Predicting house prices in King County, Washington, USA

By Panagiotis Petsas

This report is part of the course "Applied Data Science Capstone" from "IBM Data Science Professional Certificate" on Coursera.

Part 2: Data

1 Introduction

(in the previous document)

2 Data acquisition and cleaning

2.1 Data sources

We obtained data on houses in King County from Kaggle **[3]**. This dataset contains more than 21.000 house records, providing information on features like price, year build, number of bedrooms, longitude, latitude, zip code etc. The complete list of features is listed in **Table 1**.

In order to reduce computation time, we decided to work only with a specific area. After some initial data cleaning (see below), we scraped the area phone code of each house by utilizing its zip code. We maintained records with area phone code 253. These houses are located in an area near Tacoma, a medium sized city (**Figure 1**). We used 'www.getzips.com' site in order to acquire the area phone codes. This procedure reduced the house records to 2766, a more manageable size for modeling.

Figure 1 Houses in King County, Washington with area phone code as 253, marked with blue dots.

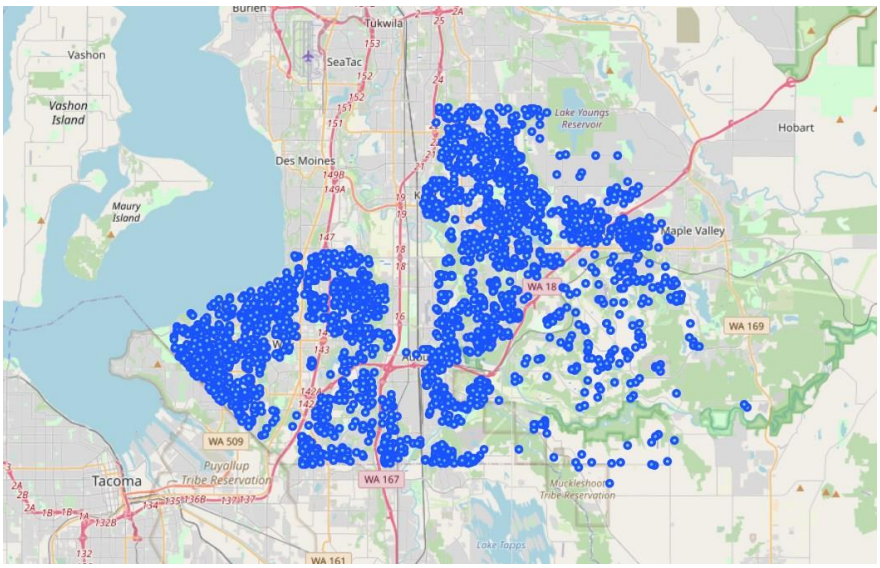

Table 1 Features in the dataset of houses in King County, Washington, along with their explanation.

Feature name	Explanation
ID	Unique id for each house record
Date	Date of home sale
Price	House price
Bedrooms	Number of Bedrooms
Bathrooms	Number of Bathrooms
Sqft_living	Apartment inter living space (in sq feet)
Sqft_lot	Lot space (in sq feet)
Floors	Number of floors
Waterfront	Binary (1 if the apartment overlooks the waterfront, 0 otherwise)
View	From 0-4 integer, a grade for house's view
Condition	From 1-5 integer, the condition of the house
Grade	From 1-13 integer, the construction and design level
Sqft_above	Apartment inter living space above ground (in sq feet)
Sqft_basement	Apartment inter living space below ground (in sq feet)
Yr_built	The year the house was built
Yr_renovated	Year of last renovation (0 if never renovated)
Zipcode	The zipcode of the area
Lat	Latitude
Long	Longitude
Sqft_living15	Apartment inter living space (in sq feet) for the 15 nearest neighboring houses
Sqft_lot15	Lot space (in sq feet) for the 15 nearest neighboring houses

Having the longitude and latitude of each house, we used Foursquare API **[4]** in order to extract the most popular venues in a proximity of 500 meters from it. We limited the API to the top 30 venues per house, in order to reduce computation time. This process derived a total of 8995 venues near the houses, providing information about their name, category and coordinates.

2.2 Data cleaning

The dataset did not contain any missing records. Some initial data cleaning was conducted by removing strange records (houses with zero bedrooms or bathrooms). The dataset contained a feature that indicated the year of renovation. Most houses were not renovated, thus they received a value of 0. For this reason, we created a new feature that indicated whether a house was renovated (1 if yes and 0 if not). After these adjustments, the next part of the cleaning process was to maintain the records with area phone code of 253.

The next part was to group the venue records based on their category. While Foursquare provides their own general categories, we preferred to classify the venues in our custom groups. We created 13 distinct groups (**Table 2**) to classify a total of 254 unique venue categories. The grouping was performed by searching on keywords within the venue category names. For example, venue category names including the word 'restaurant' would be put in the group of 'places to eat'. We ensured that categories that were grouped were removed from the grouping process, in order to avoid grouping them twice. There were some venue categories that could not be grouped based on keywords, which we grouped manually to the correct group.

In the next session, we summarized the amount of venues per group for each house. This process created 13 new features, one for each group, indicating the amount of venues of this group in the near proximity of each house. For example, a house with 2 places to eat and 3 venues related to services would get a value of two in the 'places to eat' feature, 3 in the 'venues related to services' features and 0 to the remaining 11 features related to venue groups.

2.3 Feature selection

Since our purpose is to use features to predict the house price, we must maintain features that could provide such information. We initially discarded 'id' and 'date' (of recoding the house), as they are used for reference only. Similarly we discarded the 'zipcode' and 'area phone code'

as they provide no further information. The year of renovation feature was replaced by the binary feature ‘is renovated’ stated above.

We performed pearson correlation analysis to identify linear patterns among the remaining features. We decided to discard four more features that caused high correlation, being ‘above area in sqft’ (after all, we maintained the ‘total area in sqft’, which is exactly equal to the summary of the ‘above’ and ‘basement area in sqft’), ‘grade’ (which is a subjective value anyway), ‘sqft_living15’ and sqft_lot15’ which indicate the total living and lot area of the 15 neighboring houses, and have high correlation with their respective features. With that procedure, we maintained 26 features for predicting the house prices.

Table 2 Groups of venue categories, along with some keywords used to identify the right venues.

Group	**Some keywords used**
Places to eat	Restaurant, food, pizza, burger, chicken, fish, breakfast, ...
Places to drink	Bar, wine, beer, pub, coffee, café, ...
Places with sweets	Candy, donut, dessert, pie, ice cream, ...
Amusement	Playground, arcade, casino, roller, night, ...
Facilities	Facility
Sports	Soccer, basketball, baseball, sport, gym, fit, ...
Transport	Airport, boat, ferry, bus, train, ...
Shops	Shop, market, store
Services	Service, assist, bank, doctor, police, ...
Beauty services	Spa, massage, salon, beauty, tattoo
Attractions	Art, history, attract, museum, ...
Landmarks	Mountain, lake, tree, ...
Accommodation	Hotel, motel, inn, rest, ...

[3] House prices in KC, Washington – Kaggle

(https://www.kaggle.com/harlfoxem/housesalesprediction)

[4] Foursquare API (https://developer.foursquare.com/)