Toronto Airbnb Price Prediction

Project Introduction:

Airbnb is widely used for booking short-term or even long-term rentals wherever customers need it.
In reality, most customers choose their Airbnb places mainly based on the locations and price. I came up with this machine learning model to not only help hosts to list their rentals at a reasonable price but also provide guidelines for Airbnb to check any listings price outliers.

Datasets:

The dataset is about 2019 Toronto Airbnb Listings Information which was collected from an open-scored organization insideairbnb.com (The data behind the Inside Airbnb site is sourced from publicly available information from the Airbnb API.)
The dataset can be downloaded from the Google Drive link Inside Airbnb provided in CSV format based on the city and year.
The 2019 Toronto Airbnb Listings dataset has over 21000 rows and 29 columns, providing details about each listing, including listing ID, host area, number of reviews, price, and so on. Most of the columns are straightforward to understand what it represents, but a few of them need to additional explanation:
* `description`: introduction from hosts about the general information of their listings
* `property_type`: the property types, such as Apartment, Condominium, House, and so on
* `room_type`: customer rental space as the Entire home/apt, Private Room, or Shared rooms
* `bed_type`: sleeping area as the Real Bed, Futon, Pull-out Sofa, Airbed or others
* `cancellation_policy`: including moderate, flexible, strict_14_with_grace_period

Overall Process Breakdown:

Exploratory Data Analysis
Data Cleaning
Visualization
Modeling
Comparing Results
Summary

Exploratory Data Analysis

For EDA, I focused on 3 main parts:

- the target price:

including its data distribution, linearity, skewness, and correlation coefficient relationships with other features.

- the numerical features:

checked each of their data distribution with a side by side comparing of the scatterplot between individuals vs the target, overall null values, and any potential data cleaning needs

- the categorical features:

At that stage, I didn't include the "description" column as I'll use the Natural Language Processing to further analyze it.
similar processes as plotting out the data distribution for each categorical features, numbers of null values, and any potential data adjustments.

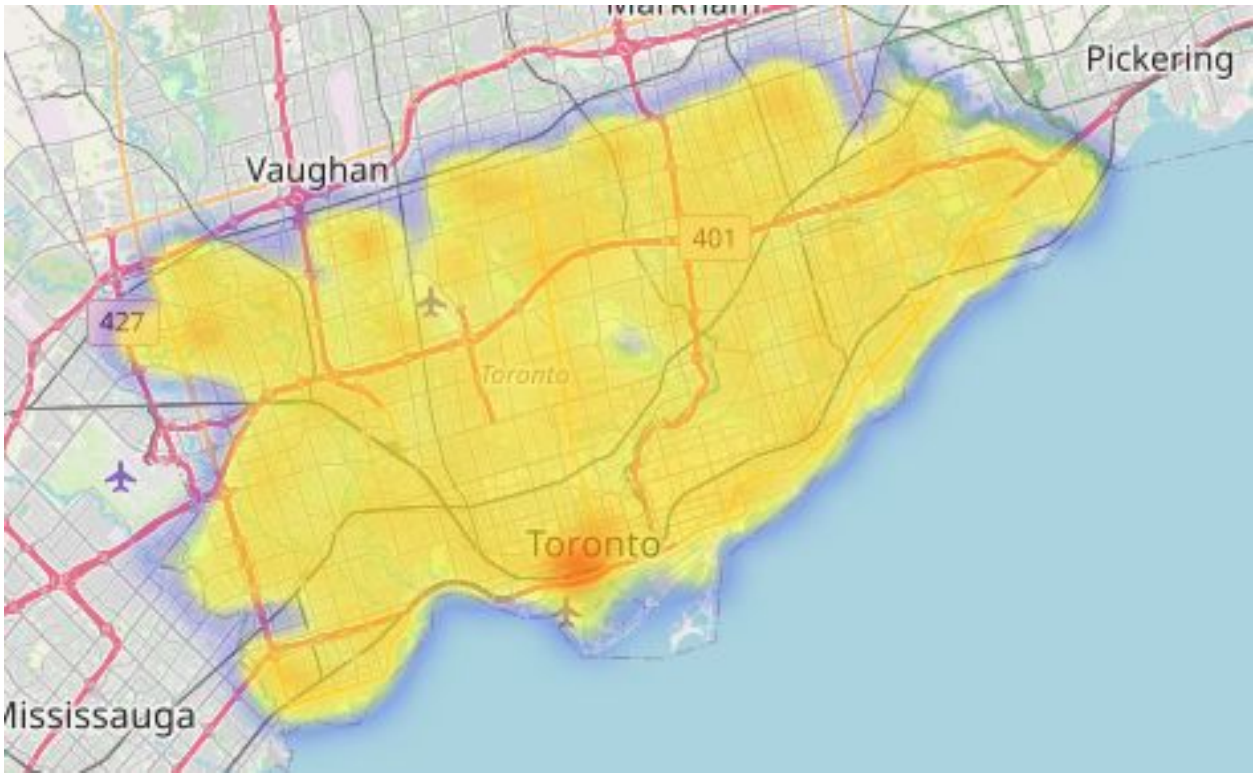
Data Cleaning / Adjustments

The Data Cleaning step included dealing with null values, duplicated rows/columns.
For adjustment parts, as:

- ➔ converted categorical based binary columns into numerical based binary columns
- ➔ dropped categorical columns where over 99% of the values are the same
- ➔ Turned Latitude & Longitude into categorical based information

Visualization

Looking for any visually relationships between the price and other features
For example the heatmap shows that downtown Toronto area had the most amount of listings



Modeling

Target selection¶

As we explored above the target price has a wide range of values and is not normally distributed. There are two main approaches for that:

- log-transform the target variable
- predict price range by range

in this part of the analysis, I focus on predicting the 2019 Toronto Airbnb listing price:

under 300 CAD

based on the mean of the price: 144 CAD and also the article published in early 2019 show that: "According to Luckey Homes, the current overall average price for renting an Airbnb apartment (entire home/apartment) in Toronto is $178 a night."

https://dailyhive.com/toronto/toronto-subway-average-airbnb-map

For features, I separated them into two sets:
one set without the description feature and another set included it.

Then split the dataset into train-test sets
Before applying machine learning models on the first feature set, I used the OneHotEncoder which is one of the feature engineering technologies to convert a few of the categorical columns (host area, property type, room type ...) into numerical columns.

For another feature set that included the description, I used the TfidfVectorizer to tokenize the words into a 0/1 based matrix.

As for the modeling, I choose to use RandomForestRegressor which is a tree-based nonlinear regression model. and then tuned the max_depth parameter to get the baseline of how the testing results could be.

Comparing Results

When it comes to comparing results from two feature sets, I used MAE (Mean absolute error) and R-squared (Coefficient of determination) to determine the performance.
Here's a quick look at the results:

Processing	MAE	R^2	Features
OneHotEncoder	29.45	0.59	158
OHE + NLP	29.23	0.59	1404

By adding the description column with over 1000 features, the results didn't have significant improvements.

Next, I used the Data Pipelines and focused on the first set of features.

as the hyperparameter tuning process:
for the models, I choose one set of
* linear models (Linear Regression, Lasso, Ridge) and added scaling and PCA steps which both could improve the performance.
and another set of
* nonlinear models (RandomForestRegressor, XGBRegressor)
with additional Neural Network regression modeling

as results:
For Price under 300 CAD:
after fitting over 1400 combinations
The nonlinear RandomForestRegressor model
with the hyperparameter optimization reached the best result:
MAE: 29.21, R-squared: 0.59
Also I pulled out the feature importance:

	Feature Importance
Private room	0.348265
bedrooms	0.111974
number_of_reviews	0.065416
Waterfront Communities-The Island	0.058346
host_listings_count	0.048873
accommodates	0.040035
bathrooms	0.039719
minimum_nights	0.034291
-79.4	0.026272
guests_included	0.022959

Summary

Total over 1400 fits and additional neural network regression model training, we reached the best test R-squared 0.59 and best test MAE 29.21 for Predicting 2019 Toronto Airbnb listing price range from 0 ~ 300 CAD.
From the feature importance from the best performance model shows that the room type, the number of bedrooms and the locations will determine the higher listing price.
Interestingly, the more reviews the higher listing price will be.
However, overall the MAE and R-squared didn't reach the ideal range as I expected.

One of the reasons could be the 2019 Toronto Airbnb dataset didn’t include a few more features that might be essential for the price prediction, such as: if the rental has wifi, the size of the bedroom, what kind of beds and so on.

next steps:
* Tuning more hyperparameter and Machine Learning models, as well as Neural Networks Regression models
* predict the listing price over 300 CAD
* predict the Log Transformation listing entire price range
* Collecting more features using Airbnb API, customize the dataset