EE 219 Project 1 - Report

January 30, 2017

Isha Verma	Heenal Doshi	Pranav Thulasiram Bhat
ishaverma@cs.ucla.edu	heenald@cs.ucla.edu	pranavtbhat@cs.ucla.edu
404761131	004758927	704741684

1 Introduction

The goal of this project was to analyze two data sets:

- Network Backup Dataset : Captures simulated traffic data on a backup system and contains information of the size of the data moved to the destination as well as the time it took for backup. We had to predict the backup size of the traffic depending on the the other variables such as file-name, day/time of backup.
- Housing Dataset : The Boston-Housing data set that contains the mortgage values of suburbs. We had to employ Linear and Polynomial Regression along with other algorithms to overcome overfitting to create a predictive model that can be used to estimate the value of owner-occupied homes.

We have used regressions to analyze both datasets. Regression is an approach for modeling the relationship between a scalar dependent variable and one or more explanatory variables / features. In this project, we have used cross-validation, regularization, random forest and neural network regression in order to predict a dependent variable present in the datasets.

2 Network Backup Dataset

The Network-Backup Dataset has information of files maintained in destination machine and it monitors and copies their changes in four hours cycle. The features captured in data set are as follows:

1. Week index : Week number (Numeric)
2. Day of the week : The day on which the backup takes place. (Categorical)
3. Backup start time : The exact time at which the backup completes (Numeric)
4. Work-Flow-ID : Identifies similar file backup operations (Categorical)
5. File Name : Name of the file being backed-up (Categorical)
6. Size of Backup (GB) : Size of the file's backup (Numeric)
7. Backup time: The duration of the backup procedure in hours (Numeric)

3 Question 1 : Types of Relationships in the Dataset

To understand the relationships between the variables in the dataset, we plotted the actual copy sizes of all files over a time period of 20 days for each workflow.

We observed that each workflow followed a pattern that repeated over a period 7 days (i.e. weekly). All files in a workflow echoed the pattern pretty closely. More details follow,

1. Work_Flow_0 : Copy size for files 0 to 5 usually stayed high during the weekdays and dipped during the weekends.
2. Work_Flow_1 : Files in this workflow (6 to 11) emulate each other very closely. Each file reaches a peak copy size of 1.7 GB on Mondays, stays around .1 GB on Tuesday and negligible on other days.
3. Work_Flow_2 : The copy size for Files (12 to 17) only varies on Thursdays and Fridays, and stays negligible on other days.
4. Work_Flow_3 : Files in these workflow (18 to 23) show a little uneven distribution (more pronounced in the below graph since the total copy size is lower). Overall they have a higher copy size on Thursdays and Fridays and stay low on other days.
5. Work_Flow_4 : Data back-up for Files (24 to 29) is mostly constant between 0.4 and 0.6 GB over the weekdays and shoots up over the weekend. Thus they show a behavior opposite to workflow 1 files.

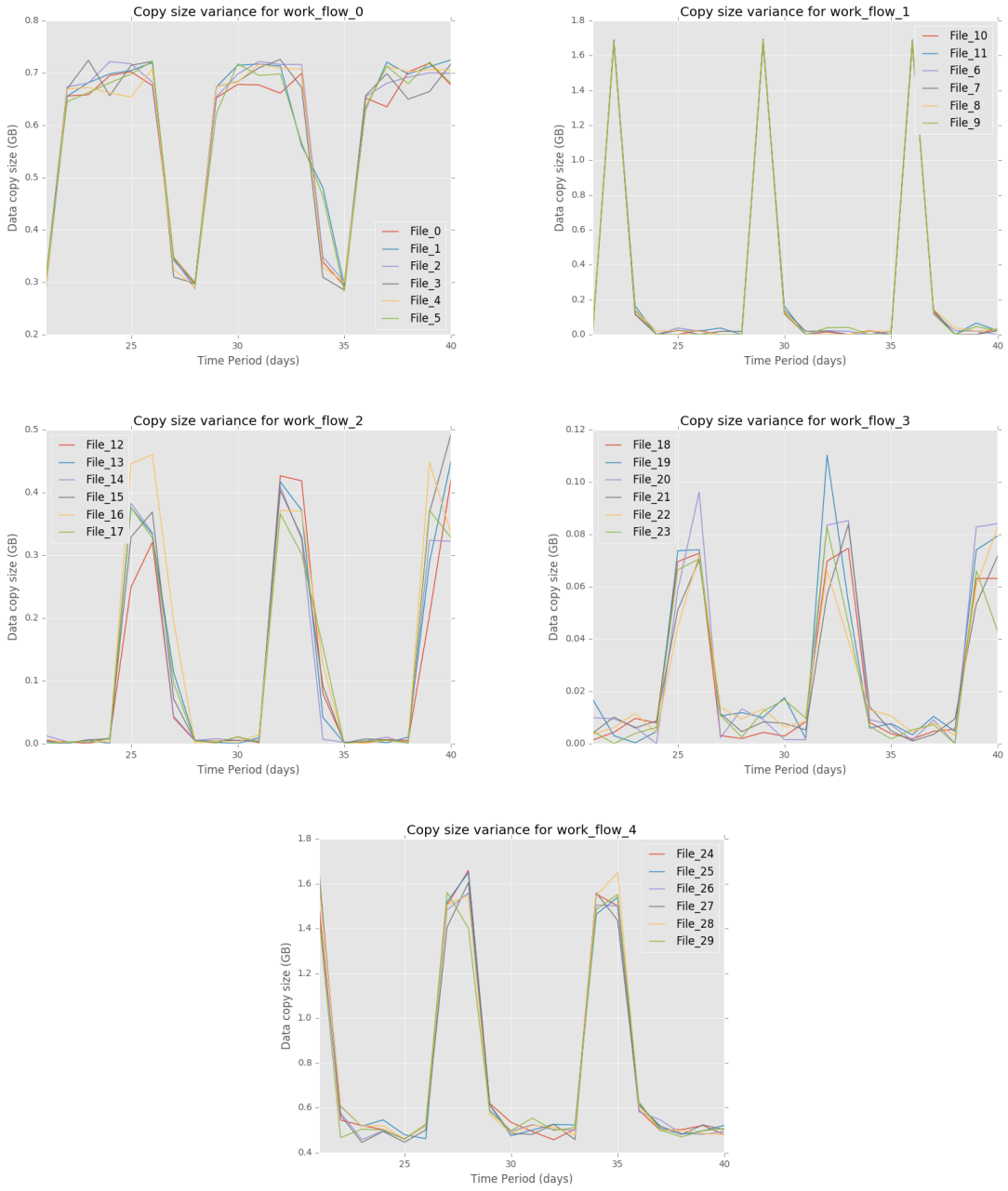

Figure 1: The variance of file copy sizes over a time period of 20 days (Day 21 - 40) for each workflow

4 Question 2a : Linear Regression

We started with a linear regression model using all variables and trained it using 10 fold cross validation. The RMSE obtained after 10-fold Cross-Validation was 0.0796

In order to identify the statistical significance of each attribute we found their p-values using the statsmodel library.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         SizeofBackupGB   R-squared:                       0.417
Model:                            OLS   Adj. R-squared:                  0.417
Method:                 Least Squares   F-statistic:                     2218.
Date:                Sun, 29 Jan 2017   Prob (F-statistic):               0.00
Time:                        22:30:33   Log-Likelihood:                 20683.
No. Observations:               18588   AIC:                        -4.135e+04
Df Residuals:                   18581   BIC:                        -4.130e+04
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept         -0.0267      0.002    -11.919      0.000        -0.031    -0.022
WeekNo             0.0001      0.000      0.842      0.400        -0.000     0.000
DayofWeek          0.0013      0.000      4.397      0.000         0.001     0.002
BackupStartTime    0.0010   8.54e-05     11.413      0.000         0.001     0.001
WorkFlowID         0.0031      0.002      1.459      0.145        -0.001     0.007
FileName       -5.717e-06      0.000     -0.017      0.987        -0.001     0.001
BackupTimeHour     0.0712      0.001    113.770      0.000         0.070     0.072
==============================================================================
Omnibus:                    17609.643   Durbin-Watson:                   0.360
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1009235.744
Skew:                           4.532   Prob(JB):                         0.00
Kurtosis:                      37.941   Cond. No.                         92.9
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 2: Linear Regression stats for Network Dataset

As seen in the above figure,

1. Attributes Day of the week, Backup start time and Backup time had p values of 0 i.e. they are significant to the estimation. We observed the same from our manual investigation too.
2. On the other hand, p-values for Week number and File name were considerably high, indicating they do not contribute much to the estimation. This was according to our expectations, as the patterns repeated every week and files belonging to same workflow showed more or less the same copy pattern and sizes too.

4.1 Fitted vs Actual Values

The graph below shows the Actual vs Fitted values of Copy Size for the Linear Regression model. The data seems to be randomly distributed about the 45 degree line. As seen, the correlation between actual and predicted values is weak i.e. the predictions can be better.

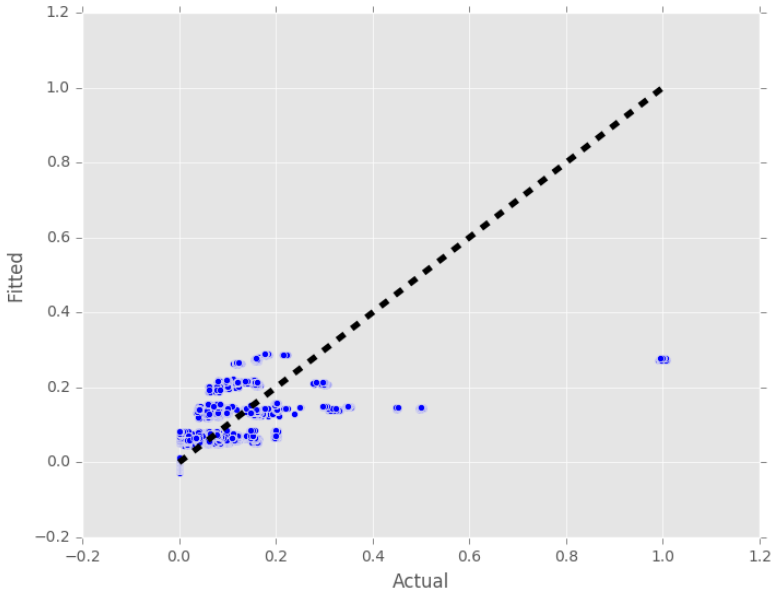

Figure 3: Fitted Values vs Actual Values for copy size

4.2 Residual vs Fitted Values

Figure 4 shows the graph for Residuals and Fitted values. The points are distributed evenly around Residual=0 and do not show any clear pattern.

Based on the above two graphs we concluded that linear regression seems to be missing the predictive power to explain training data and more complex models need to be tried.

Notes:

1. We even tried removing the two features with high p value - Week number and File Name, however the RMSE and above graphs remained almost the same.
2. The given dataset contains three categorical features - day of the week, Workflow ID and File names. We considered two approaches for dealing with these features: Enumeration and one hot encoding.

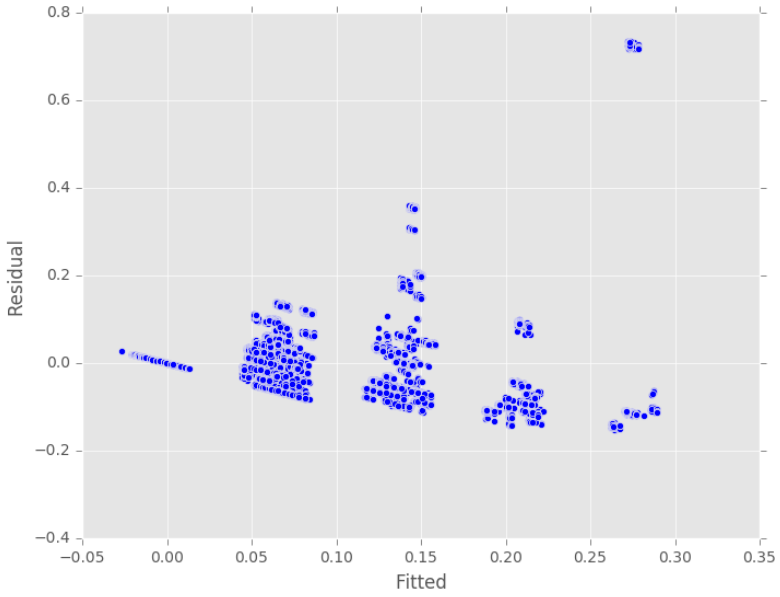

Figure 4: Residual Values vs Actual Values for copy size

One hot encoding resulted into total of 45 columns, however improved the RMSE only negligibly. We decided to continue with enumeration following few intuitions: the feature fileName did not seem to contribute much to the prediction anyway, it might be useful to preserve the linear distance between week days given by enumeration and there was a negligible improvement in rmse for a large increase in variable count.

5 Question 2b: Random Forest Regression

We used the RandomForestRegressor from sklearn's ensemble library for this part. We performed parameter tuning to identify best values for max depth and number of estimators. Keeping a constant value for number of estimators we saw the RMSE was minimum when max depth = 10 (as seen in Figure 5). Using this max depth we identified the optimal number of estimators to be 180.

With these parameters and 10 fold cross validation we observed RMSE around 0.00959.

5.1 Fitted vs Actual Values

As expected, we saw a better fit between actual values and predicted values using Random forest. Figure 7 shows the points lying close to 45 degree line, indicating the same.

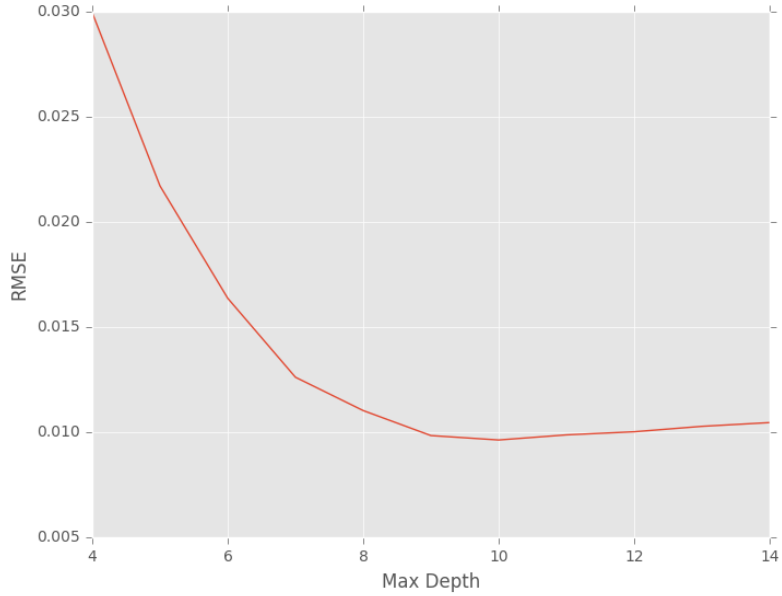

Figure 5: RMSE vs Maximum Depth of Tree

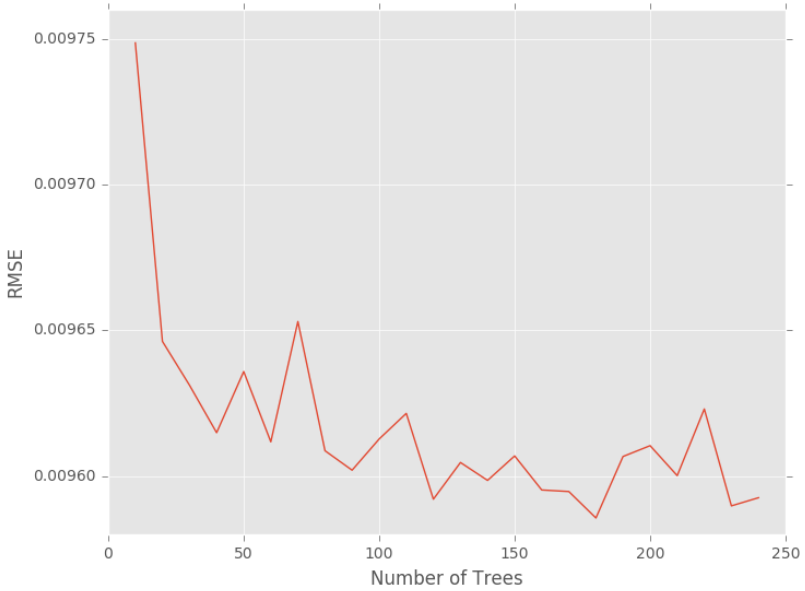

Figure 6: RMSE vs Maximum Number of Trees

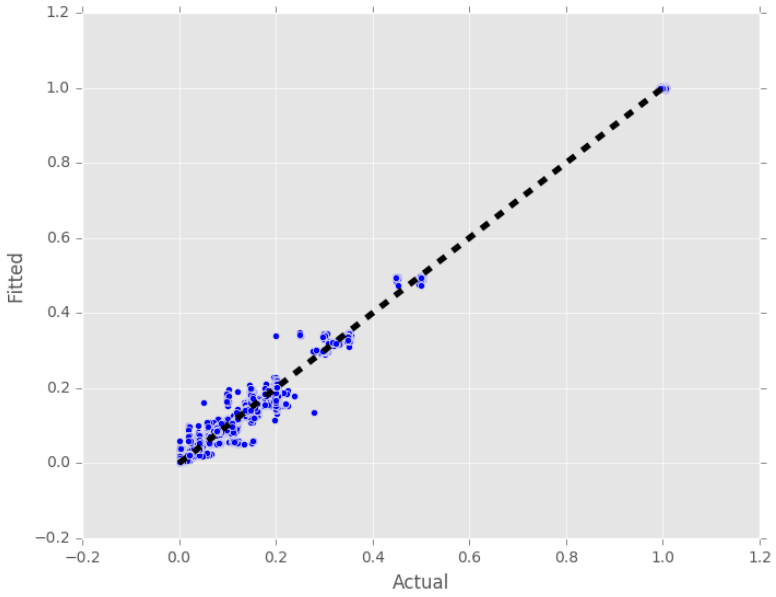

Figure 7: Fitted vs Actual Values

5.2 Residuals vs Actual Values

Figure 8 shows the plot for residuals vs fitted values. The residuals seem evenly distributed around 0, i.e. the model is correct on average for all fitted values, and do not show any clear pattern showing this particular model may not have evident room for improvement.

5.3 Interpreting Random Forest Model

We used the feature_importance_ measure provided for Random Forest Regressor by sklearn to identify the relative importance of the 6 features in our dataset.

The importance score in sorted order with 6 features was: [(0.4792, 'BackupTimeHour'), (0.3154, 'DayofWeek'), (0.0791, 'BackupStartTime'), (0.0755, 'FileName'), (0.0496, 'WorkFlowID'), (0.0012, 'WeekNo')] Thus we saw that BackupTime and week day were the most informative features - which is aligned with our intuition, time taken for backup is a good indicator of the size of file and as observed in patterns in task 1 files followed a pattern each day of the week. At the same time, week number did not seem to be contributing much to the prediction which again aligns with our observation since the pattern repeats every week. Hence we decided to remove week number feature from our consideration.

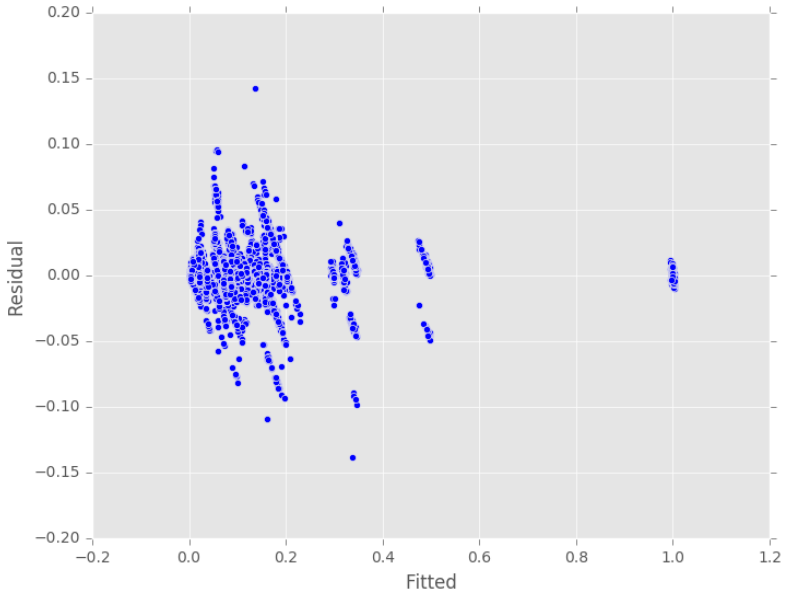

Figure 8: Residuals vs Fitted Values

File name and workflowID had similar importance score and it is possible in case of correlated features that feature picked first is given a higher score than the second even though it may not necessarily be more important that the other. We feel that may be the reason why workflow ID was ranked a little below File name. (Why they are correlated: Each workflow has a unique set of file names).
Hence we re-tested with 4 features: once removing the week number and file name and then removing the week number and the workflow ID. We observed the former gave the least RMSE, 0.0092.

5.4 Linear Regression vs Random Forest Regression

Thus the best RMSE value we obtained using Random Forest was 0.0092, significantly lower than the one observed with Linear Regression, 0.0796.

6 Question 2c: Neural Network Regression

We used Pybrain's buildNetwork and BackpropTrainer to build and train a neural network regressor. The number of nodes in input layer were 6 - our features and a single output node predicting the copy size. As a rule of thumb we selected number of hidden layers = 1 and number of hidden units in that layer to be between number of input and output nodes.

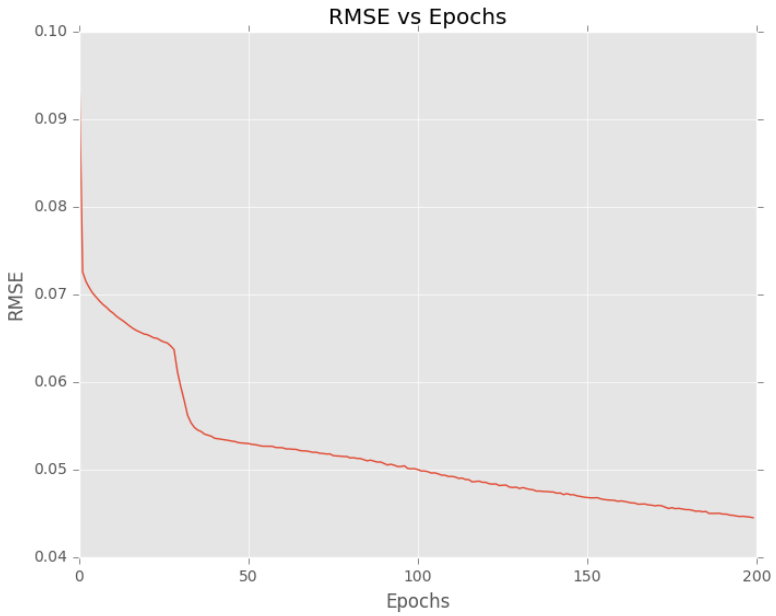

Figure 9: RMSE vs Epoch

6.1 How Parameters affect the RMSE

- We tried different values for epoch and hidden unit count to see their effect on RMSE. Figure 9 above shows the plot of RMSE vs Epochs - RMSE goes on decreasing with Epochs. Epoch (number of times the entire network is trained) = 100 with hidden units = 5 resulted into RMSE of 0.045.
- Default learning rate of 0.01 was used and the network was trained until convergence, since increasing the learning rate can possibly result into skipping of the answer. We also observed keeping bias = true gave us a better fit.
- In order to train up to convergence we used the parameter continueEpoch which sets the number of epochs that should be tested for convergence.

7 Question 3a: Linear Regression on WorkFlows

As part of Question 3, we first performed piece-wise linear regression on each of the workflows present in the dataset. To do this, we loaded the dataframe, grouped by Work-Flow-ID and ran 10-fold cross validation. The results are presented in the subsections below:

7.1 WorkFlow 0

Estimated intercept coefficient : 0.02605

RMSE Values of Estimator : 0.0294

Features	EstimatedCoefficient
0	5.94401787e-05
1	-9.84671894e-03
2	4.48020874e-03
3	0.00000000e+00
4	6.59748390e-05
5	2.60534268e-02

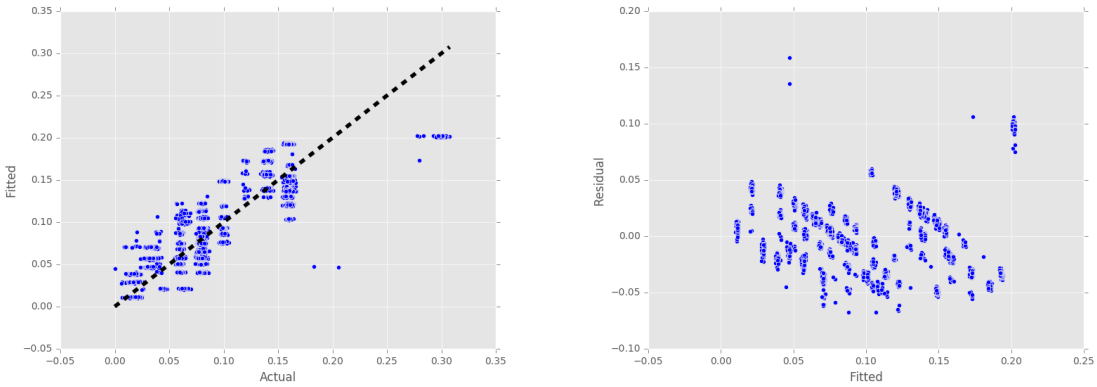

Figure 10: Fitted Values vs Actual Values and Residual Values vs Actual Values for Work-Flow-0

The RMSE value for Work-Flow-0 is clearly a huge improvement over the global RMSE value of 0.0796. There is very little variance in the predicted and fitted values. Therefore, the piece-wise fit is clearly better for Work-Flow-0.

WorkFlow 1 Estimated intercept coefficient : 0.103 RMSE Values of Estimator : 0.12

Features	EstimatedCoefficient
0	2.71199653e-04
1	-1.50916282e-03
2	3.53561392e-03
3	-2.77555756e-17
4	1.64243130e-04
5	1.27503655e-01

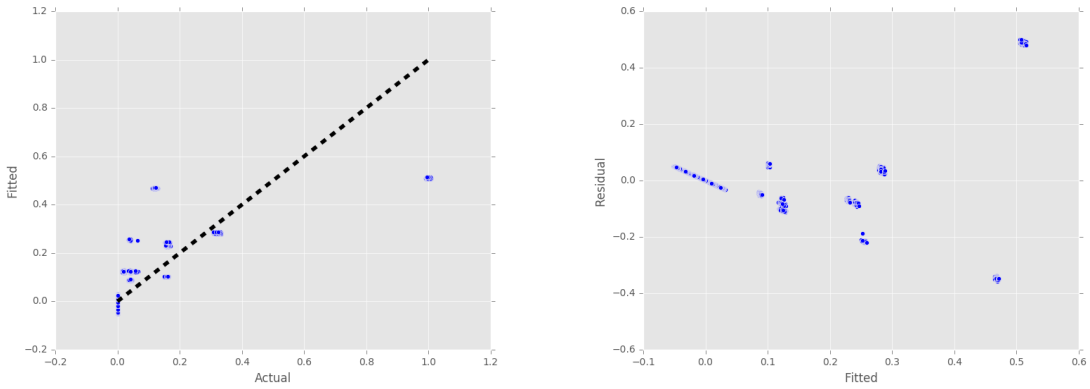

Figure 11: Fitted Values vs Actual Values and Residual Values vs Actual Values for Work-Flow-1

The piece wise fit for Workflow-1 is clearly worse than the global fit. The graphs and the RMSE value indicate this.

WorkFlow 2 Estimated intercept coefficient : 0.0255 RMSE Values of Estimator : 0.02941

Features	EstimatedCoefficient
0	1.73961525e-04
1	6.37279081e-04
2	9.91253434e-05
3	-6.93889390e-18
4	-9.33609348e-05
5	4.38423508e-02

The piece-wise fit for Work-Flow-2 is much better than the global fit. The Residual vs Fitted values show deviations in the range of -0.15 to 0.05. Clearly the standard deviation is higher.

WorkFlow 3 Estimated intercept coefficient : 0.00591 RMSE Values of Estimator : 0.02941

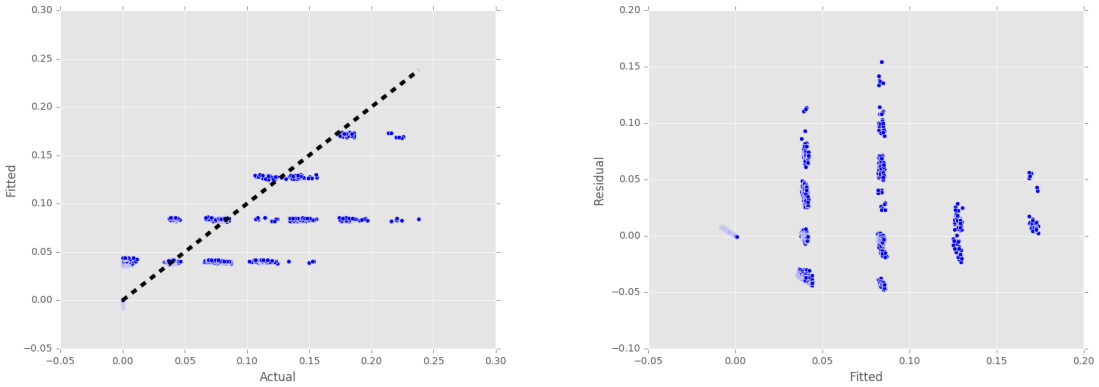

Figure 12: Fitted Values vs Actual Values and Residual Values vs Actual Values for Work-Flow-2

Features	EstimatedCoefficient
0	2.19482487e-05
1	2.41194316e-04
2	5.28018124e-05
3	1.62630326e-19
4	-4.46363557e-05
5	8.54608156e-03

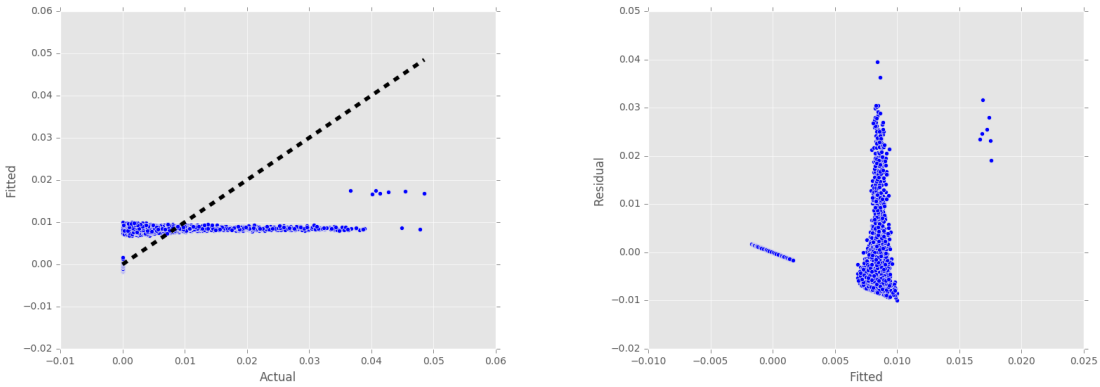

Figure 13: Fitted Values vs Actual Values and Residual Values vs Actual Values

The piece-wise fit for Work-Flow-3 gives an RMSE value of 0.00591. Clearly this fit is excellent as it provides the best RMSE value across Work-Flows. The Actual vs Fitted and Residual vs Fitted graphs also seem to indicate this.

WorkFlow 4 Estimated intercept coefficient : 0.0842 RMSE Values of Estimator : 0.02941

Features	EstimatedCoefficient
0	1.45723821e-04
1	2.87361814e-02
2	-3.56007999e-04
3	-3.46944695e-18
4	2.52616548e-05
5	2.61646914e-02

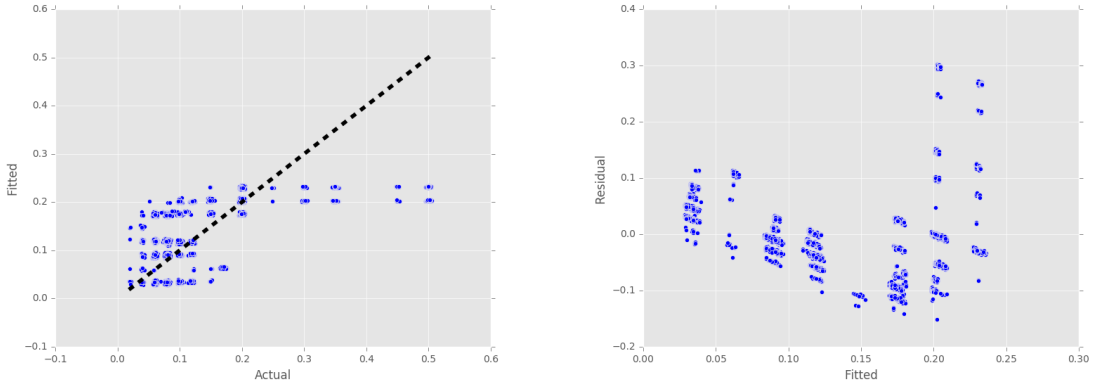

Figure 14: Fitted Values vs Actual Values and Residual Values vs Actual Values

For Work-Flow-4, the piecewise fit provides an RMSE value of 0.08357, which is higher than 0.07956. As seen from the graph of Residual Vs Fitted values the residual, the fit is actually worse than the overall fit.

Therefore, we can see that the piece wise linear regressions have improved the fits for Work-Flows 0, 2 and 3. However, the fits for Work-Flows 1 and 4 are worse than the global regressions run in Question 2a.

8 Question 3b : Polynomial Regression

As part of task 3, we fit a polynomial regression model to our data. We varied the degree of the polynomial between 1 and 9. We evaluated the RMSE values both with and without cross validation. We found that degree 5 yielded the best RMSE value. We obtained an optimal RMSE value of

In order to test our prediction model, we tried fitting a polynomial function to our variables. We tried and tested the model, by fitting the polynomial function upto power of 10. For polynomial regression we plotted RMSE vs polynomial degree and found that for degree 5 the RMSE value is minimum and then we get constant RMSE value as shown in the plot below. We used this obtained degree on our entire dataset

by splitting the dataset as 90 percent training and 10 percent testing. The input attributes of the dataset were transformed to the degree of 5 and the model was created and tested. We obtained an RMSE Value = 0.0575.

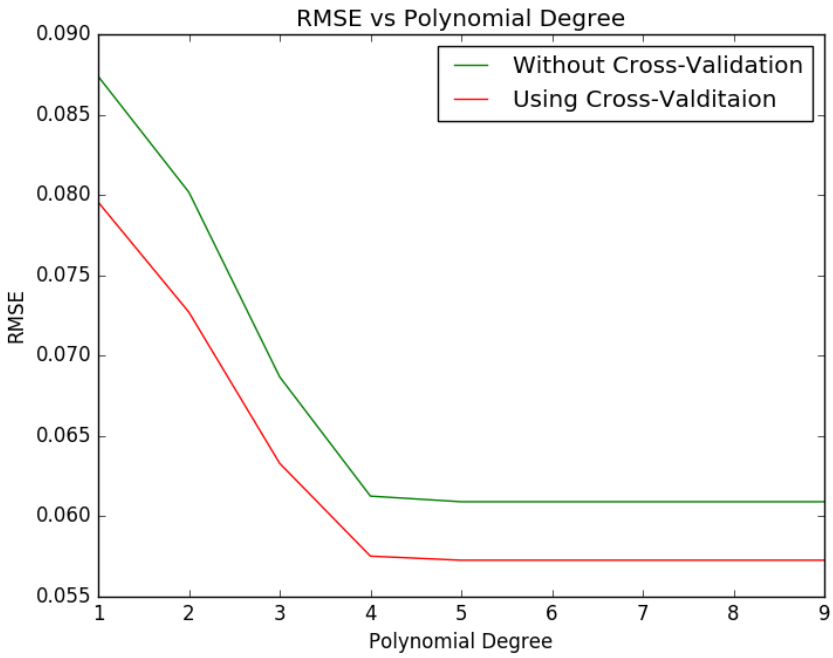

Figure 15: Polynomial Regression - RMSE vs Degree of Polynomial

8.1 How Cross-Validation Helps Control Complexity of the Model

1. We can use Cross Validation as a predictive metric of a statistical model. Statistical methods don't always yield accurate results. A high R-square value may not be the best way of concluding on the feasibility of a model, as the model may have been over-fitted using higher degrees of freedom. Thus a model made to fit on training data may not be very effective with real world data.
2. In cross validation, a part of the training data itself is used for model estimation. This makes the model less prone to over fitting. The error calculated using multiple folds of the data will provide a more conservative result. Therefore, Cross Validation provides a more unbiased measure of the error on new observations.
3. Cross Validation can be applied on several different methods of model training. The method with the least cross validation error can be chosen, thereby reducing the complexity of the model.

9 Boston Housing Dataset

The Boston-Housing Dataset has information of housing values in the suburbs of the greater Boston area. The features captured in data set are as follows:

- CRIM: per capita crime rate by town.
- ZN: proportion of residential land zoned for lots over 25,000 sq. ft.
- INDUS: proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: nitric oxides concentration (parts per 10 million)
- RM: average number of rooms per dwelling
- AGE: proportion of owner-occupied units built prior to 1940
- DIS: weighted distances to five Boston employment centers
- RAD: index of accessibility to radial highways
- TAX: full-value property-tax rate per $10,000
- PTRATIO: pupil-teacher ratio by town
- 1000(Bk 0.63)2 where Bk is the proportion of blacks by town
- LSTAT: lower status of the population
- MEDV: Median value of owner-occupied homes in $1000s

For details about the dataset look at Appendix A: Boston Housing Dataset.

10 Question 4a: Linear Regression

Linear regression was performed on the above mentioned dataset with MEDV as the dependent variable and rest other attributes as features. Ordinary least squares was used as the penalty function. As done with Network Backup dataset, we performed the 10 fold cross validation. The results obtained were as follows (using the ols function of statsmodels.formula.api).

1. The overall regression accuracy can be seen from R-square and adjusted R-square.
2. The R-square of 0.741 shows that the variance in MEDV is dependent by 74% on the remaining features. Thus the model is a good fit.
3. In the table above we notice that the p value of INDUS and AGE is very high which makes these features less significant.
4. RMSE Value of the estimator is 5.8889.

```
|                            OLS Regression Results
==============================================================================
Dep. Variable:                   MEDV   R-squared:                       0.741
Model:                            OLS   Adj. R-squared:                  0.735
Method:                 Least Squares   F-statistic:                     108.3
Date:                Sun, 29 Jan 2017   Prob (F-statistic):          5.89e-135
Time:                        15:19:55   Log-Likelihood:                -1495.5
No. Observations:                 505   AIC:                             3019.
Df Residuals:                     491   BIC:                             3078.
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     36.6349      5.102      7.180      0.000        26.610    46.659
CRIM          -0.1074      0.033     -3.270      0.001        -0.172    -0.043
ZN             0.0461      0.014      3.361      0.001         0.019     0.073
INDUS          0.0143      0.062      0.231      0.817        -0.107     0.135
CHAS           2.6711      0.861      3.102      0.002         0.979     4.363
NOX          -17.6336      3.819     -4.618      0.000       -25.137   -10.131
RM             3.7943      0.418      9.081      0.000         2.973     4.615
AGE            0.0011      0.013      0.081      0.935        -0.025     0.027
DIS           -1.4792      0.199     -7.420      0.000        -1.871    -1.088
RAD            0.3015      0.066      4.541      0.000         0.171     0.432
TAX           -0.0121      0.004     -3.202      0.001        -0.019    -0.005
PTRATIO       -0.9589      0.131     -7.329      0.000        -1.216    -0.702
B              0.0093      0.003      3.467      0.001         0.004     0.015
LSTAT         -0.5276      0.051    -10.400      0.000        -0.627    -0.428
==============================================================================
Omnibus:                      177.951   Durbin-Watson:                   1.083
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              783.814
Skew:                           1.522   Prob(JB):                     6.26e-171
Kurtosis:                       8.290   Cond. No.                      1.51e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.51e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 16: Linear Regression Results - Housing Dataset

10.1 Fitted vs. Actual Values

From the plot of Fitted Values vs Actual Values of MEDV , we can observe that majority of predicted values lie close to the regressed diagonal line. This explains the high R square value of this model. A high R square indicates that the goodness of fit is strong and thus the above plot should have majority points close to the diagonal. For a actual median value of 50, we see that the predicted values are underestimated by the model.

10.2 Residuals vs. Actual Values

From the plot of Residual Values vs. Fitted Values we can see that the residuals are close to the fitted line and a cluster can be seen in the middle of the plot. Also, there are no clear patterns visible which makes this plot an ideal one.

11 Question 4b: Polynomial Regression

We performed polynomial regression on the housing dataset starting from degree 1 to degree 10. We observed that the RMSE value decreased as we increased the degree of polynomial but increases drastically after increasing degree beyond 3. We can see the spike in the graph in Figure 19 beyond degree 3. The best

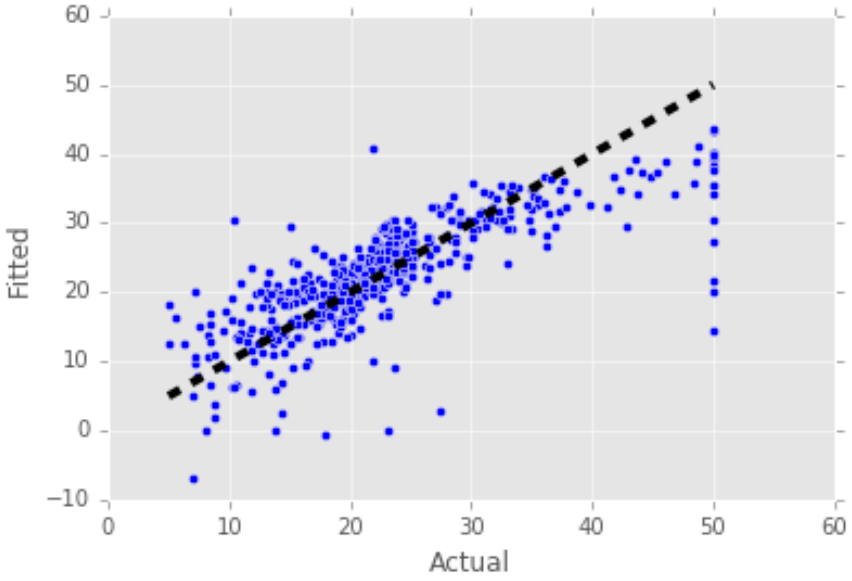

Figure 17: Linear Regression - Fitted Values vs. Actual Values of MEDV

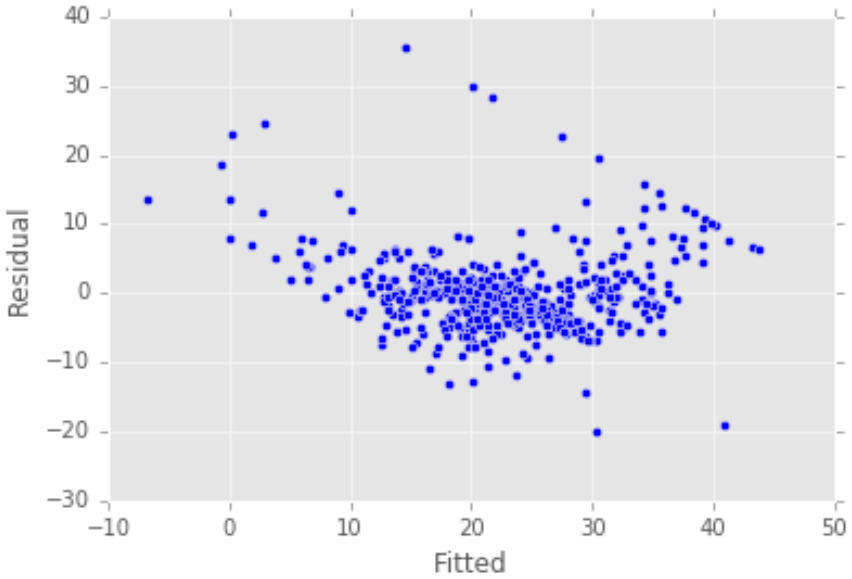

Figure 18: Linear Regression - Residual Values vs. Fitted Values of MEDV

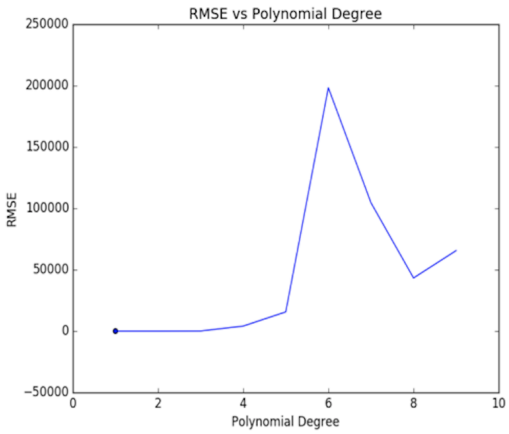

Figure 19: Polynomial Regression - RMSE vs. Degree of Polynomial

RMSE was obtained at degree 1 which was equal to 2.9512.

12 Question 5: Ridge Regression and Lasso Regularization

12.1 Ridge Regression

- To avoid overfitting by doing regularization of parameters we use ridge regression which is similar to least squares but shrinks the estimated coefficients towards zero.
- We used Ridge library along with Cross Val Predict for performing the ridge regression. We plotted the RMSE values against the alpha values in the range [1,0.1,0.01,0.001].

Best Alpha value for Ridge Regression : 1

Best RMSE for corresponding Alpha = 4.691833

12.2 Lasso Regularization

- The next regularization technique used was Lasso regularization.
- We used Lasso library along with Cross Val Predict for performing the lasso regularization. he plot of RMSE values against the alpha values of [1,0.1,0.01,0.001] looks like this:

Best Alpha value for Lasso Regularization : 0.01

Best RMSE for corresponding Alpha = 4.860201

The coefficients of parameters with un-regularized, ridge and lasso regression is compared in the table below:

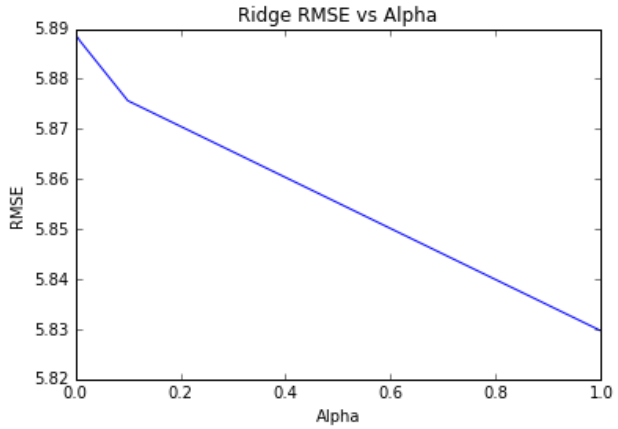

Figure 20: Plot of RMSE values against alpha values

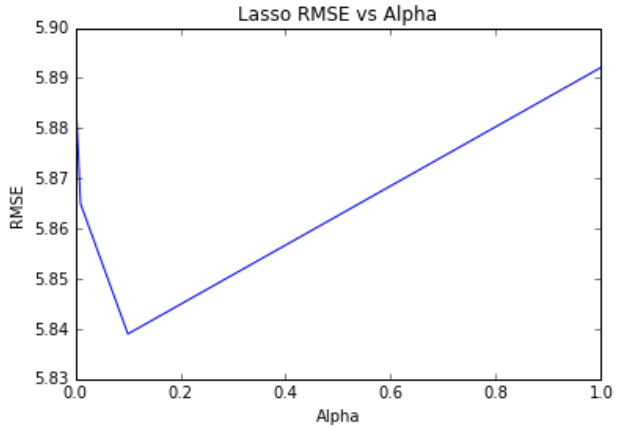

Figure 21: Plot of RMSE values against alpha values

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
Un Regularized	-0.1074	0.0461	0.0143	2.6711	-17.6336	3.7943	0.0011	-1.4792	0.3015	-0.0121	-0.9589	0.0093	-0.5276
Ridge (alpha = 1)	-0.1040	0.0471	-0.0151	2.5370	-10.6929	3.8373	-0.0049	-1.3771	0.2855	-0.0126	-0.8830	0.0096	-0.5362
Lasso (alpha = 0.01)	-0.0360	0.0130	-0.0030	2.3284	-8.4319	4.2193	-0.0	-0.7489	0.0	-0.0	-0.8263	0.0072	-0.5234

Figure 22: Parameter coefficients for different models