Austin Neighborhoods, A Look at Venues and Real Estate Prices

Jonna Pander

July 15, 2020

Problem Statement:

Can nearby venues be used to predict whether the average price per square foot for a neighborhood in Austin, Texas is, on average, less or greater than 400 dollars per square foot.

Executive Summary:

Austin is one of the hottest real estate markets in the country right now. What's driving this boom? Is it the amenities and entertainment the city has to offer or is it Californians and their silicon valley money! The goal of this project is to explore whether or not venue types in neighborhoods throughout Austin, Texas contribute to the average price per square footage of a home for that neighborhood. People who are interested in this information are those looking for what neighborhood to purchase a home in based on the right mix of price and nearby attractions for them.

Location data for 37 neighborhoods in Austin was obtained along with the average price per square foot and a compilation of venue types within a 4000 meter radius of the neighborhood's center. I ended up with a total of 247 unique venue types. Names of neighborhoods was obtained from Wikipedia via web scraping, real estate data was obtained from neighborhoods.com via web scraping, and venue data was obtained from Foursquare using an API.

I treated this as a classification problem and grouped the prices into two (2) categories, one for real estate less than and average of 400 dollars per square foot and real estate that is, on average, greater than 400 dollars per square foot.

Two uncommon libraries were used in this project. Geopy was used to pull latitude and longitude coordinates for each city and Folium was used to visualize the locations of cities in a map.

Data Sources

Location data was obtained by webscraping Wikipedia for a list of neighborhoods in Austin, Texas. (https://en.wikipedia.org/wiki/List_of_Austin_neighborhoods) Real Estate data detailing the average prices per square foot per neighborhood was obtained from neighborhoods.com (https://www.neighborhoods.com) Venues for each neighborhood were obtained from Foursquare using an API. (https://api.foursquare.com/v2/venues/explore)

Data Dictionary

Feature	Type	Dataset	Description
Neighborhood	*object*	df_comb	Name of neighborhood
Venue Type	*object*	df_comb	Name of each venue type
Price Group	*int*	df_comb	0 for price <400/sq.ft. and 1 for >400/sq.ft.

Metrics

Several models were applied to the problem. I used Principal Component Analysis (PCA), Support Vector Machines (SVM), Decision Trees, and Logistic Regression. The Decision Tree, SVM, and Logistic Regression models all scored a Testing Accuracy score of 0.875 with Logistics Regression being the least overfit to the training data. The PCA model yielded R2 Score of 0.28 and Mean Squared Error (MSE) of 0.17. Since the data set is so large with a total of 247 columns representing venue types I am using PCA with coefficients from Linear Regression to explain the correlations of venues to real estate prices.

Limitations

There were several challenges in this project. For one, getting a complete set of data was not possible. Either geopy did not return coordinates or neighborhoods.com did not have sales data. This resulted in going from and original list of 59 neighborhoods to 37. Additionally, a dataset of neighborhoods in just one city yields a small test set. This became less of an issue when I turned my target into a binary classification instead of 20 price ranges.

Conclusions

Using PCA and Linear Regression for coefficients to assess the impact of venues types on real estate prices, the venue types with the most positive correlation or effect are:

- Hotel Bar
- Coffee Shop
- Vegetarian/Vegan Restaurant
- Food Truck
- Beer Garden

Venue types with the most negative effect are:

- Discount Store
- Fast Food Restaurant
- Wings Joint
- Arts & Crafts Store
- Electronics Store

Venue types with the least correlation are:

- Student Center
- Botanical Garden
- American Restaurant
- Wine Bar
- Gastropub