Predicting market value of single family homes in Philadelphia

Background

Home property ownership is an attractive long-term investment not only because it tends to be safer than longer-term stock investments but also for their brick-and-mortar appeal[1]. In recent years, the City of Philadelphia has witnessed increased net positive migration from New York City[2] . This migration has been attributed to job growth in the six years leading up to the early 2020 COVID pandemic[2] . Another possible reason for the move south from New York to Philadelphia could relate to increased remote work and relative affordability Philadelphia has to offer[2].

Here, I aimed to develop a model that could predict property values of single family homes, and to identify which features best predict market value. This project could be of interest to current homeowners, potential future buyers, and owners who pay property (real estate tax).

Ultimately, I ended up concentrating on homes that fall "in the middle" in terms of home value, excluding those that might be very expensive or inexpensive relative to others in the city. I used several features, such as home location, size, age, distance to schools, and interior and exterior condition, to predict market value as estimated by the City of Philadelphia Office of Property Assessment[3]. All data was retrieved through OpenDataPhilly, an online portal that houses public data related to the Philadelphia region[4]. Through this portal, I downloaded public records of individual properties in the city. This dataset exists because the Office of Property Assessment uses home location and features to estimate market value, which ultimately determines property tax[5]. These resources provided an opportunity to uncover which features of homes in the city drive up market value.

Approach

OpenDataPhilly and the City of Philadelphia Office of Property Assessment provide recent public records of individual properties in Philadelphia[3,4]. I downloaded these data and metadata with more information on variables in the form of CSV files in September 2022. I also downloaded information on school locations from OpenDataPhilly[6] and added the distance to the nearest schools as additional features in the dataset.

I wrote Python code in Jupyter notebooks, primarily using Jupyter Lab, to wrangle data, train and evaluate models. In data wrangling (wrangling.ipynb), I imported the csv files into a pandas dataframe, and dropped variables that were extraneous or redundant with other variables in the dataset. I also excluded variables where more than half of the records contained missing values. The full dataset contained all types of properties (vacant land, multiuse, etc). For purposes of this project, I filtered out records so as to only include finished single family homes. For variables with more than 10% but less than 50% of missing values, I substituted in the mode for all non-missing values. I dropped records where the number of rooms, bathrooms, or stories was equal to 0, assuming that that meant that key information was missing.

I chose to focus on single family homes whose market value was estimated to fall in the middle of the distribution of all single family homes, excluding those that fell at the extremes. To do so, I limited the dataset to records whose estimated market value was between +/- 1.5*iqr (interquartile range). For this regression problem (predicting market value), I knew I wanted to

train a linear regression model to serve a point of comparison against other tree based models. Because linear regression is sensitive to outliers, I opted to exclude records where values for other continuous variables fell out of the +/1.5 iqr for that predictor.

I was curious to see if distance to the nearest high school, middle school, or elementary school related to market value. To test for this relationship, I downloaded a csv file containing the locations (latitude and longitude) and types of schools in Philadelphia, and calculated the shortest distance between each single home property in the dataset and the closest high school, middle school, and elementary school using Manhattan distance (Philadelphia is mostly organized as a grid).

With the data cleaned, I explored the distributions of all variables, and the relationship between predictors and market value (eda.ipynb). For continuous predictors (house depth, house width, total area, total livable area, and year built), I plotted scatter plots showing the value for each record against its market value. For each discrete numeric predictors (number of bathrooms, bedrooms, number of stories, interior and exterior condition rating), I plotted violin plots, with quartiles overlaid, to visually inspect if there was a relationship between these predictors and market value. I repeated this step for categorical variables (heater, type of view, street designation, basement type, garage type). As a last step in data exploration, I peeked at which predictors independently best predict market value by calculating the predictive power score, a measure that uses tree-based algorithms to assess the correlation between a predictor and target variable.

Next, I trained and evaluated several regression models (model_development.ipynb). I first split the data into a training and test (proportion = 0.3) set. Then, I fit a linear regression model to start. For this model, I scaled the predictors, and made a pipeline to choose the K best predictors for the linear regression model. All models thereafter were tree-based models. From the sklearn model, I trained a DecisionTreeRegressor, GradientBoostingRegressor, and RandomForestRegressor. I also used the LGBMRegressor from the lightgbm (“light gradient boosting model”) module[7]. This model is similar to gradient boosting models but offers higher speeds by retaining data instances with large gradients (high information gain), and randomly dropping instances with smaller information gain[8]. I chose this class of models because I wished to explore which predictors (features) were most important in predicting home value, and tree-based algorithms offer relatively high interpretability.

Summary

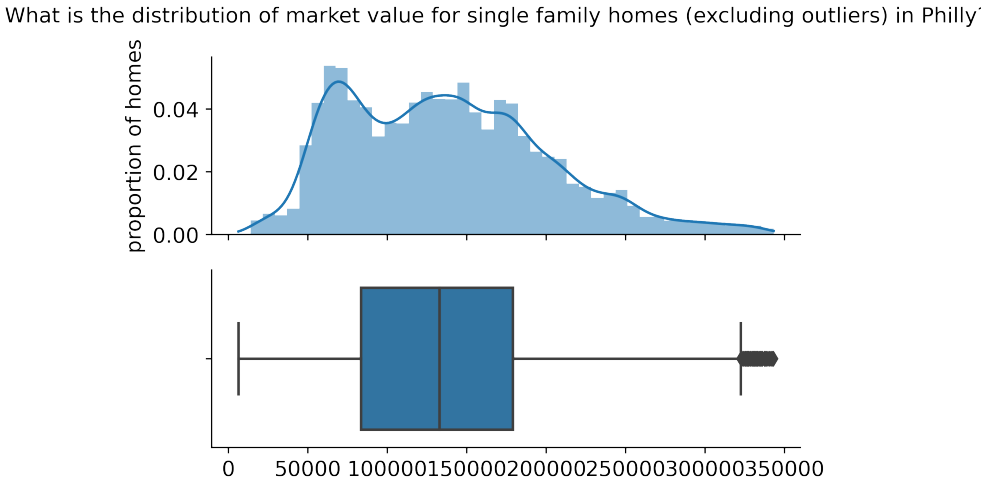

Figure 1. Distribution of market value for single property homes, excluding those that are very expensive or inexpensive.

After cleaning the data and filtering it to include only homes that fall “in the middle,” I explored the distribution and relationship of features in the dataset with the target variable, market value. The middle 50% of the filtered data set fell within a range less than $100,000. Variability in the upper 25% was greater than the lower range, spanning $164,100 (**Figure 1**).

Next, I considered the relationship between continuous, numeric variables, including the year the property was built and features related to home size

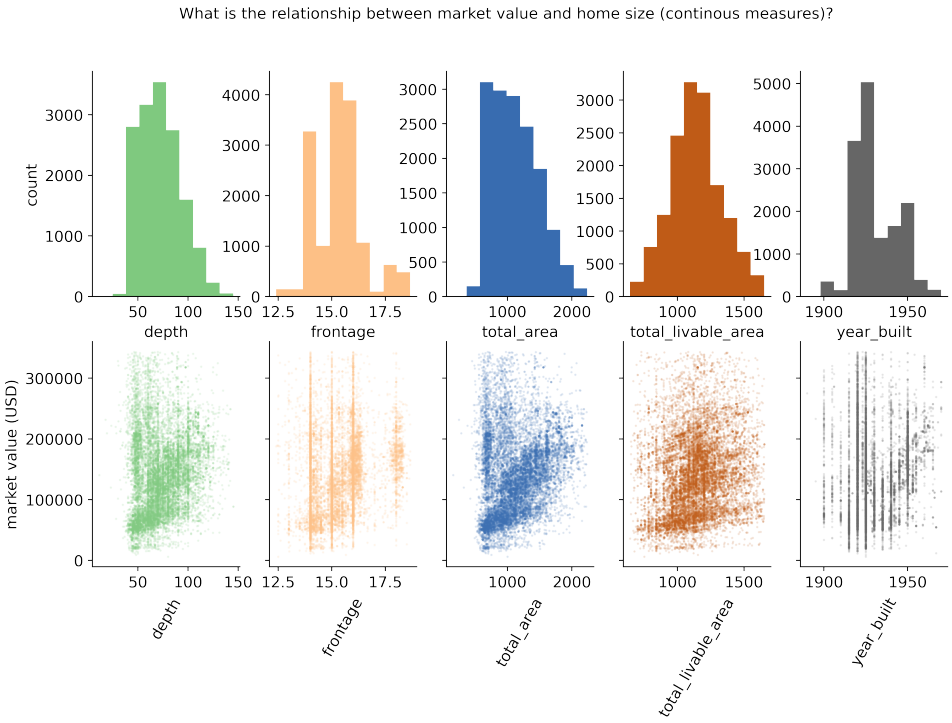

Figure 2. Distributions of (top) and relationship between depth (feet), frontage (width, feet), area (square feet), year built and market value.

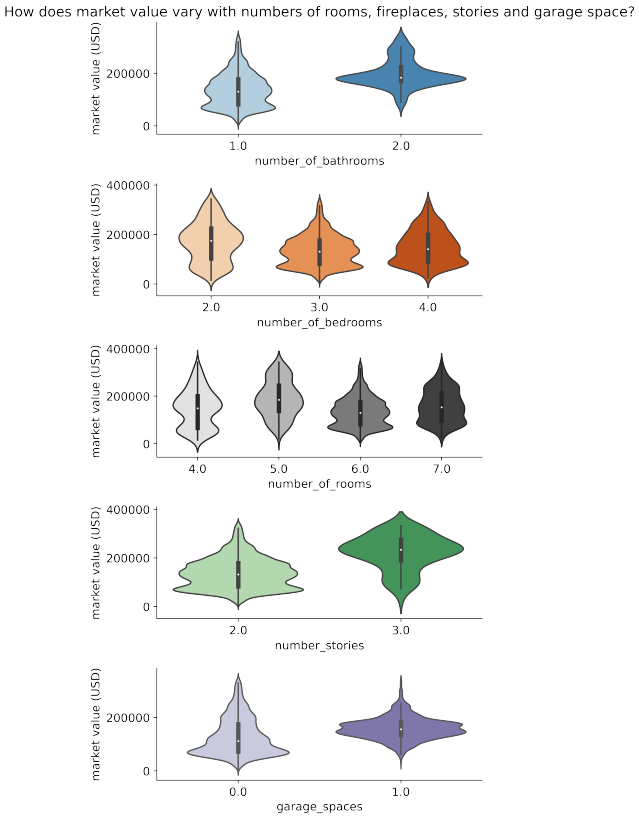

Figure 3. Relationship between market value and number of rooms, bedrooms, bathrooms, stories, and garage spaces.

(**Figure 2**). Visual inspection revealed a general positive relationship between these variables and market value. There were also several discrete numeric variables in the dataset, such as number of bedrooms, bathrooms, garage spaces, and stories. Visual inspection of market value as a function of these variables (**Figure 3**) revealed overall higher market values for properties with 3 stories (over 2), with 1 garage space (over 0), and with 2 bathrooms (over 1).

The Office of Property Assessment grades the interior condition and exterior condition of properties when making property assessments for tax estimates. In this filtered dataset, market value was higher for lower grades for both interior and exterior condition variables (**Figure 4**).

In addition to features related to the size and condition of the home, three variables related to location were included (zip code, latitude, and longitude). A map of mean market value as a function of zip code (**Figure 5**) suggests that market value varies along the map.

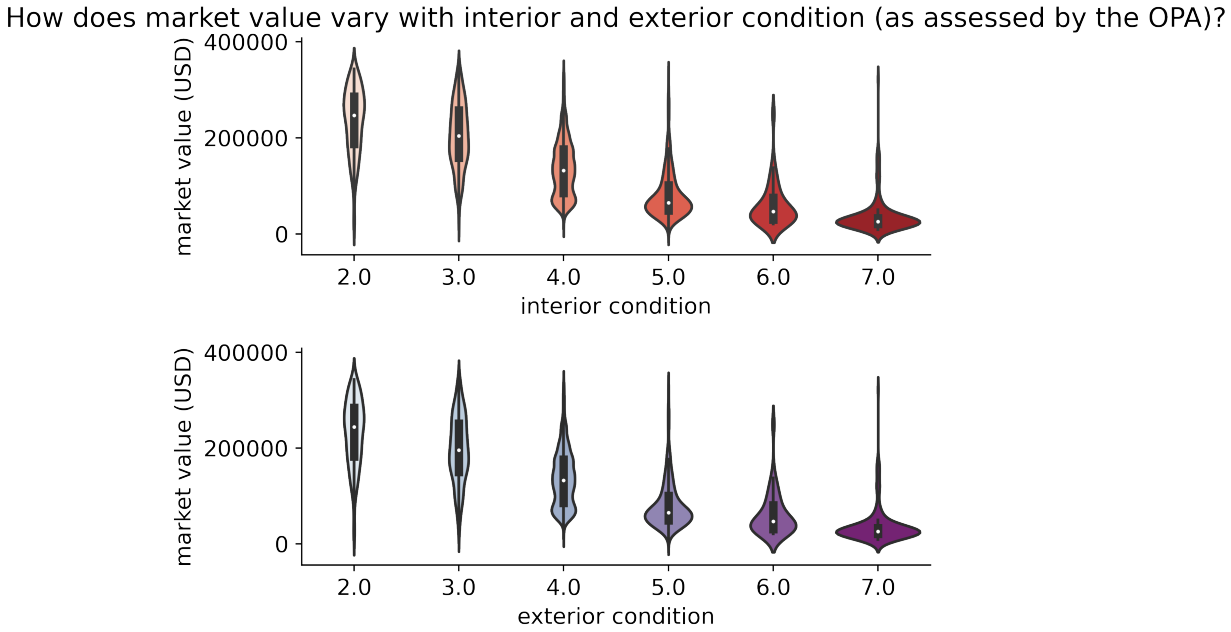

Figure 4. Relationship between market value and interior and exterior condition grades by the OPA

After exploring the features in the dataset, I selected 5 different models to train and evaluate: Linear Regression and four tree-based models (Decision Tree Regressor, Gradient Boosting Regressor, Random Forest Regressor, and Light Gradient Boosting Regressor). I trained a linear regression model to serve as a point of comparison against the more advanced tree-based models, and tree-based models because of the higher interpretability they offer (one aim of this project was to identify features that drive up market value). I split the cleaned data into training and test sets (proportion = 0.3), and ran 5-fold cross-validation to tune hyper-parameters. Then, for each model, I used the best estimator to predictor market value from features in the test set.

Because I was interested in predicting home value for 'the middle' or average property, I considered both the median absolute error and the mean square error (because the data were already filtered). The LGBMRegressor and RandomForestRegressor both came out on top, the latter with a lower median absolute error and the former with a lower mean squared error (**Figure 6**). The variability of scores across the 5-fold cross-validation were also highly similar for these two models (model_development.ipynb). The LightGBM model trained 10x faster than the RandomForestRegressor, and for that reason I chose to continue to evaluate this model.

To test how the model compared to a random model, I ran a permutation test. I shuffled the target (y) variable, keeping the training (X) data intact and trained the model as before using 5-fold cross-validation. I then computed the mean squared error of this trained shuffled model and repeated this 100 times. The mean squared error of the true model was lower than all 100 shuffled models (**Figure 7**), suggesting that the LGBMRegressor holds predictive power.

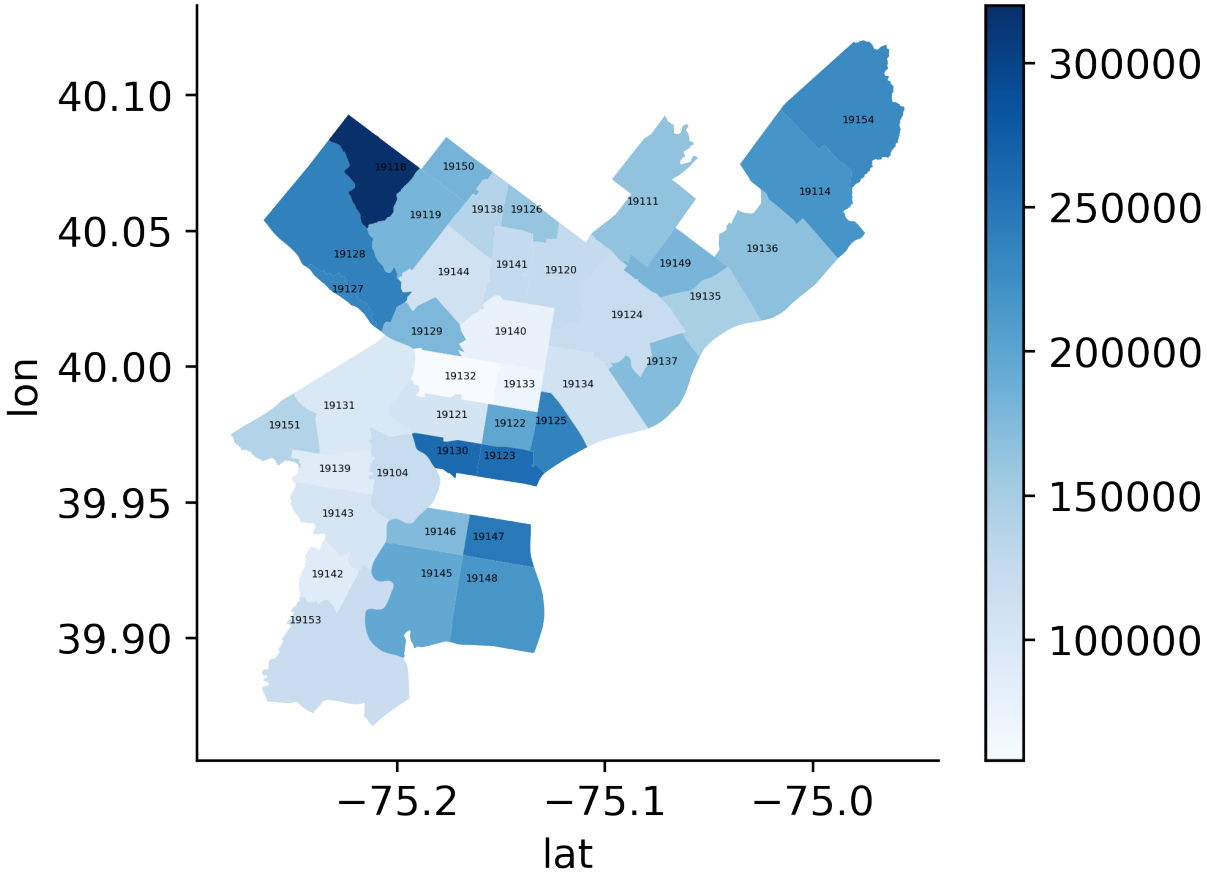

Figure 5. Mean market value (USD, color) as a function zip codes in Philadelphia. Map produced with data from an OpenDataPhilly partner [9].

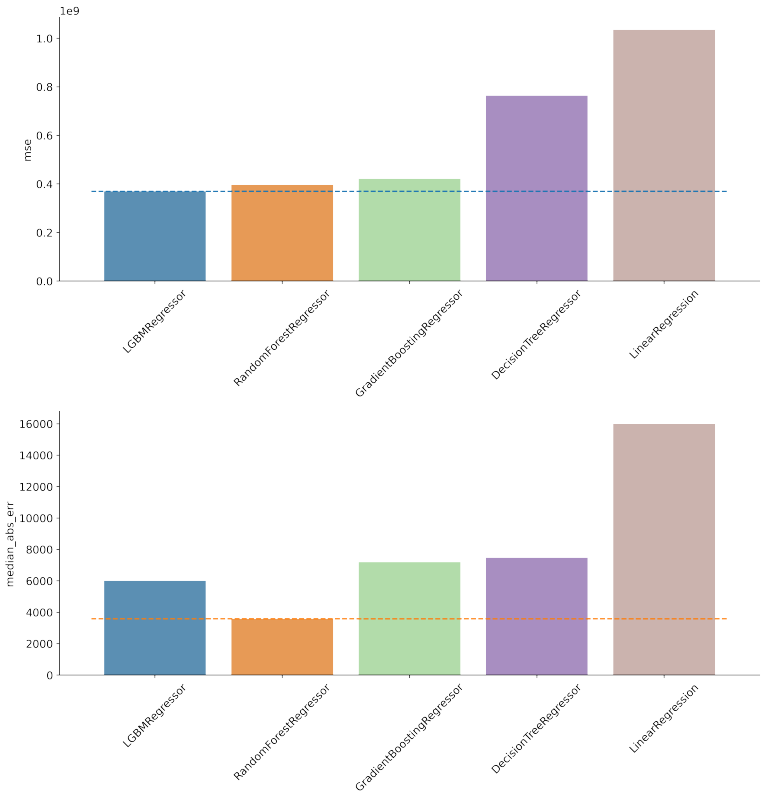

Figure 6. Mean squared error (MSE) and median absolute error (bottom) for trained models. Horizontal line shown in color of model with lowest error.

Next, I wished to identify the features that served as the best predictors for market value. Before investigating this, I trained the model on the full dataset. With this model, I plotted the features as a function of feature importance in this tree-based model (Figure 7). This revealed that the top 3 most important features were related to location (latitude, longitude, and zip code). The map in Figure 5 indicates higher mean market value towards the east relative to the west (latitude), with the exception of a region in the upper northwest of the county).

In addition to location, the total livable area and total area of the property were important features for predicting market value (**Figure 2**). Interior condition (**Figure 4**) and year built (**Figure 2**) also fell within the top 10 most important features.

Lastly, I considered properties where the model made the biggest errors in its predictions. I selected the top 1% of absolute errors (n=45), and considered the relationship between these properties actual market value with respect to others in their zip code (**Figure 7**). Within this group, there were more instances in which the model underestimated the market value.

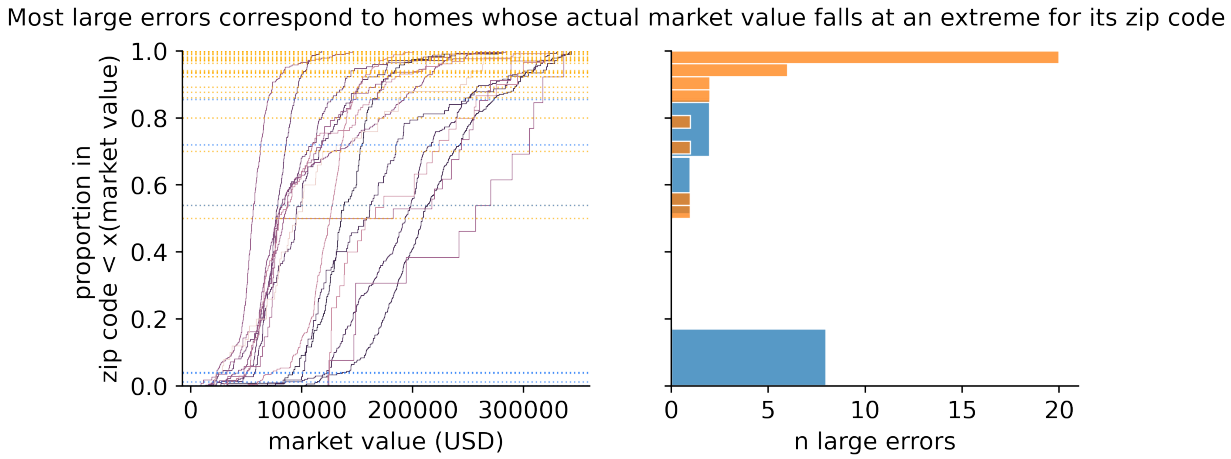

Figure 7. Investigating model errors. Percentile of large error instances based on market value distribution of market values for that zip code (left). Horizontal orange lines mark percentile for model underestimations; blue, model overestimations. Counts of zip-code based percentiles for errors (right) shows these properties tended towards the extremes in their respective zip codes.

Calculating the percentile at which actual market value fell relative to the market value distribution for that zip code revealed that these properties tended towards the extremes for their zip code, and that the model predicted values closer to the middle for those zip codes.

From these analyses, location is the best predictor of market value. For those interested in increasing the value of a home they already own, expanding the size of the home or making renovations that improve the interior condition of the property are also likely to increase market value.

References and Data Sources

1. Rao, Krishna. (2014, May 8). Is a Home a Good Investment? Zillow Research. https://www.zillow.com/research/returns-to-housing-6874/

2. Central Philadelphia Development Corporation (2020, September 17). More New Yorkers Moving to Philadelphia: Job Growth Plus Affordability Key Attractors. https://www.centercityphila.org/cpdc/cpdc-news/more-new-yorkers-moving-to-philadelphia-job-growth

3. Office of Property Assessment, City of Philadelphia. Property Records. https://www.phila.gov/property/data/#

4. OpenDataPhilly. https://www.opendataphilly.org/

5. Office of Property Assessment, City of Philadelphia. “What we do.” https://www.phila.gov/departments/office-of-property-assessment/

6. Philadelphia school locations data. https://www.opendataphilly.org/dataset/schools/resource/ 8e1bb3e6-7fb5-4018-95f8-63b3fc420557

7. LightGBM documentation. https://lightgbm.readthedocs.io/en/v3.3.2/

8. Geeksforgeeks (2021, Dec 22). LightGBM (Light Gradient Boosting Machine). https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/

9. OpenDataPhilly, geojson for Philadelphia zip code map. https://www.opendataphilly.org/dataset/zip-codes/resource/ 825cc9f5-92c2-4b7c-8b4e-6affa41396ee