q2

December 8, 2020

[50]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestRegressor
from sklearn import metrics
```

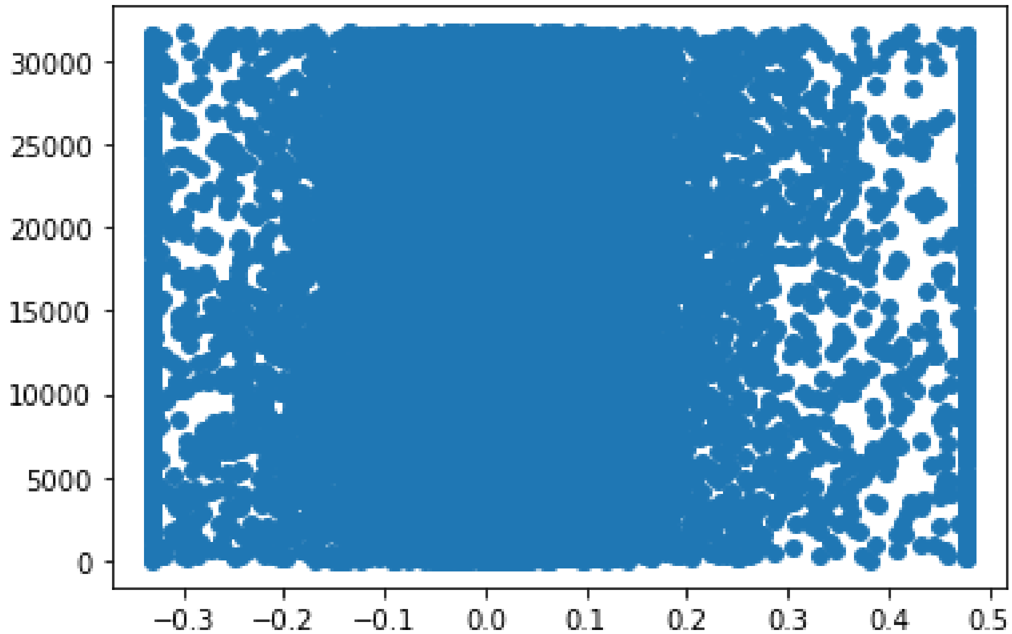
0.1 a) load/merge data and visualize logerror

[51]:
```
# load data into DataFrames
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

train_df = pd.read_csv("train.csv", keep_default_na=False, na_values=[""])
properties_df = pd.read_csv("properties.csv", keep_default_na=False,␣
↪na_values=[""])

df = pd.merge(train_df, properties_df, on='id')
# [31725 rows x 60 columns]
#X = dataset.iloc[:, [3, 4]].values
```

[52]:
```
# eliminate outliers
min_value, max_value = np.percentile(df['logerror'], [1, 99])

df['logerror'] = np.where(df['logerror']>max_value, max_value, df['logerror'])
df['logerror'] = np.where(df['logerror']<min_value, min_value, df['logerror'])
```

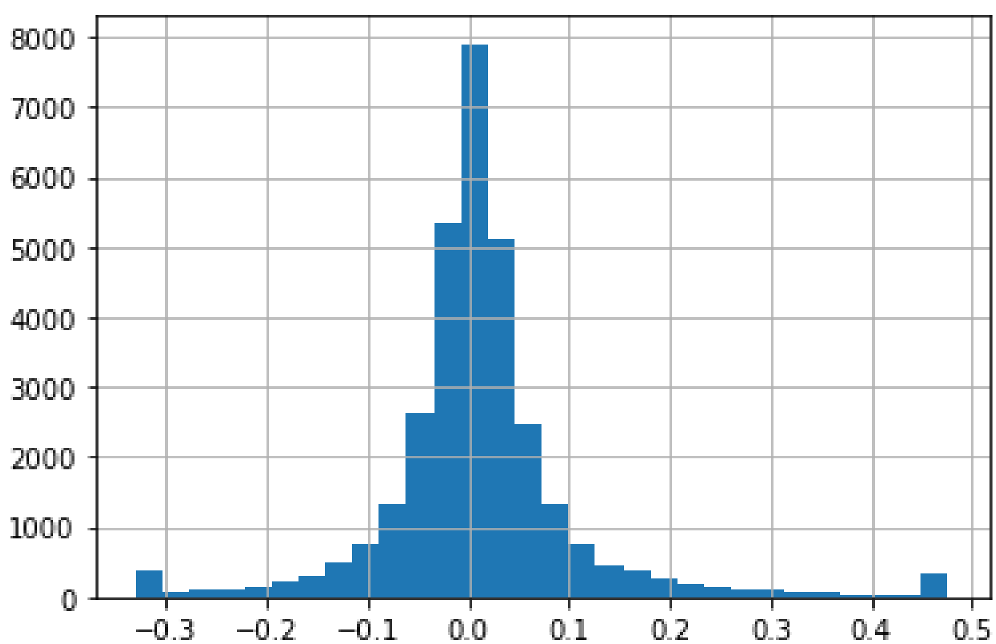
[53]:
```
# scatter of logerr
fig, ax = plt.subplots()
ax.scatter(df['logerror'], df['logerror'].index)
# plt.show()
```

[53]: <matplotlib.collections.PathCollection at 0x7fb40e1aadc0>



[54]:
```
# histogram of logerr
# # df.logerror.hist()
df.logerror.hist(bins=30)
```

[54]: <matplotlib.axes._subplots.AxesSubplot at 0x7fb40e0b20d0>



0.2 b) data cleaning

[55]:
```
# build new data frame
num = len(df)
data = pd.DataFrame({'column_name': list(df.columns), 'missing_count': list(df.
 ↪isnull().sum())})
data['missing_ratio'] = data.iloc[:,1].values/num
print(data)
```

```
                     column_name  missing_count  missing_ratio
0                             id              0       0.000000
1                       logerror              0       0.000000
2                transactiondate              0       0.000000
3          airconditioningtypeid          21563       0.679685
4       architecturalstyletypeid          31628       0.996942
5                   basementsqft          31711       0.999559
6                    bathroomcnt              0       0.000000
7                     bedroomcnt              0       0.000000
8            buildingclasstypeid          31717       0.999748
9          buildingqualitytypeid          11488       0.362112
10             calculatedbathnbr            414       0.013050
11                    decktypeid          31502       0.992971
12      finishedfloor1squarefeet          29381       0.926115
13  calculatedfinishedsquarefeet            228       0.007187
14          finishedsquarefeet12           1647       0.051915
15          finishedsquarefeet13          31711       0.999559
16          finishedsquarefeet15          30454       0.959937
17          finishedsquarefeet50          29381       0.926115
18           finishedsquarefeet6          31591       0.995776
19                          fips              0       0.000000
20                  fireplacecnt          28374       0.894374
21                  fullbathcnt            414       0.013050
22                  garagecarcnt          21280       0.670764
23               garagetotalsqft          21280       0.670764
24                hashottuborspa          30929       0.974909
25         heatingorsystemtypeid          11962       0.377053
26                      latitude              0       0.000000
27                     longitude              0       0.000000
28             lotsizesquarefeet           3522       0.111017
29                       poolcnt          25454       0.802333
30                   poolsizesum          31394       0.989567
31                  pooltypeid10          31337       0.987770
32                   pooltypeid2          31317       0.987139
33                   pooltypeid7          25862       0.815193
```

```
34     propertycountylandusecode           0    0.000000
35         propertylandusetypeid           0    0.000000
36            propertyzoningdesc       11135    0.350985
37        rawcensustractandblock           0    0.000000
38                  regionidcity         666    0.020993
39                regionidcounty           0    0.000000
40          regionidneighborhood       19082    0.601481
41                   regionidzip          12    0.000378
42                       roomcnt           0    0.000000
43                   storytypeid       31711    0.999559
44          threequarterbathnbr       27471    0.865910
45       typeconstructiontypeid       31613    0.996470
46                       unitcnt       11127    0.350733
47             yardbuildingsqft17      30814    0.971284
48             yardbuildingsqft26      31691    0.998928
49                     yearbuilt         260    0.008195
50               numberofstories       24526    0.773081
51                 fireplaceflag       31631    0.997037
52    structuretaxvaluedollarcnt         128    0.004035
53              taxvaluedollarcnt          1    0.000032
54                assessmentyear           0    0.000000
55         landtaxvaluedollarcnt           1    0.000032
56                     taxamount           1    0.000032
57             taxdelinquencyflag      31112    0.980678
58             taxdelinquencyyear      31112    0.980678
59            censustractandblock        208    0.006556
```

[56]:

```python
# fill missing data
df.update(df.fillna(df.mean(), inplace=True))
print(df)
```

```
             id  logerror transactiondate  airconditioningtypeid  \
0      14366692   -0.1684          1/1/16               1.809683
1      14739064   -0.0030          1/2/16               1.809683
2      10854446    0.3825          1/3/16               1.809683
3      11672170   -0.0161          1/3/16               1.000000
4      12524288   -0.0419          1/3/16               1.809683
…           …         …              …                     …
31720  12756771    0.0658        12/30/16               1.809683
31721  11295458   -0.0294        12/30/16               1.000000
31722  11308315    0.0070        12/30/16               1.000000
31723  11703478    0.0431        12/30/16               1.809683
31724  12566293    0.4207        12/30/16               1.809683

       architecturalstyletypeid  basementsqft  bathroomcnt  bedroomcnt  \
0                      7.453608    670.571429          3.5         4.0
1                      7.453608    670.571429          1.0         2.0
2                      7.453608    670.571429          2.0         2.0
```

```
3                      7.453608    670.571429          4.0          5.0
4                      7.453608    670.571429          1.0          1.0
…                           …            …          …          …
31720                  7.453608    670.571429          1.0          3.0
31721                  7.453608    670.571429          2.0          2.0
31722                  7.453608    670.571429          3.0          5.0
31723                  7.453608    670.571429          1.0          3.0
31724                  7.453608    670.571429          1.0          3.0

       buildingclasstypeid  buildingqualitytypeid  …  numberofstories  \
0                      4.0               5.570193  …         1.450479
1                      4.0               5.570193  …         1.450479
2                      4.0               7.000000  …         1.450479
3                      4.0               1.000000  …         1.450479
4                      4.0               7.000000  …         1.450479
…                      …                      …  …              …
31720                  4.0               7.000000  …         1.450479
31721                  4.0               7.000000  …         1.450479
31722                  4.0               4.000000  …         1.450479
31723                  4.0               7.000000  …         1.450479
31724                  4.0               7.000000  …         1.450479

       fireplaceflag  structuretaxvaluedollarcnt  taxvaluedollarcnt  \
0                  1                    346458.0           585529.0
1                  1                     66834.0           210064.0
2                  1                     55396.0           105954.0
3                  1                    559040.0          1090127.0
4                  1                     56233.0            70316.0
…                …                         …                 …
31720              1                     65728.0           307167.0
31721              1                     40163.0            50203.0
31722              1                    248378.0           331525.0
31723              1                     17520.0            39934.0
31724              1                     66258.0           163037.0

       assessmentyear  landtaxvaluedollarcnt  taxamount  taxdelinquencyflag  \
0                2015               239071.0   10153.02                 NaN
1                2015               143230.0    2172.88                 NaN
2                2015                50558.0    1443.69                 NaN
3                2015               531087.0   13428.94                 NaN
4                2015                14083.0     913.17                 NaN
…                 …                     …         …                 …
31720            2015               241439.0    4038.70                 NaN
31721            2015                10040.0    1263.39                   Y
31722            2015                83147.0    6461.79                 NaN
31723            2015                22414.0     627.91                 NaN
31724            2015                96779.0    2560.96                 NaN
```

```
       taxdelinquencyyear  censustractandblock
0               13.314845         6.048996e+13
1               13.314845         6.059040e+13
2               13.314845         6.037140e+13
3               13.314845         6.037260e+13
4               13.314845         6.037570e+13
…                     …                  …
31720           13.314845         6.037550e+13
31721           15.000000         6.037900e+13
31722           13.314845         6.037900e+13
31723           13.314845         6.037230e+13
31724           13.314845         6.037540e+13

[31725 rows x 60 columns]
```

0.3 c) univariate analysis

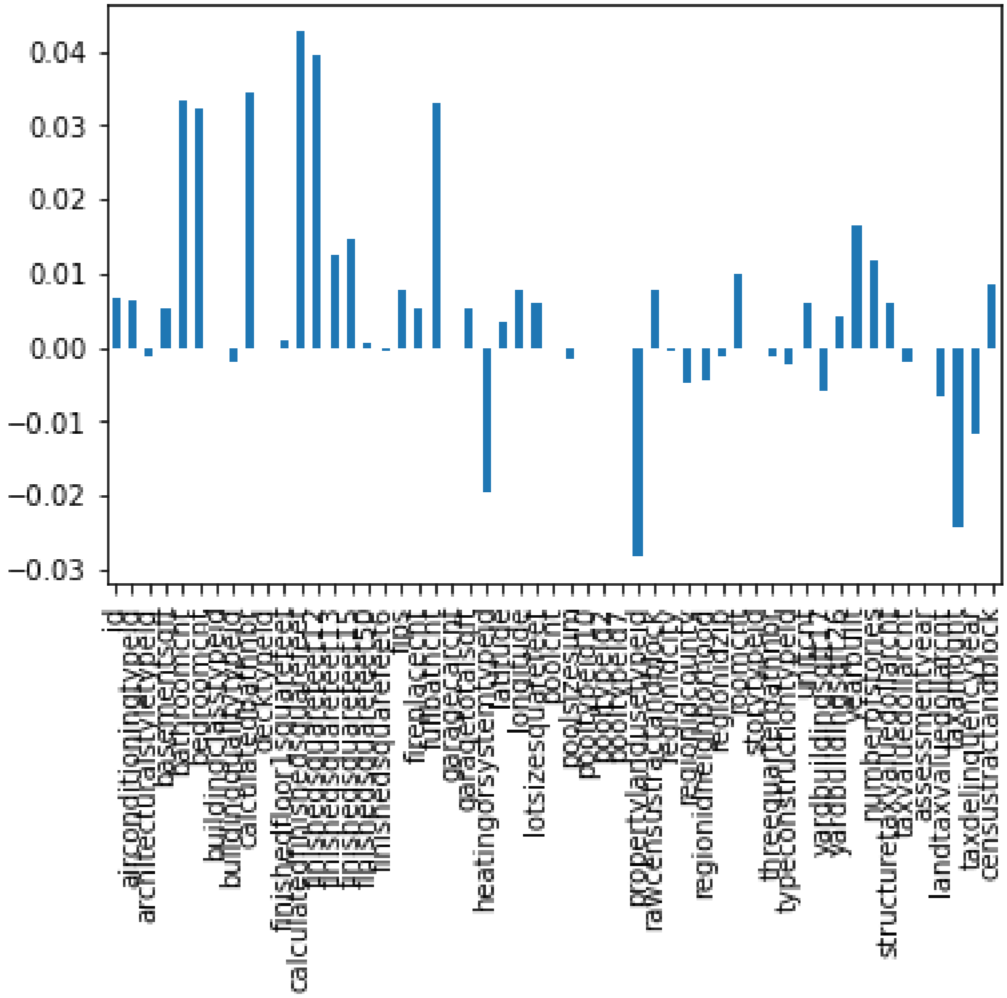
[57]:

```
# make bar chart
correlation = df.corr()['logerror']
neg = correlation.index.isin(['logerror'])
correlation = correlation[~neg]
print(correlation)
correlation.plot.bar()
```

```
id                              0.006562
airconditioningtypeid           0.006328
architecturalstyletypeid       -0.001234
basementsqft                    0.005239
bathroomcnt                     0.033445
bedroomcnt                      0.032168
buildingclasstypeid                  NaN
buildingqualitytypeid          -0.001840
calculatedbathnbr               0.034345
decktypeid                           NaN
finishedfloor1squarefeet        0.000807
calculatedfinishedsquarefeet    0.042841
finishedsquarefeet12            0.039504
finishedsquarefeet13            0.012608
finishedsquarefeet15            0.014687
finishedsquarefeet50            0.000621
finishedsquarefeet6            -0.000656
fips                            0.007863
fireplacecnt                    0.005099
fullbathcnt                     0.032986
garagecarcnt                   -0.000039
garagetotalsqft                 0.005227
heatingorsystemtypeid          -0.019511
```

```
latitude                        0.003277
longitude                       0.007782
lotsizesquarefeet               0.006093
poolcnt                              NaN
poolsizesum                    -0.001442
pooltypeid10                         NaN
pooltypeid2                          NaN
pooltypeid7                          NaN
propertylandusetypeid          -0.028459
rawcensustractandblock          0.007821
regionidcity                   -0.000542
regionidcounty                 -0.004874
regionidneighborhood           -0.004454
regionidzip                    -0.001253
roomcnt                         0.009762
storytypeid                          NaN
threequarterbathnbr            -0.001133
typeconstructiontypeid         -0.002361
unitcnt                         0.005964
yardbuildingsqft17             -0.006088
yardbuildingsqft26              0.004131
yearbuilt                       0.016442
numberofstories                 0.011565
structuretaxvaluedollarcnt      0.005950
taxvaluedollarcnt              -0.002060
assessmentyear                       NaN
landtaxvaluedollarcnt          -0.006758
taxamount                      -0.024282
taxdelinquencyyear             -0.011826
censustractandblock             0.008389
Name: logerror, dtype: float64
```

[57]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb410869280>
```



1 explain reason

This bar plot above indicates that there is generally either a positive correlation and negative correlation between the logerror and each variable. The few variables without correlation values indicate that the association between those variables and logerror is not obvious or is hardly observable. The variables with positive correlation values indicate that large values of logerror corresponds to large values of those variables, and vice versa. While the variables with negative correlation values indicate that large values of logerror corresponds to small values of those variables, and vice versa.

1.1 d) non-linear regression model

[58]:
```
# drop categorical features
# ("hashottuborspa", "propertycountylandusecode", "propertyzoningdesc",␣
↪"fireplaceflag", "taxdelinquencyflag")
# drop "id" and "transactiondate"
df.drop(['id','transactiondate'], axis=1, inplace=True)
df.fillna(df.mean(), inplace=True)
print(df)
```

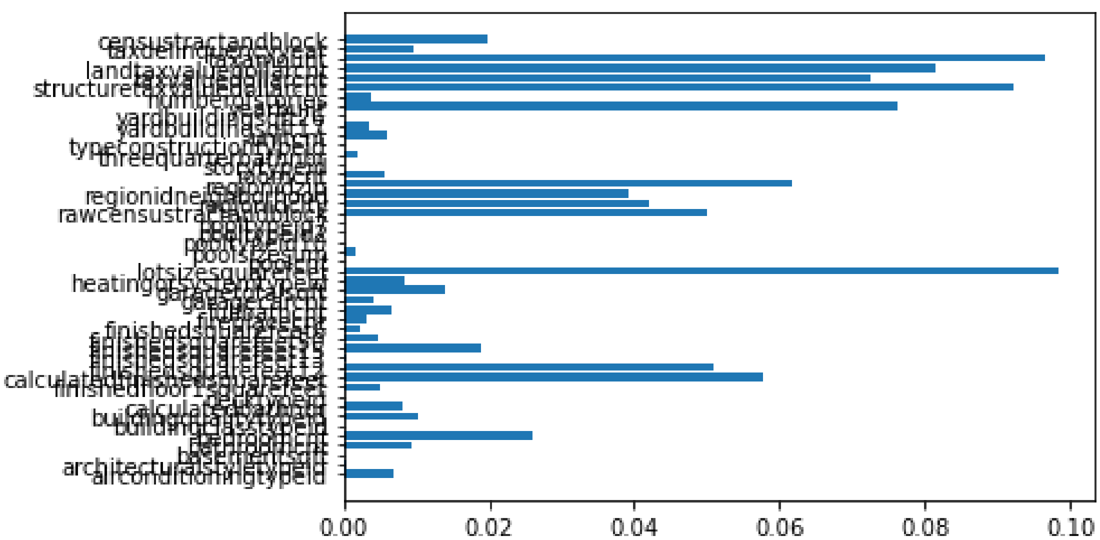
```
       logerror  airconditioningtypeid  architecturalstyletypeid  \
0       -0.1684               1.809683                  7.453608
1       -0.0030               1.809683                  7.453608
2        0.3825               1.809683                  7.453608
3       -0.0161               1.000000                  7.453608
4       -0.0419               1.809683                  7.453608
…           …                    …                       …
31720    0.0658               1.809683                  7.453608
31721   -0.0294               1.000000                  7.453608
31722    0.0070               1.000000                  7.453608
31723    0.0431               1.809683                  7.453608
31724    0.4207               1.809683                  7.453608

       basementsqft  bathroomcnt  bedroomcnt  buildingclasstypeid  \
0        670.571429          3.5         4.0                  4.0
1        670.571429          1.0         2.0                  4.0
2        670.571429          2.0         2.0                  4.0
3        670.571429          4.0         5.0                  4.0
4        670.571429          1.0         1.0                  4.0
…             …            …           …                    …
31720    670.571429          1.0         3.0                  4.0
31721    670.571429          2.0         2.0                  4.0
31722    670.571429          3.0         5.0                  4.0
31723    670.571429          1.0         3.0                  4.0
31724    670.571429          1.0         3.0                  4.0

       buildingqualitytypeid  calculatedbathnbr  decktypeid  …  \
0                   5.570193                3.5        66.0  …
1                   5.570193                1.0        66.0  …
2                   7.000000                2.0        66.0  …
3                   1.000000                4.0        66.0  …
4                   7.000000                1.0        66.0  …
…                       …                 …           …  …
31720               7.000000                1.0        66.0  …
31721               7.000000                2.0        66.0  …
31722               4.000000                3.0        66.0  …
31723               7.000000                1.0        66.0  …
```

```
31724               7.000000                1.0        66.0  …

       numberofstories  fireplaceflag  structuretaxvaluedollarcnt  \
0             1.450479              1                    346458.0
1             1.450479              1                     66834.0
2             1.450479              1                     55396.0
3             1.450479              1                    559040.0
4             1.450479              1                     56233.0
…                  …             …                         …
31720         1.450479              1                     65728.0
31721         1.450479              1                     40163.0
31722         1.450479              1                    248378.0
31723         1.450479              1                     17520.0
31724         1.450479              1                     66258.0

       taxvaluedollarcnt  assessmentyear  landtaxvaluedollarcnt  taxamount  \
0               585529.0            2015               239071.0   10153.02
1               210064.0            2015               143230.0    2172.88
2               105954.0            2015                50558.0    1443.69
3              1090127.0            2015               531087.0   13428.94
4                70316.0            2015                14083.0     913.17
…                    …             …                     …        …
31720           307167.0            2015               241439.0    4038.70
31721            50203.0            2015                10040.0    1263.39
31722           331525.0            2015                83147.0    6461.79
31723            39934.0            2015                22414.0     627.91
31724           163037.0            2015                96779.0    2560.96

       taxdelinquencyflag  taxdelinquencyyear  censustractandblock
0                     NaN           13.314845         6.048996e+13
1                     NaN           13.314845         6.059040e+13
2                     NaN           13.314845         6.037140e+13
3                     NaN           13.314845         6.037260e+13
4                     NaN           13.314845         6.037570e+13
…                   …                  …                   …
31720                 NaN           13.314845         6.037550e+13
31721                   Y           15.000000         6.037900e+13
31722                 NaN           13.314845         6.037900e+13
31723                 NaN           13.314845         6.037230e+13
31724                 NaN           13.314845         6.037540e+13

[31725 rows x 58 columns]
```

[68]:
```
# split and train
X = df.select_dtypes('float').values[:,1:]
y = df.select_dtypes('float').values[:,0]
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,␣
 ↪random_state=0)
# Already normalized data**
# sc = StandardScaler()
# X_train = sc.fit_transform(X_train)
# X_test = sc.transform(X_test)

regressor = RandomForestRegressor(n_estimators=20, random_state=0)
regressor.fit(X_train, y_train)
y_pred = regressor.predict(X_test)
```

[69]:
```
# report importances and mse
plt.barh(df.select_dtypes('float').columns[1:], regressor.feature_importances_)
plt.show()

print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
```




```
Mean Squared Error: 0.010773740173083314
```

1.2 e) KFold

[70]:
```
# KFold, k = 5
X_cv = df.select_dtypes('float').values[:500,1:]
y_cv = df.select_dtypes('float').values[:500,0]
model = RandomForestRegressor(n_estimators=20, random_state=1)
```

```
cv = KFold(n_splits=5, random_state=1, shuffle=True) #default is 5 (redundant)
scores = cross_val_score(model, X_cv, y_cv, scoring='neg_mean_squared_error',␣
↪cv=cv, n_jobs=-1)
print('Mean Squared Error_CV:', np.mean(np.absolute(scores)))
```

```
Mean Squared Error_CV: 0.014292075903614952
```

[71]:

```
# Run d2 for 100 times
import random

for i in range(100):
    seed = random.randint(0,99)
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,␣
↪random_state=seed)

#     sc = StandardScaler()
#     X_train = sc.fit_transform(X_train)
#     X_test = sc.transform(X_test)

    regressor = RandomForestRegressor(n_estimators=20, random_state=seed)
    regressor.fit(X_train, y_train)
    y_prd = regressor.predict(X_test)
    print('random_seed: ',seed,'MSE:', metrics.mean_squared_error(y_test,␣
↪y_prd))
```

```
random_seed:  49 MSE: 0.010374547672181215
random_seed:  64 MSE: 0.010468170321508635
random_seed:  27 MSE: 0.010738558893648883
random_seed:  49 MSE: 0.010374547672181215
random_seed:  40 MSE: 0.010652433594255526
random_seed:  29 MSE: 0.010615661260027666
random_seed:  49 MSE: 0.010374547672181215
random_seed:  27 MSE: 0.010738558893648883
random_seed:  18 MSE: 0.010661646303790831
random_seed:  75 MSE: 0.010056237810833929
random_seed:  72 MSE: 0.010346488533182234
random_seed:  7 MSE: 0.010596493018438952
random_seed:  91 MSE: 0.010704934797804322
random_seed:  43 MSE: 0.010759998565866298
random_seed:  38 MSE: 0.010741879316798836
random_seed:  77 MSE: 0.010849221227234375
random_seed:  84 MSE: 0.010447766787825772
random_seed:  69 MSE: 0.010847690855126105
random_seed:  49 MSE: 0.010374547672181215
random_seed:  9 MSE: 0.010424234396700042
random_seed:  61 MSE: 0.010612658205591279
random_seed:  23 MSE: 0.010710105239384803
```

```
random_seed:  93 MSE: 0.010859594909032997
random_seed:  98 MSE: 0.010582599725900184
random_seed:  31 MSE: 0.010980465063084267
random_seed:  35 MSE: 0.010644306478161933
random_seed:  83 MSE: 0.010759286545648037
random_seed:  48 MSE: 0.010803725674328008
random_seed:  30 MSE: 0.01087953889017812
random_seed:  29 MSE: 0.010615661260027666
random_seed:  66 MSE: 0.010624729789045708
random_seed:  13 MSE: 0.010856756087537044
random_seed:  66 MSE: 0.010624729789045708
random_seed:  32 MSE: 0.01085643501810313
random_seed:  53 MSE: 0.010413282115624773
random_seed:  29 MSE: 0.010615661260027666
random_seed:  35 MSE: 0.010644306478161933
random_seed:  36 MSE: 0.011157417113462553
random_seed:  6 MSE: 0.010933839022101186
random_seed:  88 MSE: 0.010751002850993351
random_seed:  99 MSE: 0.010971654293104337
random_seed:  25 MSE: 0.010753152963904783
random_seed:  7 MSE: 0.010596493018438952
random_seed:  35 MSE: 0.010644306478161933
random_seed:  40 MSE: 0.010652433594255526
random_seed:  9 MSE: 0.010424234396700042
random_seed:  46 MSE: 0.010307968058643967
random_seed:  38 MSE: 0.010741879316798836
random_seed:  19 MSE: 0.010716984213912935
random_seed:  57 MSE: 0.010768774084516734
random_seed:  90 MSE: 0.010381302678966874
random_seed:  34 MSE: 0.011247187826096326
random_seed:  91 MSE: 0.010704934797804322
random_seed:  92 MSE: 0.010708301653129328
random_seed:  26 MSE: 0.01083831778438862
random_seed:  84 MSE: 0.010447766787825772
random_seed:  19 MSE: 0.010716984213912935
random_seed:  60 MSE: 0.010591942493656225
random_seed:  15 MSE: 0.010554539942102875
random_seed:  83 MSE: 0.010759286545648037
random_seed:  22 MSE: 0.010287365101144821
random_seed:  90 MSE: 0.010381302678966874
random_seed:  89 MSE: 0.010908196113060164
random_seed:  72 MSE: 0.010346488533182234
random_seed:  91 MSE: 0.010704934797804322
random_seed:  78 MSE: 0.011250073584658234
random_seed:  77 MSE: 0.010849221227234375
random_seed:  23 MSE: 0.010710105239384803
random_seed:  69 MSE: 0.010847690855126105
random_seed:  19 MSE: 0.010716984213912935
```

```
random_seed:  73 MSE: 0.010286551802982683
random_seed:  53 MSE: 0.010413282115624773
random_seed:  51 MSE: 0.010775947624611643
random_seed:  76 MSE: 0.01061688470814898
random_seed:  64 MSE: 0.010468170321508635
random_seed:  86 MSE: 0.011083256725539977
random_seed:  84 MSE: 0.010447766787825772
random_seed:  25 MSE: 0.010753152963904783
random_seed:  99 MSE: 0.010971654293104337
random_seed:  41 MSE: 0.010952345344469934
random_seed:  93 MSE: 0.010859594909032997
random_seed:  29 MSE: 0.010615661260027666
random_seed:  16 MSE: 0.010477144279381526
random_seed:  60 MSE: 0.010591942493656225
random_seed:  96 MSE: 0.010924333303800698
random_seed:  87 MSE: 0.010554557590616645
random_seed:  58 MSE: 0.010330998457447272
random_seed:  60 MSE: 0.010591942493656225
random_seed:  29 MSE: 0.010615661260027666
random_seed:  70 MSE: 0.010577889454047267
random_seed:  61 MSE: 0.010612658205591279
random_seed:  6 MSE: 0.010933839022101186
random_seed:  56 MSE: 0.010982175440280791
random_seed:  63 MSE: 0.010534524775824654
random_seed:  7 MSE: 0.010596493018438952
random_seed:  62 MSE: 0.010164430815652757
random_seed:  49 MSE: 0.010374547672181215
random_seed:  70 MSE: 0.010577889454047267
random_seed:  16 MSE: 0.010477144279381526
random_seed:  9 MSE: 0.010424234396700042
```

There is not a huge different between the MSE with different random seeds. Cross_validation would be good here because it would build K different models so we are able to make prediction on all the data. Confidence in the alogrithm would also be built with cross-validation as confidence in the algorithm is currently weak.

[ ]:

```

```

[ ]:

```

```