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*

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.

Keywords: house prices, machine learning, caret, shiny

Introduction

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

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.

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 then evaluated with a density plot and 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).

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

Label	Description
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          title              final_price          list_price
```

Min. : 0 Length:15234 Min. : 103000 Min. : 104900
1st Qu.: 5678 Class :character 1st Qu.: 535000 1st Qu.: 529000
Median : 9804 Mode :character Median : 715000 Median : 699900
Mean : 9520 Mean : 882714 Mean : 875093
3rd Qu.:13668 3rd Qu.: 989000 3rd Qu.: 969900
Max. :17543 Max. :13180000 Max. :13180000
bedrooms bathrooms sqft parking
Length:15234 Min. : 1.000 Min. : 250 Min. : 0.000
Class :character 1st Qu.: 1.000 1st Qu.: 650 1st Qu.: 1.000
Mode :character Median : 2.000 Median : 900 Median : 1.000
Mean : 2.122 Mean :1116 Mean : 1.559
3rd Qu.: 3.000 3rd Qu.:1300 3rd Qu.: 2.000
Max. :14.000 Max. :4374 Max. :11.000
NA's :4521
description mls type full_link
Length:15234 Length:15234 Length:15234 Length:15234
Class :character Class :character Class :character Class :character
Mode :character Mode :character Mode :character Mode :character
full_address lat long city_district
Length:15234 Min. :43.59 Min. :-79.62 Length:15234
Class :character 1st Qu.:43.65 1st Qu.:-79.45 Class :character
Mode :character Median :43.69 Median :-79.40 Mode :character
Mean :43.70 Mean :-79.39
3rd Qu.:43.76 3rd Qu.:-79.34
Max. :43.84 Max. :-79.12

mean_district_income district_code final_price_transformed final_price_log Min. : 25989 Min. : 1.0 Min. :2.380 Min. :11.54
1st Qu.: 34904 1st Qu.: 39.0 1st Qu.:2.390 1st Qu.:13.19
Median : 50580 Median : 76.0 Median :2.391 Median :13.48
Mean : 56066 Mean : 71.3 Mean :2.391 Mean :13.54
3rd Qu.: 67757 3rd Qu.:101.0 3rd Qu.:2.392 3rd Qu.:13.80
Max. :308010 Max. :140.0 Max. :2.397 Max. :16.39
bedrooms_ag bedrooms_bg
Min. :0.000 Min. :0.0000
1st Qu.:1.000 1st Qu.:0.0000
Median :2.000 Median :0.0000
Mean :2.336 Mean :0.5396
3rd Qu.:3.000 3rd Qu.:1.0000
Max. :9.000 Max. :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].