Finding the best Airbnb for Travelers

CMPT 353 D100

Project Members: Shan Ying (Felicity) Yang – 301305759
Dipansh Arora – 301329048
Instructor: Greg Baker
TA(s): Ali Arab, Ghazal Saheb Jam

Date: August 12, 2021

Finding the best Airbnb for Travelers

Abstract

The project aims to find the best Airbnb for travelers to stay in Vancouver. The project's scope is to suggest Airbnb listings to users based on outdoor amenities near the listings. We used the provided "amenities-vancouver.json.gz" and the Airbnb's listings dataset to acquire the amenities around each Airbnb listing. We also explored if the price of listings correlates to listings' total nearby amenities using supervised machine learning techniques.

Introduction

The problem we found for travellers is that not many travelling websites or applications would provide the option to tell them the different outdoor amenities around their place of stay. Most of the hotel and Airbnb listings only provide the indoor amenities, but not the shops nor restaurants around their stay. Therefore, we thought of resolving this issue by finding the outdoor amenities around the Airbnb listing and providing an amenities rating. Since we want to develop a more user-centred program, we also ask users to input their preference on the Airbnb's price range, number of bedrooms, and number of accommodates.

Furthermore, we performed three regression techniques to analyze if there exists a stronger correlation between Airbnb listing's price and nearby amenities. As a result, we will produce two CSV files, one containing the complete filtered listing information and the other containing the result from the machine learning models.

Data Analysis and Cleaning

We used Vancouver's OSM (amenities dataset) and the detailed Airbnb's Vancouver Listings (listings dataset). The OSM data was used for filtering out the necessary outdoor amenities around all of Vancouver's Airbnb listings. Only the most-needed outdoor amenities are kept in the OSM dataset.

After reducing the size of the OSM dataset by filtering out the unnecessary items, we began performing the three regression techniques on the listing data with the total number of amenities appended to each listing tuple. The three regression techniques were k-nearest neighbours, random forest, and gradient boosting regressors.

The Listings Dataset had 74 total columns and much information that we did not need. We found all the column names using the pandas ".column" method. With the list of amenities

around each listing, we calculated the amenities score by examining the total number of nearby amenities and appended a column to the listing data. The listings dataset was first filtered based on the user requirements of the number of accommodates, the number of bedrooms needed and the budget.

Moreover, we appended a dictionary of amenities within 1km of each Airbnb listing data. The joining of the two datasets is only the first step of data cleaning for our program. The provision of amenities around each listing will still be an overly large dataset for users to interpret. Therefore, to provide user-centred analysis, we further narrow the filtering by asking users to provide a text file of their listings' preference of price range and number of bedrooms and accommodations. We then extract the inputs from the user input file to perform data cleaning the Airbnb listing data.

Data Analysis Techniques

The listings dataset was first filtered based on the user's input of the number of accommodates, the number of bedrooms needed and the budget. The strings inside the text files are converted to the corresponding type and values with the handle_input() function.

Since we want to know the specified amenities around all Airbnb listings, we created an algorithm that calculates the number of different amenities within 1km of the listing's coordinates. Then, we calculated the amenity score for each listing based on its different amenities nearby.

1. Amenity dataset consisted of 'lat', 'lan', 'timestamp', 'amenity', 'name' and 'tags'. First, we removed unnecessary columns (timestamp, tags) and then checked for NaN values. We further analyzed the amenity dataset using the .value_counts method on the amenity column and found all the different amenities.
2. The num_amenities function returns a dictionary of the number of amenities in a 1km radius of each Airbnb listing. This function is simplified with the help of the haversine_distance function, taking the listing data's coordinates and the filtered OSM DataFrame as parameters.
3. Using the haversine_distance function, we accept the filtered OSM DataFrame and the latitude and longitude data from the filtered AirBnb's listing DataFrame. The amenities' latitudes and longitudes are also extracted from the DataFrame inside the haversine_distance function. All of the latitudes and longitudes are converted into radians.
4. The following haversine formula is applied to find the distance between the two given points:

$$d = 2r \arcsin\left(\sqrt{hav(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\, hav(\lambda_2 - \lambda_1)}\right)$$
$$= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{(\varphi_2 - \varphi_1)}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{(\lambda_2 - \lambda_1)}{2}\right)}\right)$$

The haversine distance is then multiplied by the earth's radius in kilometres (6371km) to return the actual circle distance between the two points on earth.

5. The amenities_score function is relatively simple after we have appended the amenities dictionary to each Airbnb listing. We iterate through the filtered listing DataFrame and increment the score based on the total number of different amenities around the location.

We further analyzed our data with regression techniques from supervised machine learning. The original Airbnb listing data was initially filtered with the function clean_data_ML. We dropped unnecessary columns, corrected the wrong data types (i.e., converting from strings to floats), and fit label encoder with the columns using LabelEncoder.fit_transform(). After data filtering for the regression algorithm was completed, we obtained the training and testing data using the train_test_split function in the Sklearn model, which splits the data arrays into two subsets. The X data is the filtered listing data without the price column, and the y data is the price column of the listing data. The models are applied to the listing data without the amenities initially. Then the models are applied to the listing data with the additional nearby amenities column added to it. Upon successfully splitting the data arrays, we fitted the models after calling the regressor functions from the Sklearn model in the following format:

- knn = KNeighborsRegressor(n_neighbors=50)
- rf = RandomForestRegressor(100, max_depth=40)
- gb = GradientBoostingRegressor()

The coefficient of determination (R^2) for each model is obtained for analysis.

Findings

Eventually, we narrowed down the filtering of the listings data after handling the input text file from the user.
Then, we sorted the listings by the calculated amenities score that we calculated. We managed to compute the two CSV files as output: the "ML_Price_Prediction.csv" and "AirBnb_search_results.csv," which represent results of the machine learning model scores and the sorted Airbnb listings.

The following table displays the computed coefficient of determination for the original and the filtered Listing Datasets using the three different regression models:

Table 1. Sample Output of Coefficient of Determination for Listing Datasets

Regressors	Original Listing Dataset (without Amenity Scores)	Filtered Listing Dataset (with Amenity Scores)
K-Nearest Neighbors	0.05235971270926676	0.08927844411212049
Random Forest	0.6421580362281065	0.6691652824194065
Gradient Boosting	0.659758407071633	0.6788327640120857

The coefficient of determination represents the square of the correlation between the predicted and actual prices. A higher coefficient implies a better fit for the regression models. The above table sample output shows that the filtered listing data has a higher coefficient of determination (R^2) than the original listing dataset from all models.
From observation, we concluded that nearby outdoor amenities are more correlated to price prediction. However, other factors, such as fancy or luxury interiors, could impact the price of a listing.

Limitations and Conclusions

One limitation we had in the project is that we are not analyzing enough data to support the claim in our findings. Since we are only analyzing the correlation between outdoor amenities and Airbnb listing prices, we neglected many other features that may have influenced the listing price. Another limitation we face in conducting this project is providing enough visualizations to present our data. Our data findings are all represented in tables and CSV formats, which could have been improved by showing the found correlations graphically and thus encouraging better interpretation of the results.

We have tried multiple approaches to find a meaningful relationship between outdoor amenities and Airbnb listings in Vancouver. Eventually, we narrowed it down to the final version, where we examine the relationship and provide suggestions to users according to the outdoor amenity scores. If time permits, we would further clean and select the appropriate data for analyzing the relationship between listing price and listing's outdoor amenities. We would also further polish the display of our data and findings, such as creating graphs using matplotlib for the regression scores.

Project Experience Summary

Dipansh:

- Performed Data analysis and Data cleaning on both datasets.
- Performed feature engineering to find the number of amenities in 1km radius of each listing.
- Prepared data for machine learning models.
- Performed parameter tuning to improve the performance of the machine learning models.
- Analysed the results of the machine learning models.
- Documented the conclusions and contributed to the project report.

Shan Ying (Felicity) Yang:

Finding Airbnb listings according to surrounding amenities for travelers *August 2021*

- Proposed project outline and milestones
- Refined the research question for the project
- Cleaned datasets
- Prepared input and output data
- Analyzed findings from machine learning models
- Conducted project report covering all requirements and analysis

References

http://insideairbnb.com/get-the-data.html
https://www.kite.com/python/answers/how-to-filter-a-pandas-dataframe-with-a-list-by-%60in%60-or-%60not-in%60-in-python
https://stackoverflow.com/questions/4913349/haversine-formula-in-python-bearing-and-distance-between-two-gps-points