Capstone project 2020-06

Predicting real estate prices

Introduction

When potential clients are going through real estate listings the only reference easily available is the average price in a given city or a district which is not always helpful because there are many factors to be considered when buying an apartment.

The goal of this project is to predict prices of apartments and to identify which ones have most attractive price compared to the predicted price.

Target audience - The final result may be of value for people who are looking to buy an apartment as well as people who are selling an apartment and would like to know what price to ask for.

City of choice - I've analysed the city of Poznan in Poland, because this is where i live and therefore may one day need results of this project. The same concept could be used for any other city.

Data

three sources of the data were combined

- Foursquare – needed to identify what kind of venues are within a distance of each apartment. With information on number of “Arts & Entertainment”, “College & University”, “Food”, “Nightlife Spot”.
- Data with real estate listings - in the city of Poznan wer downloaded and used with information about - no of rooms, built year, size. A custom page scraping code had to written to get the data as there was no available API.
- Location data - to be able to explore venues around apartments, I also needed location data. This information was retrieved using OpenCage Geocoding API to get the longitude and latitude.

Exploratory data analysis

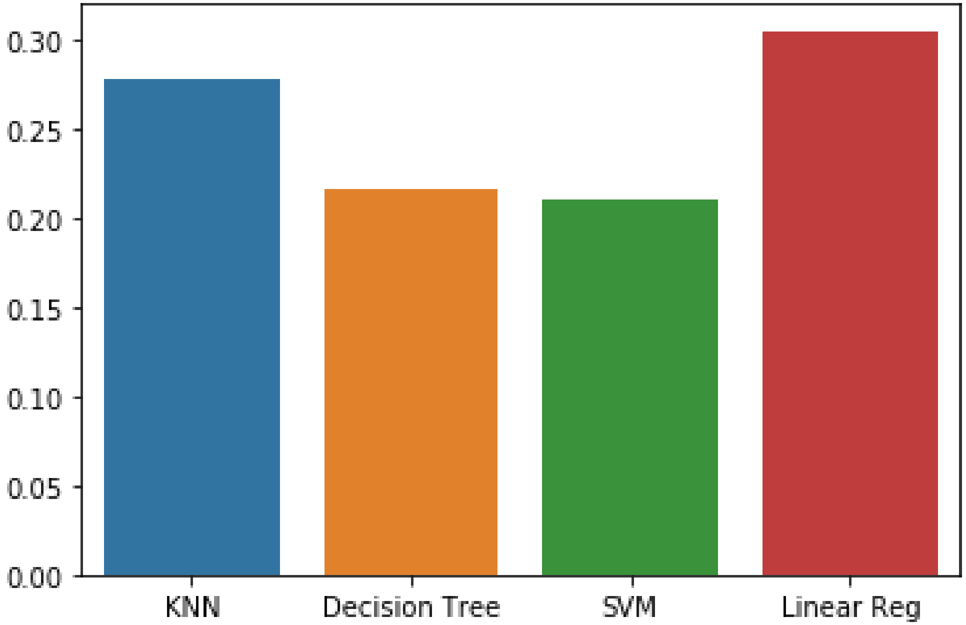
Table below presents which columns were dropped and which were later used for regression. The reason for removing columns is that after exploring the data for some columns there was no visible correlation between the data and apartment prices.

Used for regression	Dropped
for_renovation, for_completion, building_type, rooms, rent, built year, size, floors, floor no, distance, Arts & Entertainment, College & University, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places	has_balcony, has_basement, has_garage, has_garden, has_terrace, has_lift, is_two_storey, Shop & Service, Travel & Transport

Methodology

Different types of regression models used

- K Nearest Neighbor (KNN)
- Decision Tree
- Support Vector Machine
- Linear regression

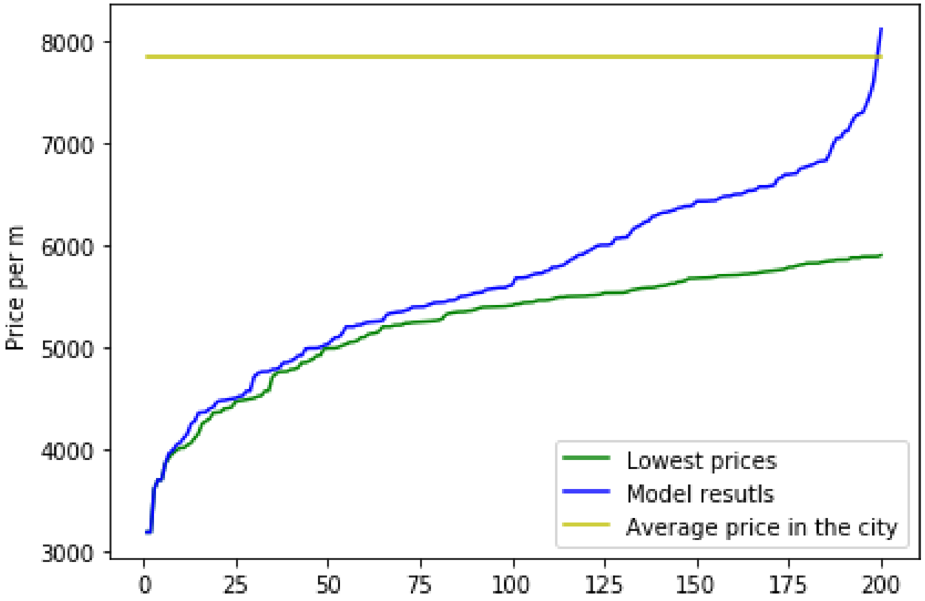


It was validated how good they are at predicting prices by using R2 score. The winner model was **Linear regression.** But R2 was very low even for best model which suggests that there is more information influencing prices of apartments.

Results

model results vs. simply sorting by lowest price

- I’ve compared 10% of listings with biggest difference in price (compared to y_hat) against the list of listings with lowest prices. A result was a list in which **almost 60% of listings were the same**.
- When I've visualized the results, we can see that model is more likely to suggest listings with prices closer to average price in the city than simply listing apartments with lowest prices.



Discussion

- Based on the results of the project we can see that there is a number of information impacting the price but missing in the data. Unfortunately I was not able to parse this information because it was either unavailable at all or could be inferred only from picture or text descriptions.
- More information could be added like: standard of the apartment, building condition, how well rooms are arranged.
- Missing data which were replaced during preparation could be filled in.
- Both of the above would have to be done manually but this should improve results of models.
- In future an ability to add data incrementally about new listings and venues would be a plus as this project was based only on active listings. My hope is that with more historical data, we could get better predictions.

Conclusion

- Based on the Linear Regression model and Foursquare, geo, real estate data, I was able to identify potential good deals in a price ranges closer to average price of an apartment in the city. This can be considered as satisfactory result as such information is not available when filtering only by lowest price. Unfortunately model is not very good at predicting prices so further work is needed – most likely manually correcting the data – to achieve better results.
- The goal of this project was to suggest potential good deals, but making the final decision on purchase consists of more criteria that need to be evaluated by the buyer.