Project 4 Report

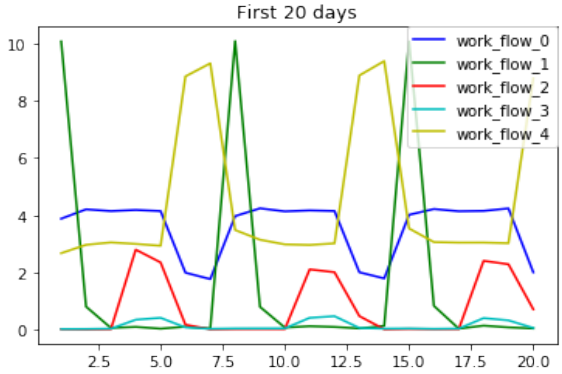
Mahmoud Essalat, Sherry He, Weitang Sun, Siqi Huang
ID: 005034839, 805040110, 904946260, 504490530

Dataset 1

Question 1.

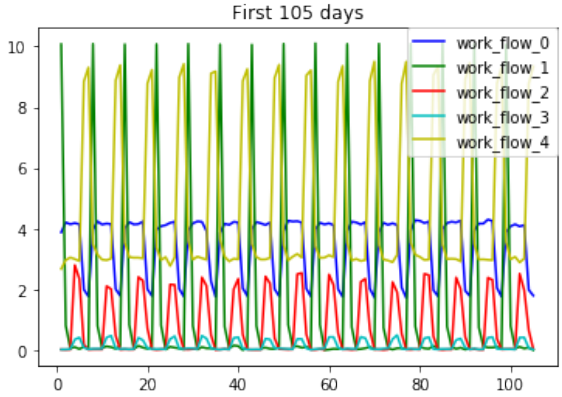
a)

The backup sizes for workflows (0-5) in the first twenty-day period is plotted in the following.



b)

The backup sizes for workflows (0-5) in the first 105-day period is plotted in the following.



c)

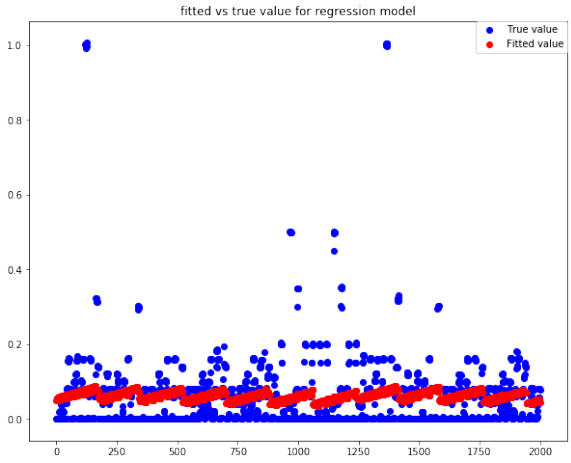
Both plots show that the backup sizes for each workflow repeats the same pattern every seven days.

Question 2.

a)

By using the function convert to scalar, we convert categorical feature such as day of the week, file name, work-flow-ID, into one dimensional numerical value, then directly fit these features into a basic linear regression model. The 10-fold average training RMSE is 0.103585. The 10-fold average test RMSE is 0.103676.

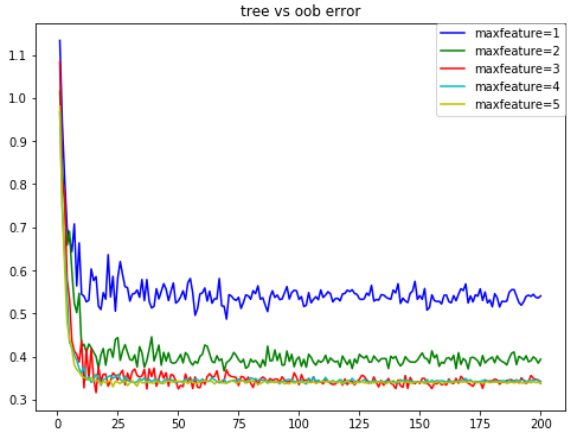
The following two scatter plots are the plot of fitted values versus fitted values and the plot of residual values versus fitted values. Since there are more than 18000 data points in the dataset, it's hard to visualize them all in one plot. Therefore, we choose the first 2000 points.

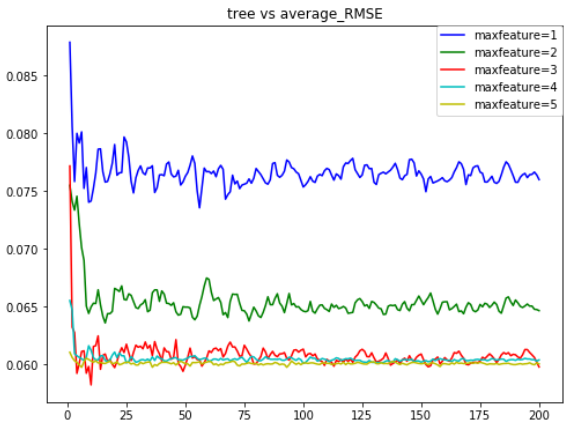


b)

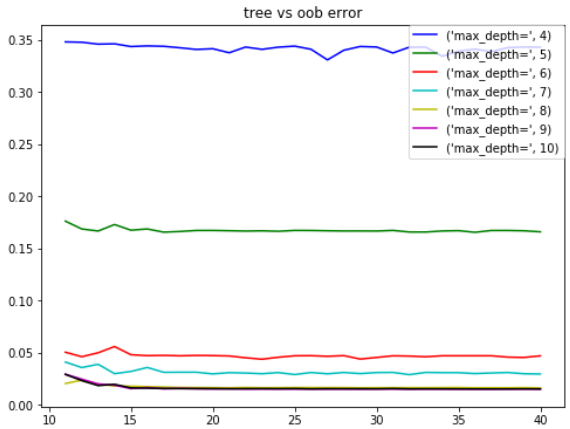
i) Here we use Random Forest Algorithm to fit dataset 1. For initial models, out-of-bag-error is 0.34195; 10-fold average training RMSE is 0.01156; 10-fold average test RMSE is 0.01299

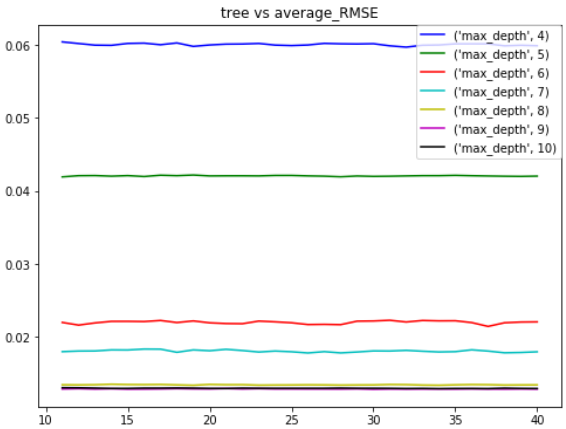
ii)





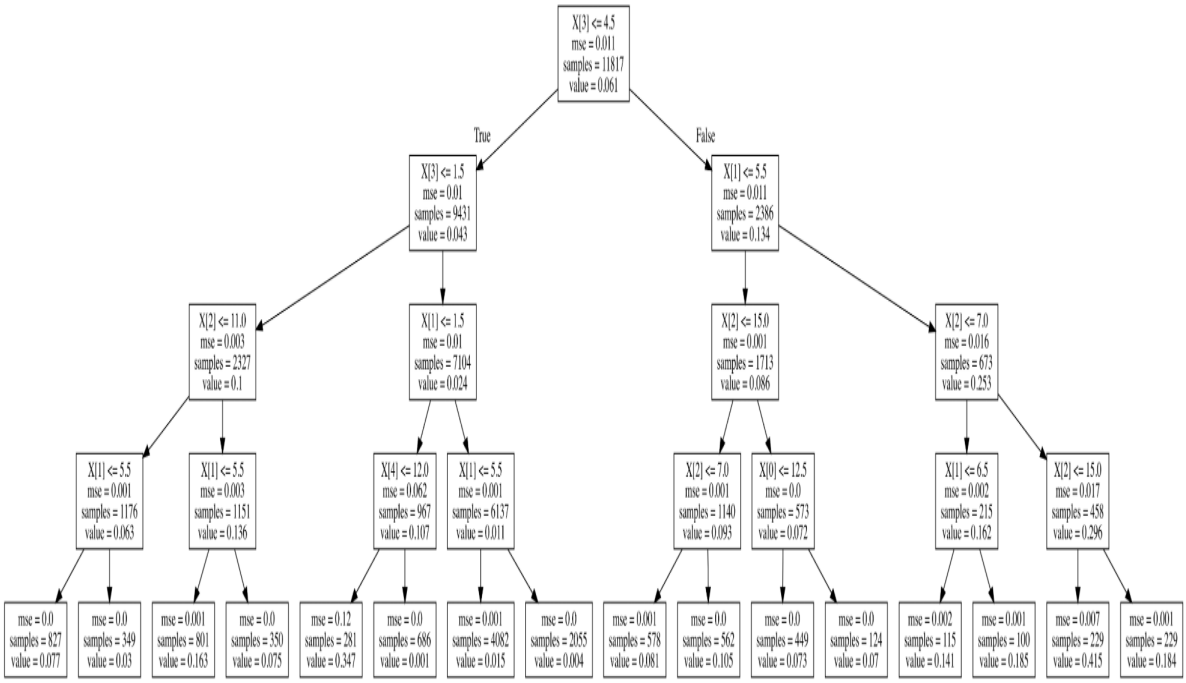
iii) We pick max_depth as another parameter and search from depth 4 to 10. From the figures we see the best parameter combination is: n_estimators=20, max_depth=9, max_features=5



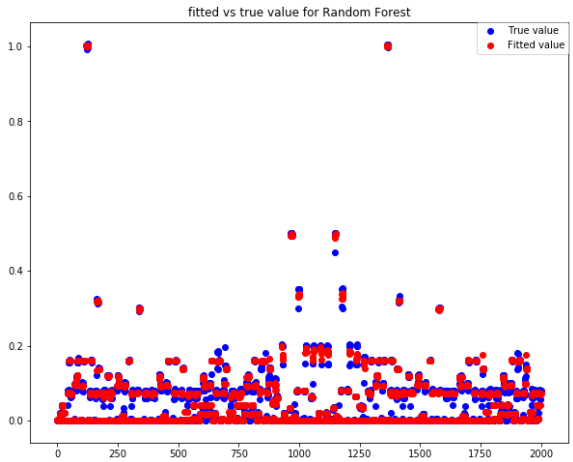


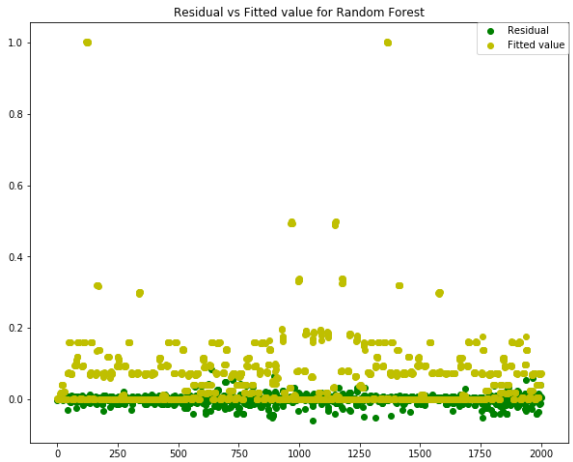
iv) The five features which are the week index, day of the week, backup start time, workflow ID, and file name, have the importance (from left to right) are [0.0017 0.1965 0.3989 0.2210 0.1819]. Hour of day is the most important feature.

v)



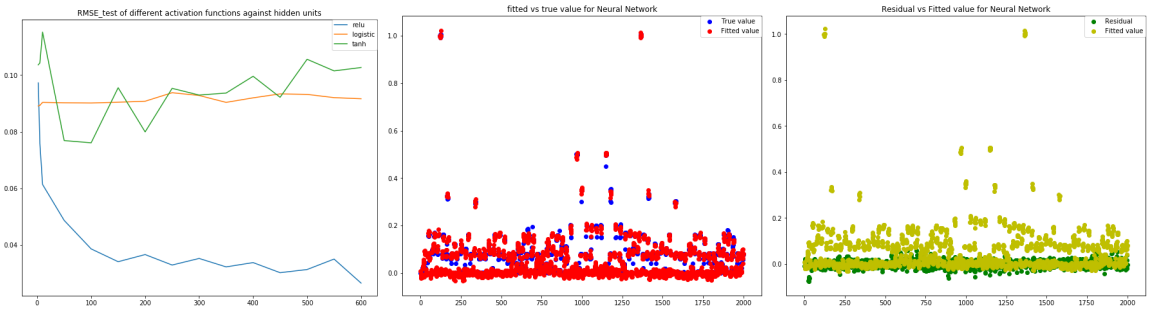
Root node is the node at the top X[4], however this is not the most important feature. The reason is that random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. Therefore, the Root node is not necessarily the most important feature.





c)

For neural network regression, average training RMSE is 0.0911, and average test RMSE is 0.1032. Best combination of parameters is hidden_layer_sizes=600 and activation='relu'



d)

i)

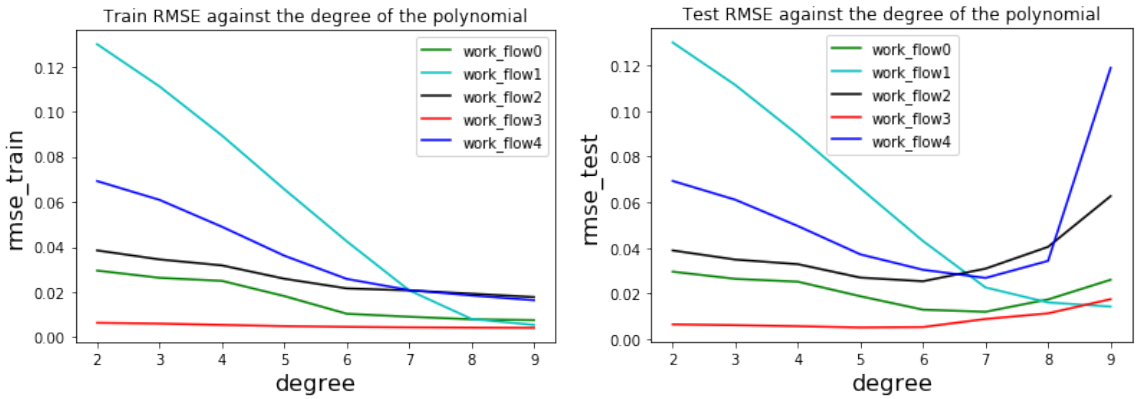
By using linear regression model to predict the back up size on each individual workflow, we are able to find the following results.

	Training RMSE	Test RMSE
Work Flow 0	0.035836	0.035887
Work Flow 1	0.148766	0.148919
Work Flow 2	0.042909	0.043067
Work Flow 3	0.007244	0.007261
Work Flow 4	0.085922	0.085991

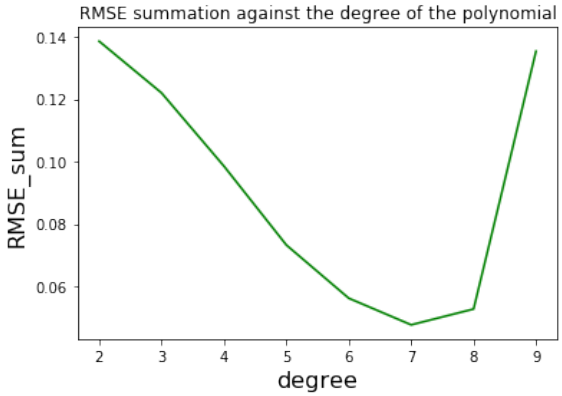
From the train-RMSE and test-RMSE results, we can see that workflows (0, 2, 3, 4) improves the fitting, however while work flow1 doesn't fit really well.

ii)

	Training RMSE	Test RMSE	Training RMSE	Test RMSE
	2 Degrees of Polynomial		3 Degrees of Polynomial	
Work Flow 0	0.029519	0.029540	0.026310	0.026388
Work Flow 1	0.129844	0.130097	0.111134	0.111459
Work Flow 2	0.038460	0.038860	0.034470	0.034852
Work Flow 3	0.006380	0.006426	0.006024	0.006079
Work Flow 4	0.069192	0.069341	0.060862	0.061113
	4 Degrees of Polynomial		5 Degrees of Polynomial	
Work Flow 0	0.024962	0.025151	0.018209	0.018755
Work Flow 1	0.089392	0.089623	0.065558	0.066162
Work Flow 2	0.031824	0.032836	0.025908	0.026937
Work Flow 3	0.005453	0.005622	0.025908	0.005064
Work Flow 4	0.048941	0.049574	0.004901	0.037128
	6 Degrees of Polynomial		7 Degrees of Polynomial	
Work Flow 0	0.010377	0.012865	0.009060	0.011939
Work Flow 1	0.042484	0.043003	0.020753	0.022609
Work Flow 2	0.021626	0.025341	0.020727	0.030863
Work Flow 3	0.004587	0.005237	0.004367	0.008778
Work Flow 4	0.025825	0.030372	0.020890	0.026785
	8 Degrees of Polynomial		9 Degrees of Polynomial	
Work Flow 0	0.007927	0.017353	0.007564	0.025990
Work Flow 1	0.008073	0.016038	0.005429	0.014216
Work Flow 2	0.019281	0.040424	0.017781	0.062682
Work Flow 3	0.004240	0.011253	0.004144	0.017470
Work Flow 4	0.018471	0.034276	0.016388	0.118946

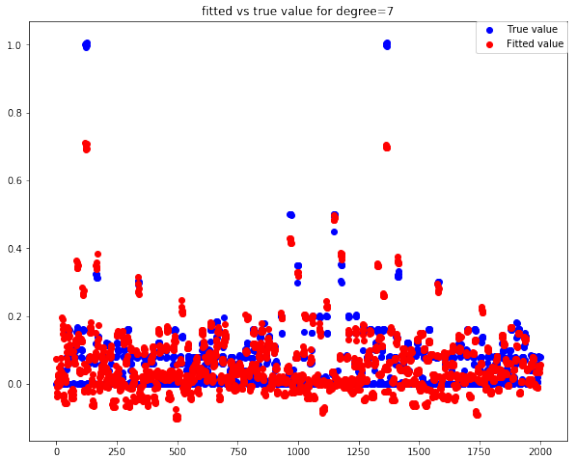


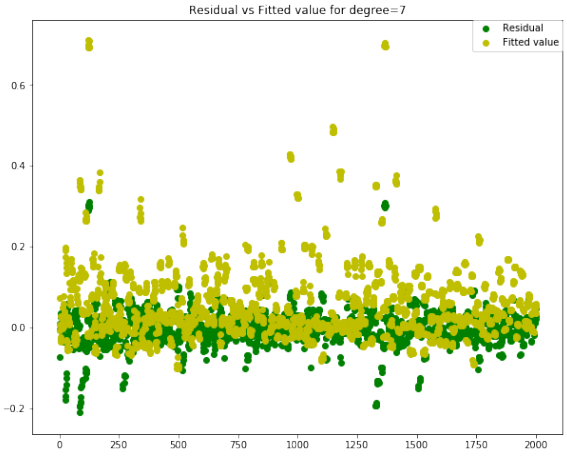
Using a polynomial function of our variables allows us to improve the fit. Figure shows that the Train-RMSE across degree varies from 2 to 9. Figure also shows that the Test-RMSE across degree varies from 2 to 9. We conclude that for workflow 0, the best polynomial degree is 7; for workflow 1, the best polynomial degree is 9; for workflow 2, the best polynomial degree is 6; for workflow 3, the best polynomial degree is 5; for workflow 4, the best polynomial degree is 7.



Combining the RMSE values on all workflow, we use the plot to show the polynomial function under degree 7 gives lowest RMSE error overall.

The following two scatter plots are the plot of fitted values versus fitted values and the plot of residual values versus fitted values. Since there are more than 18000 data points in the dataset, it's hard to visualize them all in one plot. Therefore, we choose the first 2000 points.

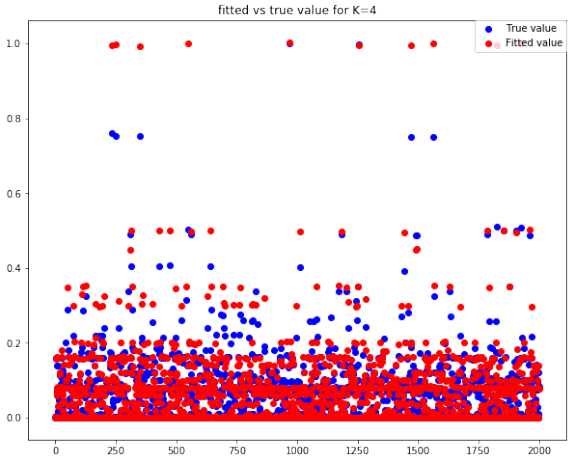


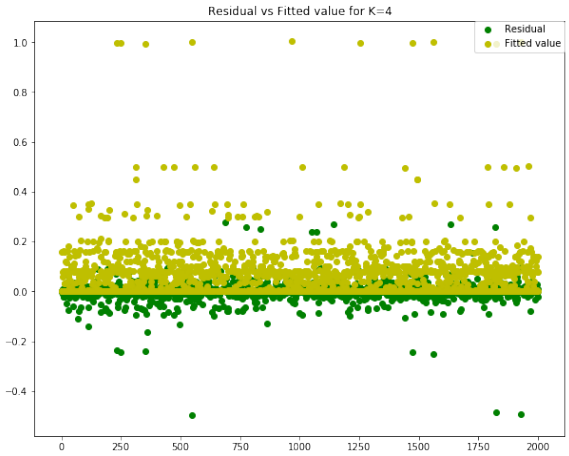


The higher degree of polynomial function will allow the model to train really well. We can observe this fact from the decreasing trend of train RMSE. However, a higher degree can cause overfitting, which leads to a huge jump in RMSE for test set after degree 8. By using 10-fold cross validation, it can help us in gauging the effectiveness of our model's performance. In other words, we can find a best degree function to our model with lower RMSE without overfitting.

(e)

By printing out training RMSE and test RMSE for k from 1 to 50, we can see that the minimum test RMSE is 0.03434 when k is 4.





Question 3.

	Training RMSE	Test RMSE
A) Linear Regression (scalar encoded categorical data)	0.10359	0.10368
B) Random Forest (scalar encoded categorical data)	0.01156	0.01299
C) Neural Network (one-hot encoded categorical data)	0.09115	0.10321
D)	0.01515	0.02019
E) K-nearest neighbor	0.02734	0.03474

Generally, Random Forest has the best result. Since all the experiments include categorical features, we can conclude that Random Forest is also the best at handling categorical features.

Dataset 2

Question 2.

a)

We set MEDV as the target variable and the rest 13 variables as the features, and fit in a linear regression model. The training RMSE is 4.487 and testing RMSE is 5.335.

b)

OLS Regression Results

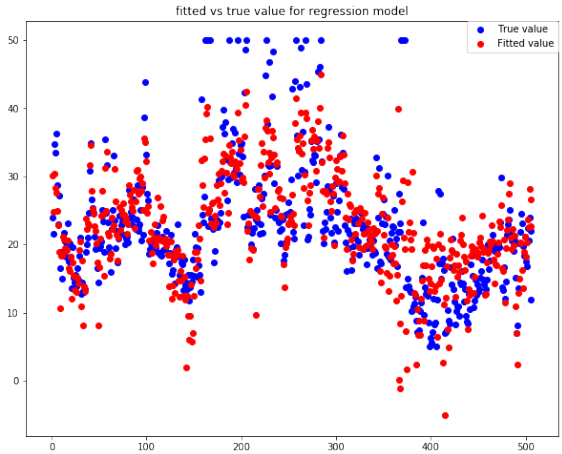
Dep. Variable:	y	R-squared:	0.741
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	108.1
Date:	Wed, 06 Mar 2019	Prob (F-statistic):	6.72e-135
Time:	11:02:16	Log-Likelihood:	-1498.8
No. Observations:	506	AIC:	3026.
Df Residuals:	492	BIC:	3085.
Df Model:	13		
Covariance Type:	nonrobust		

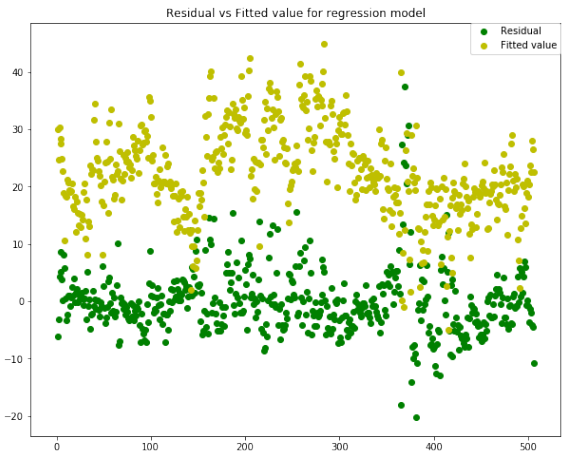
	coef	std err	t	P>|t|	[0.025	0.975]
const	36.4595	5.103	7.144	0.000	26.432	46.487
x1	-0.1080	0.033	-3.287	0.001	-0.173	-0.043
x2	0.0464	0.014	3.382	0.001	0.019	0.073
x3	0.0206	0.061	0.334	0.738	-0.100	0.141
x4	2.6867	0.862	3.118	0.002	0.994	4.380
x5	-17.7666	3.820	-4.651	0.000	-25.272	-10.262
x6	3.8099	0.418	9.116	0.000	2.989	4.631
x7	0.0007	0.013	0.052	0.958	-0.025	0.027
x8	-1.4756	0.199	-7.398	0.000	-1.867	-1.084
x9	0.3060	0.066	4.613	0.000	0.176	0.436
x10	-0.0123	0.004	-3.280	0.001	-0.020	-0.005
x11	-0.9527	0.131	-7.283	0.000	-1.210	-0.696
x12	0.0093	0.003	3.467	0.001	0.004	0.015
x13	-0.5248	0.051	-10.347	0.000	-0.624	-0.425

Omnibus:	178.041	Durbin-Watson:	1.078
Prob(Omnibus):	0.000	Jarque-Bera (JB):	783.126
Skew:	1.521	Prob(JB):	8.84e-171
Kurtosis:	8.281	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.

By using 10 fold cross validation, we obtained the *F* score and *p* values in each fold and conclude that NOX, RM, DIS, RAD, PTRATIO, and LSTAT are the most important features. CRIM, ZN, CHAS, TAX, and B are relatively less important features. The rest features are not that important. 10-fold training RMSE is 3.456 and 10-fold testing RMSE is 5.015





Question 3.

For Ridge regularizer, we used different alpha values to test our model and we found that when alpha = 0.0001, RMSE is the lowest.

Alpha	Training Set	Testing Set	Alpha	Training Set	Testing Set
0.0001	4.66553	9.43403	1	5.52672	15.2821
0.001	4.66556	9.45295	1.5	5.84254	16.5710
0.01	4.66788	9.63021	2	6.09582	17.5530

0.1	4.74545	10.8235	3	6.48046	18.9944
0.5	5.12665	13.4235			

For Lasso regularizer, we used different alpha values to test our model and we found that when alpha = 0.0001, RMSE is the lowest.

Alpha	Training Set	Testing Set	Alpha	Training Set	Testing Set
0.0001	4.66557	9.45415	1	9.11536	27.5216
0.001	4.66855	9.67031	1.5	9.11536	27.5216
0.01	4.85090	11.7926	2	9.11536	27.5216
0.1	5.92013	17.8369	3	9.11536	27.5216
0.5	9.11536	27.5216			

The coefficients of the three models are presented in the following table:

Ridge, Alpha = 0.0001, 10-fold training RMSE = 4.6259, test RMSE = 5.8906, coefficients are presented in the following

-0.1056 0.0491 0.0320 2.5139 -17.6164 3.8184 0.0106 -1.4356 0.3609 -0.0154 -0.9121 -0.0099 -0.5550

Lasso, Alpha = 0.0001, 10-fold training RMSE = 4.6259, test RMSE = 5.8881, coefficients are presentedin the following

-0.1049 0.0488 0.0294 2.5161 -17.4741 3.8227 0.0102 -1.4333 0.3563 -0.0152 -0.9097 -0.0098 -0.5544

Unregularized coefficients are presented in the following

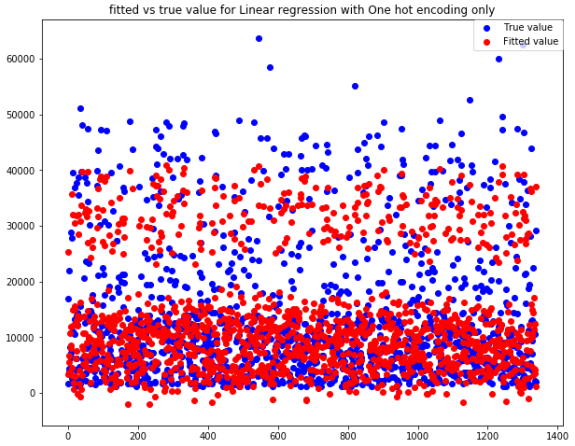
-0.1056 0.0491 0.0322 2.5129 -17.6280 3.8177 0.0106 -1.4361 0.3615 -0.0155 -0.9123 0.0099 -0.5551

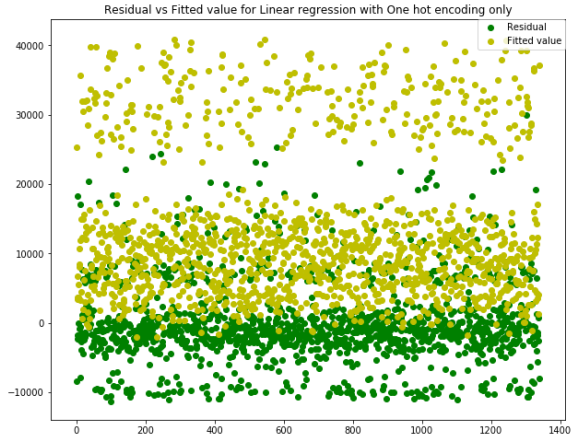
Dataset 3

Question 1.

a)

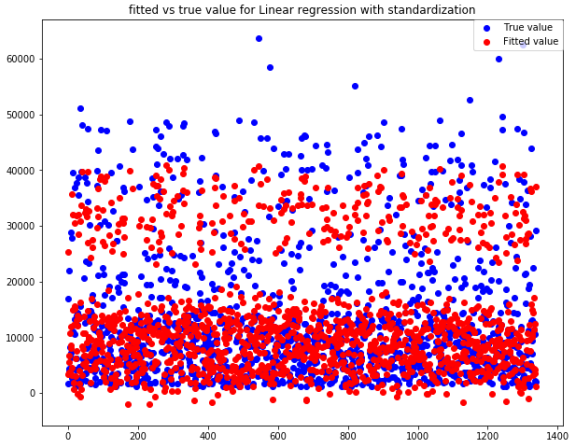
Feature encoding: 10-fold training RMSE is 6039.57, and 10-fold testing RMSE is 6081.96.

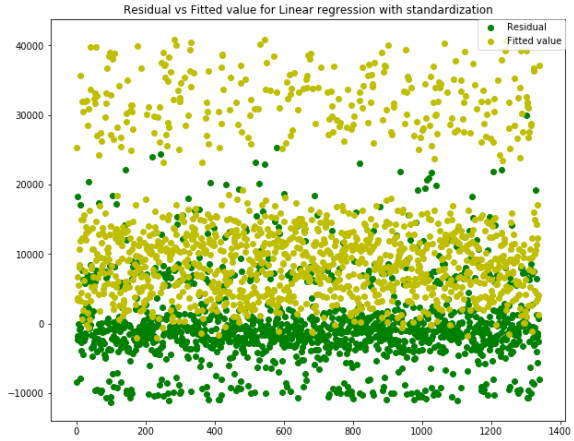




b)

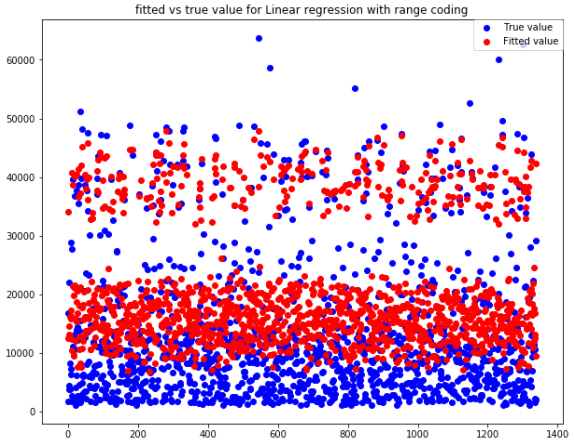
Standardization: 10-fold training RMSE is 6039.57, and 10-fold testing RMSE is 6081.96.





c)

10-fold training RMSE is 6198.08, and 10-fold testing RMSE is 6239.97.



From the RMSE numbers, we can see that feature encoding and standardization perform similarly well.

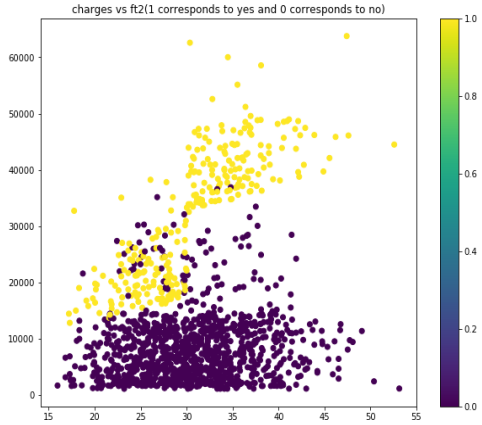
Question 2. Correlation Exploration

a)

Most important two features by *f*-regression are ft1 and ft5, and most important two features by mutual info regression are ft1 and ft5.

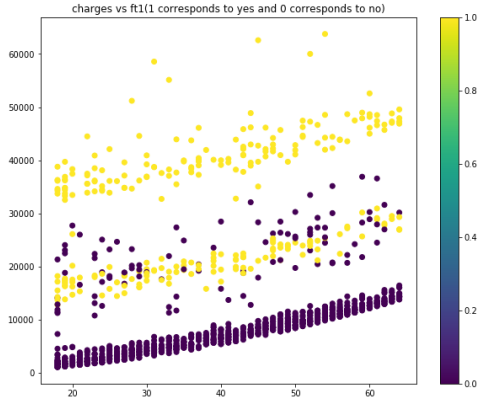
b)

We can see that with ft2 and ft5, data are nicely separated.



c)

With charges vs ft1, data are nicely separated as well.

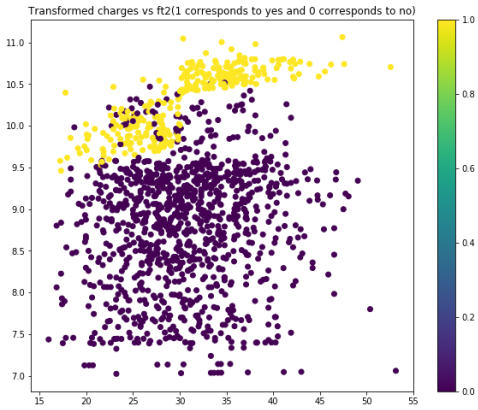


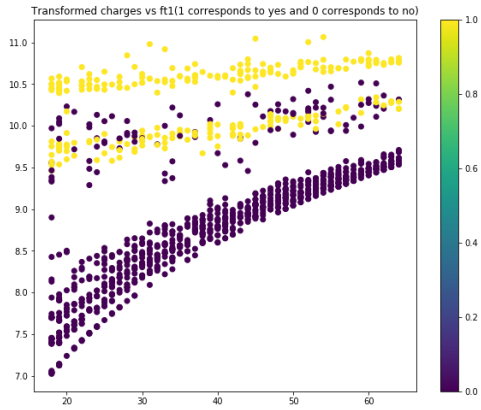
Question 3*. Modify the target variable*

a)

10-fold training RMSE is 18846.87, 10-fold testing RMSE is 18960.56, and the whole dataset RMSE is 8363.97. With the target variable transformed, RMSE actually increases and performance degrades.

b)





Question 4. *Bonus questions*

a)

Using polynomial feature encoding, we observe an improvement of the performance. Using order 2 polynomial, training RMSE is 5802.59, testing RMSE is 5984.33. However, using order 3 polynomial, training RMSE increases to 6850.02, and testing RMSE increase to 7313.77.

b)

i)

Three types of models are applied to improve the results.

	Training RMSE	Test RMSE
A) Random Forest	5125.91	5370.09
B) Neural Network	5748.41	5845,07
C) Gradient boosting	3937.39	4528.34

ii)

By modifying the hyper-parameter of each model, we can a general improvement.

	Training RMSE	Test RMSE
A) Random Forest	4131.89	4499.54
B) Neural Network	5486.67	5619.26
C) Gradient boosting	3915.10	4530.75

Conclusion.

In dataet1, majority of the features are used for scalar encoding. We used on different models to test the performance. Since the range of value in each feature is compact, the overall RMSE is relatively low. We conclude the random forest model is the best for handling categorical data. Random forest has the benefits of generation an internal unbiased estimate of the generalization error as the forest building progresses over other models. More importantly, it can perform efficiently on large databases. The improvements on dataset3 are dramatic with random forest model. From the dataset2, we observe that some of the features in Dataset2 have wider range of values compare to others. Since we didn't exclude some of the outliers in some features, it can have small impact on our RMSE error. That's why using linear regression model in dataset2 generates higher RMSE error than dataset1. Also, from the experiments, we find regular linear regression model has better results than adding addition regularizers. It is making sense for the reason that we pick the smallest alpha.