Logo

Description automatically generated with medium confidence

**ALY6040**

**Final Project : EDA**

Course: ALY6040 21402 Sec 02 Winter 2022 Data Mining Applications

Professor: Justin Grosz

Assignment: Module 4 Final Project Milestone 2 Draft Report

Submitted By

Saurabh Harne (harne.s@northeastern.edu)

Roshan Ambhore (ambhore.r@northeastern.edu)

Abhigna Ramamurthy (ramamurthy.a@northeastern.edu)

Benazeer Shabbir Shaikh (shaikh.b@northeastern.edu)

# **Introduction**

The Project covers analysis of the vital data set  “Seattle King County Housing”. The real estate property has always been a basic need for an individual. With the advent of increasing population, there arises a need to accurately understand the contributing factors, impacting variables responsible for the property pricing. As a result, we as the Analysts through this dataset are analysing the Seattle King County Housing dataset, because in real estate, the property's condition, location, and other qualitative factors largely affect the potential income of a property.

The Objective of the Project is to analyse and obtain deeper insights about the Seattle King County Housing dataset and accurately understand the factors, the contributing and impacting variables responsible for the property pricing in the King County. The data analysis will assist in understanding the preferences of the Customer Segment, their choices while buying Houses, the contributing and impacting variables responsible for the property pricing in the King County. This information will assist the Business Real Estate Agencies and Companies in understanding the most influential variables impacting the pricing of the Housing property, using this details the Company can also determine the Clients' needs, Propose Economical Solutions that suit them and accurately estimate the Future Property value.

**Exploratory Data Analysis**

We will start data profiling of the Raw data, the “Seattle King County Housing” dataset contains 21,613 observations with 21 attributes which has 15 Integer variables, 5 floating and 1 datetime type variable. We perform Data profiling, Data Cleansing and Exploratory Data Analysis in 2 step processes.

**Data Pre-processing**: In this part, we will perform Data Profiling and Data Cleansing to improve the Data Quality and overall productivity of the data set. During Data Profiling we find that there are no missing values in the dataset. However there exists duplicity of 177 Ids, on further investigation we find these Id’s are changes in prices of the same houses at a different point of time, therefore to maintain data consistency, and prevent data redundancy we drop the old records and maintain only the latest pricing records of the Housing dataset. Further, we see that the attribute “Date” has redundant data of ‘T000000’ associated with the time factor H:M:S ,therefore we perform formatting and cleansing of the date values in the respective Date, Month , Year and Day format. Furthermore, to determine the House Age of the property we perform feature engineering, the attribute “Age” of the building is calculated from difference between the year built and 2015(latest year of the buildings). We also created a categorical variable of bins of Ages such as '<1', '1-5', '6-10', '11-25', '26-50' ,'51-75', '76-100', and '>100' to determine the definite contributing and impacting Age Group for House pricing analysis. Similarly Prices of the houses are also binned into categories to understand the Price group distribution for the King County dataset. The Price Group includes categories such as 'upto 250k', 'upto 500k', 'upto 750k','upto 1mil','upto 2mil', and 'more than 2 mil'. To accurately understand Price per square feet of the Houses in the King County we calculate “Price per square feet” through price of the house divided by the square feet of the lot. Now we can say that our data is clean, but we will first have to explore the data to find hidden patterns that will be of great help to our analysis.

**Exploratory Data Analysis:** Further, in exploratory data analysis we will discover the relationships between variables, the distributions of the various features, and display some summary statistics of the data set. From the Boxplot, we check for outliers, we see that there are outliers for the variable 'Bedroom', the bedroom data has an outlier with 33 bedrooms, on investigating further the large number of bedrooms does not match the total square foot mentioned for the house, hence this erroneous entry is dropped. We also see that there are houses with zero bedrooms and bathrooms which can be considered as error values, because the common profile of houses in the dataset have at least 1 bedroom and 1 bathroom. As the data distribution for Bedroom and Bathroom is skewed we replace the 0 bedroom and bathroom values with their respective median values.

From the Histograms, we see the distribution of the numerical columns, the attribute Price is right skewed distribution, indicating that majority of the House prices are less than 1.3M with highest number of houses under category of 250K to 500K, this shows that Majority of the buyers prefer houses in this price range. The Houses have Average Age distribution between 26 to 50 years with absence of Waterfront and View feature, which entails that Waterfront and View is not a feature that many buyers consider before buying a home in King county. As most of the houses have an average distributions of 3-4 bedrooms and 1-2.5 bathrooms, we can infer that Buyer in king county have preference of houses with 3 to 4 bedrooms and 1 to 2.5 bathrooms. From the Bar plot it is evident that the age group of the house doesn't really have a positive or negative correlation with the price, the house that's on the lower price tend to be older in average. From the Barplot of Price\_Group with Basement square foot and Housing Grade, it is seen that Base square foot has positive effect on price along with grade value of the house, which means, bigger the basement or higher the overall grade of the house, higher are the House prices. From the Line chart, we see that House Price Sales in the month of April are the highest and gradually decreases over the next months. (Refer below fig 1 )

Chart, line chart

Description automatically generated

**Fig 1: Line chart of average sale price per month**

From the Bar plot for House prices with Zip Codes, we found that the Zip codes- 98039 (Medina), 98004 (Bellevue), 98040 (Mercer Island) are the top 3 zip codes with the “Highest Average Housing Prices” having House prices between 1.1 M to 2.18 M. On the Contrary the Zip codes- 98002 (Auburn), 98168 (Seattle), 98032 (Kent) as the top 3 zip codes with the “Lowest Average Housing Prices” having House prices between 230 – 250 K. The similarities persisting between top three Zip codes are that they are close to water bodies and surrounded by ample of companies however, the bottom three Zip codes are suburban areas away from the downtown. (Please refer below Fig 2)

From the Correlation plot, it is evident that high positive correlation exists between “Price” and “Square ft living”, followed by “grade”, “sqft\_above”, “sqft\_living15” therefore we can say that increase in the features results in proportional increase in the house prices. Month and day of the house is negatively correlated with house price. We also checked for Multicollinearity, through Variance Inflation Factor, the variables “Sqft above” resulted to highest intercorrelation with other variables with variance greater than 10. As the Multicollinear variables undermines the statistical significance of an independent the attribute “Sqft above” is dropped. The VIF values are recomputed and finally the Model without Multicollinearity (no variables having VIF greater than 10) is obtained.

In further analysis, we would like to explore data modelling techniques to accurately predict the influential and impacting variables that affect the prices of the houses.

Chart, bar chart

Description automatically generated

**Fig 2: Average house price by zip codes**

**Logistic Regression Modelling**

**Model Building:** We are using Logistic Regression classification model to predict the pricing of the houses using the dependent attribute price. The dependent variable Price is labelled “0” for Low price value houses (<$750,000) and “1” for High price value houses (>=$750,000). The dataset is available in .csv format and the model is built in Python. The model is aimed to assist the real estate business owners to understand the most significant factors contributing to pricing of the houses in King County. We split the dataset into two segments, with one segment having all the independent predictor variables and the other with the dependent response variable price. The dataset was further split into two parts, 80% of the original data for training and 20% of rest of the data for testing purposes. As the dataset has attributes with differing scales, we standardised the dataset before performing any modelling. The scaled data was then used for training and testing independent attributes. After splitting, we therefore go ahead to fit our Logistic Regression model using Logit function on the data, and assess the Logit Regression Model results, determine the impact coefficients, and accurately predict the House pricing significant driving attributes, which influences and contributes to the maximum High price value houses and Low price value houses.

**Model analysis and output:** The logistic regression model output displays significant variables which can be determined by the Pr(>|t|) , p-value, and also depicts the positive and negative impact over dependent variable price through the coefficients. The Significance of the variable is determined if the P-Value is less than the significance level ( 0.05 A significance level of 0.05 indicates a 5% risk) then it terms that the model fits the data well. The significant attributes obtained using Logit Regression Model are “Waterfront” and “Sqft living”. The attribute “Waterfront” has a coefficient of 4.10 interprets, closeness of the house to the waterfront increases the likelihood of the House price value being highly priced i.e. “(>=$750,000”. Secondly, the attribute “sqft\_living” has a coefficient of 0.9173 which indicates an increase in sqft living of the house increases the likelihood of the house price value being highly priced i.e “(>=$750,000”) by 91.73 percent in King County.

A confusion matrix, also known as an error matrix, is a summarized table used to assess the performance of a classification model. The number of correct and incorrect predictions are summarized with count values and broken down by each class. As seen in the Confusion Matrix result, our Model has an accuracy of 89.7%. Our model predicted that 378+104=482 houses to be high priced category when there were actually 378+339=717 houses as high priced category.

The Accuracy of the Model is 89.7%. We have got Precision of 78.4% it means when our model makes a prediction, it is precise 78.4% of time. In our prediction case, when our Logistic Regression model predicted that Houses with high “Sq ft living” will result in high prices of the houses “(>=$750,000”) in the Market it means 78.4% times the House values are highly priced. We also got 52.7% recall, which means if the house is of higher value the model can predict it 52.7% of the times. The mean squared error (MSE) is 0.103 it tells how close a regression line is to a set of points, as we know the lower the MSE, the better is the forecast.

**Decision Tree Model**

**Model building:** We are using Decision Tree Classification Model in order to predict the price category of the house in King County. The dependent variable price with labels of “0” for low price (<$750,000) and high price value (>=$750,000) classification is used. A decision tree classifier is a binary tree where predictions are made by traversing the tree from root to leaf at each node, we go left if a feature is less than a threshold, right otherwise. For Model Creation, we will split our dataset into two segments, with one segment having all the Independent or the Predictor variables together and the other segment with the Response or the Dependent Variable. Further the data segment will be split into two parts.ie. 80% training data and 20% testing data. After splitting, we therefore go ahead to fit our tree model on the data, and assess the tree Plot results, and accurately predict the House pricing significant driving attributes, which influences and contributes to the maximum High price value houses and Low price value houses in King County.

**Model analysis and output:** The significant variable obtained is “sqft\_living”, as it is the Root node of the model which implies that it is the best predictor variable. The Decision Tree visualization obtained reveals that “sqft\_living” resulted is Class 1 is high value price house category. This result is aligned with the output obtained in the logistic regression previously. Now we have fit(), to Predict in the Decision Tree is simply to follow the path in the constructed tree-shape decisions to the leaf node, and return the value of that node as we define in the fit() function.

With a maximum depth of 4 levels, we achieved a Decision Tree with Accuracy of 88.8%.

We have got Precision of 79.8% it means when our model makes a prediction, it is precise 79.8% of time. In our prediction case, when our Decision Tree model predicted that Houses with high “Sq ft living” will result in high prices of the houses “(>=$750,000”) in the Market it means 79.8% times the House values are highly priced. We also got 44.2% recall, which means if the house is of higher value the model can predict it 44.2% of the times. The mean squared error (MSE) is 0.112 it tells how close a regression line is to a set of points, as we know the lower the MSE, the better is the forecast. This can be noted that the MSE value in decision tree model is higher than that of logistic regression model. The accuracy, precision and recall have reduced in the decision tree when compared to logistic regression model which implies that the Decision tree model predicts less accurately than the Logistic Regression Model.

**Random Forest Model**

**Model building:** With a reduced performance of decision tree, it can be said that model might have overfitted the results to the training dataset. Hence considering an ensemble Decision Tree model like Random Forest would help in further investigation. The goal is to predict the price category of the houses in King County. The dependent variable price with labels of “0” for low price (<$750,000) and high price value (>=$750,000) classification is used. The model is split into segments of independent attributes and dependent variable price and further split into training and test data, similar to previously shown. With an n\_estimator of 5 the model results are obtained to determine the impact variables using feature importance by random forest and and accurately predict the House pricing significant driving attributes, which influences and contributes to the maximum High price value houses and Low price value houses in King County.

**Model analysis and output:** Random forests are a bit harder to decrypt to find significant variables, but we have used feature importance methodologies to understand which variables are consistently in each decision tree. We can see that the most important feature is “sqft living”, which was also identified as important predictor in Decision Tree and Logistic Regression. The model yielded an Accuracy of 88.9%, and Precision of 69.7%, which means when our model makes a prediction, it is precise 69.7% of the times. In our prediction case, when our Random Forest model predicted that Houses with high “Sq ft living” will result in high prices of the houses “(>=$750,000”) in the Market it means 79.8% times the House values are highly priced. The model has 59% recall, which means if the houses are high priced, in the test dataset, random forest model can identify 59% of the times. The mean squared error of 0.111 which shows how close a regression line is to a set of points. The results are very similar to that of decision tree. All the metrics indicates a lower performance than Logistic Regression hence can conclude that Random Forest Model predicts less accurately than the other Models.

**Benchmarking Metrics**

The benchmarking metrics obtained from the models are listed in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **MSE** | **Speed** |
| Logistic Regression | 89.7% | 78.4% | 52.7% | 0.103 | 0:00:00.147571 |
| Decision tree | 88.8% | 79.8% | 44.2% | 0.112 | 0:00:10.738927 |
| Random Forest | 88.9% | 69.7% | 59% | 0.111 | 0:03:54.156974 |

Table 1: Comparison of model metrics.

From the above result, It is evident that the optimum accuracy is obtained from the Logistic Forward Stepwise Regression Model with optimum Accuracy and Precision, minimum Time taken by the Model for execution (calculated via stop watch) and the lowest MSE value, as we know the lower the MSE, the better is the forecast. Although the Accuracies of Decision Tree and Random Forest Models are very close to the Logistic Regression Model, they both have high MSE value which is crucial in this case as we need to understand the Price value Category of the houses in King County and the features that affect its categorization. Therefore from the above result, It is evident that the optimum accuracy is obtained from the Logistic Regression Model with accurate precision and the lowest MSE value, as we know the lower the MSE, the better is the forecast. Therefore, considering the above factors I would recommend the Business to use the Logistic Regression Model. Using the Significant variables obtained in Logistic Regression Model the Real Estate Business Team can understand the most influential variables impacting to the High pricing (“>750,000”)or the Low pricing (“<750,000”) of the Housing property, can also determining the clients' needs , propose economical solutions that suit them and accurately estimate the future property value in the King County.

# **Conclusion**

From the Model Analysis results, we can conclude that ideal Accuracy, MSE and precision is obtained from the Logistic Regression Model. The attributes “Waterfront” and “Sqft\_living” act as the most important driving factors which influences the High pricing (“>750,000”) or the Low pricing (“<750,000”) of the Housing property in King County, Washington.

We can see from the Model results that proximity to "Waterfront" is one of the most significant variable influencing the High pricing (“>750,000”) of the Housing property in King County, Washington. One probable reason for High pricing for Waterfront houses could be Waterfront View or Sustainable climatic conditions, or High Supply and Demand due to proximity to Water sports. For example, Customers with finance availability (“>750,000”) can be recommended for Houses with waterfront view because the waterfront property is limited in supply with high buyer demand, with appreciation rate to be more steady over time, On the contrary the Customer with finance availability (“<750,000”) can be recommended for purchasing home away from the Waterfront because it will be more economical investment which is vital for majority of the consumers.

Secondly, we see that “Square foot living area” highly influences the High pricing (“>750,000”) of the Housing property in King County, Washington which signifies that, higher the square foot living area higher are the house prices. As a result, it is critical for real estate industry stakeholders to first assess the needs of their customers and then offer price per square foot living area options based on Zip codes as per their financial resources. For example, for a Customer with a large family, a larger living area with a lower overall housing price is often more appealing. In such cases, Real estate agents may advise purchasing houses in Zip codes with low average housing prices rather than Zip codes with higher housing price averages. Our recommendation to the upper management of real estate agencies or businesses would be to optimize the House search for clients by recommending that they purchase similar-sized Dwelling properties in lower-cost Zip codes.

On final thoughts, this information will assist the real estate Business Stakeholders in devising strategies as per the Business clients' requirements ,their financials abilities, and benefit the organisation by proposing them with economical solutions.

# **References**

Burhanykiyakoglu. (2019, May 23). *Predicting House prices*. Kaggle. Retrieved March 5, 2022, from <https://www.kaggle.com/burhanykiyakoglu/predicting-house-prices>

Geomms. (n.d.). *Geomms/predicting-Seattle-housing-prices*. GitHub. Retrieved March 5, 2022, from <https://github.com/geomms/Predicting-Seattle-Housing-Prices>

Plotly Python Graphing Library. (n.d.). Plotly: The front end for ML and data science models. <https://plotly.com/python/>

Why Are Waterfront Properties More Expensive? - Coastal Dream Life. (n.d.). Coastal Dream Life. <https://coastaldreamlife.com/why-are-waterfront-properties-more-expensive/#:~:text=Waterfront%20properties%20are%20more%20expensive,demand,%20naturally%20boosting%20the%20prices.>

Should I use a decision tree or logistic regression for classification? (n.d.). Data Science Stack Exchange. <https://datascience.stackexchange.com/questions/6048/should-i-use-a-decision-tree-or-logistic-regression-for-classification>

# 

# **Appendix**

Fig 1: Density plot of price

A picture containing text, transport, sailing vessel

Description automatically generated

Figure 2: Number of houses per price group

Chart, bar chart

Description automatically generated

Figure 3: Age of houses

Chart, bar chart

Description automatically generated

Figure 4: Model 2 - Decision Tree model results

Diagram

Description automatically generated with low confidence