Identifying a Potential Housing Bubble in Pennsylvania

Ilkin Bayramli, Wanxin Ye Dhungjoo Kim, Jialan Zhu

February 24, 2019

1 Executive Summary

In this report, we outline our team's findings on existence of a potential housing bubble in the state of Pennsylvania. We developed a Machine Learning model to explain the variance between different housing groups in

2 Technical Exposition

2.1 Tools:

We conducted our analysis using *Python* and *R*. Packages used included *numpy, pandas, matplotlib, sklearn* for Python and *tidyverse, ggplot2, dplyr, lubridate*.

2.2 EDA and Feature Engineering

Our team performed EDA (Exploratory Data Analysis) in order to impute missing values, remove outliers, and investiate feature correlations in the dataset. We also conducted feature engineering with Python to remove unnecessary features and combine highly correlated variables to create new features.

2.3 Supervised Machine Learning

We built two essential ML models for this project. The first model was used to solve the regression problem and predict the house prices. For housing prices, we tried to use both linear (Lasso) and nonlinear models (Random forest and XGboost). From the prediction vs. real data comparison figures, we found out that the model perform poorly at low house price region. Therefore, we decided to implement a second model to solve the classification problem and investigate whether the house going to be a distressed sell or not. The model has accuracy as high as 0.95 in both train and test datasets.

3 Analysis

3.1 Observation

While constructing a scatterplot (Figure 1) to determine the relationship between the sale price and the assessed real value of houses in states of Pennsylvania (PA), Massachusetts (MA), and Rhode Island (RI), we noticed a lack of a unifying trend in the Pennsylvanian housing market. While in MA and RI, the price - assessed value relationship follows a single strict linear trend, in PA such a relationship does not exist. Instead, it can be inferred from the scatterplot that in PA there exists two housing groups (let's call them group A and group B) that follow different linear value-price trends. While the group

A exhibits a trend very similar to statewide trends in MA and RI, the group B forms a trend of its own. Although group A follows a nearly one-to-one price-assessed value relationship, group B demonstrates a noticeable gap between sale price and assessed value of the houses. We hypothesize that the behaviour of Group B indicates an over-inflated housing market and a housing bubble.

Figure 1: Scatter Plot of Sale Price and Assessed Total Value Relationship in States of MA, PA, RI

3.2 Geographic Distribution of Group A and B

Our team tracked down the geographic distribution of Group A and Group B houses. By plotting (Figure 2) the assessed real value - acres gap against acres, we were able to observe a negatively proportional relationship. In other words, the houses in over-inflated markets occupy less amount of acres. It can be inferred from this that over-inflated houses are mostly in urban regions. To confirm this, we graphed the distribution houses on map by the size of the real value - acres gap and found out that these houses are predominantly concentrated around Philadelphia and Pittsburgh.

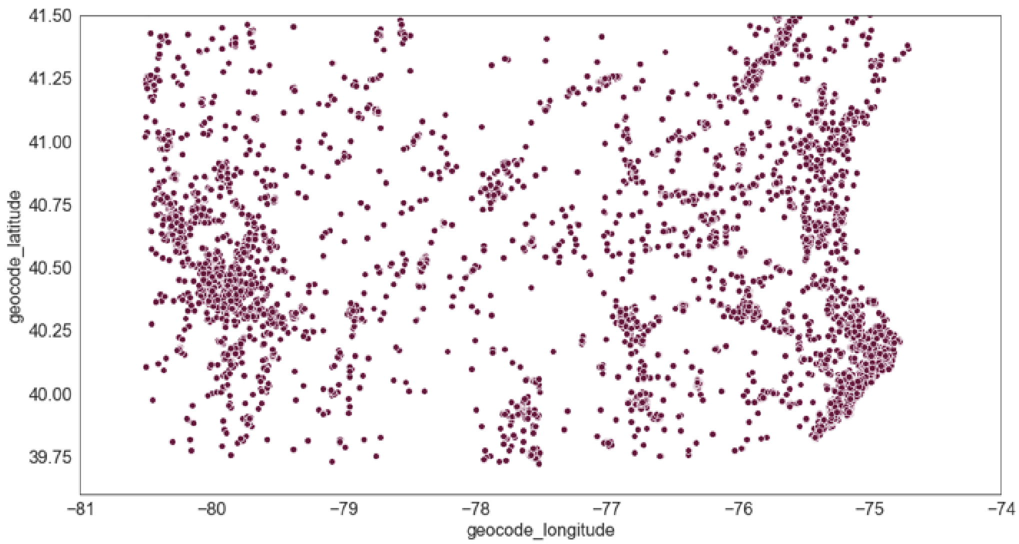

Figure 2: Geographic distribution of cities with highest value - price gap

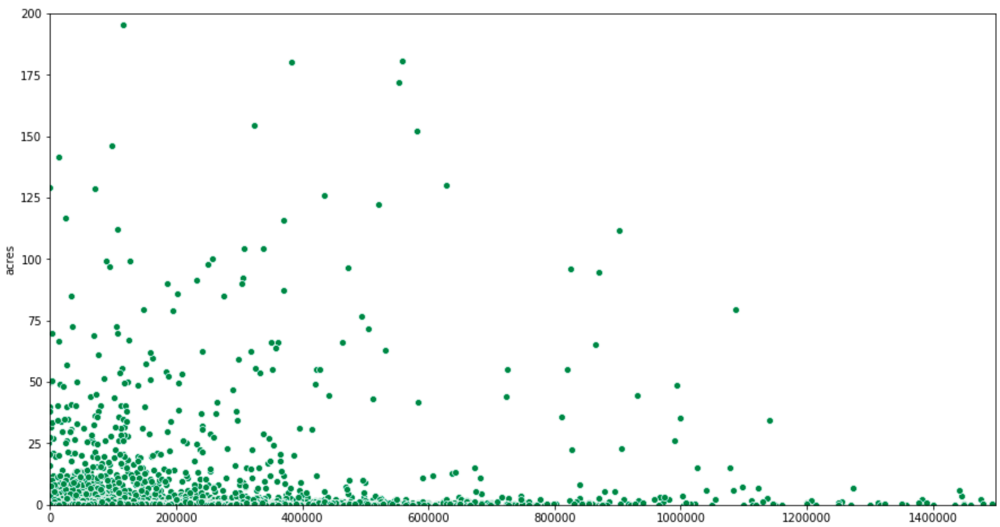

Figure 3: Assessed Real Value - Acres Difference (x axis) versus Acres (y axis)

3.3 Houses traded at a higher price

We also transformed the Citizenbank dataset to obtain a list of houses that have been traded for the highest price (), and unsurprisingly, those were concentrated around Philadelphia as well. As the graph above shows, there is a ring of houses that have been traded for a price between USD 600,000 and 1,000,000.

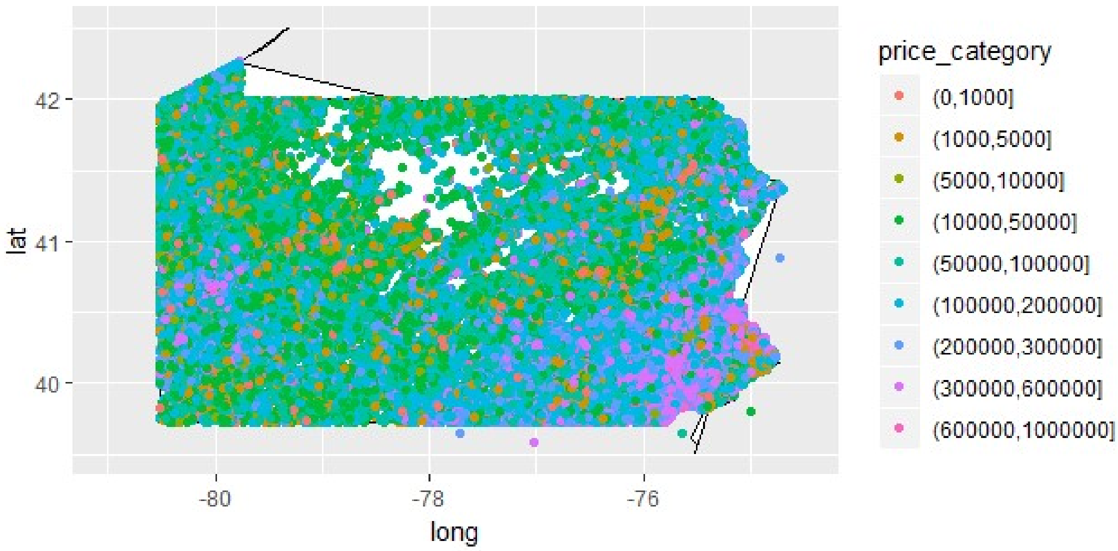

Figure 4: Geographic Distribution of Houses by Price

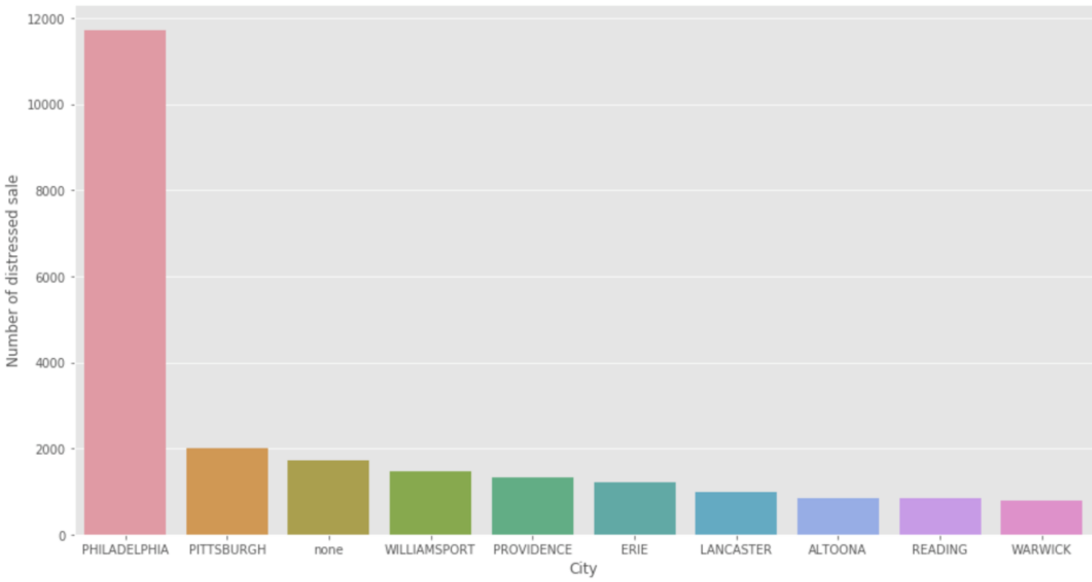

Figure 5: Number of Distressed Sales by City

4 Predictive Model

Our team was able to come up with linear and non-linear models that predict the house predict with 95 percent accuracy.

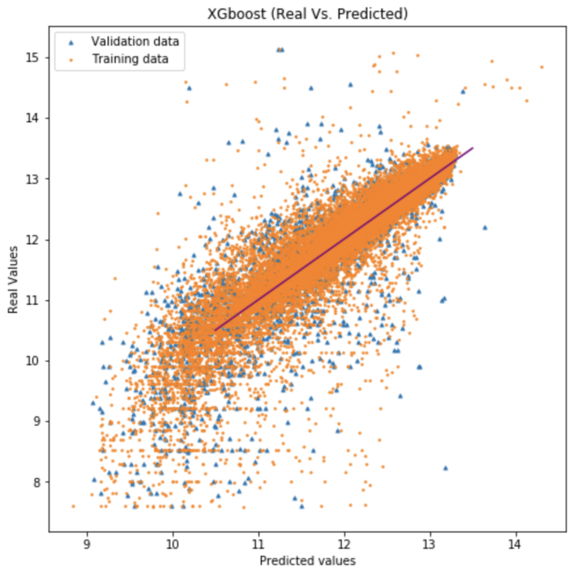

Figure 6: Non-linear Model)

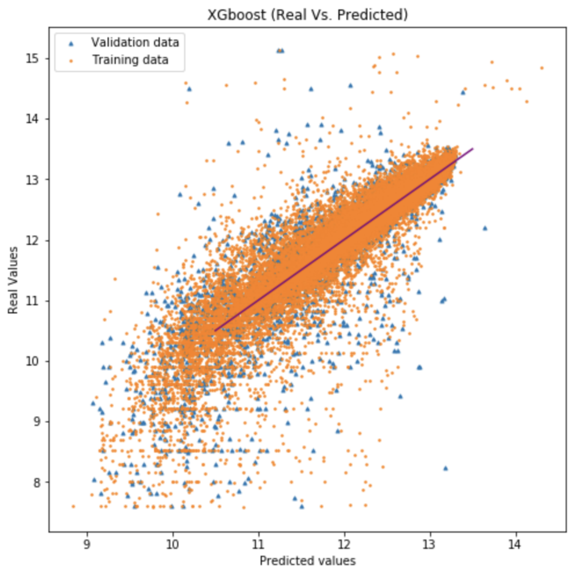

Figure 7: Linear Model)

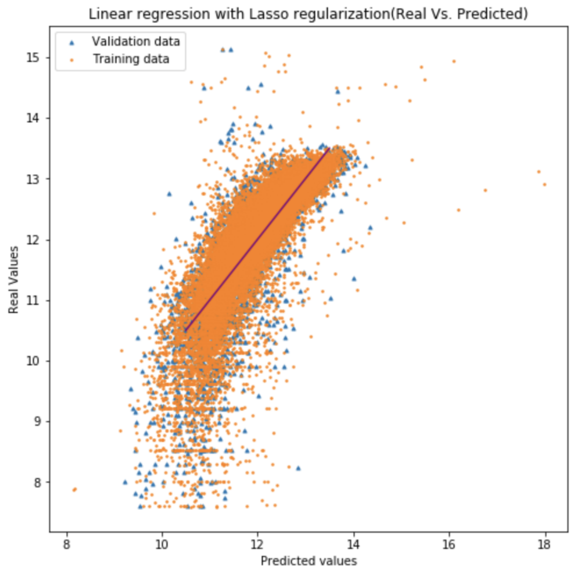

Figure 8: Linear Model