Predictive Housing Price Model using Data Mining and Machine Learning

Dani Renteria, Jesus Cornejo, Runpeng Li
COMP 395 Data Mining - December 2023

ABSTRACT

In recent years, the real estate market has undergone a transformative shift with an increasing demand for precise housing price predictions. Property transactions are no longer solely influenced by traditional factors; instead, there is a need now for tailored pricing models that account for a diverse array of variables. Recognizing this new digital landscape, our project aims to fulfill this demand by developing a predictive model rooted in data mining and machine learning techniques.

INTRODUCTION

This study concentrates its efforts on King County, Washington, aiming to provide localized insights into the dynamic housing market of this region. The predictive model will harness clustering data mining techniques to sift through extensive housing datasets, identifying patterns, correlations, and influential factors that impact property prices. Machine learning algorithms will be employed to analyze historical data, enabling the model to adapt and learn from evolving market dynamics. By considering variables such as location, and property features, the model seeks to deliver accurate and personalized predictions.

1. Preprocessing

In the initial phase of our data preprocessing task, we focused on enhancing data quality. This involved addressing missing values, removing unnecessary attributes for streamlined relevance, and reformatting certain features for improved clarity and consistency. This meticulous cleanup forms a robust foundation for subsequent analyses, ensuring a comprehensive and refined dataset. With these changes we are well-equipped for advanced analytical tasks, and confident that our preprocessing efforts have fortified the data integrity, this will lead to more accurate models and insightful outcomes.

```
<bound method NDFrame.head of           id       date   price
0  7129300520  10-13-2014  221900  Affordable Price Range        3
1  6414100192  12-09-2014  538000     Premium Price Range        3
2  5631500400  10-25-2016  180000  Affordable Price Range        2
3  2487200875  12-09-2014  604000     Premium Price Range        4
4  1954400510  02-18-2015  510000     Premium Price Range        3>
```

Figure 1 - First 5 rows in the data set

2. Clustering Task 1 *(CT1 - Geographical)*

For this task, we used clustering to segment the houses in King County into different regions based on their geographic coordinates (longitude and latitude). The goal of this task was to understand the type of planning that may take place when deciding where to locate emergency services. For part 2 of this task, we created a new nominal attribute for price that would group houses into 5 distinct categories ***(see figure 2)***. Categories were created based on median income in the county and the most common mortgage type for homes.

Median income in the state of Washington = 82,400 USD

Recommended **percentage of income** devoted to mortgage (25%) = 20,600 USD per year

Average **30-year fixed rate** (Most common option for US home buyers) of **8.507%** for the state of Washington.

Low-Price Range = < 149,999

Affordable-Price Range = 150,000-249,999 (*Range = 100,000*)

High-Price Range = 250,000-449,999 (*Range = 200,000*)

Premium-Price Range = 450,000-799,999 (*Range = 350,000*)

Luxury-Price Range = >800,000

Figure 2 - New nominal attribute for price

Once the categories were made [this step was done through Weka] we created a visualization of the data along with their geographical location to uncover any patterns in price throughout the county.

2.2 Results and Discussion *(CT1)*

Emergency Service Location

As stated before, the aim of this clustering task was to analyze optimal emergency services placement within the county. Properly locating emergency services within a city is critical for quicker response times and effective crisis management.
Using a cluster analysis task, such as this one, can help city planners create more strategic placement which would ensure timely assistance, minimize casualties, and property damage.
We wanted each centroid within the cluster to have a maximum range of roughly two square miles. Of course this can be set to new updated values based on certain restrictions or requirements within a city.
We then ran the K-means clustering on the data set which produced ***figure 2.1***.

Future adaptations of this task can be used to plan a variety of things including amenities such as parks, restaurants, hotels and much more.

Price

This task aims to show a geographic visualization on price within King County. This is useful in visualizing which parts of the county have higher house price ranges. This data can be useful for homeowners and buyers who are looking to move or sell their homes. This task helps understand patterns within the county to help us analyze where in the county do we see a spike in home value. From our findings - ***figure 2.2 -*** we can see that the downton Seattle area typically sees homes in the Luxury price range (*>800,000 USD*) while cheaper and more affordable homes are typically located farther south.

This type of visualization can be adapted to uncover a multitude of patterns and trends.

Figures *(CT1)*

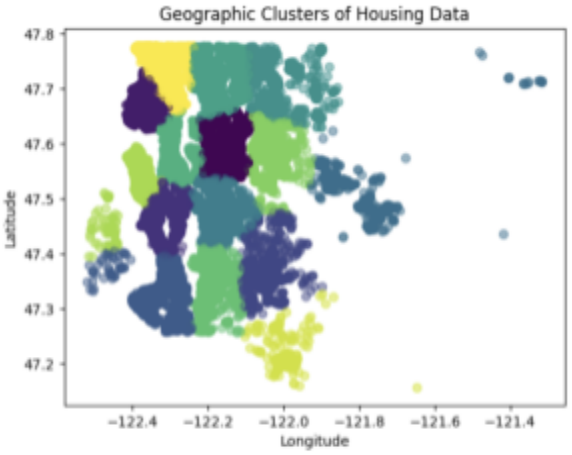

Figure 2.1 - Results from part 1 of CT1

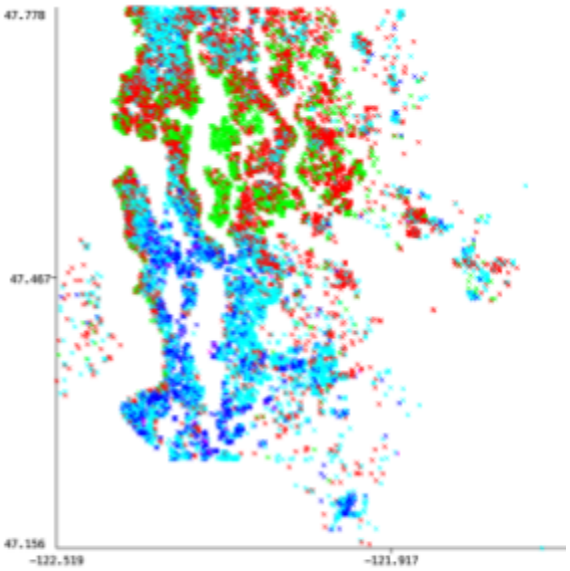

Figure 2.2 - Results from part 2 of CT1

Color legend

3. Clustering Task 2 *(CT2 - Characteristics)*

This second clustering task aimed to cluster the houses based on common characteristics. We used price(Nom), sqft_living, sqft_lot, condition, grade, and zip code as primary features shown in ***figure 3.1***. We

chose features that we believed would impact price the greatest. We used the K-means algorithm once more through Weka.

3.2 Results and Discussion *(CT2)*

Part of our process required choosing the optimal amount of clusters that would result in more distinct groupings, after trial and error our group decided that 3 produced the best results. From ***Figure 3.2,*** The results show that a majority of the houses in the area are of three main price categories (High, Premium, Luxury) demonstrating that this county is very expensive to live in.

The rest of our analysis attempted to figure out what attributes affect price the greatest. From this task we were able to uncover that size of the property, condition, and grade were the biggest factors contributing to high prices in the area.

Figures *(CT2)*

```
Scheme:       weka.clusterers.SimpleKMeans -init 0 -
Relation:     raw-data-DM
Instances:    21613
Attributes:   22
              price (Nom)
              sqft_living
              sqft_lot
              condition
              grade
              zipcode
Ignored:
              id
              date
              price
              bedrooms
              bathrooms
              floors
              waterfront
              view
              sqft_above
              sqft_basement
              yr_built
              yr_renovated
              lat
              long
              sqft_living15
              sqft_lot15
Test mode:    evaluate on training data
```

Figure 3.1 - Feature selection for clustering

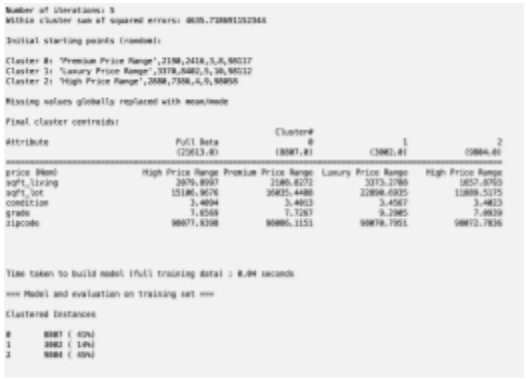

Figure 3.2 - Results from K-means clustering

4. Machine Learning Task *(ML)*

This task centers on leveraging housing data to train a predictive model for house prices in King County. By employing a Linear Regression approach, we aimed to discover patterns within the dataset and establish correlations between various housing features and their corresponding prices. Our objective was to develop a reliable model capable of accurately predicting house prices based on key attributes within King County, WA.

Results and Discussion *(ML)*

Our model uses a Python library called sklearn to create and train the linear regression model. The model learns the coefficients and the intercept of the line that best fits the training data. The coefficients in ***figure 4.1*** show how much each feature affects the price, and the intercept shows the base price when all the features are zero.

To interpret our coefficients, we must look at the values and signs. The values tell us how much the price changes when the attribute changes by one unit, holding all other attributes constant. The signs tell us whether the change is positive or negative. In our case, the sign tells us whether the price increases or decreases when the attribute increases.

For example, the coefficient for bedrooms is 3.73e-04, meaning that for each additional bedroom, the price decreases by 0.000373 times holding all other attributes constant. The coefficient for grade is 1.08e+05, which means that for every unit increase in

grade, the price increases by 108,000.00 dollars when holding all other attributes constant. Our study indicated that Condition and Grade are the biggest indicators when considering price. However, it's important to note that some of these results are unexpected and warrant further investigation to understand exactly why these have the effect they do. For example, as noted before, bedrooms and bathrooms have negative correlation with price, but using a basic understanding of real estate, we know that result doesn't really hold in the real world.

Additionally, as shown in ***figure 4.3***, we see the residuales of our model - a plot showing the difference between our results minus expected results. We see that overall there is a normal distribution centered at zero with some over estimation occurring.

Future adaptations of this model can improve on the training set in hopes of uncovering more accurate results.

Figures *(ML)*

Attribute	Coefficient
Bedrooms	-3.73e+04
Bathrooms	-1.23e+04
sqft_living	2.13e+02
sqft_lot	-2.54e-01
Floors	-1.98e+04
Condition	**6e+04**
Grade	**1.08e+05**
yr_renovated	8.13e+01
Zipcode	5.83e+02

Figure 4.1 - Regression coefficients

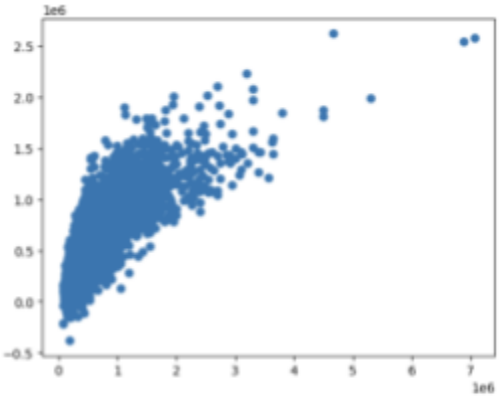

Figure 4.2 - Model predictions

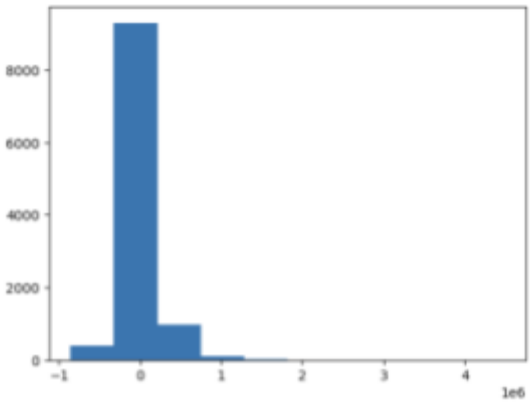

Figure 4.3 - Histogram of the residuals

5. Conclusion

Our study addresses the new digital landscape of the real estate market, responding to the recent demand for precise housing price predictions. Focused on King County, Washington, our data mining and machine learning-based predictive model goes beyond analyzing traditional influences on property prices. By analyzing extensive housing datasets with clustering techniques, we unveil patterns and influential factors that contribute to the constantly changing real estate market.

By considering location and property features, our approach offers custom, accurate predictions, reflecting a progressive shift toward data-driven strategies.