Enkhchimeg Tsendnyam

**Housing Price Prediction**

***Intro/background of the problem***

Buying a house can be one of the most important but daunting decisions in life . It is considered a key aspect of the American dream. There are so many factors that one has to consider before making a big decision such as buying a house. Depending on what the person’s or family’s financial situation is, how many kids and pets they have, what their interests are, they might plan on buying a house with larger number of bedrooms, bathrooms, large backyard and a waterfront in a location they want. Some people might prefer a small house with low maintenance, or a condo in a city center. Retirees might be looking at getting away from the city life and moving to the suburbs or country for more space and relaxed atmosphere. Some people might be looking to buy a vacation or a second home. Every home buyer’s situation, wants and needs are different. A list of preferences & an ability to afford a house can greatly affect the buyer’s choice and decision on buying a home.

Real estate market often follows the cycles of economy and there are few key factors that impact the housing prices, availability and investment potential. Some of the key factors that affect the housing prices include demographics, interest rates, government policies and health of the economy. For example, millennials who were born between 1980 and 1997 are an example of a demographic trend and can significantly impact the real estate market. According to the 2019 Zillow real estate market report, an average age of first-time homebuyer in the U.S. was 34. If that number stays the same in 2021, considering that an average millennial turning 34 this this year we can also guess that majority of the first-time buyers will be the average 34 your-old millennials for 2021. Interest rate also has a major impact on the real estate market. It can create higher or lower demand. If the interest rates are low, more people will qualify to obtain mortgages which can create higher demand and higher prices. Government legislations such as tax credits, deductions and subsidies can have a sizable impact on the property demand and prices for at least for a short-term. Furthermore, the overall health of the economy and its indicators such as the gross domestic product (GDP), employment data, manufacturing sectors, prices of goods & services, inflation can have impact on the real estate market. These are some of the external factors.

So, one might ask, what about other factors such as what the house itself has to offer? How much of an impact does things like location, number of the bedrooms, bathrooms, square footage, lay of the house have impact on the housing price? Have you thought about being able to predict the housing price for certain locations, zip codes while considering your needs and wants? It certainly is feasible based on a historical housing market data. What are some of the biggest factors that impact the housing prices? One might say location and location and location. We all know that a 3-bedroom condo in New York city is going to cost much higher than a 3-bedroom condo in Bloomington, Indiana. Conventional wisdom says as the houses get bigger, their cost goes up. Is that really true? How much of a correlation is there between the size of the house and the cost of the house? Do other factors such as view, waterfront and size of the lot have any impact on the housing cost? Do houses that had renovation sell for higher prices? Do younger houses sell for higher prices than older houses? To answer some of these questions, I looked at the 2014 Washington Housing dataset from Kaggle and did a thorough cleaning, exploratory data analysis & prediction. The housing price will be the target variable for the prediction analysis.

***Exploratory Data Analysis(EDA)***

The 2014 Washington housing dataset included samples of the houses that were on the market for sale in Washington in 2014. It includes information such as housing price, number of bedrooms, number of bathrooms, living square feet, lot square feet, basement square feet, number of floors, waterfront info, view ratings, year the houses were built, year the houses were renovated, and city and zip code of the houses. Below are the variables and what they describe:

* Date: Date when the house is ready for sale.
* Price: Price of the house to be sold.
* Bedrooms: No. of bedrooms in the house.
* Bathrooms: No. of bathrooms in the house.
* Sqft\_living: Squarefoot of Living in the house.
* Sqft\_lot: Squarefoot of Floor in the house.
* Floors: Floors on which living area located.
* Waterfront: If waterfront available in front of house.
* View: View from the house.
* Condition: Condition of the house.
* Sqft\_above: Squarefoot above is the space available at roof.
* Sqft\_basement: Squarefoot basement is the space available at the basement.
* Yr\_built: In which year the house is built.
* Yr\_renovated: Year of renovation.
* Street: On which street house is located.
* City: City in which the country is located.
* Statezip: Zip code of the area in which house is located.
* Country: Country is US.

When the data was first loaded, it had 4600 rows and 18 rows. Through basic statistical analysis, I discovered that there are houses with $0 costs, houses with 0 bedrooms & houses with 0 bathrooms, which did not really make sense. The first data cleaning step included removing those data errors which were about total of 49 rows. The next major step included checking the data types, converting some quantitative variables that are float & object formats to integers. Two of the variable’s “date” & “statezip” had to be split, generating new features in the dataset. There were no missing or duplicate values in the dataset. There were several variables that had outliers detected with the help of outlier detection methods such as boxplot, z-score & IQR range. The largest outliers were removed from price, sqft\_living, sqft\_lot, & sqft\_above variables to help with better prediction accuracy results. After removing the outliers, I had 4526 rows in my dataset and my dataset was ready for EDA.

Through basic EDA & visualizations, I revealed average house that was being sold in Washington had 3-bedrooms, 2.5 bathrooms, 1 floor house with 2,114 square footages. Also, houses that were built after 2000s had the highest frequency compared to the houses that were built between 1900 and 2000. Average square foot of the floor was 13,698 and average square foot available at roof was 1,808. Majority of the houses in the dataset have 3 points in condition out of 5( 5 reflects the highest condition & 1 reflects the lowest condition). Although the average housing prices varied from one Washington city to next, average(mean) housing price was $541,000 in the Washington area. The standard deviation of the housing price was $331,000 which indicates the typical distance between Washington housing price tend to fall from the average Washington housing price. Looking at the housing stock in the dataset, over 30% of the houses that were being sold was in Seattle.

Looking at the correlations between all variables in the dataset, a few strong correlations stood out. Square foot above & square foot living had a strong positive correlation of 0.87. This totally makes sense and it would not be coincidence as more square foot living a house has, a more square foot space roof will have available. The next positive correlation was between price & square feet living with a positive correlation of 0.68. This correlation validates the general conception that everyone has, the more square foot of living space there is, the higher the house costs. Lastly, square foot above & number of bathrooms had a positive correlation of 0.68. In our dataset, square foot above reflects the amount of space available at the roof and there is not really a good explanation on why these two variables would be positively correlated. Therefore, this last correlation example is a mere coincidence and true testament to “correlation does not always mean causation”. The highest negative correlation was -0.4 and it was between year the houses were built & the condition of the house. Although it is not a strong negative relationship, it supports the general assumption that the older the house is, the worse the condition will be.

***Methods***

Prior to do doing prediction analysis, the preprocessing of the Washing housing dataset included encoding the categorical variables such as “city” and “street” to quantitative variables. By using dimensionality reduction, I removed some of the unnecessary variables such as “state”, “country” and “date” from the dataset. Then the dataset was split into training & test with 70:30 ratio. As previously mentioned, the target variable that is being predicted is a quantitative variable housing “price”. Therefore, a number of regression supervised learning machine learning models were used to predict the housing prices. Machine Learning models that were used to predict the housing prices in Washington include:

* Linear Regression
* Gradient Boosting Regressor
* Random Forest Regressor
* Ransac Regressor
* Decision Tree Regressor

The training data was normalized and fit to each model and housing prices were predicted using each model. The model evaluation metrics used for evaluating the regression results are R2, MSE, RSME & MAE. Since the business objective more aligned with predicting the housing prices, regression problem & its commonly used metrics were the most appropriate.

* R2 – coefficient of determination; proportion of the variance in the dependent variable that is predictable from the independent variables
* MSE – mean squared error; estimates the average squared difference between the estimated values and the actual value
* RSME – root mean squared error; measure of the differences between values predicted by the model
* MAE – mean absolute error; measure of error between paired observations expressing the same phenomenon

***Results***

Using 5 different machine learning models resulted in 5 different R2, MSE, RSME, & MAE numbers. When it comes to determining how well the model fits the dependent variables, Decision Tree Regressor model excelled with the highest R2 of 99%, followed by Random Forest Regressor Model with R2 of 95%. In other words, when Decision Tree Regressor model is deployed, 99% of our housing price can be explained by the variations in the dependent variables. However, when it comes to determining how close the forecasts are to actual values, Gradient Boosting Regressor did the best, with the lowest RSME of 187,551, followed by Random Forest Regressor with RSME of 189,891. Since RSME is an absolute measure of fit and describes how accurately the model predicts the housing prices, it is the most important criterion for this project.

Linear Regression: R2 = 57%; MSE= 50,899,518,034; RMSE = 225,609; MAE = 149,848

Gradient Boosting Regressor: R2 = 82%; MSE = 35,175,600,034; RSME = 187,551; MAE = 112848

Random Forest Regressor: R2 = 95%; MSE = 36,058,731,494; RSME = 189,891; MAE = 110,817

Ransac Regressor: R2 = 44%; MSE = 59,174,480,876; RSME = 243,258; MAE = 163,667

Decision Tree Regressor: R2 = 99%; MSE = 73,953,445,813; RSME = 271,943; MAE = 161,269

***Discussion/Conclusion/Executive Summary***

Have you wondered what truly affects housing prices other than outside factors such as economy, interest rates and government policies? Have you ever wondered what will be the average housing prices for your chosen location in the future? Is there a correlation between number of bedrooms a house has and its housing price? If so, the 2014 Washington housing dataset gives a unique perspective on housing prices for future home buyers. This dataset/report includes over 4000 houses that were being sold in the state of Washington in 2014. It also had many “house” related info such as year the house was built, location, size, number of bedrooms and bathrooms and conditions of the houses. Basic exploratory analysis on this dataset provided that average housing price in Washington was $541,000 in 2014 and over 30% houses that were being sold were in Seattle. As intuitive as it is, the size of the house & the price of the house had strong correlation. As the house gets bigger, the cost goes up. Also, as the older the house was, the worse the conditions of the house were. No other meaningful conclusive correlations were drawn. The best machine learning model for predicting houses were evaluated based on how accurately it forecasts the housing prices. As a result, 82% of the housing prices in Washington can be explained by changes in variables such as number of bedrooms, number of bathrooms, square foot of the house, square foot of the lot, number of floors, condition of the house, year the house was built, year the house was renovated, and whether the house has a waterfront or a view.

***Acknowledgements***

DSC 630 Predictive Analytics class was my 7th class that I took at Bellevue University as part of the M.S. in Data Science program. I was able to tie a lot of the concepts I’ve learned up to this point and demonstrate it on my final project on “Housing Price Prediction.” This dataset that Zillow provided through Kaggle was very interesting and offered a lot to anyone who is interested in analyzing housing prices. I would like to thank prof. Brett Werner for guiding me and providing meaningful insights throughout my project milestones. I also would like to thank to my classmates who provided peer review feedbacks and engaged in provocative class discussions.

***References:***

Shree, S. (2018, August 26). House price prediction. Retrieved April 20, 2021, from <https://www.kaggle.com/shree1992/housedata>

Nguyen, J. (2021, May 19). *4 Key Factors That Drive the Real Estate Market*. Investopedia. https://www.investopedia.com/articles/mortages-real-estate/11/factors-affecting-real-estate-market.asp.

Fontinelle, A. (2021, May 6). *First-Time Homebuyer's Guide*. Investopedia. <https://www.investopedia.com/updates/first-time-home-buyer/>.

Fitzgerald, R. (n.d.). *Everything You Need to Know About Millennial Home Buyers (2021)*. Raleigh, NC Real Estate - Raleigh Realty. https://www.raleighrealtyhomes.com/blog/everything-you-need-to-know-about-millennial-home-buyers-2019.html.

SnowBrains. (2021, January 11). Brain post: How much time does the average American Spend Outdoors? Retrieved April 18, 2021, from <https://snowbrains.com/brain-post-much-time-average-american-spend-outdoors/>

State of Homelessness: 2020 Edition. (2021, February 09). Retrieved April 18, 2021, from <https://endhomelessness.org/homelessness-in-america/homelessness-statistics/state-of-homelessness-2020/>

How housing instability impacts individual and FAMILY Well-Being: Hud user. (n.d.). Retrieved April 18, 2021, from <https://www.huduser.gov/portal/pdredge/pdr-edge-featd-article-020419.html>