Predicting Housing Prices Using Machine Learning

Divyansh Shah
Faculty of Business
Humber Institute of Technology and Learning
Toronto, Canada
n01472284@humbermail.ca

Subhanjan Das
Faculty of Business
Humber Institute of Technology and Learning
Toronto, Canada
n0143147 43@humbermail.ca

Yuvraj Shand
Faculty of Business
Humber Institute of Technology and Learning
Toronto, Canada
n01479401@humbermail.ca

Abhi Tyagi
Faculty of Business
Humber Institute of Technology and Learning
Toronto, Canada
n01474042@humbermail.ca

Hardi Patel
Faculty of Business
Humber Institute of Technology and Learning
Toronto, Canada
n01480409@humbermail.ca

***Abstract*—The housing market is undoubtedly one of the most highly invested asset class of our generation and hence it is critically significant to keep enhancing the currently used models and algorithms. In this study we examine the key variables that influencing the house prices and forecast these prices with the help of Machine Learning models like Multiple Linear Regression and k-NearestNeighbours (kNN) we also evaluate the performances of our models. Although the difference is not significant, it was discovered that the multiple liner regression model consistently outperformed the k-NN algorithm. Though only applicable for simple predictions, the models described in our work yielded exceptionally low Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).**

***Keywords*—Multiple Linear Regression, k-NearestNeighbours, Machine Learning.**

I. INTRODUCTION

The purchase of a property is easily the most significant financial investment for the majority of the population. purchase typically follows an extensive decision-making process that involves enormous amounts of research[1]. For the majority of home buyers, this research proves to be a challenge due to the lack of information and knowledge regarding the housing market and the overall economic state[1]. Properties and assets are generally overestimated due to this lack of knowledge and a multitude of other factors which induce an imbalance in property prices and the housing market[1]. Predicting the prices of these properties is hence an extremely challenging research avenue the house market is influenced by several mutually correlated factors[1]. Human behavior also plays an indistinguishable role in determining the price of a property in an area or a locality[1]. Price prediction models and algorithms are of critical significance as extremely high value transactions are dependent on the models, due to integration of Machine Learning and Artificial Intelligence with the banks, asset management firms and big businesses[1]. It is therefore indispensable to conduct substantial studies by adopting newer methodologies and approaches instead of the contemporary methods[1].

In this paper, we attempt to construct realistic models using regression and evaluate their performances and efficiencies, to accurately estimate the value of real estate[3]. We also discovered several relevant factors that directly influence the price of a property and to what extent.

II. DATA

The dataset consists of 21 columns and 21613 rows that combine a total entry of 453,873 entries.

Table 1. Data Description

Variable	Description	Data Type
price	Sale Price for the houses	Numeric
bedrooms	No. of bedrooms in the house	Ordinal
bathrooms	No. of bathrooms in the house	Ordinal
sqft_liv	Size of the living room	Numeric
sqft_lot	Size of thhe entire lot	Numeric
floors	Type of flooring in the house	Ordinal
waterfront	Property facing water body	Categorical
view	Rating of the view from the house	Ordinal
condition	The condition of the house at the time of viewing	Ordinal
grade	Rating of building	Ordinal

	construction and design	
sqft_above	Aside from the basement, square feet above ground	Numeric
sqft_basmt	square feet underground	Numeric
yr_built	How old the house is	Numeric
yr_renov	When was the house renovated	Numeric
squft_liv15	Average size of living area	Numeric
squft_lot15	Area of land lots	Numeric

There are 15 integer type columns, 5 are float type and there is an object column. We do not have any NULL values.

ID, date, zip code, lat, long columns have been removed as they are not useful for the modeling and would have consumed additional memory.

The following table describes the minimum, mean, standard deviation, median, maximum for the continuous variable in the Housing dataset:

	price	sqft_living	sqft_lot	sqft_abov	sqft_baser	sqft_living	sqft_lot15
Min	75,000	290	520	290	-	399	651
MEAN	540,088	2,080	15,107	1,788	292	1,987	12,768
Std	367,127	918	41,421	828	443	685	27,304
Median	450,000	1,910	7,618	1,560	-	1,840	7,620
Max	7,700,000	13,540	1,651,359	9,410	4,820	6,210	871,200

Figure 1: Description of numerical columns

III. WORKING

This work aims to build a linear regression model and a k-Nearest Neighbors model to predict the housing prices and suggest which among the two models gives better predictions. The dataset consists of 21 columns and 21613 records, with a brief overview of the dataset we were able to assess those 5 columns i.e., ‘id’, ‘date’, ‘zipcode’, ‘lat’, and ‘long’ do not contribute much to the price of the house in the current dataset.

The next step was to remove or correct any outliers that existed in the dataset. It consisted of only one outlier that existed in the ‘bedrooms’ column. The house with id 2402100895 had a typographical error and the value of that parameter was changed and not removed.

As there were no categorical columns with good correlation with the output variable ‘price’, no dummy variables were created. Few of the houses were renovated that could have influenced their price. Thus, a new variable was created that represented if the house was renovated or not. If the house was renovated, it had value as 1 else 0 and the original column ‘yr_renovated’ was deleted from the dataset.

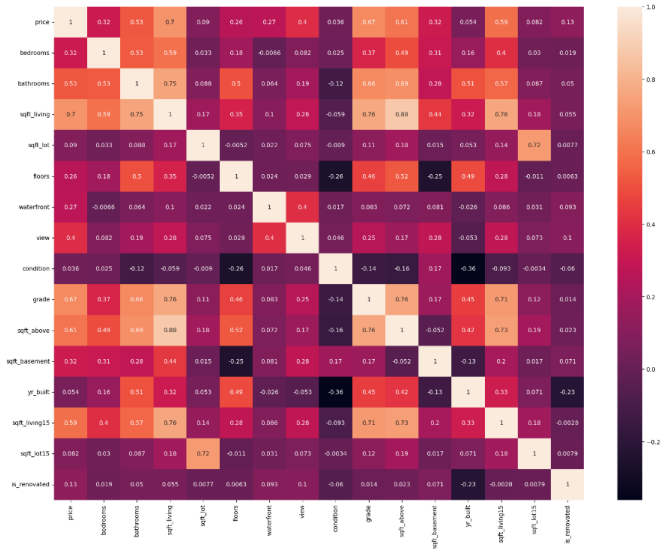

Figure 2: Correlation Heat Map

Before performing any regression to predict variables, the dataset went through an exhaustive search algorithm using backward elimination, forward selection, and stepwise selection. This allowed us to split the dataset into training and validation dataset. Python function was created to fit and find the AIC score[2]. This was done to predict which variables would be beneficial for the regression model. The backward elimination method suggested to remove ‘sqft_living’ and ‘floors’. However, the correlation of ‘sqft_living’ with ‘price’ column is very high. Thus, it cannot be removed. The next step was using the forward elimination method and this method suggested to not include ‘sqft_basement’ and ‘floors’ in the further analysis. Similarly, stepwise selection method also suggested the same result.

Following the suggestions, the linear regression model was fitted using 60% training data and 40% testing data. The trained data was further used to predict the price of the houses in the testing dataset.

Next , K-Nearest Neighbours Model was used to predict the price of the houses. In the model, the value of K was taken as 5 and 10.

IV. RESULT

Table 1. Results

	Linear Regression	**K Nearest Neighbours (K = 5)**	**K Nearest Neighbours (K = 10)**
Mean Absolute Error	151751.64	158357.06	153539.86
Mean Squared Error	58443560506.06	70345440921.10	68241853246.20

Root Mean Squared Error	241751.03	265227.15	261231.41

As mentioned in table 1, it depicts that the linear regression model has the lowest MAE, MSE and RMSE at 151751.64, 58443560506.06 and 241751.03 respectively. The model is considered a good model if their corresponding MAE and RMSE are low. Therefore, this model is best suitable for predicting the price of the houses.

Among the k-NN model, the k=10 regression model had the lowest RMSE at 261231.41, thus confirming that with 10 nearest neighbours the model performed better.

ACKNOWLEDGMENT

We extend our sincere gratitude to Humber Institute of Technology and Learning for guiding us throughout the process of this work.

REFERENCES

[1] O'Farrell, S. (2018). House Price Prediction. Comparison of Data Mining Models to Predict House Prices.

[2] Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). *Data mining for Business Analytics: Concepts, techniques and applications in Python*. John Wiley & Sons, Inc.

[3] Qingqi Zhang, "Housing Price Prediction Based on Multiple Linear Regression", Scientific Programming, vol. 2021, Article ID 7678931, 9 pages, 2021