EE219 Project 4: Regression Analysis

Team members:

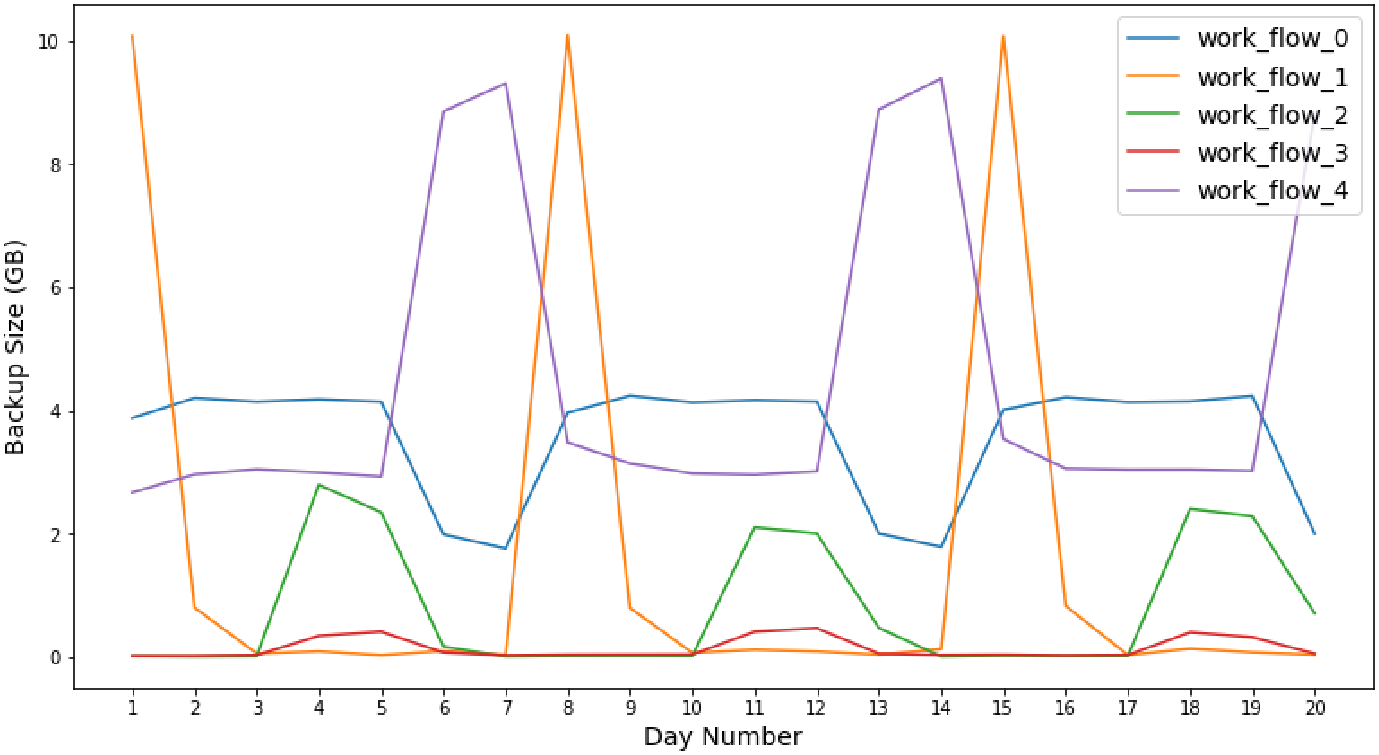
- **Yin Fei** 404284074
- **Yichen Wu** 504294181
- **Siyuan Chen** 405024391
- **Ruchen Zhen** 205036408

Required Packages:

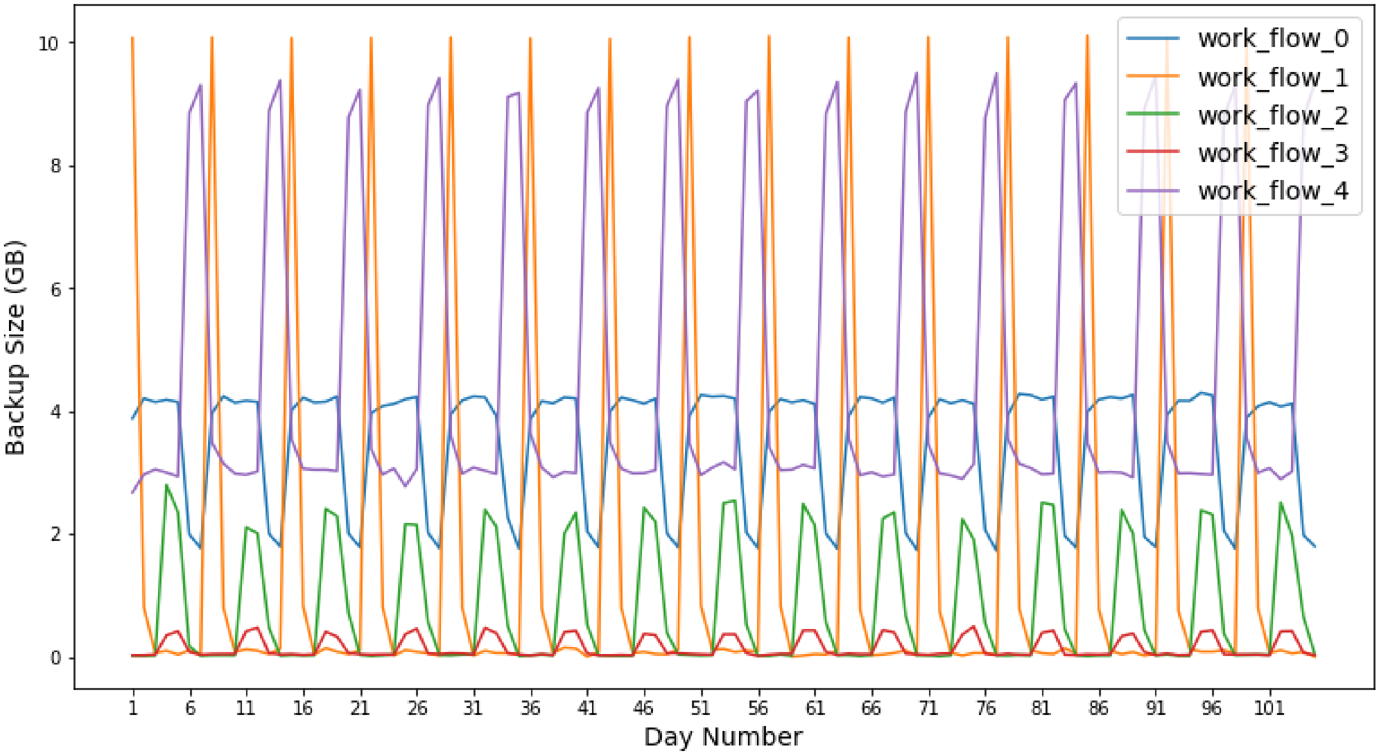
python 3.6
numpy v1.14.0
scikit-learn v0.19.1
scipy v1.0.0
matplotlib v2.1.2
pandas v0.22.0
graphviz v2.38.0

Part 1) Load the dataset.

(a) For a twenty-day period (X-axis unit is day number) plot the backup sizes for all workflows (color coded on the Y-axis)



(b) Do the same plot for the first 105-day period.

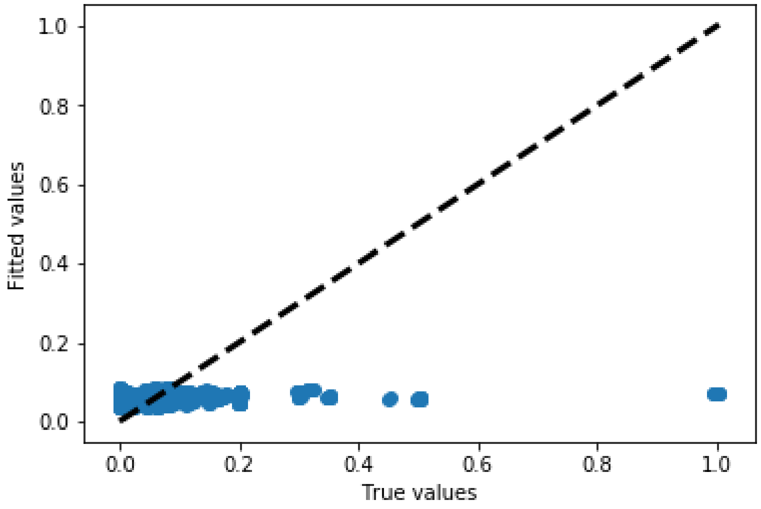


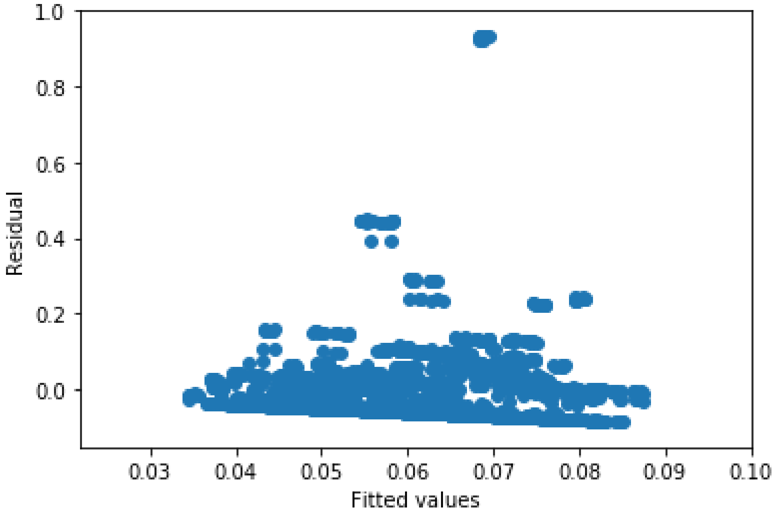
We can see an obvious periodic pattern of the data on the graph. Periods of all 5 workflows are the same, which is approximately 7 days, the number of days of a week. The peaks of workflow 4 appear when the troughs of workflow 0 appear. The peaks of workflow 2 and 3 appear at the same time, leading the peak of workflow 4 by approximately 2 days. The peaks of workflow 1 are lagging the peaks of workflow 4 by approximately 1 day.

Part 2) Predict the backup size of a file given the other attributes.

a) Fit a linear regression model.

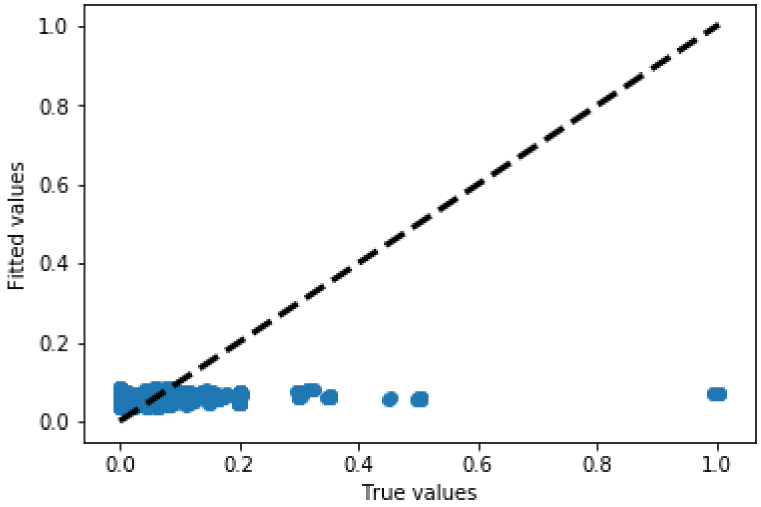
i) Naive linear regression

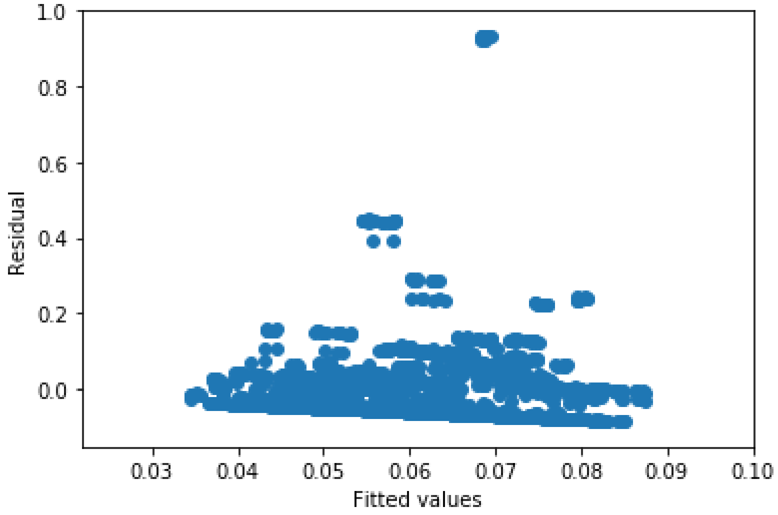




Train set average RMSE: 0.103585393643
Test set average RMSE: 0.103675847676

ii) Data Preprocessing: with standardized feature data

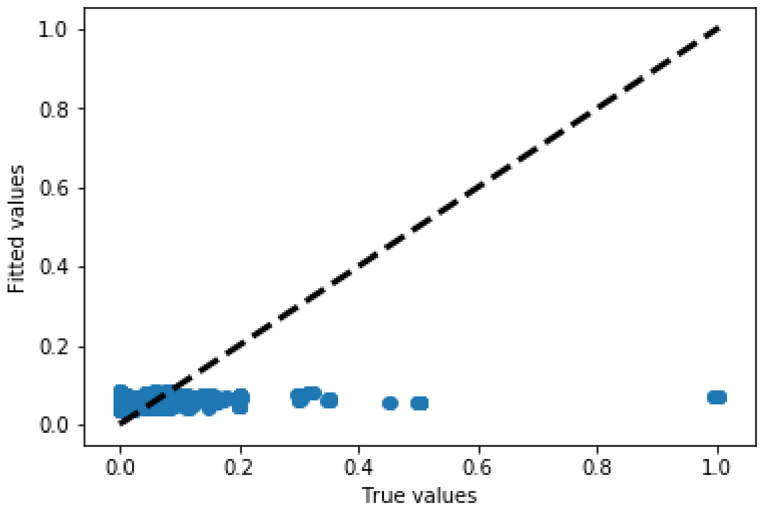


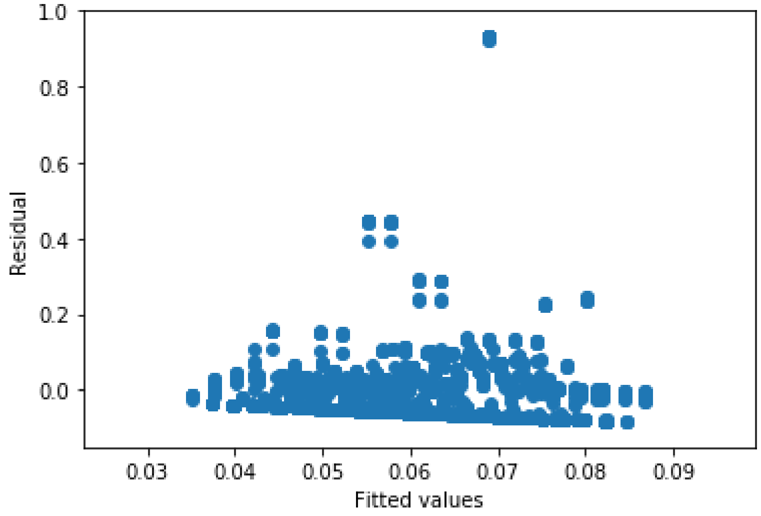


The fitting result does not change as shown in the plots. After the standardization, the RSME for trin set and test set does not change.

Train set average RMSE: 0.103585393643
Test set average RMSE: 0.103675847676

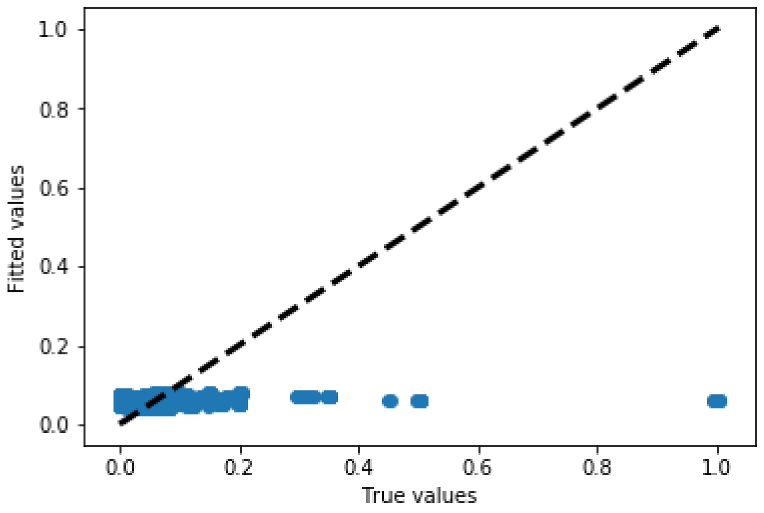
iii) Feature Selection: select three most important features

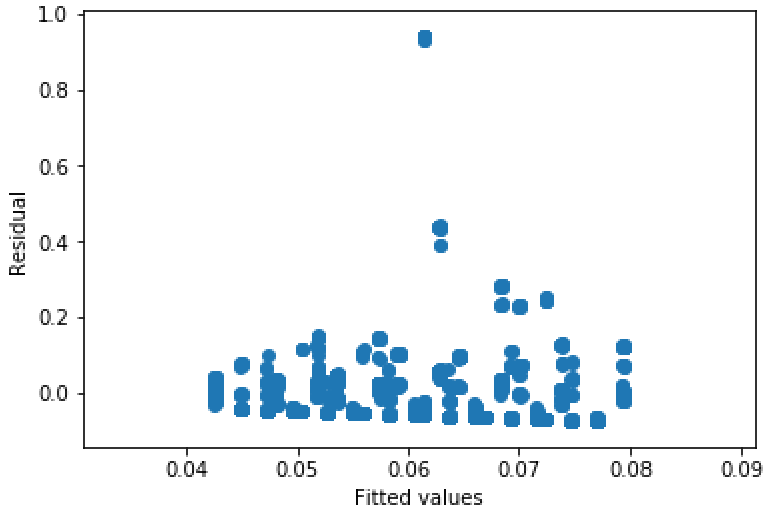




For f-regression, our result is: [false true true true false], which means that the second, the third and the fourth variables are the most important three variables.

Train set average RMSE: 0.103585682142
Test set average RMSE: 0.103670661831



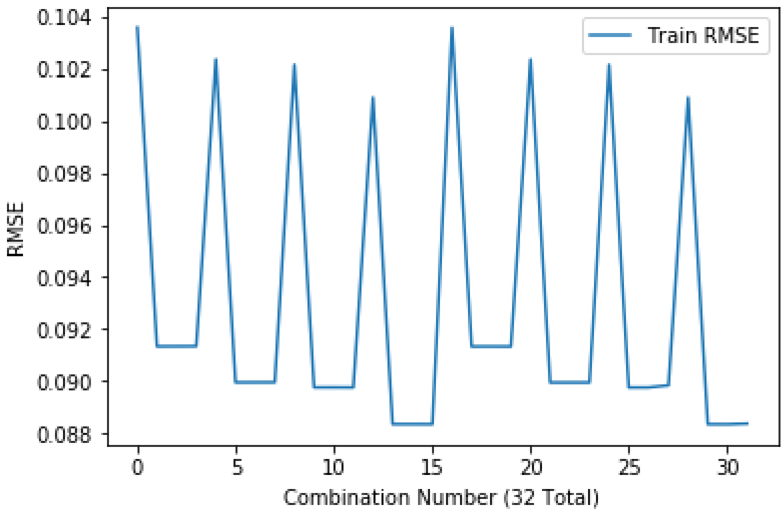


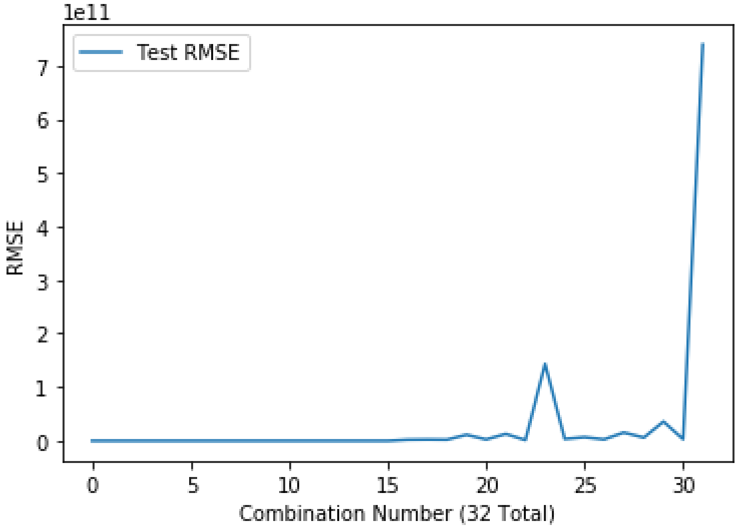
For mutual information regression, our result is: [false false true true true], which means that the third, the fourth and the fifth variables are the most important three variables.

Train set average RMSE: 0.103694528194
Test set average RMSE: 0.103772293071

The performance of prediction does not improve when we choose three most important variables to train the model.

iv) Feature Encoding:





The combination which achieves the best performance is [false true true true true], of which the true values correspond to “day of the week”, “backup start time-hour of day”, “work-flow ID” and “file name”. One-Hot-Encoding performs better than scalar encoding because for variables like “day of week”, Monday and Sunday are one day apart, but the scalar encoding gives them value 1 and 7, whose difference is 6. Similar reason is applied to other three variables in this combination.

v) Controlling ill-conditioning and over-fiting:

Significant increases in test RMSE compared to train RMSE are found in some combinations. The reason might be overfitting.

With Ridge Regularizer

```
With ridge regression, best alpha
and best combination is: (3, (False, True, True, True, False))
['Day of Week', 'Backup Start Time - Hour of Day', 'Work-Flow-ID']
Test RMSE: 0.0883677426258
```

With Lasso Regularizer

```
With ridge regression, best alpha
and best combination is: (0.0004572473708276177, (True, True, True, True, True))
['Week #', 'Day of Week', 'Backup Start Time - Hour of Day', 'Work-Flow-ID', 'File Name']
Test RMSE: 0.0884676783956
```

coefficients of best unregularized model

```
[ 1.96145460e+10 1.96145460e+10 1.96145460e+10 1.96145460e+10 1.96145460e+10 1.96145460e+10
1.96145460e+10 -2.89569745e+10 -2.89569745e+10 -2.89569745e+10 -2.89569745e+10 -2.89569745e+10
-2.89569745e+10 1.65056477e+10 -4.39256988e+10 -1.67303354e+10 -2.26495354e+09 -5.20965060e+10
-2.98779213e+10 -2.98779213e+10 -2.98779213e+10 -2.98779213e+10 -2.98779213e+10 -2.98779213e+10
3.05534252e+10 3.05534252e+10 3.05534252e+10 3.05534252e+10 3.05534252e+10 3.05534252e+10
3.35806181e+09 3.35806181e+09 3.35806181e+09 3.35806181e+09 3.35806181e+09 3.35806181e+09
-1.11073201e+10 -1.11073201e+10 -1.11073201e+10 -1.11073201e+10 -1.11073201e+10 -1.11073201e+10
3.87242323e+10 3.87242323e+10 3.87242323e+10 3.87242323e+10 3.87242323e+10 3.87242323e+10
8.94069672e-06]
```

coefficients of best regularized model (Ridge)

```
[ 3.92669163e-02 -1.28487144e-02 -2.02445927e-02 -5.24244158e-03 -5.69974195e-03 3.27354301e-03
1.49503134e-03 -2.01859243e-02 -2.10402115e-02 7.78402332e-03 3.34228557e-02 -1.98663507e-03
2.00589195e-03 3.88382162e-02 -1.37318537e-02 -4.01695350e-02 -5.71824730e-02 7.22456456e-02
1.12166970e-05 5.37811040e-05]
```

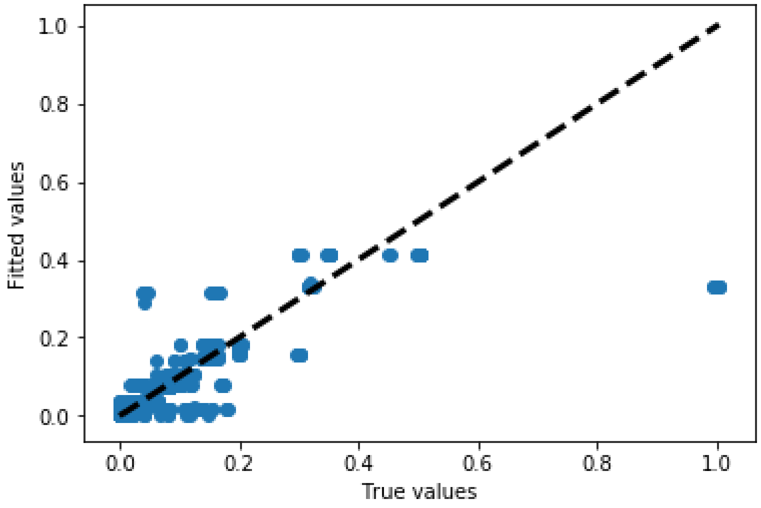
coefficients of best regularized model (Lasso)

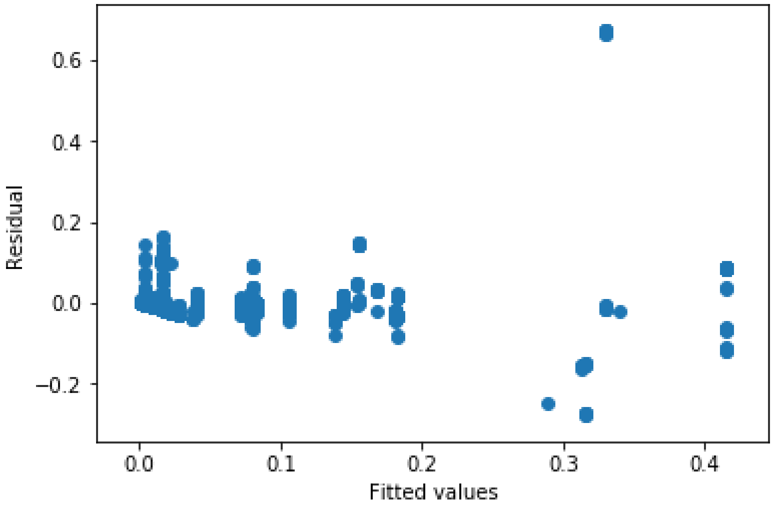
```
[-0. -0. 0. -0. 0. -0. 0. 0. 0. 0. -0. 0. -0. -0. -0. 0.04001671 -0.00550534 -0.01322969 -0. -0. 0.00401878
0.00223829 -0.01755974 -0.01823346 0.00504044 0.0307037 -0. 0. 0.0499828 -0. -0.02400039 -0.04072367
0.08463377 0. 0. 0. 0. 0. 0. 0. -0. 0. -0. -0. 0. -0. -0. -0. -0. -0. -0. -0. -0. -0. -0. -0. -0. 0. 0. 0. 0. 0. 0. ]
```

By comparing the regularized models and unregularized models, we can find that the coefficients of regularized models are smaller than that of the best unregularized model.

b) Use a random forest regression model for this same task.

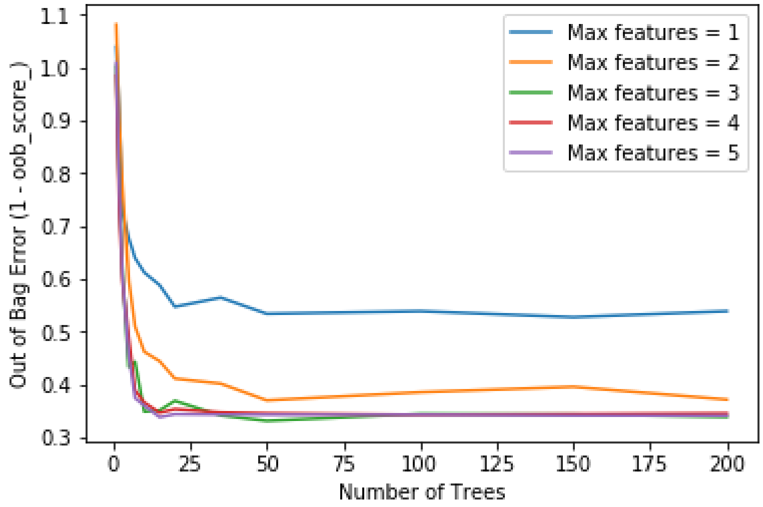
i) Out of bag error

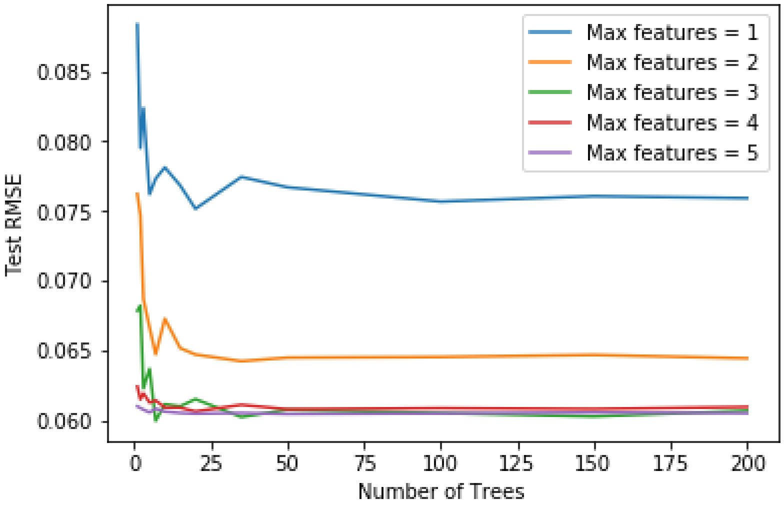




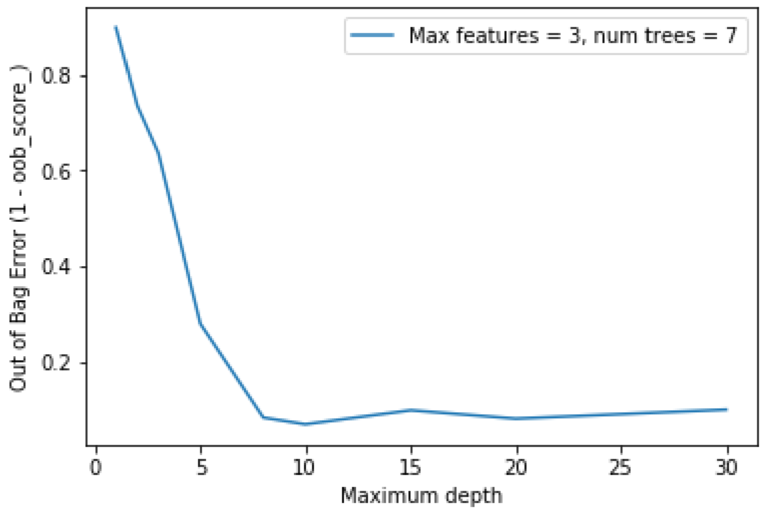
Train set average RMSE: 0.0606309314186
Test set average RMSE: 0.0607161064476
Out of bag error: 0.342134001925

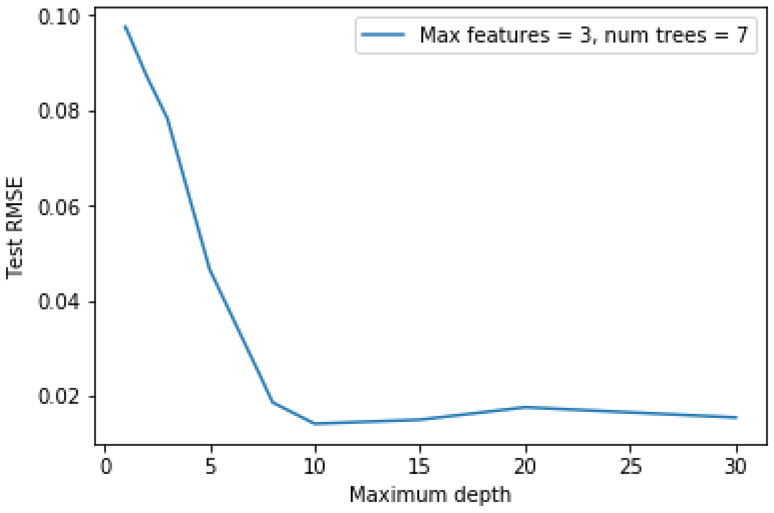
ii) Sweep over number of trees from 1 to 200 and maximum number of features from 1 to 5





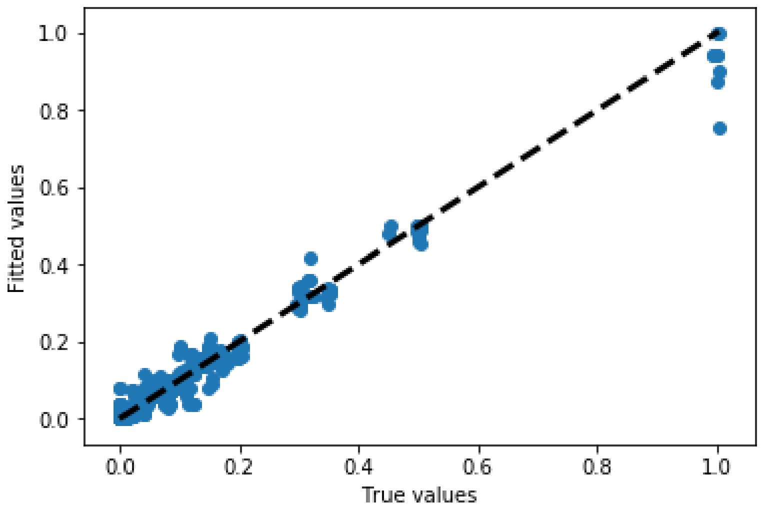
iii) Pick another parameter to experiment on — max_depth

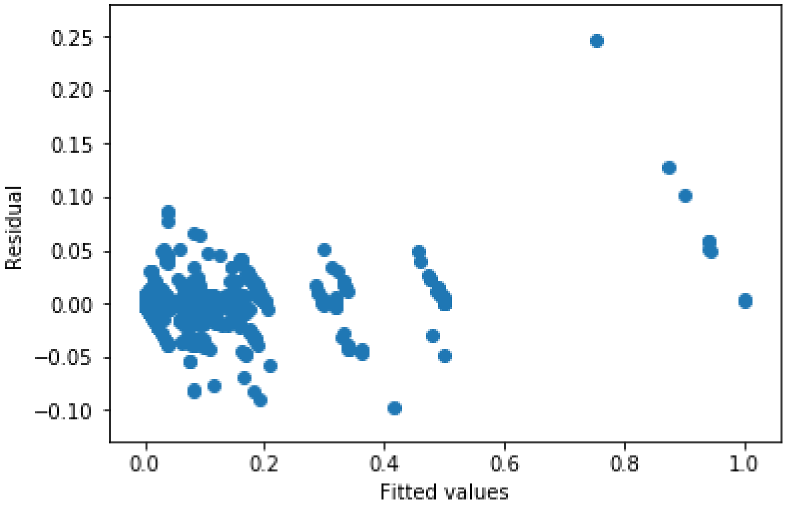




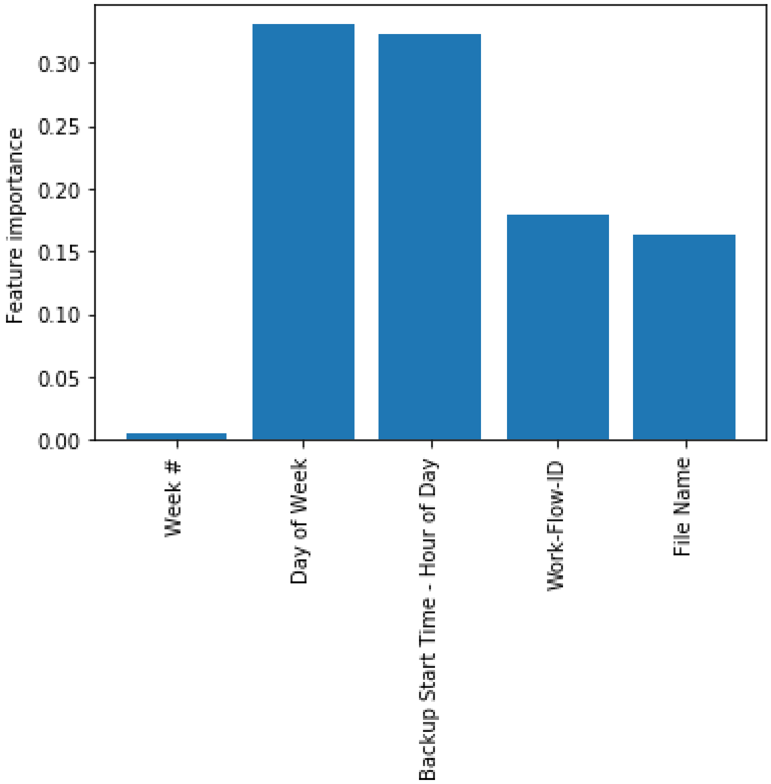
The other variable we pick to achieve the best performance is the maximum depth.

iv) Report the feature importance from the best random forest regression.

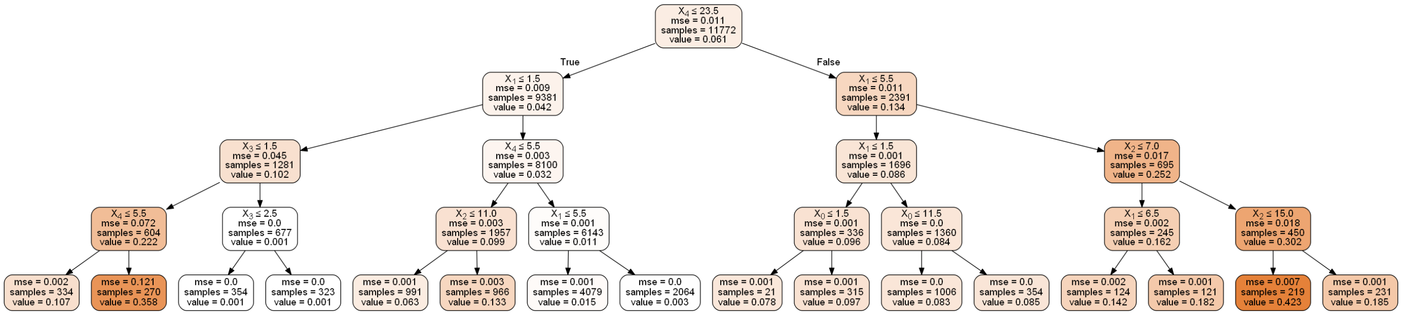




```
Train set average RMSE: 0.0121801827691
Test set average RMSE: 0.0142939499261
Feature importance:
[ 0.00474529 0.33020781 0.32231875 0.17955937 0.16316878]
```

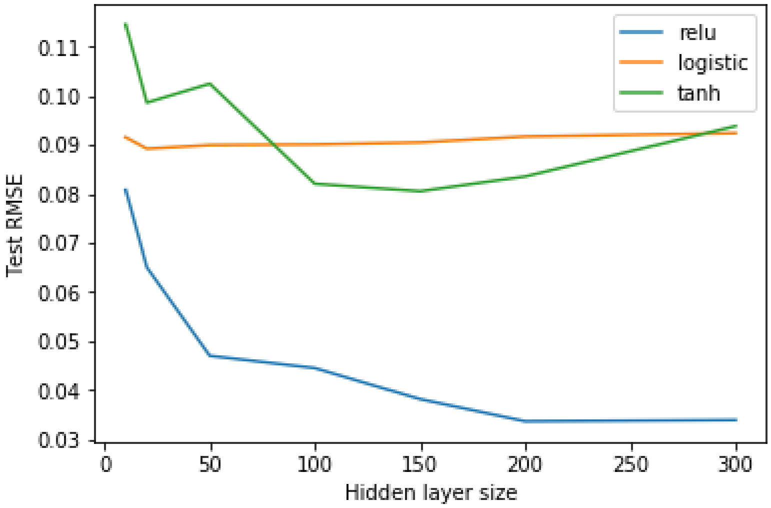


v) Visualize your decision trees.


$X_4 \leq 23.5$ mse = 0.011 samples = 11772 value = 0.061
True
False
$X_1 \leq 1.5$ mse = 0.009 samples = 9381 value = 0.042
$X_1 \leq 5.5$ mse = 0.011 samples = 2391 value = 0.134
$X_3 \leq 1.5$ mse = 0.045 samples = 1281 value = 0.102
$X_4 \leq 5.5$ mse = 0.003 samples = 8100 value = 0.032
$X_1 \leq 1.5$ mse = 0.001 samples = 1696 value = 0.086
$X_2 \leq 7.0$ mse = 0.017 samples = 695 value = 0.252
$X_4 \leq 5.5$ mse = 0.072 samples = 604 value = 0.222
$X_3 \leq 2.5$ mse = 0.0 samples = 677 value = 0.001
$X_2 \leq 11.0$ mse = 0.003 samples = 1957 value = 0.099
$X_1 \leq 5.5$ mse = 0.001 samples = 6143 value = 0.011
$X_0 \leq 1.5$ mse = 0.001 samples = 336 value = 0.096
$X_0 \leq 11.5$ mse = 0.0 samples = 1360 value = 0.084
$X_1 \leq 6.5$ mse = 0.002 samples = 245 value = 0.162
$X_2 \leq 15.0$ mse = 0.018 samples = 450 value = 0.302
mse = 0.002 samples = 334 value = 0.107
mse = 0.121 samples = 270 value = 0.358
mse = 0.0 samples = 354 value = 0.001
mse = 0.0 samples = 323 value = 0.001
mse = 0.001 samples = 991 value = 0.063
mse = 0.003 samples = 966 value = 0.133
mse = 0.001 samples = 4079 value = 0.015
mse = 0.0 samples = 2064 value = 0.003
mse = 0.001 samples = 21 value = 0.078
mse = 0.001 samples = 315 value = 0.097
mse = 0.0 samples = 1006 value = 0.083
mse = 0.0 samples = 354 value = 0.085
mse = 0.002 samples = 124 value = 0.142
mse = 0.001 samples = 121 value = 0.182
mse = 0.007 samples = 219 value = 0.423
mse = 0.001 samples = 231 value = 0.185

The root node of the decision tree that we pick is feature 4, which is work-flow ID. It is not the most important feature reported by regressor, which is feature 2, day of week.

c) Use neural network regression model

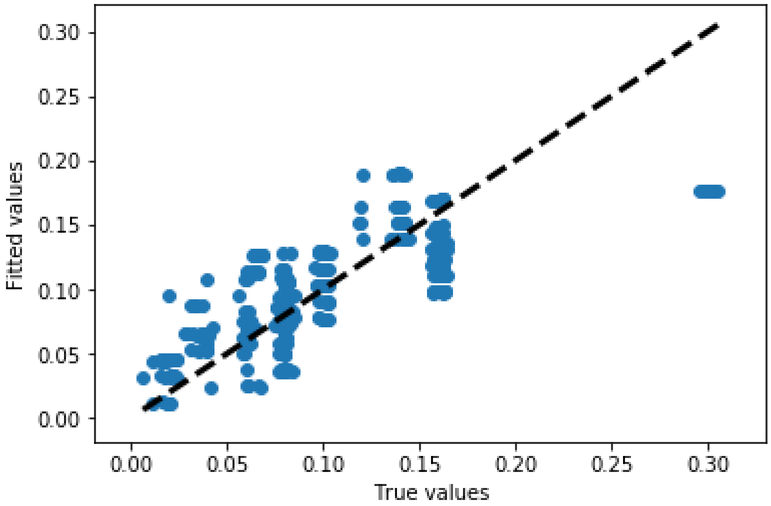


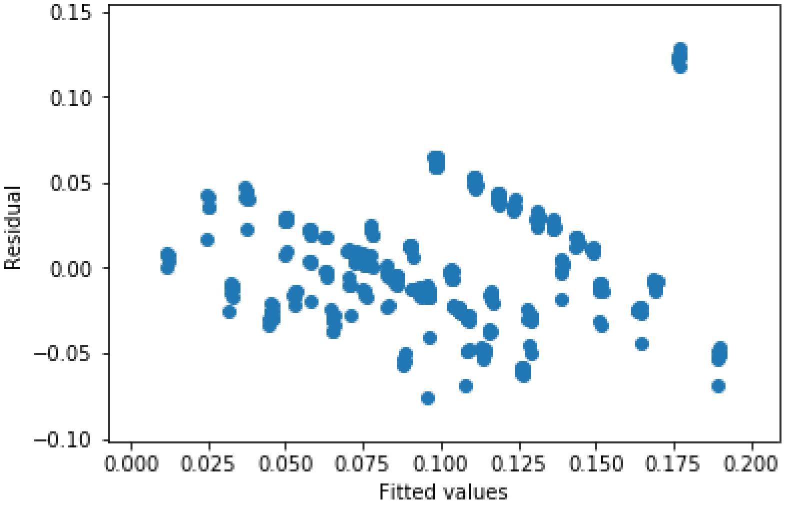
The best combination is 300 hidden layer units with activity function relu.

d) Predict the Backup size for each workflows separately

i) With linear regression model

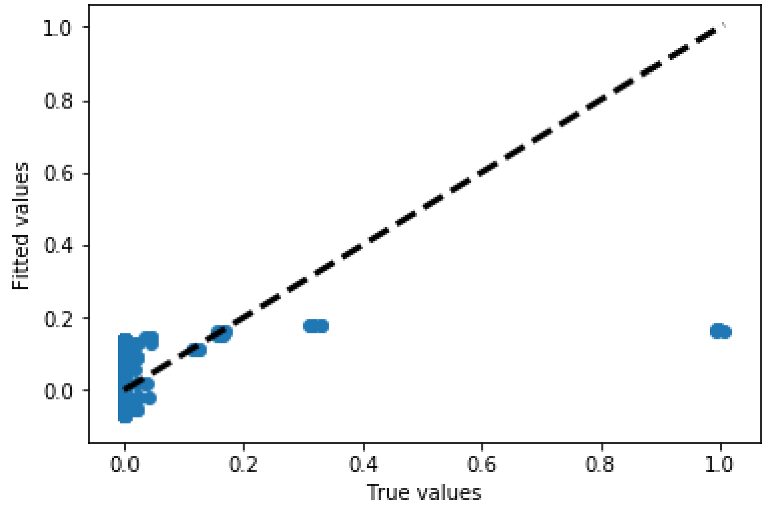
work_flow_0

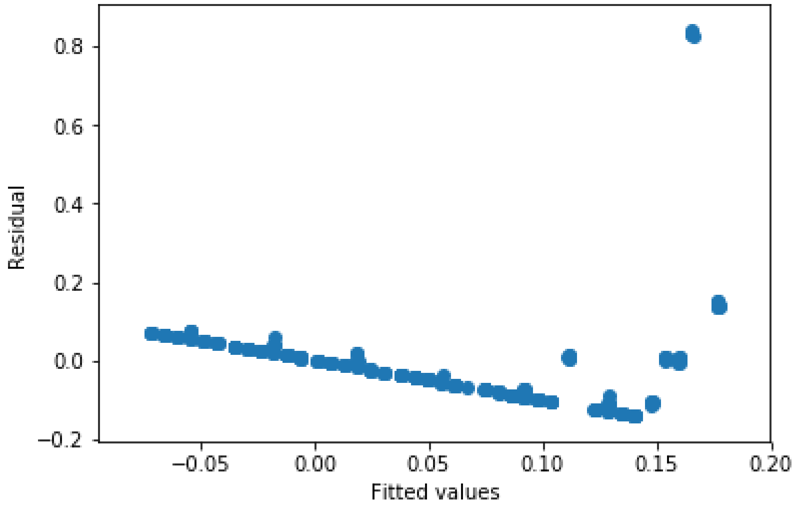




Train set average RMSE: 0.0358355207799
Test set average RMSE: 0.0358869702489

work_flow_1

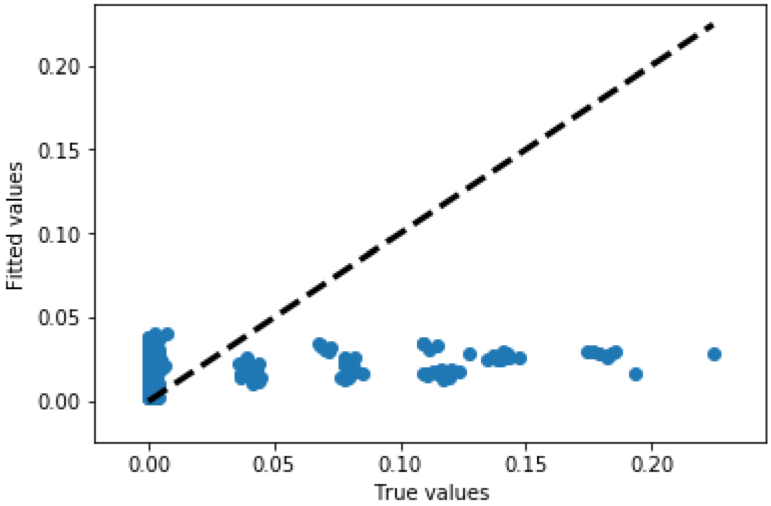


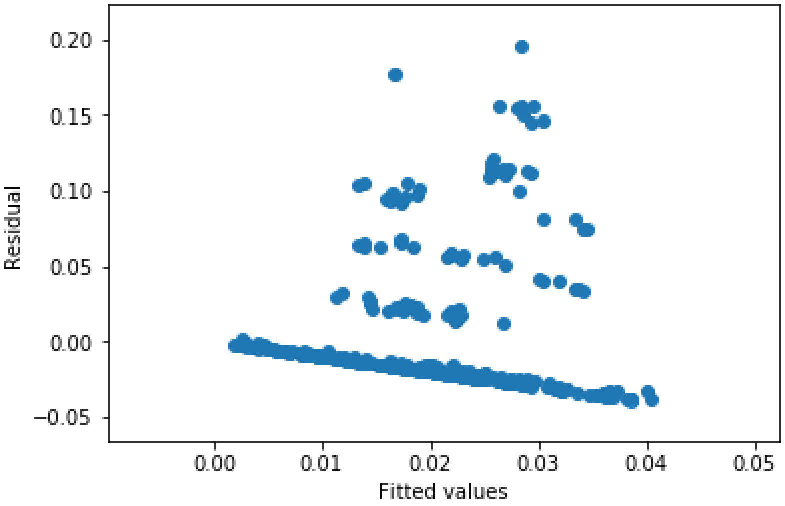


Train set average RMSE: 0.148766030563

Test set average RMSE: 0.148918602014

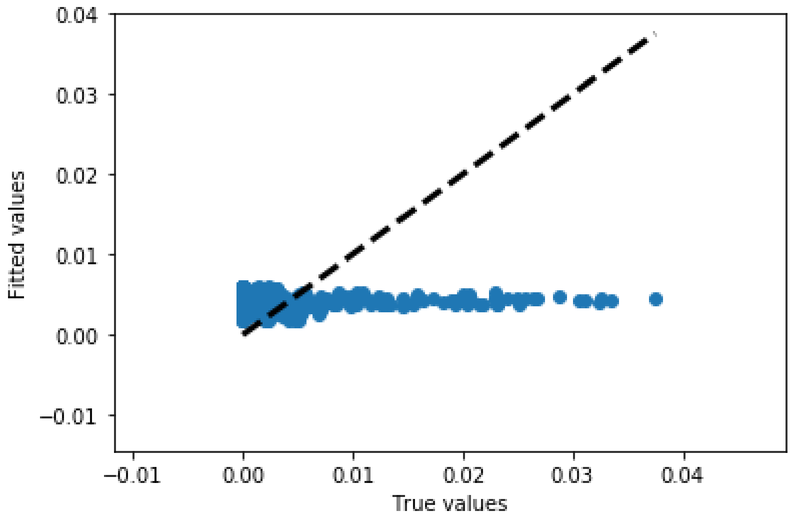
work_flow_2

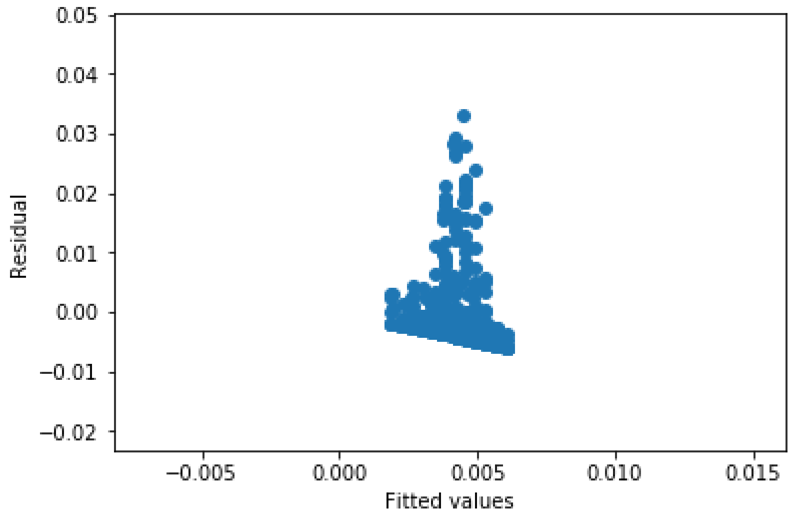




Train set average RMSE: 0.0429093206391
Test set average RMSE: 0.0430669058479

work_flow_3

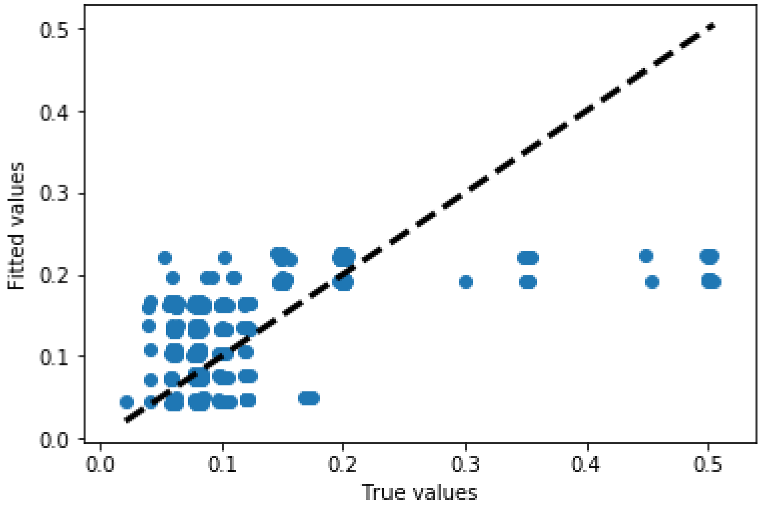


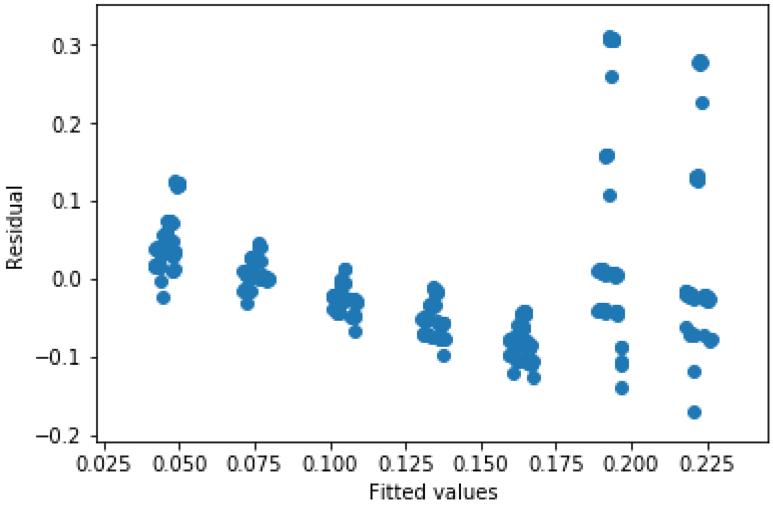


Train set average RMSE: 0.00724387887388

Test set average RMSE: 0.0072608942421

work_flow_4

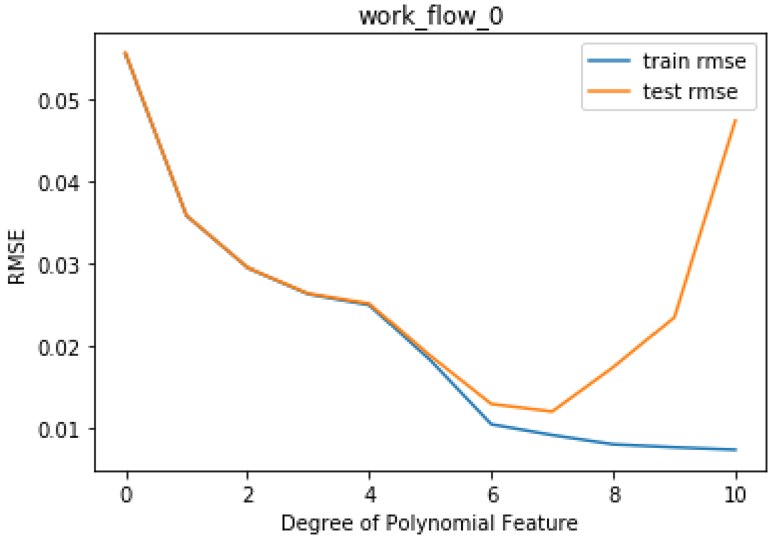


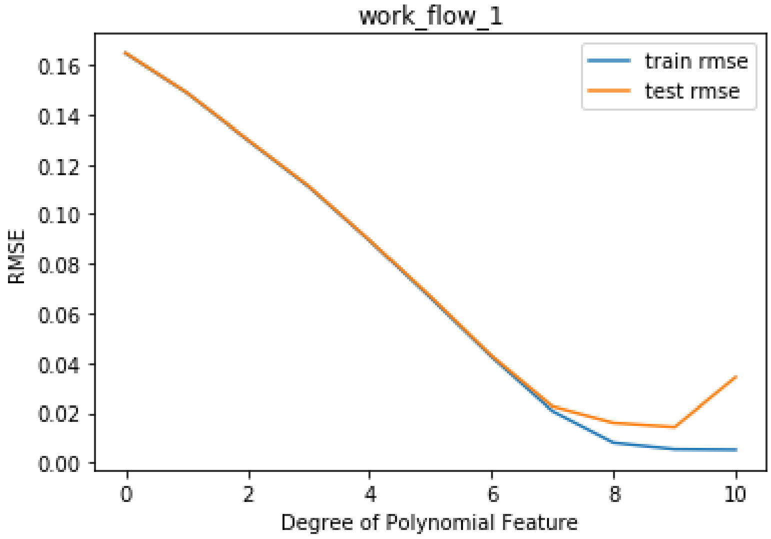


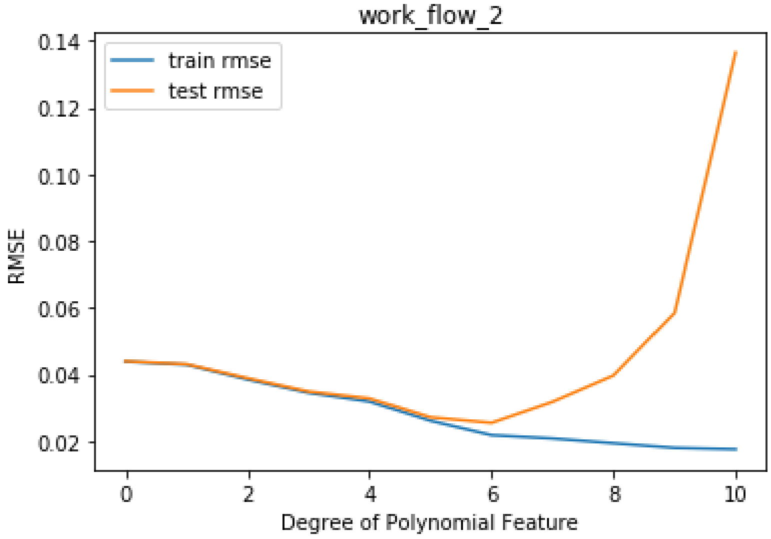
Train set average RMSE: 0.0859219367933
Test set average RMSE: 0.0859906141157

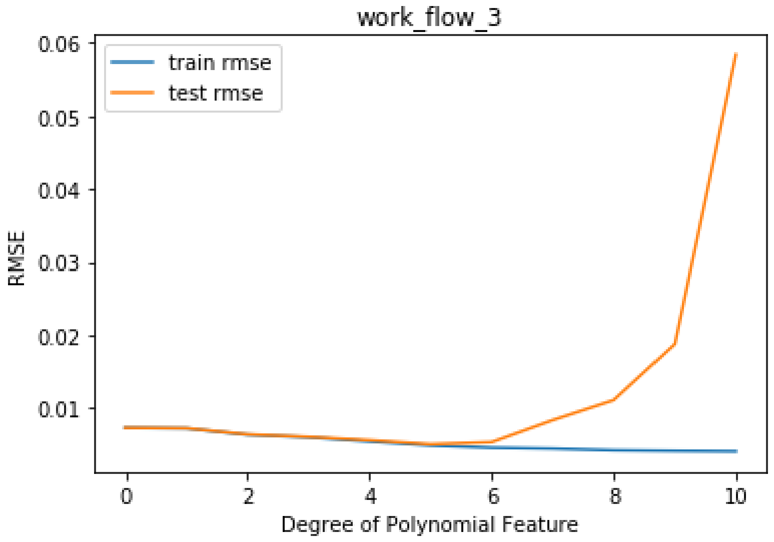
When predicting the backup size for each workflow separately, we found that the performance of linear regression is improved except for workflow 1.

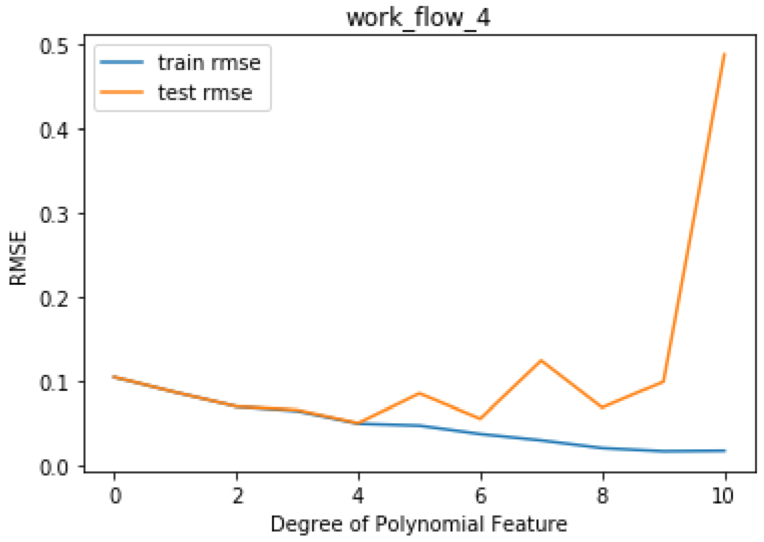
ii) Polynomial regression.


work_flow_0
train rmse
test rmse
RMSE
0.05
0.04
0.03
0.02
0.01
0
2
4
6
8
10
Degree of Polynomial Feature


work_flow_1
train rmse
test rmse
RMSE
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
2
4
6
8
10
Degree of Polynomial Feature


work_flow_2
train rmse
test rmse
RMSE
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0
2
4
6
8
10
Degree of Polynomial Feature

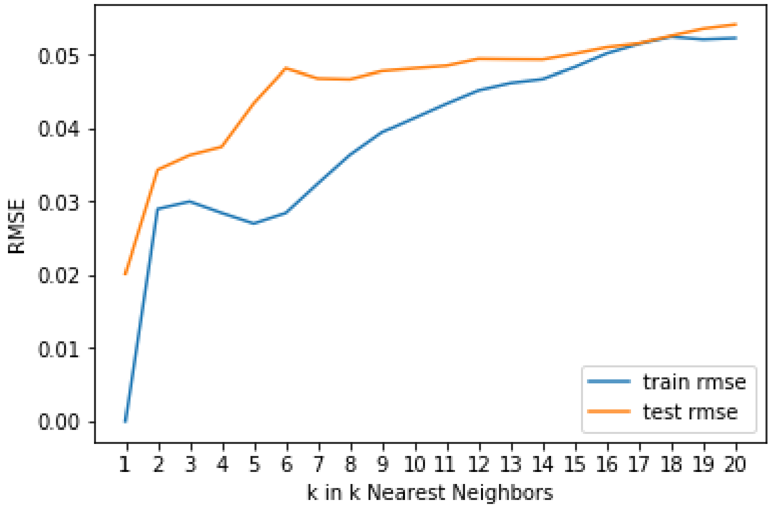




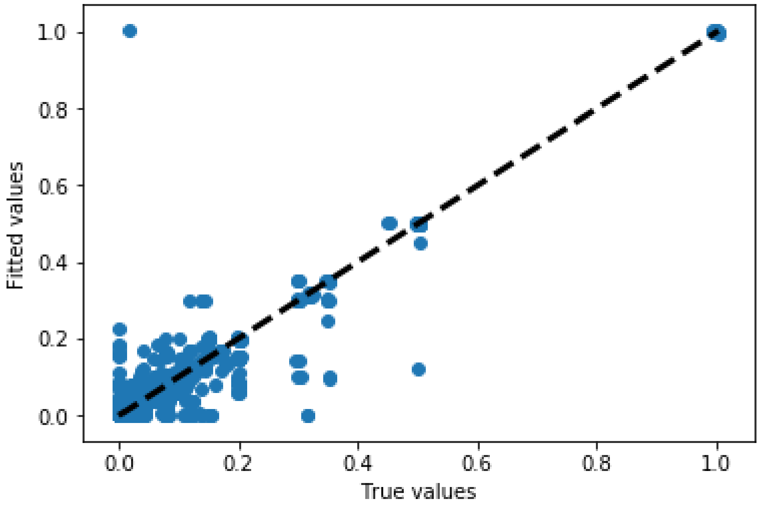
Beyond the degree 6, the generalization error of the model gets worse.

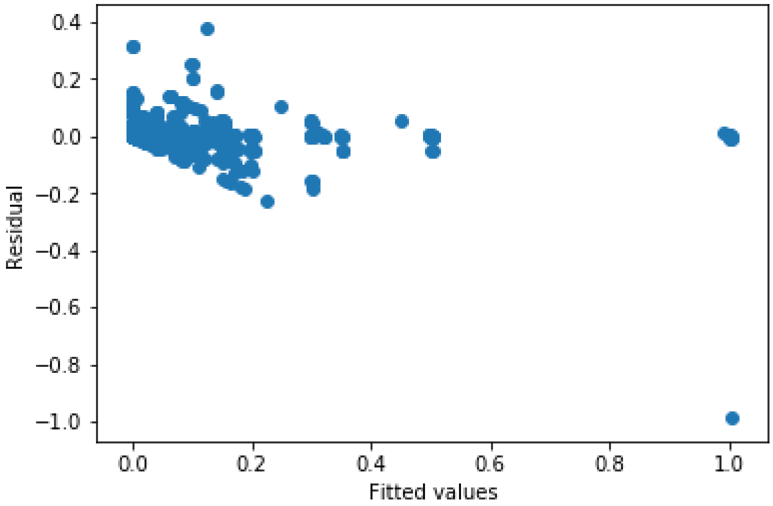
The threshold for workflow 0, 2 and 3 is 6. The threshold for workflow 1 is 8, and the threshold for workflow 4 is 7.
Cross validation increases the test error when the model is complex, which helps to identify overfitting problems and thus, improve the regression model.

e) Use kNN regression and find the best parameter



Best kNN result comes with k = 1





Train set average RMSE: 0.0
Test set average RMSE: 0.020165775873

Part 3) Compare the regression models

By comparing the RMSE for all the regression models that we used, random forest model (random tree 7, maximum depth 10, maximum feature 3) generate the best result.
Linear regression model does a bad job compared to other regression models. By One-Hot-Encoding, we can improve the performance of linear regression model, but still not good enough.
Neural Network regression takes good care of categorical features, while it is also good at handling sparse data when combined with 300 hidden layer units with activity function of relu.
Predicting the backup size for each workflow separately can generally improve the performance of most of the regression models.