**Prediction of Immobile Market in Montreal - Data Mining**

Huachao Li

University of Colorado Boulder  
Laval, Québec, Canada  
+1 (123) - 456 - 8888

huli7852@colorado.edu

**ABSTRACT**

As inflation aggravates and the number of immigrants increases drastically in the country, house prices have been skyrocketing and augmenting drastically from year to year. Under such circumstances, it is desirable to develop a prediction model to be able to predict future house prices in major cities in Canada. Because of differences between cities, Montreal, the city where I dwell in, was chosen to be the subject of research for this data mining project.

Our objective is to gather trustworthy and consistent data, which show temporal evolution and variation in price regarding the geographical location on Earth, and to train a robust model which can predict housing prices for a given location in the near future. The reason why we stress “the near future” is that housing market is volatile and fast-changing, therefore the learnt features might not apply to another period, if the conjecture of the housing market has witnessed too many vicissitudes.

To simply this project, we will ignore the impact of time on the house prices and only look at the impact of geographical locations. In terms of data, a dataset is available from Kaggle, and we potentially can add more data to this dataset by searching for more data on the internet. Some data preprocessing and cleaning might be necessary before the data can be utilized to train our prediction model. Before conducting machine learning algorithms, it is desired to play with the data by applying a series of functions from libraries in Python such as pandas and numpy, and to visualize them, to grasp the tangible features even before building the model. Domain knowledge in the housing market can be integrated with our observations, with which we will already have a basic control over the direction of the project.

To obtain the best model, parameter tuning and learning rate analysis need to be performed; since there are no other features except for geographical location among the predictors, there is no need to filter out features. Since the predictors and the dependent variable, which is the price are all numeric, it is apparent that regression will be used as the training method. Considering that many regression models are available, we might want to train different regression models, comparing their respective outcomes to choose the best one. To evaluate those models, we need to obtain the R square values and compare them among all the models.

**Keywords**

Machine learning; regression; prediction

# INTRODUCTION

# To begin with, the problem here is to train a model that can efficiently and effectively predict housing prices in Montreal with merely the geographical location information. Dataset with latitude, longitude, number of bedrooms and price will be used, where the first three variables will be the predictors used in the regression model, with which the price is predicted.

# As the economic situation in Canada worsens and immigration gets out of control, the house prices in major cities are reaching rather high levels, and it concerns researchers and habitants to know the house market better and evaluate a house for which the price is unknown. Therefore, this project is of utmost importance, because it is the first step to build a robust model to predict house prices. Since it is a simplified version of a final prediction model that can be put in use for economists and brokers, we can say that our work is crucial for building a more robust and more complex model in the future.

# Our contribution is this project is potentially adding more data to the already existing dataset, training a prediction model that is at the same time efficient, effective and interpretable, evaluating it and planning the next steps to build a more mature and more robust model

# Related Work

Before the start of the project, a lot of datasets were already available on the internet, among which we have found the dataset of Montreal house prices on Kaggle. Apart from that, many analyses have been conducted to understand the data and models have been trained and evaluated on the Kaggle website. However, it is possible to add more data to the dataset and perform more cleaning on it to improve the quality of the data; in addition, improvements can be made on the models, by tuning the parameters using grids and by selecting the optimal training data percentage. We will also discuss future possibilities of making use of such a model, adding more features to it, to build a model that is complex enough to explain the oscillations in the housing markets, with regard to time, location, and other parameters.

# Proposed Work

## Enlargement of Dataset

## We wanted to obtain data from the already existing dataset from Kaggle, which contains four columns, i.e., longitude, latitude, number of bedrooms and price. However, it is decided that this process could lead a considerable waste of time; thereby, we decide that we no longer will include this in the scope of this project.

## Data cleaning and visualization

## We have performed preprocessing and filtering on the data.

## We produced box plots of the price against number of bedrooms, longitude, latitude respectively, and to rule out outliers, we apply the rule of 1.5 quantile interval, which should improve the quality of the rest of the data. We have included a scaler in the pipeline as well, so that relationships between the numeric variables are easily identified.

## In order to understand the relationships inside the data and to obtain a global vision about the whole project, we carried out statistical analysis, including the analysis of the covariances between all pairs among the four values. As is shown in Figure 1, it is observed that the price can be related to the number of bedrooms in a way, and it might be possible that the price is not strongly linked to either longitude or latitude, which is surprising, because this is opposite to what we assumed in the beginning.

## Une image contenant texte, capture d’écran, carré, diagramme Description générée automatiquement

Figure 1. Correlation Heatmap for All Variables

## Data Warehousing

We have created a data warehouse which helps us stock and manage historical data, facilitating future use of data.

## Training of Model

After we had the data ready, we performed the conventional supervised regression model.

According to our plot of MSE vs training data set percentage, it is constated that as the training data percentage increases, MSE decreases, which indicates an improvement of the model. When the percentage is at around 0.4-0.5, we can observe a desirable MSE. Therefore, we chose the 0.55 as the test size for the training of the model.

Afterwards, we tried different regression algorithms, including linear regression, decision tree regression and random forest regression, among other regression algorithms. After training all of them respectively, we could decide on the best algorithm to use, which was the linear model. It is noted that random forest regression was able to give a very low MSE and a very high R square score, which made them the best model candidates.

Une image contenant texte, diagramme, ligne, Tracé

Description générée automatiquement

Figure 2. Plot MSE vs Training Set Percentage

We performed parameter tuning as well, improving the performances of the two models.

# Evaluation

To evaluate the models, like what is mentioned above, MSE and R square scores are our main tools. It is desirable to run different regression models with different parameters on the training set, to evaluate their respective performances and to select the model or models that have the lowest level of MSE and highest level of R square score.

# Discussion

As is shown in table 1, among all the models, linear regression, decision tree and random forest seemingly yield very ideal results.

Table 1. MSE and R Square Score Comparison

|  |  |  |
| --- | --- | --- |
|  | MSE | R Square |
| Linear | 0,001 | 0,999 |
| Decision Tree | 0,096 | 0,922 |
| Random Forest | 0,093 | 0,924 |

Since the linear model has a considerably lower MSE and a remarkably high R square score, compared to the other two models, which perform also very well, it is determined that the house price is subject to a linear relationship regarding the number of bedrooms, the latitude and the longitude.

This is very interesting, as during the data visualization process, in the correlation heatmap, the covariance appeared very low between the geographical coordinates and the price, whilst here it seems that the price is highly correlated with the three predictors.

If we compare our model with what is done by the other on Kaggle, we can see that our model outperforms their models by a lot, which confirms the great success of our data mining project.

Despite the temporary success, we must keep in mind that the model we have constructed is rather simple, and is built on many assumptions, which means that it probably will not apply when more complex factors are considered. For future research based on this model, we are bound to introduce a wide spectrum of variables, to improve the performance of our current model.

# Conclusion

As a conclusion, without doubt, the construction of our model is a great success. During the data cleaning step, we have ruled out outliers according to the box plots and our domain knowledge, which ensured the high quality of the original data. In addition, we scaled the data, which facilitated the machine learning process. During the training process, to determine the best training percentage, we carried out a learning curve analysis and the best training percentage was determined to be 0.45. With this training percentage, we could split the data optimally, apply different models with different hyperparameters to obtain the respective MSE and R Square scores. The latter was used to evaluate the models, with the goal to identify the best performing model, which was the linear regression model, having a MSE level of almost 0 and an R square score that was close to 1.

From what was observed, we could conclude that house price in Montreal follows a linear relationship regarding the houses’ longitude, latitude and number of bedrooms, which can be confirmed with our domain knowledge.gg

This successful model building gave us hints for future steps to build a model that is more mature with more complex parameters.

# Acknowledgements

My thanks to ACM SIGCHI for the template for this report. My utmost appreciation for professor Lv from University of Colorado Boulder.

# REFERENCES

1. https://www.kaggle.com/datasets/rnunez1975/montrealproperties