“Fixer - Upper” Mining

Venkat Sai Charan Bandi

December 9, 2020

Abstract

Real Estate Properties across the world, change their value rapidly based on many factors, including economy, property specifications, location, etc. One of the biggest challenges for individuals working as a Real Estate professional or a beginner investor is to find and select the best pool of properties to invest in. This project aims to push forward this procedure with efficiency and accuracy using specific data mining techniques exploring a specific problem of “fixer-upper” and with the help of proper investment and effort can be improved in many ways and this boosts the price of the property. This work explains the development of a framework that would collect, analyze and pick out the best properties that are mainly under this category.

1 Introduction

The increasing number of real estate properties in the modern world has been a promising market for individuals and families requiring shelter and investors to make money in the market, but this has also resulted in unprecedented skyrocketing in the price of these properties over the last few decades. U.S. Housing Market’s Combined Value Hits $33.6 Trillion in 2020 [5], this also leads to many people not able to afford and keep up with the income to buy such houses. In contrast to the traditional purchase of new and maintained houses, an alternative is to invest in houses that have been around for a lot of years are degrading in value allowing for a cheap purchase and some initial investment to fix all the broken and necessary features to keep up with the times. These Houses are often referred to as “Fixer Uppers”, and could be a great way to save money and prove to be a valuable way to invest. This project deals with using Data Mining Techniques as described in the Methods and Techniques section of this paper while discussing the results and analysis of the model in the Discussion and Results section.

2 Problem Statement

A translation barrier to find and understand the real estate business leaves a lot of individuals unable to make proper investments, while resources like time, energy, and a lot of money are being misused. Unlike traditional house listings, there is no separate board to identify and chart the Fixer Upper properties because there is no automated approach to find them electronically. These comparisons and analytical data are mostly local to an individual and lead to ultimately not making a profit by being limited to their scope and knowledge. This project starts with the idea to find a model that allows estimating the most accurate price of any given house specification while also using a UI based framework that allows an individual with any skill-set to help give them the best tools to make a real estate investment with the best properties available on the market that are considered 'Fixer-Uppers' identified by using multiple data mining techniques in real-time.

3 Literature Review

This project can be viewed as a solution to multiple underlying set of problems related to the real estate industry and basic principles to enable property flipping while maintaining profits with bare risks associated with budgeting or unexpected expenses.

One of the main motivation for this project is direct ongoing research work on real estate analysis by Dr.Setareh Rafatirad [2] on exploring a variety of Machine Learning and Deep Learning techniques for housing based data. While most real estate price prediction methods are vastly used with various techniques, some of the work is done in [3] to establish basic models and perform standard regressions. A 70% Rule, stating Our ARV(After repaired Values) should not exceed 65 to 70%, gives us enough leeway for things to go wrong or compensation for extra incurred losses.

4 Methods and Techniques

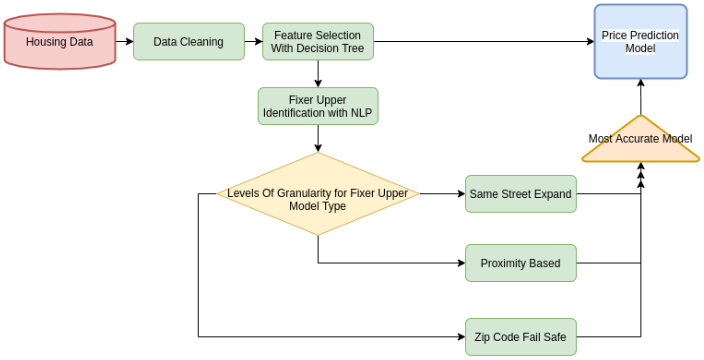

Housing Data
Data Cleaning
Feature Selection With Decision Tree
Price Prediction Model
Fixer Upper Identification with NLP
Most Accurate Model
Levels Of Granularity for Fixer Upper Model Type
Same Street Expand
Proximity Based
Zip Code Fail Safe

Figure 1: Flowchart for the proposed Framework

For this project, the proposed framework is shown in Fig.1 and this section describes how each phase of this methodology is developed

4.1 Data Cleaning

Data used here, as mentioned in the Data set Subsection of this paper, is taken to perform the initial corpus of this project. We import our data as a pandas Data Frame to work with powerful data structures throughout the task. We initially remove all data that is inconsistent and has ‘Nan’ or ‘NULL’ or ‘No Data’ fields since we have an abundance of data and having precise details of a house can get us accurate models. The Data set is part of a bigger project and hence consists of all kinds of data that are not required for our purpose, such as coordinates, dates listed, house ID, etc. The following step is to identify features that can either be misleading or unnecessary for our model. For example, we don’t need to use the description of the house for calculating pricing, etc. Hence we drop these columns from our frame.

4.2 Feature Selection

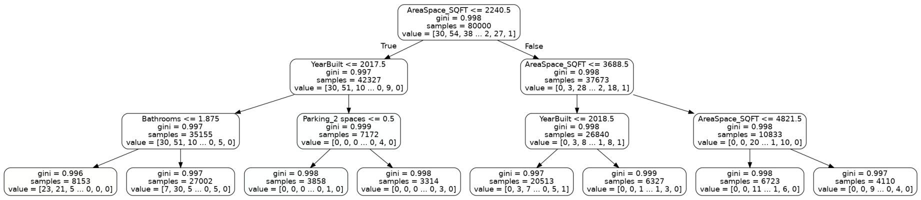


Figure 2: Decision Tree with GINI index

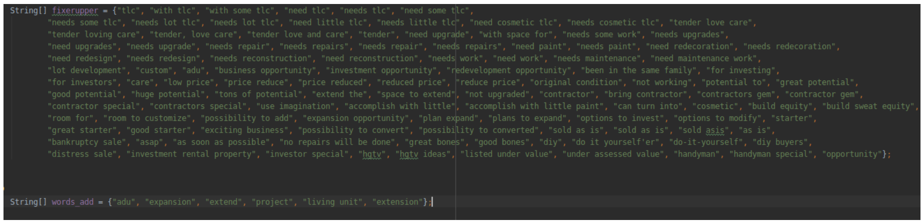
When the importance of the price of a house is determined by categorical features such as amenities it provides and the geographical location of the property, we will consider this as a multi class classification problem and features are well representing in decision trees based on the GINI index of the contribution they provide to the model. Hence for this project, I used Decision Tree Classifier from ‘Sklearn’ [4]. We will now divide the data into two types, numerical and categorical. Numerical data such as the number of bedrooms and size of the house are numerical and can be adjusting in the model but features such as type of heating or cooling, rich localities, or parking types are one hot encoded using the built-in pandas get_dummies function to convert them accordingly. After training the classifier and Upon adjusting the pruning of the tree for visual purposes, the tree looks like Fig.2.

From the analysis of this tree and using the Sklearn’s ’compute feature importance’s’ function we can get some of the most important features listed rank wise as : AreaSpace_SQFT, Bathrooms, Bedrooms, YearBuilt, Cooling, Locality.

Upon analyzing the numbers, I began to understand that the single most important feature is the area space with 0.56 importance and followed by the combined total of bathroom and bedroom score of 0.24 and hence I followed these weights to gather weights of importance to my model that can predict prices later.

4.3 Property Identification with NLP

The next step once we are finalized with the prediction model is to perform the main purpose of this project to identify The fixer Upper Properties of this Project. To maintain a comparison of keywords, I came up with a dictionary of n-grams from research listings in websites as shown in Fig.3 that would allow me to match the house to a category of 'Fixer Upper' and use cosine similarity to add them to the fixer-upper data set. While also making another set of keywords that could help describe the scope of remodeling of the house, such as room to expand or ability to have an additional living unit this could help boost the value of the house by increasing our most important feature that is the area space.



```
String[] fixerupper = {"tlc", "with tlc", "with some tlc", "need tlc", "needs tlc", "need some tlc",
        "needs some tlc", "needs lot tlc", "needs lot tlc", "need little tlc", "needs little tlc", "need cosmetic tlc", "needs cosmetic tlc", "tender love care",
        "tender loving care", "tender, love care", "tender love and care", "tender", "need upgrade", "with space for", "needs some work", "needs upgrades",
        "need upgrades", "needs upgrade", "needs repair", "needs repairs", "needs repair", "needs repairs", "need paint", "needs paint", "need redecoration", "needs redecoration",
        "need redesign", "needs redesign", "needs reconstruction", "need reconstruction", "needs work", "need work", "needs maintenance", "need maintenance work",
        "lot development", "custom", "adu", "business opportunity", "investment opportunity", "redevelopment opportunity", "been in the same family", "for investing",
        "for investors", "care", "low price", "price reduce", "price reduced", "reduced price", "reduce price", "original condition", "not working", "potential to", "great potential",
        "good potential", "huge potential", "tons of potential", "extend the", "space to extend", "not upgraded", "contractor", "bring contractor", "contractors gem", "contractor gem",
        "contractor special", "contractors special", "use imagination", "accomplish with little", "accomplish with little paint", "can turn into", "cosmetic", "build equity", "build sweat equity",
        "room for", "room to customize", "possibility to add", "expansion opportunity", "plan expand", "plans to expand", "options to invest", "options to modify", "starter",
        "great starter", "good starter", "exciting business", "possibility to convert", "possibility to converted", "sold as is", "sold as is", "sold asis", "as is",
        "bankruptcy sale", "asap", "as soon as possible", "no repairs will be done", "great bones", "good bones", "diy", "do it yourself'er", "do-it-yourself", "diy buyers",
        "distress sale", "investment rental property", "investor special", "hgtv", "hgtv ideas", "listed under value", "under assessed value", "handyman", "handyman special", "opportunity"};



String[] words_add = {"adu", "expansion", "extend", "project", "living unit", "extension"};
```

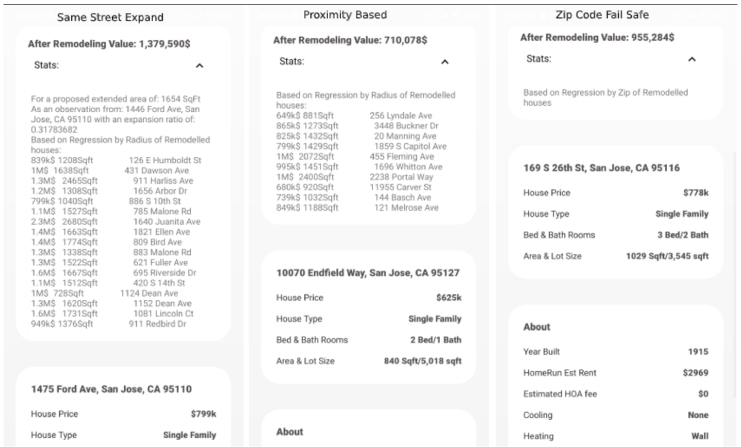
Figure 3: Dictionary of Fixer Upper Keywords

4.4 Levels of Granularity in the model

To most accurately maintain the accuracy of the data and to keep with the constant changes in the value of real estate each time frame, we cannot stick to using data that generalizes the purpose repeatedly. To overcome this, picking models and parameters according to the most precise and accurate data available gives us the safest investment. One such attempt to solve this issue is to come up with an algorithm that switches between the type of houses used to evaluate the model based on several features, they are described below and shown in Fig.4

4.4.1 Same Street Expand

Addresses on the same street can be identified by taking apart the street name after the house number. If there exists more than one house within our data set that belongs in the same street, using "Address Matching with Google Maps API" [1] then we get the average area space of the house on the same street



Same Street Expand

After Remodeling Value: 1,379,590$

Stats:

```
For a proposed extended area of: 1654 SqFt
As an observation from: 1446 Ford Ave, San
Jose, CA 95110 with an expansion ratio of:
0.31783682
Based on Regression by Radius of Remodelled
houses:
839k$ 1208Sqft        126 E Humboldt St
1M$  1638Sqft        431 Dawson Ave
1.3M$  2465Sqft        911 Harliss Ave
1.2M$  1308Sqft        1656 Arbor Dr
799k$ 1040Sqft        886 S 10th St
1.1M$  1527Sqft        785 Malone Rd
2.3M$  2680Sqft        1640 Juanita Ave
1.4M$  1663Sqft        1821 Ellen Ave
1.4M$  1774Sqft        809 Bird Ave
1.3M$  1338Sqft        883 Malone Rd
1.3M$  1522Sqft        621 Fuller Ave
1.6M$  1667Sqft        695 Riverside Dr
1.1M$  1512Sqft        420 S 14th St
1M$  728Sqft        1124 Dean Ave
1.3M$  1620Sqft        1152 Dean Ave
1.6M$  1731Sqft        1081 Lincoln Ct
949k$ 1376Sqft        911 Redbird Dr
```

1475 Ford Ave, San Jose, CA 95110

House Price	**$799k**
House Type	**Single Family**

Proximity Based

After Remodeling Value: 710,078$

Stats:

```
Based on Regression by Radius of Remodelled
houses:
649k$ 881Sqft        256 Lyndale Ave
865k$ 1273Sqft        3448 Buckner Dr
825k$ 1432Sqft        20 Manning Ave
799k$ 1429Sqft        1859 S Capitol Ave
1M$  2072Sqft        455 Fleming Ave
995k$ 1451Sqft        1696 Whitton Ave
1M$  2400Sqft        2238 Portal Way
680k$ 920Sqft        11955 Carver St
739k$ 1032Sqft        144 Basch Ave
849k$ 1188Sqft        121 Melrose Ave
```

10070 Endfield Way, San Jose, CA 95127

House Price	**$625k**
House Type	**Single Family**
Bed & Bath Rooms	**2 Bed/1 Bath**
Area & Lot Size	**840 Sqft/5,018 sqft**

About

Zip Code Fail Safe

After Remodeling Value: 955,284$

Stats:

Based on Regression by Zip of Remodelled houses

169 S 26th St, San Jose, CA 95116

House Price	**$778k**
House Type	**Single Family**
Bed & Bath Rooms	**3 Bed/2 Bath**
Area & Lot Size	**1029 Sqft/3,545 sqft**

About

Year Built	**1915**
HomeRun Est Rent	**$2969**
Estimated HOA fee	**$0**
Cooling	**None**
Heating	**Wall**

Figure 4: Levels of Granularity

and if the average is more than the property we are looking at within relatively same total lot size, we propose to the user the area can be expanded by building or remodeling new structures or extend the construction to get more area for the property directly leading in a guaranteed way of increasing the price of the house. We then update our data set with the proposed house to match pricing with our model later.

4.4.2 Proximity Based

With our limited data set of housing data, if we fail to identify any houses in the same street from the method above or if the houses are incomparable in the total lot size they provide, we will go with trying to identify all data points within the geographical area that is relatively close to the property we are looking at. Usually from testing, better results are observed within 2 to 3 miles of radius. We can use the coordinates mentioned in the data set or use [1] to tag the distance metrics. We calculate the house price after getting the prices of all the remodeled houses near the area and then comparing the fixer-upper price based on a set threshold.

4.4.3 Zip Code Fail Safe

A fail-safe measure of predicting the house price in case the first two most accurate methods fail, this is a general approach for price prediction that has been tried and tested in [3] to generate results over time. This can be used as a last resort, but given the size of the data set and testing, there were very few times this was called, meaning mostly has sufficient data to self-sustain with the above two methods.

4.5 Model Selection and Prediction

To estimate the price of the Fixer-Uppers after remodeling or to get a value to the amount of return of Investment one might be expecting, we use our pre-trained model as described above to combine regression techniques from granularity and use our Decision Tree Approach to arrive at accurate numbers of expectancy.

5 Discussion and Results

5.1 Datasets

The raw data is scraped from listings of Zillow [6], an online real estate marketplace, collected over the last 3 years with a web scraper that crawls through house listings and stored in a web server with node JS API routing, MongoDB storage, and downloaded as a CSV file for this project. It Has a total of 24 Features mainly with important features such as: Address, AreaSpace_SQFT, Bathrooms, Bedrooms, Cooling, Description, HOAFee, Heating, Locality, Lot, Parking, Price, State, Type, YearBuilt, Zip Code Data is located at http://129.174.126.176:8080/api/ on the GMU server and can be accessed using GMU VPN. There are 1,378,574 total listings collected to the date of writing, but for this project, I scaled down the size for better run times and got rid of older and inconsistent data.

5.2 Evaluation Metrics

The evaluation of this project mainly has two parts to it, one being the actual model of the house price prediction and the other being the Fixer Upper Analysis.

5.2.1 Price Prediciton Model

To test the performance of this model i used the Sklearn train_test_split to divide the data in multiple parts and evaluate the model on it's accuracy. A good way to measure accuracy is to use precision_recall_fscore_support, but in the real world the most accurate results needn't be in the exact dollar value, rather the closeness the price of the house. Upon cross Validation Testing and pruning and tuning the hyper-parameters the average Precision is around: 0.9983126799496166

which i believe is very accurate given that all features such as description, etc are not taken into consideration.

5.2.2 Fixer Upper Analysis

To create a metric for evaluation based on Fixer Uppers", since we are predicting the value of the house on no existing data set or values, we can follow a manual method of finding the closest properties as shown in Fig.5 and comparing the price history of the house before remodelling and then comparing it after, to get the increase of price and calculate the change of percentage for our property. This could be expansion or the radius based regression. We will identify the most recently remodeled houses to get the most probable estimate of the current selling price and then compare them against our model. Chart results within this for randomly selected data sets across the fields and make calculations for accuracy.

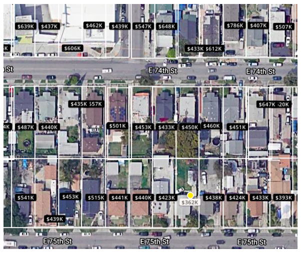
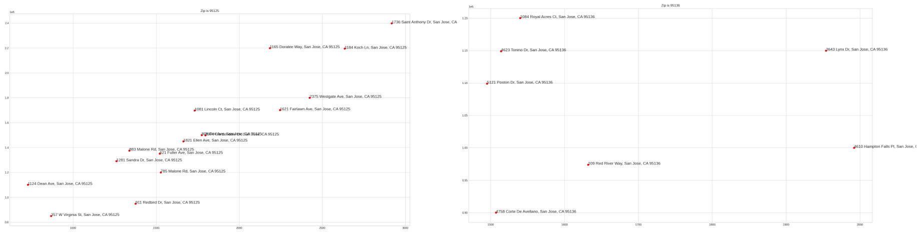


Figure 5: Nearby House Sale History

5.3 Experimental Results



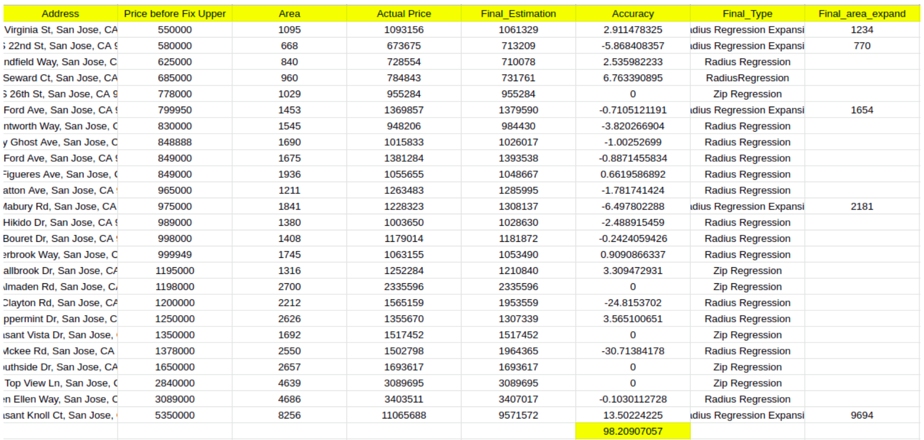
(a) Proper Fitting Line

(b) Irregular Line

Figure 6: Comparison of Line Fitting

While the performance of the Price prediction model is fairly accurate as mentioned in the metrics section consistently giving the most accurate pricing stats for each property as described in 5.2.1, one of the main issues i ran into in this project is the failure to fit a line fitting properly in the regression. For example as shown in the Fig.6. For some Properties and its proximity data points, the line fits linearly on the graph Fig.6a most of the time when we use the features mentioned, but the issue arises due to some outlier's in the houses that are fitted with sophisticated features, such as high end furnishings, expensive flooring etc, this graph is irregular as seen in Fig.6b. This should be improved in future implementation with proper data set and identification as to why these said data points are much expensive or cheaper than normally predictable.

The main experimental analysis of this project, the identification of Fixer Upper Properties identifies about with around 82% accuracy when manually tested against 20 properties to identify if the said property is a Fixer Upper or



Address	Price before Fix Upper	Area	Actual Price	Final_Estimation	Accuracy	Final_Type	Final_area_expand
Virginia St, San Jose, CA	550000	1095	1093156	1061329	2.911478325	ıdius Regression Expansi	1234
; 22nd St, San Jose, CA 9	580000	668	673675	713209	-5.868408357	ıdius Regression Expansi	770
ndfield Way, San Jose, C	625000	840	728554	710078	2.535982233	Radius Regression	
Seward Ct, San Jose, CA	685000	960	784843	731761	6.763390895	RadiusRegression	
; 26th St, San Jose, CA 9	778000	1029	955284	955284	0	Zip Regression	
Ford Ave, San Jose, CA 9	799950	1453	1369857	1379590	-0.7105121191	ıdius Regression Expansi	1654
ntworth Way, San Jose, C	830000	1545	948206	984430	-3.820266904	Radius Regression	
y Ghost Ave, San Jose, C	848888	1690	1015833	1026017	-1.00252699	Radius Regression	
Ford Ave, San Jose, CA 9	849000	1675	1381284	1393538	-0.8871455834	Radius Regression	
Figueres Ave, San Jose, C	849000	1936	1055655	1048667	0.6619586892	Radius Regression	
atton Ave, San Jose, CA	965000	1211	1263483	1285995	-1.781741424	Radius Regression	
Mabury Rd, San Jose, CA	975000	1841	1228323	1308137	-6.497802288	ıdius Regression Expansi	2181
Hikido Dr, San Jose, CA 9	989000	1380	1003650	1028630	-2.488915459	Radius Regression	
Bouret Dr, San Jose, CA	998000	1408	1179014	1181872	-0.2424059426	Radius Regression	
erbrook Way, San Jose, C	999949	1745	1063155	1053490	0.9090866337	Radius Regression	
allbrook Dr, San Jose, CA	1195000	1316	1252284	1210840	3.309472931	Zip Regression	
lmaden Rd, San Jose, CA	1198000	2700	2335596	2335596	0	Zip Regression	
Clayton Rd, San Jose, CA	1200000	2212	1565159	1953559	-24.8153702	Radius Regression	
ppermint Dr, San Jose, C	1250000	2626	1355670	1307339	3.565100651	Radius Regression	
ısant Vista Dr, San Jose,	1350000	1692	1517452	1517452	0	Zip Regression	
Mckee Rd, San Jose, CA	1378000	2550	1502798	1964365	-30.71384178	Radius Regression	
uthside Dr, San Jose, CA	1650000	2657	1693617	1693617	0	Zip Regression	
Top View Ln, San Jose, C	2840000	4639	3089695	3089695	0	Zip Regression	
en Ellen Way, San Jose, C	3089000	4686	3403511	3407017	-0.1030112728	Radius Regression	
ısant Knoll Ct, San Jose,	5350000	8256	11065688	9571572	13.50224225	ıdius Regression Expansi	9694
					98.20907057		

Figure 7: Performance Results of Fixer Upper

not, while the accuracy of the Fixer Upper after pricing will be subjective and personal to choice and bids at the time of sale, based on the remodelled houses around the proximity or in the same street will always be fitted accordingly with out model and evaluates at an accuracy of 98.2% as shown from calculations of testing in Fig.7 most of the times except when the algorithms fails to identify the address or due to improper input data. We can closely observe the results when we chart them on a graph Fig.8 to compare the actual price of the house before remodelling and the manual comparison of houses that it was sold after remodelled and the price our algorithm came up with before knowing the results. Then we calculate the change in percentage to get the accuracy metrics shown in Fig.7

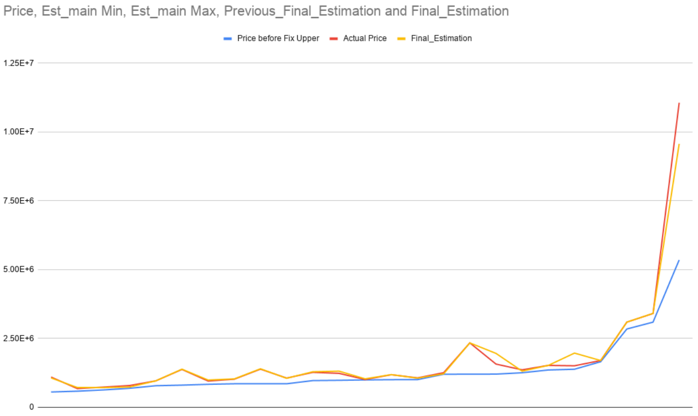


Figure 8: Comparison Graph between model and training values

6 Conclusion

The main low of this project remains the uncertainty of a particular individual during the time of sale and the failure to put a price on a fixer upper since it depends on style and opinion to each person. While Classifying the features of a house with a computational classifier such as Decision Tree Classifier based on the most important features helps me learn what the most important features are that push the price of a house up the most. This implementation of a specific Fixer Upper Problem can lead to much further exploration than a singular Webscrape as i did, and could be much better if i use various other metrics and scores that other websites could provide and i could combine them based on geographical data.

6.1 Directions for Future Work

A model is only as good as its data is, and a lot of the features that determine the price of a house depends on the precision and consistency of data available. A path this project could improve on is to build a custom data set using a feature extraction methodology that describes the data of an already existing description to evaluate the house, with features like furnishing, condition, maintenance, pipelines, water quality, nearby school ratings, etc.

Public data available on housing regulation vary per state rules and implementation of the feasibility of expansion based on legality allowed would improve the profit margin by a lot.

References

[1] *Google Maps API.* URL: https://developers.google.com/maps/.

[2] *Housing Analytics.* URL: http://mason.gmu.edu/~srafatir/pages/research.html.

[3] C. R. Madhuri, G. Anuradha, and M. V. Pujitha. "House Price Prediction Using Regression Techniques: A Comparative Study". In: *2019 International Conference on Smart Structures and Systems (ICSSS).* 2019, pp. 1–5. DOI: 10.1109/ICSSS.2019.8882834.

[4] *sklearn DecisionTreeClassifier.* URL: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html.

[5] *U.S. Housing Market Combined Value Hits $33.6 Trillion in 2020.* URL: https://www.worldpropertyjournal.com/real-estate-news/united-states/los-angeles-real-estate-news/real-estate-news-zillow-housing-data-for-2020-combined-housing-market-value-in-2020-us-gdp-china-gdp-rising-home-value-data-11769.php.

[6] *Zillow.* URL: https://www.zillow.com/.