Real-Estate Market Model

Bilal Ayyache
University Of Guelph
Guelph, Ontario
bayyache@uoguelph.ca

Alex Vos
University Of Guelph
Guelph, Ontario
vosa@uoguelph.ca

ABSTRACT

In this paper, the Real-Estate Market in Canada is being modelled through the use of 4 different models. 4 models are used in order to obtain an accurate approximation of a property depending on certain specifications set by the user. The model uses BeautifulSoup to scrape kijiji in order to obtain information to train the model giving it the ability to output a predicted price depending on the data inputted into the model. Data such as what province is the property located in, what city, number of Bedrooms, Bathroom, size of land, and price of property is collected, filtered, analyzed and finally used to train the model. The model takes a user input of an information list on desired prediction. This information list includes number of rooms, number of washrooms, size of property, and location.

CCS CONCEPTS

• **Computing methodologies → Agent / discrete models**.

KEYWORDS

KNN Regression, Decision Tree Regression, Gradient Boosting, Decision Tree, Predicted Price, Real price, Residual Counts, Correlation Matrix

ACM Reference Format:
Bilal Ayyache and Alex Vos. 2018. Real-Estate Market Model. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/1122445.1122456

1 PROBLEM STATEMENT

When Looking into buying a new home, typically an overriding concern is: Am I paying too much? This question is often difficult to answer due to the fact that it's hard to keep track of all the houses available on the market. A second, and related concern, is: Which house with similar specifications are available in a specific location? This information can help the buyer get a feel for what else is available on the market and provide an indication of the value of the estate currently under consideration.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

2 SOLUTION OVERVIEW

Through this project, the main goal is to design a model that can be used to predict the price of real estates. Being able to predict a house price in a specific location can set the user a step ahead in making the right purchase decision. Using this model, a user will be able to input year of build, square feet, number of rooms and washrooms. In return the model will output the best price for such specifications. The main objective of this model is to find the best price of an estate depending on specific factors such as number of bedrooms, bathrooms, house type and how many square-feet the house is. Such model can be used to compare prices of real estates in cities of Ontario. This model outputs accurate approximation on what the price range of a certain specification of a house would be. In this project, models such as the KNN Regression, simple linear regression, Multiple linear regression, Decision tree and gradient boosting will be used to model achieves accurate results.

3 RELATED WORK

House Price Affecting Factors

There are several factors that affect house prices. In his research Rahadi, et al. divide these factors into three main groups, there are physical condition, concept and location. Physical conditions are properties possessed by a house that can be observed by human senses, including the size of the house, the number of bedrooms, the availability of kitchen and garage, the availability of the garden, the area of land and buildings, and the age of the house [15], while the concept is an idea offered by developers who can attract potential buyers, for example, the concept of a minimalist home, healthy and green environment, and elite environment.

Location is an important factor in shaping the price of a house. This is because the location determines the prevailing land price [16]. In addition, the location also determines the ease of access to public facilities, such as schools, campus,

hospitals and health centers, as well as family recreation facilities such as malls, culinary tours, or even offer a beautiful scenery [17], [18].

Hedonic Pricing

Hedonic pricing is a price prediction model based on the hedonic price theory, which assumes that the value of a property is the sum of all its attributes value [20]. In the implementation, hedonic pricing can be implemented using regression model. Equation 1 will show the regression model in determining a price

$$y = ax_1 + bx_2 + cx_3 \ldots + nx_n \quad (1)$$

Where, y is the predicted price, and x1, x2, xi are the attributes of a house. While a, b, ... n indicate the correlation coefficients of each variables in the determination of house prices.

4 METHODOLOGY

To model the market, elements of the model will be used to analyze and study the behavior of the system. After collecting information, 4 different models will be trained to predict the price of the specified real estate. To compare these models, variance score and root mean square error will be compared. The model with the lowest RMSE and highest Variance score will be used to output the approximated price. KNN Regression, simple linear regression, Multiple linear regression, Decision tree and gradient boostingâĹŮ model approach will be used to predict the price of the specifications inputted into the model.

KNN Regression

K nearest neighbors is a simple algorithm that was used to store all available cases and predict the numerical target based on a similarity measure. KNN has been used in statistical estimation and pattern recognition in the beginning of 1970 as a non-parametric technique.
Data is usually split into three parts when using KNN regression: training, validation and Testing set, but for simplicity and scope of this project, model will be trained and tested with 20% in test size. A simple implementation of KNN regression is to calculate the average of the numerical target of the K nearest neighbors.

Decision Tree Regression

Decision tree builds regression or classification models in the form of a tree structure. The decision Tree regression down the dataset collected from kijiji into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

A decision node usually has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. In this model's case there was only one branch for simplicity

Gradient Boosting

Gradient Boosting is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms which convert weak learners to strong ones. Boosting is based on the question posed by Kearns and Valiant (1988, 1989): Can a set of weak learners create a single strong learner? A weak learner is defined to be a classifier which is only slightly correlated with the true classification. Using boosting, scores were improved.

Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

5 MODEL RESULTS

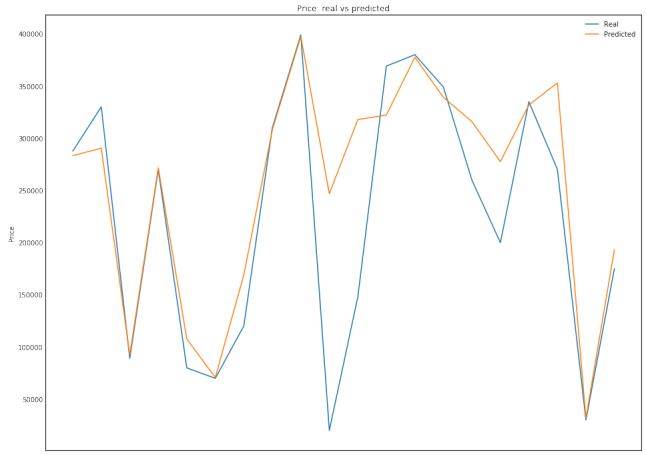

Figure 1: Real Vs Predicted Price [Real: Blue Predicted: Orange].

It is noticeably clearly that the two line (real vs predicted) fit each other well, with some small differences which proves improvement compared with the first model.
Table 1 highlights the model used, variance obtained using the model, and the RMSE Value:

Table 1: Model Results (Gradient Boosting recorded best score)

Model	Variance	RMSE
KNN	0.56	37709.67
Multiple Regression	0.62	34865.07
Gradient Boosting	25176.16	
Decision Tree	0.63	34551.17

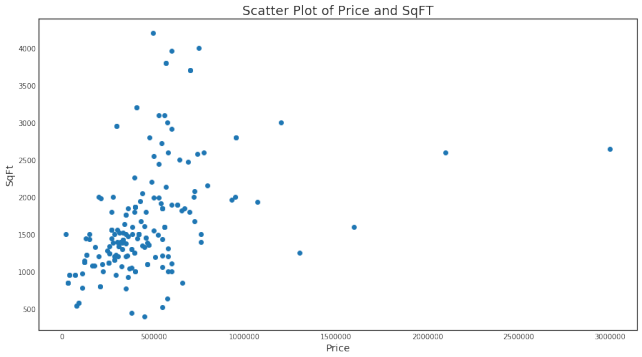

Figure 2: SqFT Vs Price graph.

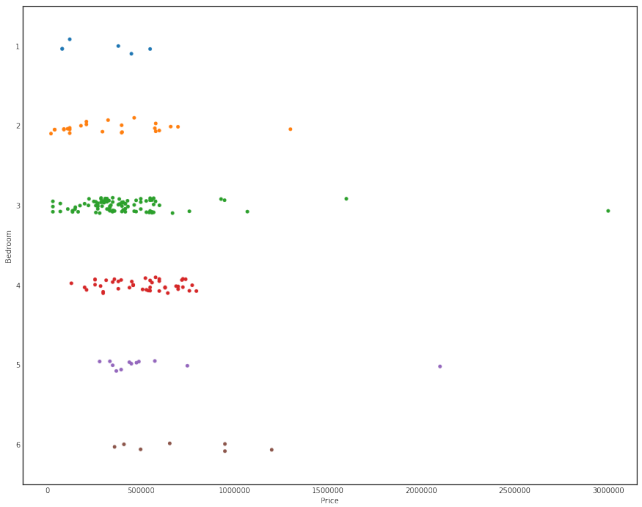

Figure 3: Number of bedrooms Vs Price [Real: Blue Predicted: Orange].

Figure 2 and 3 describe the behavior of the system as the property price increase. The estate price increase respectively as size of estate increases, and more explicitly it is clear that the more the rooms and bathrooms in a house, the price augment, while in the other side a small estate still has a low price, and this is totally logical since whenever size of house increase their price starts to increase.

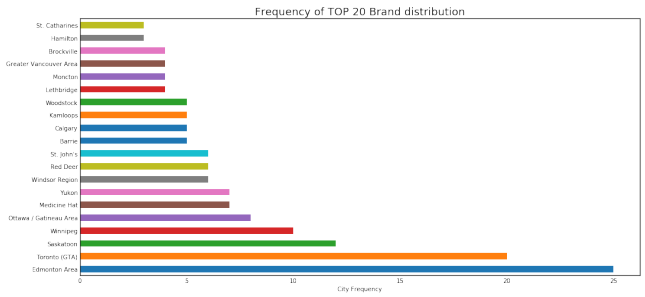

Figure 4: Frequency Of top 20 City Distribution.

Figure 4 clearly presents that most of the ad publications are coming from Edmonton Toronto. Having Toronto on the top 20 distribution is quite normal due to the geographic distribution of the population. It seems like the model collected multiple duplicates in Edmonton. This would change every time you run the model. To solve this problem, a function should be programmed to delete duplicates
furthermore such graph can describe the economic position of those cities in Canada beyond the other ones

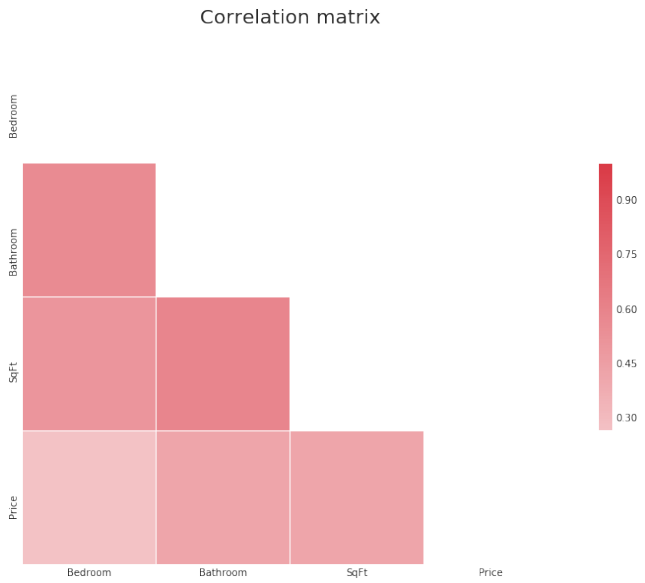

Figure 5: Model Correlation Matrix.

The correlation matrix shows that there is a strong relationship between Sqft and bathroom. This seems accurate because in bigger homes we ussually have more bathrooms. SqFt and price is an another strong correlation. bigger property means larger sqFT value. The correlation matrix can be used to prove that the chosen variables are a good choice or not. Using the correlation matrix, we see that our variables correlate to each other. This helps in building a stronger and a more accurate model.

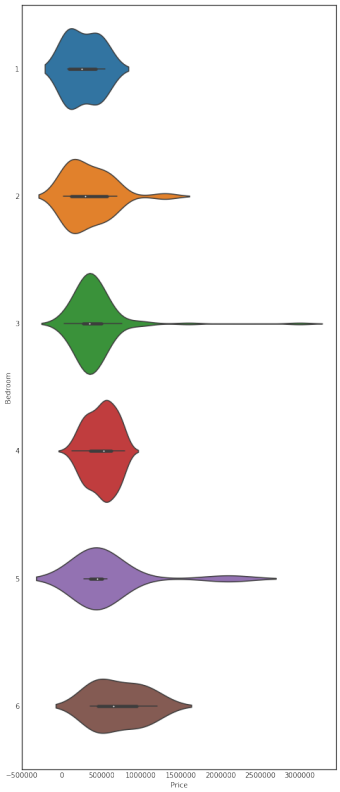

Figure 6: Violin Plot.

The figure above is a Violin plot. The violin plot shows the full distribution of the data (Price vs Num of bedrooms).

6 DISCUSSION

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) (or sometimes root-mean-squared error) is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. The RMSD represents the sample standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation and are called prediction errors when computed out-of-sample. The RMSD serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSD is a measure of accuracy, to compare forecasting errors of different models for a particular data and not between data sets, as it is scale-dependent. By comparing the Gradient Boosting Model with the KNN Regression we can see that the RMSE was reduced from 37709 to 25176.16 and a variance increase from 0.56 to 0.80.

7 FUTURE FIXES AND IMPLEMENTATION

The model is fully functional but results is not accurate. There are some issues that creates inaccuracy in result. Issues include:

- The model collects information from kijiji properly, but due to inconsistency in ads information it made it difficult for the model to process the information. Manual filtration was required to run the model. To fix this issue the model should scrape a more accurate website that is specialized for real estates. This would provide more variables to model with, therefore accuracy would be improved drastically. another solution to this problem is programming a better filtration process
- When collecting ads details, duplicates were not deleted. duplicates reduced the accuracy of the model as the same information was saved multiple times. To solve this issue a function would be programmed to delete duplicates.
- Information is not accurate as only 200 samples were used. This is due to the above issues causing the function that scrapes the website very slow. We couldn't figure out why scraping took a very long time to get completed

8 CONCLUSION

using the data that was collected we were able to produce accurate enough results proving that the concept would work if a different website was used.

In this model, Data was used to train the model on predicting the price of a specific requirement of a real estate property. This model was based on a gradient boosting model. The model recorded a variance and RMSE of .80 and 34551.17 after analyzing 200 Ads. The modelâĂŹs accuracy increases over time as proven in the result section.

Most of the challenges faced during the design process in this project were mostly technical as this was the first time both team members worked with python.

9 ACKNOWLEDGMENTS

A special thanks to Alan Downey as the code we wrote was based on his work in his book Think Complexity 2 . We would also likeextend our thanks to Thorsteinn Hjortur Jonsson, Matthew Saunders, and Hannah Szentimrey for their continued leadership, motivation, and inspiration during the modelling complex systems course at the University Of Guelph.This project wouldn't have been possible without their education and patience.

REFERENCES

[1] Jerome H.Friedman (1999). A gradient boosting Machine. Gradient Function Approximation, 10(3), 1âĂŞ50.
[2] Olaf Gefeller (2017). Predictive Modelling Based on Statistical Learning in Biomedicine. Comput Math Methods Med.
[3] RW Swain, KE Kilpatrick, JJ Marsh 3rd (1977). Implementation of a model for census prediction and control. Health Serv Res, 12(4).
[4]Allen B.Downey Think Complexity: Exploring Complexity Science in Python. Green Tea Press, Needham, Massachusetts, 2016, pp. 141-151.
[5] Joshua M. Epstein and Robert Axtell. Growing Artificial Societies: Social Science from the Bottom Up. The MIT Press, 1996

10 APPENDICES