EE 219 Project I - Report

Muralidharan, Vignesh `UID: 904729596`
Muthappan, Chidambaram `UID: 704774938`
Jeyakumar, Jeya Vikranth `UID: 404749568`

January 30, 2017

Abstract

The purpose of this project is to explore basic regression models on two data sets, along with some techniques to handle over-fitting, namely cross-validation and regularization.

1 Question 1

The given Network Backup dataset has several features like Week number, Day of the week, backup start time, work-flow-ID and file name. To develop prediction models for predicting the size of the data at a given time and the time a backup process might take, it is necessary to develop prediction models which considers the other attributes as the predictors. To plot the size of backup for

a period of 20 days, we have to combine "Day of Week" and "Week number" together and thereby convert "Day of Week" to a numerical value. The basis of our conversion is by referencing "Monday" to one and thereby incrementing each day by 1. Let "Day of Week" be 'd', Weeknumber be 'w'.

$$Days = (w - 1) * 7 + d \tag{1}$$

Based on the value obtained, we a plot between "Days" and "Size of Backup" for the five different workflows is obtained, which can be seen below.

1.1 Plots between Days and Size of Backup

So, it can be seen that size of backup is typically high during weekdays and low during weekends.

From figure 1, for Workflow 0, it can be seen that Size of backup in GB is:

- High during weekdays (Monday - Friday).
- Low during the weekends (Saturday, Sunday).
- Reaches its peak on Tuesday.

From figure 2, for Workflow 1, it can be seen that Size of backup in GB is:

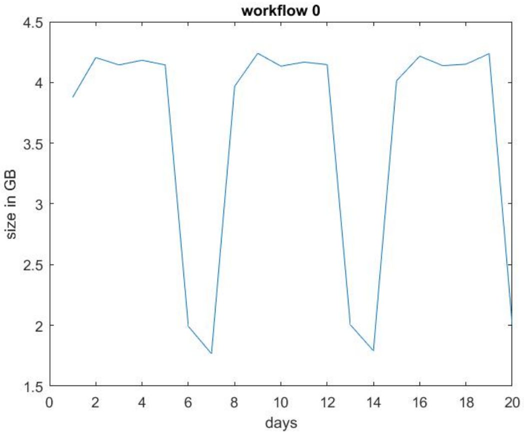


Figure 1: Workflow 0

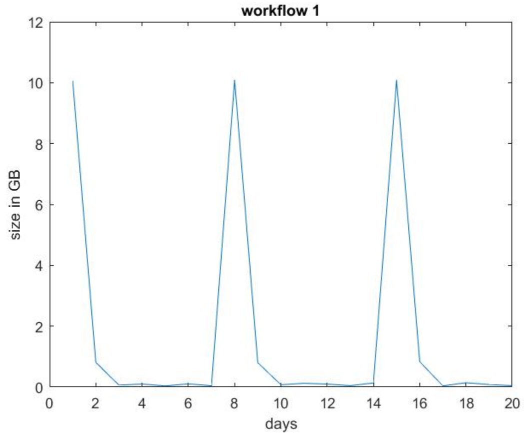


Figure 2: Workflow 1

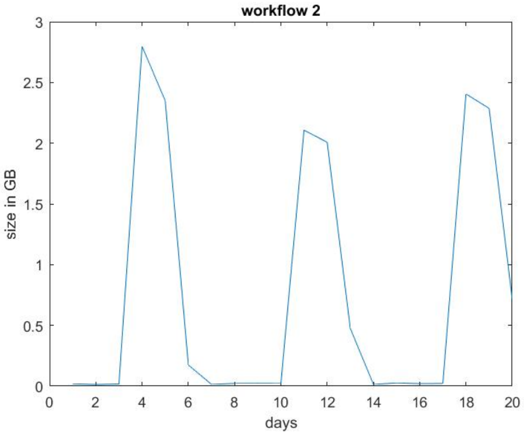


Figure 3: Workflow 2

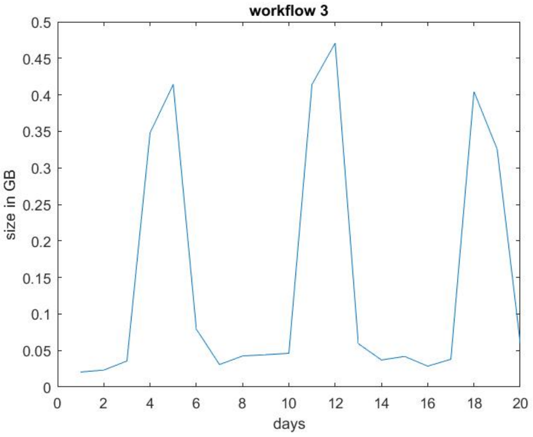


Figure 4: Workflow 3

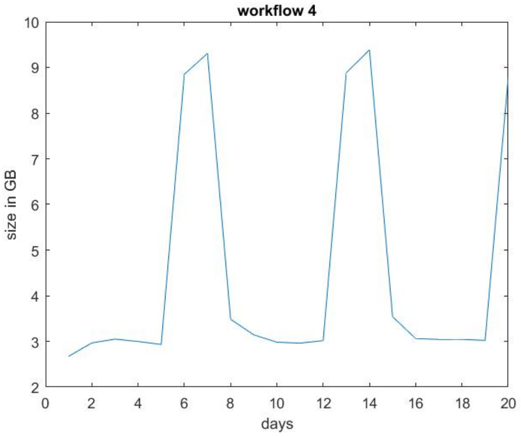


Figure 5: Workflow 4

- Low between Tuesday and Sunday.
- High only on Monday; Size decreases from Monday to Tuesday and increases from Sunday to Monday
- Reaches its peak on Monday.

So, it can be seen that size of backup is is very high on Monday and decreases starting Monday.

From figure 3, for Workflow 2, it can be seen that Size of backup in GB is:

- Zero on Monday, Tuesday and Wednesday.
- High on Thursday and Friday.
- Reaches its peak on Thursday.

So, it can be seen that the size of backup is high only on Thursday and Friday. On all other days, there is nearly zero or no backup needed.

From figure 4, for Workflow 3, it can be seen that Size of backup in GB is:

- Low on Monday, Saturday and Sunday.
- Increases between Wednesday and Thursday, decreases between Thursday and Friday.
- Reaches its peak between Thursday and Friday.

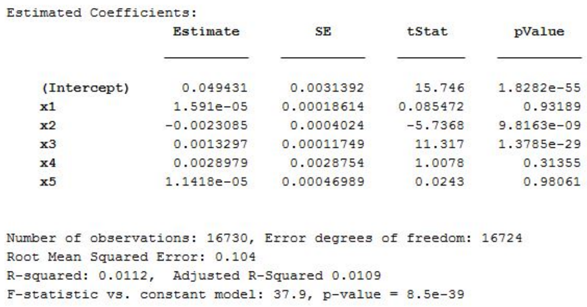
So, it can be seen that size of backup is low on weekends and around it and increases during middle of the week.

From figure 5, for Workflow 4, it can be seen that Size of backup in GB is:

- Low during weekdays (Monday - Friday).
- High during the weekends (Saturday, Sunday).
- Reaches its peak on Sunday.

So, it can be seen that size of backup is low during weekdays, increases around the beginning of weekend and decreases from Sunday to Monday.

Therefore, it can be seen that the distribution of backup size is varying differently based on the days of the week for different workflows .



```
Estimated Coefficients:
                    Estimate        SE          tStat       pValue
                   __________    __________    ________    __________

    (Intercept)      0.049431     0.0031392      15.746    1.8282e-55
    x1              1.591e-05    0.00018614    0.085472       0.93189
    x2             -0.0023085     0.0004024     -5.7368    9.8163e-09
    x3              0.0013297    0.00011749      11.317    1.3785e-29
    x4              0.0028979     0.0028754      1.0078       0.31355
    x5             1.1418e-05    0.00046989      0.0243       0.98061


Number of observations: 16730, Error degrees of freedom: 16724
Root Mean Squared Error: 0.104
R-squared: 0.0112,  Adjusted R-Squared 0.0109
F-statistic vs. constant model: 37.9, p-value = 8.5e-39
```

Figure 6: Model 1

2 Question 2

2.1 Part a

To fit a linear regression model, consider 'y' as "Size of backup" and

- x1 - Week number
- x2 - Day of week
- x3 - Backup start time
- x4 - Work-Flow-ID
- x5 - File Name

$$y = 1 + x1 + x2 + x3 + x4 + x5 \tag{2}$$

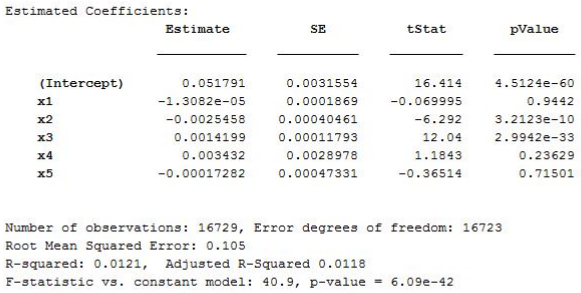
Based on this particular mathematical model, cross-validation is performed. Cross-validation, sometimes called rotation estimation, is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

"Day of week", which is basically x2 (in our case) can be converted from character type to an integer. To do this conversion, we assigned "Monday" to 1 and incremented each of the other days by 1.

"File name" can be converted from a string to integer by finding the length of the string and thereby extracting the last two characters of the string type. The extracted value is thereby stored as an integer type and used for further calculations. Thereafter, by training the cross-fold, we plotted "10-fold cross validated MSE" vs "Learning cycle".

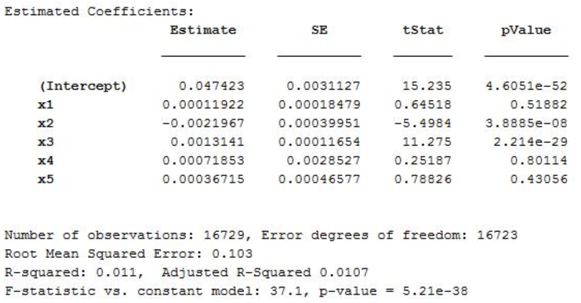
The 10 models are as shown below.

From Figure 6 to Figure 15, 'pvalue' and 'tStat' are to be interpreted to identify the significant terms.



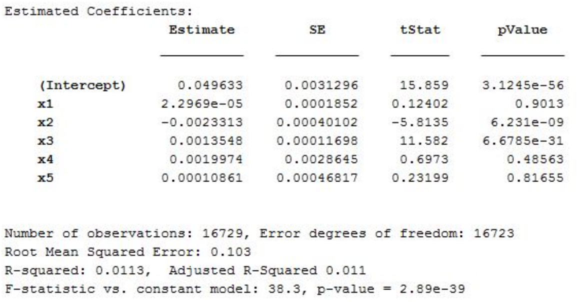
```
Estimated Coefficients:
                   Estimate        SE          tStat        pValue   
                  ___________    __________    _________    __________

    (Intercept)      0.051791     0.0031554       16.414    4.5124e-60
    x1            -1.3082e-05     0.0001869    -0.069995        0.9442
    x2             -0.0025458    0.00040461       -6.292    3.2123e-10
    x3              0.0014199    0.00011793        12.04    2.9942e-33
    x4               0.003432     0.0028978       1.1843       0.23629
    x5            -0.00017282    0.00047331     -0.36514       0.71501


Number of observations: 16729, Error degrees of freedom: 16723
Root Mean Squared Error: 0.105
R-squared: 0.0121,  Adjusted R-Squared 0.0118
F-statistic vs. constant model: 40.9, p-value = 6.09e-42
```

Figure 7: Model 2



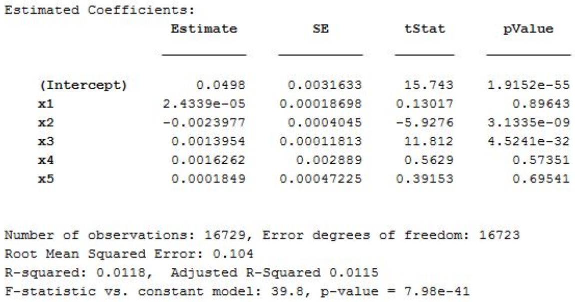
```
Estimated Coefficients:
                    Estimate         SE         tStat       pValue  
                   __________    __________    _______    __________

    (Intercept)      0.047423     0.0031127     15.235    4.6051e-52
    x1             0.00011922    0.00018479    0.64518       0.51882
    x2             -0.0021967    0.00039951    -5.4984    3.8885e-08
    x3              0.0013141    0.00011654     11.275     2.214e-29
    x4             0.00071853     0.0028527    0.25187       0.80114
    x5             0.00036715    0.00046577    0.78826       0.43056


Number of observations: 16729, Error degrees of freedom: 16723
Root Mean Squared Error: 0.103
R-squared: 0.011,  Adjusted R-Squared 0.0107
F-statistic vs. constant model: 37.1, p-value = 5.21e-38
```

Figure 8: Model 3



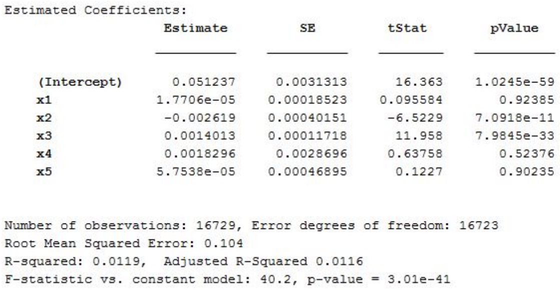
```
Estimated Coefficients:
                    Estimate         SE         tStat       pValue  
                   __________    __________    _______    __________

    (Intercept)      0.049633     0.0031296     15.859    3.1245e-56
    x1             2.2969e-05     0.0001852    0.12402        0.9013
    x2             -0.0023313    0.00040102    -5.8135     6.231e-09
    x3              0.0013548    0.00011698     11.582    6.6785e-31
    x4              0.0019974     0.0028645     0.6973       0.48563
    x5             0.00010861    0.00046817    0.23199       0.81655


Number of observations: 16729, Error degrees of freedom: 16723
Root Mean Squared Error: 0.103
R-squared: 0.0113,  Adjusted R-Squared 0.011
F-statistic vs. constant model: 38.3, p-value = 2.89e-39
```

Figure 9: Model 4



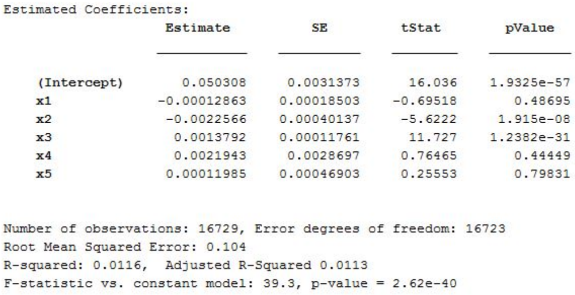
```
Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                  __________    __________    _______    __________

    (Intercept)       0.0498     0.0031633     15.743    1.9152e-55
    x1            2.4339e-05    0.00018698    0.13017       0.89643
    x2            -0.0023977     0.0004045    -5.9276    3.1335e-09
    x3             0.0013954    0.00011813     11.812    4.5241e-32
    x4             0.0016262      0.002889     0.5629       0.57351
    x5             0.0001849    0.00047225    0.39153       0.69541


Number of observations: 16729, Error degrees of freedom: 16723
Root Mean Squared Error: 0.104
R-squared: 0.0118,  Adjusted R-Squared 0.0115
F-statistic vs. constant model: 39.8, p-value = 7.98e-41
```

Figure 10: Model 5



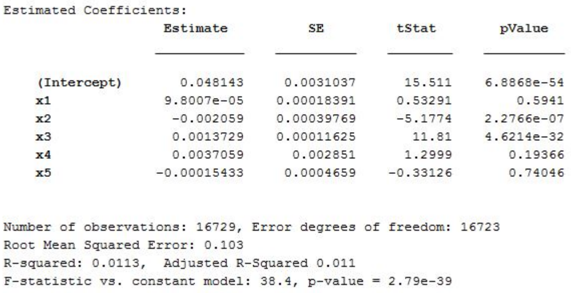
```
Estimated Coefficients:
                   Estimate        SE          tStat       pValue
                  __________    __________    ________    __________

    (Intercept)     0.051237     0.0031313      16.363    1.0245e-59
    x1            1.7706e-05    0.00018523    0.095584       0.92385
    x2             -0.002619    0.00040151     -6.5229    7.0918e-11
    x3             0.0014013    0.00011718      11.958    7.9845e-33
    x4             0.0018296     0.0028696     0.63758       0.52376
    x5            5.7538e-05    0.00046895      0.1227       0.90235


Number of observations: 16729, Error degrees of freedom: 16723
Root Mean Squared Error: 0.104
R-squared: 0.0119,  Adjusted R-Squared 0.0116
F-statistic vs. constant model: 40.2, p-value = 3.01e-41
```

Figure 11: Model 6



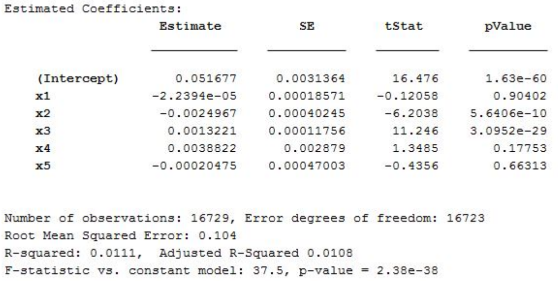
```
Estimated Coefficients:
                   Estimate          SE         tStat       pValue
                  ___________    __________    ________    __________

    (Intercept)      0.050308     0.0031373      16.036    1.9325e-57
    x1            -0.00012863    0.00018503    -0.69518       0.48695
    x2             -0.0022566    0.00040137     -5.6222     1.915e-08
    x3              0.0013792    0.00011761      11.727    1.2382e-31
    x4              0.0021943     0.0028697     0.76465       0.44449
    x5             0.00011985    0.00046903     0.25553       0.79831


Number of observations: 16729, Error degrees of freedom: 16723
Root Mean Squared Error: 0.104
R-squared: 0.0116,  Adjusted R-Squared 0.0113
F-statistic vs. constant model: 39.3, p-value = 2.62e-40
```

Figure 12: Model 7



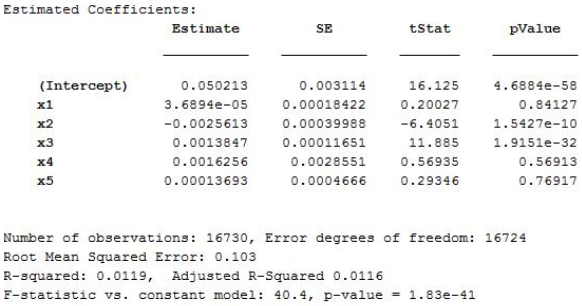
```
Estimated Coefficients:
                   Estimate          SE         tStat       pValue
                  ___________    __________    ________    __________

    (Intercept)      0.048143     0.0031037      15.511    6.8868e-54
    x1             9.8007e-05    0.00018391     0.53291        0.5941
    x2              -0.002059    0.00039769     -5.1774    2.2766e-07
    x3              0.0013729    0.00011625       11.81    4.6214e-32
    x4              0.0037059      0.002851      1.2999       0.19366
    x5            -0.00015433     0.0004659    -0.33126       0.74046


Number of observations: 16729, Error degrees of freedom: 16723
Root Mean Squared Error: 0.103
R-squared: 0.0113,  Adjusted R-Squared 0.011
F-statistic vs. constant model: 38.4, p-value = 2.79e-39
```

Figure 13: Model 8



```
Estimated Coefficients:
                   Estimate        SE         tStat       pValue  
                  ___________  __________  ________  __________

    (Intercept)      0.051677   0.0031364    16.476     1.63e-60
    x1            -2.2394e-05  0.00018571  -0.12058      0.90402
    x2             -0.0024967  0.00040245   -6.2038   5.6406e-10
    x3              0.0013221  0.00011756    11.246   3.0952e-29
    x4              0.0038822    0.002879    1.3485      0.17753
    x5            -0.00020475  0.00047003   -0.4356      0.66313


Number of observations: 16729, Error degrees of freedom: 16723
Root Mean Squared Error: 0.104
R-squared: 0.0111,  Adjusted R-Squared 0.0108
F-statistic vs. constant model: 37.5, p-value = 2.38e-38
```

Figure 14: Model 9



```
Estimated Coefficients:
                   Estimate        SE         tStat       pValue  
                  ___________  __________  ________  __________

    (Intercept)      0.050213    0.003114    16.125   4.6884e-58
    x1             3.6894e-05  0.00018422   0.20027      0.84127
    x2             -0.0025613  0.00039988   -6.4051   1.5427e-10
    x3              0.0013847  0.00011651    11.885   1.9151e-32
    x4              0.0016256   0.0028551   0.56935      0.56913
    x5             0.00013693   0.0004666   0.29346      0.76917


Number of observations: 16730, Error degrees of freedom: 16724
Root Mean Squared Error: 0.103
R-squared: 0.0119,  Adjusted R-Squared 0.0116
F-statistic vs. constant model: 40.4, p-value = 1.83e-41
```

Figure 15: Model 10

In technical terms, a 'p-value' is the probability of obtaining an effect at least as extreme as the one in your sample data, assuming the truth of the null hypothesis. A term is said to be significant if it has p-value lesser than 0.05.

Similarly, 'tStat' measures a probability that a parameter value is significant. A parameter is said to be statistically significant if there is sufficient evidence that the true value of the parameter does not equal zero. If the tStat value is 2 or greater, then there is a 95% probability or better that the parameter estimate does not equal zero.

From the above figures (6-15), only x2 and x3 are significant, because, they have tStat value greater than 2 and pvalue lesser than 0.05.

Therefore, "Day of week" and "Backup start time" are the significant features among the given parameters.

The graphs plotting "Fitted values and actual values scattered plot over time" and "Residuals vs fitted values plot" are shown in Figures 16 and 17.

2.2 Part b

We started our analysis on MATLAB and tried to interpret the outputs of random forest model. Though we could set the "Number of trees" as 20, setting the "Depth of each tree" to 4 was comparatively difficult. We tried to use

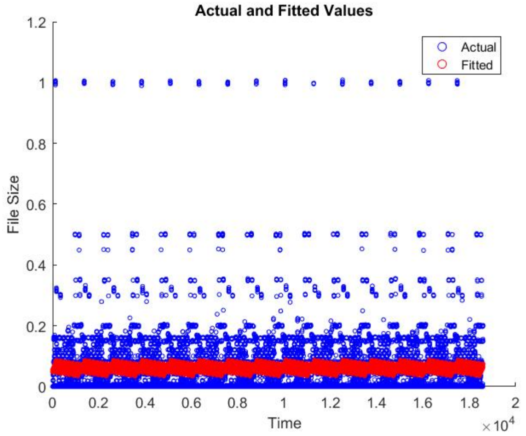


Figure 16: Actual and Fitted vs Time

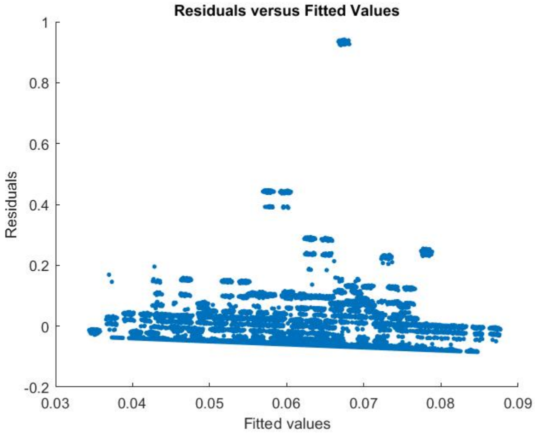


Figure 17: Residuals vs Fitted

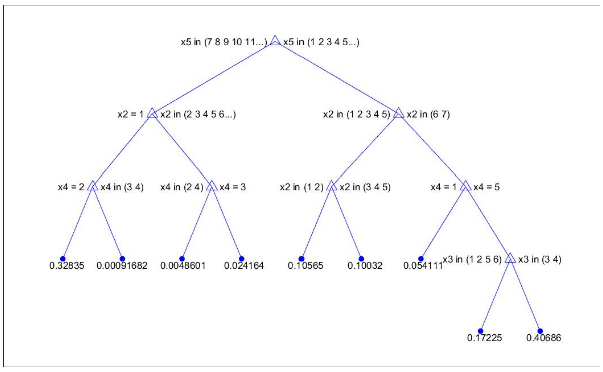


Figure 18: Tree

”fitrensemble” and ”TreeBagger” commands to make a fit. However, they did not have the option to restrict the ”Depth of each tree” to 4. Therefore, we used Python and incorporated ”sklearn.ensemble” for importing the RandomForestRegressor.

```
RANDOM FOREST

RMSE: 0.06091572092037555

IMPORTANCE:
X1 : 4.81178920e-06
X2 : 2.72347256e-01
X3 : 1.49992806e-01
X4 : 2.28570219e-01
X5 : 3.49084907e-01
```

Figure 18 shows the tree model generated. Based on the python code that we ran, the RMSE values for the random forest regression model were much better than the linear regression model. The RMSE value for the random forest regression model was found to be around 0.06, whereas the RMSE value for linear regression model was around 0.104.

In addition to it, the features that are significant in the dataset are ”Day of Week”, ”Backup start time”, ”Workflow ID” and ”File Name”. ”Week number” is the least significant among them. ”Day of Week”, ”Workflow ID” and ”Filename” are comparatively less significant than ”Backup Start Time” as this is a tree model where initial segregation is based on the Workflow ID”. So, features relating to it are relatively less significant than ”Backup Start Time”.

2.3 Part c

To predict the backup sizes of each of the workflows separately, it is necessary to fit a piece-wise linear regression model.

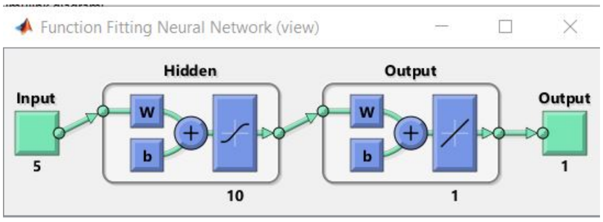


Figure 19: Neural Network Model

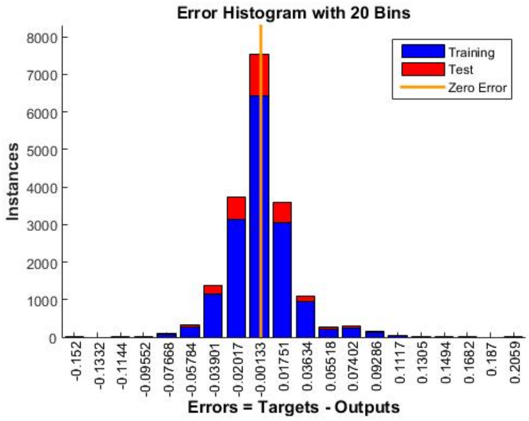


Figure 20: Error Histogram with 20 bins

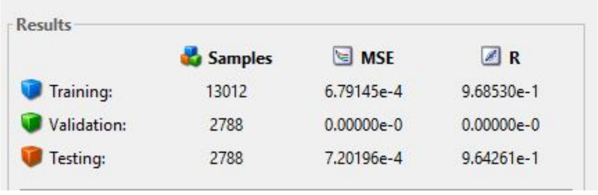
We established a neural network model generated by neural network toolbox. For 5 inputs (features in the dataset), we have used 10 hidden nuerons, as shown in Figure 19.

This particular model is generated on Matlab with 70% testing, 15% validation and 15% testing.

For different instances, the errors are plotted on a histogram, with training, test and zero error for difference between targets and outputs. The histogram is as follows:

The Mean Square Error (MSE) and R values are depicted as shown in Figure 21. It can be noted that MSE and R value for validation are zero. This means that the errors are not present for the validation part.

The major parameters are the number hidden nuerons and the bayseian


Results

	Samples	MSE	R
Training:	13012	6.79145e-4	9.68530e-1
Validation:	2788	0.00000e-0	0.00000e-0
Testing:	2788	7.20196e-4	9.64261e-1

Figure 21: Results

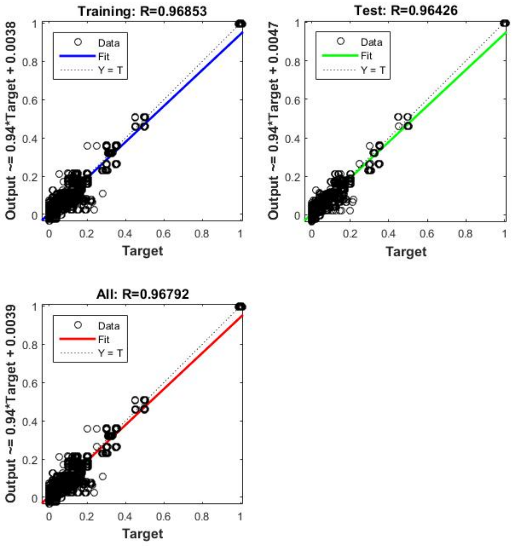


Figure 22: Results

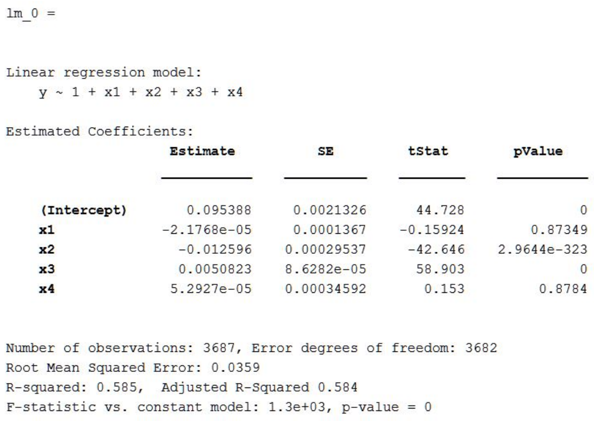
model that has reduced the error and given us a good fit. The number of hidden nuerons was chosen to be twice the number of inputs and the bayseian model that we chose is suitable. In bayseian regularization, the algorithm typically takes more time, but can result in good generalization for difficult, small or noisy datasets. Training stops according to adaptive weight minimization (regularization). Also, we plotted graphs for our nueral network model and identified that most of the points match well with the "Fit"", with a deviation of around 3.6%. Hence, our model fits very well.

3 Question 3

In order to predict the backup size for each of the workflows separately, we first segregated the features of the dataset as explained in Question 2.

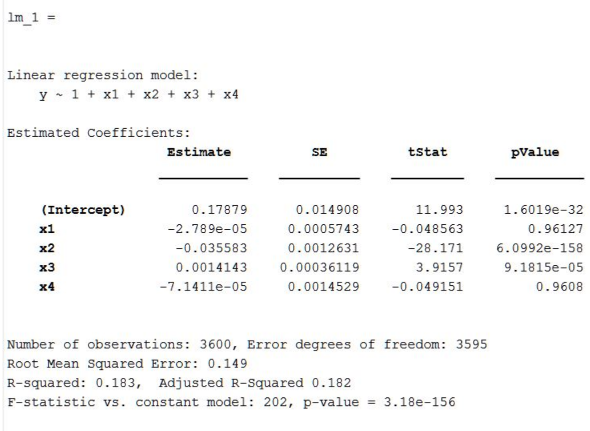
After training the model's indices for a fold, We used commands on MATLAB to avoid overfitting. To "test" for under- and over-fitting, it can be helpful to plot your regression/classification error by some measure of complexity. For under-fitting degrees of complexity, training and testing errors will both be high. For over-fitting degrees of complexity, training errors will be low and testing errors will be high.

An alternative graph to use is a "learning curve" which plots regression/classification error by the number of training examples used. For under-fitting, training and testing errors will converge with larger training sets (and both be high). For over-fitting, the gap between training and testing errors will remain large (with the training error being smaller).



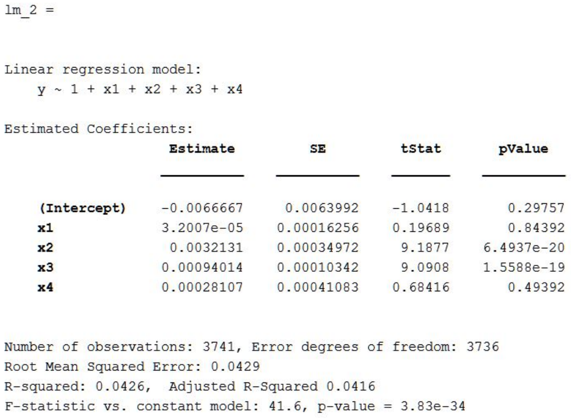
```
lm_0 =


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                    Estimate         SE          tStat        pValue  
                   ___________    __________    ________    __________

    (Intercept)       0.095388     0.0021326      44.728             0
    x1             -2.1768e-05     0.0001367    -0.15924       0.87349
    x2               -0.012596    0.00029537     -42.646    2.9644e-323
    x3               0.0050823    8.6282e-05      58.903             0
    x4              5.2927e-05    0.00034592       0.153        0.8784


Number of observations: 3687, Error degrees of freedom: 3682
Root Mean Squared Error: 0.0359
R-squared: 0.585,  Adjusted R-Squared 0.584
F-statistic vs. constant model: 1.3e+03, p-value = 0
```

Figure 23: Backup size prediction model for workflow 0



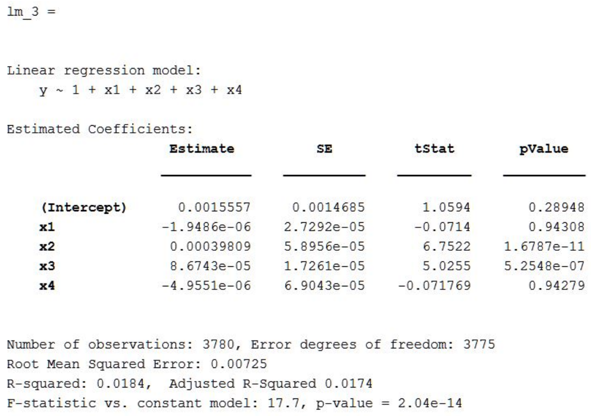
```
lm_1 =


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                    Estimate         SE          tStat        pValue  
                   ___________    __________    _________    __________

    (Intercept)        0.17879      0.014908       11.993     1.6019e-32
    x1              -2.789e-05     0.0005743    -0.048563        0.96127
    x2               -0.035583     0.0012631      -28.171    6.0992e-158
    x3               0.0014143    0.00036119       3.9157     9.1815e-05
    x4             -7.1411e-05     0.0014529    -0.049151         0.9608


Number of observations: 3600, Error degrees of freedom: 3595
Root Mean Squared Error: 0.149
R-squared: 0.183,  Adjusted R-Squared 0.182
F-statistic vs. constant model: 202, p-value = 3.18e-156
```

Figure 24: Backup size prediction model for workflow 1



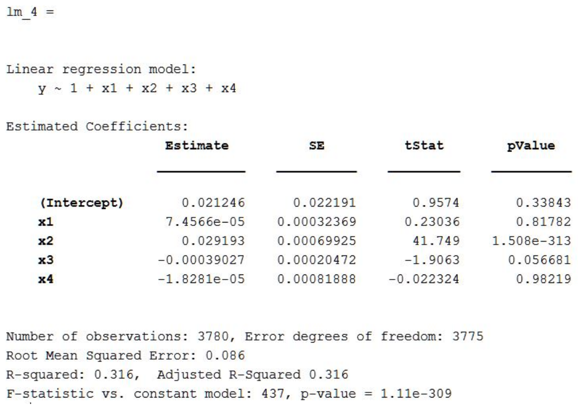
```
lm_2 =


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                    Estimate        SE         tStat      pValue  
                   __________    __________    _______    __________

    (Intercept)    -0.0066667     0.0063992    -1.0418       0.29757
    x1             3.2007e-05    0.00016256    0.19689       0.84392
    x2              0.0032131    0.00034972     9.1877    6.4937e-20
    x3             0.00094014    0.00010342     9.0908    1.5588e-19
    x4             0.00028107    0.00041083    0.68416       0.49392


Number of observations: 3741, Error degrees of freedom: 3736
Root Mean Squared Error: 0.0429
R-squared: 0.0426,  Adjusted R-Squared 0.0416
F-statistic vs. constant model: 41.6, p-value = 3.83e-34
```

Figure 25: Backup size prediction model for workflow 2



```
lm_3 =


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                     Estimate         SE          tStat       pValue  
                    ___________    __________    ________    __________

    (Intercept)       0.0015557     0.0014685      1.0594       0.28948
    x1              -1.9486e-06    2.7292e-05     -0.0714       0.94308
    x2               0.00039809    5.8956e-05      6.7522    1.6787e-11
    x3               8.6743e-05    1.7261e-05      5.0255    5.2548e-07
    x4              -4.9551e-06    6.9043e-05   -0.071769       0.94279


Number of observations: 3780, Error degrees of freedom: 3775
Root Mean Squared Error: 0.00725
R-squared: 0.0184,  Adjusted R-Squared 0.0174
F-statistic vs. constant model: 17.7, p-value = 2.04e-14
```

Figure 26: Backup size prediction model for workflow 3



```
lm_4 =


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                     Estimate          SE          tStat       pValue  
                    ___________    __________    _________    __________

    (Intercept)        0.021246      0.022191       0.9574       0.33843
    x1               7.4566e-05    0.00032369      0.23036       0.81782
    x2                 0.029193    0.00069925       41.749    1.508e-313
    x3              -0.00039027    0.00020472      -1.9063      0.056681
    x4              -1.8281e-05    0.00081888    -0.022324       0.98219


Number of observations: 3780, Error degrees of freedom: 3775
Root Mean Squared Error: 0.086
R-squared: 0.316,  Adjusted R-Squared 0.316
F-statistic vs. constant model: 437, p-value = 1.11e-309
```

Figure 27: Backup size prediction model for workflow 4

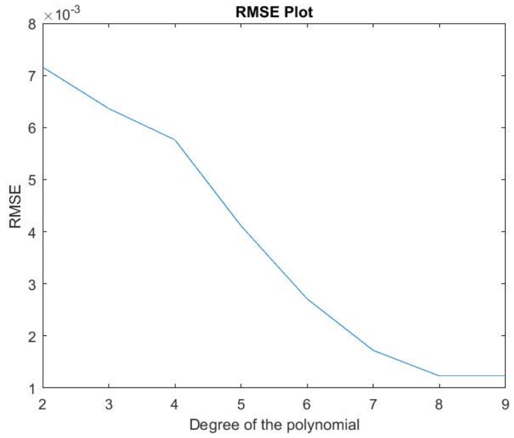


Figure 28: RMSE vs Degree of Polynomial

By avoiding overfitting, complexity of the cross-validated model can be reduced. Since only 90% of the model is trained, there are lesser data points compared to the total number of data points. It evaluates the model only for 90% of the model.

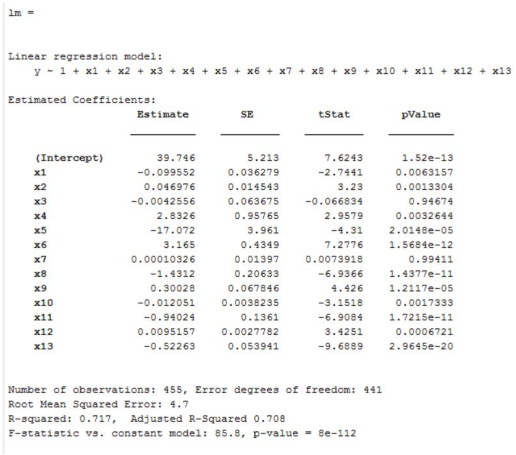
By plotting the RMSE of the trained model against the degree of the polynomial, it is necessary to fit a fixed training and test set, and then for the average RMSE using cross validation.

Plot of RMSE vs Degree of polynomial shows that the degree saturates after 8. Till degree number equal to 8, RMSE decreases exponentially with the degree of polynomial. So, a threshold on the degree of the fitted polynomial beyond which the generalization error of your model gets worse is 8.

4 Question 4

For this question we use the Boston Housing Dataset. This dataset gives the values of the houses in the suburbs of the greater Boston area. This dataset has 500 data points with 14 columns. The first 13 columns are considered as the predictors and the 14^{th} column which contains the median value of the occupied homes is considered the predictor's output.

- x1 - CRIM: per capita crime rate by town
- x2 - ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
- x3 - INDUS: proportion of non-retail business acres per town
- x4 - CHAS: Charles River dummy variable(1/0)
- x5 - NOX: nitric oxides concentration (parts per 10 million)
- x6 - RM: average number of rooms per dwelling



```
lm =


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13

Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                  __________    _________    _________    __________

    (Intercept)       39.746        5.213       7.6243      1.52e-13
    x1             -0.099552     0.036279      -2.7441     0.0063157
    x2              0.046976     0.014543         3.23     0.0013304
    x3            -0.0042556     0.063675    -0.066834       0.94674
    x4                2.8326      0.95765       2.9579     0.0032644
    x5               -17.072        3.961        -4.31    2.0148e-05
    x6                 3.165       0.4349       7.2776    1.5684e-12
    x7            0.00010326      0.01397    0.0073918       0.99411
    x8               -1.4312      0.20633      -6.9366    1.4377e-11
    x9               0.30028     0.067846        4.426    1.2117e-05
    x10            -0.012051    0.0038235      -3.1518     0.0017333
    x11             -0.94024       0.1361      -6.9084    1.7215e-11
    x12            0.0095157    0.0027782       3.4251     0.0006721
    x13             -0.52263     0.053941      -9.6889    2.9645e-20


Number of observations: 455, Error degrees of freedom: 441
Root Mean Squared Error: 4.7
R-squared: 0.717,  Adjusted R-Squared 0.708
F-statistic vs. constant model: 85.8, p-value = 8e-112
```

Figure 29: Best fitted linear regression model

- x7 - AGE: proportion of owner-occupied units built prior to 1940
- x8 - DIS: weighted distances to five Boston employment centers
- x9 - RAD: index of accessibility to radial highways
- x10 - TAX: full-value property-tax rate
- x11 - PTRATIO: pupil-teacher ratio by town
- x12 - Bk is the proportion of blacks by town
- x13 - percentage lower status of the population
- y - Median value of owner-occupied homes

4.1 Part a - Fitting Linear Regression Model

For the given housing dataset 10-fold cross validation is performed and the best Linear Regression model is found to predict the median value of the occupied houses. This can be done by finding the model which gives minimum MSE from the Cross validated MSE vs Learning Cycle plot. Figure 30 shows this plot and as we can see model 5 gives the least MSE and hence it is chosen as the best plot. Figure 29 shows the best fitted Linear Regression model with MEDV as the target variable and the other attributes as the features. From the output we can infer that the predictors x3 and x7 have p-value <0.05 and abs(t-stat) >2 and hence they are insignificant in predicting the mean value of the houses.

The scatter plot of the actual values and the fitted values are plotted in Figure 31 and the residual values are plotted against the fitted values in Figure 32. From these figures we find that the predicted value is almost same as the actual value and the MSE is 4.75

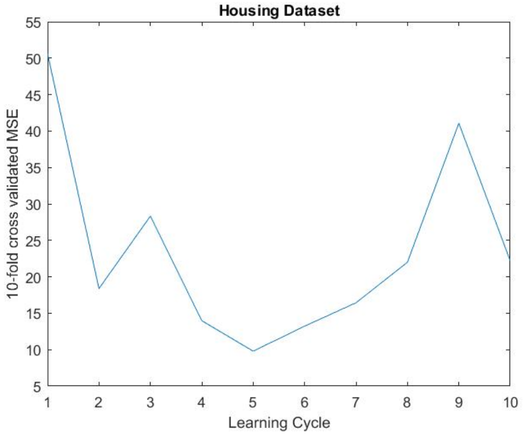


Figure 30: Cross Validated MSE vs Learning Cycle

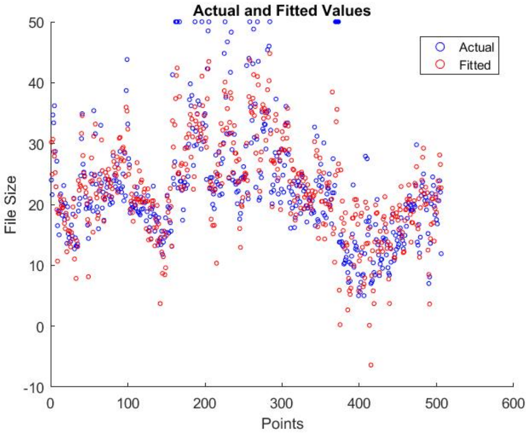


Figure 31: Scatter plot of Actual and Fitted values

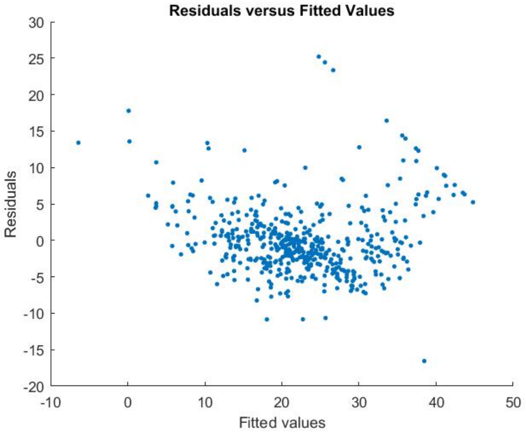
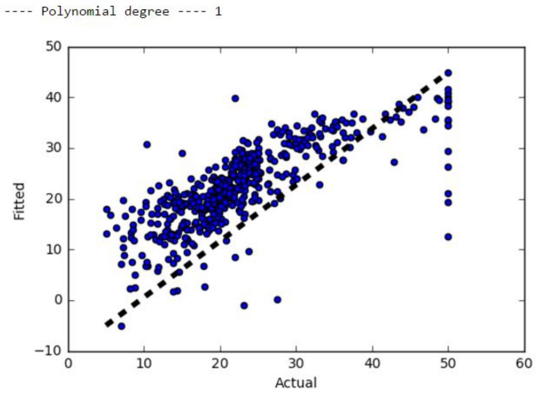
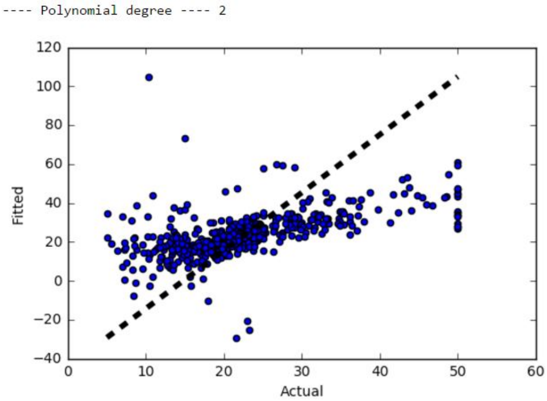


Figure 32: Residuals vs Fitted Values



```
---- Polynomial degree ---- 1
```

Figure 33: Polynomial Degree: 1



```
---- Polynomial degree ---- 2
```

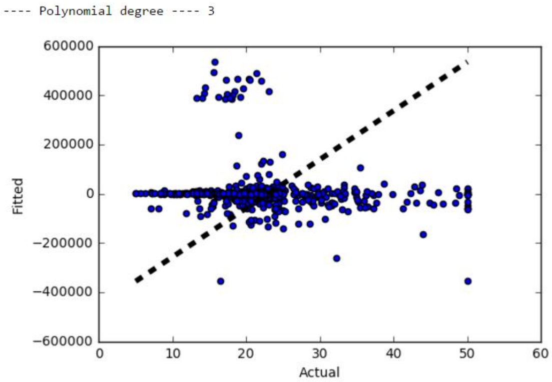
Figure 34: Polynomial Degree: 2

4.2 Part b - Polynomial Regression

To repeat the same steps for a polynomial regression function to find the optimal degree of fit, we trained the model in a way similar to question 2 and found the mean squared error values. The plots show the different figures for degrees ranging from 1 to 7.

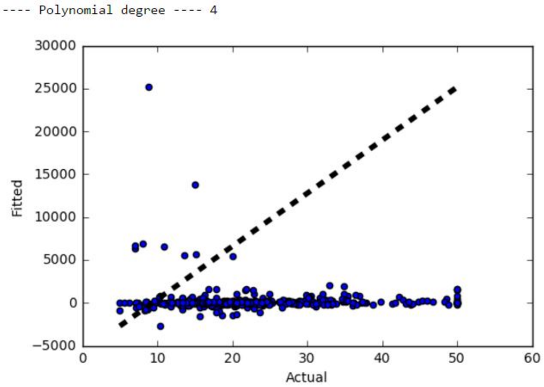
The given dataset has 13 features. So, while trying to fit a polynomial, linear regression would possibly be a better fit because only a few features associated with it will scatter the points away from the fitting model.

From figures 32 to 38, it is evident that the polynomial with degree 1 is more the most optimal fit as it has lesser deviation from the fit. Only a few of the 13 features are used in the model. As the degree of polynomial increases, a larger number of features are used from the dataset thereby increasing the number of points that deviate from the fit.



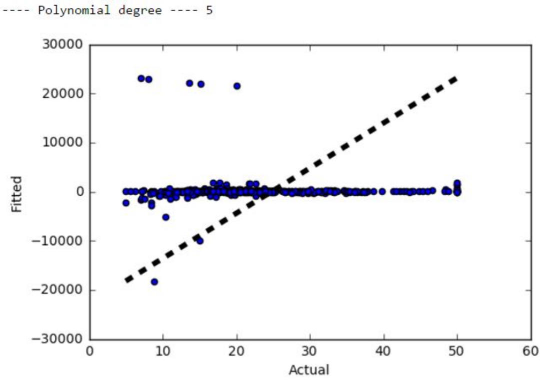
```
---- Polynomial degree ---- 3
```

Figure 35: Polynomial Degree: 3



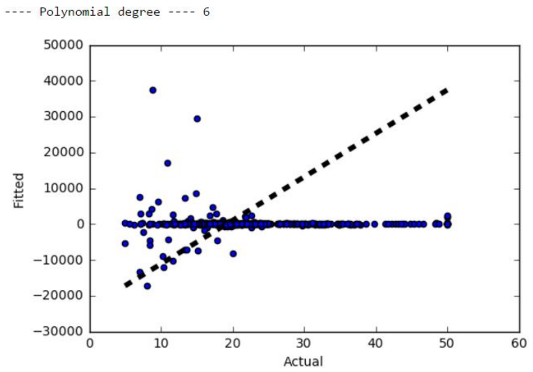
```
---- Polynomial degree ---- 4
```

Figure 36: Polynomial Degree: 4



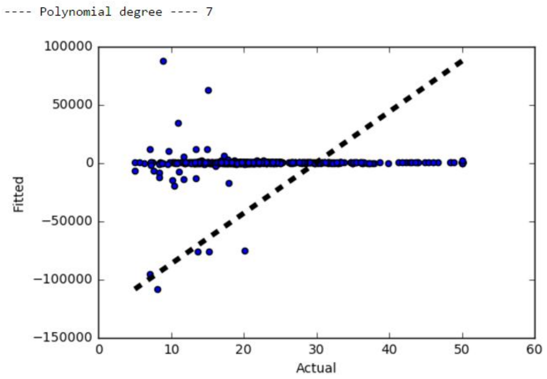
```
---- Polynomial degree ---- 5
```

Figure 37: Polynomial Degree: 5



```
---- Polynomial degree ---- 6
```

Figure 38: Polynomial Degree: 6



```
---- Polynomial degree ---- 7
```

Figure 39: Polynomial Degree: 7

5 Question 5

5.1 Part A

Ridge regression is used on the Boston housing data set. Ridge regression is l2 regularization technique applied to impose penalty on the size of the regression coefficients along with sum of residuals. On applying the given alpha values 1, 0.1,0.01,0.001 with a 10 fold cross validation on the data set it can be observed that best alpha value is 1.

```
RIDGE REGRESSION

Best Alpha value :1

Coefficient values:
[ -1.04003113e-01  4.71233616e-02 -1.51224538e-02  2.53703506e+00
  -1.06929400e+01  3.83737125e+00 -4.96597699e-03 -1.37719890e+00
   2.85568045e-01 -1.26146317e-02 -8.83050341e-01  9.66388220e-03
  -5.36241553e-01]

Best RMSE : 4.691833312145027
```

From the coefficient values we understand that according ridge regression features AGE and B are not important in determining MEDV.

5.2 Part B

Lasso Regularization is l1 regularization where the l1 norm is applied to impose penalty on the size of the regression coefficients along with sum of residuals. The alpha values given the question are 1, 0.1,0.01,0.001. A 10 fold cross validation is used to avoid over fitting.It is interesting that even in Lasso the best observed alpha value is 1.

```
LASSO REGRESSION

Best Alpha value :1

Coefficient values:
[ -3.60861533e-02  1.30015641e-02 -3.08477438e-03  2.32846703e+00
  -8.43197310e+00  4.21933905e+00 -0.00000000e+00 -7.48930829e-01
   0.00000000e+00 -0.00000000e+00 -8.26381598e-01  7.24980052e-03
  -5.23461136e-01]

Best RMSE : 4.860201523597031
```

On observing the coefficient values it is very clear that Lasso eliminates the following features by assigning a 0 as the coefficient: AGE, RAD and TAX.