Predicting Pittsburgh Single Family Home Prices

Cole Thomas (nhthomas)
Sai Rajuladevi (srajulad)
Kraig Sheetz (ksheetz)
Hannah Fairfield (hfairfie)

Problem Statement

- Given current real estate market difficulties, use **non-conventional data** to predict home prices

- Input -> Output

- Criteria for Success

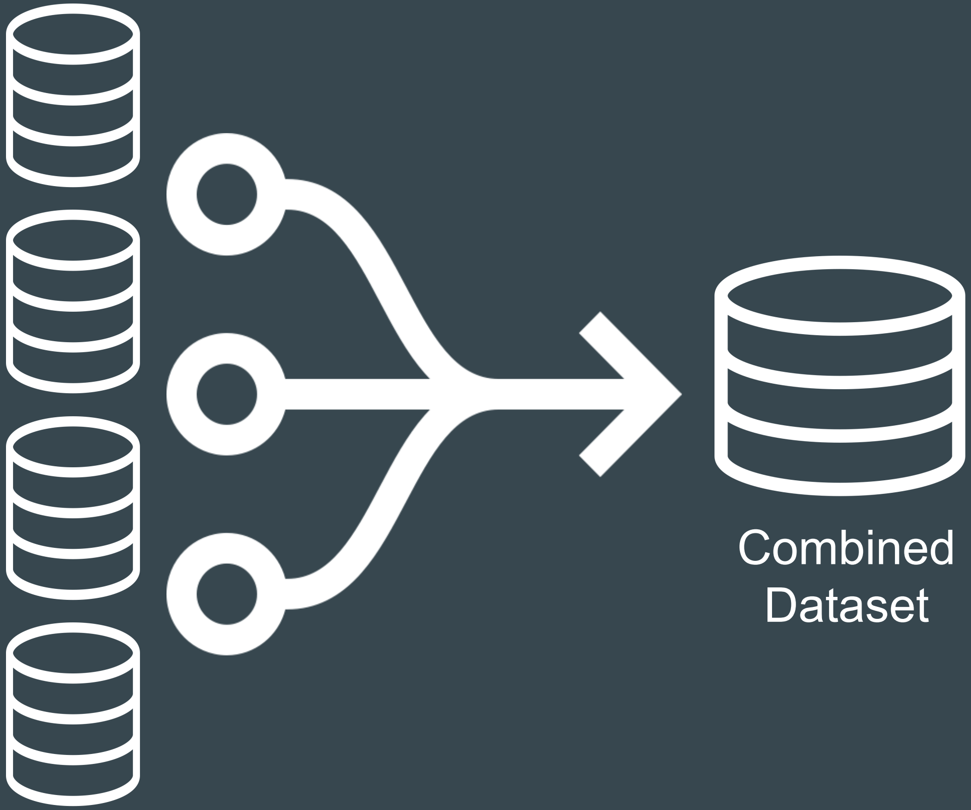
External Data Sources

Home Data

Crime Stats

School Stats

ACS Survey



Features

Redfin

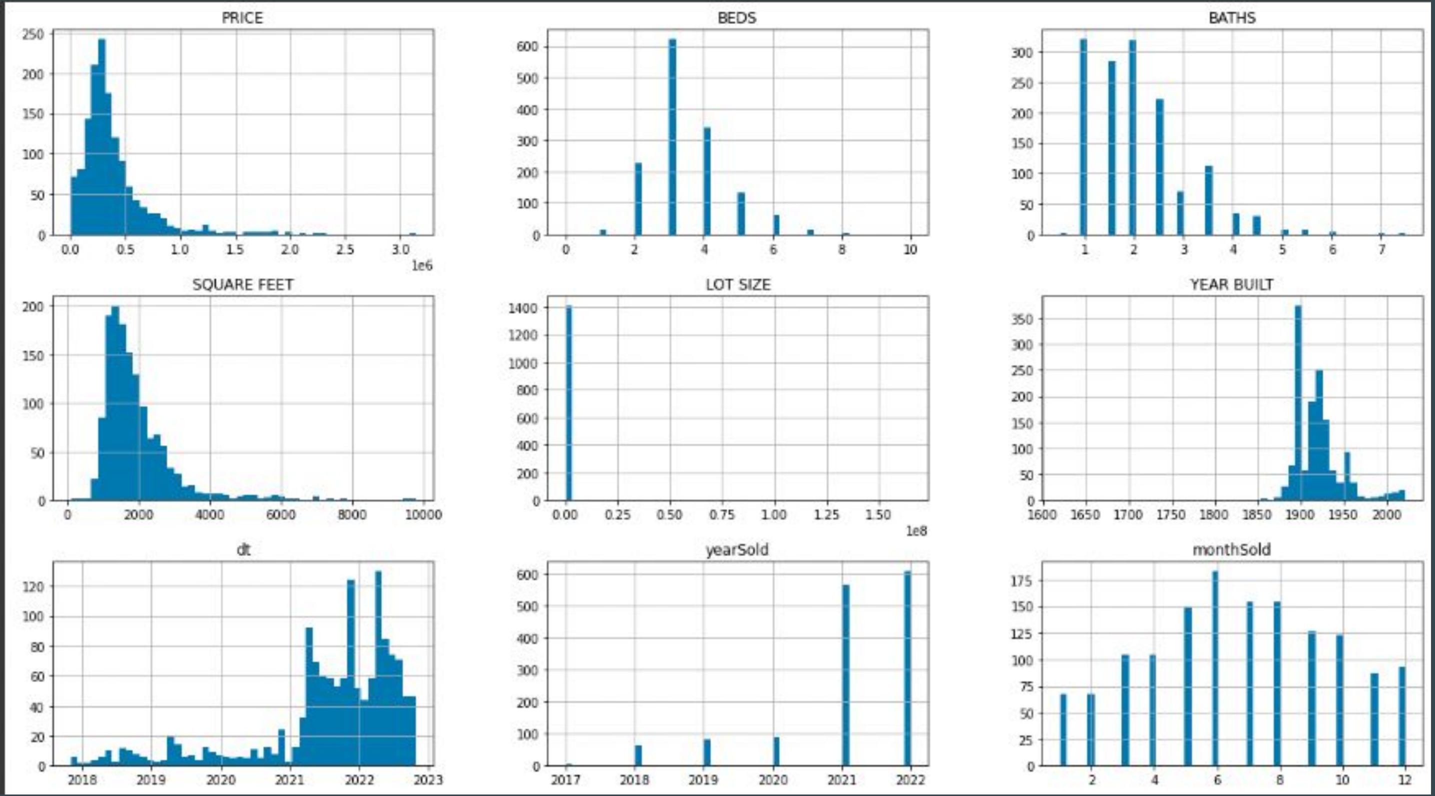
Redfin Dataset Description

- 1989 x 29

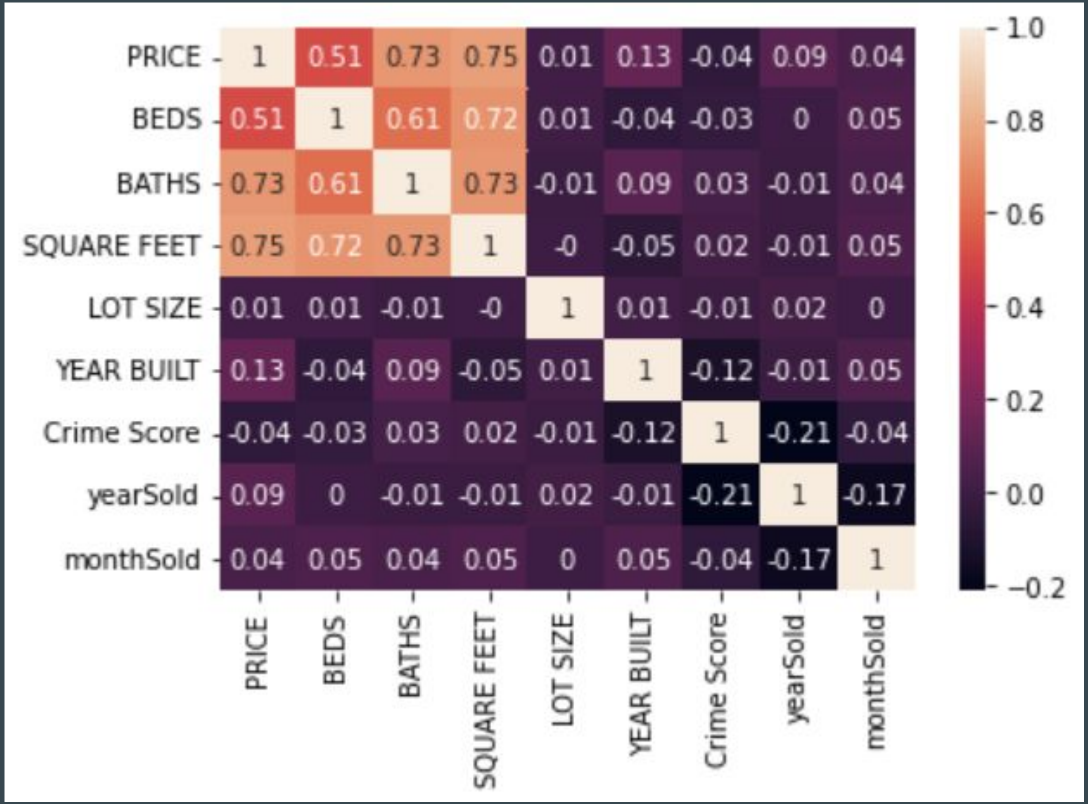
- Categorical:
  - Nominal: 13
  - Ordinal: 2

- Numerical:
  - Discrete: 4
  - Continuous: 10

Redfin Dataset Features



Redfin Data Exploration



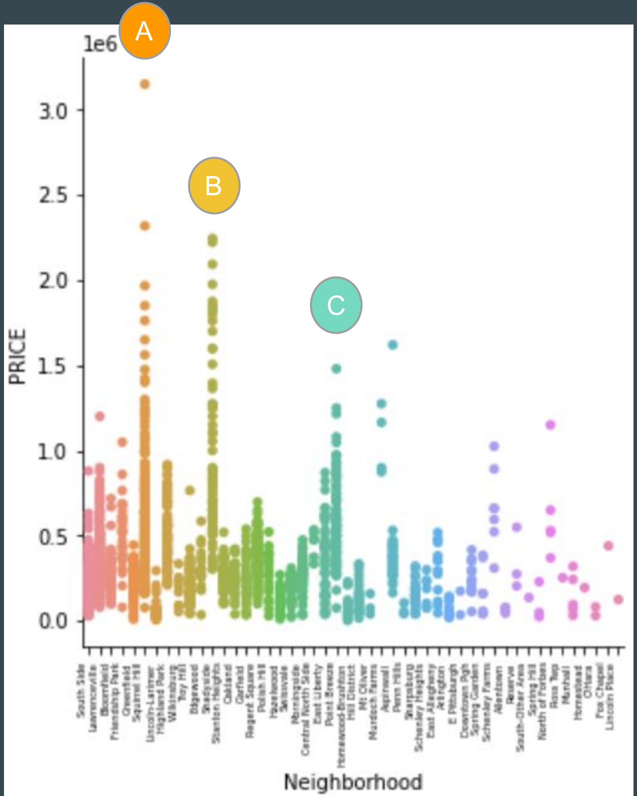
Baseline Model

Baseline Model – Features

Beds	*Price*	Year Sold
Square Feet	Bath	Month Sold
Neighborhood	Lot Size	One-Hot Encode Neighborhood
Location	Year Built	

1989 x 55

- A Squirrel Hill
- B Shadyside
- C Point Breeze



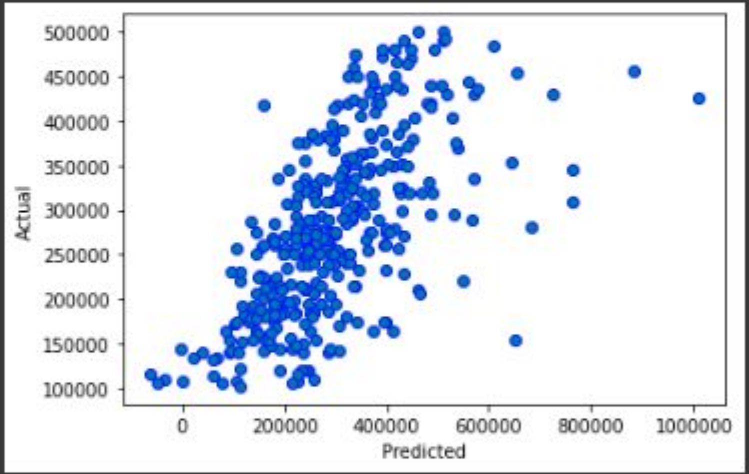
Baseline Model – Preliminary Results

- Linear Regression

- Split 75% train (1491) 25% test (498)

```
Redfin Linear Regression  results  Test

RMSE:  137,752.83
MAE:   93,381.03
R^2:   0.74
```



The Approach – Search for New Features

Redfin

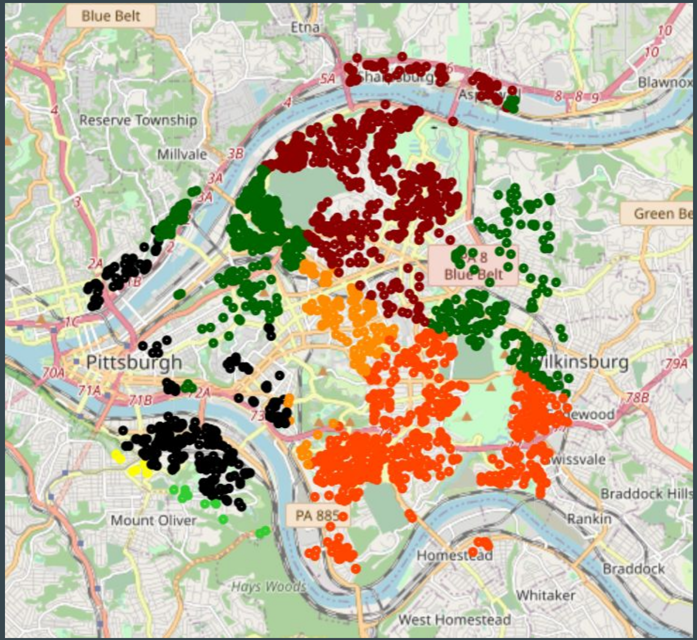
Local schools

New Features - Local Schools

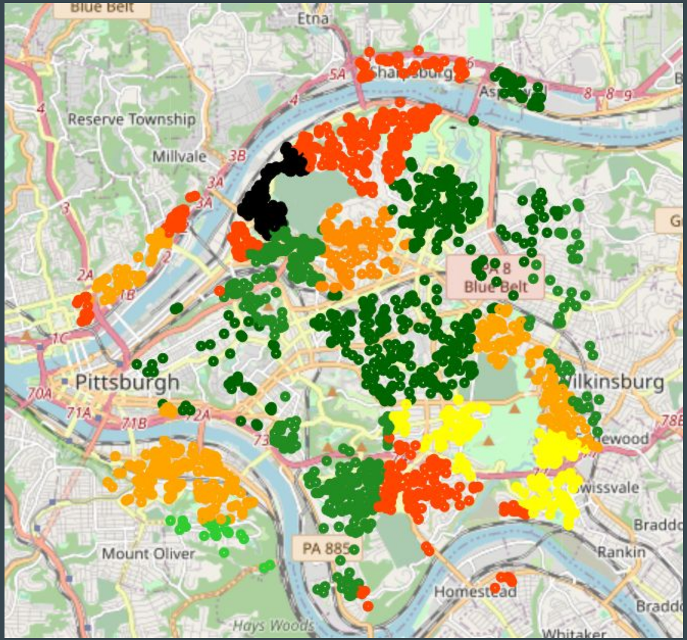
NCES National Center for Education Statistics

New Features - Local Schools

High Schools



Elementary/Middle Schools



Score
90-100
80-90
70-80
60-70
50-60
40-50
30-40
20-30
10-20
0-10

Features

Redfin

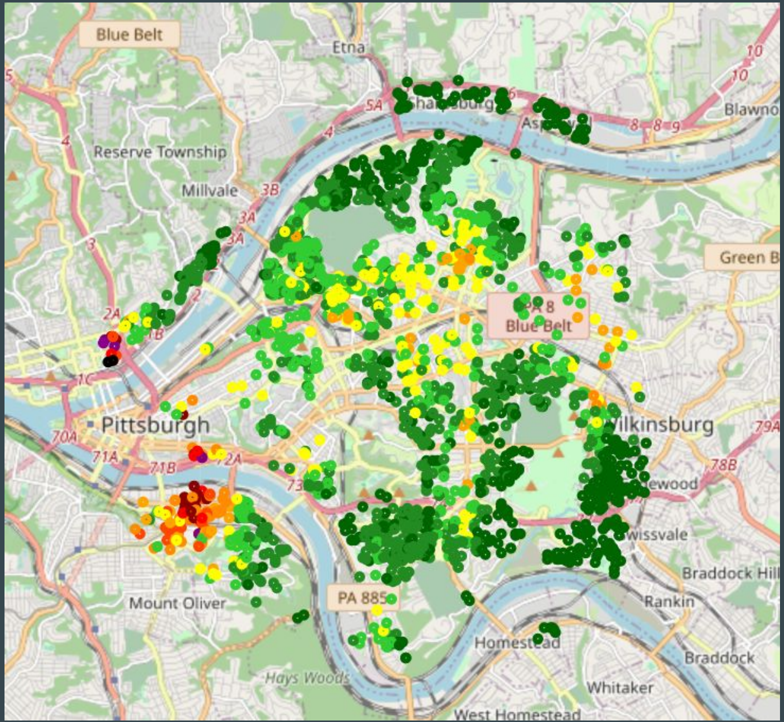
Local schools

Crime Data

New Features - Crime

WESTERN PENNSYLVANIA REGIONAL DATA CENTER

New Features - Crime



Score = Max(0, 1 - (SQRT(Distance/Max Radius)))

Score

0-10	
10-20	
20-30	
30-40	
40-50	
50-60	
60-70	
70-80	
80-90	
90-100	

Features

Redfin

Local schools

Crime Data

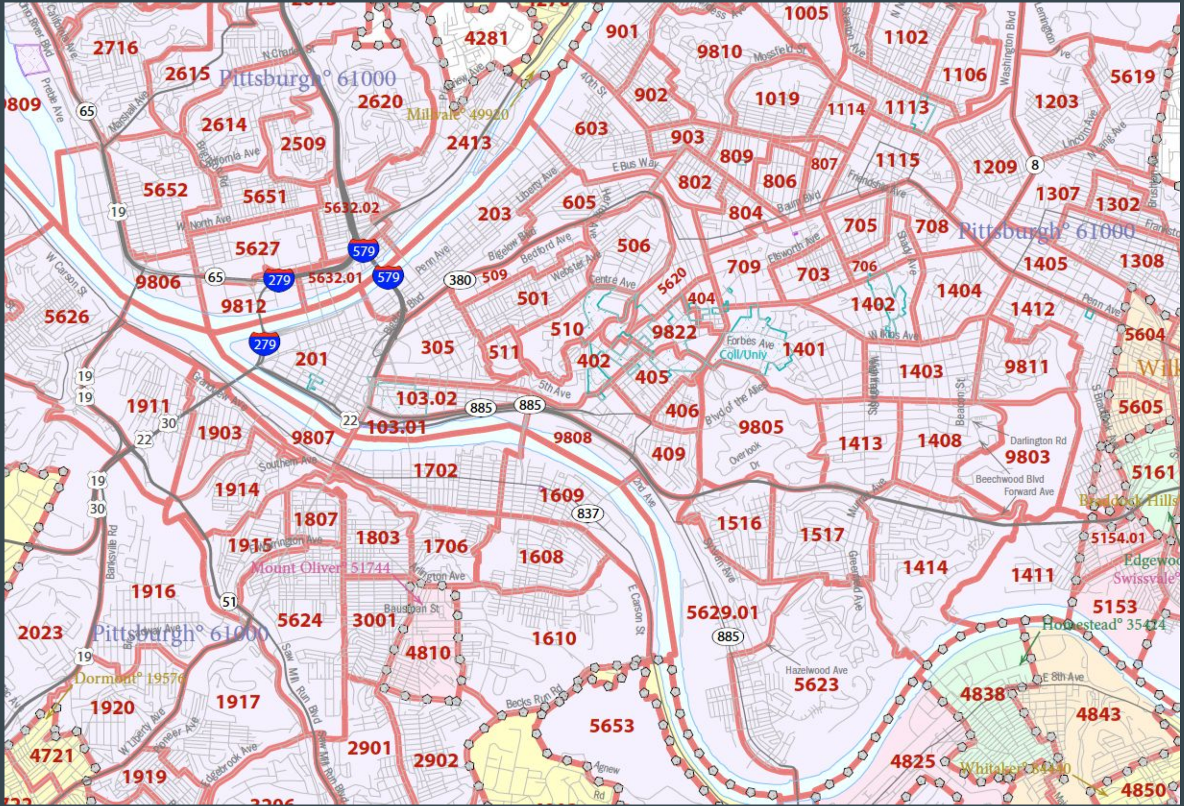
Census data

New Features - Census Data

American Community Survey

Detailed 5 Year Estimates

What is a Census Tract?



Source -> https -> //www2.census.gov/geo/maps/DC2020/PL20/st42_pa/censustract_maps/c42003_allegheny/DC20CT_C42003.pdf

Example Census Features

B25109_001E
housing_OwnerOccupiedMedianValue

B25111_001E
renting_MedianRentValue

B15012_009E
bachelors_STEM

B19001_017E
income_200KOrMore

B15003_025
Eeducation_DoctorateDegree

……. And more

Data Preparation

Data Preparation and Feature Engineering

R	S	T	U
joinKey	age_Median	housing_OwnerOccupiedMedianValue	renting_MedianRentValue
42003090100	33.2	250000	1470
42003110600	42.7	322200	935
42003562300	54.4	79400	664
42003424000	43.1	78300	820
42003516200	37.5	216500	903
42003090100	33.2	250000	1470
42003140300	40.3	450800	1682
42003141300	33.2	275800	1116
42003140500	35.1	-666666666	1340
42003130700	36.6	43800	574
42003090200	37.1	250800	981

Remove Outliers: Census

-666666666 means that the estimate could not be computed because there were an insufficient number of sample observations

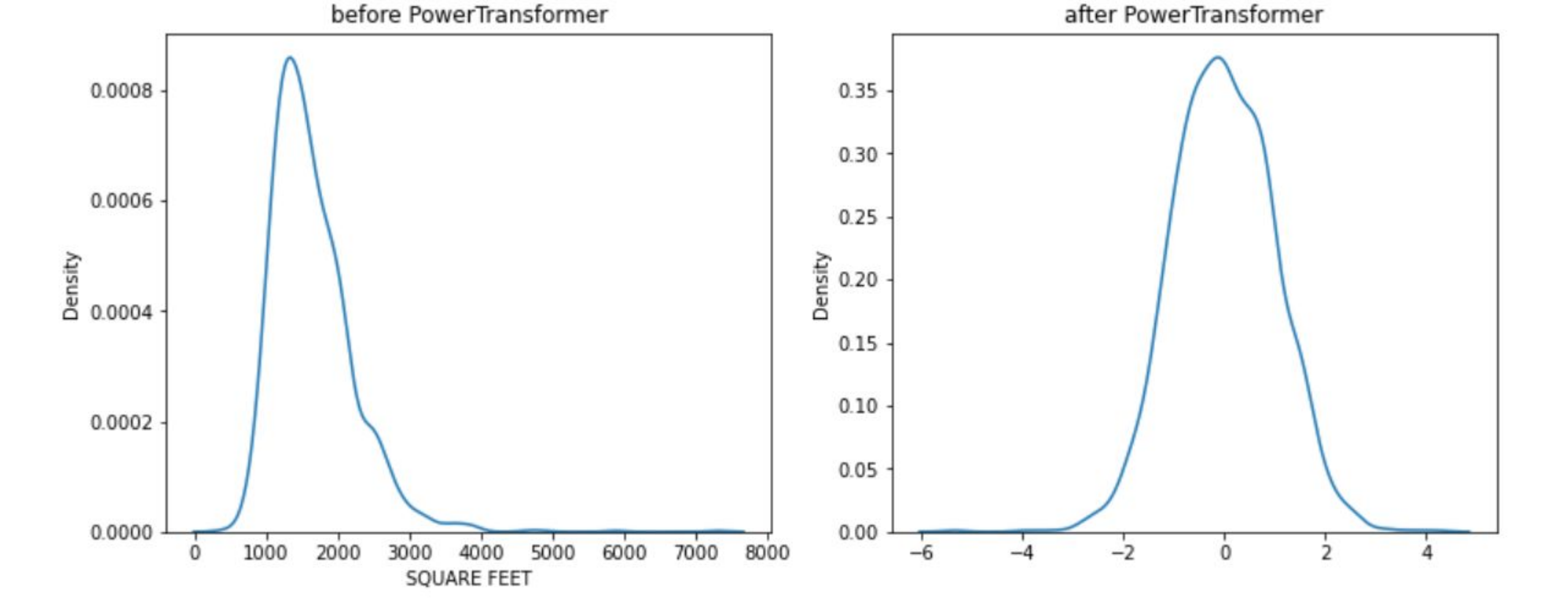
% STEM Bachelor's Degrees =

# of People with STEM Degrees

÷

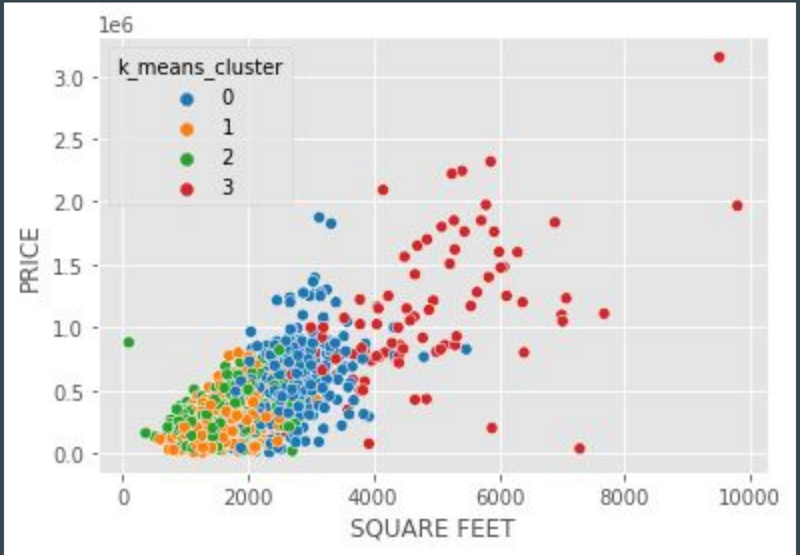
Total Bachelors Degrees

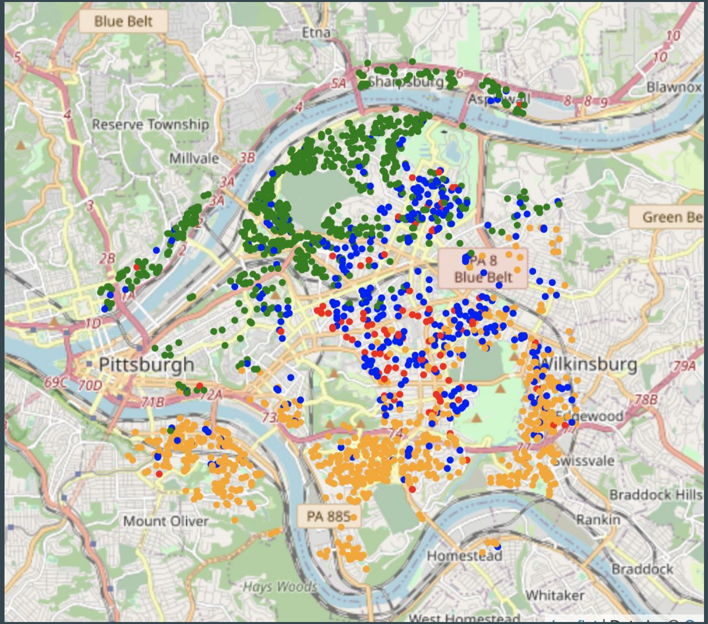
Data Preparation - Transform and Standardize



Modeling - ML

K-Means Experiment– Unsupervised EDA





Added : Latitude and Longitude Vars

Modeling Approach – Supervised

Regression Models with
GridSearchCV

AutoML tool developed by
Amazon

Target -> **Property Prices**

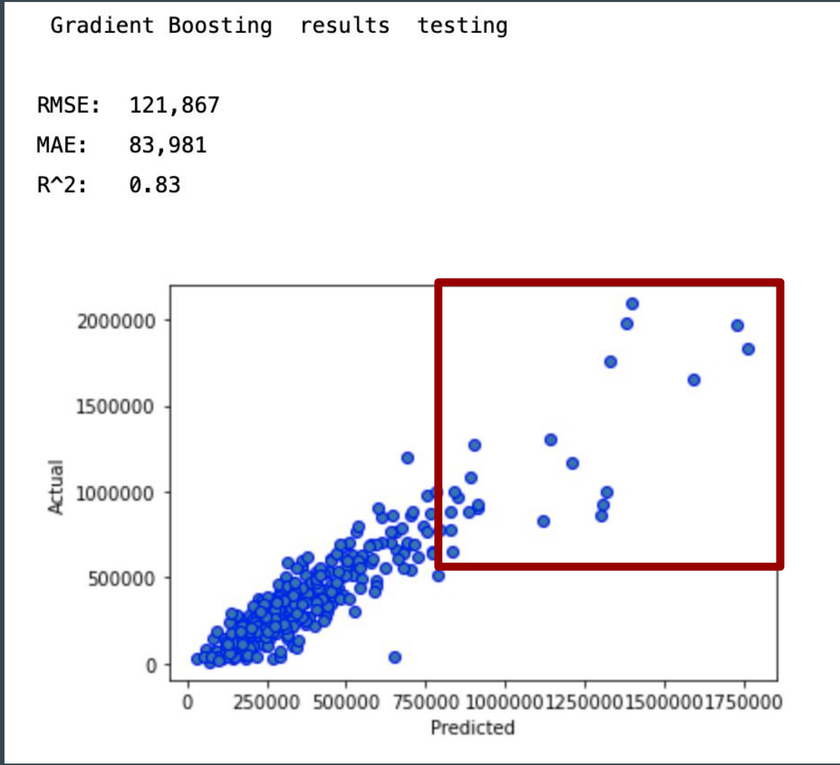
Modeling Approach

Models Evaluated:

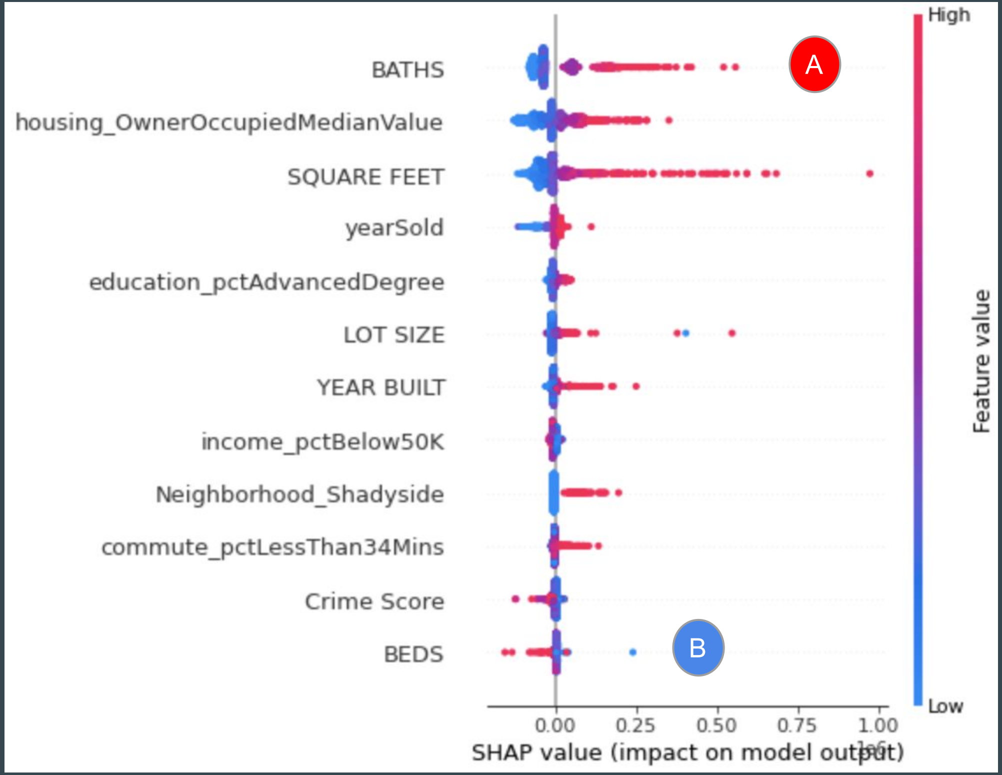
Lasso
Ridge
Elastic Net
Kernel Ridge
Bayesian Ridge
Decision Tree
Random Forest
Gradient Boosting
Multilayer Perceptron
Stochastic Gradient Descent

Scikit Learn Champion Model

- Gradient Boosting Champion Model
- MAE and RMSE
- High leverage testing points



Scikit Learn Champion Model



A BATHS is most important feature

B High BEDS -> lowers home price

Modeling Approach

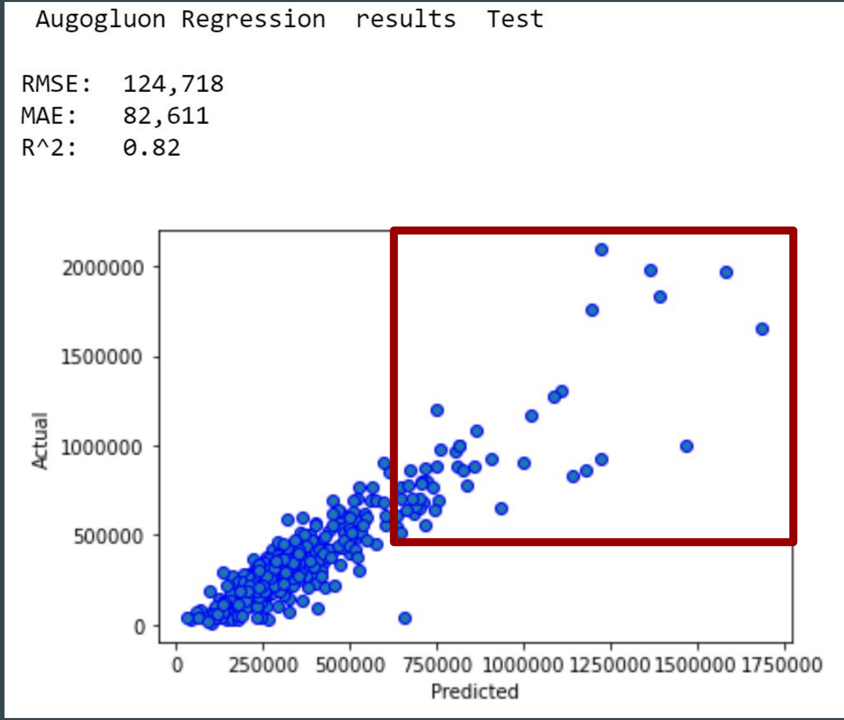
AutoGluon

Models Evaluated:

XGBoost
CatBoost
Extra Trees
Light GBM
K Neighbors
Random Forest
Neural Net Fast AI
Light GBM Xtreme
Weighted Ensemble

AutoGluon Champion Model (Weighted Ensemble)

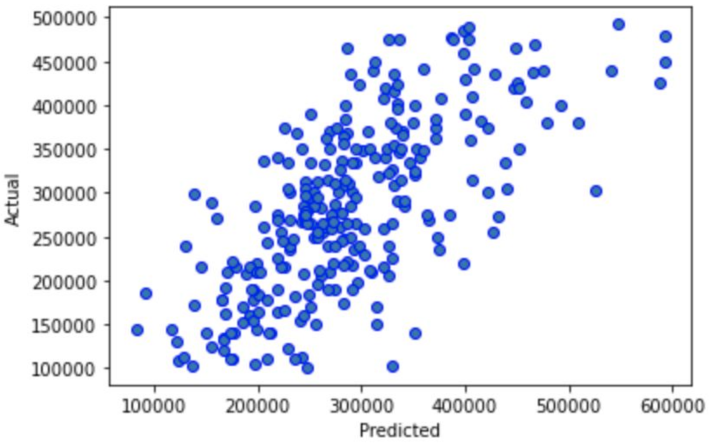
- Large residuals for properties priced over $500,000



Limiting Results to Homes to $100K - $500K

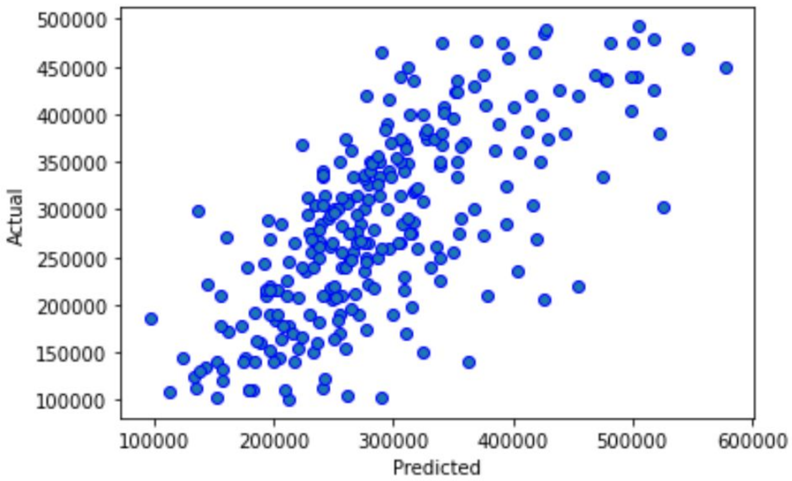
Gradient Boosting Regression results Test

RMSE: 74,712
MAE: 58,903
R^2: 0.42



Augogluon Regression results Test

RMSE: 72,130
MAE: 57,051
R^2: 0.46



Results Comparison

Solution Effectiveness

	Baseline Linear Reg	SK Learn GB Regression	AutoGluon Ensemble
R Squared	0.74	0.83	0.82
MAE	$93,381	$83,981	$82,611
RMSE	$137,752	$121,867	$124,718

Next Steps

Next Steps

- Modeling
  - Computer Vision
  - NLP on Property Descriptions

- Business
  - Expansion to other cities
  - Property investment opportunities

Lesson Learned

- Real World Variability

- New Information is helpful

References

Redfin - https://www.redfin.com/

Pittsburgh School Data - https://nces.ed.gov/ccd/districtsearch/district_detail.asp?ID2=4219170

Pittsburgh Crime Data - https://data.wprdc.org/dataset/uniform-crime-reporting-data

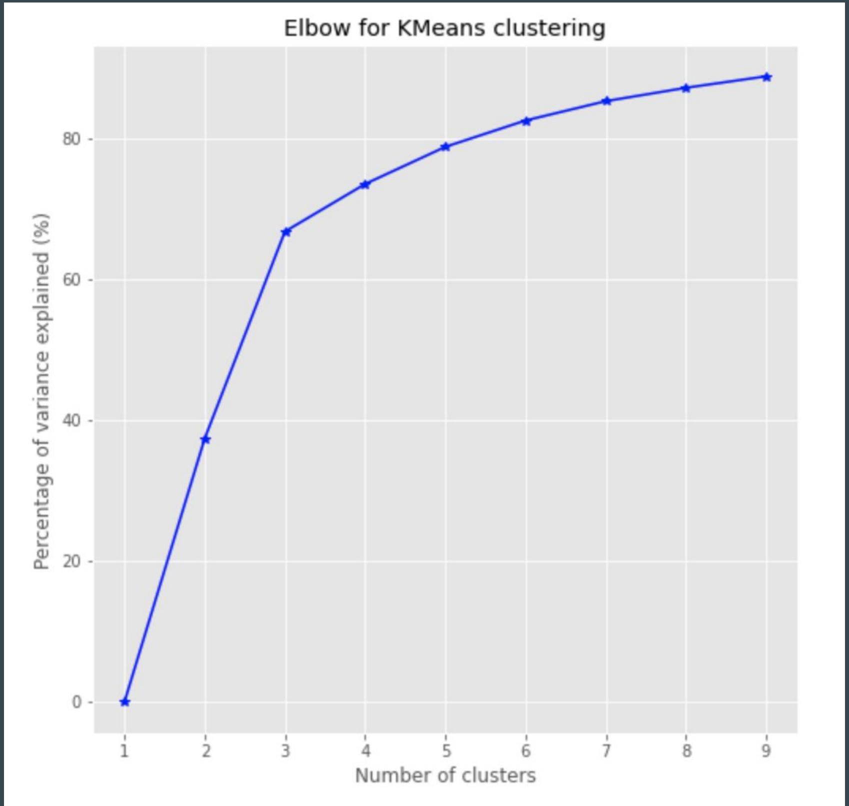
American Census Survey - https://www.census.gov/programs-surveys/acs

Pittsburgh Census Data - https://api.census.gov/data/

Redfin Description

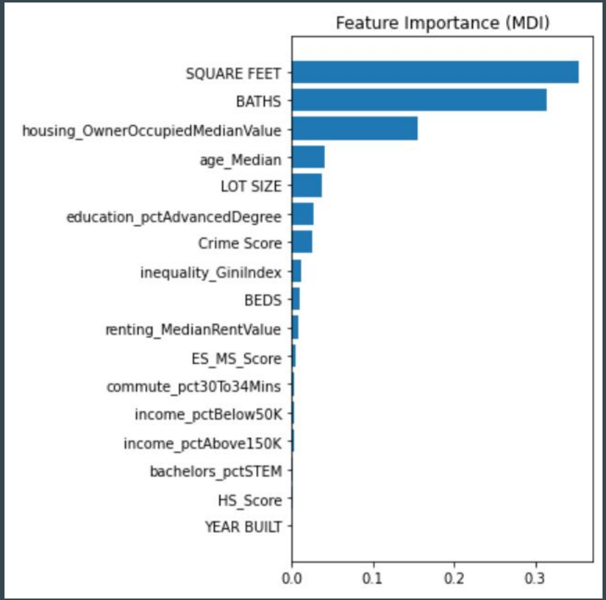
	PRICE	BEDS	BATHS	SQUARE FEET	LOT SIZE	YEAR BUILT	yearSold	monthSold
count	1.989000e+03	1989.000000	1989.000000	1989.000000	1.989000e+03	1989.000000	1989.000000	1989.000000
mean	3.634715e+05	3.356461	2.032931	1868.541981	8.697352e+04	1919.804424	2021.212670	6.885872
std	2.836188e+05	1.101095	0.952755	916.721625	3.706572e+06	28.070127	1.002271	3.052285
min	3.000000e+03	0.000000	0.500000	100.000000	4.300000e+01	1620.000000	2017.000000	1.000000
25%	1.990000e+05	3.000000	1.500000	1288.000000	1.742000e+03	1900.000000	2021.000000	5.000000
50%	2.950000e+05	3.000000	2.000000	1632.000000	2.857000e+03	1915.000000	2021.000000	7.000000
75%	4.420000e+05	4.000000	2.500000	2146.000000	4.356000e+03	1930.000000	2022.000000	9.000000
max	3.150000e+06	10.000000	7.500000	9800.000000	1.653102e+08	2022.000000	2022.000000	12.000000

Elbow Plot for K-Means



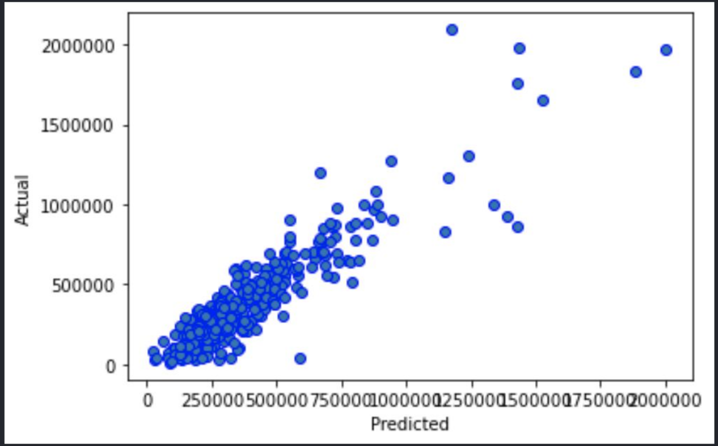
Scikit Learn Champion Model

Gradient Boosting Champion Model




```
Gradient Boosting  results  testing

RMSE:  124,103
MAE:   83,081
R^2:   0.83
```



AutoML Model Summary

	model	score_test	score_val	pred_time_test	pred_time_val	fit_time	pred_time_test_marginal	pred_time_val_marginal
0	KNeighborsDist_BAG_L1	-0.009254	-94503.542167	0.081001	0.079997	0.007003	0.081001	0.079997
1	ExtraTreesMSE_BAG_L1	-32336.244507	-88013.193618	0.340000	0.103001	0.418531	0.340000	0.103001
2	RandomForestMSE_BAG_L1	-32655.529895	-88931.898728	0.323603	0.142521	0.588086	0.323603	0.142521
3	LightGBMLarge_BAG_L1	-35722.490630	-90949.892843	0.255069	0.055046	111.492148	0.255069	0.055046
4	XGBoost_BAG_L1	-48967.557468	-90132.450736	0.463506	0.063006	29.379372	0.463506	0.063006
5	WeightedEnsemble_L2	-50821.279852	-82663.896414	1.560961	0.514040	230.363301	0.010003	0.001000
6	CatBoost_BAG_L2	-54514.490200	-85647.385207	2.529183	0.850601	324.738806	0.046061	0.018027
7	ExtraTreesMSE_BAG_L2	-55107.740047	-86190.300607	2.721129	0.958577	279.030031	0.238007	0.126003
8	RandomForestMSE_BAG_L2	-56376.698633	-87558.457622	2.786123	0.965562	279.304029	0.303001	0.132987
9	WeightedEnsemble_L3	-56626.070952	-83862.584294	3.549666	1.301103	447.021861	0.007999	0.001000
10	XGBoost_BAG_L2	-56730.190265	-89000.715616	2.671115	0.902619	303.194270	0.187993	0.070045
11	LightGBMXT_BAG_L2	-57072.669427	-85215.750754	2.840119	0.924579	360.717721	0.356997	0.092004
12	LightGBM_BAG_L2	-60027.630535	-88874.699325	2.530123	0.851135	294.348327	0.047001	0.018560
13	NeuralNetFastAI_BAG_L2	-60871.639785	-86811.756196	2.900602	1.064069	318.197394	0.417480	0.231495
14	LightGBM_BAG_L1	-63967.901916	-87909.295392	0.062057	0.018015	18.406258	0.062057	0.018015
15	LightGBMXT_BAG_L1	-65497.239355	-86761.932689	0.121130	0.030996	18.518195	0.121130	0.030996
16	NeuralNetFastAI_BAG_L1	-67531.790104	-85544.692577	0.624768	0.226999	38.902772	0.624768	0.226999
17	CatBoost_BAG_L1	-75030.357181	-88919.960555	0.128991	0.017002	60.877663	0.128991	0.017002
18	KNeighborsUnif_BAG_L1	-77579.979509	-95760.004733	0.082997	0.095992	0.007006	0.082997	0.095992

Variables Selected

B01002_001E -> age_Median
B25109_001E -> housing_OwnerOccupiedMedianValue
B25111_001E -> renting_MedianRentValue
B08134_001E -> commute_Total
B08134_007E -> commute_30to34mins
B15012_001E -> bachelors_Total
B15012_009E -> bachelors_STEM
B15003_001E -> education_Total
B15003_023E -> education_MasterDegree
B15003_024E -> education_ProfessionalDegree
B15003_025E -> education_DoctorateDegree
B19083_001E -> inequality_GiniIndex

B19001_001E -> income_Total
B19001_002E -> income_LessThan10K
B19001_003E -> income_10Kto15K
B19001_004E -> income_15Kto20K
B19001_005E -> income_20Kto25K
B19001_006E -> income_25Kto30K
B19001_007E -> income_30Kto35K
B19001_008E -> income_35Kto40K
B19001_009E -> income_40Kto45K
B19001_010E -> income_45Kto50K
B19001_014E -> income_100Kto125K
B19001_015E -> income_125Kto150K
B19001_016E -> income_150Kto200K
B19001_017E -> income_200KOrMore

Source -> https -> //www2.census.gov/geo/maps/DC2020/PL20/st42_pa/censustract_maps/c42003_allegheny/DC20CT_C42003.pdf