

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

Chih Huang

huangcs
@seas.upenn.edu

Ignacio Navarro

inavarro
@seas.upenn.edu

Nikhil Ramesh

nramesh
@seas.upenn.edu

Abstract

Valuating real estate properties is driven by a series of direct factors (total square area, number of bedrooms, garage spaces, etc) and indirect factors (crime rate, quality of school district, etc). However, private sellers tend to focus more on the former when setting a listing price for their property. In particular, there are many geospatial features indirectly related to the property that sellers might easily overlook. Proximity to a church, a river, or a police station are all factors that play a role, albeit minor, in the valuation of a property. In this paper we determine by running a series of models on a dataset consisting on properties in South Philadelphia if geographical features do indeed play a decisive role in the valuation of a property.

1 Introduction

1.1 Motivation

Traditionally, real estate valuation has concentrated on assessing the features directly surrounding the property to come up with a price. Straight-forward features such as number of bedrooms, whether the property has a pool or a garage space, or even the number of bathrooms play a major contribution on the price. And while that makes perfect sense, private sellers may not fully realize that the property's surroundings can also play a role. Points of interest like supermarkets, museums, parks, cafes, or police stations can affect house prices but many times these may be overlooked.

1.2 Problem Definition

With all of this in mind the natural question becomes: given all the geospatial data related to a

property, can a machine learning model predict the real value of a property as a **regression** problem? And if so, can one use this to find investment opportunities in the real estate market? Throughout this paper we will answer the former question. The latter is a simple corollary of the former. Indeed, if there is a model that can accurately predict house prices given geospatial and direct features, one can use these results to find investments opportunities by comparing the listing price on properties currently on the market with what the best model predicts the value is and seeing if the property is really undervalued.

In other words, let $M : X \rightarrow \mathbb{R}^+$ be an accurate model (we will define what constitutes “accurate” to our use case in the next section), where X is the feature space that constitutes direct and indirect features related to a property. Given a property p with a listing price of s_p and a set of features x_p , we will buy the property if

$$M(x_p) > s_p \quad (1)$$

because p is undervalued.

1.3 Related Work

A lot of work has been done in the last few years on predicting house prices using machine learning. Traditional models have relied in hedonic regression, a type of regression which breaks the property apart into its direct factors in order to establish a relationship between the factors and the price of the property. A paper using this method have been presented by (?) focusing on properties in Singapore with property sales between 1995 and 2014.

Our main source for our project is (?), in which they analyze a dataset consisting of properties in Madrid, Spain. There are similarities to what we're trying to investigate: the paper models the question as a regression problem and uses similar

models to the ones we train. The main difference, however, is that they only consider direct features for their models. The main novelty in our project is to combine these features with geospatial features. As far as we have researched, no work has investigated our proposal.

2 Data Acquisition and Exploration

2.1 Acquisition

The acquisition of the dataset for this project consisted on finding basic property information for Philadelphia that included direct features, and merging this dataset with all the geospatial information related to the property. The first part was relatively easy, as <https://www.opendataphilly.org/> provided basic information for over 250,000 properties in Philadelphia. The main challenge was in fact finding the geographical features. Our approach to solve this was to find APIs that offered nearest distance to points of interests. Once we found some reliable APIs for this task (e.g. GoogleMaps API, TomTom API, or FourSquare API), we began the process of merging the data to our property dataset.

The main obstacle of this part, and probably of the project, was the API rate limit each service offered. Conservatively speaking, we needed to make

$$\underbrace{250,000}_{\text{properties}} \times \underbrace{24}_{\text{geo features}} = 6,000,000 \text{ API calls}$$

and unfortunately most of the services offered 1,000 API calls limit per day. The solution to this part was to drastically reduce the size of the dataset to around 8,000 properties in South Philadelphia, and even this amount posed some architectural challenges: we distributed the API calls for each member of the group and remerged them later. We encourage the reader to look at the code dealing with this on the project's [repo](#), in particular the `properties/data` package.

2.2 Features

The list of final features that we feed our model are in Table ?? for a total of 5 categorical and 28 numerical.

2.3 Exploration

Once we acquired the data the next step was to explore it to fully understand it. Some significant results to show are the following.

| Feature | Type |
|---------------------------|-------------|
| Fireplaces | categorical |
| Garage Spaces | categorical |
| Bathrooms | categorical |
| Bedrooms | categorical |
| Stories | categorical |
| Total Area | numerical |
| Total Livable Area | numerical |
| Latitude | numerical |
| Longitude | numerical |
| Nearest Museum | numerical |
| Nearest Gas Station | numerical |
| Nearest Coffee Shop | numerical |
| Nearest Stadium | numerical |
| Nearest Food | numerical |
| Nearest Bar | numerical |
| Nearest Gym | numerical |
| Nearest Bridge | numerical |
| Nearest Garden | numerical |
| Nearest Park | numerical |
| Nearest River | numerical |
| Nearest City Hall | numerical |
| Nearest Police Station | numerical |
| Nearest Hospital | numerical |
| Nearest Elementary School | numerical |
| Nearest Church | numerical |
| Nearest Bank | numerical |
| Nearest Supermarket | numerical |
| Nearest Pharmacy | numerical |
| Nearest Bus Stop | numerical |
| Nearest Metro Station | numerical |
| Nearest Train Station | numerical |
| Nearest University | numerical |
| Nearest Laundromat | numerical |

Table 1: Features used

The distribution of the properties by sale price in South Philadelphia can be seen in Figure ?? . Clearly the distribution is skewed with a mean sale price of \$324,662.



Figure 1: Distribution by sale price

We can also get a sense of where the properties are located by plotting a heatmap of South Philadelphia based on sale price (Figure ??).

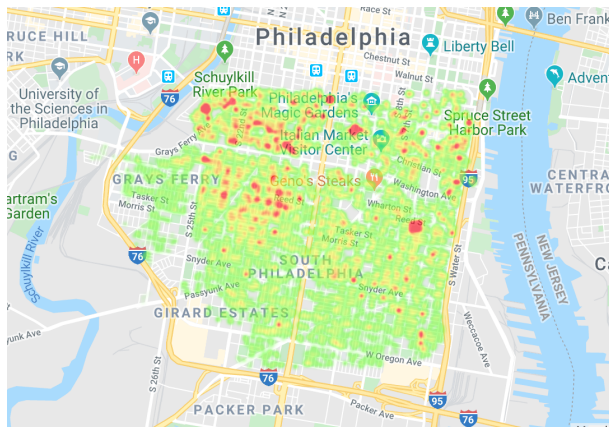


Figure 2: Heatmap of South Philadelphia (hotter means higher sale price)

Another important step was to see the relationship between the features. To do this we used a correlation matrix which can be seen in Figure ?? . For instance, there is a strong correlation between the number of bathrooms and the total livable area. There's also some interesting relationships, e.g., there is negative correlation between the total livable area and the nearest laundromat, which makes (some) sense, as bigger houses in richer neighbor-

hoods don't need laundromats.

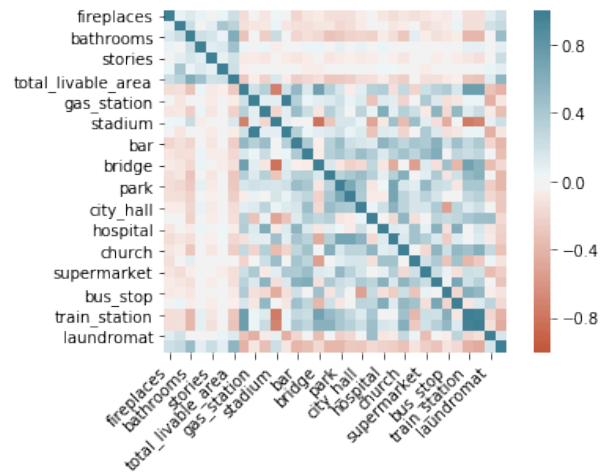


Figure 3: Correlation matrix

3 Models

3.1 Experimental Models Setup

3.2 Results and Findings

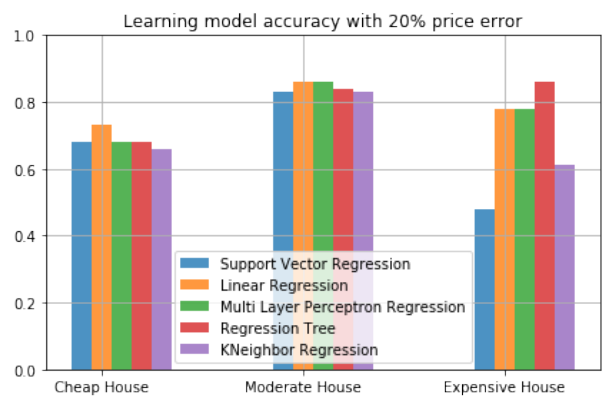


Figure 4: Accuracy comparison for five learning models which are used to predict the cheap, moderate, expensive house.

Using the geospatial features and direct factors (number of bedrooms, spaces, etc) to train five learning model: Support Vector Regression, Linear Regression, Multi-Layer Perceptron Regression, Regression Tree and KNeighbor Regrsssion, the accuracy with 20% tolerance for each model are listed in Figure ?? . According to the result, the five models have almost the same accuracy to predict cheap house and moderate house. However, there are only three models: Linear Regression, Multi-Layer Perceptron Regression and Regression Tree, have better accuracy to predict expensive house.

In order to find which is the best performance model, the three best learning model: Linear Regression, Multi-Layer Perceptron Regression and Regression Tree will be used to predict the properties price with lower tolerance for further analysis. And the comparison between prediction data and test data will be discussed in next section.

3.3 Find the Best Model

In previous analysis, the three models with higher accuracy are Linear Regression, Multi-Layer Perceptron Regression and Regression Tree will be selected to predict the properties price with 15% tolerance. The result is shown in Figure ??.

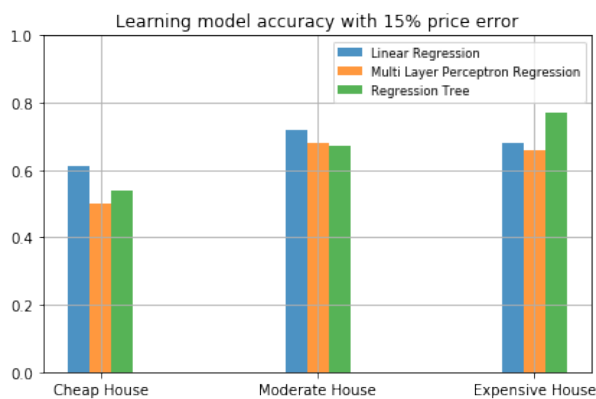
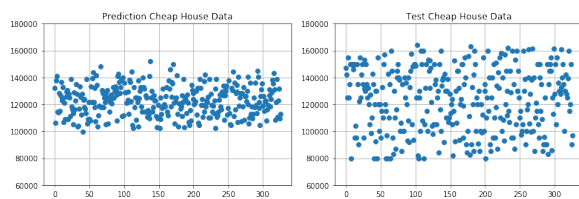


Figure 5: Accuracy comparison for three learning models which have better performance in previous prediction result.

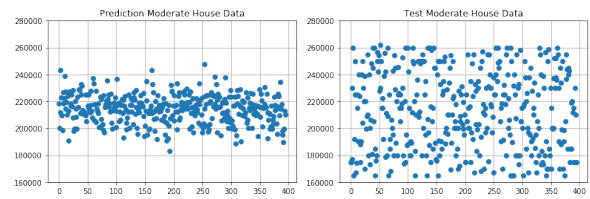


(a) Prediction data for cheap house (b) Test data for cheap house

Figure 6: Prediction and test data distribution comparison for cheap house by using linear regression model.

The accuracy decreasing with the tolerance 15% for three models, and it is obvious that Linear Regression and Regression Tree model have better accuracy performance than Multi-Layer Perceptron Regression model. In order to find which one is the best model, analyze the prediction data and test data distribution to see which model can

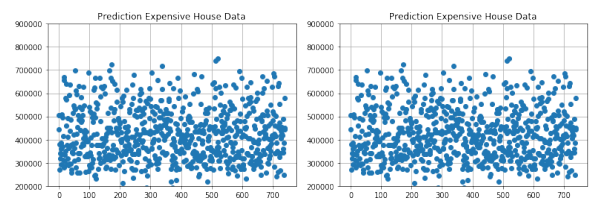
mimic the test data distribution. First, the prediction data distribution are show in Figure ??, Figure ??, Figure ??



(a) Prediction data for moderate house (b) Test data for moderate house

Figure 7: Prediction and test data distribution comparison for moderate house by using linear regression model.

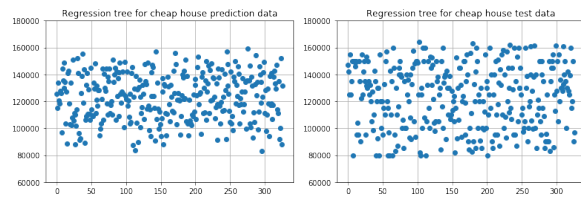
According the result which are shown in the figure, Linear Regression model predict the cheap and moderate properties price in a small range which can't represent the true test data. Only expensive house result have the similar distribution between prediction and test data.



(a) Prediction data for expensive house (b) Test data for expensive house

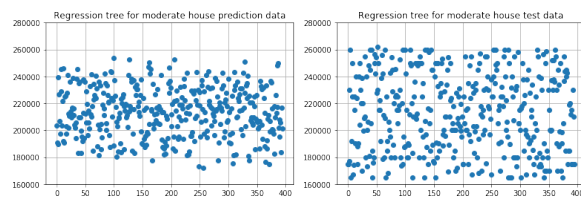
Figure 8: Prediction and test data distribution comparison for expensive house by using linear regression model.

After comparing the distribution for prediction data and test data by using Linear Regression model, then examine the data distribution result by using regression tree learning model. The data distribution are listed in Figure ??, Figure ??, Figure ??, they are very similar between prediction data and test data in three group house. Even the accuracy seems no difference by comparing linear regression and regression tree model, the prediction data have a better prediction properties price range by using regression tree model.



(a) Prediction data for cheap house (b) Test data for cheap house

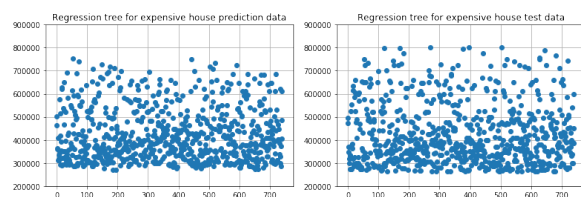
Figure 9: Prediction and test data distribution comparison for cheap house by using regression tree model.



(a) Prediction data for moderate house (b) Test data for moderate house

Figure 10: Prediction and test data distribution comparison for moderate house by using regression tree model.

By examining the two best model and analyze the data distribution for prediction and test data, regression tree model can capture the properties price distribution characteristics. Therefore, the best model to predict the house price will be regression tree model by using our features to train.



(a) Prediction data for expensive house (b) Test data for expensive house

Figure 11: Prediction and test data distribution comparison for expensive house by using regression tree model.

4 Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris

nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

4.1 Future Work