

RealEstate Investment Prediction

A Project Report

Presented to

The Faculty of the College of

Engineering

San Jose State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science in Software Engineering

By

Kunjan Malik, Praveen Nayak, Yadnyshree Savant, Shivam Shrivastav

December 2021

Abstract

The House Price Prediction is a popular tool for estimating house price fluctuations. Because Housing prices are significantly connected with other characteristics such as location, region, and population, predicting individual house prices requires information other than HPP. Although a huge number of articles have used typical machine learning algorithms to properly predict house prices, they rarely analyze the performance of different models and ignore the less popular yet sophisticated models.

As a result, this study will use both classic and advanced machine learning methodologies to investigate the differences between numerous advanced models in order to investigate the diverse influences of features on prediction methods. This research will also present an optimistic result for housing price prediction by thoroughly validating numerous strategies in model implementation on regression.

Introduction

In several nations, such as the United States, House Price Prediction (HPP) is commonly used to measure price fluctuations in residential housing. The HPP is a repeat-sales index that is weighted and evaluates average price changes in repeat sales or refinancings on the same properties. This data was gathered by looking at mortgage and HOA fees in the state of Iowa. And our forecasts are based on these two variables. Provide knowledge to an investor or buyers, whether to invest in a housing property or not. This decision has to be taken by considering various features such as, Selling Price of the property, Zestimate Price, Proximity Ranking to various Schools, Crime Rate in the area, Walk Score etc.

Machine learning has become a critical prediction approach in recent years, thanks to the growing trend toward Big Data, because it can estimate property prices more correctly based on their qualities, regardless of previous year's data. Several studies have looked into this issue and demonstrated the effectiveness of the machine learning approach.

Related Work

We used the CRISP-DM methodology.

1.Business understanding

Provide knowledge to an investor or buyers, whether to invest in a housing property or not. This decision has to be taken by considering various features such as, Selling Price of the property, Zestimate Price, Proximity Ranking to various Schools, Crime Rate in the area, Walk Score etc.

2.Data understanding

Using IOWA Dataset. We have performed Data cleaning, data preparation and eliminated unwanted columns and considered only the required features such as, address, zip code, area, bedrooms, bathrooms, year built, price, zestimate, zestimate rent. Apart from this data, we scrapped several additional dataset from above-mentioned websites to enrich the initial dataset, amalgamate it and improve the feature set and visualization to deduce the best model.

3.Data preparation – How do we organize the data for modeling?

Calculate Feature importance, gini score.

Feature transformation: Transform features and add new features to dataset via amalgamations.

Data distribution

- Dimensionality Reduction via PCA
- Implement 3 amalgamations:
- First dataset Enrichment
- Second data Enrichment
- Third data enrichment

4.Modeling - Implement different ML Algorithms to build models and refine data narrative.

Modeling - Define a Golden cluster and use Fractal Clustering to find it based on the business case.

Apply various classifiers and regressors algorithms

Classifier:

Nearest Neighbors, Linear SVM, RBF SVM, Decision Tree, Random Forest, Neural Net, AdaBoost, Naive Bayes, QDA

Regression:

GradientBoostingRegressor, RandomForestRegressor, LinearRegression, SVR, DecisionTreeRegressor, AdaBoostRegressor, GaussianProcessRegressor, LogisticRegressor

5. Evaluation - Compare relevant tasks in the same table

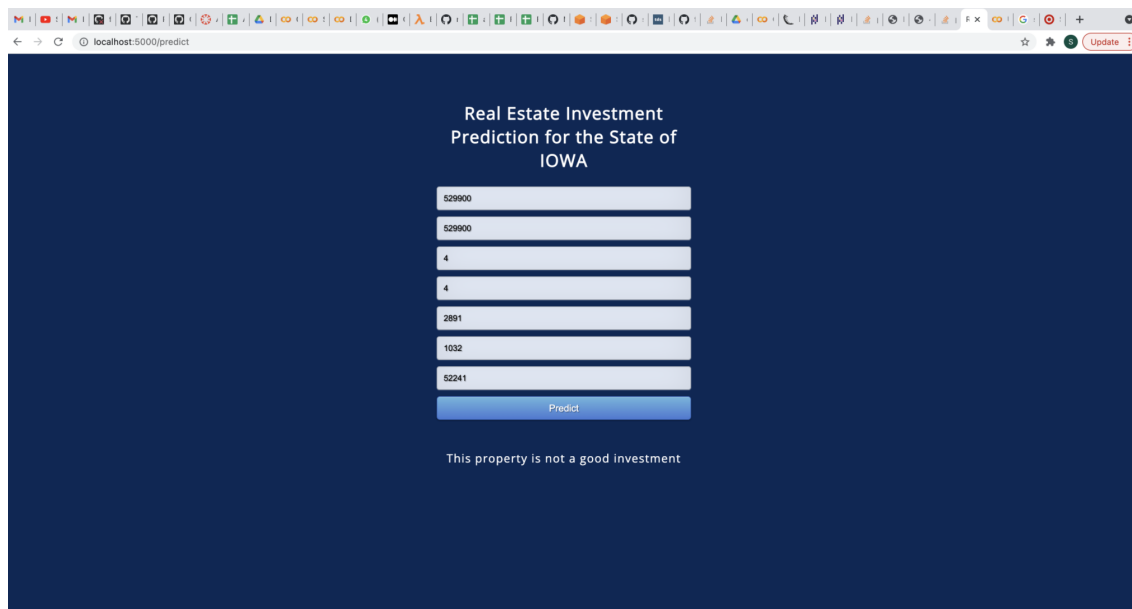
Interpret results of each algorithm.

Suggest Latent Variables or Latent Manifolds, add them to the features and see how prediction results change.

Use appropriate metrics for measuring models and compare them in a table: regression metrics and/or classification metrics (confusion matrix, F1 and R2 score).

6. Deployment

Build a flask API to predict the investment property.



Real Estate Investment
Prediction for the State of
IOWA

529900
529900
4
4
2891
1032
52241
Predict

This property is not a good investment

Data :

Data for the project is

https://drive.google.com/file/d/1P7yK2RzmLvVY9-avomTVvvNxLkPT_Pdl/view?usp=sharing

We had to do the data preprocessing and cleaning. We dropped the features that were not important. We checked if any columns or rows were missing. Few of the features data type were casted to float like the area was in string so we converted it to float. Below are the features we utilized and their respective data-types.

```
Data columns (total 14 columns):
#      Column                Non-Null Count  Dtype
---  -
0      address                641 non-null    object
1      latitude                641 non-null    float64
2      longitude               641 non-null    float64
3      price                   641 non-null    float64
4      bathrooms               641 non-null    float64
5      bedrooms                641 non-null    float64
6      area                    641 non-null    float64
7      zestimate               641 non-null    float64
8      rent_zestimate          641 non-null    float64
9      listing_type            641 non-null    object
10     input                   641 non-null    object
11     property_url            641 non-null    object
12     listing_url             641 non-null    object
13     zipcode                 641 non-null    object
dtypes: float64(8), object(6)
```

Data Scraping: In addition to the above data set we have also scrapped data the

Zillow: <https://www.zillow.com>

Walk Score: <https://www.walkscore.com>

School Proximity Rank: <https://www.niche.com/places-to-live/z/>

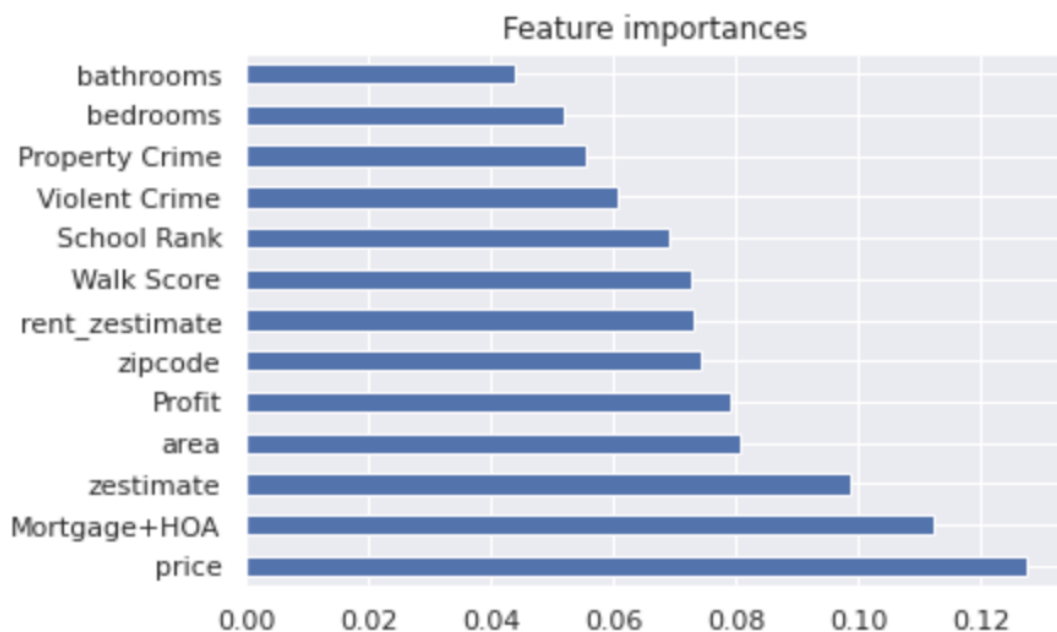
Crime rates: <https://247wallst.com/city/>

Data Cleaning was applied on the scrapped data and then amalgamated to the initial dataset. Finally, we have used the below features to perform the prediction.

#	Column	Non-Null Count	Dtype
0	zipcode	641 non-null	float64
1	bedrooms	641 non-null	float64
2	bathrooms	641 non-null	float64
3	area	641 non-null	float64
4	price	641 non-null	float64
5	zestimate	641 non-null	float64
6	rent_zestimate	641 non-null	float64
7	HOAHOA	641 non-null	float64
8	Walk Score	641 non-null	float64
9	School Rank	641 non-null	float64
10	Violent Crime	641 non-null	float64
11	Property Crime	641 non-null	float64
12	Mortgage fees per month	641 non-null	float64

Methods :

1. Feature Importance: Based on the Gini score, we computed the important features for our data set.



2. Clustering

We used K-means clustering to visualize the data and identify the outliers in the dataset.

We have also calculated the silhouette score and on the basis of silhouette score after several iterations we were able to get the golden cluster.

3. We have also applied fractal clustering with various latent variables.

- Rent v/s (Mortgage+HOA)
- House prediction v/s school proximity
- Zestimate v/s max. walk score
- Zestimate and crime rate

4. We applied the below Classifier to evaluate our model

Nearest Neighbors, Linear SVM, RBF SVM, Decision Tree, Random Forest, Neural Net, AdaBoost, Naive Bayes, QDA.

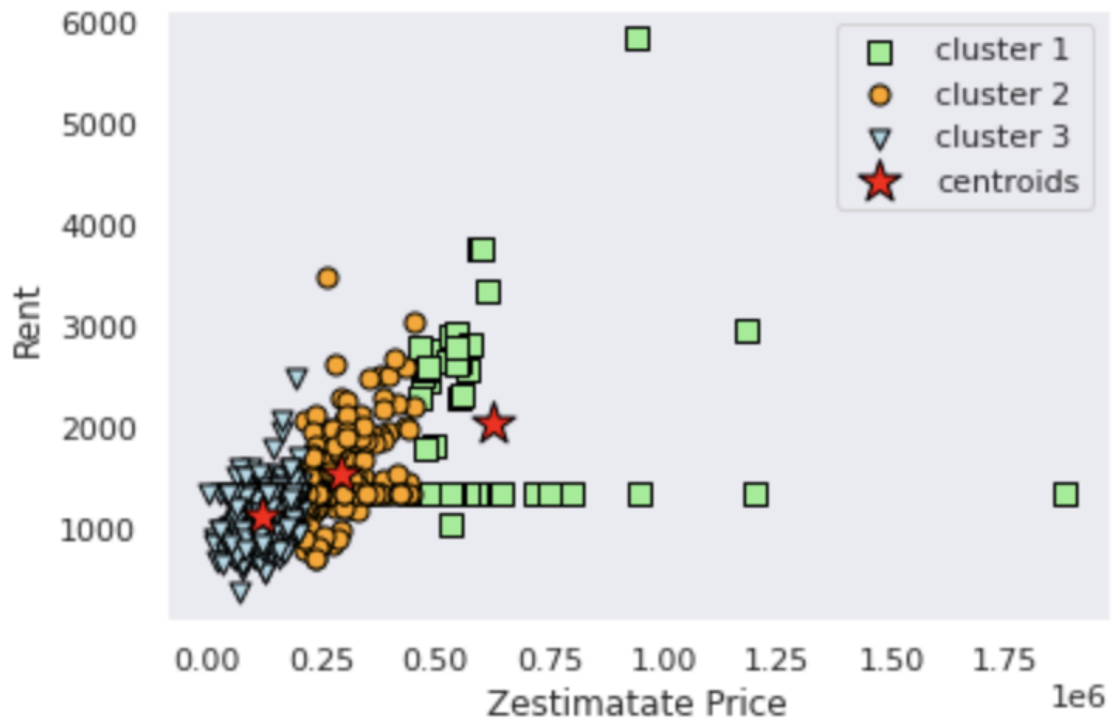
5. We used the below regressor to evaluate which regressor gave us the high accuracy:

GradientBoostingRegressor, RandomForestRegressor, LinearRegression, SVR, DecisionTreeRegressor, AdaBoostRegressor, GaussianProcessRegressor, LogisticRegressor.

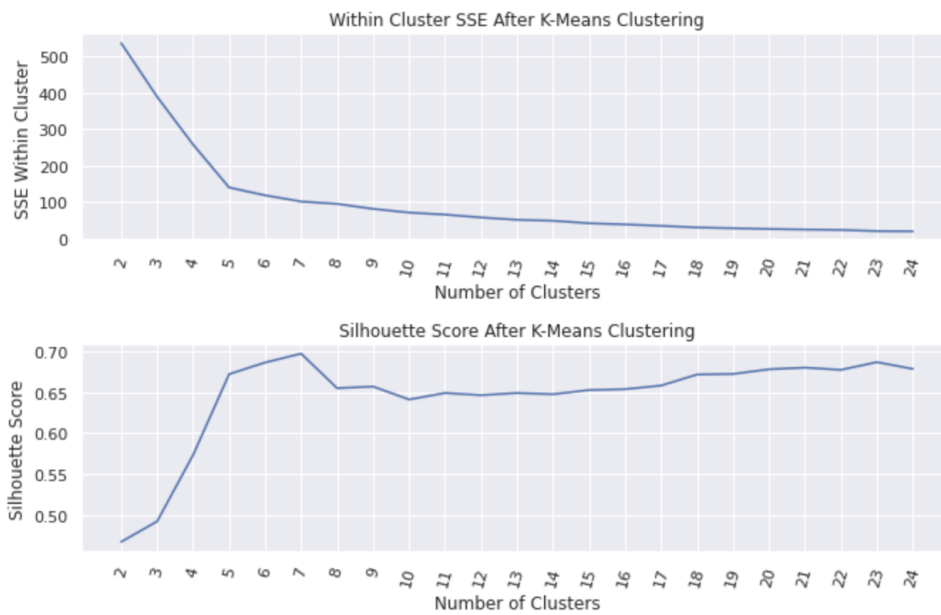
Experiments and Results

To predict the actual and predicted average house prices for the house, trained and split the final data set(80:20) with the important features and then first performed Linear regression separately to predict the house price. This was done by training the model based on the final housing dataset.

Results from K-mean clustering Rent v/s Zestimate



Results from Fractal clustering



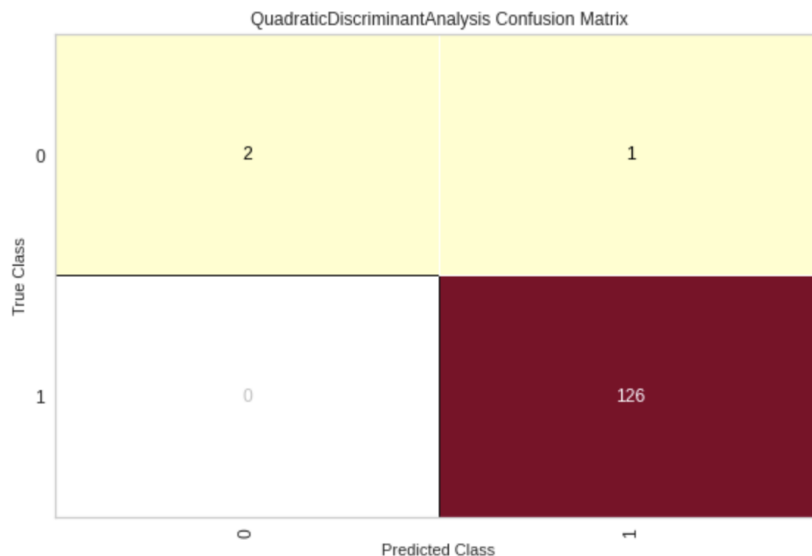
Results from the Regression:

We found that the best regressor for the model is Logistic regressor with an accuracy of 99.22%

```
R2 SCORE = 1.00,  
regressors = GradientBoostingRegressor, Score (test, accuracy) = 49.21,  
R2 SCORE = 0.91,  
regressors = RandomForestRegressor, Score (test, accuracy) = 90.27,  
R2 SCORE = 0.37,  
regressors = LinearRegression, Score (test, accuracy) = 25.27,  
R2 SCORE = 0.77,  
regressors = SVR, Score (test, accuracy) = 59.60,  
R2 SCORE = 1.00,  
regressors = DecisionTreeRegressor, Score (test, accuracy) = 49.21,  
R2 SCORE = 0.98,  
regressors = AdaBoostRegressor, Score (test, accuracy) = 95.39,  
R2 SCORE = 0.37,  
regressors = GaussianProcessRegressor, Score (test, accuracy) = 25.27,  
R2 SCORE = 0.38,  
regressors = LogisticRegressor, Score (test, accuracy) = 99.22,  
-----  
Best --> regressors = LogisticRegressor, Score (test, accuracy) = 99.22
```

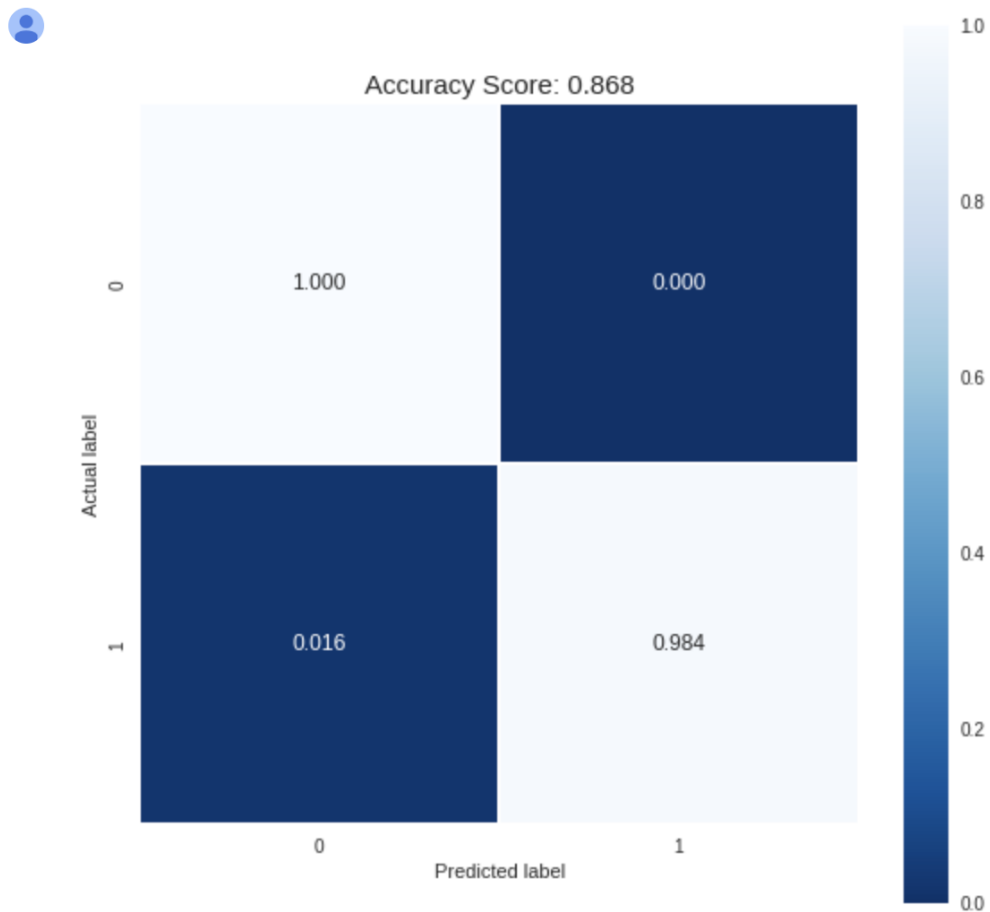
Results from the classifier

The best classifier identified for our model is the Neural Net with 99.22% accuracy.



```
-----  
Best --> Classifier = Neural Net, Score (test, accuracy) = 99.22  
-----
```

Confusion Matrix



Conclusion:

Provide knowledge to an investor or buyers, whether to invest in a housing property or not. This decision has to be taken by considering various features such as, Selling Price of the property, Zestimate Price, Proximity Ranking to various Schools, Crime Rate in the area, Walk Score etc.

Data narrative : Using IOWA Dataset, we performed Data cleaning, data preparation and eliminated unwanted columns and considered only the required features such as, address, zip code, area, bedrooms, bathrooms, year built, price, zestimate, zestimate rent. Apart from this

data, We scrapped several additional dataset from above-mentioned websites to enrich the initial dataset, amalgamate it and improve the feature set and visualization to deduce the best model.

Visualizations:

- Performed Visualizations of data preparation using First, Second and Third Data Enrichment.
- Calculated Feature importance, gini score.

Web Scrapping : we scrapped several additional dataset from below websites to enrich the initial dataset, amalgamate it and improve the feature set and visualization to deduce the best model.

Final Scrapped Dataset file stored in drive is 'DM_final_scrapped_data.csv'.

Zillow: <https://www.zillow.com>

Walk Score: <https://www.walkscore.com>

School Proximity Rank: <https://www.niche.com/places-to-live/z/>

Crime rates: <https://247wallst.com/city/>

Data Preparation: Performed Feature transformation: Transform features and add new features to dataset via amalgamations.

Data Distribution : Did Dimensionality Reduction via PCA and Implemented 3 amalgamations Performed the First, Second and Third Data Enrichment by merging Zillow data, Walk Score Data, School Rank Data and Crime rates data into the Housing dataset in DataFrame itself.

Feature Transformation : Cleaned the final dataset and scrapped dataset, performed feature transformation by checking null values, label encoding, and stored in dataframe for further processing.

Feature Importance : With the help of Correlation Coefficient Heatmap and gini score, I deduced House price, Mortgage+HOA, Zestimate Price, area, zestimate rent, zip code, crime rate, school proximity rank and, WalkScore as the most important features from the final dataset.

Clustering : Used ScikitLearn K-Means clustering for the dataset. To observe the dataset, Created clusters for the various feature combinations. Below mentioned are my observations:

- Rent and HOA+Mortgage Fees : Profit on House rent deducting the HOA + Mortgage fees seems to be a good factor on deciding the profitable investment scope. Houses with more rent and low HOA+Mortgage Fees tend to have high prices but high profit.
- House Price and zip code : Zip Codes have various properties ranging from lowest to highest price.
- zestimate rent and zestimate price : Properties having less zestimate price have the low rent and properties having the high zestimate price have high rent.
- School Proximity Rank as Zestimate price : Properties with high nearby Schools ranking have high property prices and properties nearby schools with low rank have less prices.
- Crime rate and Zestimate price : Properties in areas with Higher crime rates have the lower zestimate price and properties in areas with lower crime rate have the lower zestimate price.

Linear Regression : To predict the actual and predicted average house prices for the house, trained and split the final data set(80:20) with the important features and then first performed Linear regression separately to predict the house price. This was done by training the model based on the final housing dataset.

Fractal Clustering : To suggest a suitable property for an investor to invest on or not, we performed Fractal Clustering on various feature combinations like Rent vs Mortgage+HOA, House Price and Best School Proximity Rank, Zestimate price and maximum Walk Score, Zestimate price and crime Score

Logistic Regression : A good investment property has a positive rental income every month after deducting (Mortgage+HOA fees). Thus, we performed Logistic Regression and created a new feature in the final dataframe as 'Invest_or_Not' based on the value where $\text{Rent} > \text{Mortgage} + \text{HOA}$, which means we get a profit if our rent is much greater than (Mortgage+HOA fees). After training and predicting the model, I got the accuracy of 98%. Thus, it means most of the properties available in provided IOWA dataset is ideal for an investor to make an investment.

Algorithms: we performed algorithms to get the best Classifier and Regressor among below to get the best accuracy of the model to help find the properties to invest on. Classifiers like Nearest Neighbors, Linear SVM, RBF SVM, Decision Tree, Random Forest, Neural Net, AdaBoost, Naive Bayes, QDA and Regression like

GradientBoostingRegressor,RandomForestRegressor,LinearRegression,SVR,DecisionTreeRegressor,AdaBoostRegressor,GaussianProcessRegressor, LogisticRegressor.

Pickle and Load Prepared a model to load the best model for future reference and Link best model, also built Confusion Matrix and checked Variance in prediction quality.

References

- [1]<https://towardsdatascience.com/house-price-prediction-with-zillow-economics-dataset-18709abff896>
- [2]<https://yalantis.com/blog/predictive-algorithm-for-house-price/>
- [3]https://www.researchgate.net/publication/342782159_Prediction_of_real_estate_prices_with_data_mining_algorithms