Electric Vehicle Presence Discovery

Krystin Sinclair

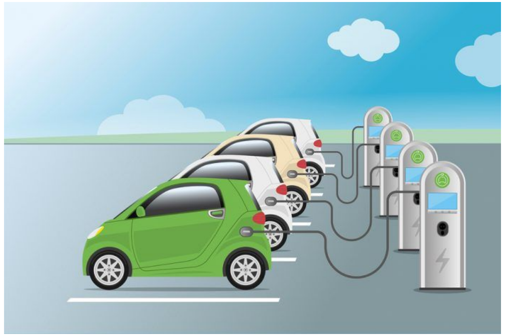
Which homes have Electric Vehicles?

US has 1 million Electric Vehicles in 2018 and there is an expansion in electric vehicle production

Electric Vehicles require energy to charge

Utilities companies need to know how much energy homes will be consuming

Knowing who has this type of car can inform utilities



"Power Supply for Electric Car Charging. Electric Car Charging St" Frontera, 09/26/2017, https://frontera.net/news/global-macro/the-5-biggest-electric-vehicle-manufacturers-in-brics-nations/

*Joselow, Maxine. "The U.S. Has 1 Million Electric Vehicles, but Does It Matter?" *Scientific American* 12.10.2018

*Colias, Mike. "Ford to Expand Electric-Vehicle Production at Michigan Plant" *Wall Street Journal* 20.3.2019

"

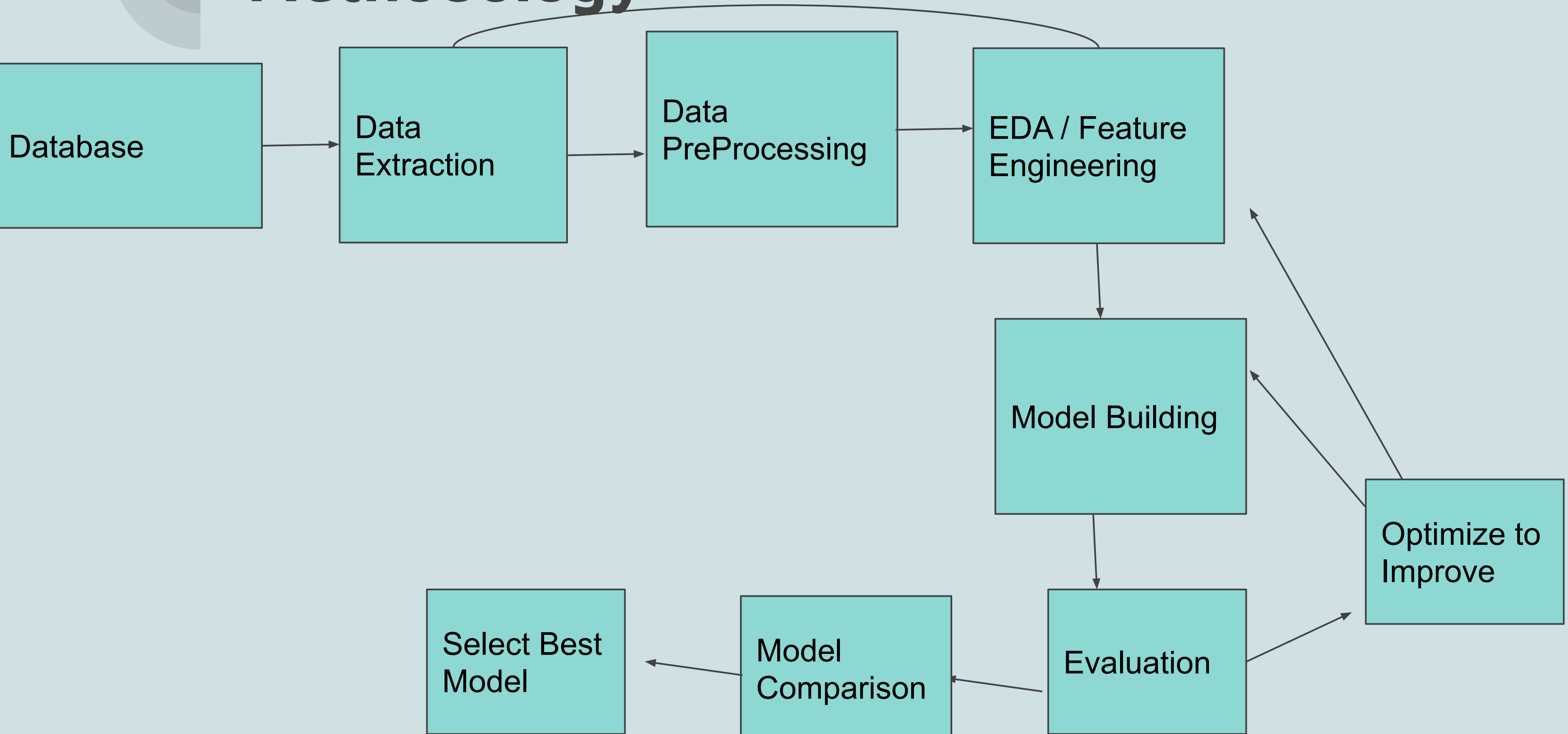
Data

Data Port’s Pecan Street

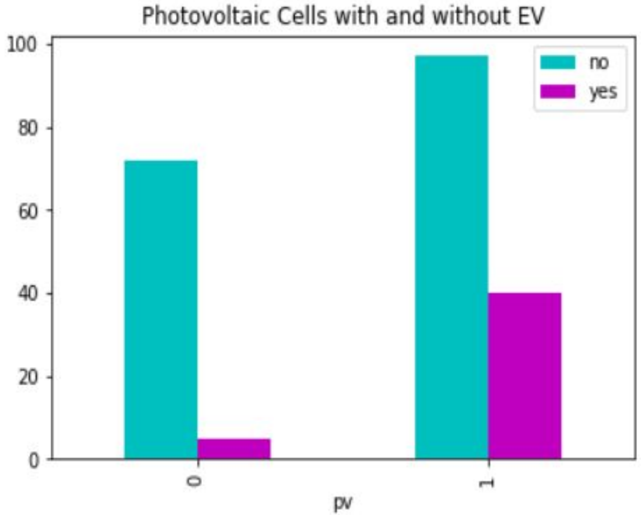
- Target: EV
- Electricity
  - All Dataid that joined program prior to 1/1/2016 and stayed through 12/31/2018
  - Grouped electricity egauge by Dataid
- Dataid information
  - House construction year
  - PV
  - Square Footage
  - Building Type
  - Location

Source: Pecan Street Inc

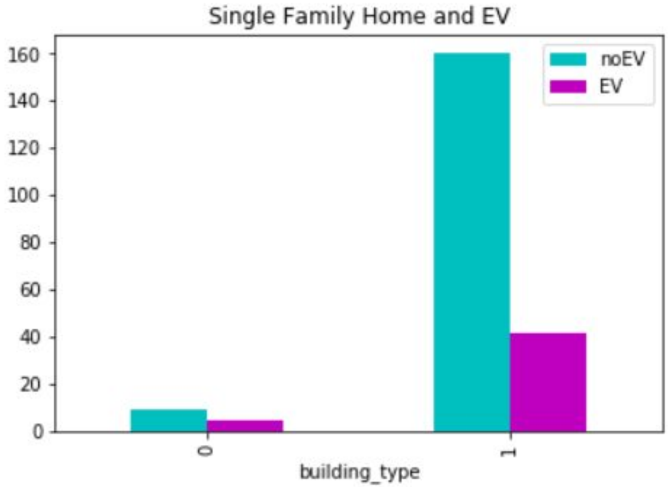
Methodology



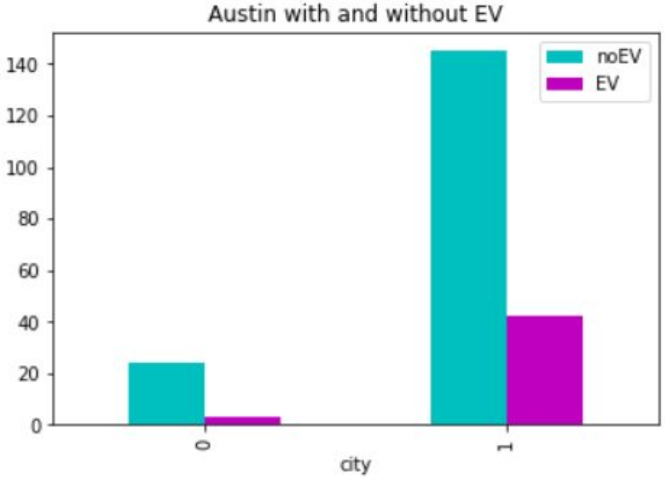
Exploratory Data Analysis: Categorical



Photovoltaic Cells with and without EV

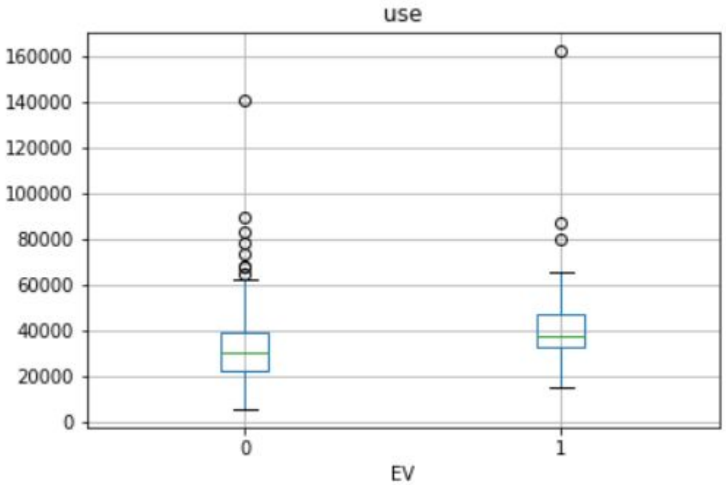


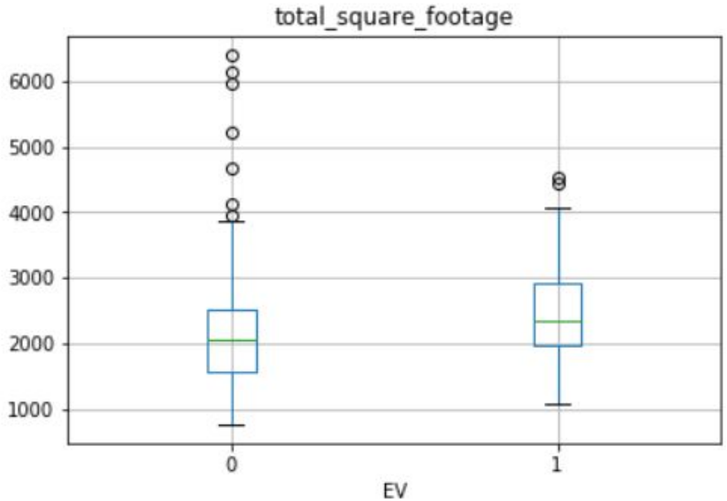
Single Family Home and EV

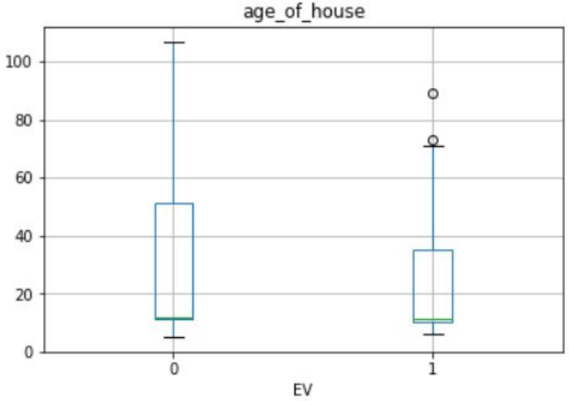


Austin with and without EV

Exploratory Data Analysis: Continuous







What makes EV homes different?

	EV	nonEV
Mean average annual energy use	14,454	11,037
Total Square Footage	2475	2191
Mean House Age	25(1994)	31(1989)

Odds Ratios

The odds of having an electric vehicle among those with single family home are .57 times the odds of having an electric vehicle among those with our housing types.

The odds of having an electric vehicle among those with PV are 5.93 times the odds of having an electric vehicle among those without PV.

The odds of having an electric vehicle among those that live in Austin are 2.3 times the odds of having an electric vehicle among those that live elsewhere.

Feature Engineering

DF1 = Original

*DF2 = removed outliers, converted construction year to age of home, total energy use for 3 years

DF3 = dropped use and sqft and created use/sqft

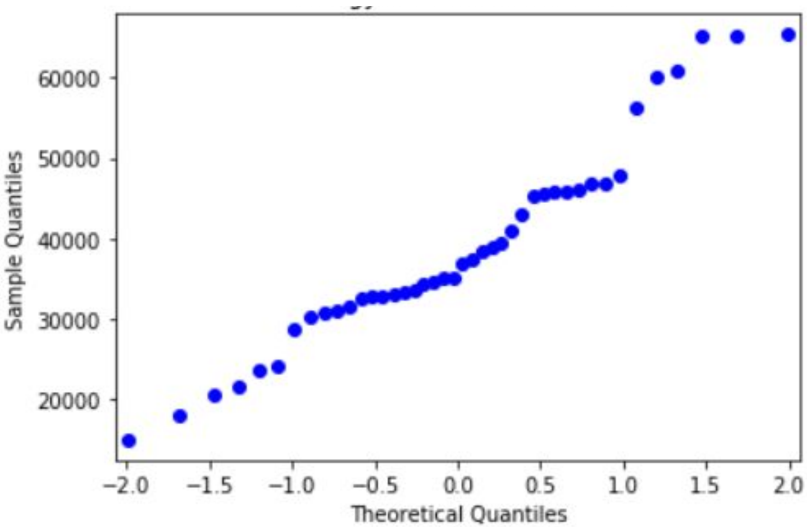
T-test on mean of EV and nonEV house energy use

Random Forest to see feature importance

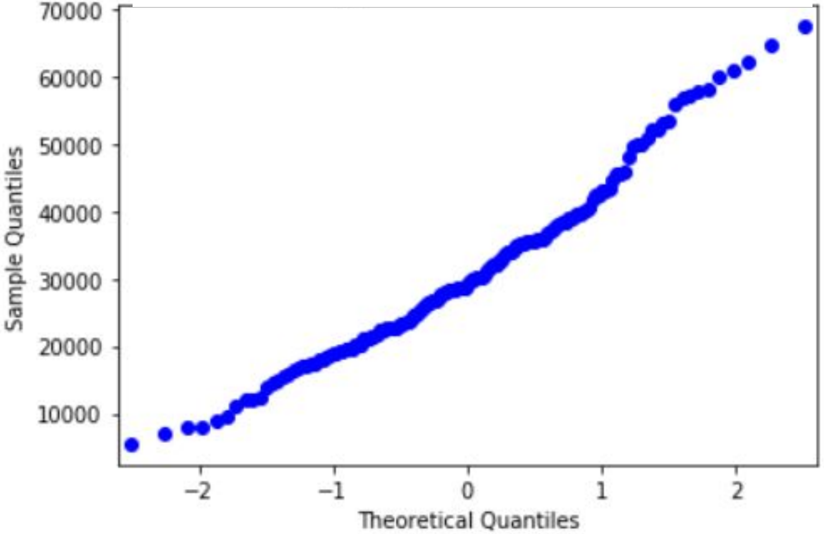
*DF2 Best Features - tried different ratios from oversampling using random forest accuracy to identify best

T- test on total energy use

Energy Use Homes with EV

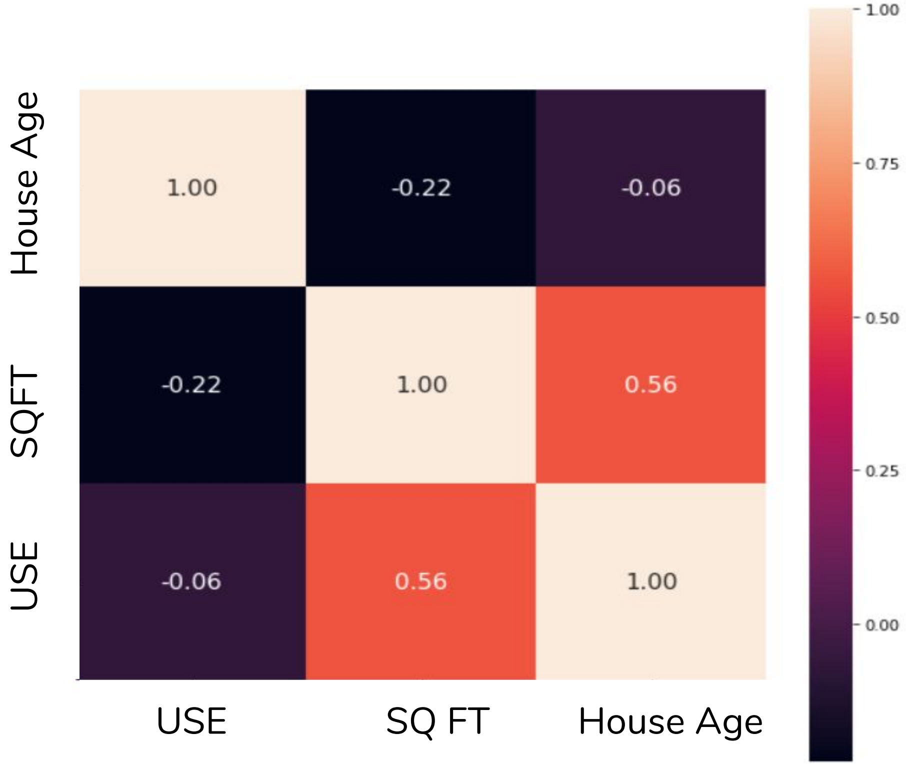


Energy Use Homes without EV

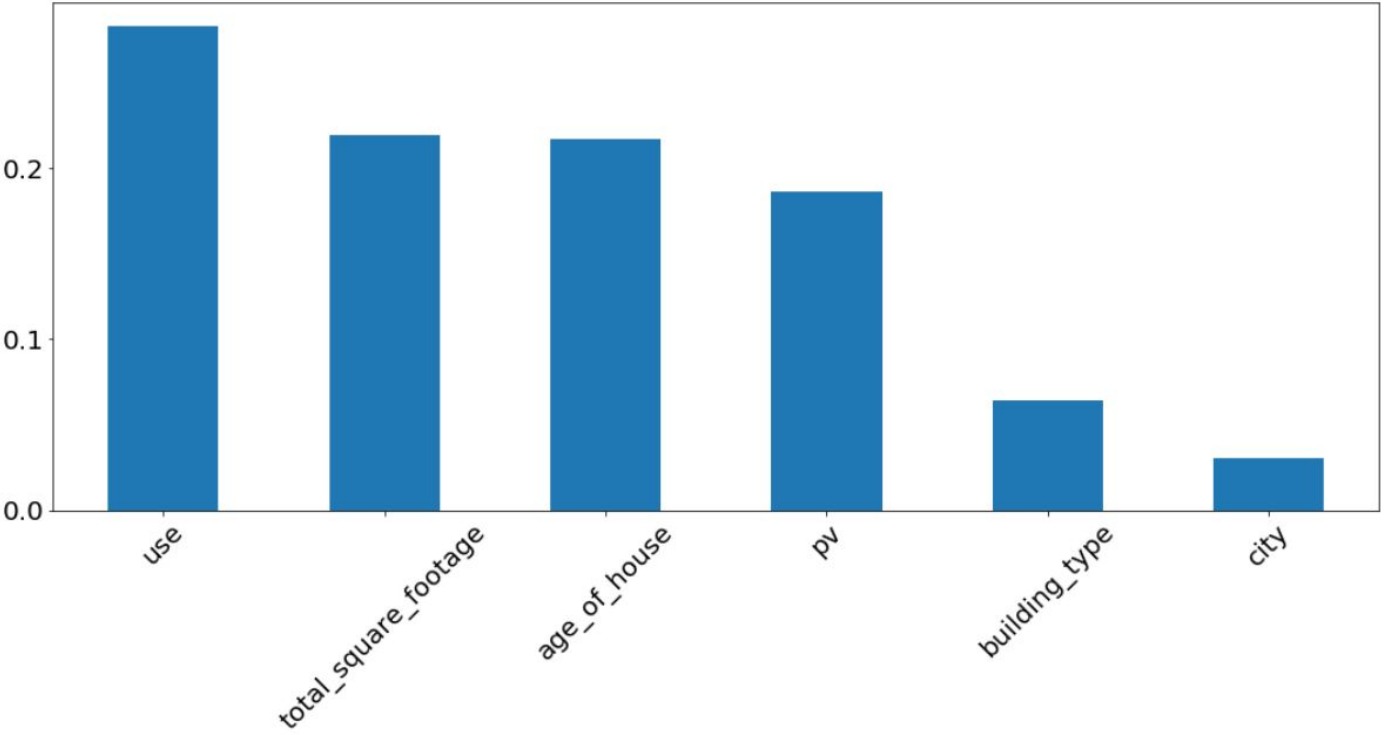


```
T value = 3.2134282629049014, p Value =  0.0020466070952163836
```

Correlation



Feature Importance



Models

Pipeline

- Standardize
- feature selection using feature importance from random forest threshold,
- Classifiers
  - Random Forest, Logistic Regression, SVM, Neural Network, KNN
- 5-fold cross validation
- hyperparameter tuning
  - Scored
    - Accuracy
    - Recall
    - Precision
    - Specificity
    - f1

Models - Oversampling in preprocessing

Classifier	Accuracy	Specificity	Sensitivity	Recall	precision
Random Forest	81%, 97%	96%, 100%	94%, 96%	94%, 96%	96%, 100%
Logistic Regression	71%, 76%	68%, 71%	74%, 81%	74%, 81%	70%, 74%
KNN	79%, 95%	100%, 100%	89%, 92%	89%, 91%	100%, 100%
SVM	83%, 96%	95%, 93%	99%, 100%	99%, 100%	95%, 94%
Neural Network	75%, 81%	69%, 73%	85%,88%	85%, 88%	73%77%

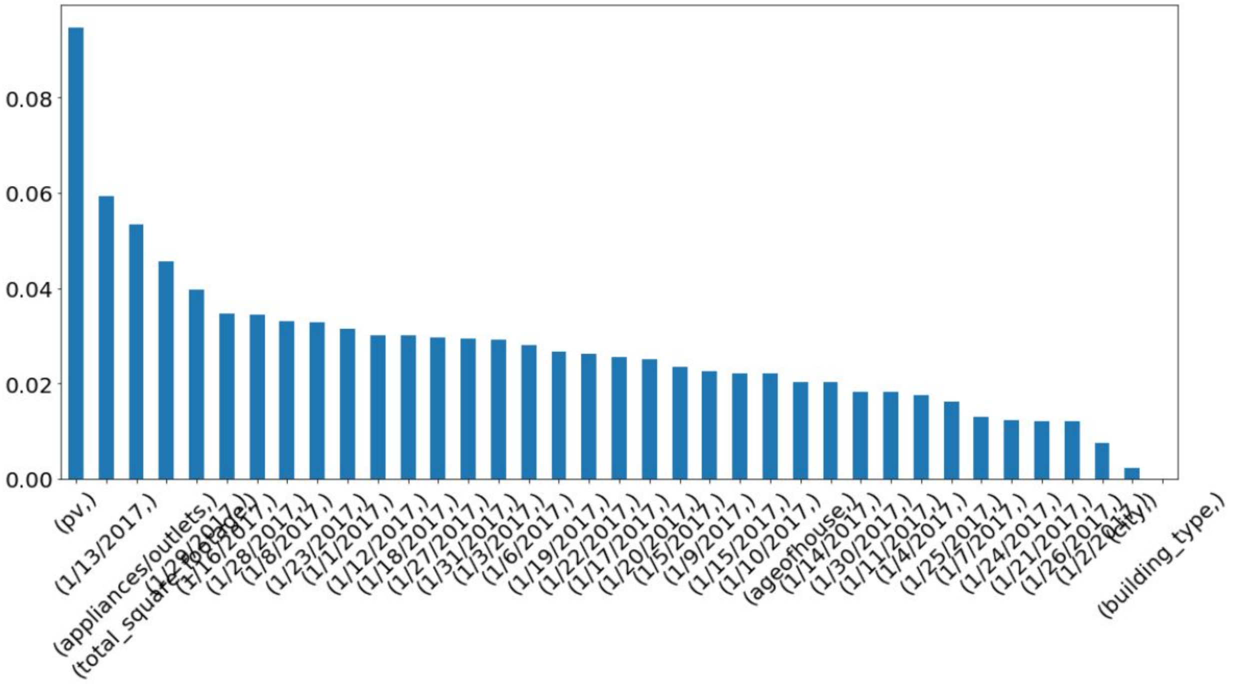
Oversampling only on train dataset

Classifier	Accuracy	Specificity	f1	Recall	precision
Random Forest	81%, 74%	95%, 85%	93%, 33%	92%, 30%	95%, 36%
Logistic Regression	72%, 67%	62%, 67%	77%, 47%	86%, 69%	69%, 36%
KNN	75%, 70%	100%, 84%	92%, 25%	85%, 23%	100%, 27%
SVM	80%, 70%	95%, 83%	97%, 17%	100%,15%	95%, 20%
Neural Network	78%, 69%	70%, 71%	83%, 45%	92%, 61%	75%, 36%

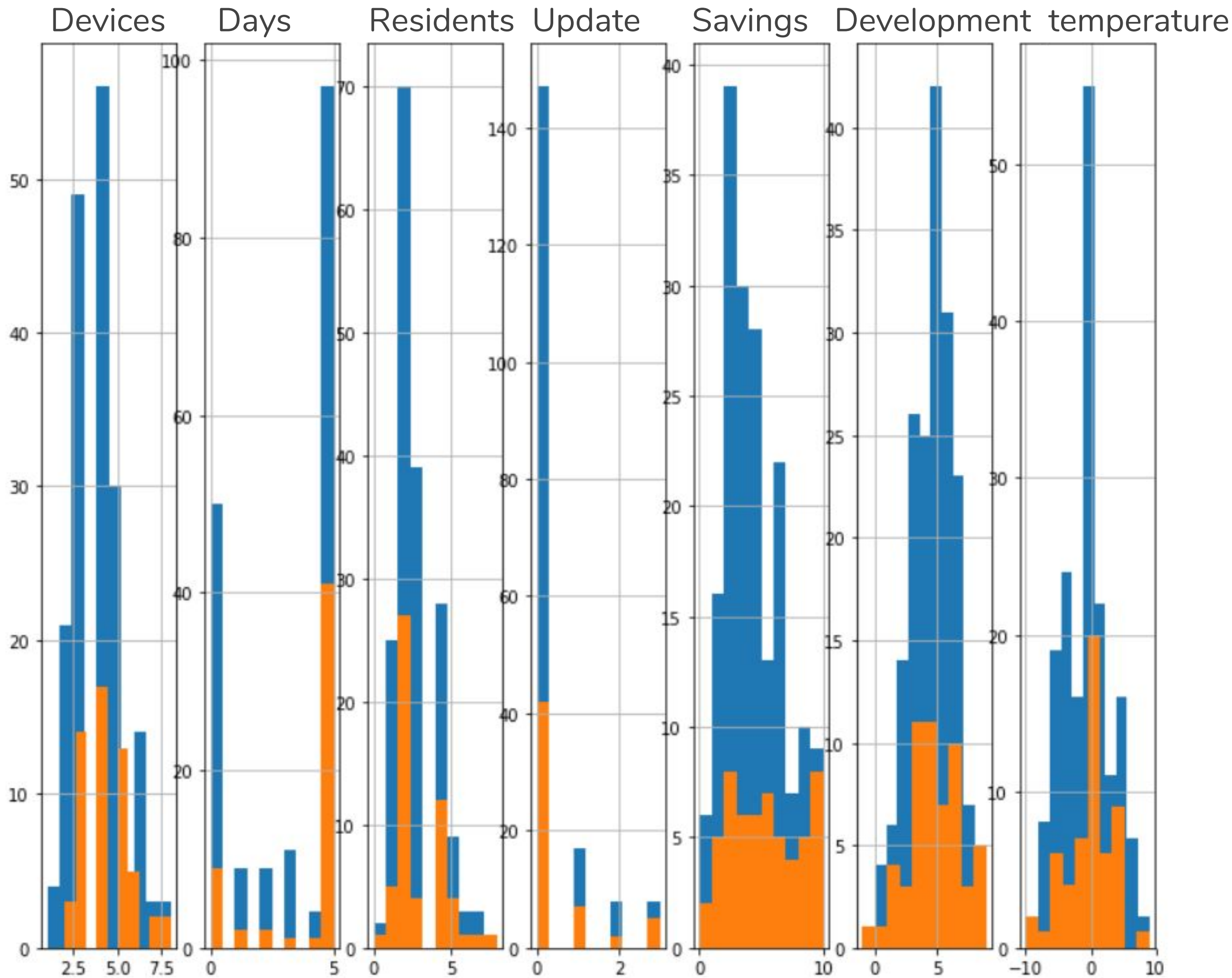
Undersampling only on train dataset

Classifier	Accuracy	Specificity	f1	Recall	precision
Random Forest	48%, 52%	48%, 47%	76%, 76%	93%, 69%	64%, 26%
Logistic Regression	52%, 56%	52%, 53%	68%, 68%	76%, 69%	61%, 28%
KNN	62%, 52%	82%, 51%	75%, 75%	83%, 54%	69%, 23%
SVM	48%, 53%	48%, 47%	63%, 63%	69%, 77%	57%, 28%
Neural Network	62%, 65%	62%, 63%	75%, 75%	83%, 69%	69%, 33%

DF4 - daily energy use (Jan 2017) number of appliances and outlets undersampling

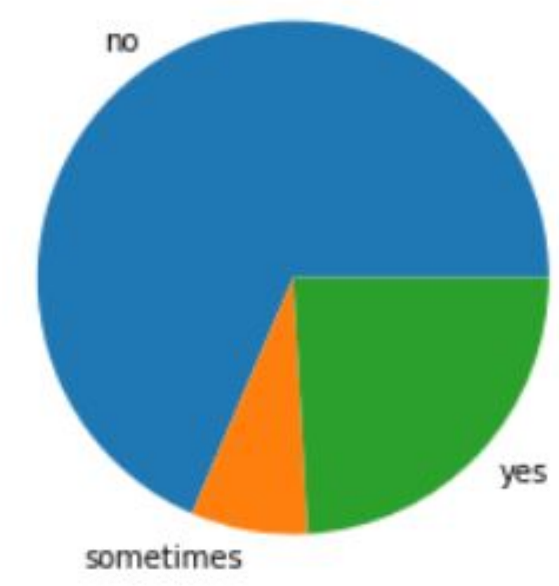


Train overfits with 100% accuracy and the test does not predict that anyone has EV



Devices Days Residents Update Savings Development temperature

Do you charge your electric vehicle at home?



Survey Results

- Most EV owners do not charge their electric vehicles at home
- Only 13 program participants said that they always charge their EV at home

Future Works

Demographics

- New features to understand EV homes better

Compare homes before and after EV purchase

Business Purpose

Introduce to utilities companies

- Provide insight into consumers
- Incorporate into method of predicting overall energy usage

Data Limitations & Lessons Learned

Class Imbalance

- Oversampling Techniques - used SMOTE package - causes bias

Time Component

- Converted dataframe from messy pecan street to typical classification problem

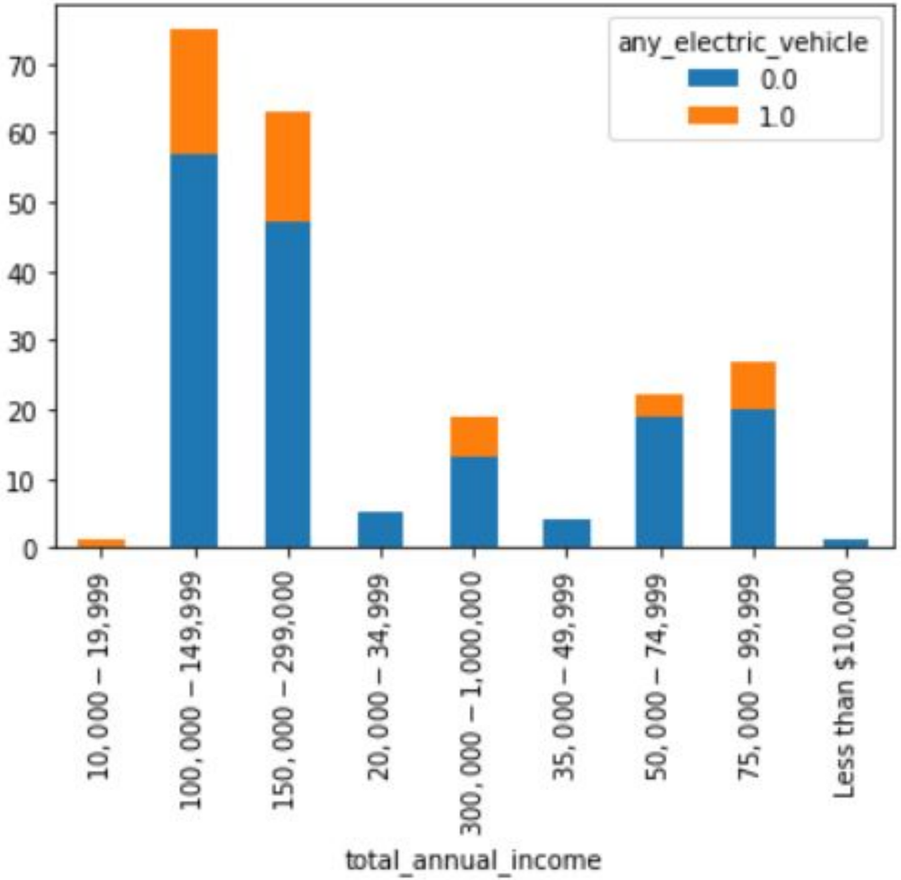
Data Science Lifecycle

- # Iterations of models and features

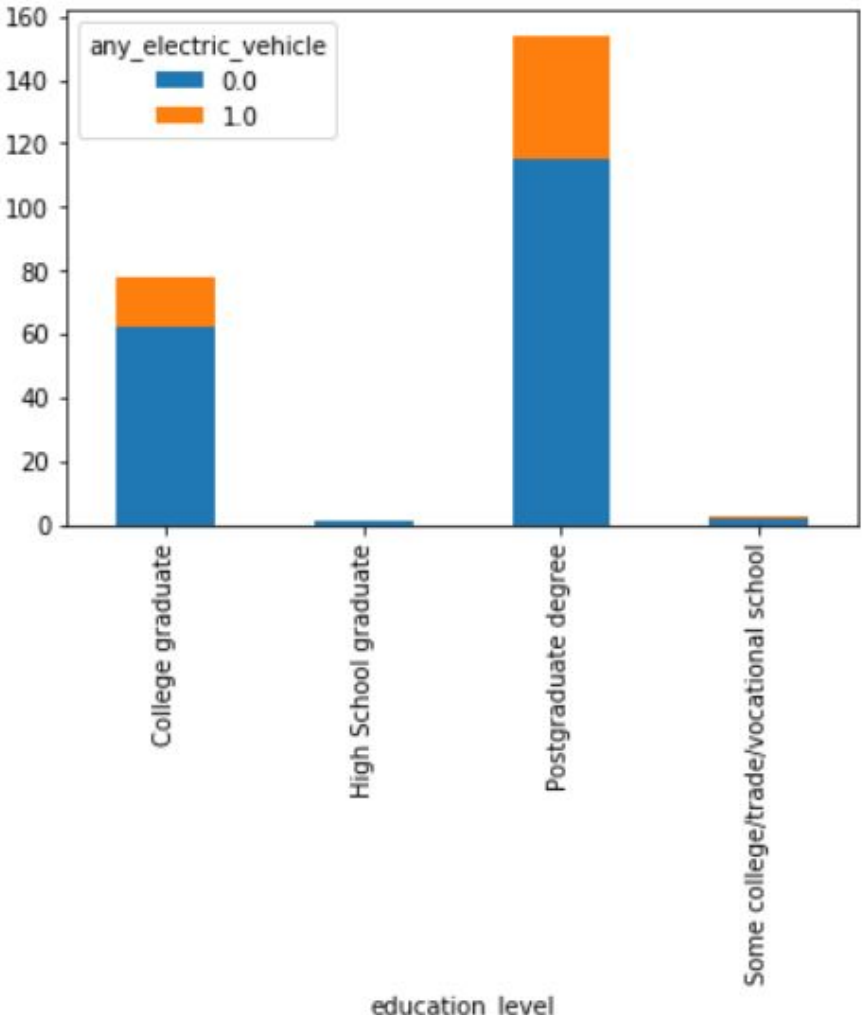
Electric Vehicle Presence Discovery

Appendix

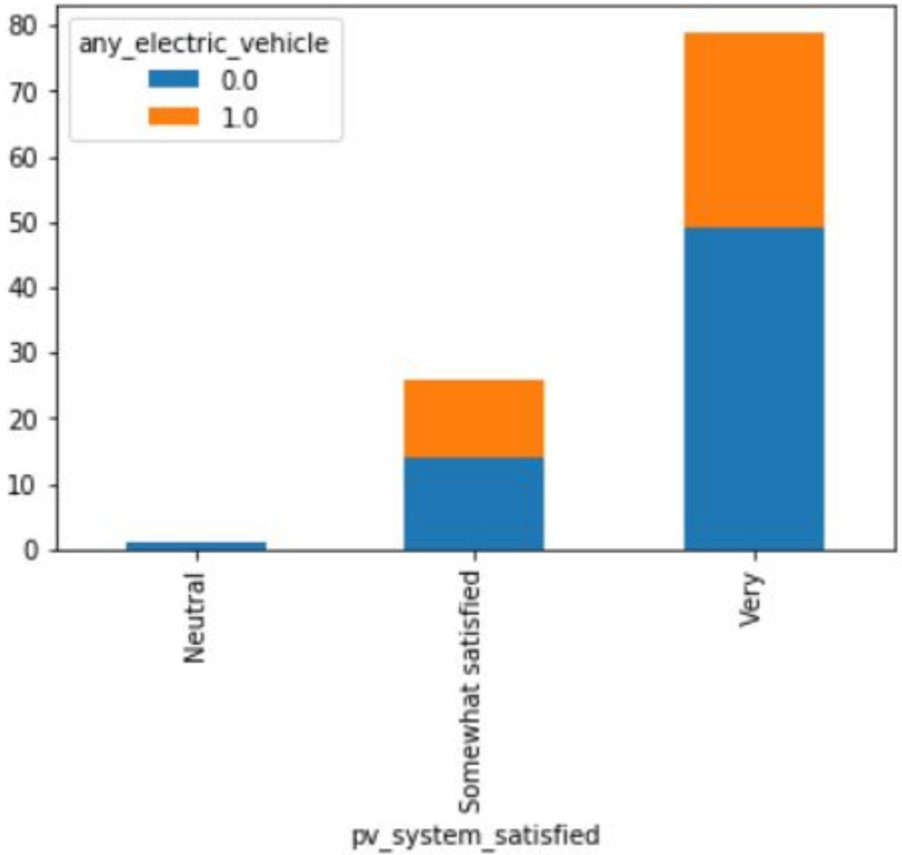
Income


any_electric_vehicle
0.0
1.0
0
10
20
30
40
50
60
70
10, 000 − 19, 999
100, 000 − 149, 999
150, 000 − 299, 000
20, 000 − 34, 999
300, 000 − 1, 000, 000
35, 000 − 49, 999
50, 000 − 74, 999
75, 000 − 99, 999
Less than $10,000
total_annual_income

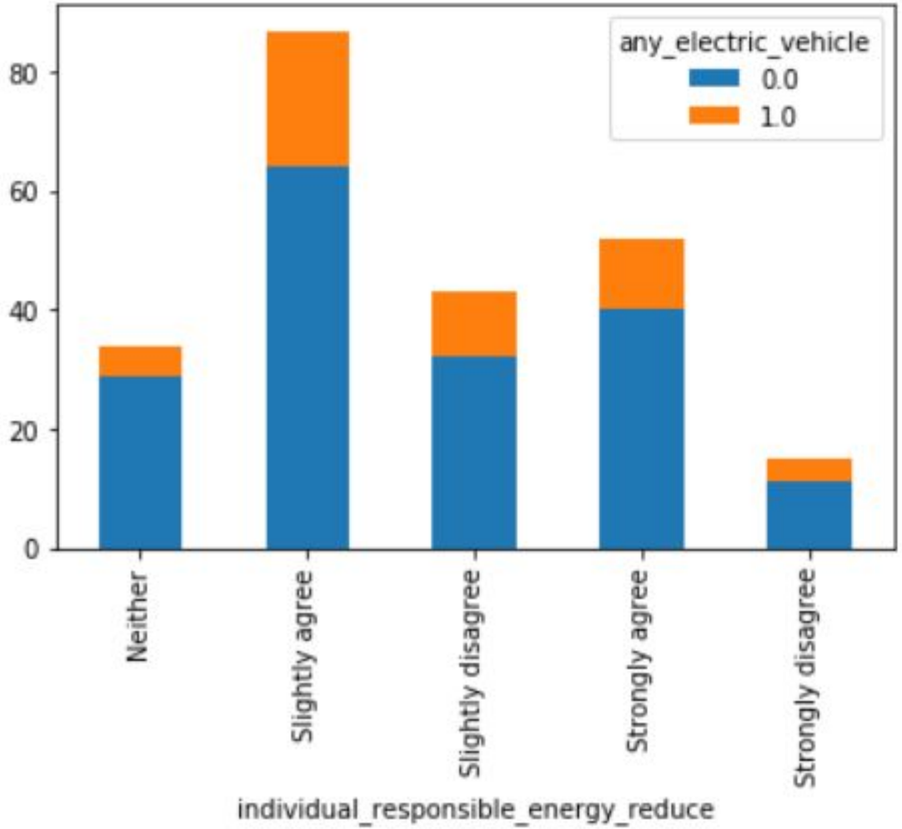
Education Level



PV Satisfaction



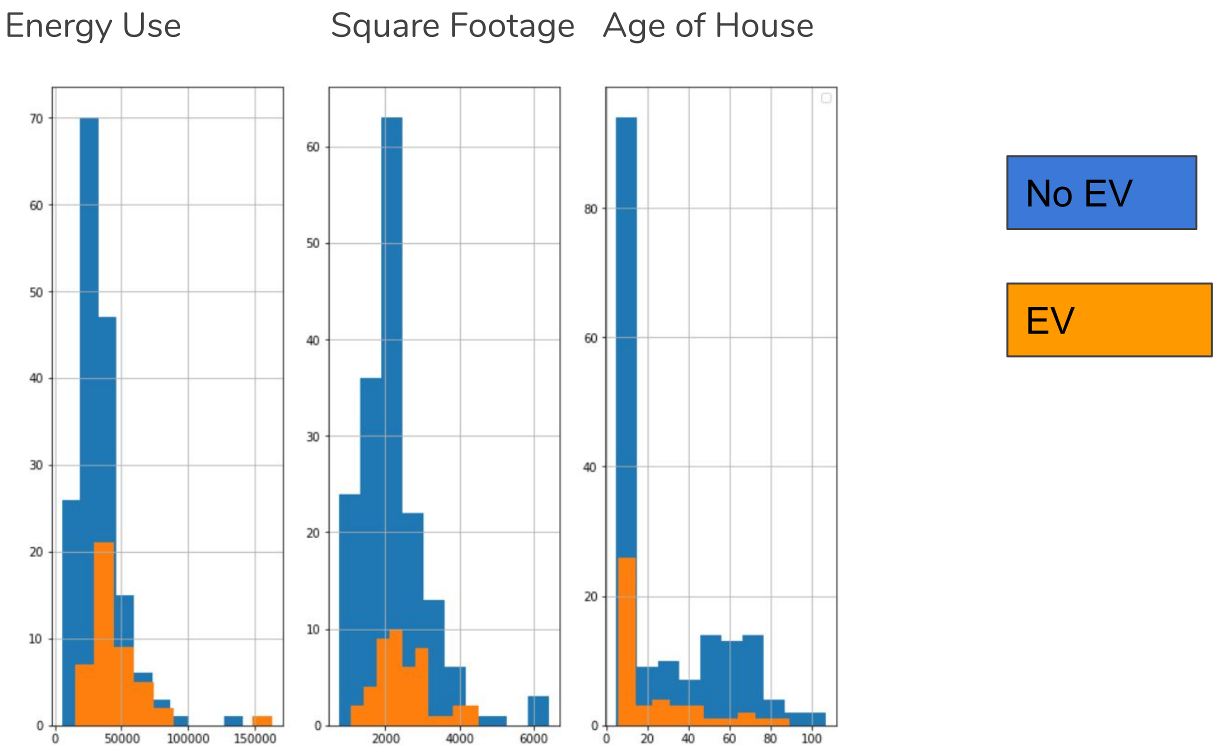
Individuals Responsible for energy reduction



Models prior to train test split

Classifier	Accuracy	Specificity	f1	Recall	precision
Random Forest	82%	96%	95%	94%	96%
Logistic Regression	71%	68%	72%	75%	70%
KNN	78%	100%	95%	90%	100%
SVM	83%	95%	97%	99%	95%
Neural Network	75%	69%	80%	85%	74%

Exploratory Data Analysis Continuous



Energy Use

Square Footage

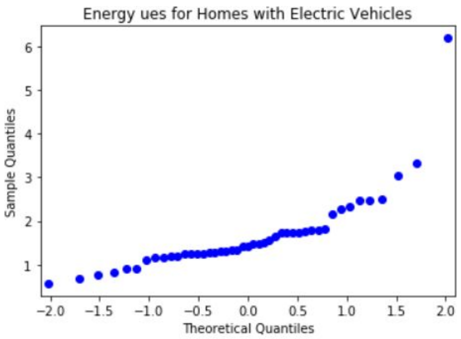
Age of House

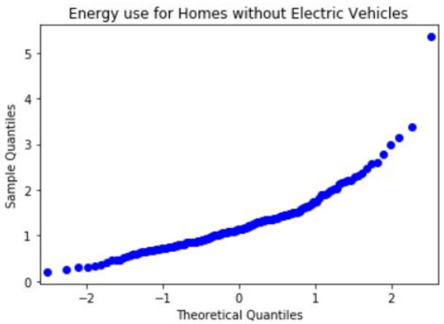
🟦 No EV

🟧 EV

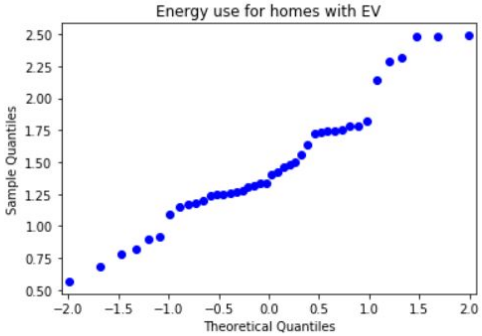
T-test on average hourly data

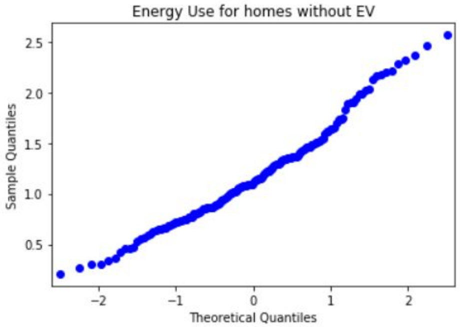
Full Data





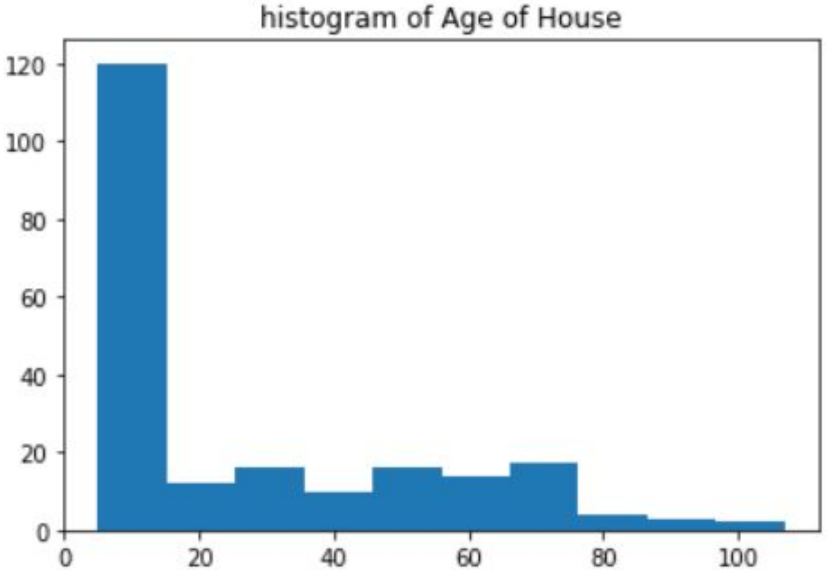
Removed Outliers

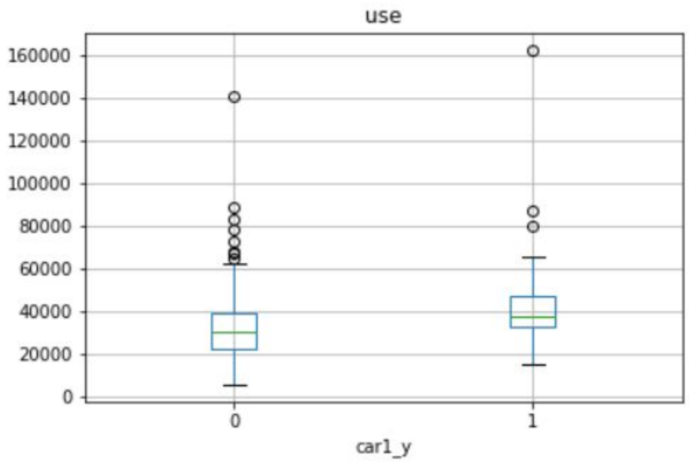


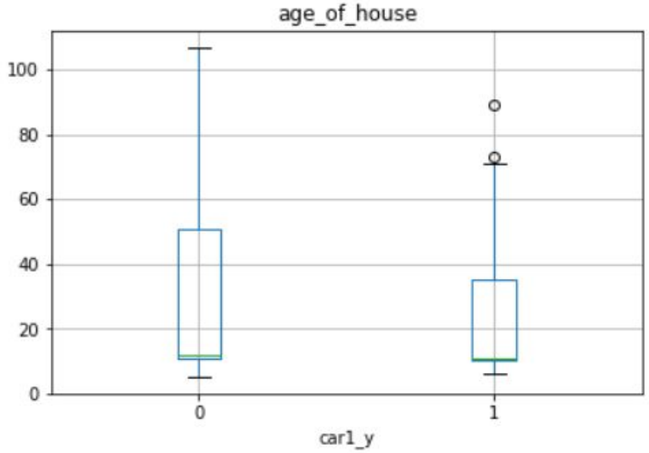


```
T Value = 2.7922847933801256, p Value = 0.007123712930791133
```

Feature Engineering

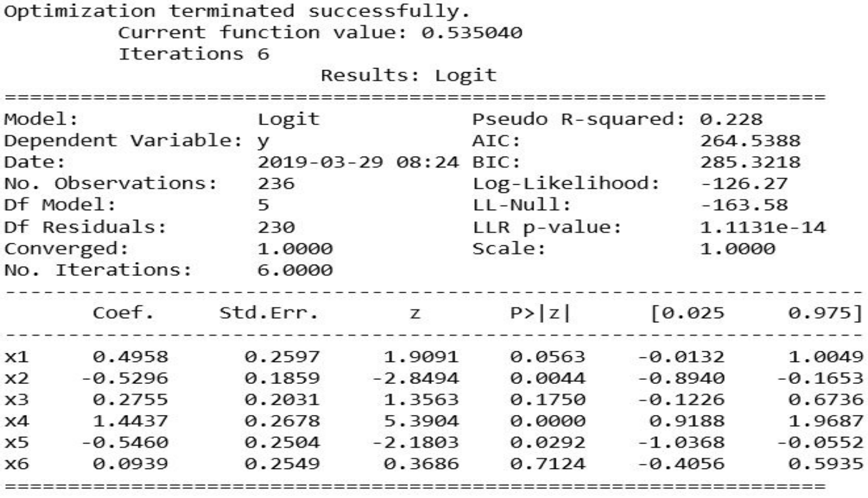

histogram of Age of House


use


age_of_house

Logistic Regression and Random Forest

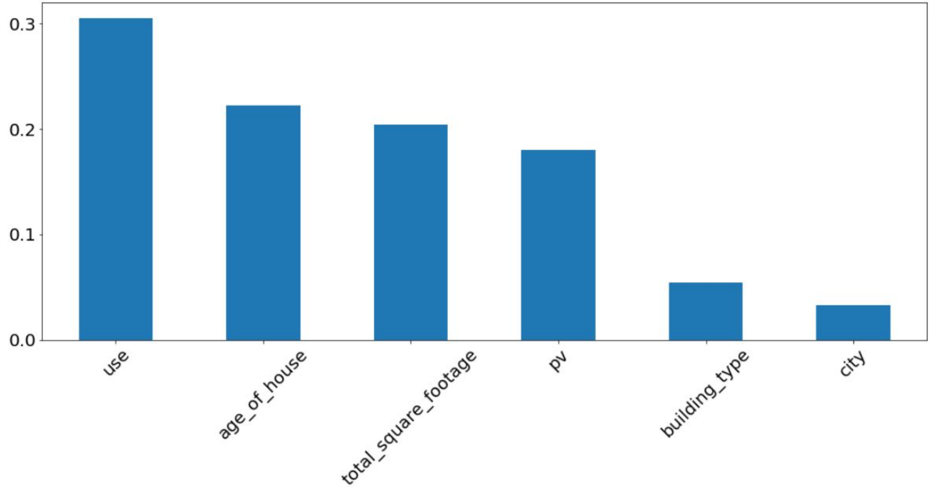
Using SMOTE (randomstate=12, ratio =1.0)



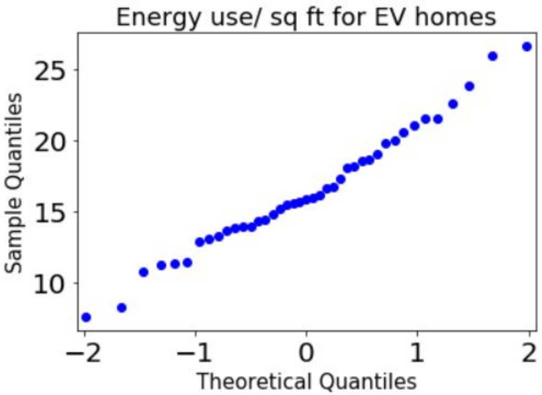
```
Optimization terminated successfully.
         Current function value: 0.535040
         Iterations 6
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.228
Dependent Variable: y                AIC:              264.5388
Date:               2019-03-29 08:24 BIC:              285.3218
No. Observations:   236              Log-Likelihood:   -126.27
Df Model:           5                LL-Null:          -163.58
Df Residuals:       230              LLR p-value:      1.1131e-14
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
--------------------------------------------------------------------
       Coef.     Std.Err.       z      P>|z|     [0.025     0.975]
--------------------------------------------------------------------
x1     0.4958     0.2597     1.9091    0.0563   -0.0132     1.0049
x2    -0.5296     0.1859    -2.8494    0.0044   -0.8940    -0.1653
x3     0.2755     0.2031     1.3563    0.1750   -0.1226     0.6736
x4     1.4437     0.2678     5.3904    0.0000    0.9188     1.9687
x5    -0.5460     0.2504    -2.1803    0.0292   -1.0368    -0.0552
x6     0.0939     0.2549     0.3686    0.7124   -0.4056     0.5935
=================================================================
```

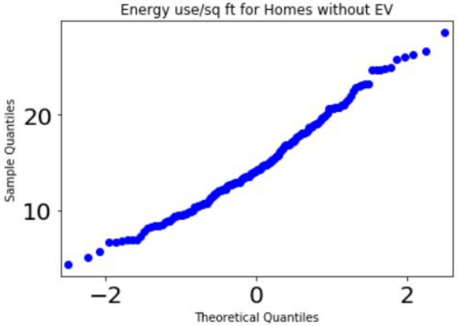
Random Forest: Accuracy on train: 0.945

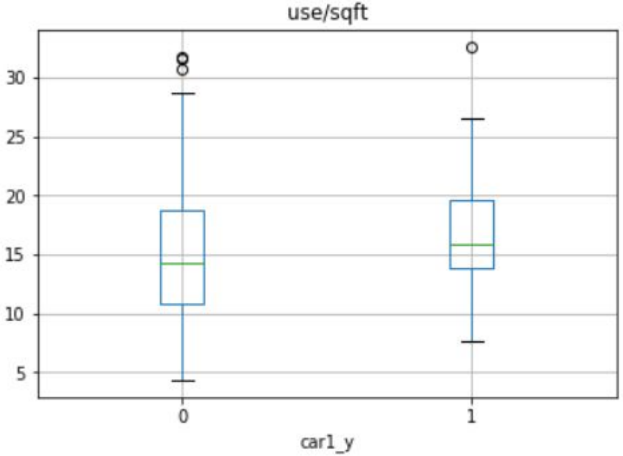
Accuracy on test: 0.662



T-test on energy use/ square foot



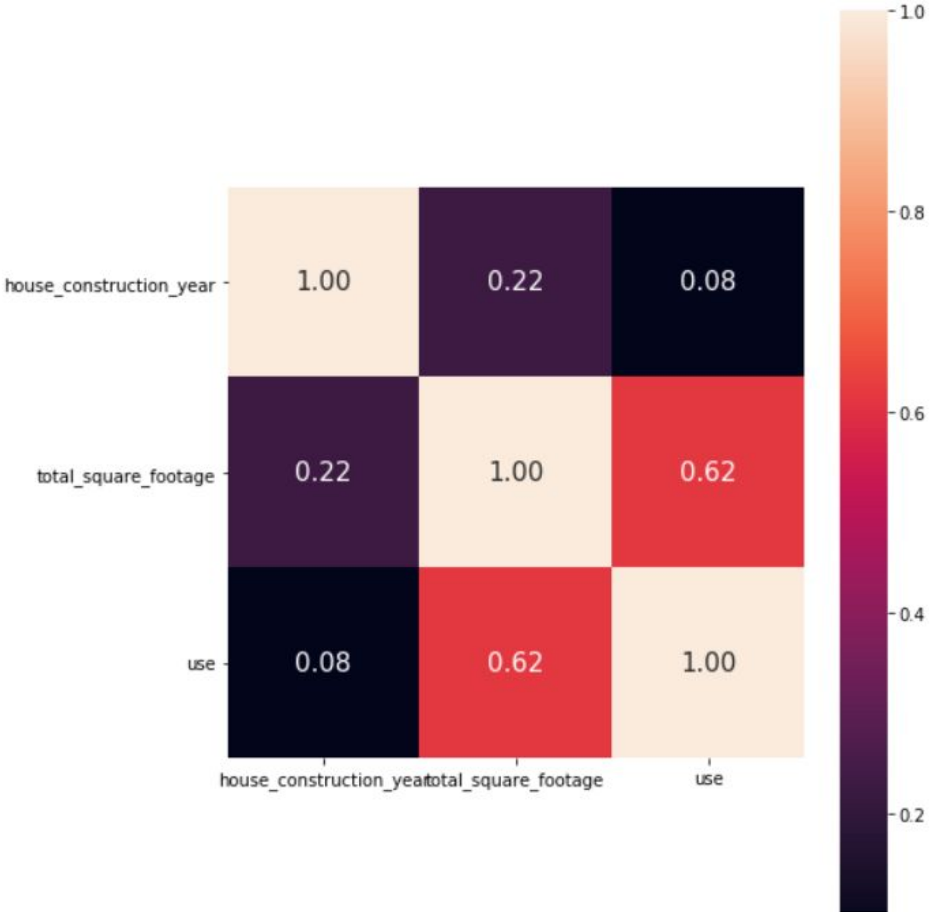




T value = 2.05

P value = 0.04

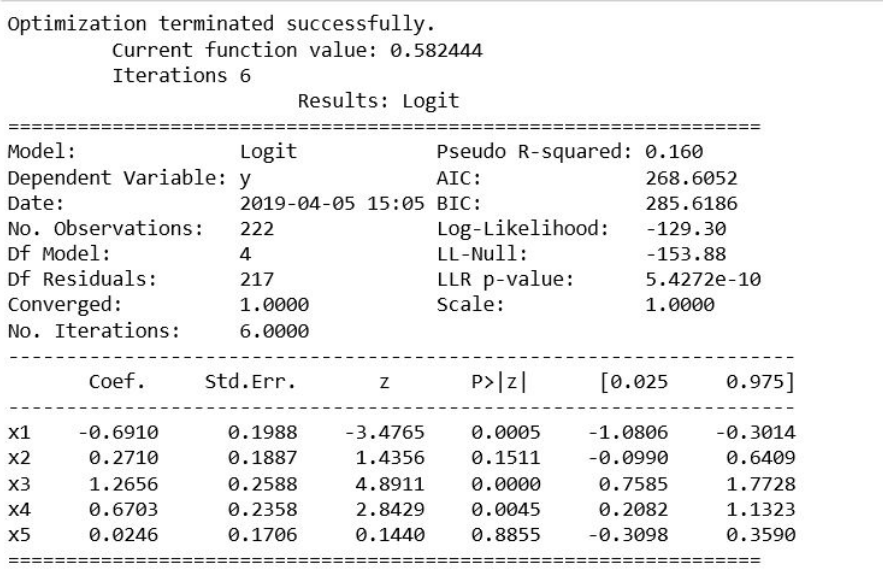
Correlation



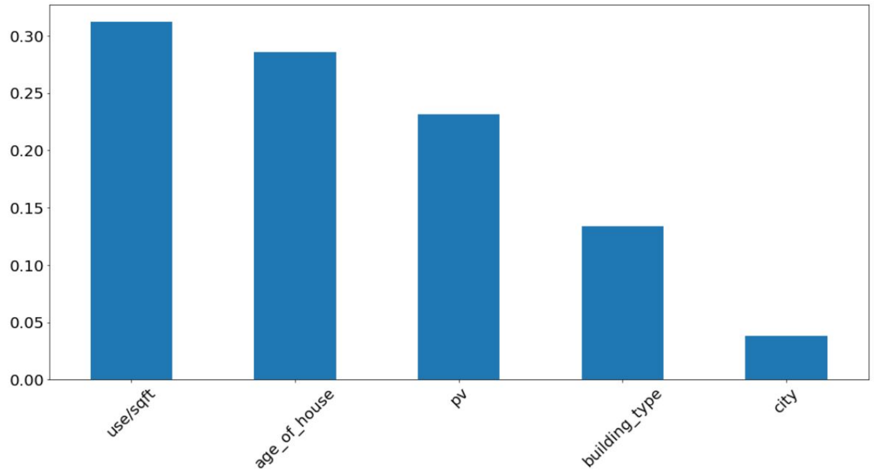
Logistic Regression and Random Forest

Using DF3 SMOTE (random_state =12, Ratio=1

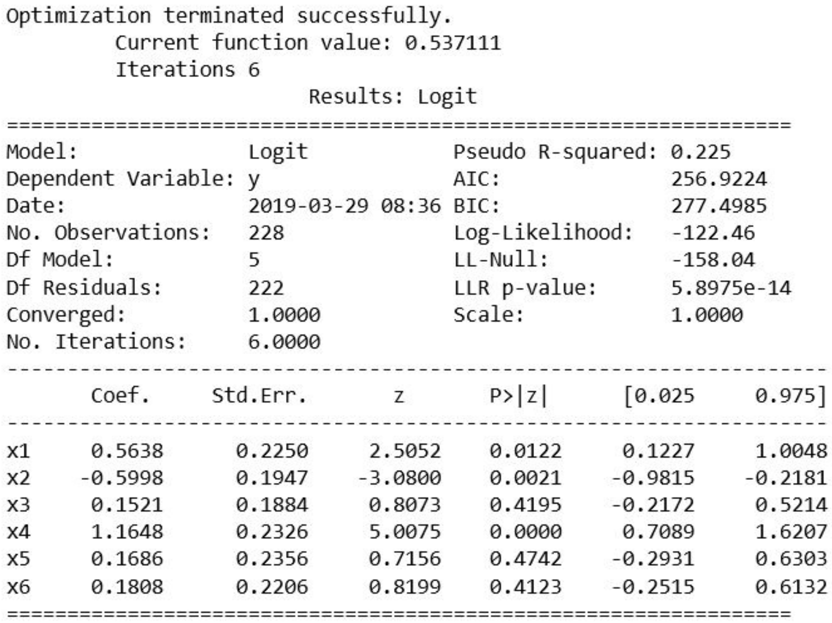
Random Forest: Accuracy on train: 0.923
Accuracy on test: 0.672



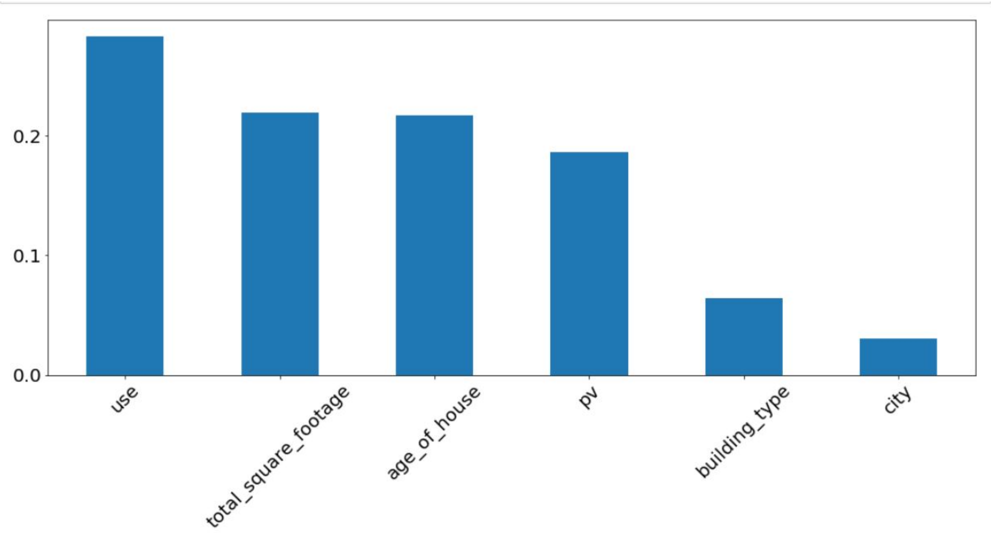
```
Optimization terminated successfully.
         Current function value: 0.582444
         Iterations 6
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.160
Dependent Variable: y                AIC:              268.6052
Date:               2019-04-05 15:05 BIC:              285.6186
No. Observations:   222              Log-Likelihood:   -129.30
Df Model:           4                LL-Null:          -153.88
Df Residuals:       217              LLR p-value:      5.4272e-10
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
--------------------------------------------------------------------
       Coef.    Std.Err.      z      P>|z|     [0.025    0.975]
--------------------------------------------------------------------
x1   -0.6910     0.1988   -3.4765   0.0005   -1.0806   -0.3014
x2    0.2710     0.1887    1.4356   0.1511   -0.0990    0.6409
x3    1.2656     0.2588    4.8911   0.0000    0.7585    1.7728
x4    0.6703     0.2358    2.8429   0.0045    0.2082    1.1323
x5    0.0246     0.1706    0.1440   0.8855   -0.3098    0.3590
=================================================================
```



DF2 = Best Model



```
Optimization terminated successfully.
         Current function value: 0.537111
         Iterations 6
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.225
Dependent Variable: y                AIC:              256.9224
Date:               2019-03-29 08:36 BIC:              277.4985
No. Observations:   228              Log-Likelihood:   -122.46
Df Model:           5                LL-Null:          -158.04
Df Residuals:       222              LLR p-value:      5.8975e-14
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
--------------------------------------------------------------------
       Coef.     Std.Err.       z       P>|z|      [0.025     0.975]
--------------------------------------------------------------------
x1     0.5638      0.2250     2.5052    0.0122     0.1227     1.0048
x2    -0.5998      0.1947    -3.0800    0.0021    -0.9815    -0.2181
x3     0.1521      0.1884     0.8073    0.4195    -0.2172     0.5214
x4     1.1648      0.2326     5.0075    0.0000     0.7089     1.6207
x5     0.1686      0.2356     0.7156    0.4742    -0.2931     0.6303
x6     0.1808      0.2206     0.8199    0.4123    -0.2515     0.6132
=================================================================
```

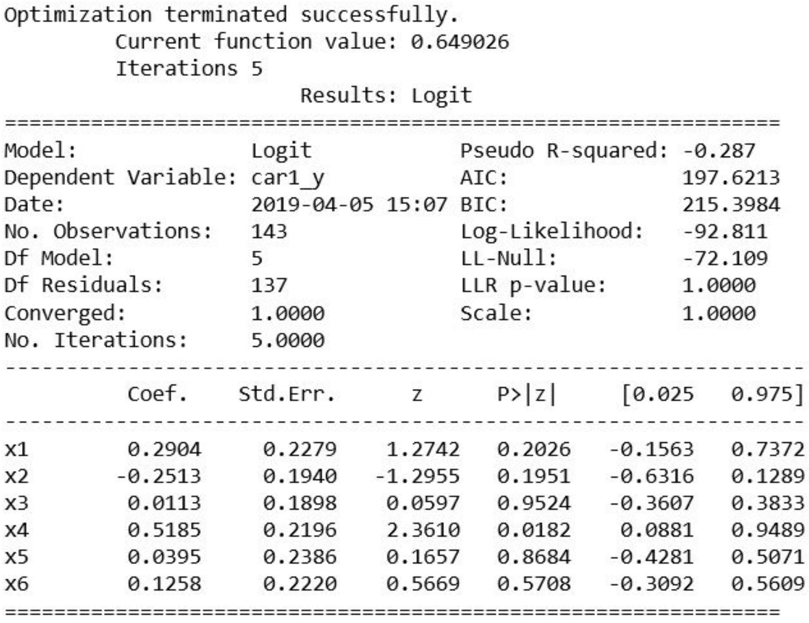


Logistic Regression and Random Forest

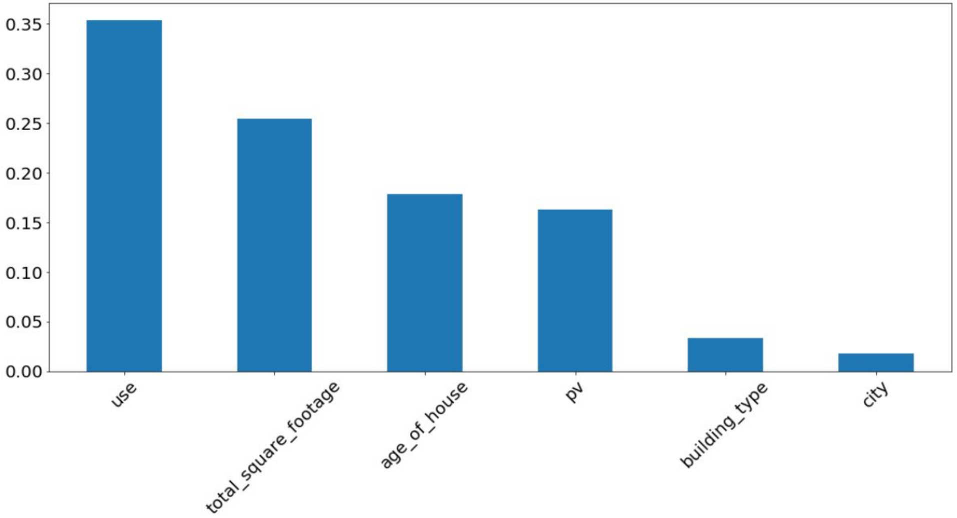
Using DF2 80,20

Random Forest: Accuracy on train: 0.860

Accuracy on test: 0.774



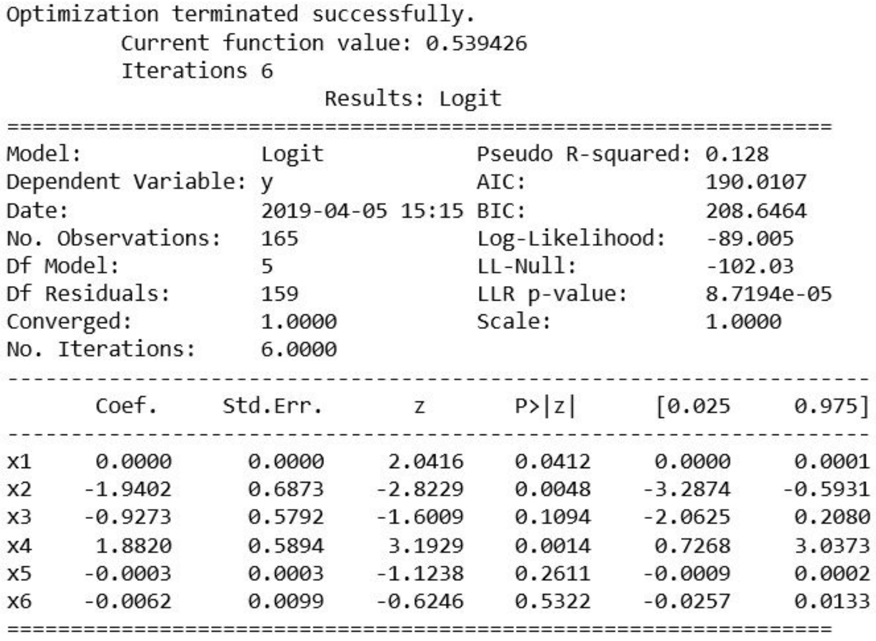
```
Optimization terminated successfully.
         Current function value: 0.649026
         Iterations 5
                        Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: -0.287
Dependent Variable: car1_y           AIC:              197.6213
Date:               2019-04-05 15:07 BIC:              215.3984
No. Observations:   143              Log-Likelihood:   -92.811
Df Model:           5                LL-Null:          -72.109
Df Residuals:       137              LLR p-value:      1.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     5.0000
------------------------------------------------------------------
          Coef.    Std.Err.      z      P>|z|     [0.025   0.975]
------------------------------------------------------------------
x1        0.2904    0.2279    1.2742   0.2026   -0.1563   0.7372
x2       -0.2513    0.1940   -1.2955   0.1951   -0.6316   0.1289
x3        0.0113    0.1898    0.0597   0.9524   -0.3607   0.3833
x4        0.5185    0.2196    2.3610   0.0182    0.0881   0.9489
x5        0.0395    0.2386    0.1657   0.8684   -0.4281   0.5071
x6        0.1258    0.2220    0.5669   0.5708   -0.3092   0.5609
=================================================================
```



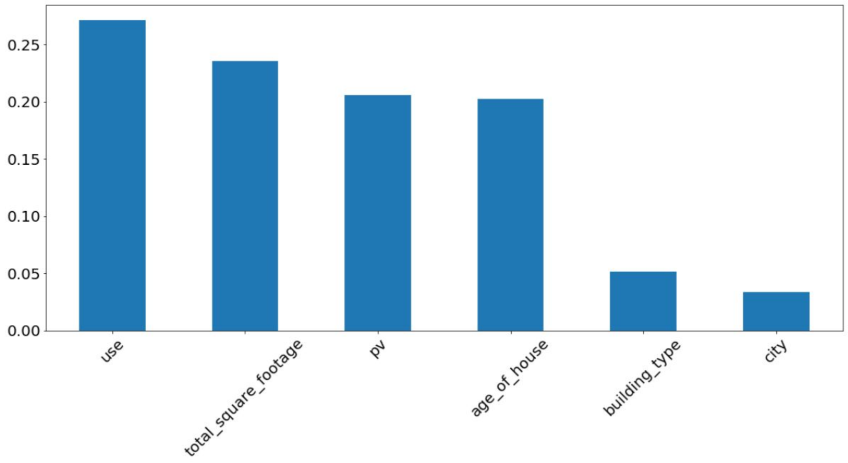
Logistic Regression and Random Forest

Using DF2 70,30

Random Forest: Accuracy on train: 0.909
Accuracy on test: 0.790



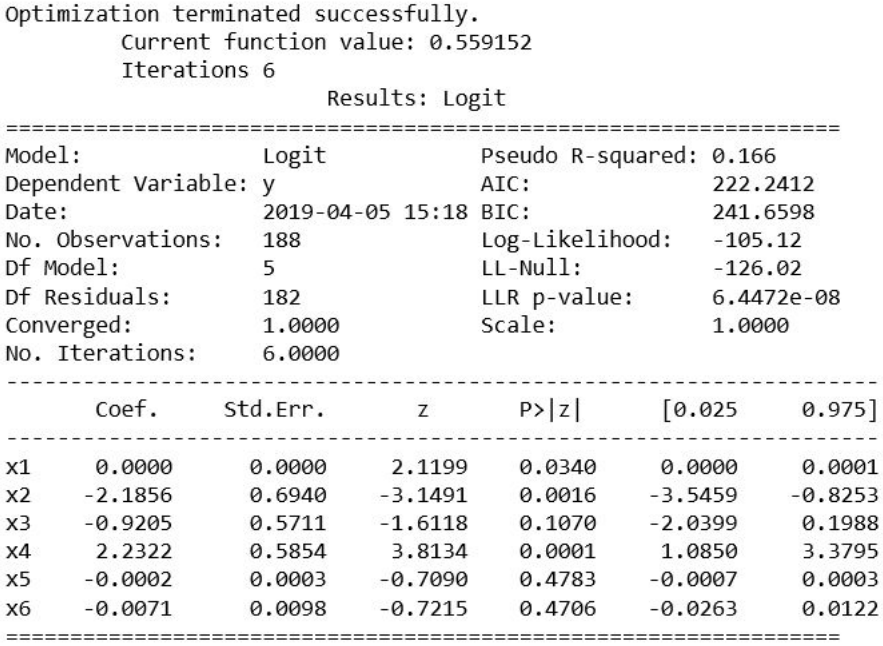
```
Optimization terminated successfully.
         Current function value: 0.539426
         Iterations 6
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.128
Dependent Variable: y                AIC:              190.0107
Date:               2019-04-05 15:15 BIC:              208.6464
No. Observations:   165              Log-Likelihood:   -89.005
Df Model:           5                LL-Null:          -102.03
Df Residuals:       159              LLR p-value:      8.7194e-05
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
--------------------------------------------------------------------
       Coef.     Std.Err.      z       P>|z|      [0.025     0.975]
--------------------------------------------------------------------
x1     0.0000     0.0000     2.0416    0.0412     0.0000     0.0001
x2    -1.9402     0.6873    -2.8229    0.0048    -3.2874    -0.5931
x3    -0.9273     0.5792    -1.6009    0.1094    -2.0625     0.2080
x4     1.8820     0.5894     3.1929    0.0014     0.7268     3.0373
x5    -0.0003     0.0003    -1.1238    0.2611    -0.0009     0.0002
x6    -0.0062     0.0099    -0.6246    0.5322    -0.0257     0.0133
=================================================================
```



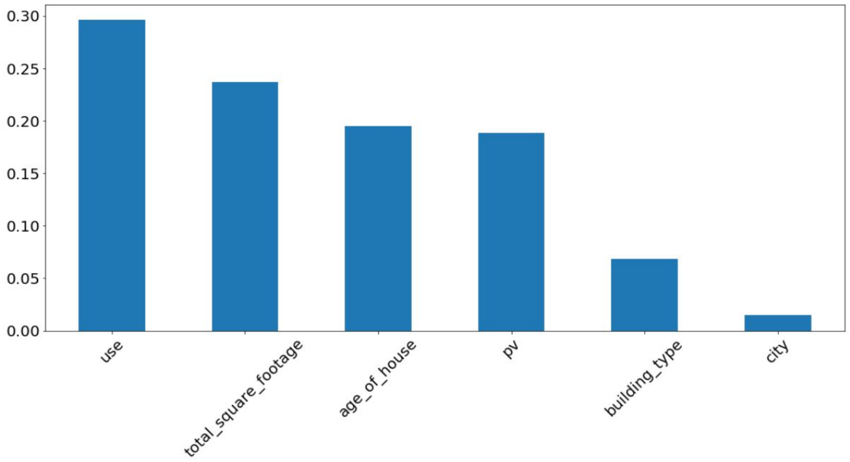
Logistic Regression and Random Forest

Using DF2 60,40

Random Forest: Accuracy on train: 0.904
Accuracy on test: 0.790



```
Optimization terminated successfully.
         Current function value: 0.559152
         Iterations 6
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.166
Dependent Variable: y                AIC:              222.2412
Date:               2019-04-05 15:18 BIC:              241.6598
No. Observations:   188              Log-Likelihood:   -105.12
Df Model:           5                LL-Null:          -126.02
Df Residuals:       182              LLR p-value:      6.4472e-08
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
--------------------------------------------------------------------
       Coef.     Std.Err.        z       P>|z|      [0.025     0.975]
--------------------------------------------------------------------
x1     0.0000     0.0000      2.1199    0.0340     0.0000     0.0001
x2    -2.1856     0.6940     -3.1491    0.0016    -3.5459    -0.8253
x3    -0.9205     0.5711     -1.6118    0.1070    -2.0399     0.1988
x4     2.2322     0.5854      3.8134    0.0001     1.0850     3.3795
x5    -0.0002     0.0003     -0.7090    0.4783    -0.0007     0.0003
x6    -0.0071     0.0098     -0.7215    0.4706    -0.0263     0.0122
=================================================================
```


use
total_square_footage
age_of_house
pv
building_type
city