CSCI 5455: Data Mining

Mahsa Rahimian
College of Engineering, Design & Computing
The University of Colorado Denver
Colorado, USA.
mahsa.rahimian@ucdenver.edu

Rumana Sultana
College of Engineering, Design & Computing
The University of Colorado Denver
Colorado, USA.
rumana.sultana@ucdenver.edu

Project Title— House Price prediction using Data Mining

.

Role:

Mahsa Rahimian: Defining Problem Statement and Background/ data collecting/ studying methododology/ defining lessons learned/ presentation and report preparing

Rumana Sultana: Planning appropriate tools to use/ data analysis and preprocessing /Project coding and implementation/ results analysis/ finalizing lessons learned/ presentation and report preparing

I. Problem Statement and Background

Nowadays, everything is shifting to automated systems and Real estate industry needs to shift to automated system as well. The present method is that investors approach real estate agents and based on agent's suggestion they start their investments. This method can be very risky to customers since there is a high probable that agent predicts wrong estates. Accordingly, this industry needs an automated and updated system. Data mining methods and algorithms can assist investors to invest in a proper estate based on their priorities and requirements.

Our goal in this project is to predict efficient house pricing for investors based on their priorities and budgets. Our project's strategy is based on analyzing previous market trends, so upcoming developments for future prices will be predictable more precisely. In this paper, we will use different data mining strategies and linear regression algorithms to predict prices by analyzing current house prices, and then predict the future prices.

We will use Zillow housing price dataset as our data source. Zillow is one of the largest online real estate market company in the United State. The dataset has data on monthly housing prices for zip code or metro or state wise of the USA. Those data should be enough to reach our goal.

The informal success measures that I can mention in this section is since we are using linear regression algorithm, we can help to fulfill customers by increasing the accuracy of estate choice and reducing the risk of investing in an estate. By using this algorithm, there are some ways to add some features to make the system more widely acceptable. In-depth details of every property will be added to provide ample details of a desired estate. This will help the system to run on a larger level. More items such as recession that affect the house prices can provide results with more accuracy.

This project can impact anybody who is looking for house, real estate's companies, and those who wants to invest on proper properties. It can broadly work on two important phases which is ranking a group of customers who are looking for ideal area, and at the same time predict the most suitable area according to their interest and requirement. There are a lot of relative works that are using linear regression. It helps establishes the relationship strength between dependent variable and another changing independent variable known as label attribute and regular attribute respectively. Regression displays continuous value of the dependent variable i.e. label attribute that is used for prediction.

II. The Data Source(s) Used

A. Dataset

We used Zillow Home Value Index (ZHVI) [1] dataset that is a smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type. It reflects the typical value for homes in the 35th to 65th percentile range.The dataset has different type of time series housing price data per month from january 2000 to October 2021. Some characteristics of the data set is given below:

- **Downloaded Link:** https://www.zillow.com/research/data/
- **Data and Annotation Files:**

a. HOME VALUES

-Zip code based House data: Data information consist of housing prices for every zip code (which are 30480 zip codes total) from January 2000 to October 2021 for every month (262 months total).

-Metro based House data: Data information consist of housing prices for every metropolitan name (which are

908 metro name total) from January 2000 to October 2021 for every month (262 months total).

-State based House data: Data information consist of housing prices for every state name (which are 51 state name total) from January 2000 to October 2021 for every month (262 months total).

-State based House data: Data information consist of housing prices for every state name (which are 51 state name total) from January 2000 to October 2021 for every month (262 months total).

b. HOME VALUES FORECASTS

-Foretasted House data: The Zillow Home Value Forecast (ZHVF) is the one-year forecast of the Zillow Home Values Index (ZHVI), which is above. ZHVF is created using the all homes, mid-tier cut of ZHVI and is available both raw and smoothed and seasonally adjusted.

III. Methodology and data analysis

Linear Regression is the best way to model our dataset and the reason is because the data follow a highly linear relationship - all we have to do is select features that represent that linear relationship best.

In this project the first step that we are performing is data pre-processing. The Data information consist of housing prices for every zip code/Metro/State zip/metro/state in rows and month in columns and finally every cell contain a house price. These variables, which served as features of the dataset, were then used to predict the the price per month of each region.

A. Data Visualization

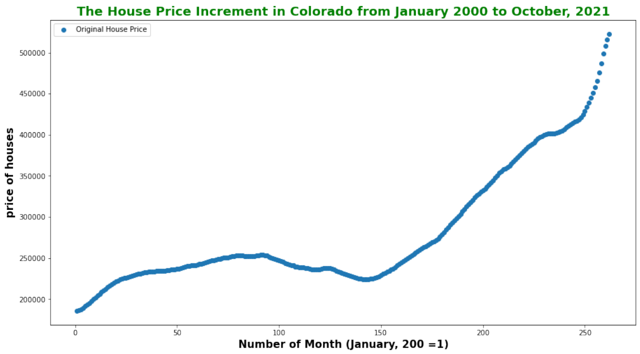


Figure 1: Increment in house price in Colorado from January 2000 to now

The next step is exploratory data analysis which is an essential step before building our regression model. In this step we can discover the implicit patterns of the data, which in turn helps choose appropriate machine learning approaches. In this section, we can determine our data based on some of the highlighted features. Figure-1 features houses price increases in Colorado during the period of January 2000- October 2021. As you can see in this picture, the prices of house are increasing almost linearly that is helpful to predict the future house price that should be higher than older price. Moreover, all the sates or zipcode doesn't contain same price or price increment for houses. That means there are states or zipcode that are more expensive and less expensive that is clearly visible in Figure-2.

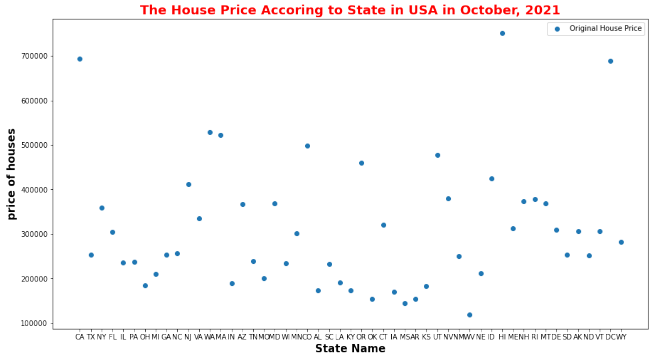


Figure 2: Statewide plot of house price on October, 2021

Finally, we have plotted all the metro based house price data from our dataset on Figure-3.

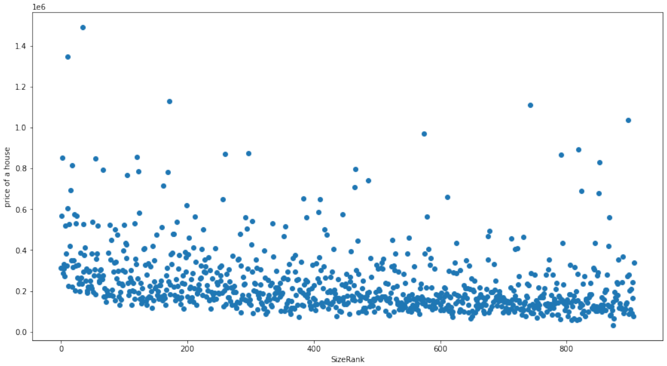


Figure 3: Scatter plot of house price based on Metro

The next step is the model selection. Before building models, the data should be processed accordingly so that the models could learn the patterns more efficiently. Specifically, numerical values were standardized, while categorical values were one-hot-encoded. We have visualized our dataset using data mining tools to decide the best fitted model for our dataset. First we visualize the Pearson correlation for all features with the average price between January 2000 and October 2021 that is shown in Figure-4. A distribution plot of house price is visualized on Figure-5. As the number of samples for zip code and metro data, we have visualized a box plot for statewide data based on average house price of all metropolitan in Figure-6. This data visualization helps us to understand which strength of data selection will be more accurate. We decide to fit our data not in average price rather with the all monthly prices provided into the dataset. That is why we proceeded to visualizing the Pearson correlation of all the months house prices in the dataset and the last 24 months house prices that are shown in Figure-7 and Figure-8. Then we proceeded to data preparation and development of our model based on these data visualizations and features characteristics.

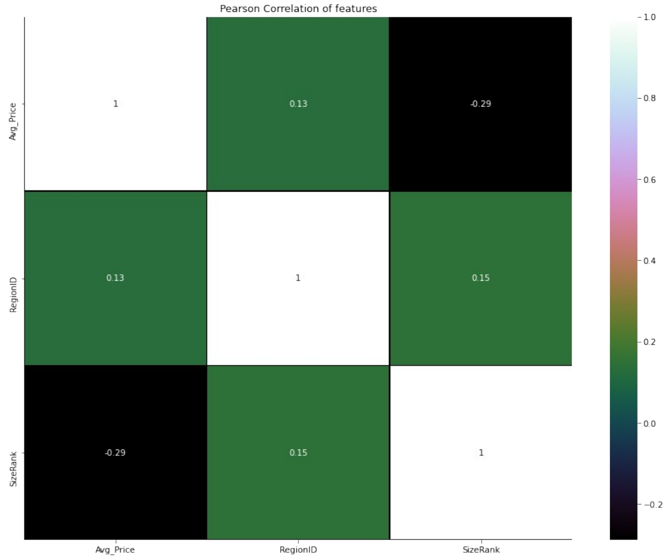


Figure 4: Pearson Correlation between features with average price (Metro)

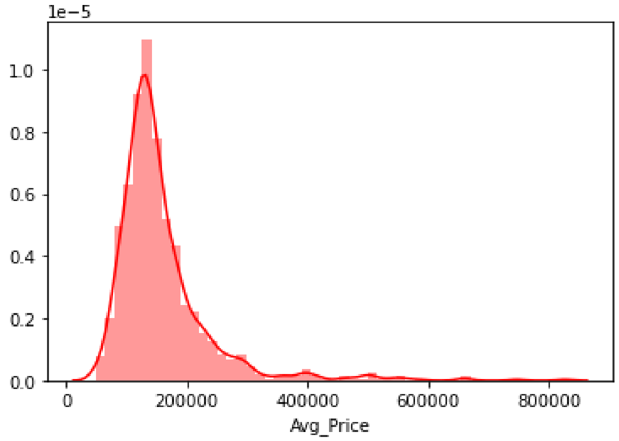


Figure 5: Distribution plot on average price in Metro data

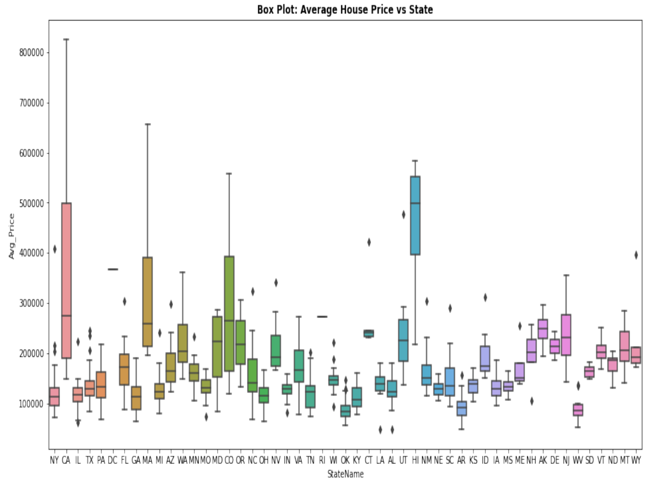


Figure 6: Box plot of average house price of metropolitan data (Statewide)

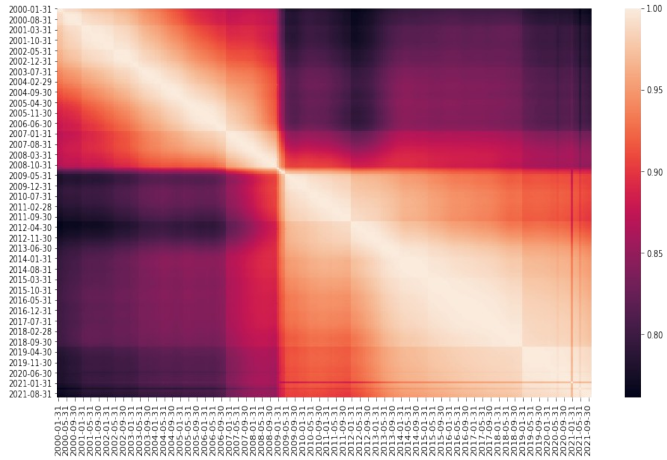


Figure 7: Correlation between all months from 2000 to 2021

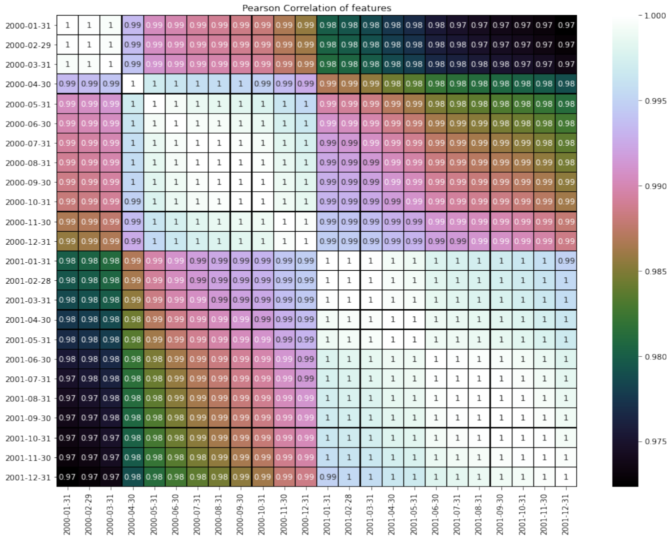


Figure 8: The last 24 months house prices correlation

B. **Data Preparation**

We have prepared our dataset from the original raw data that contains 30480 rows for zipcode-based dataset, 908 rows or metro-based dataset and 51 rows for State-based data-

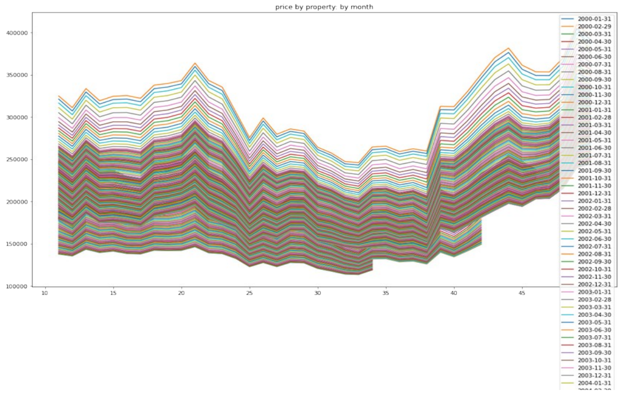
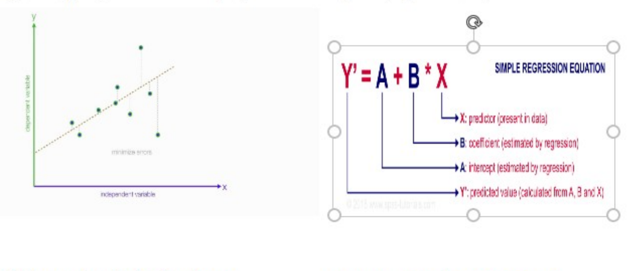


Figure 9: All months' price plot after data preparation (State based)

set and 271, 267 and 267 columns for consequently.   We kept only the suitable identification id and the house prices for all month. Just for an example, we have shown the plot of our prepared dataset in Figure-9 for the statewise dataset for clear visibility.

C. Model Development and Used Techniques

Now the question is: how can we predict the future price of the house based on a region? Our data mining solution came by using the linear regression model. Linear regression is a most well established and easy-understood algorithm in statistics and data science. It is a linear model that assumes a linear relationship between a dependent continuous variable(Y) with one or more independent variables (X). Linear regression models predict values within a continuous range (e.g., sales, price) instead of classifying them into categories (e.g., home, car).


y
dependent variable
minimize errors
independent variable
x
Y' = A + B * X
SIMPLE REGRESSION EQUATION
X: predictor (present in data)
B: coefficient (estimated by regression)
A: intercept (estimated by regression)
Y': predicted value (calculated from A, B and X)
© 2015 www.spss-tutorials.com

[4] Source: http://mybooksucks.com

[4] Source: https://spss-tutorials.com

Figure 10: Linear regression model

There are two types of linear regression model- simple linear regression and multivariable regression. As we are using Zillow housing price dataset that has every month price of a region(zip code/metro/state) from January 2000 to October 2021, we found from our data analysis that the house price values have a linear increasing relationship from month to month. This strategy motivates us to use simple linear regression to solve our problem. We used python and other necessary libraries for data preprocessing like importing dataset, handling missing data, splitting dataset into training, and testing or scaling of data and for building training and testing the linear regression model. We used Anaconda navigator and Spyder or Jupyter notebook for coding.

D. Algorithm and Primary Analysis of the Algorithm

The linear regression model algorithm for our project that includes the following steps:

1. Calculate x_mean and y_mean.
2. Calculate the difference between (x[i], x_mean) and (y[i],y_mean)
3. Calculate the square of the difference between (x[i], x_mean) and sum
4. Calculate the products of the difference between (x[i], x_mean) and (y[i],y_mean) and sum
5. Calculate the coefficient, B by the equation B =(diffx * diffy)/(diffx *diffx)
6. Calculate the intercept A by the equation A=y_mean-(B* x_mean)
7. Finally calculate the prediction by the equation: Y_pred=B*x+A and draw the regression line.

To analyze primarily, we took the first 31 months (about 2 and a half years) data for Colorado from the state-wise dataset of Zillow housing price dataset and followed the above algorithm to draw the regression line. We got a promising result for our analysis that is shown in Graph (Figure 11). The x-axis represents the months, and the y-axis represents the price of house w.r.t. month. The red line indicates the actual prices whereas the green line indicates the predicted prices. It seems that the prediction error is exceptionally low with the linear regression model. From our hand calculation, we got B(coefficient/slope) =654.3, A(Intercept)=140905.11. For example, to predict the 35th month house price, the generated Y=654.3 * (35) + 140905.11 = 163805.54 which is near to the actual value of 165233 in dataset.

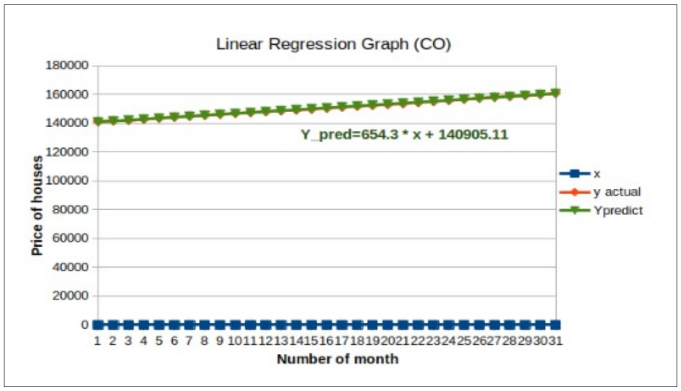


Figure 11: Graph of preliminary hand calculation analysis

The preliminary analysis is done with a small dataset. However, it is highly relevant to our project as it helps us to choose the algorithm and evaluate the possible outcome and proceed to final implementation. The final project is implemented with sufficient data, and we achieved more accurate result with the huge dataset.

IV. Implementation and Results

The main method we used closed form linear regression from sklearn libraries. Before fitting the data to the model, we have set our x-train and y-train dataset. In our case, we made the month number as a x-train dataset and the house price is as y-train dataset. We didn’t split the dataset into test and train because we want to predict the price of the house for a upcoming date. So, any upcoming date is our test data. For one date it will produce all the predicted house price for every zipcode or metro or state. We calculate the passing monthe between two dates taking the base 0ctober 2000 and represent all months with a number. In case of testing we apply the same method, too. We fitted a whole month data of all the region to the regression model and defined the coefficient, intercept and prediction price. Finally we used these coefficient and intercept to predict house price for an unseen date. We got around 45% accuracy for linear regression model for our dataset. We also used gradientboostingregreesion model in our dataset and we got 99% above accuracy for this model with the training dataset. However, The testing dataset didn't show a promising result with this model that needs more works for perfection. In Figure-12 we shown the result we got from linear regression model prediction with metro-based training dataset.

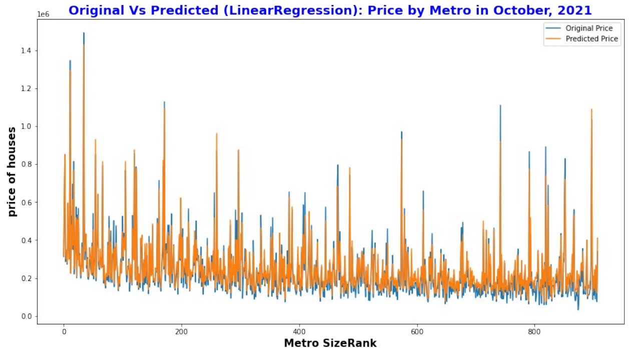


Figure 12: Original vs. Predicted result on Metro-based data.

With the same dataset, we got unexpectedly well result for the model gradientboostingregreesion that is shown in Figure-13.

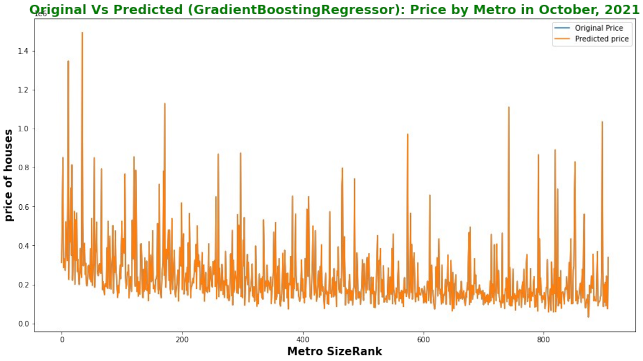


Figure 13: Original vs. Predicted result GradientBoostingregression on metro-based data

Then we have listed the most expensive five metropolitan and less expensive five metropolitan and shown in a bar plot based on the gradientboostingregreesion model result in Figure-14 and Figure-15.

Table 1: Most and Less expensive house price

Most Expensive		Less Expensive	
Predicted	**Original**	**Predicted**	**Original**
1493019.88	1493020.00	31775.00	31775.00
1346994.95	1346995.00	58553.00	58553.00
1128825.92	1128826.00	58558.00	58558.00
1110212.99	1110213.00	59964.00	59964.00
1035467.98	1035468.00	62956.99	62957.00

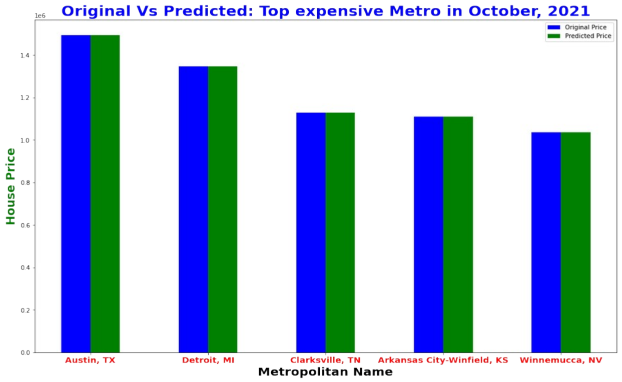


Figure 14: Most expensive metropolitan (predicted vs. original)

The metro name we got for most expensive house price are : Austin, TX, Detroit, MI, Clarksville, TN, Arkansas City Winfield, KS and Winnemucca, NV. The metro name we got for less expensive house price are : Parsons, KS, Van Wert, OH, Fort Morgan, CO, Bainbridge, GA and Jackson, WY. The less expensive metro house price is shown in Figure-15.

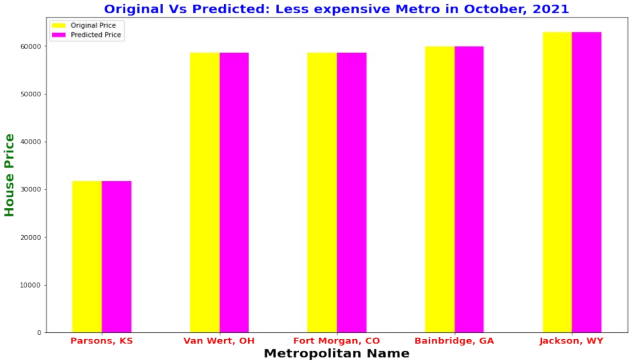


Figure 15: Less expensive metropolitan name (predicted vs. original)

Then we have taken the foretasted dataset from Zillow that has the price change factors on the last year's listed house price. We used this dataset to calculate the foretasted house price for October 2022. We then test our model with the date '2022-10-31' to predict the house price of October 2022. Finally we compare our result with the foretasted dataset that is shown in a plot Figure-16. We plot the most expensive

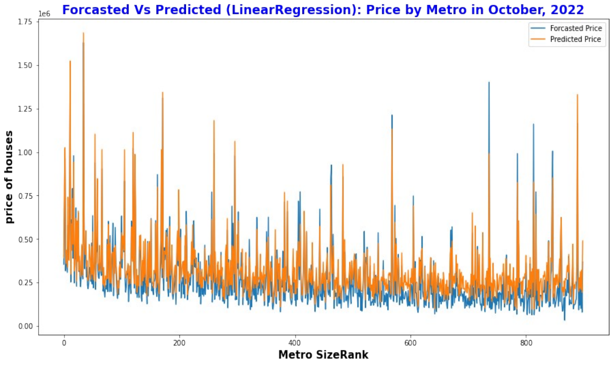


Figure 16: Predicted vs. Original house price on October 2022

metro for the result on Figure-17.

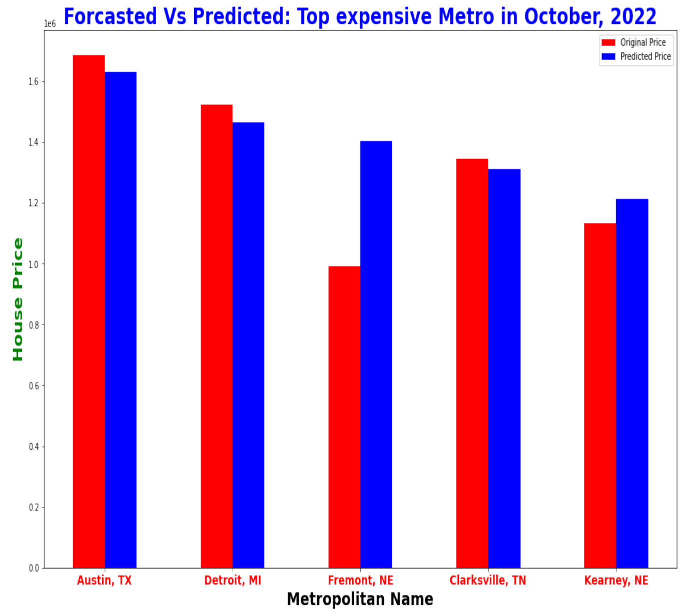
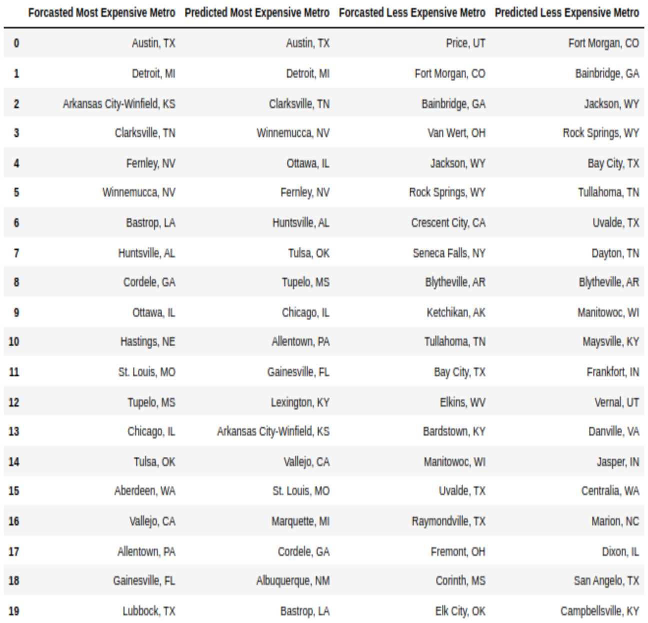


Figure 17: Top 5 metro house price on October, 2022

We have create a chart in Figure 18 with the top 20 metro both for higher and lower in price that we got from our prediction and foretasted price by Zillow for a comparison. We found both the list almost contain the similar metro name.



	Forcasted Most Expensive Metro	Predicted Most Expensive Metro	Forcasted Less Expensive Metro	Predicted Less Expensive Metro
0	Austin, TX	Austin, TX	Price, UT	Fort Morgan, CO
1	Detroit, MI	Detroit, MI	Fort Morgan, CO	Bainbridge, GA
2	Arkansas City-Winfield, KS	Clarksville, TN	Bainbridge, GA	Jackson, WY
3	Clarksville, TN	Winnemucca, NV	Van Wert, OH	Rock Springs, WY
4	Fernley, NV	Ottawa, IL	Jackson, WY	Bay City, TX
5	Winnemucca, NV	Fernley, NV	Rock Springs, WY	Tullahoma, TN
6	Bastrop, LA	Huntsville, AL	Crescent City, CA	Uvalde, TX
7	Huntsville, AL	Tulsa, OK	Seneca Falls, NY	Dayton, TN
8	Cordele, GA	Tupelo, MS	Blytheville, AR	Blytheville, AR
9	Ottawa, IL	Chicago, IL	Ketchikan, AK	Manitowoc, WI
10	Hastings, NE	Allentown, PA	Tullahoma, TN	Maysville, KY
11	St. Louis, MO	Gainesville, FL	Bay City, TX	Frankfort, IN
12	Tupelo, MS	Lexington, KY	Elkins, WV	Vernal, UT
13	Chicago, IL	Arkansas City-Winfield, KS	Bardstown, KY	Danville, VA
14	Tulsa, OK	Vallejo, CA	Manitowoc, WI	Jasper, IN
15	Aberdeen, WA	St. Louis, MO	Uvalde, TX	Centralia, WA
16	Vallejo, CA	Marquette, MI	Raymondville, TX	Marion, NC
17	Allentown, PA	Cordele, GA	Fremont, OH	Dixon, IL
18	Gainesville, FL	Albuquerque, NM	Corinth, MS	San Angelo, TX
19	Lubbock, TX	Bastrop, LA	Elk City, OK	Campbellsville, KY

Figure 18: List of Top20 metro name from prediction and forecasting

Finally, we have plotted the top 20 metro from our prediction compared with top 20 metro from Zillow foretasted metro that is shown in Figure-19. Moreover, we also plotted the bottom 20 metro from our prediction and Zillow forcast in Figure-20.

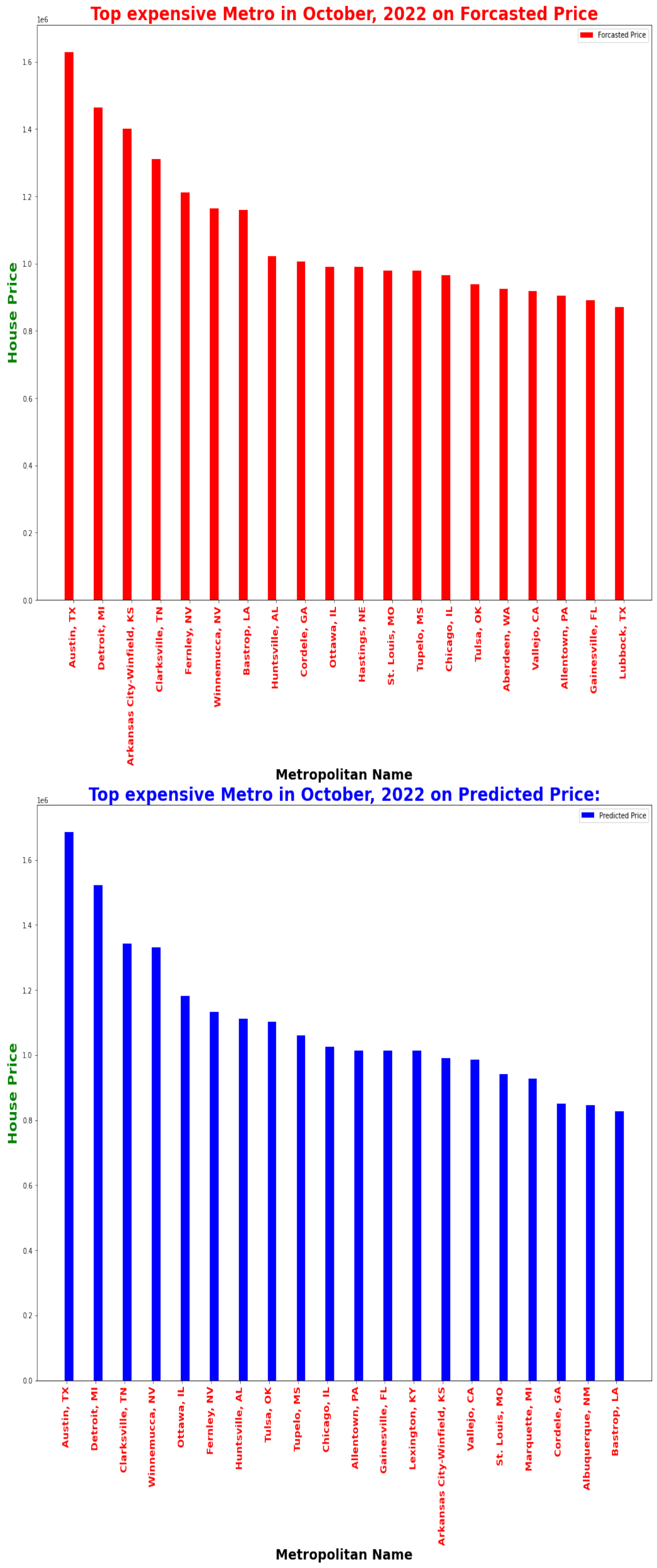


Figure 19: Top20 metro from prediction and forecasting

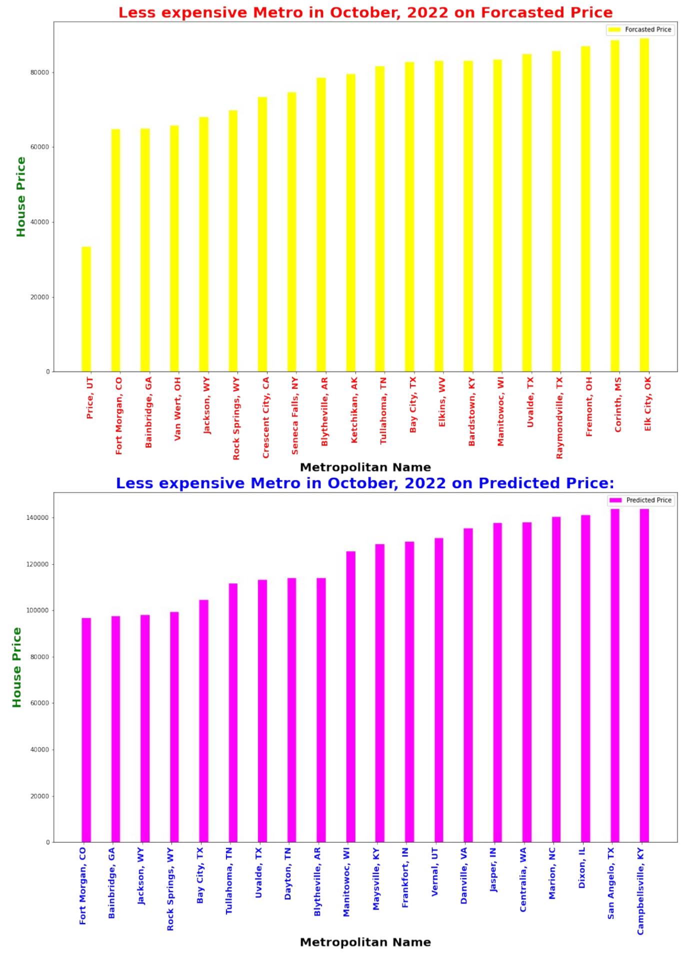


Figure 20: Bottom20 metro from prediction and forecasting

V. Description of Tools is Using

The tools we used through our project are Pandas (for data preparation), Python3, matplotlib and Linear Regression and GradientboostingRegression for network training and evaluation, NumPy, SciPy, sklearn, h5py, etc. The reason of choosing the tools is:

- **MATplotlib** has user-friendly plotting toolbox for data visualizations
- **Python3 with Jupyter Notebook** is very enriched with machine and deep learning libraries.
- **NumPy** makes easy complex machine and deep learning numerical operations with large dataset.
- **SciPy** contains different modules for optimization, linear algebra, integration and statistics that is very helpful for data analysis.
- **Scikit-learn**, a machine learning library for Python has various algorithms like support vector machine, random forests, and k-neighbors, and it supports Python numerical and scientific libraries like NumPy and SciPy.

We used other necessary libraries as per our requirement. We have used the easier and complex libraries that can help us to understand the data set and how to process a big data set using available tools.

VI. Lesson learned

We have gained many knowledge by doing this project. First of all I come to familiar with housing data set of time series and how to process it to use in a regression model. We have learned how to use the available tools to build model. We have learned data pre-processing, joining, merging, deleting, missing values handling and many other techniques. We have learned many important data mining techniques along with median, mean, binning, data warehouse, frequent pattern mining, and other mining learning from our course content. We also got familiar with visualization of a dataset in many aspects of plotting like scatter plot, line plot, bar chart, Pearson correlation plot, distribution plot and others. The project is very interesting. I will continue the project to get a final outcome.

VII. Future work and challenges

First of all our priority future work is to fine tune the trained model to achieve more accurate output. Currently, we got only 44% accuracy with our linear model and the prediction is below the original price. That is the main challenge we are feeling to work more. In addition, our 2nd model didn't work good for unseen data. Hence we want to find out the reason of this and fix the model in accordance to new testing data. We will work to develop it and modify it in future for my own interest.

VIII.Conclusion

A home is the heaven for a man when it is comfortable, reasonable and with a beautiful surrounding. An home and outside environment can increase mental and physical health with happiness. There is not doubt that finding a proper housing is highly impotence for human life. Therefore, we set our goal of Data Mining Solution is to be able to predict with some accuracy (currently 45% for linear model) the future home prices for potential investors. By using historical data, along with current home selling prices, we can predict future market trends and provide investors with information they need to make decisions that align with their goals.

References

[1] "Housing data," Zillow Research, 25-Mar-2021. [Online]. Available: https://www.zillow.com/research/data/. [Accessed: 14-Sep-2021].

[2] N. Bhagat, A. Mohokar, and S. Mane, "House price forecasting using data mining," International Journal of Computer Applications, vol. 152, no. 2, pp. 23–26, 2016.

[3] V. Valkov, "Predicting house prices with linear Regression: Machine learning from SCRATCH (PART II)," Medium, 05-

Jul-2019. [Online]. Available: https://towardsdatascience.com/predicting-house-prices-with-linear-regression-machine-learning-from-scratch-part-ii-47a0238aeac1. [Accessed: 14-Sep-2021].

[5] T. Gupta, “Data preprocessing in Python,”Medium, 25-Dec-2020. [Online]. Available: https://towardsdatascience.com/data-preprocessing-in-python-b52b652e37d5. [Accessed: 14-Sep-2021].

[6] Online source. Link: https://github.com/jaskirat111/Housing-Price-Prediction-using-Advanced-ML-Algorithms

[7] Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, *174*, 433-442.