Using Geospatial Data to Find Investment Opportunities in the Real Estate Market

Nikhil Ramesh, Chih-Sheng Huang, Ignacio Navarro

October 24, 2019

1 Introduction

Real estate is a trending field in computation as predicting valuations of properties and getting in early on a market can lead to high monetary returns in the future.

People can use many online resources to see housing prices like Zillow and Trulia when they want to buy a house or invest in a property. However, most people don't know whether or not the property is being overvalued or undervalued. The property price will be effected by many factors like the distance to central city, the population, is a good school near the house, is near to supermarket, etc. We want to analyze those features and add the factors to our machine learning model, then use our model to predict that is the house price overvalued or undervalued.

2 Objective

In real estate, housing prices are usually determined by various internal factors relating to the house as well as external factors such as neighborhood, nearby amenities, etc. Our goal is to identify more external factors that may have a subconscious effect on house valuation / final selling price and use this to our advantage. Specifically, we will try to find houses that are being undervalued or overvalued based on their listing price.

To do this, we can hone in on a specific city (ex. Philadelphia) and select houses that have sold during a time period; this can be the final housing valuation that a buyer decided was the worth of the house upon purchase. We can then do some data exploration and feature engineering to expand the feature space of this housing training set based on factors we believe affect sale price. Finally, we can train a model on this engineered dataset to predict the housing price of unsold houses and compare this predicted price to the listed price to see if we should purchase the house at an undervalued rate or if the seller is overvaluing the property.

3 Project Roadmap

1. **Acquiring data and data exploration**: one of the main challenges in the project will be to acquire and clean the data for further exploration. Real estate data is scattered in many places, and although it is straightforward to acquire data explicitly related to the property (e.g. number

of bedrooms, bathrooms, square feet, parking), it is more complex to acquire geospatial data related to the property. By geospatial data we mean any data externally related to the property: proximity to a river, nearby schools, number of coffee places, gas stations, etc. This will directly require a lot of data exploration to see what geospatial data is correlated to house prices.

2. **Feature engineering**: Once we have acquired the domain knowledge of the data, we will need to tailor this to create features that make our models work. We will have a heavy mix of numerical (both discrete and real), and categorical features to work with, so representing them in one way or another will impact the performance of our models.

3. **Running and evaluating experiments**: Since we are shaping our thesis as a regression problem, we will train our data with standard regression algorithms such as SVM, Perceptron, regression trees, random forests, etc. We will finally compare all of them to find out the best model for our purposes.