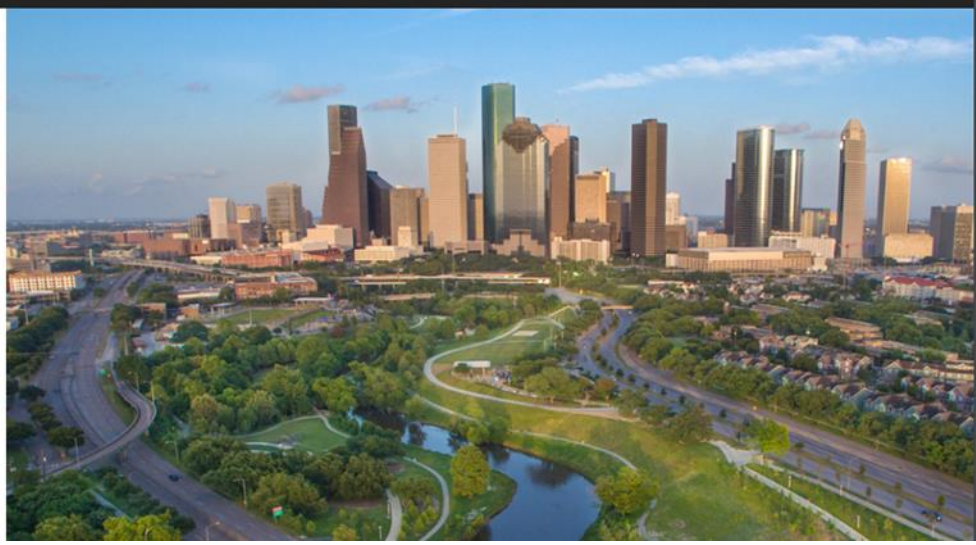
Real estate Predictions in Houston TX

by Randy Miskuski



Background:

The purpose of this project is to effectively build a tool that can give homeowners more informed buying decisions. Furthermore, such a successful tool would be used directly in line with an investing strategy as a supplementary tool on mitigating the risk of an investment.

The passion of such a project was born back in January 2020, when I actively started looking for a land development project in Sacramento California, and Houston Texas. After further consideration, Houston Texas proved to be a more lucrative decision, so I focused my efforts in this city. Building a proposal felt rudimentary and didn't have anything unique that would separate me from competing investment groups. My digital decision is owed to a friend of mine who is a Web Developer at Stack. He had recommending doing something digital that allows me to work remote, and also giving me the freedom to work such a project like this along with a career.

After further research of my own, I had understood that Data Science would leverage my existing Engineering and Statistics knowledge and introduce me to Machine Learning where I could develop tools to give me a competitive edge. Upon entering BrainStation, I knew that I wanted a tool that could compliment my existing work, and give me something that not as many people are using. Machine Learning. Starting this project, I knew my bounds and what I needed to do, so finding the specific Houston data that was current, was 100% necessary.

Data Acquisition:

After searching the internet high and low, I could not find the data I was looking for. It wasn't until the very last day where I had agreed, if I had no success, I would change my Capstone project, I found it. The data I found was publicly found from https://github.com/ellenrud84/RealEstateApp. The had done a web scrape sourcing 19,036 rows of data from 2019 Houston Texas Real-estate database.

Our database consists of 19,077 rows and 18 columns where data related to single family properties in 8 selected Zip Codes in Houston Texas. The data contains property description such as id, account, appraised value, percent change, new owner date, square feet, acreage, latitude, longitude, zip code, neighborhood code, offence count, school id, school type, school rating, flood risk, & sales by neighborhood. We will be looking how each of these features are correlated with each other, and determine what factors influence the price of a home to go up or go down. We will be using linear regression and a decision tree regressor to model our variance. Then, by using that knowledge predict the evaluated home price, and determine if the value will go up by the following year.

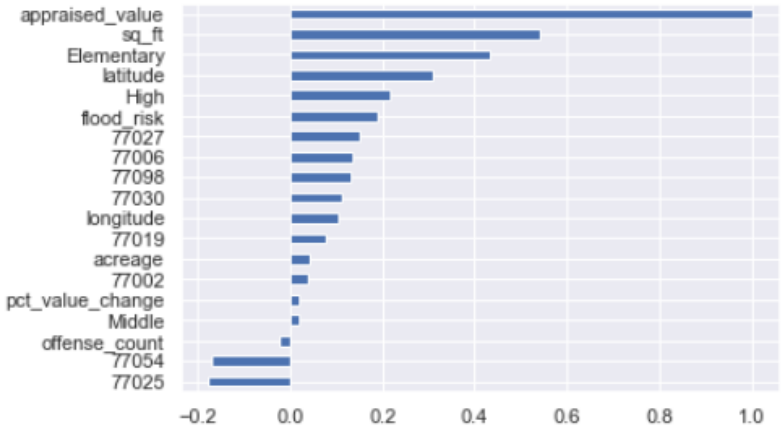
It should be noted that this project is a Proof of Concept and is not representative of the population of Houston TX. Our dataset has a maximum appraised value of 450,000 USD which is not truly the maximum in Houston. In **Results and Interpretation** of this report, on our Homoscedasticity scatter plot, our “buckshot” that we anticipated has an irregular border in our scatterplot and that is due to our house price being capped. Had the appraised value been represented by the population of Houston TX, we would have not seen the cropped section in our scatterplot, and we would have had the entire representation of our “buckshot”. However, this project will be a Proof of Concept and will be further evaluated in Phase 2 of our analysis completed later this year. But for now, we will assume that the highest priced home in Houston TX based on our sample is 450,000 USD.

Modeling

The first model that we approached was a using a statistical approach to look at the correlation between our features. And what we found was that square feet, Elementary, and Latitude were our top 3 predictors with the highest correlation.

Out[104]:

	correlation	p-value
appraised_value	1.000000	0.000000e+00
pct_value_change	0.019612	1.178740e-01
sq_ft	0.540807	0.000000e+00
acreage	0.042160	7.714365e-04
latitude	0.308615	2.222927e-140
longitude	0.105895	2.527719e-17
offense_count	-0.022050	7.871085e-02
flood_risk	0.188875	3.784285e-52
Elementary	0.431612	6.240121e-287
High	0.217820	3.649611e-69
Middle	0.018070	1.496420e-01
77002	0.039950	1.440664e-03
77006	0.134796	3.584739e-27
77019	0.075697	1.505055e-09
77025	-0.176021	2.023862e-45
77027	0.150689	1.297057e-33
77030	0.110476	1.000622e-18
77054	-0.169652	2.849691e-42
77098	0.130591	1.367436e-25



Square footage was to be expected when looking at our correlation, because the larger the home, the more money it will cost. The other correlation that was found was found to be predictive of Appraised Value was Elementary. This is likely to be linked with the performance of the Elementary school, and lining that to they type of neighborhood that the relative home is associated with. For example: An elementary school with a high rating 95/100 is likely going to be in a nicer higher income neighborhood versus an Elementary school with a lower rating like 65/100. This information is predictive of the appraised value of a home.

```
X.corrwith(y)
```

pct_value_change	0.019612
sq_ft	0.540807
acreage	0.042160
latitude	0.308615
longitude	0.105895
offense_count	-0.022050
flood_risk	0.188875
Elementary	0.431612
High	0.217820
Middle	0.018070
77002	0.039950
77006	0.134796
77019	0.075697
77025	-0.176021
77027	0.150689
77030	0.110476
77054	-0.169652
77098	0.130591

When We look at the Correlation with respect to our target variable (appraised_value) we see that our top 3 coefficients are Square Feet (0.5408) “Moderate Correlation”, Elementary School Rating (0.4316) “Weak Correlation”, and Latitude (0.3086) “Weak Correlation”. Typically Latitude and Longitude would be considered non-valuable in ML models, but since Houston is an Oceanic Metropolitan Hub, proximity to ocean is quite important.

After fitting using Order of Least Squares (OLS) and fitting it to our X, and y variables, we were able to achieve a variance explained, (or R Squared) of 0.583. Often times scaling data can prove to be beneficial to a models performance, but in our instance our data was proved to be normally distributed, by plotting a histogram of our residuals, and performing a Q-Q plot. Both suggested the distribution of our Data to be normal.

We applied a Standard Scaler to the data because MinMax Scaler would not be appropriate due to the distribution of the data. And after plotting the difference between the scaled data, and the non-scaled data. It seems that there is not large difference in the shape of the graphs. And sure enough, once we score the data, we get an R Squared of 0.5857.

The next model we conducted was Principal Component Analysis. Before optimizing, the model gave us a variance explained, or an R Squared of 0.5946. However, in order to optimize the models performance, we plotted the proportion of the Variance explained against the number of PC’s and found that the “elbow” point at which we want to chose as our number of PC’s was best found a value of 14 PC’s out of the 18pc’s we had started with. By running PCA with a (PC = 14) we were able to get an R Squared value of 0.5880. By reducing the number of PC’s we had actually inhibited our models performance.

We then used a Decision Tree Regressor to model our data. Though decision trees are prone to overfitting, we found that at a depth of 6, proved to give us the best result, just before we see our train data beginning to overfit to the noise in our data. After using a Decision Tree Regressor, our model had given us a variance explained of 0.7684.

The Last Model that we used was a Random Forest Regressor. When optimizing a Random Forest, we adjust the range of trees in our forest, and the depth of our trees. Trying a variety of depths, we found that using a forest containing 5 trees proved to give us the best variance explained in our model. We were able to achieve an R Squared of 0.8451

Now that we have proved Random Forest Regressor to be the best performing model, Phase 2 of this project will entail finding more data from Houston TX, and with the same features, see if we can accurately predict the price of a home using the same ML practices. Results will be coming this summer to finalize the capstone project.