DATS 6202

Machine Learning I

Dr. Amir Jafari

Individual Final Report on

Maryland Property Data

by

Tyler Wallett

Introduction: 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 was 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 correspondingly.

Description of your individual work: Since I am the only member of my team, my work revolved around taking care of all of the areas of the project. From the pre-processing of the data to the interpretation of the model results, I attempted to devise a well-structured project from start to finish.

Describe the portion of the work that you did on the project in detail:

- Motivation: Developed the motivation behind the analysis by reading an online article titled *Getting ahead of the market: How big data is transforming real estate* by McKinsey & Company.
- Dataset & S.M.A.R.T. Question: Read once, twice and three times the data dictionary to inform myself of the features of the dataset. Performed the necessary and corresponding pre-processing of the dataset. And developed the S.M.A.R.T. question to establish a main goal for the project.
- EDA: Explored the dataset using subplots of various seaborn histogram plots, one geoplot by zip codes and a seaborn countplot of the grade structures.
- Feature Selection: Developed an appropriate correlation matrix and eigensystem analysis by conditional values. Researched on different methods and developed a Lasso Regression and Recursive Feature Elimination using a Random Forest Regression from the book *Machine Learning Pocket Reference* by Matt Harrison.
- Pre-processing: Transformed the selected dependent and independent variables to normal distributions by means of Scikit-Learn's QuantileTransformation method.
- Models: 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.
- Use Case: Developed a use case to properly assess the practicality behind the models approximation of consideration prices.
- Conclusions: Developed different conclusions from each part of the analysis, namely: EDA, Feature Selection, and the models.
- Limitations: Brainstormed some of the limitations and aspirations for the next project when it comes to the Maryland Property Dataset.

Results: 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.

Summary and conclusions:

- From the EDA: Distributions were highly-right skewed. On average
- From the Feature Selection: Best model would include all variables, but to avoid collinearity "Land value" and "Land improvements" should be used.
- From the Modeling: Scikit-learn MLPRegressor can predict approximately 85% of the variance in consideration.

Limitations:

- Lots of redundant Id's, geographical and boolean features.
- Lack of traditional features (e.g. amount of bedrooms, historical values)
- Lack of reproducibility.
- The definition of consideration.

Percentage copied from the internet: Lasso Regression and Recursive Feature Elimination were copied from the book *Machine Learning Pocket Reference* by Matt Harrison. ChatGPT was of help in creating the 'Average consideration prices by zip codes' visualization and the two animations for the multilayer perceptrons. So, exactly 700 lines of code were written 80 lines were written by ChatGPT (56 lines animation and 24 geopandas plot) and 34 lines were copied from the book (10 RFE & 24 Lasso Regression), about (114/700 * 100) 16.29% of the project was referenced.

References:

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:

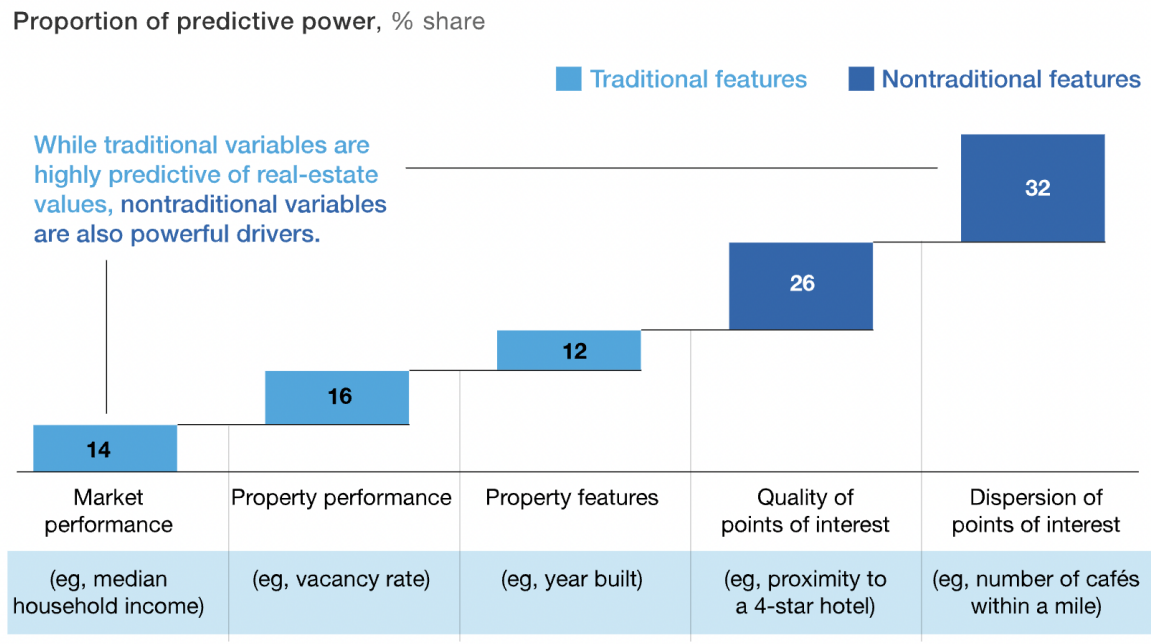
Harrison, Matt. *Machine Learning Pocket Reference: Working with Structured Data in Python*. O'Reilly Media, Inc., 2019.

Appendix:

(App. 1)

Exhibit 1

Nearly 60 percent of predictive power can come from nontraditional variables.



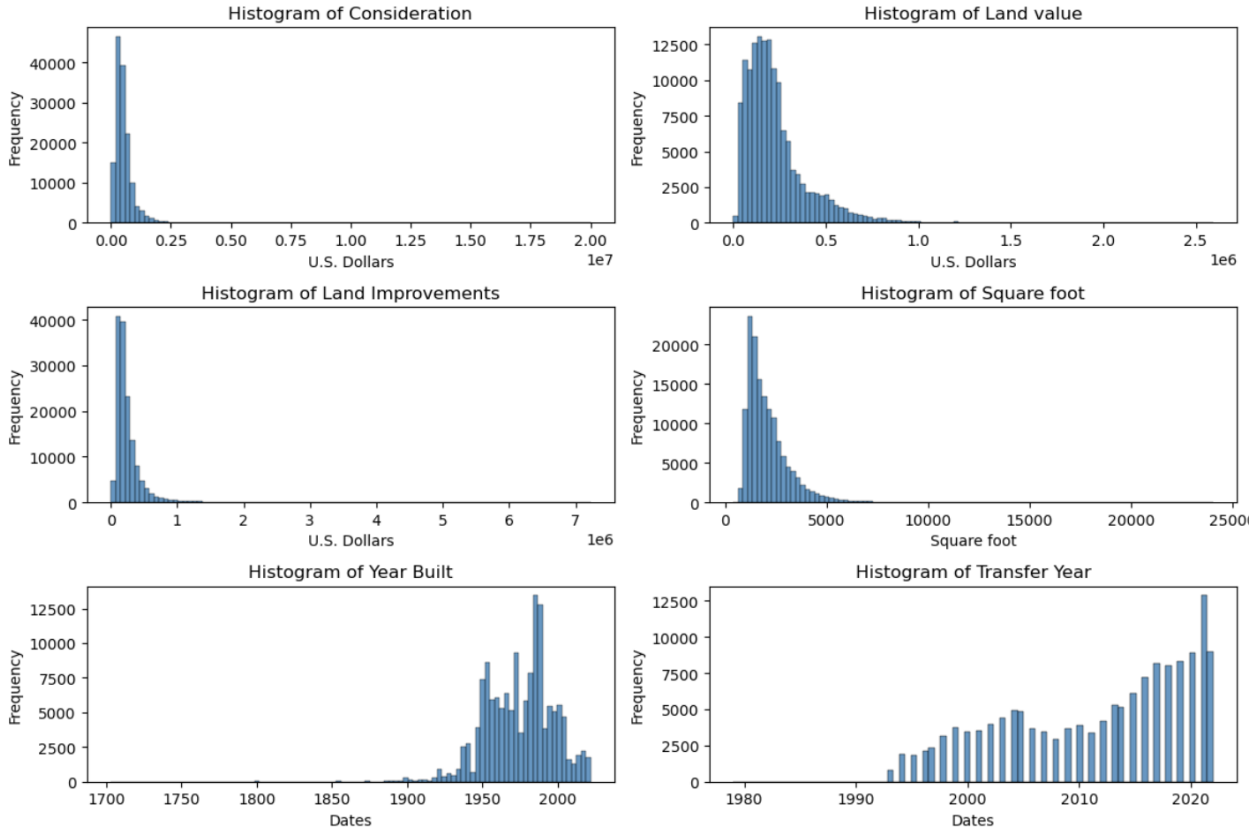
McKinsey&Company

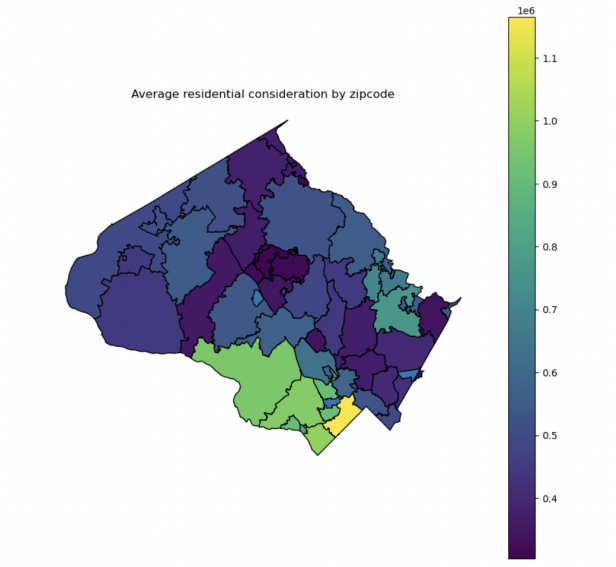
(App. 2)

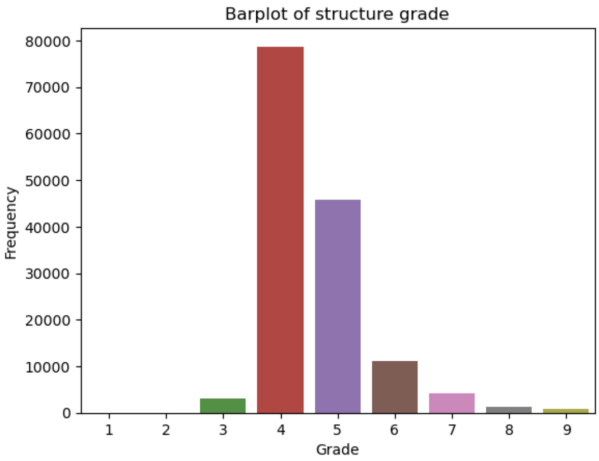
	Id	Longitude	Latitude	Address	Zipcode	Grade	Year_built	Sqft	Trade_date	Consideration	Land_value
588336	160100000033	-77.168495	39.207629	21411 WOODFIELD RD	20882.0	3.0	1936.0	1064.0	20190702.0	260000.0	231200.0
588338	160100000066	-77.166819	39.207889	8120 BRINK RD	20882.0	2.0	1923.0	864.0	20190815.0	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

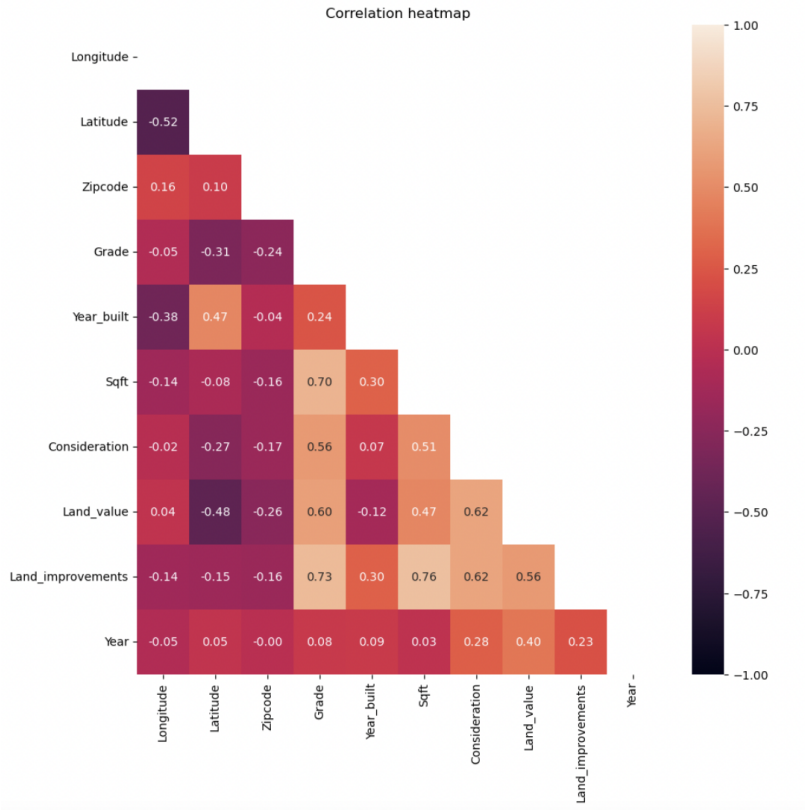
145133 rows × 11 columns

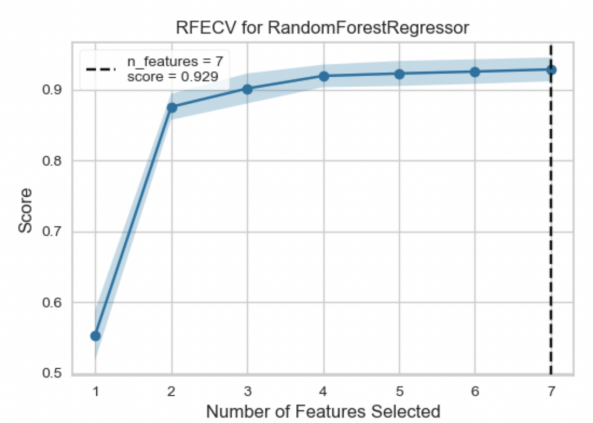
(App. 3)




Average residential consideration by zipcode
1e6
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4


Barplot of structure grade
Frequency
80000
70000
60000
50000
40000
30000
20000
10000
0
1
2
3
4
5
6
7
8
9
Grade


Correlation heatmap
Longitude
Latitude
Zipcode
Grade
Year_built
Sqft
Consideration
Land_value
Land_improvements
Year
-0.52
0.16 0.10
-0.05 -0.31 -0.24
-0.38 0.47 -0.04 0.24
-0.14 -0.08 -0.16 0.70 0.30
-0.02 -0.27 -0.17 0.56 0.07 0.51
0.04 -0.48 -0.26 0.60 -0.12 0.47 0.62
-0.14 -0.15 -0.16 0.73 0.30 0.76 0.62 0.56
-0.05 0.05 -0.00 0.08 0.09 0.03 0.28 0.40 0.23
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

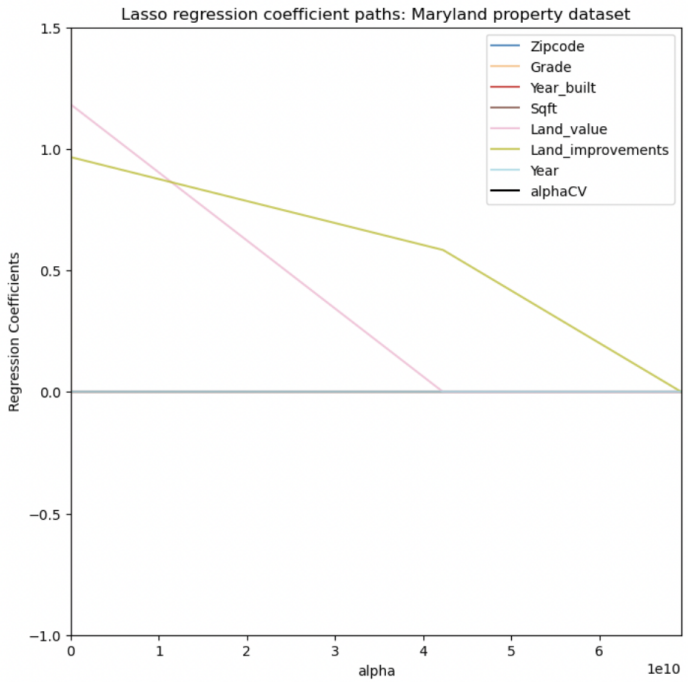



```
top_features
```
✓ 0.0s

```
Index(['Land_improvements', 'Land_value', 'Grade', 'Year'], dtype='o
```

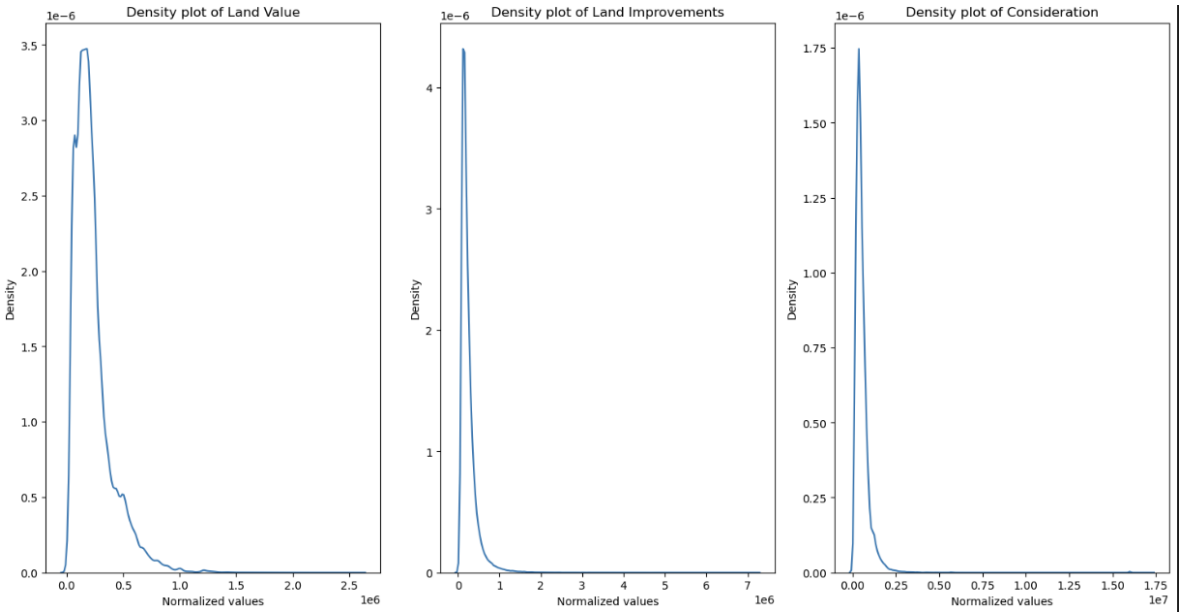
```
top_features
```
✓ 0.0s

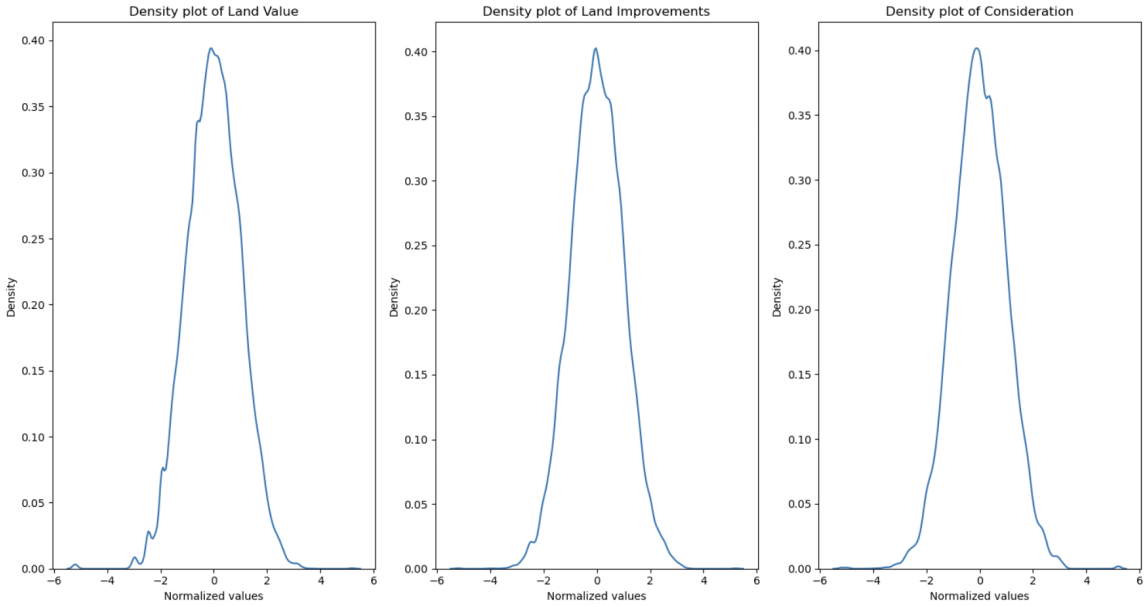
```
Index(['Land_improvements', 'Land_value'], dtype='object')
```



```
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
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−6
−4
−2
0
2
4
6

(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)

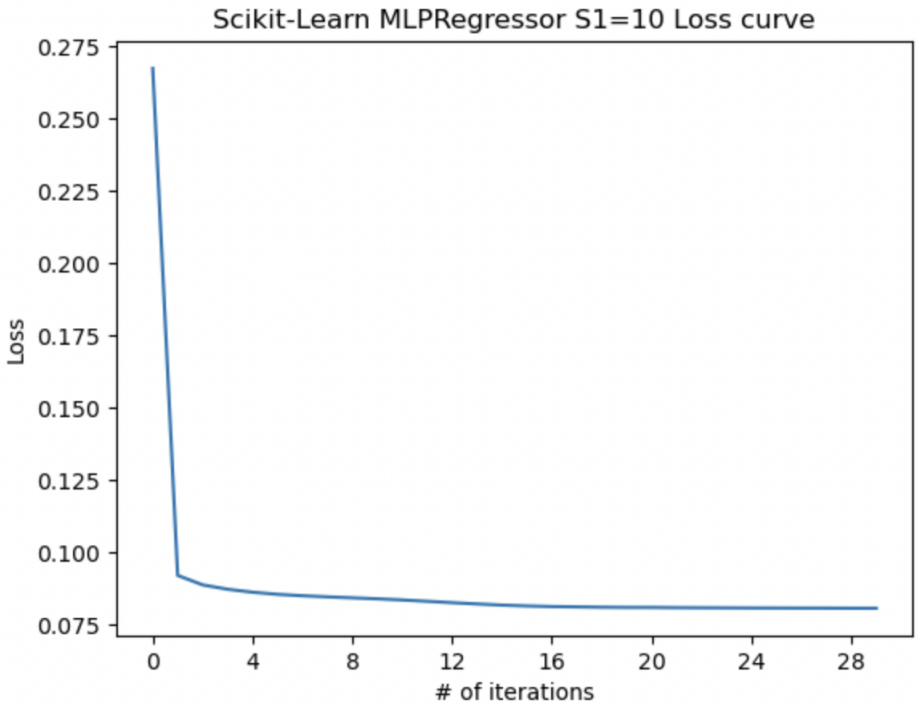
# of	2	3	6	**10**	100

Neurons					
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$
```