King County Housing Price Prediction

Muhammad Imran Sharif
Brian Belt
Carter Powell
Kuan Ruei Chiang

Agenda

- Introduction
- Analysis
- Results
- Summary
- Limitations
- Future Work

Introduction

- Housing prices have been on the rise and continue to rise, so knowing which features of housing contribute to higher pricing can impact homeowner decision making

- The primary goal of the project is to analyze the attributes of housing data to determine which parts of a property contribute to its price

Importance/Relevance

- Real Estate Industry
  - Accurate pricing of properties is essential for both buyers and sellers, and a predictive model can help them make more informed decisions
- Real Estate Agents
  - Real estate agents can use analysis to provide more accurate valuations to their clients. This can help them win more listings and improve their reputation as trustworthy and knowledgeable professionals
- Community Impact
  - Accurate housing price predictions can also have a positive impact on communities. By providing accurate valuations of properties, this analysis can help prevent underpricing or overpricing of properties

Questions

1. What aspect of the property brings value?

2. Do renovations have effect on property value?

3. What attributes when combined lead to the highest property value?

4. Why are view and grade important aspects in affecting price?

Approach

1. Preprocess any missing values and outliers that may skew the results of the data
2. Derive association rules based on relevant attributes
3. Classify data by price, grade, and zip code in order to get an understand of housing in the King County area
4. Perform clustering on data to determine which attributes contribute to property price similarity

Dataset

- Publicly available dataset from Kaggle for King County, Washington, USA
- Dating from May 2014 to May 2015
- The dataset contains ~21,600 instances and 21 attributes
- This dataset originated from the King County Department of Assessments, a government agency responsible for maintaining property records in King County, Washington

Attribute Descriptions

id:	Unique identifier for each property
date:	Date when the property was sold
price:	Sale price of the property in US dollars
bedrooms:	Number of bedrooms in the property
bathrooms:	Number of bathrooms in the property
sqft_living:	Square footage of the interior living space of the property
sqft_lot:	Square footage of the lot on which the property is built
floors:	Number of floors in the property
waterfront:	A binary variable indicating whether the property has a view of the waterfront or not

view:	An index from 0 to 4 of how good the view of the property is
condition:	An index from 1 to 5 of the condition of the property
grade:	An index from 1 to 13 of the overall grade given to the property based on King County grading system
sqft_above:	Square footage of the interior living space above ground level
sqft_basement:	Square footage of the interior living space below ground level
yr_built:	The year the property was built
yr_renovated:	The year when the property was last renovated (0 if never)
Zip code:	The zip code of the location of the property
lat:	Latitude of the location of the property
long:	Longitude of the location of the property

Agenda

- Introduction
- Analysis
- Results
- Summary
- Limitations
- Future Work

Preprocessing

- Missing Values
  - There were few missing values as the dataset was very complete
  - Example: “yr_renovated” was 0 if it had never been renovated
- Outliers
  - We visualized the distribution of the target variable (sale price) and identified any outliers. We then used various techniques to deal with outliers, such as removing them or transforming them to be within an acceptable range
  - Example: One house had a reported 33 rooms, and 1.5 baths

Association Rules

- On preliminary analysis
    - Setting Minimum Support to 0.85 and Minimum Confidence to 0.9
    - All attributes included in the association
    - Best rule found was connecting the worst viewing properties to those not on waterfront (Conf 1)
    - Most rules discovered connected view, waterfront and yr_renovated
    - For further analysis some of the data will need to be temporarily purged

Association Rules

- After further preprocessing
  - The Minimum Support is 0.1 and Minimum Confidence is 0.8
  - Removed the previously repeating attributes since the rules are self explanatory
  - More interesting results were shown in the new associations
  - If a house has 2.5 Bathrooms and 2 floors it most likely will not have a basement (Conf 0.9)
  - If a house has 1-1.75 Bathrooms it most likely has 1 floor (Conf 0.81)
- Price predictions based on association
  - Most price based associations do not have enough support to back them traditionally
  - With how different and specific real estate pricing is in our data set, to associate them we needed to condense the prices into similar price ranges

Association Rules

- Pricing Based Association Rules
  - Bedroom/Bathroom Counts
    - 0.8-2.4 Bathrooms typically under $837,500 (Conf 0.94)
    - 1.1-3.3 Bedrooms typically under $837,500 (Conf 0.96)
  - Square Footage
    - 820-1350 Square Feet typically under $380,000 (Conf 0.66)
    - 1880-2410 Square Feet typically between $380,000 and $685,000 (Conf 0.52)
  - View
    - Any house under $685,000 most likely will not have a good view (Conf 0.92)
    - Any house under $990,000 most likely will not be waterfront (Conf 0.99)

Data Classification

- To try to predict the quality of the houses
    - Classified by View
    - Classified by Grade

```
=== Summary ===

Correctly Classified Instances       19662               90.9899 %
Incorrectly Classified Instances      1947                9.0101 %
Kappa statistic                          0.1654
Mean absolute error                      0.0442
Root mean squared error                  0.1576
Relative absolute error                 60.0306 %
Root relative squared error             82.1819 %
Total Number of Instances            21609

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.999    0.893    0.911      0.999    0.953      0.302    0.947     0.994     0
                 0.012    0.000    0.333      0.012    0.023      0.061    0.979     0.334     1
                 0.088    0.001    0.766      0.088    0.158      0.251    0.950     0.444     2
                 0.104    0.001    0.779      0.104    0.184      0.280    0.977     0.484     3
                 0.147    0.000    0.922      0.147    0.254      0.366    0.993     0.643     4
Weighted Avg.    0.910    0.805    0.893      0.910    0.875      0.296    0.949     0.942

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 19473     3     6     4     1 |     a = 0
   325     4     2     1     0 |     b = 1
   870     3    85     4     0 |     c = 2
   440     1    12    53     3 |     d = 3
   259     1     6     6    47 |     e = 4
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     1
                 0.000    0.000    ?          0.000    ?          ?        1.000     0.230     3
                 0.276    0.000    1.000      0.276    0.432      0.525    0.998     0.673     4
                 0.281    0.000    0.907      0.281    0.429      0.502    0.995     0.787     5
                 0.775    0.032    0.717      0.775    0.745      0.718    0.975     0.832     6
                 0.846    0.115    0.839      0.846    0.842      0.730    0.942     0.921     7
                 0.783    0.078    0.797      0.783    0.790      0.709    0.947     0.888     8
                 0.815    0.034    0.767      0.815    0.790      0.761    0.975     0.891     9
                 0.844    0.006    0.883      0.844    0.863      0.856    0.993     0.938     10
                 0.829    0.001    0.943      0.829    0.882      0.882    0.998     0.960     11
                 0.833    0.000    0.893      0.833    0.862      0.862    1.000     0.954     12
                 0.692    0.000    1.000      0.692    0.818      0.832    0.999     0.853     13
Weighted Avg.    0.810    0.077    ?          0.810    ?          ?        0.955     0.899

=== Confusion Matrix ===

    a    b    c    d    e    f    g    h    i    j    k    l   <-- classified as
    1    0    0    0    0    0    0    0    0    0    0    0 |    a = 1
    0    0    0    2    1    0    0    0    0    0    0    0 |    b = 3
    0    0    8    1   20    0    0    0    0    0    0    0 |    c = 4
    0    0    0   68  136   38    0    0    0    0    0    0 |    d = 5
    0    0    0    0 1579  435   20    3    0    0    0    0 |    e = 6
    0    0    0    4  433 7595  751  172   17    4    4    0 |    f = 7
    0    0    0    0   29  907 4754  333   38    3    4    0 |    g = 8
    0    0    0    0    3   68  376 2131   30    6    1    0 |    h = 9
    0    0    0    0    0    8   61  107  957    1    0    0 |    i = 10
    0    0    0    0    0    0    6   30   32  329    0    0 |    j = 11
    0    0    0    0    0    0    0    3    8    4   75    0 |    k = 12
    0    0    0    0    0    0    0    0    2    2    0    9 |    l = 13
```

Data Classification

- Results from the classification
    - The View and Grade attributes were the most accurate to classify
    - The View ties in closely to the Zip_Code and Waterfront attributes and was 92% correct with NaiveBayes
    - The Grade relies heavily on multiple other attributes and is homeowner opinion based
        - These other attributes may include anything that a homeowner may seek in a property
        - Ex: Waterfront, Yr_Renovated, Bathrooms, Basement, Etc
        - Using Naive Bayes we were able to correctly classify 81% of the properties grades

Data Clustering

- K-Means Clustering
  - Results differed from our original thoughts
  - With the limited data in waterfront and view it makes sense with it not mattering as much
  - Neither did condition which was a major factor
  - Factors that contributed to a higher price in house includes
    - Number of bedrooms and bathrooms
    - Square feet of living space
    - Year property was built
    - Zip Code

```
kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 153129.0

Initial starting points (random):

Cluster 0: 450000,3,1.75,1540,9154,1,0,0,3,8,1983,0,98074
Cluster 1: 399000,2,1,1120,8661,1,0,0,3,7,1946,0,98125

Missing values globally replaced with mean/mode

Final cluster centroids:
                             Cluster#
Attribute       Full Data           0           1
                (21609.0)   (11177.0)   (10432.0)
=================================================
price              350000      450000      325000
bedrooms                3           4           3
bathrooms             2.5         2.5           1
sqft_living          1300        2100        1010
sqft_lot             5000        5000        5000
floors                  1           2           1
waterfront              0           0           0
view                    0           0           0
condition               3           3           3
grade                   7           8           7
yr_built             2014        2014        1968
yr_renovated            0           0           0
zipcode             98103       98052       98115
```

Data Clustering

- Farthest First
    - This clustering algorithm showed much more promise in what causes price discrepancy
    - More Bedrooms/Bathrooms, much more square footage, higher graded, new properties
    - The most interesting takeaway from this clustering algorithm stems from the zip code
        - 98004 was where more expensive properties are found (Right in the heart of Seattle)
        - 98198 was where cheaper properties are found (Further from downtown Seattle)

```
FarthestFirst
==============

Cluster centroids:

Cluster 0
        1965000.0 5.0 3.75 3940.0 1.5 0.0 3.0 9.0 1951.0 98004.0
Cluster 1
        570000.0 1.0 1.0 720.0 1.0 1.0 4.0 6.0 1905.0 98198.0
```

Agenda

- Introduction
- Analysis
- Results
- Summary
- Limitations
- Future Work

Results

- Association
  - If a house has 1 - 1.75 Bathrooms it most likely has 1 floor
  - Any house under $990,000 most likely will not be waterfront
- Classification
  - The View and Grade attributes were the most accurate to classify
  - Using Naive Bayes we were able to correctly classify 81% of the properties grades
- Clustering
  - Number of bedrooms, bathrooms, square footage, zip code, and year built had the most effect on price
  - Houses in the heart of Seattle were found to be more expensive, whereas the further you got out of the city, the lower the houses cost

Agenda

- Introduction
- Analysis
- Results
- Summary
- Limitations
- Future Work

Summary

- There are a lot of factors that influence the pricing of houses, in which most are homeowner opinion based
- Flooding or cultural building techniques may affect the statistics on basement representation in Seattle
- Housing in metropolitan areas is expensive (average of $540,000)
- Clustering gave more accurate and representative results than classification or association
- Renovations have the largest correlation on housing price

Agenda

- Introduction
- Analysis
- Results
- Summary
- Limitations
- Future Work

Limitations

- Data
  - The dataset contains data for a specific geographic area and time period. The results may not be generalizable to other locations or time periods
- Features
  - While the dataset contains several important features that affect housing prices, there may be other variables that are not included in the dataset, such as crime rates, access to public transportation, and nearby amenities. Adding these variables to the model could improve its accuracy
- Lack of domain expertise
  - The project does not incorporate domain expertise in the real estate industry, which could have helped to identify additional features that affect housing prices or to provide context for the results

Agenda

- Introduction
- Analysis
- Results
- Summary
- Limitations
- Future Work

Future Work

- Collect more data
    - Obtaining more data from other locations and time periods could help to develop a more comprehensive understanding of the factors that affect housing prices
- Incorporate additional features
    - Adding new features to the dataset such as crime rates, proximity to public transportation, and nearby amenities could help to improve accuracy
- Explore different algorithms
    - Linear Regression and other numerical analysis methods could be used in the future to obtain more accurate housing price prediction

Cited Works

- Kaggle
  - https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

Questions?