**CE 475 COURSE PROJECT REPORT**

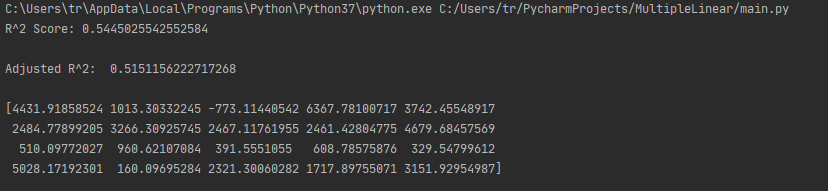
**Student Name/Surname:** Ali Efe BARUŞ

**Student Number:**20180601008

**1-Multiple Linear Regression**

**Methodology:** In this trial, I decided to try Multiple Linear Regression which assumes there is a simple relationship between all data points. I try this approach even I know the between all data points are reliable and multiple linear regression has a low performance but I want to create comparable “R^2” and “Adjusted R^2” scores and predicted y values.

**Results for the first Trial (Predicted Y Values and R^2 Scores):**



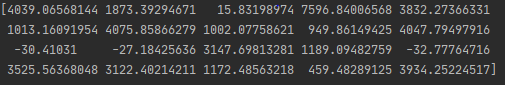
**2-Random Forest:**

**Methodology:** I decided to test all of my data as both train and test and I implement 2 random forests which distinguished by their max-features which I learned from Sci-Kit Learn’s random forest regressor documentation. I used the Auto feature and the Sqrt Feature (Also mentioned in our course slides as the best choice) because I try other features and I get the two highest performances in these results. I choose max-depth 6 because in this model we have 6 dimensions. However, the results of this trial surprised me because the R^2 scores are way much higher than I expected.

**Results for Second Trial (R^2 Scores):**



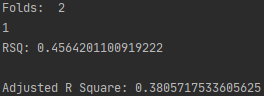
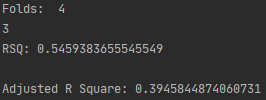
**Predicted Y-Values:**

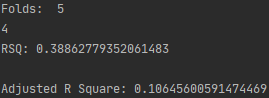
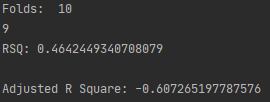


When I see these results, I am shocked because it is too high comparing to Multiple Linear Regression. I clearly understand that I cannot trust these results because it is unreliable due to the lack of implementation of validation techniques

**3-Adding K-fold Cross-Validation:**

**Methodology:** Due to the results of the first trial, I decided to add K-fold cross-validation to my model to separate my data as train and test data.

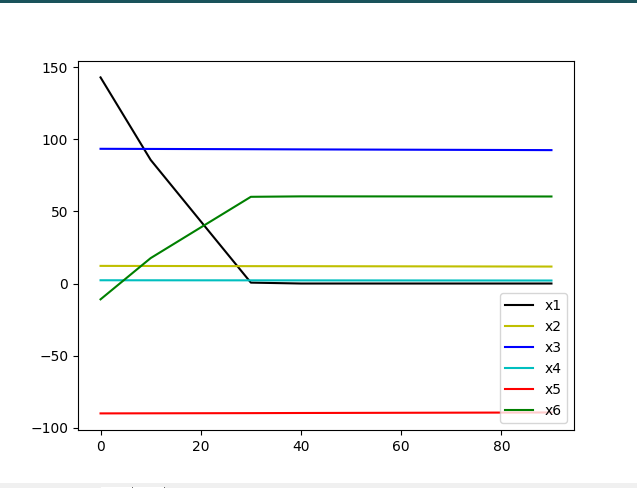
 ,

I decided to split and test my data on 2-4-5 and 10 folds. According to the results, 4-fold gives the best solution but adjusted R^2 scores slightly decrease due to loss of some training data. 2 folds gives this result because our test data is insufficient to test for our training data.5 folds is insufficient because I get better results when I split data into 4 folds. When trying to split data into 10 folds our training data is not enough to teach and predict the model. The common thing in every model is the minimum MSE value to choose the best fold for test data is the last fold for every trial.

**4- Lasso Regression:**

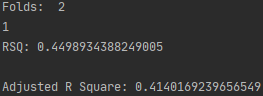
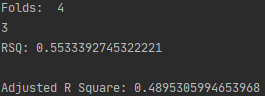
**Methodology:** In the last trial, getting a slight change does not increase our performance much but makes our prediction more reliable. However, our aim is increasing performance with reliability so I decided to go with the Lasso Regression which is introduced in our course because I am more familiar with the Lasso regression and when implementing Ridge Regression, the slope just gets closer to 0 and never shrink to 0 but in Lasso Regression the slope can shrink to 0. With lasso, I can eliminate some of these variables which shrinks to 0. I examine the library documentation of Sci-kit Learn to implement the Lasso Regression correctly

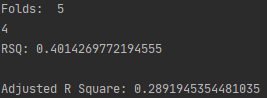
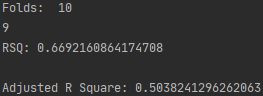


From these results, we can understand that x1, x2, x4 variables have no effect or minimal effect on the model. I decided to leave them and continue with x3, x5, x6.

**5-Dimension Reduction Test:**

**Methodology:** I decided to test the same model used in the third trial with three variables (x3, x5, x6) which is obtained in lasso regression. I test the new data with the same folds.

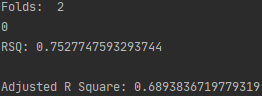
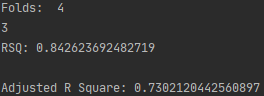
 

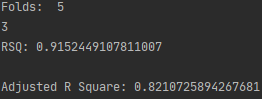
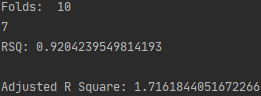
 

These outputs showed that Surprisingly,10-folds has the best performance in this trial because it has the worst performance in the last trial. The performance of other folds has increased. 4-folds is the second powerful choice for the data. Even an increase in the performance of 2 and 5 folds is observed there is not enough increase in choosing these folds. The best fold for the test data is still the same for all of the folds which is the last fold for every trial.

**6-Creating Polynomial Model:**

**Methodology:** According to a slide in our course slides (Course slides, nonlinear.pdf, Slides (6-7)), we can multiply the dimensions to create new dimensions and exclude coefficients to create another polynomial regression model and treat the model as a multiple linear regression model. In this trial, I use this information and I calculate the square of (x3+x5+x6+1) identity and exclude the coefficients of each variable to increase dimensionality and fit the data to a polynomial model. Also, I add (x3\*x5\*x6) to adding a new dimension and I test it the same multiple linear regression model with k-fold cross-validation.

According to these results, 5 folds are the perfect fold decision for this model. The perfect fold for 5-folds is the third fold which corresponds to from 60 to 80 of the first 100 rows of all data because the maximum R^2 scores obtained when we use this data for the test data. We cannot use 10 folds because even obtained R^2 scores bigger than 5-folds’ scores, 10-folds’ scores show the model an overfit.

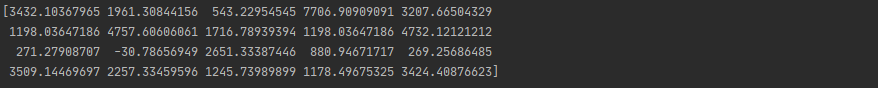
**7-Random Forest 2:**

**Methodology:**  I tried once more because I change my model to the polynomial model and I find the best fold for test data so I expected different results other than my second trial. For a better comparison, I choose my max depth, n estimators, and maximum feature same.

**Results for the last Trial (R^2 Scores):**



**Predicted Y-Values:**



As I expected this model gives less performance but I know that this model’s predictions are more reliable than other trial. Predicted Y-values obtained in this trial are the final form of the Y-values from 100 to 120 of data but they may vary among multiple runs because of the randomness.

**Implementation:**

**First Trial:** This trial was carried out because we want a comparable R^2 scores and Y predictions it is obtained with the help of Multiple Linear Regression.

**Used Library(s):** NumPy, Pandas

**Second Trial:** This trial has done for a find out a shortcut way with Random Forests however it helps to comparison the last result’s reliability also showed the importance of elimination, validation techniques for machine learning.

**Used Library(s):** NumPy, Pandas, Sci-kit Learn(sklearn.ensemble)

**Third Trial:**  The aim of this trial to adding cross validation technique into my model to find out the best fold size and best fold for the test data

**Used Library(s):** NumPy, Pandas

**Fourth Trial:** This trial’s purpose is to eliminate some of the variables which has no or minimum effect on the model by using lasso regression

**Used Library(s):** NumPy, Pandas, Matplotlib (matplotlib. pyplot)

**Fifth Trial:** This is the control of the Fourth Trial in order to compare the results with the Third Trial.

**Used Library(s):** NumPy, Pandas

**Sixth Trial:** In this trial, according to the information given in the course slides, I am adding new dimensions by multiplying columns themselves and the other columns in the model and I am creating a polynomial model by taking the square of (x3+x5+x6+1) ^2 identity. And I am testing the model with the same technique

**Used Library(s):** NumPy, Pandas

**Seventh Trial:** In this trial, I am demonstrating a new Random Forest with the same parameters for a different model Even these results are lower than my Second Trial’s result I know that these results are more reliable.

**Used Library(s):** NumPy, Pandas, Sci-kit Learn (sklearn.ensemble)

**Conclusion:**

In this work, I am trying to predict the last twenty Y-values by demonstrating some of the topics that we cover this semester during this course. The way that I followed throughout the project is explained clearly in the project.

First of all, I used the Multiple Linear Regression to have benchmark for other demonstrations. Then. I used Random forest for same model the results showed that without separation of data the agent memorizes the all of the data and gives nearly perfect results. However, the results Becomes unreliable. At the next step I was turned back to my Multiple Linear regression and I added K-fold Cross Validation into my model. Then I added the Lasso regression to eliminate the variables that cannot participate the performance. After the testing, I have figured out that some of the variables does not provide any improvements for a model. I learned that eliminating those variables maybe more beneficial than keeping it After that I decided to increase my dimensionality by multiplying variables with themselves and each other. Except those I added (x3\*x5\*x6) because I want to see the cross section of those graphs as a dimension. After that I implement a random forest algorithm which is similar to the previous one except the data set and folds. I learned that even performance is lower than the first random tree the second random forest is more reliable because it uses elimination and validation techniques that increases reliability.

In conclusion, it is a demonstration of machine learning techniques that we learn in the course throughout the semester. I think it is very instructive for me who wants to learn machine learning