DATS 6202

Machine Learning I

Dr. Amir Jafari

Group 10 Final Report on

Maryland Property Data

by

Tyler Wallett

<u>Introduction</u>: The project aspired to approximate the consideration of residential properties in Montgomery County, Maryland. To this extent, an appropriate exploratory data analysis (EDA), feature selection, and modeling were created to properly approximate these residential considerations. Lastly, two extra sections were devised: use cases and limitations; to properly interpret the practicality and handicaps of the modeling results. The work was done solely by me with some assistance from outside sources that are cited. The reason behind choosing this topic was to challenge the conception behind the article *Getting Ahead of the Market: How Big Data Is Transforming Real Estate* by McKinsey & Company where non-traditional features were shown to explain roughly 60% of the variation in the predictive power (App. 1) Therefore, my goal is to attempt to answer this question: Can we successfully approximate consideration prices of residential properties from Montgomery County, Maryland **only using traditional features**?

<u>Description of the dataset</u>: The public data was gathered from the Maryland State Department of Assessments and Taxation (SDAT) and contains roughly 2.4 million properties in the state of Maryland, and was last updated on January 7, 2023. The dataset also contains 134 features describing each property parcel (64 numeric - 70 categorical) The subset of the dataset we are interested in, 'Land usage' residential and 'County' Montgomery County, contains roughly 145,000 residential properties (App. 2)

<u>Description of the machine learning network and training algorithm or other algorithms that you used</u>: I developed a Statsmodels OLS regression model, Scikit-Learn's Random Forest regression models with different 'max_depths', a "homemade" multilayer perceptron regression model with different amounts of neurons in the hidden layer, and a Scikit-Learn multilayer perceptron regression model with also different hidden layer sizes.

<u>Theoretical description experimental setup</u>: I am going to use the data from the best features from the EDA and Feature Selection process to pass them through Scikit-Learn's 'train_test_split' function, and pre-process them using Scikit-Learn's 'QuantileTransformation' function to properly fit the models (App. 3 & 4) I will train the machine learning algorithm on the 'X_train' and 'Y_train' variables, to then produce predictions and a R^2 score from 'X_test' and 'Y_test', respectively. Then, I will judge the performance of each machine learning model from the R^2 score, mean squared error along with a 95% confidence interval and using the loss_curve method.

<u>Actual description experimental setup</u>: The linear regression model from statsmodels yielded an R^2 score of 0.836, and as for the Random Forest regression model with "max_depth" 2 had 0.6712, 3 had 0.7734, 4 had 0.8161, 6 had 0.8476, and 10 had 0.8503 (App. 5) These scores were compared with the neural networks models as "benchmarks". As for the "homemade" multilayer perceptron the different R^2 scores were: 2 neurons hidden layer 0.8377, 3 neurons hidden layer 0.8482, 6 neurons hidden layer 0.8495, 10 neurons hidden layer 0.8504, and 100 neurons hidden layer 0.8469 (App. 6) The "homemade" neural network appears to outperform the linear regression with just 2 neurons and the Random Forest regressor with its best model: 10 neurons. As for Scikit-Learn's multilayer perceptron the results were: 2 neurons hidden layer 0.8447
, 3 neurons hidden layer 0.8502, 6 neurons hidden layer 0.8504, 10 neurons hidden layer 0.8506, and 100 neurons hidden layer 0.8430 (App. 7) Once again, the model outperformed both

benchmarks and the "homemade" neural network with 10 neurons, which makes it the best model to approximate consideration prices.

<u>Results</u>: From the various models that were implemented, I was able to successfully capture 85% of the variance in the prices of consideration, or a 0.85 R^2 score, by using Scikit-Learn's multilayer perceptron regression model with a structure of 2-100-1. Additionally, the mean squared error with 95% confidence interval for this model was of 0.15 ∓ 0.01, and the error of the post-processed model with 95% confidence interval was of approximately 16,000$ ∓ 4,000$ (App. 8)

<u>Summary and conclusions</u>:

- <u>From the EDA</u>: Distributions were highly-right skewed. On average consideration is higher when residences are close to Washington D.C.
- <u>From the Feature Selection</u>: Best model would include all variables, but to avoid collinearity only "Land value" and "Land improvements" should be used.
- <u>From the Modeling</u>: Scikit-learn MLPRegressor can predict approximately 85% of the variance in consideration.

<u>Limitations</u>:

- Lots of redundant Id's, geographical and boolean features.
- Lack of traditional features (e.g. amount of bedrooms, historical values)
- Lack of reproducibility (e.g. the analysis will work for Maryland however it will not work for Massachusetts)
- The definition of consideration: the value expected by the owner of a property.

<u>References</u>:

1. Motivation:

   Asaftei, Gabriel Morgan, et al. "Getting Ahead of the Market: How Big Data Is Transforming Real Estate." *McKinsey & Company*, McKinsey & Company, 8 Oct. 2018, https://www.mckinsey.com/industries/real-estate/our-insights/getting-ahead-of-the-market-how-big-data-is-transforming-real-estate.

2. Dataset & S.M.A.R.T. Question:

   https://hub.arcgis.com/datasets/maryland::maryland-property-data-parcel-points/about

3. EDA & Models:

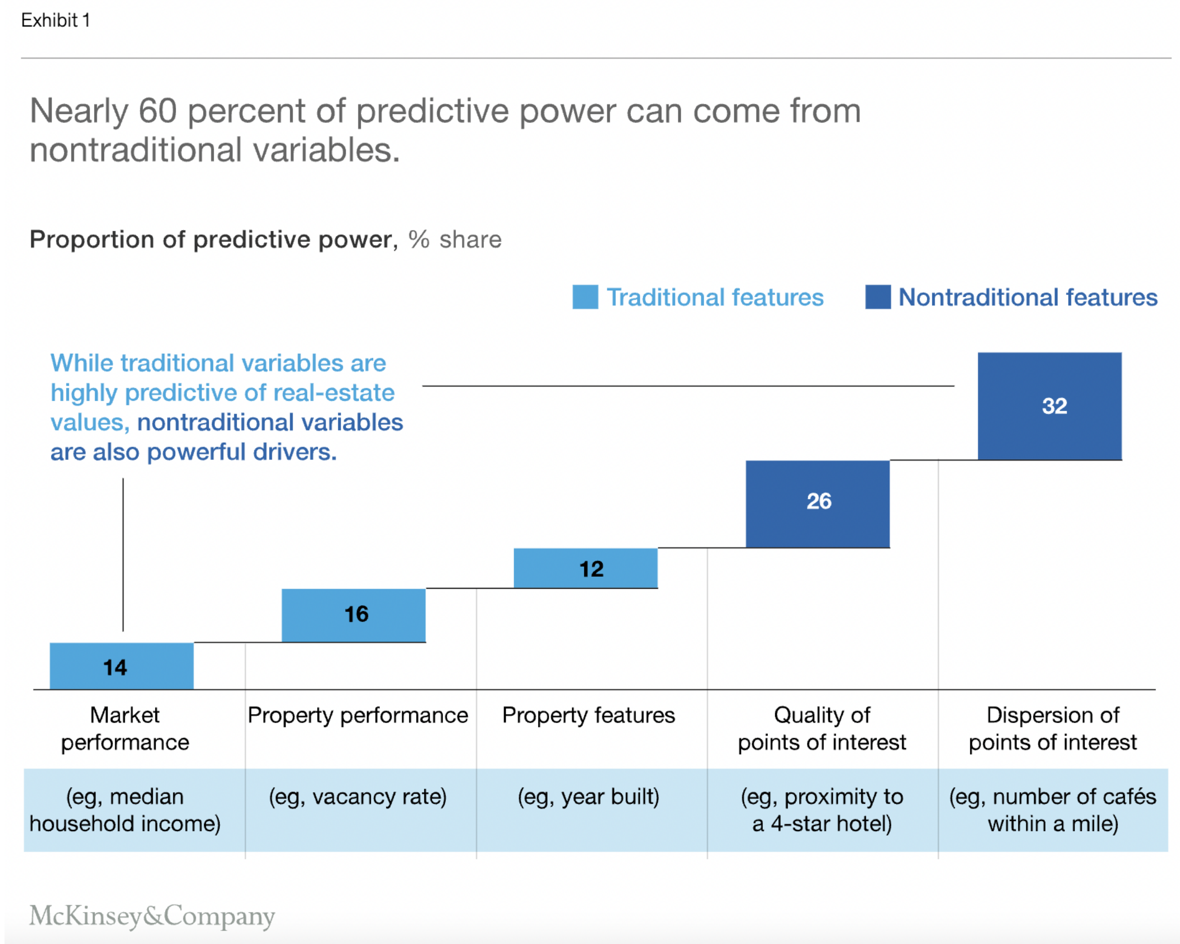
   ChatGPT. (2023, April 14)

4. Feature Selection:

   Harrison, Matt. *Machine Learning Pocket Reference: Working with Structured Data in Python*. O'Reilly Media, Inc., 2019.

<u>Appendix</u>:

(App. 1)

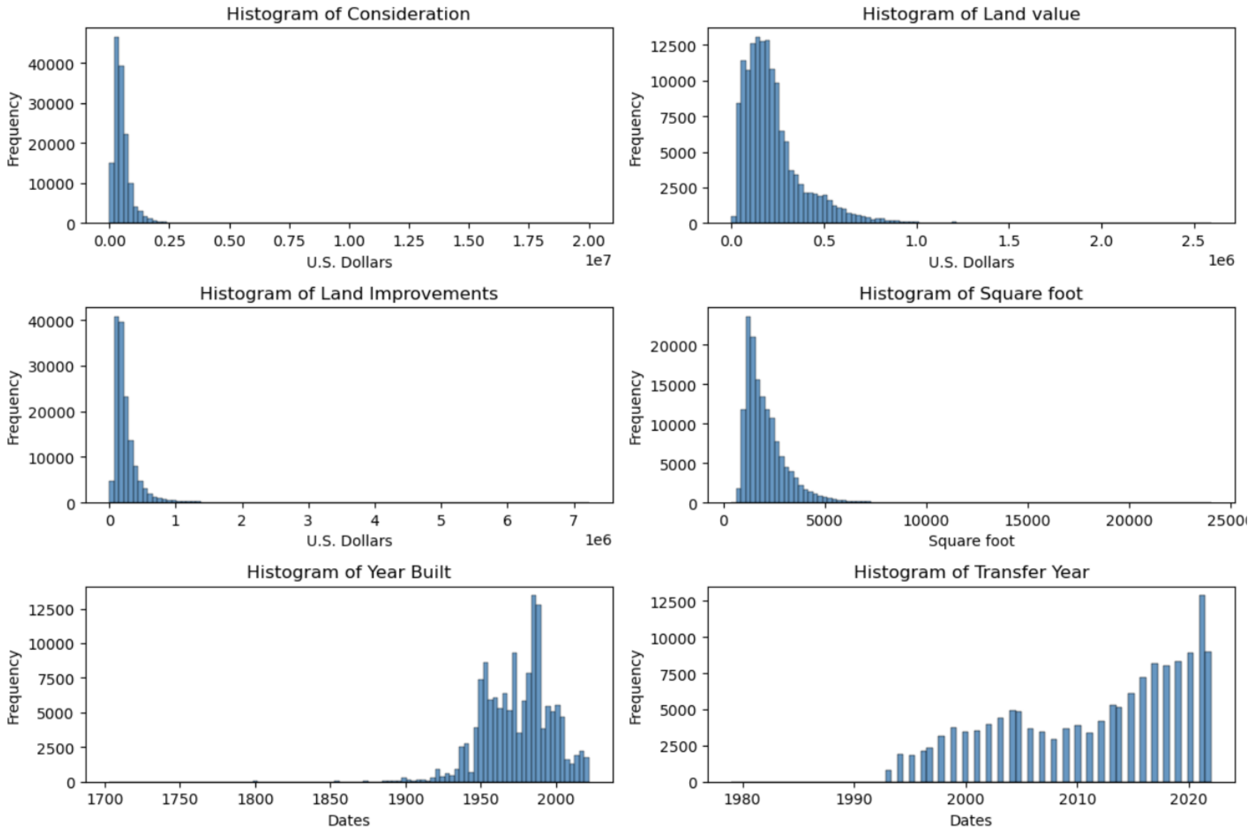


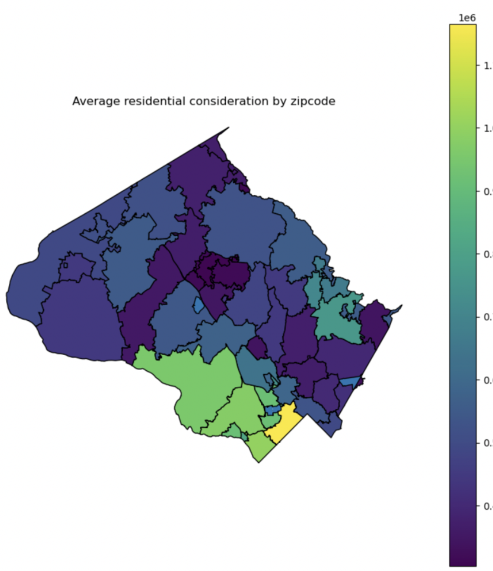
(App. 2)

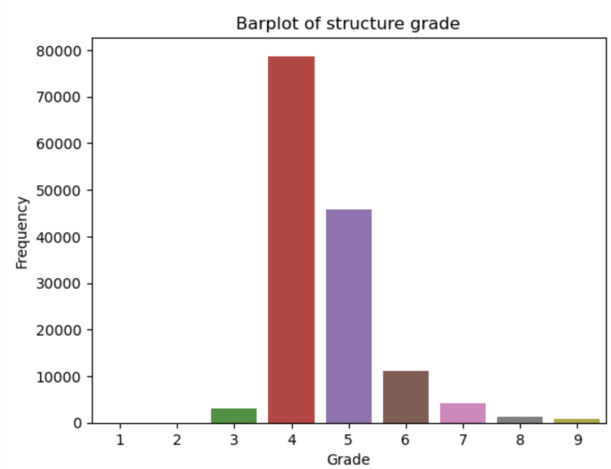
	Id	Longitude	Latitude	Address	Zipcode	Grade	Year_built	Sqft	Trade_date	Consideration	Land_value
588336	160100000033	-77.168495	39.207629	21411 WOODFIELD RD	20882.0	3.0	1936.0	1064.0	20190702.0	260000.0	231200.0
588338	160100000066	-77.166819	39.207889	8120 BRINK RD	20882.0	2.0	1923.0	864.0	20190815.0	100000.0	174700.0
588342	160100000113	-77.178153	39.198018	8615 LOCHAVEN DR	20882.0	5.0	1840.0	2968.0	19931220.0	355000.0	120970.0
588343	160100000124	-77.141826	39.200285	6934 WARFIELD RD	20882.0	4.0	1978.0	1896.0	20180509.0	475000.0	244800.0
588349	160100000204	-77.122566	39.257719	24501 HIPSLEY MILL RD	20882.0	4.0	1913.0	2552.0	20170516.0	525000.0	201500.0
...	...	...	...	...	...	...	...	...	...	...	...
934877	161303841203	-77.070689	39.029270	10543 SAINT PAUL ST	20895.0	6.0	1893.0	2209.0	20200506.0	855000.0	498100.0
934879	161303841794	-77.078684	39.025855	10311 DETRICK AVE	20895.0	5.0	1951.0	1549.0	20200602.0	760000.0	508800.0
934883	161303842446	-77.071083	39.020108	10031 FREDERICK AVE	20895.0	7.0	2020.0	2674.0	20210419.0	1598500.0	486700.0
934967	161303856077	-77.034956	39.081966	13837 ALDERTON RD	20906.0	6.0	2021.0	2964.0	20220407.0	823480.0	169300.0
934975	161303856157	-77.036462	39.080937	13708 ALDERTON RD	20906.0	6.0	2021.0	2964.0	20220407.0	811066.0	160500.0

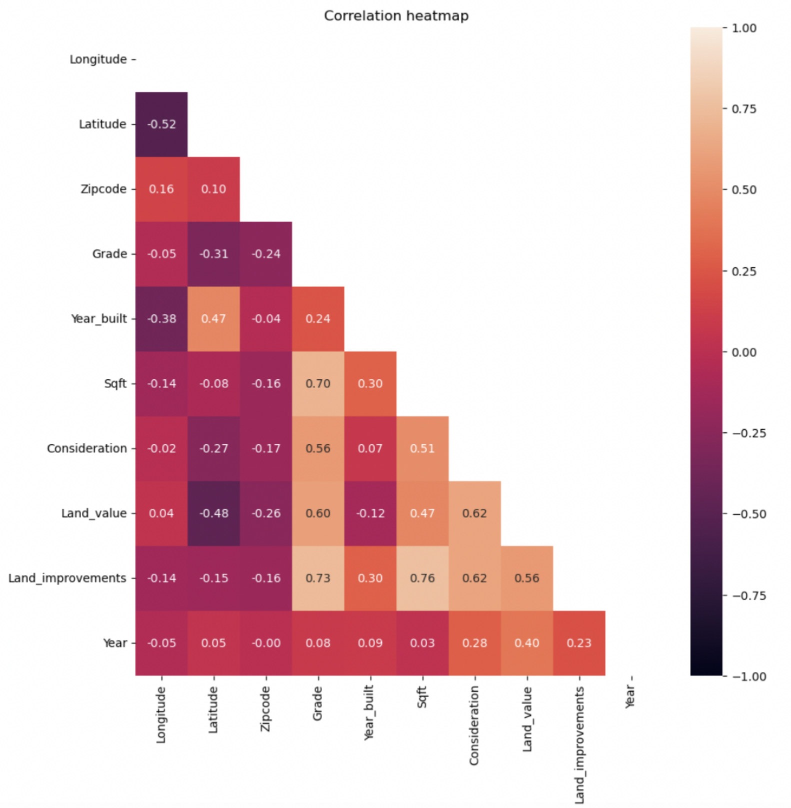
145133 rows × 11 columns

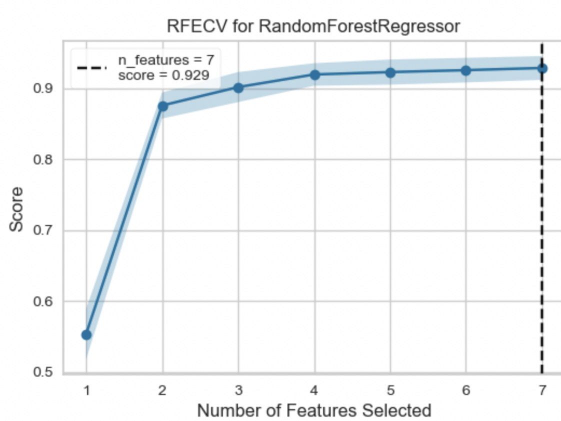
(App. 3)


Histogram of Consideration
Frequency
U.S. Dollars
1e7
Histogram of Land value
Frequency
U.S. Dollars
1e6
Histogram of Land Improvements
Frequency
U.S. Dollars
1e6
Histogram of Square foot
Frequency
Square foot
Histogram of Year Built
Frequency
Dates
Histogram of Transfer Year
Frequency
Dates


Average residential consideration by zipcode
1e6

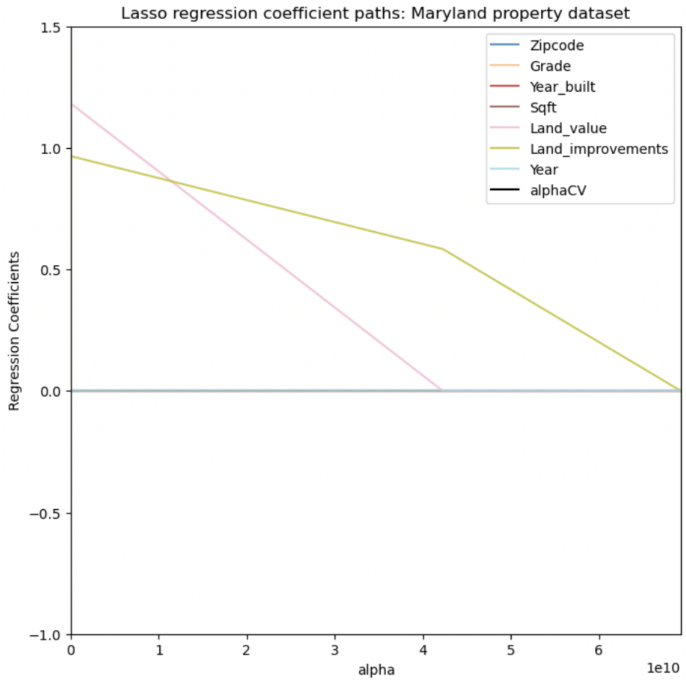

Barplot of structure grade
Frequency
Grade





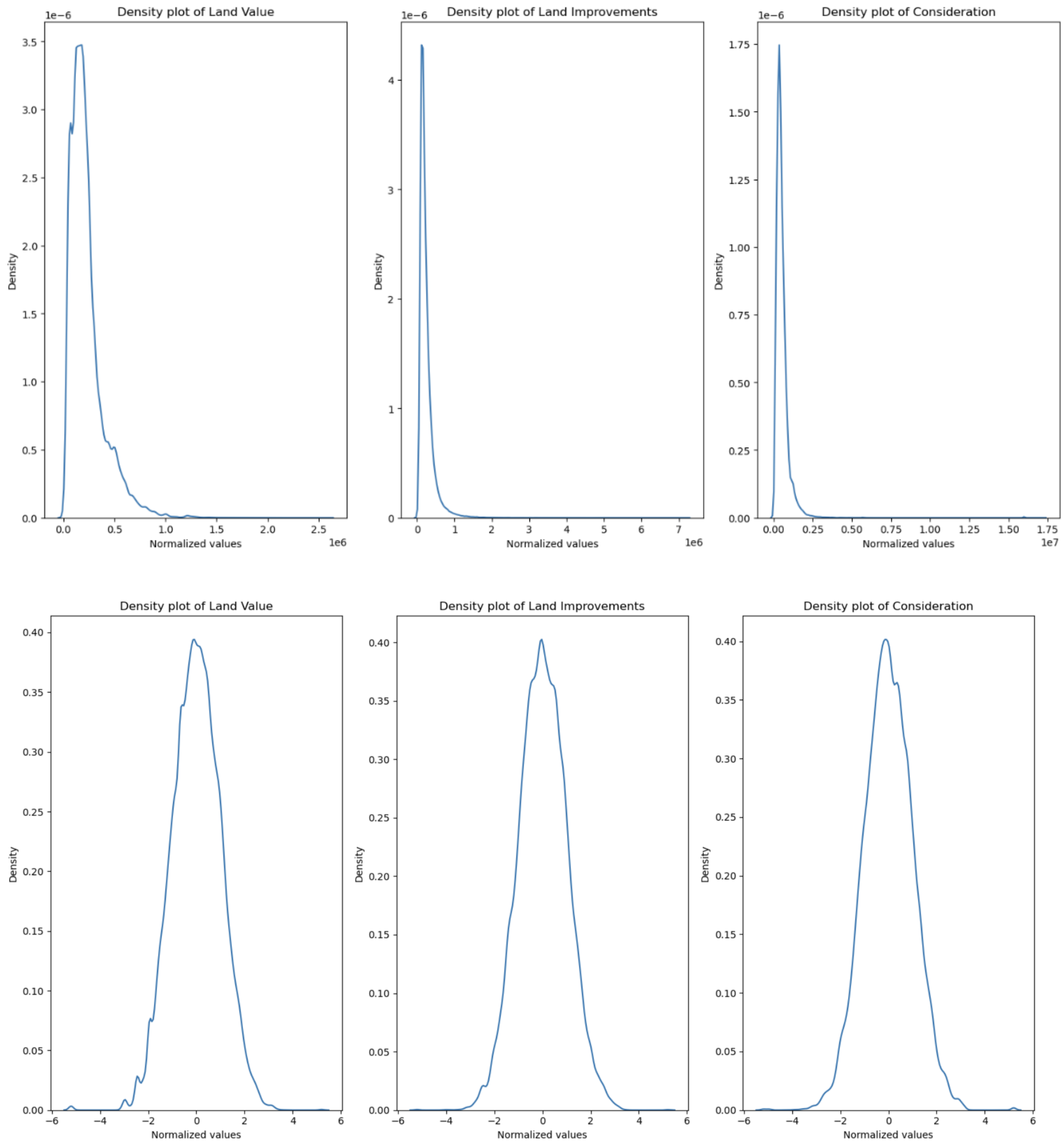

```
top_features
✓ 0.0s
Index(['Land_improvements', 'Land_value', 'Grade', 'Year'], dtype='o
```

```
top_features
✓ 0.0s
Index(['Land_improvements', 'Land_value'], dtype='object')
```




```
Initial conditional number: 765411.58

Conditional number without regressor `Zipcode`:759817.36
Decrease in conditional number: 5594.22

Conditional number without regressor `Grade`:30342.73
Decrease in conditional number: 729474.63

Conditional number without regressor `Year_built`:724.26
Decrease in conditional number: 29618.47

Conditional number without regressor `Sqft`:360.23
Decrease in conditional number: 364.03

Conditional number without regressor `Year`:3.29
Decrease in conditional number: 356.94
```

(App. 4)


Density plot of Land Value
Density plot of Land Improvements
Density plot of Consideration
Density
Normalized values
1e−6
1e6
1e7
Density plot of Land Value
Density plot of Land Improvements
Density plot of Consideration
Density
Normalized values

(App. 5)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.836
Model:                            OLS   Adj. R-squared:                  0.836
Method:                 Least Squares   F-statistic:                 2.959e+05
Date:                Sun, 30 Apr 2023   Prob (F-statistic):               0.00
Time:                        16:56:46   Log-Likelihood:                -61575.
No. Observations:              116106   AIC:                         1.232e+05
Df Residuals:                  116103   BIC:                         1.232e+05
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0029      0.001      2.370      0.018       0.000       0.005
x1             0.5447      0.002    362.557      0.000       0.542       0.548
x2             0.4903      0.001    330.051      0.000       0.487       0.493
==============================================================================
Omnibus:                    40698.761   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          7769026.670
Skew:                          -0.523   Prob(JB):                         0.00
Kurtosis:                      43.060   Cond. No.                         1.98
==============================================================================
```

Max_depth	2	3	4	6	**10**
R^2	0.6712	0.7734	0.8161	0.8476	**0.8503**

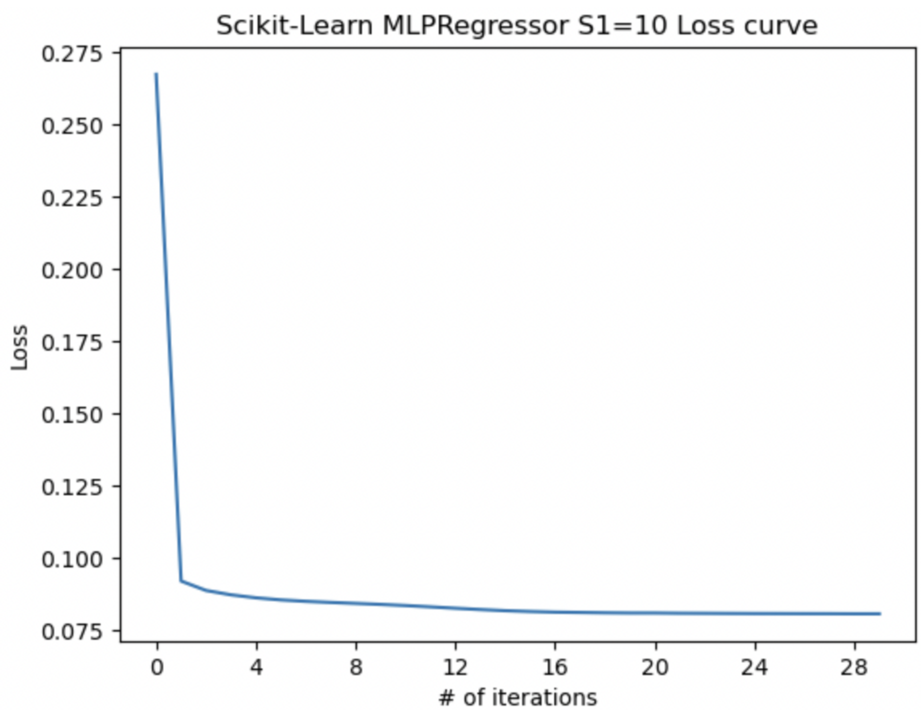
(App. 6)

# of Neurons	2	3	6	**10**	100
R^2	0.8377	0.8482	0.8495	**0.8504**	0.8469

(App. 7)

# of Neurons	2	3	6	**10**	100
R^2	0.8447	0.8502	0.8504	**0.8506**	0.8430

(App. 8)



```
Mean error_sq: 0.15
Upper confidence: 0.16
Lower confidence: 0.14
```

```
Mean error: 16379.23$
Upper confidence: 20373.81$
Lower confidence: 12384.65$
```