Capstone project IBM Data Science

1 Introduction

House prices have huge margins and prices are often very volatile. Predicting house prices using data can lead to a more efficient housing market. The goal of this research will be to quantify subjective house prices. Real estate agents often give certain properties to houses, e.g. number of bedrooms, bathrooms, square meters and location. With data becoming more publicly available, distances to certain venues and number of venues in the area can be incorporated in the pricing models. The goal of this research is to find out which type of venues or whether venue data increase the price of a house in King County, USA. To be more precise, we will determine the effect of the number of venues within each neighborhood on the price.

2 Data

The data that are used is the King County data set[1]. The dataset countains house prices, the date the house was sold, number of bedrooms, bathrooms, square foot of the house, living room and lot, if the house has a view to a waterfront, if the house has been viewed, general conditions, grades based on King County grading system, year the house was built, year the house was renovated, and finally some location data. We will also be using data from Foursquare[2] to find venue types in the area.

2.1 Neighborhoods

Figure 1 shows the top ten most expensive neighborhoods of King County. As we can see, Medina is the most expensive neighborhood with an average price of 1,800,000 dollars. We dive deeper into the features of every neighborhood. Table 1 shows the top ten neighborhoods including their average number of floors, bedrooms, bathrooms and grade. We see small dispersions accross neighborhoods in their features that are correlated with the price. We might find added value in venue data.

2.2 Venue data

We retrieve venue data by looking at venues in a radius of a 1000 meters from the center of each neighborhood. This gives us an idea of how crowded the center is, and what kind of venues are centered in each neighborhood. There are a total of 257 unique venue categories to be found accross all neighborhoods. The exact categories can be found in Appendix A.

3 Methodology

The goal of this research is to see whether venue data have predictive power in estimating house prices. We combine data from the King County data set and retrieve the neighborhood the house is located. Furthermore, we will analyse each neighborhood and find the number of

[1] Data is retrieved from https://data.kingcounty.gov

[2] Data retrieved from https://foursquare.com/

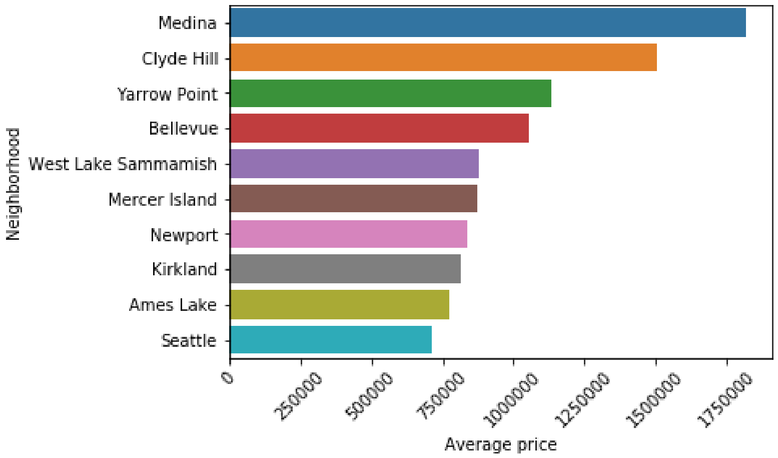

Figure 1: Top 10 most expensive neighborhoods.

Table 1: Some features of the top ten neighborhoods

Neighborhood	Price ($\times 10^5$)	Floors	Bedrooms	Bathrooms	Grade
Medina	18.17	1.47	3.86	2.90	9.17
Clyde Hill	15.06	1.38	4.07	2.67	8.76
Yarrow Point	11.33	1.56	3.59	2.41	8.57
Bellevue	10.54	1.43	3.66	2.41	8.49
West Lake Sammamish	8.770	1.54	3.79	2.57	8.75
Mercer Island	8.741	1.45	3.63	2.29	8.11
Newport	8.381	1.28	3.72	2.40	8.19
Kirkland	8.145	1.50	3.52	2.35	8.05
Ames Lake	7.771	1.84	3.79	2.81	9.15
Seattle	7.101	1.70	3.22	2.09	7.72

The features price in dollars, floors, bedrooms, bathrooms and grade denote the average value.

venues within a category inside the neighborhood. We group the categories of even further and look specifically at the following categories: restaurants, studios and bars. To predict house prices we perform different regression techniques, such as linear regression, ridge regression and random forest model. The data is separated in a training set, which contains 80% of the data and a test set, which contains 20% of the data. To conclude whether venue data improve the prediction, we compare mean squared errors of the test set.

4 Results

Table 2 shows the results of three different models. We find lower mean squared errors (MSE) accross all models in out-of-sample data. This indicates that venue data indeed has forecasting power for house prices. Furthermore, we find a positive sign for the number of restaurants and bars, but a negative sign for the number of studios.

5 Discussion

The reason for the different sign in the categories could be that studios are often to be found in the outskirt neighborhoods, while restaurants and bars might be found in central neighborhoods.

With this knowledge, one could predict the prices of new houses that are going to be build in which area. Cities are getting bigger and bigger and house prices are often very unstable when they are built in new areas. Using venue data we can get a better estimate of the true value of a house in a certain location.

Table 2: Mean squared error of different models

Model	Without venue features ($\times 10^{10}$)	With venue features ($\times 10^{10}$)
Linear regression	4.012	3.672
Ridge regression	4.012	3.672
Random forest	2.259	2.105

This table shows the mean squared error (MSE) for three different model in out-of-sample data.

6 Conclusion

Our results show that indeed venue data can be exploited to predict house prices. We showed that restaurants and bars have a positive effect on the price, while the number of studios in the city centre have a negative effect on the price. Possible reasons could be that certain venue categories are often located in areas that are further away from the centre, causing the house prices to be lower. There is a lot of future research to be done with this data. One could for example look at more categories. This will give us a fuller picture of all possible venue categories effects on the price of a house. Another possibility is that one could ignore categorising houses in neighborhoods. We could for example look at the latitude and longitude of each house and calculate the distance to certain venues, or look at the number of certain types of venues in a radius of the house. Unfortunately, this was not possible in our research due to lack of resources and computing power. All in all, predicting house prices is an important field of research which allows to create a more efficient housing market.

A Venue categories

This is a list of all venue categories: Food, Rock Club, Bar, Convenience Store, Pizza Place, Construction Landscaping, Flower Shop, Home Service, Gift Shop, Farm, American Restaurant, Coffee Shop, Brewery, Sushi Restaurant, Ice Cream Shop, Thai Restaurant, Italian Restaurant, Café, Park, Dive Bar, Automotive Shop, Vietnamese Restaurant, Storage Facility, Sandwich Place, Grocery Store, Noodle House, Bank, Fast Food Restaurant, Mexican Restaurant, Performing Arts Venue, Breakfast Spot, Recreation Center, Japanese Restaurant, Thrift / Vintage Store, Gas Station, Smoke Shop, Football Stadium, Adult Boutique, Steakhouse, Sporting Goods Shop, Boutique, Toy / Game Store, Dessert Shop, Wine Shop, Gym, Sports Bar, Shopping Mall, Seafood Restaurant, Bagel Shop, Bakery, Hotpot Restaurant, Furniture / Home Store, Cupcake Shop, Spa, Hotel Bar, Wine Bar, Womens Store, Hotel, Burger Joint, Botanical Garden, Jazz Club, Dumpling Restaurant, Pet Store, Gym / Fitness Center, Bookstore, Cocktail Bar, Bubble Tea Shop, Stationery Store, Kitchen Supply Store, Salon / Barbershop, Museum, Lingerie Store, Brazilian Restaurant, Cycle Studio, Electronics Store, Poke Place, Donut Shop, Movie Theater, French Restaurant, Mediterranean Restaurant, Chinese Restaurant, Library, Clothing Store, Art Museum, Lake, Beer Store, Food Drink Shop, Pool, Whisky Bar, Gaming Cafe, Diner, Pub, Farmers Market, Gastropub, Rental Car Location, Supermarket, Pharmacy, Massage Studio, Music Venue, Taco Place, Mattress Store, Golf Course, Juice Bar, Video Store, Market, Playground, Garden, Baseball Field, Deli / Bodega, Butcher, New American Restaurant, Credit Union, Australian Restaurant, Bistro, Hawaiian Restaurant, Discount Store, Arts Crafts Store, Dog Run, Skate Park, Athletics Sports, Lebanese Restaurant, Mobile Phone Shop, Theater, Soccer Field, Shopping Plaza, Nail Salon, Food Truck, Liquor Store, Record Shop, Shoe Repair, Shipping Store, Trail, Business Service, Moving Target, IT Services, Department Store, Asian Restaurant, Locksmith, Marijuana Dispensary, College Basketball Court, Hardware Store, Middle Eastern Restaurant, Other Repair Shop, Skating Rink, Optical Shop, Dance Studio, Indian Restaurant, Auto Dealership, Korean Restaurant, College Bookstore, Parking, Planetarium, Tennis Court, College Baseball Diamond, College Track, Cafeteria, Gymnastics Gym, Restaurant, Karaoke Bar, Snack Place, Martial Arts Dojo, Surf Spot, ATM, Memorial Site, Basketball Court, Fried Chicken Joint, Cosmetics Shop, Beach, Intersection, Kids Store, "Mens Store", Bus Station, Print Shop, RV Park, Harbor / Marina, Camera Store, Bowling Alley, Bike Shop, Pier, Auto Garage, Hot Dog Joint, Greek Restaurant, Wings Joint, Tanning Salon, Video Game Store, Lounge, Warehouse Store, Outdoor Sculpture, Boat or Ferry, Bay, Beer Bar, Bus Stop, Residential Building (Apartment / Condo), Mini Golf, Insurance Office, Art Gallery, Garden Center, Office, Amphitheater, Nature Preserve, Water Park, Theme Park Ride / Attraction, Rest Area, BBQ Joint, Recording Studio, Child Care Service, Shoe Store, Comic Shop, Big Box Store, Dim Sum Restaurant, Accessories Store, Supplement Shop, Community Center, Frozen Yogurt Shop, Indie Movie Theater, Outlet Mall, "Dentists Office", English Restaurant, Chocolate Shop, Irish Pub, Financial or Legal Service, Yoga Studio, Hobby Shop, Salad Place, German Restaurant, Other Great Outdoors, Arcade, Tea Room, Vegetarian / Vegan Restaurant, Burrito Place, Southern / Soul Food Restaurant, Photography Studio, Police Station, Light Rail Station, Airport, Summer Camp, Boxing Gym, Resort, Shop Service, Airport Terminal, Herbs Spices Store, Event Space, Track, Drugstore, Music Store, Latin American Restaurant, Rental Service, Motorcycle Shop, Bus Line, State / Provincial Park, River, Casino, Indoor Play Area, Field, Miscellaneous Shop, Food Service and Scenic Lookout