THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

ML I Final Term Project on *Maryland Property Data*

By Tyler Wallett

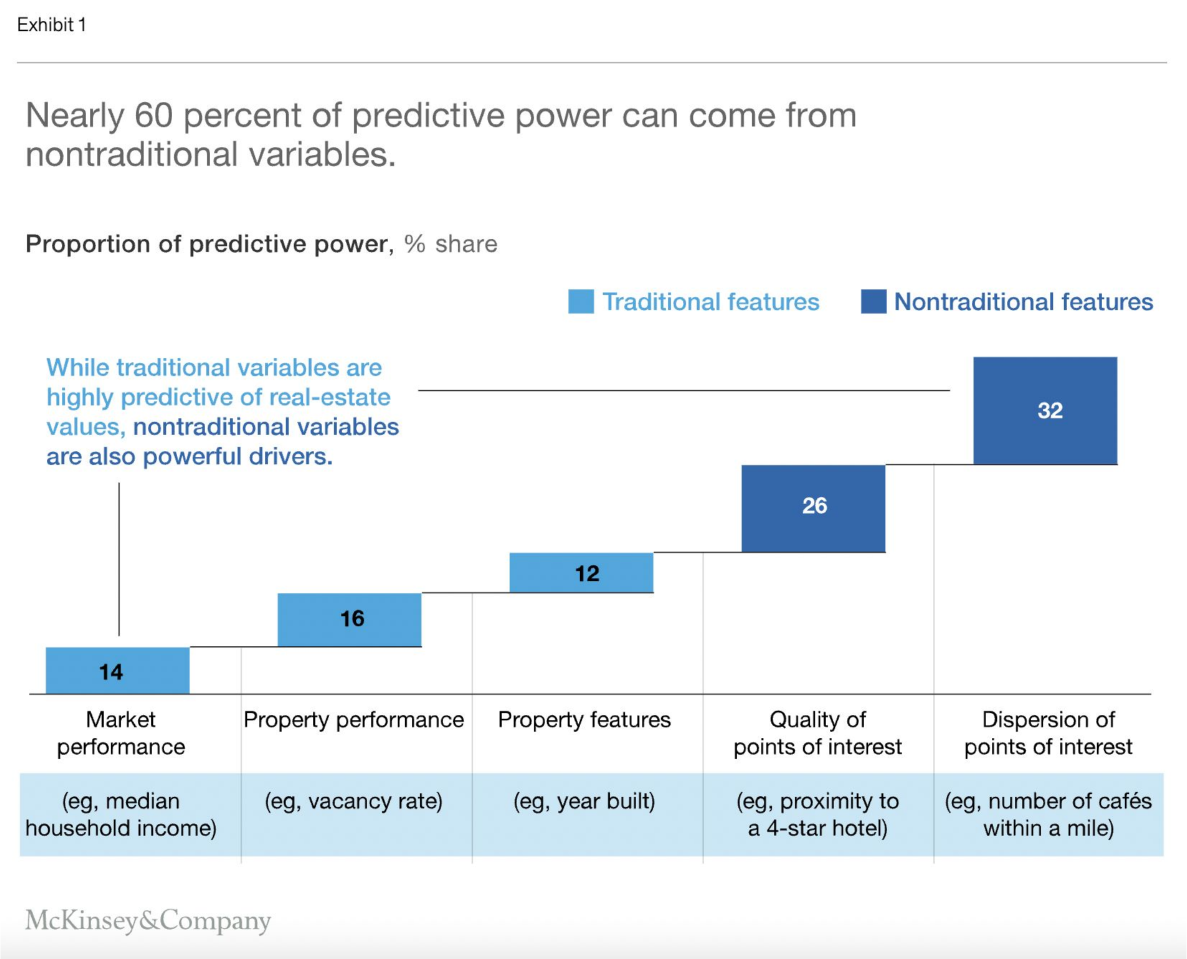
Outline:

1. Motivation
2. Dataset & S.M.A.R.T. Question
3. EDA
4. Feature Selection
5. Pre-processing
6. Models
7. Use case
8. Conclusion
9. Limitations
10. References

Motivation

Motivation

“Many real estate firms have long made decisions based on a combination of **intuition and traditional, retrospective data**. Today, **a host of new variables** make it possible to paint more vivid pictures of a location’s future risks and opportunities.” (McKinsey)



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

Dataset

- Public data is gathered from the Maryland State Department of Assessments and Taxation (SDAT)
- Approximately 2.4 million unique property parcels
- 134 features describing each property parcel (64 numeric - 70 categorical)
- Last updated on January 7, 2023

Maryland
DEPARTMENT OF ASSESSMENTS AND TAXATION

Dataset

After setting ‘County’ Montgomery County and ‘Land Usage’ as Residential, and removing:

- Unnecessary Categorical features
- Redundant location features
- Redundant ID features
- Redundant Boolean features

Dataset

Dependent Var.

Id	Longitude	Latitude	Address	Zipcode	Grade	Year_built	Sqft	Trade_date	Consideration	Land_value	Land_improvements	Year	Month	Day	transfer_date	geometry
160100000033	-77.168495	39.207629	21411 WOODFIELD RD	20882.0	3.0	1936.0	1064.0	20190702.0	260000.0	231200.0	76200.0	2019	07	02	2019-07-02	POINT (-77.16850 39.20763)
160100000066	-77.166819	39.207889	8120 BRINK RD	20882.0	2.0	1923.0	864.0	20190815.0	100000.0	174700.0	14500.0	2019	08	15	2019-08-15	POINT (-77.16682 39.20789)
160100000113	-77.178153	39.198018	8615 LOCHAVEN DR	20882.0	5.0	1840.0	2968.0	19931220.0	355000.0	120970.0	157880.0	1993	12	20	1993-12-20	POINT (-77.17815 39.19802)
160100000124	-77.141826	39.200285	6934 WARFIELD RD	20882.0	4.0	1978.0	1896.0	20180509.0	475000.0	244800.0	50500.0	2018	05	09	2018-05-09	POINT (-77.14183 39.20029)
160100000204	-77.122566	39.257719	24501 HIPSLEY MILL RD	20882.0	4.0	1913.0	2552.0	20170516.0	525000.0	201500.0	267700.0	2017	05	16	2017-05-16	POINT (-77.12257 39.25772)

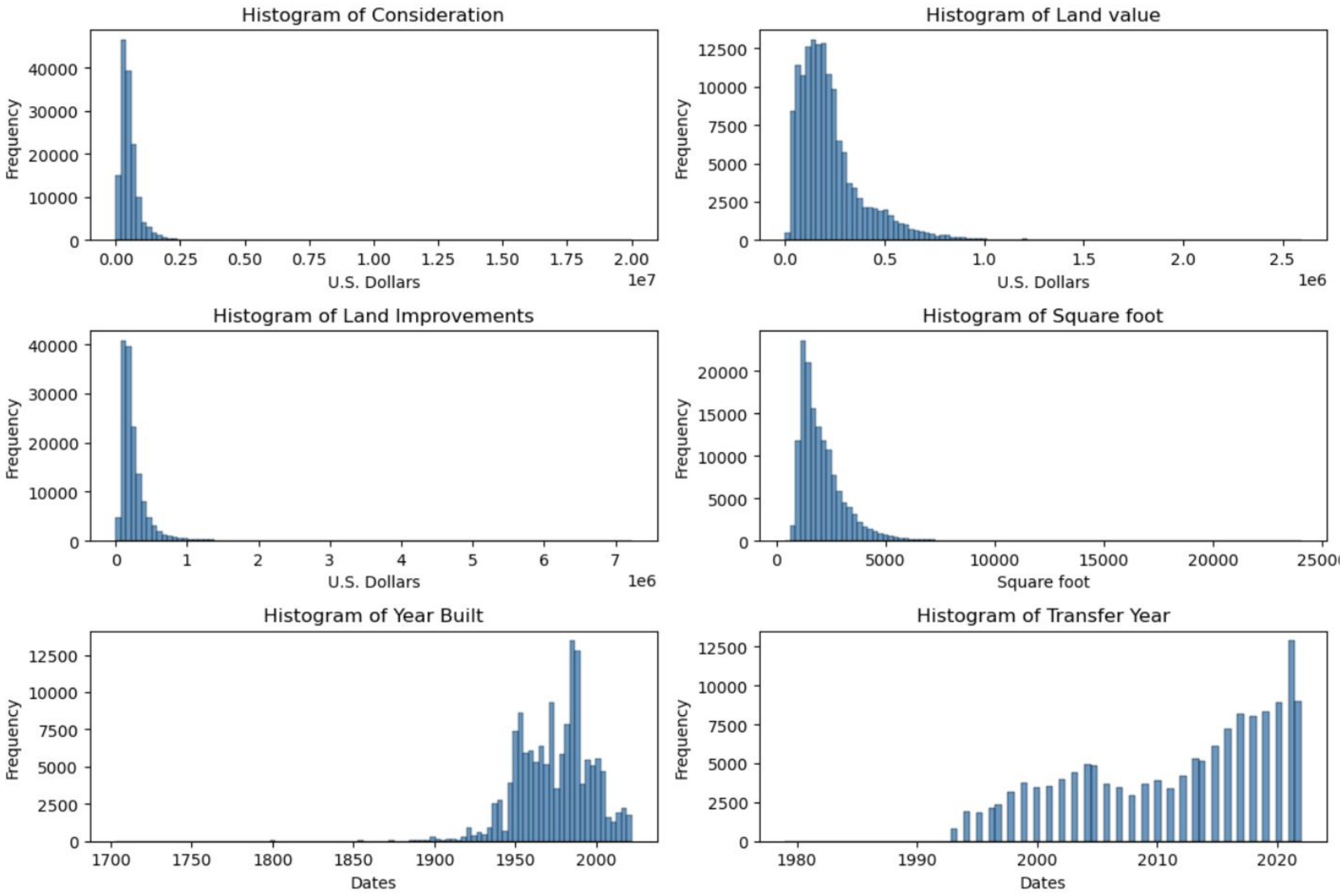
145133 rows × 17 columns

S.M.A.R.T. Question

*Can we successfully approximate consideration prices of residential properties from Montgomery County, Maryland **only using traditional features**?*

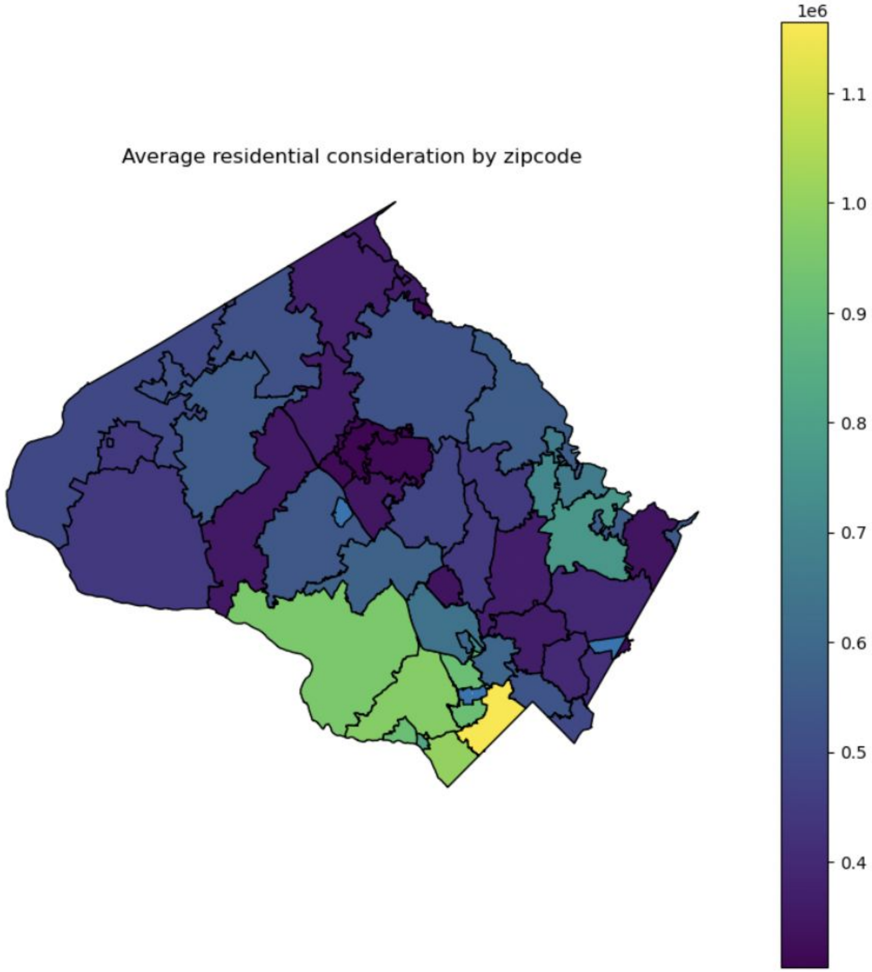
Exploratory Data Analysis (EDA)

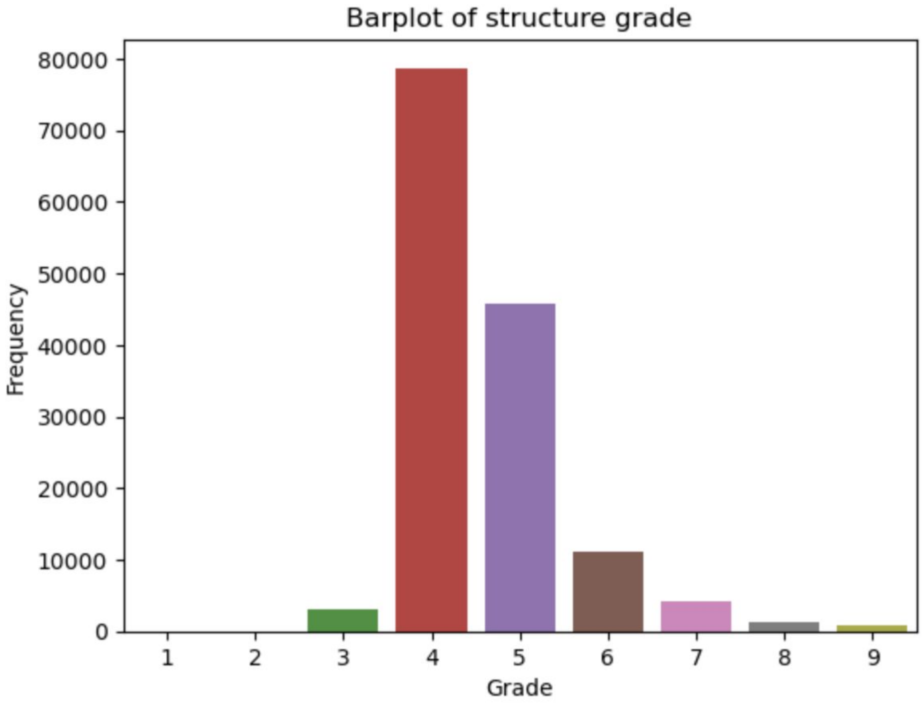
EDA



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

EDA





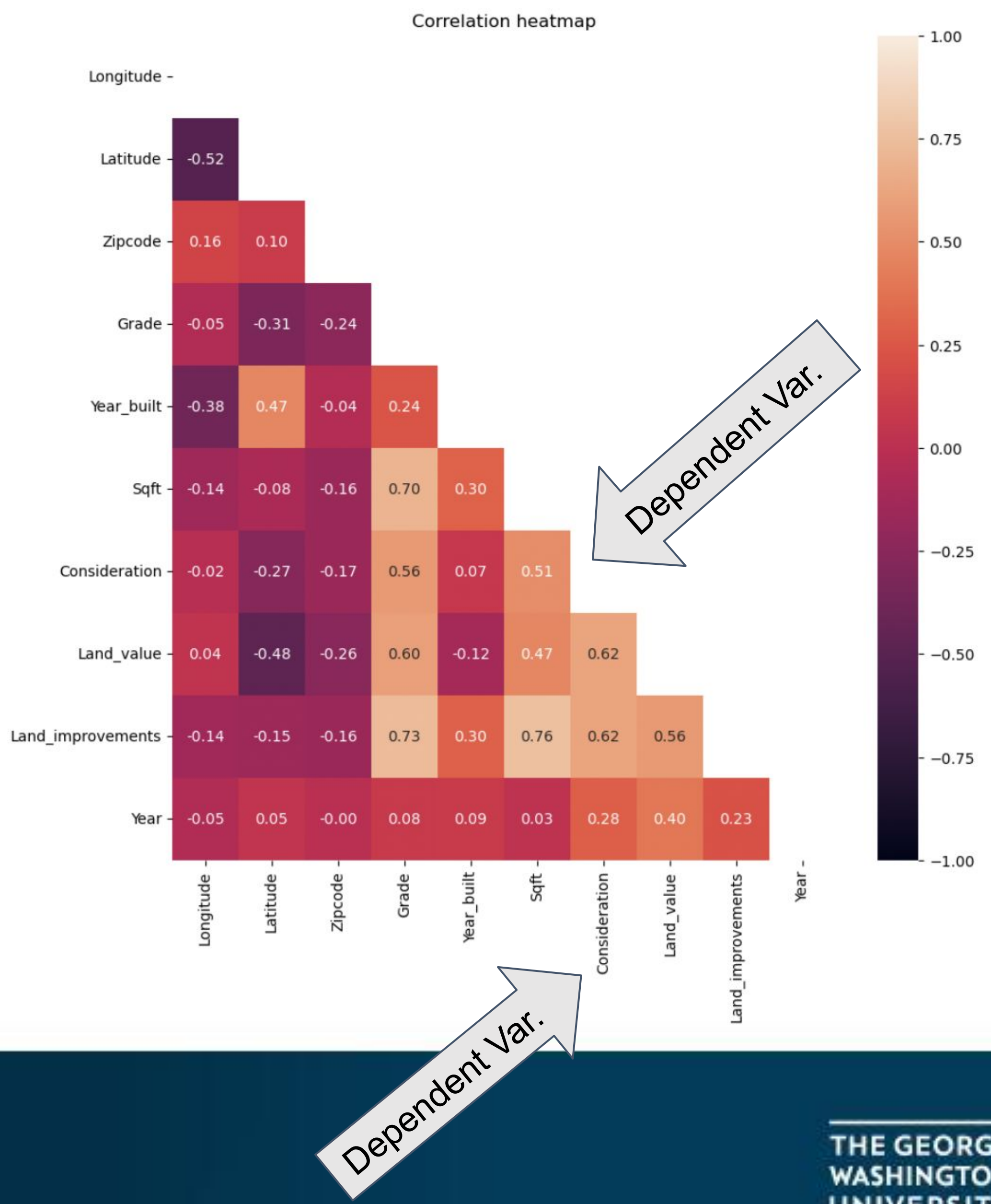
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Feature Selection

Correlation Heatmap

High positive correlations with:

- Land value
- Land improvements
- Grade
- Sqft



THE GEORGE WASHINGTON UNIVERSITY

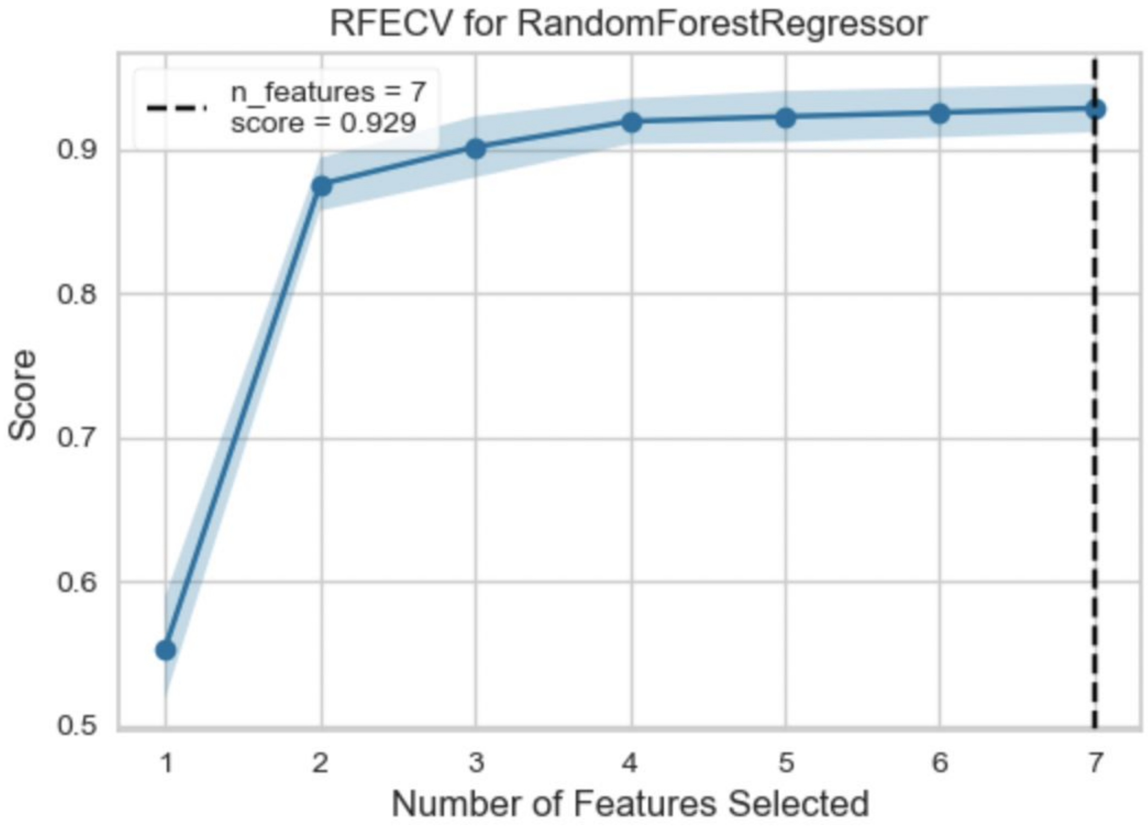
WASHINGTON, DC

RFE Random Forest

- Number_estimations = 100
- Cross validation = 5

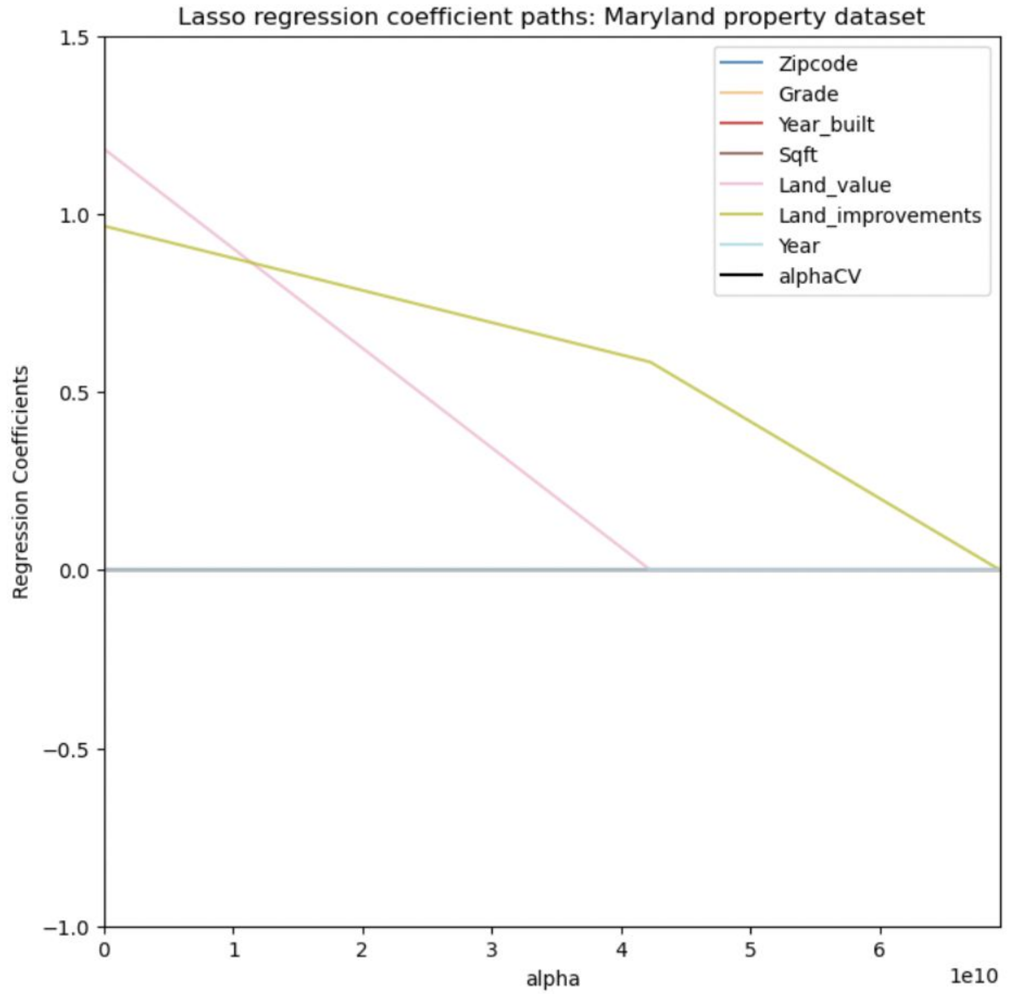
```
top_features
✓ 0.0s
Index(['Land_improvements', 'Land_value', 'Grade', 'Year'], dtype='o
```

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



Lasso Regression

- When added a regularization penalty to each coefficient, the last to converge to zero are “Land Value” and “Land Improvements”.



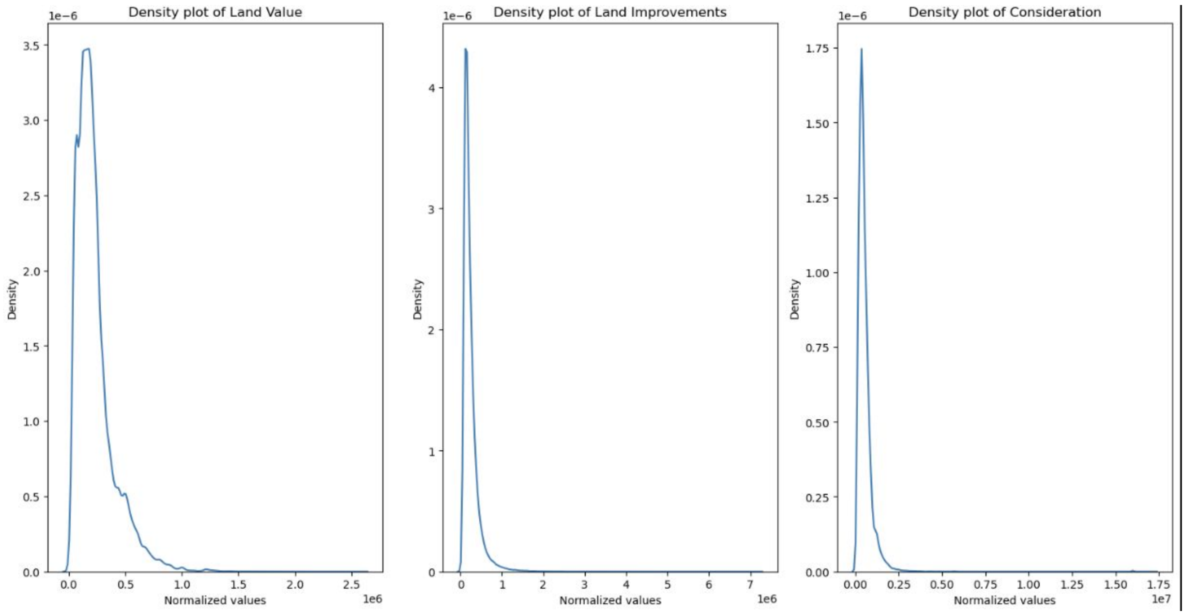
Conditional Values

- The degree of collinearity is very small when only “Land Value” and “Land Improvements” are considered.

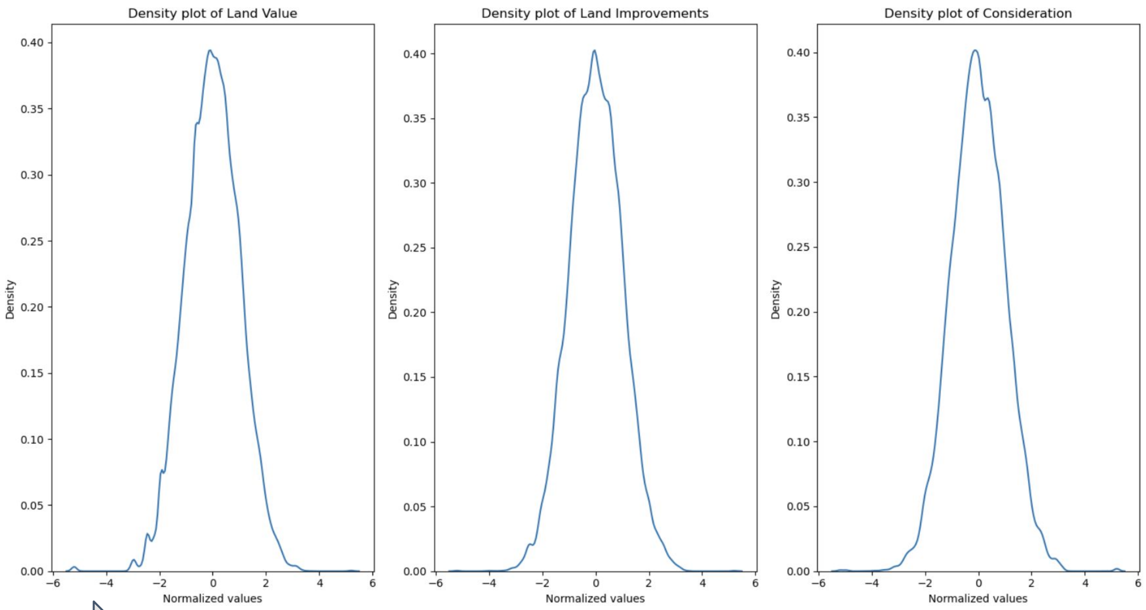
```
Initial conditional number: 765411.58

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
```

Pre-processing

Scikit Learn - Pre-processing



QuantileTransformation(“normal”)



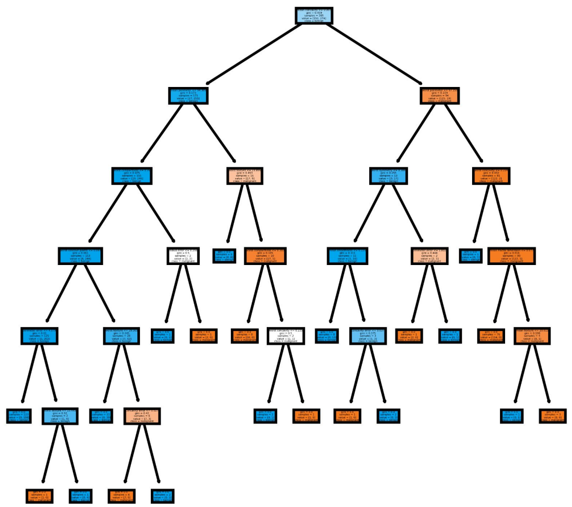
Models

Statsmodels - OLS Benchmark

```
                            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
==============================================================================
```

Random Forest Regression Benchmark

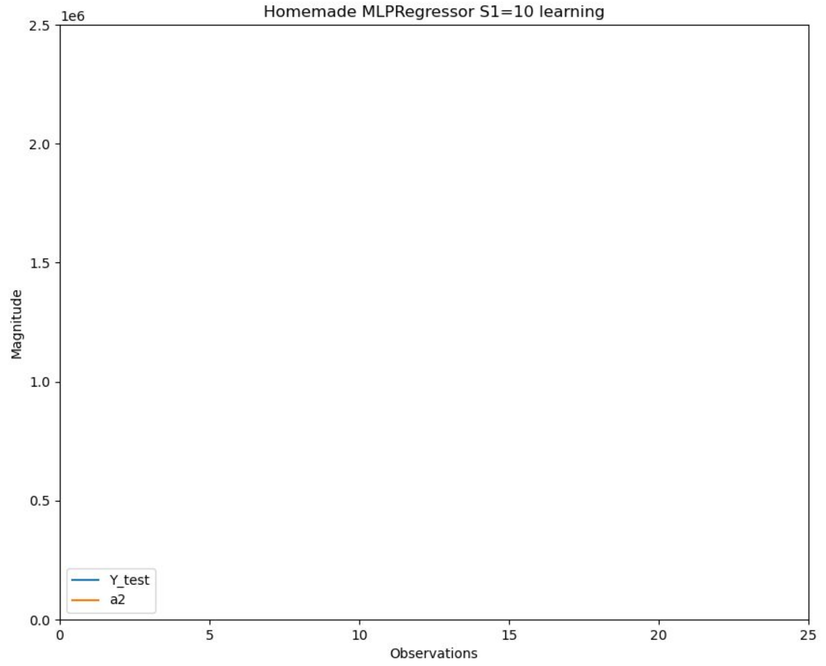
- Max_depth: The maximum depth of the tree.



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

“Homemade” MLPRegressor

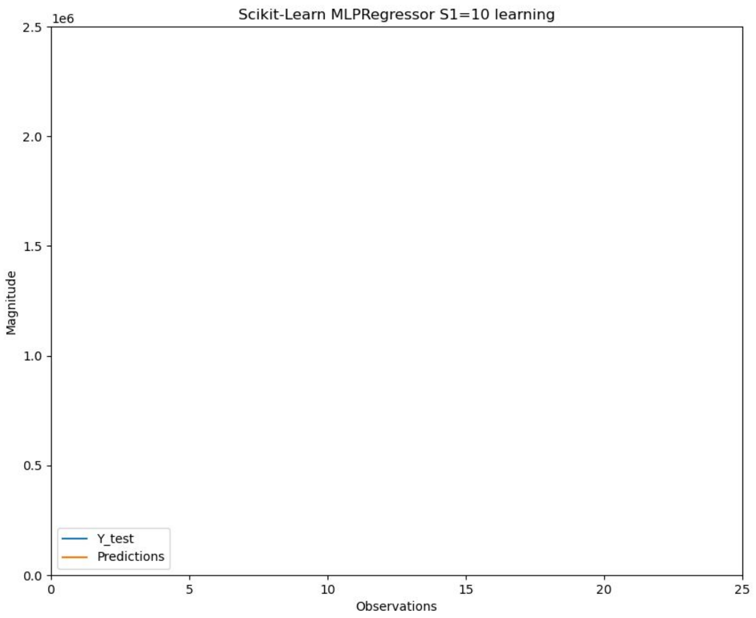
- 2-S1-1
- Epochs = 10
- α = 0.001
- n1 => logsig()
- n2 => purelin()



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

Scikit Learn - MLPRegressor

- 2-S1-1
- Epochs = 200 (default)
- α = 0.0001 (default)
- n1 => logistic()

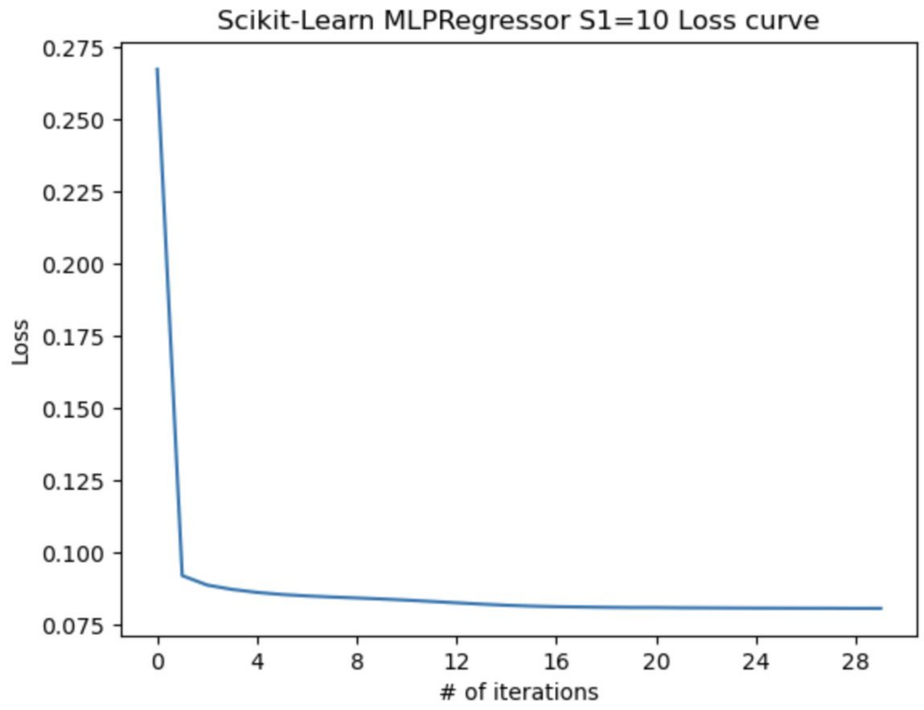


# of Neurons	2	3	6	**10**	100
R^2	0.8447	0.8502	0.8504	**0.8506**	0.8430

Scikit Learn - MLPRegressor 2-10-1

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

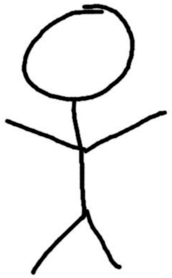
```
Mean error: 16379.23$
Upper confidence: 20373.81$
Lower confidence: 12384.65$
```



Use case

Use case

Homeowner is looking to get an estimate of their consideration for their property.



Transfer year 2015:

- Land value: 563,000$
- Land improvements: 183,900$

Approximation of consideration

825,000$

UCL 845,000$

LCL 805,000$

Conclusions

Conclusions

- From the **EDA**: Distributions are highly-right skewed. On average consideration is higher when residences are close to Washington D.C.
- From the **Feature Selection**: Best model would include all variables, but to avoid collinearity “Land value” and “Land improvements” should be used.

Conclusions

*Can we successfully approximate consideration prices of residential properties from Montgomery County, Maryland **only using traditional features**?*

Yes, by **only using traditional features** with a Scikit-learn MLPRegressor we can approximate 85% of the variance in consideration.

Limitations

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.

As this project comes to a close,
Our skills and knowledge have surely grown,
From data wrangling to model selection,
We've tackled challenges with conviction.

Through long hours and sleepless nights,
We've worked to make our results just right,
And now we stand at the end,
With new skills and insights to extend.

So let's celebrate this final slide,
And the knowledge that we'll carry with pride,
For our journey may end here today,
But our passion for learning will forever stay.

Thank you!

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.