Predicting House Prices in Toronto: A Machine Learning Approach

Albina Cako, BSc *York University, Certificate in Machine Learning*
Colin Green, BSc *York University, Certificate in Machine Learning*
Lucy Zhang, BSc *York University, Certificate in Machine Learning*
Sean X. Zhang, MSc *York University, Certificate in Machine Learning*

The project focus on building a machine learning model for predicting house price in Toronto. Although House Price Index (HPI) is commonly used for estimating the changes in housing price but it is not perfect for individual housing price prediction due to high correlation of housing price and other factors such as house location, income distribution, area, and etc. This project evaluates four different model and chooses the best one for deployment with ShinyApp. The app created for both buyer and seller to have an estimated house price due the location and different attributes of the house, and help users to make decisions.

Keywords: house prices, machine learning, caret, shiny

Introduction

The House Pricing Prediction app is created for estimate the house price for both buyer and seller based on different factors such as total Sqft, house locations, etc. The deployment was constructed using ShinyApp. An user friendly app for both buyer and seller, with simple click of factors users will get an estimated housing price. The app can be used for individual buyers who want to know the final price of the houses they are interested or for individual sellers to know what is the best listing price. The app uses regression model for prediction, which was trained by the data set of Toronto housing price. The housing price is strongly correlated with other factors, for increasing the model accuracy decreasing errors, it is important to try different factors and combinations. This project will comprehensively validate four different models: decision tree, random forest, K nearest neighbors, and gradient boosting machine. This report will go through data analysis, modeling implementation and provide an optimistic result for housing price prediction.

Objective

The objective of this project was to evaluate the application of machine learning algorithms to predict house prices in the Greater Toronto Area, and apply

Background

Purchasing a house is a big life decision for every individual and needs a considerable amount of research. Everyone has different purpose of buying houses, someone would prefer by the house at the best rate for living now, someone would buy houses for future investment. Selling the houses is also very important and needs to do research and decide what is the best leasing price. Commonly, people will ask advise from various websites, real estate agents or realtors before purchasing or leasing; However, due to the trend towards big data, house pricing prediction can be done by using machine learning strategies base on large amount of data from previous years more correctly. House Price Index (HPI) can measure the price changes of residential housing as a percentage change, In Canada the new Housing Price Index is calculated monthly by Statistics Canada. HPI is useful but because it is a rough indicator calculated from all transactions, it is inefficient for predicting a specific house with its attributes. The purpose of this project is to create an app for both buyers and sellers can easily check the predicted list price or final price based on the attributes of the house such as locations, square foot, number of bedrooms, etc.

Methodology

Data Preprocessing

The housing dataset, originally shared on Github[1], was extracted from Zoocasa.com in the summer of 2019. The dataset contained all completed property sales in the city of Toronto within an approximately 1-year span. We performed several data exploration and cleaning steps to prepare this data for modeling.

Missingness

We assessed the dataset for missing values. Many parametric machine learning models do not accept missing data, and the accuracy of even non-parametric models are often negatively impacted by missingness. Thus, missing values ought to be either imputed or removed before data modeling. We then determined whether missing data was Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR). Should the data be MCAR, then it is acceptable to simply remove each observation that is missing, as doing so would not introduce bias to the remaining observations. However, if there was a correlation between missingness and other data features, then imputation must be performed. Missingness correlation was assessed using the missing_compare() function from the finalfit library, which applies the Kruskal Wallis correlation test for numerical variables and chi-squared correlation test for categorical variables to determine correlation. Using the MICE package in R, we then applied the following imputation methods: 1) simple, which imputes a value from a simple random sample from the rest of the data; 2) mean, which imputes the average of all observations; 3) random forest, which applies a random forest algorithm; and 4) CART, which imputes by classification and regression trees. The distribution of the imputed data were evaluated with a density plot and an imputed dataset was chosen based on best fit.

Data Curation

Modeling

As the data contained a mix of categorical and numerical variables and did not satisfy many requirements of parametric models, such as variable independence, and normally distributed data. Thus, several parametric models were used. We trained four different non-parametric models using k-fold cross validation. The models were then tuned using various grid searches to improve the accuracy. The final model was chosen based on three metrics: Root Mean-Squared Error (RMSE), Pearson correlation (Rˆ2), and Mean Average Error (MAE).

Decision Tree Model

Random Forest Model

Decision Tree Model

Gradient Boosting Model

Results

The original housing dataset contained 21 variables and 15234 observations. Table 1 defines each variable of the dataset.

Table 1: Data Dictionary

Label	Description
title	Title of the listing
final_price	Final price of the property
list_price	Listing price of the property
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms
sqft	Area of property in square feet
parking	Number of parking spaces
description	Verbatim text description of the property
mls	MLS ID
type	Property type
full_link	URL to listing
full_address	Full address of the property
lat	Latitude
long	Longitude
city_district	Toronto district to which property belonged to
mean_district_income	Average household income of district
district_code	Numerical code of the district
final_price_transformed	Box-Cox transformation of final price
final_price_log	Log transformation of final price
bedrooms_ag	Number of bedrooms above ground
bedrooms_bg	Number of bedrooms below ground

Data Exploration

```
data <- read.csv('houses_edited.csv')
numeric_cols <- list('final_price', 'list_price', 'bathrooms', 'sqft', 'parking', 'lat', 'long',
predictor_cols <- list('bathrooms', 'sqft', 'parking', 'lat', 'long', 'mean_district_income', 'b
```

index Min. 1st Qu. Median Mean 3rd Qu. Max. 0 5678 9804 9520 13668 17543 title Length Class Mode 15234 character character

final_price Min. 1st Qu. Median Mean 3rd Qu. Max. 103000 535000 715000 882714 989000 13180000

list_price Min. 1st Qu. Median Mean 3rd Qu. Max. 104900 529000 699900 875093 969900 13180000

bedrooms Length Class Mode 15234 character character

bathrooms Min. 1st Qu. Median Mean 3rd Qu. Max. 1.000 1.000 2.000 2.122 3.000 14.000

sqft Min. 1st Qu. Median Mean 3rd Qu. Max. NA's 250 650 900 1116 1300 4374 4521

parking Min. 1st Qu. Median Mean 3rd Qu. Max. 0.000 1.000 1.000 1.559 2.000 11.000

description Length Class Mode 15234 character character

mls Length Class Mode 15234 character character

type Length Class Mode 15234 character character

full_link Length Class Mode 15234 character character

full_address Length Class Mode 15234 character character

lat Min. 1st Qu. Median Mean 3rd Qu. Max. 43.59 43.65 43.69 43.70 43.76 43.84

long Min. 1st Qu. Median Mean 3rd Qu. Max. -79.62 -79.45 -79.40 -79.39 -79.34 -79.12

city_district Length Class Mode 15234 character character

mean_district_income Min. 1st Qu. Median Mean 3rd Qu. Max. 25989 34904 50580 56066 67757 308010

district_code Min. 1st Qu. Median Mean 3rd Qu. Max. 1.0 39.0 76.0 71.3 101.0 140.0

final_price_transformed Min. 1st Qu. Median Mean 3rd Qu. Max. 2.380 2.390 2.391 2.391 2.392 2.397

final_price_log Min. 1st Qu. Median Mean 3rd Qu. Max. 11.54 13.19 13.48 13.54 13.80 16.39

bedrooms_ag Min. 1st Qu. Median Mean 3rd Qu. Max. 0.000 1.000 2.000 2.336 3.000 9.000

bedrooms_bg Min. 1st Qu. Median Mean 3rd Qu. Max. 0.0000 0.0000 0.0000 0.5396 1.0000 6.0000

Discussion

Acknowledgements

The authors would like to thank Hashmat Rohian, adjunct faculty at York University for supervision of the project. We also thank Slava Spirin for the original extraction of the Toronto Housing dataset [1]. Finally, we thank Steve V. Miller for creation of the manuscript template in R Markdown [2].

References

1. Spirin, S. (2020). Slavaspirin/toronto-housing-price-prediction Available at: https://github.com/slavaspirin/Toronto-housing-price-prediction [Accessed October 26, 2020].

2. Miller, S.V. An r markdown template for academic manuscripts. Steven v. Miller. Available at: http://svmiller.com/blog/2016/02/svm-r-markdown-manuscript/ [Accessed October 26, 2020].