Using public transport data to predict hotel prices

Hira Piracha

PROJECT SUMMARY: June 2020

Capstone Project Summary
Business Analytics

Student Name	Hira Aamir Piracha
Student ID	1904185
Company Name	Datapolis
Project Category	Machine Learning
Faculty Supervisor	Miklos Koren

The project I chose was predicting hotel prices by selecting some feature of public transport data. The planned scope of the project was to select three to four European cities for which I hotel prices from different booking sites was already available to me.

Data Sources

1) Hotels data
   I used the data that was compiled by Professor Gabor Bekes. It was available to me as two csv file and contains data about hotels and about hotel prices downloaded for different cities at different point in time from a rice comparison website. It was altered slightly to ensure confidentiality.

2) Public transport data
   Countries publish public transport data online that is available for free in the form of General Transit Feed Specification GTFS files. GTFS allow public transit agencies to publish their transit data. A GTFS feed is composed of a series of text files collected in a ZIP file. This can vary for cities from six to up to thirteen files. Each file models a particular aspect of transit information.

Data Cleaning

All the data cleaning was done in Python to stay consistent with the programming language that Datapolis is working with and to make it easily reproducible for different cities in the future.

For hotels data, I merged the dataset on hotels id. Filtered data for 2018 for the cities Paris, Berlin and Vienna. I removed 5 star hotels as it is a potential cause for higher prices and kept only the fields that would be relevant for further analysis.

For public transport data I downloaded GTFS feed from 2018 for the same cities. I merged all the text files for every city on their unique ID and had 3 data frames for each city. I then sampled 10000 rows for each city to be able to narrow down the sample upon suggestion from the data scientist supervisor at Datapolis.

Explorative Data Analysis

The next step included doing a detailed exploratory data analysis to see general features of public transport data that could help in feature engineering. This included comparison of number of stops in each city, number of routes available,

how traffic changes in a city during the day, that is, the pay rush hours, and the number of public transport operating per hour foe each stop. This part of my project too was done in python and I plotted some graphs to be able to draw a comparison between different cities.

Geospatial Mapping

The two datasets had no common field on which they could be merged for analysis.

Hotels data had information about which neighbourhood each hotel in the city belonged to. Every city was divided into 10-30 neighbourhoods.

The shape.txt file in public transport data had the latitude and longitude of each bus/train/tram stop.

In the hotels data, I assigned a midpoint and geocoded each neighbourhood and assigned it one latitude/longitude points.
For the public transport data, I assigned each stop from the city to a neighbourhood that was minimum distance from the midpoint of the neighbourhood latitude/longitude to the latitude/longitude of the stop. Every single stop was assigned to one neighbourhood.
From this I was able to merge to two datasets to move on to regression and price prediction.

Feature Selection

Initially I wanted to a seasonal analysis but public transport schedule does not change according to schedule. After exploratory data analysis I settled on using the number of stops in a particular neighbourhood to see how that impacts changes in hotel prices.

Regression Analysis

For the regression and predictive part of my project, I shifted to coding in R.

I carried out regression analysis to describe the relationship between the independent variables (number of stops) and the dependent variable (price) and understand how price of hotels in these cities changes with the number of stops in each neighbourhood

I plotted three regression plots to understand the trend: loess regression model, piecewise linear spline and polynomial regression model.

All three models showed a similar trend – the price of hotels increases as number of stops in the neighbourhood increase up to a certain point (around 200 to 300) stops and after that it either decreases very slightly or does not change at all.

Predictive Analytics

I built different models to see how different variable would influence change in prices. Since I had a few variables I created log and square columns to work with. I set up six models with nine variables in total. I then separated a holdout set. Created a holdout set which contained was 20% of the observations from the data at random. Set the random number generator that will make results reproducible. The workout set or the training data contains 80% of the total observations which I will use to train my models. Based on this result, out of all the models, modellog2 performed the best and had the lowest RMSE of 31.93913. This is the model that included stars, ratings, number of stops as the variables used.

I used thos predictors for CART, Random Forest and GBM models. RSME was used to select the best model.

Conclusion

Finally, I estimated my models on the 20% holdout data separated earlier. We can see the the GBM model performs only slightly better than the CART and Random Forest models on the holdout set as seen from the comparison table below:

MODEL	HOLDOUT RMSE
GBM	29.57643
CART	30.36978
Random forest (smaller model)	30.24622

Summary

I think from the R2 of my models, I can conclude that given the data I had, density of public transport data estimated from number of stops cannot explain increase in hotel prices. The attribution that I do see due to higher number of stops could also be due to other confounding variables that play a part in increase in hotel prices that may be related to public transport data also. For example, the city center may have more stops as well as higher priced hotels but we cannot attribute higher prices to the number of stops.

External validity of my model, therefore, is extremely low.
Furthermore, I only had a few variables to work with
Changes in rental prices due to covid-19 not taken into account. This might be a problem when doing analysis for 2020 data.

Next Steps

To further study how public transport data plays a role in prices, it would be interesting to use the same models on rental properties holding constant for apartment size. This data was not available to Datapolis when I started my project.