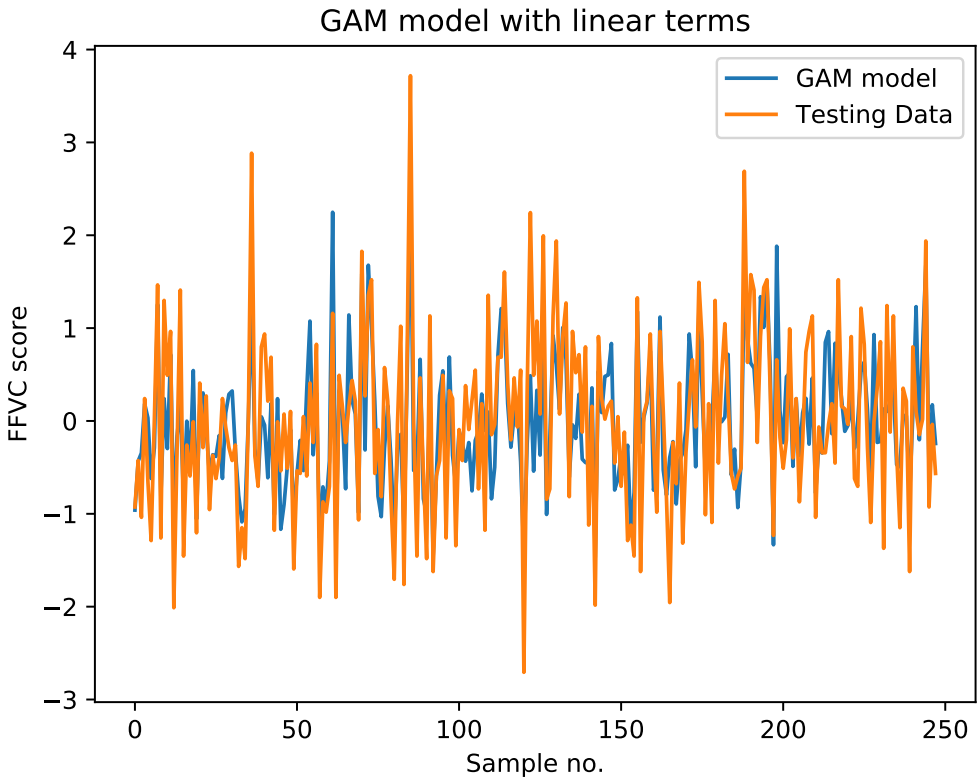


Figure 5: Figure illustrating the true FFVC scores vs the predicted scores using a GAM with only linear terms.

Spline terms allowed:

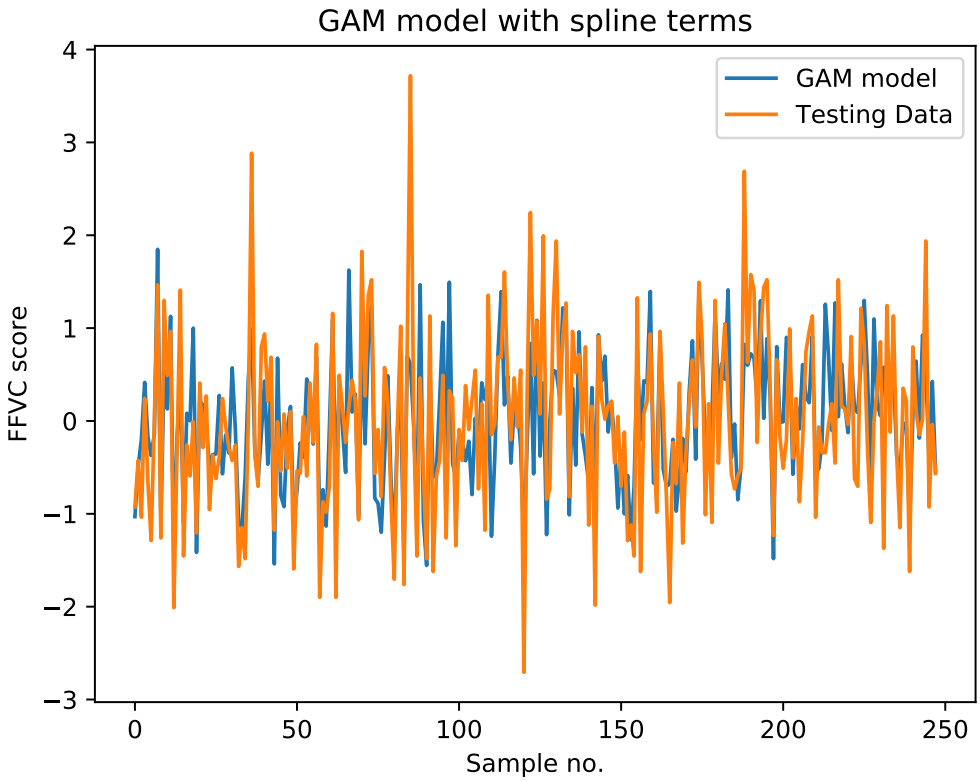


Figure 6: Figure illustrating the true FFVC scores vs the predicted scores using a GAM with splined, linear and polynomial terms.

The 'only linear' GAM produced an MSE of 0.43756327

The 'splines and polynomial degrees allowed' GAM produced an MSE of 0.45348646.

These results are very chance-dependent, but in general the results are not too far off.

6.

7.

Exercise 2

1.

Figure 7 illustrates the cumulative gains curve generated by the k-Nearest Neighbor method using $k = 10$ and 33% testing data.

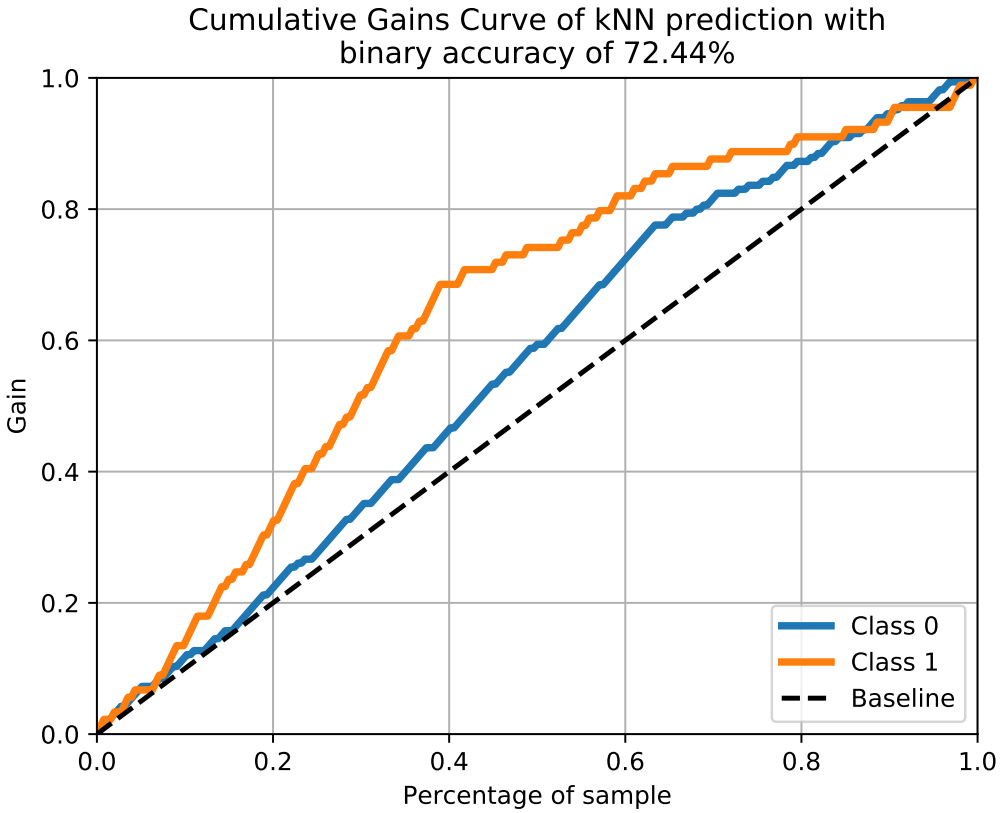


Figure 7: Figure illustrating the averages of the R^2 scores of multiple bootstrap samples. The optimal α parameter is found to be $\alpha = 0.02$.

Figure 8 illustrates an analysis of which k value produces the optimal score:

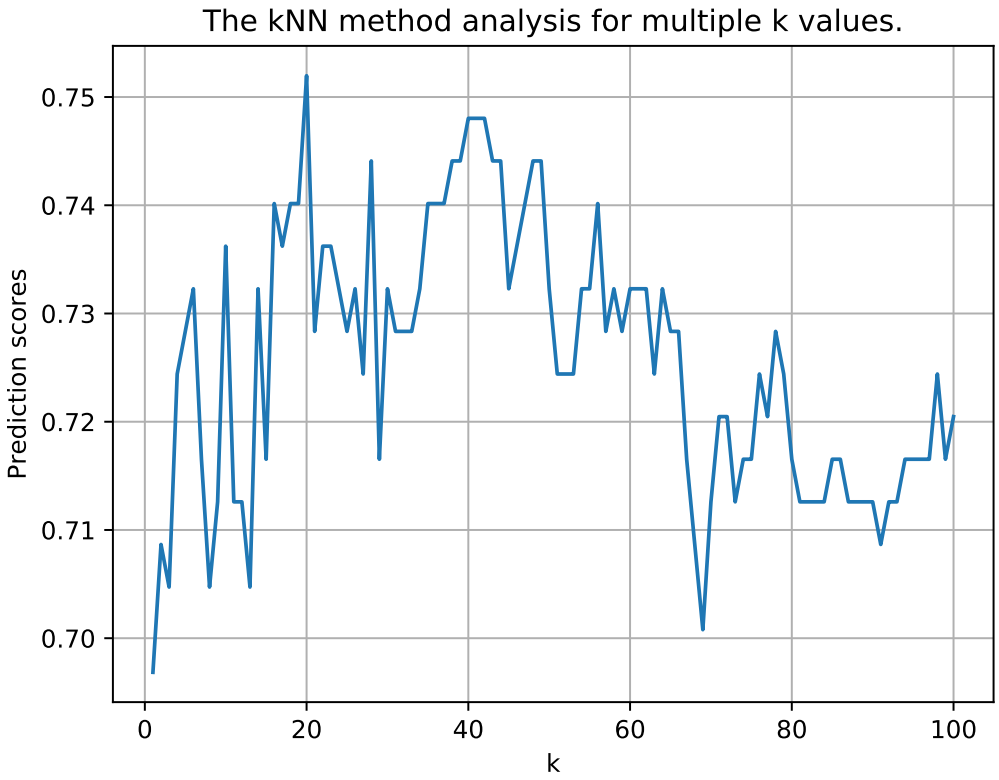


Figure 8: Figure illustrating the averages of the R^2 scores of multiple k values.

2.

Figure 9 illustrates the prediction results for the GAM method using splines only.

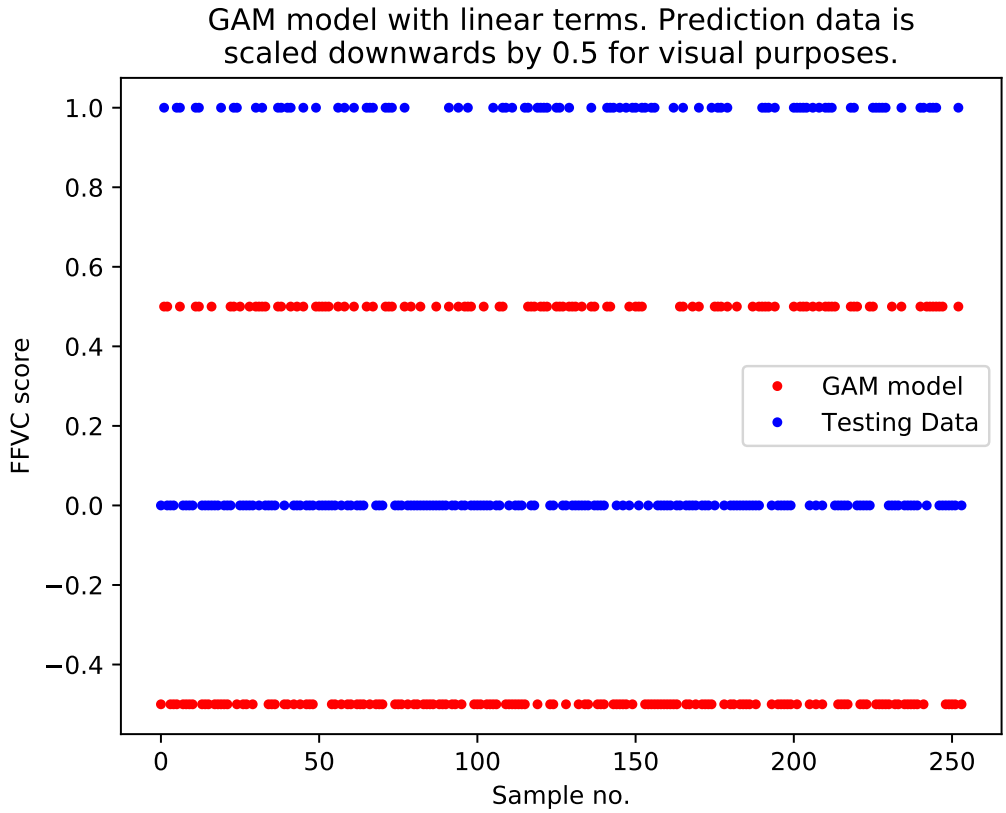


Figure 9: GAM results for exercise 2.

3.

This was implemented into the python program using sklearn's functionalities. Bagging with both voting and averaging was unfortunately not accomplished, however.

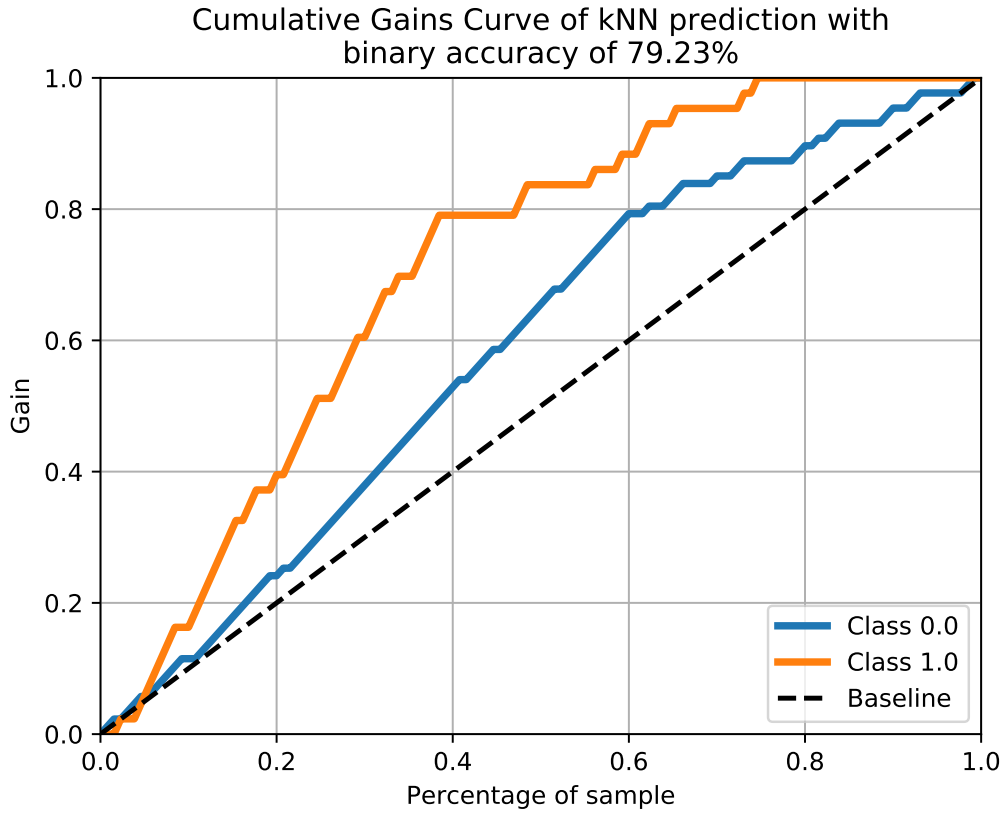
4.

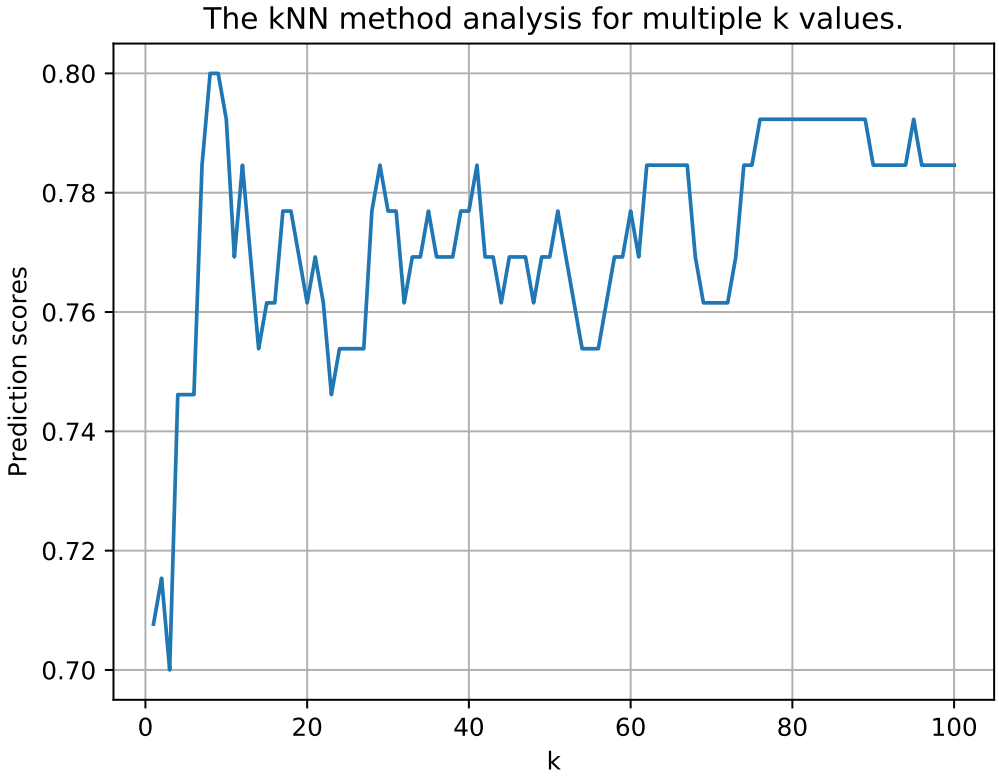
Decision Tree score: 0.6654 Bagging score: 0.7244 Random Forest score: 0.7244 Neural Network score: 0.7047 ADA Boost score: 0.7047

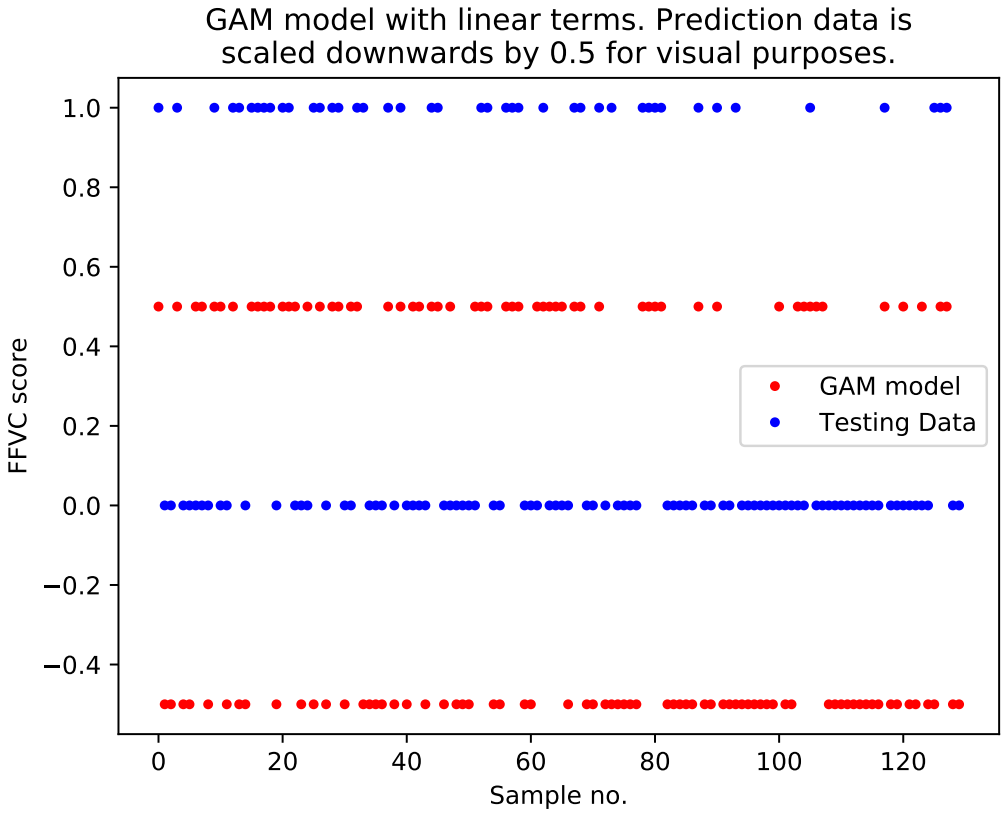
If I were to choose between these models, I would choose the Neural Network method. This
This choice is due to the ...

5.

Following are the new figures, just with the outliers removed. The figures illustrate the same phenomenon as before:


Cumulative Gains Curve of kNN prediction with
binary accuracy of 79.23%
Gain
Percentage of sample
Class 0.0
Class 1.0
Baseline
0.0
0.2
0.4
0.6
0.8
1.0


The kNN method analysis for multiple k values.
Prediction scores
k
0.70
0.72
0.74
0.76
0.78
0.80
0
20
40
60
80
100



Following are the results again:

Table 2: Scores for all the methods with outliers removed.

Scheme	Score
Decision Tree score:	0.6692
Bagging score:	0.7385
Random Forest score:	0.7923
Neural Network score:	0.7462
ADA Boost score:	0.7385

References

McLeod, S. A. (2019, May 20). What a p-value Tells You About Statistical significance. Simply Psychology. https://www.simplypsych
value.html