Math 564 Project

PREDICTING HOUSING PRICE IN CHICAGO, ILLINOIS

Mathew Athoopallil: mathoopallil@hawk.iit.edu
Hailey Huong Nguyen: hnguyen18@hawk.iit.edu
Mikhail Rybalchenko: mrybalchenko@hawk.iit.edu
with equal contribution to the project.

November 28, 2018

Contents

Abstract

In this project, price of houses in Chicago, Illinois region are predicted. First, data pertaining to specific characteristics of the house is gathered. This is combined with data from other sources, such as, socioeconomic conditions of the regions, crime-rates in different localities and distance from nearest CTA 'L' station. Exploratory analysis is then performed to identify trends, correlation and derive other interesting insights from the data. Following this, machine learning models are built to accurately predict the price of a house, and the performance of the models are measured. Model selection is performed based on the performance metric to identify the best performing model and validate the accuracy of the predictions on new data. We conclude that tree-based XGBoost model predicts the house price most accurately. We also observe some important predictor variables, including square footage of the house, the number of baths in the house and the socio-economic condition of the communities.

Introduction

Predicting housing price has been a popular topic in the research industry. One of the popular datasets that is used frequently in teaching machine learning techniques is the Boston Housing Dataset. It is considered a standard set of data to try different concepts in statistical learning. This topic is also appearing in many machine learning competitions, as the ability of accurately predicting housing prices can help real estate agents to grow their businesses. Deriving insights about characteristics that influence the price of the property and accurately estimating the price of properties can be used by various organizations including but not limited to, the insurance industry, public development authorities, and even an average citizen.

Given the high demand as well as for the purpose of applying techniques in applied statistics course, we decided to work on the housing price prediction issue. Instead of using a readily available dataset online, we collected and created our own dataset for analysis. In this project, data which describes the various features of a properties was leveraged to make predictions on the pricing of a house in the Chicago area. External factors which influence the price of a house such as socioeconomic condition of the communities, crime rates, will be observed.

The data is first subjected to cleaning where missing values and erroneous data is imputed or dropped. Then variable transformations and variable selection was performed. Following this, the data from all the sources were combined. Exploratory analysis was then performed to gain insights from the data and also to determine whether feature engineering needed to be performed. This data is then split into two subsets for training the machine learning model and to test the performance of the model.

Finally, various machine learning techniques are applied, such as Linear methods, tree-based methods to accurately predict the model. The performance of the model was measured using the root mean squared error (RMSE) value of the testing data.

As in any machine learning problem, there is a trade-off between complexity and prediction accuracy, and that was also encountered in this project. The models that are easy to interpret in our case tends not to give us good predictions. Currently our models use a large number of predictor variables to obtain high prediction accuracy. However, this could present a particular challenge if we want to provide a standard model that can help predict prices in many areas nation-wide at an acceptable accuracy score. The reason is that the data which is available for the Chicago area may not be

available at a national level and hence could impact the performance of the model.

Through the implementation of this project we have worked on an entire cycle of a data science project right from the data gathering, data cleaning, exploratory data analysis to feature engineering, variable selection techniques and finally model building and making prediction.

This paper consists of many sections, including data preparation, exploratory data analysis, modeling, and conclusion, which explain in detail the steps performed at each stage of the project.

Data Preparation

1. Data Description

The dataset for this project is comprised of many sub datasets that were acquired from multiple sources. Details of these datasets are discussed below.

In order to achieve our goal, the general information about the houses for sale were collected, and this is referred as the main dataset:

- *Price* - listing price
- *Beds* - number of beds
- *Baths* - number of baths
- *Squarefeet* - square footage
- *Propertytype* - Condo/Single Family residential/Townhouse
- *Location* - Community belonging
- *Yearbuilt* - house year built
- *Latitude* - latitude
- *Longitude* - longitude

Crime data for 2018 was merged with the main dataset and the crime rate per 1000 of people in population per community (*crime_per_*1000) was calculated in order to, first, understand how crime rate affects the price of property, second, get an idea of the safest communities.

Furthermore, data on school ratings was added. In the dataset, the performance of each public school in the Chicago area is rated as Level 1, Level 2 or Level 3. Based on the definition of each level in the website, Level 1 indicates the highest performing schools, while Level 2 indicates a middle-performing school that needs improvement. Finally, Level 3 implies the lowest performing schools. As this dataset contains the rating of each school and that it would be merged with our main dataset from Redfin [1], the data was grouped by community and 2 variables below were computed to include in our model.

- *percent_level*1*_school* - Percent number of Level 1 schools in the area
- *percent_level*2*_school* - Percent number of Level 2 schools in the area

Additionally, Manhattan distance from each house to the closest CTA L station (*min_dist_cta*) and number of CTA 'L' stations in 1-mile radius (*num_cta*_1*mile*) were computed.

Finally, socioeconomics data per community area was joined with the main dataset. Even though one normally don't consider such information when choosing a house, such data can help build a profile of the community and explain variability of housing prices. The following variables were added to the dataset (per community level):

- $life_exp$_2010 - life expectancy, 2010 data
- *unemployment* - unemployment rate
- *perc_housing_crowded* - Percent occupied housing units with more than one person per room
- *perc_household_below_poverty* - Percent of households living below the federal poverty level
- *perc*_16$plus_unemployed$ - Percent of persons over the age of 16 years that are unemployed
- *perc*_25$plus_no_schools_diploma$ -Percent of persons over the age of 25 years without a high school education
- *perc_under*18*_over*64 - Percent of the population under 18 or over 64 years of age (i.e., dependency)

- *income_per_capite* - Community Area Per capita income is estimated as the sum of tract-level aggregate incomes divided by the total population
- *hardship_index* - Score that incorporates each of the six selected socioeconomic indicators

In our project, the following data sources were used:

- Homes for sale for Chicago, Illinois area [1];
- 2010 Census Data Summarized to Chicago Community Area [2];
- Census Data - Selected socioeconomic indicators in Chicago, 2008 – 2012 [3]
- Public Health Statistics- Life Expectancy By Community Area [4]
- Public Health Statistics- Selected public health indicators by Chicago community area [5]
- CTA - System Information - List of 'L' Stops [6]
- Chicago crime data for 2018 [7]
- Chicago Public Schools - Progress Report Cards (2011-2012) [8]

2. Data Cleaning and Standardization

The original housing data which was extracted from the Redfin website[1] and consisted of 27 variables and 9716 observations. There were missing values in 12 of the variables. The data was first filtered on the type of houses. In this project, we consider three types of houses, "Condo/Co-op", "Single Family Residential", "Townhouse". The non-relevant variables and variables having large number of missing values were removed. There were 15 variables in total.

Duplicated observations and outliers were removed. The missing values in square feet were imputed using the aggregated mean of the square feet over the combination of beds and baths. Observations were further filtered on the communities in the Chicago region.

The missing values in *year_build* variable were imputed using the mice package in R which uses Gibbs sampling technique and classification regression trees(CART) as the method. Considering *year_built* variable being an important one in our model, it

was converted to *HouseAge* to make this variable continuous. As the housing data includes all the available properties in the market at the time of access, we set the current year to 2018 and computed the age of each property.

The community number variable might be an important factor to explain our housing prices, but it has too many categories. Thus, we decided to group the extreme cases to reduce the dimension for the model. Our *NeighRich* variable has three categories 0,1, and 2. Category 2 indicates rich neighborhoods, and category 0 means poor neighborhoods. The rest was grouped into category 1. To select communities for category 2 or 0, mean and median housing prices for each community were computed and compared. The top n communities that have both mean and median to agree on the high price were selected for category 2, while the category 0 is comprised of communities that have low housing prices where both their mean and median are matching in the bottom n.

For the Crime data, the counts of the incidents per community level were obtained. and then crime rate per 1000 people of the community population was calculated.

Life expectancy, unemployment and other socioeconomic metrics are provided by community level. They are merged into the main dataset.

To calculate proximity to the CTA 'L' stations, and see the correlation between this variable and the housing price, Manhattan (Taxicab) distance [9] was obtained. Given that the City of Chicago is divided into squares and most of the streets are going North to South or vice versa, Manhattan distance is a reasonable approximation for calculating distances between two points. The curvature of the Earth is also taken into account[10]. For this purpose a function "manhattanDist" was written. Metric was calculated for each observation and added as a new variable *min_dist_cta*. Furthermore, the number of CTA 'L' stations in 1-mile radius using the same function "manhattanDist" was obtained and added as a new variable *num_cta_1mile*.

For the educational data, after removing NA records, the data was grouped by community with counts for level 1, level 2, and level 3 schools in each community. With a total public schools in each community, we computed the percent of level 1 schools, percent of level 2 schools, and percent of level 3 schools. Only percent of level 1 schools and percent of level 2 schools were attained in our model.

After merging all data, a full data examination process was conducted, and our response variable (*Price*) was observed not normally distributed. Under the assumption of OLS, this variable is assumed to follow a normal distribution. Therefore, a log transformation was applied to make the variable normally distributed [fig 1].

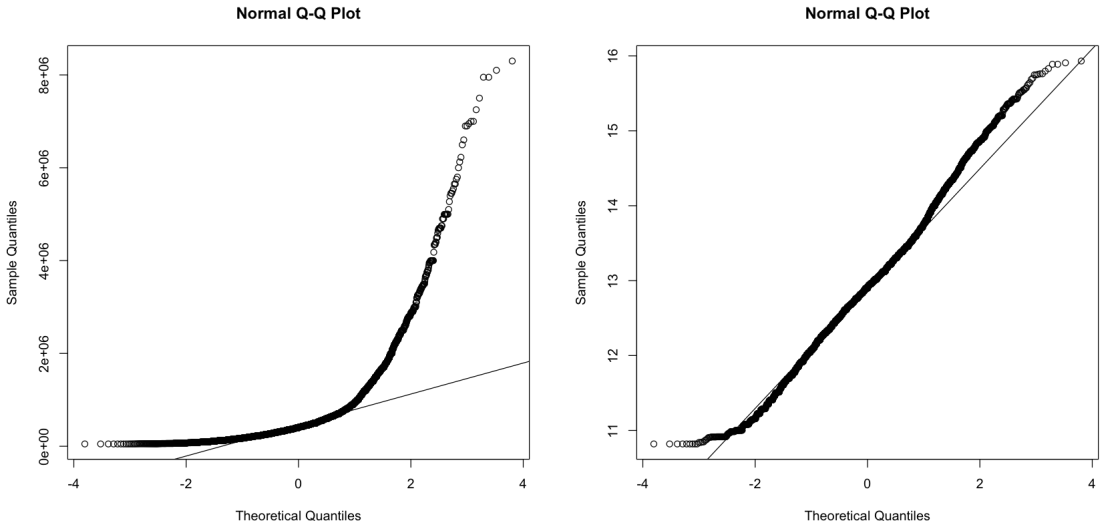

Figure 1: Q-Q Plot of PRICE variable before and after log transformation.

Exploratory Data Analysis

Price has a high positive correlation with the $square_feet$ (0.73). In addition, some of the explanatory variables were observed collinear. For example, $life\_exp\_2010$ and *income_per_capita* have a correlation of 0.81; unemployment and $perc\_16plus\_unempl$ have a correlation of 0.98 and high correlations with other variables. Analysis for multicollinearity should be performed [fig 2].

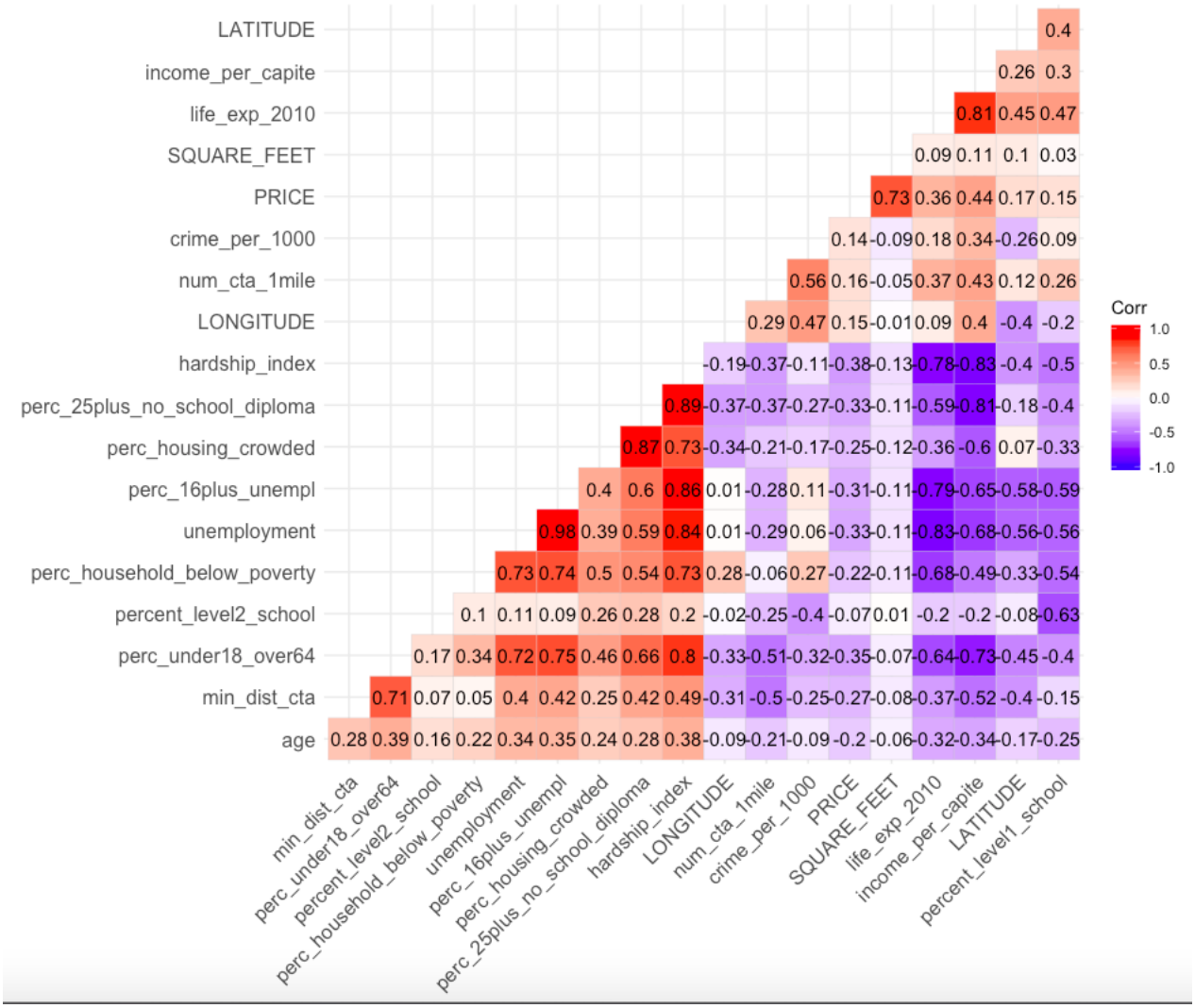

Figure 2: Correlation plot of numeric variables.

Housing price is right-skewed, which can be clearly seen from the histogram of price distribution. Given that price is positive and can not be equal to zero, log-transformation of price was used for normalization [fig 3].

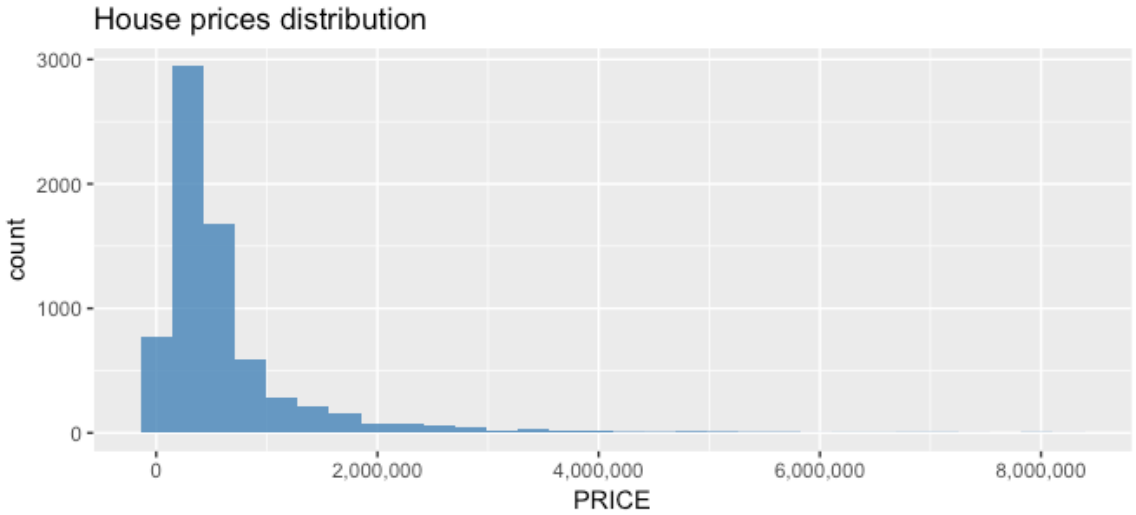

Figure 3: Histogram of the house prices.

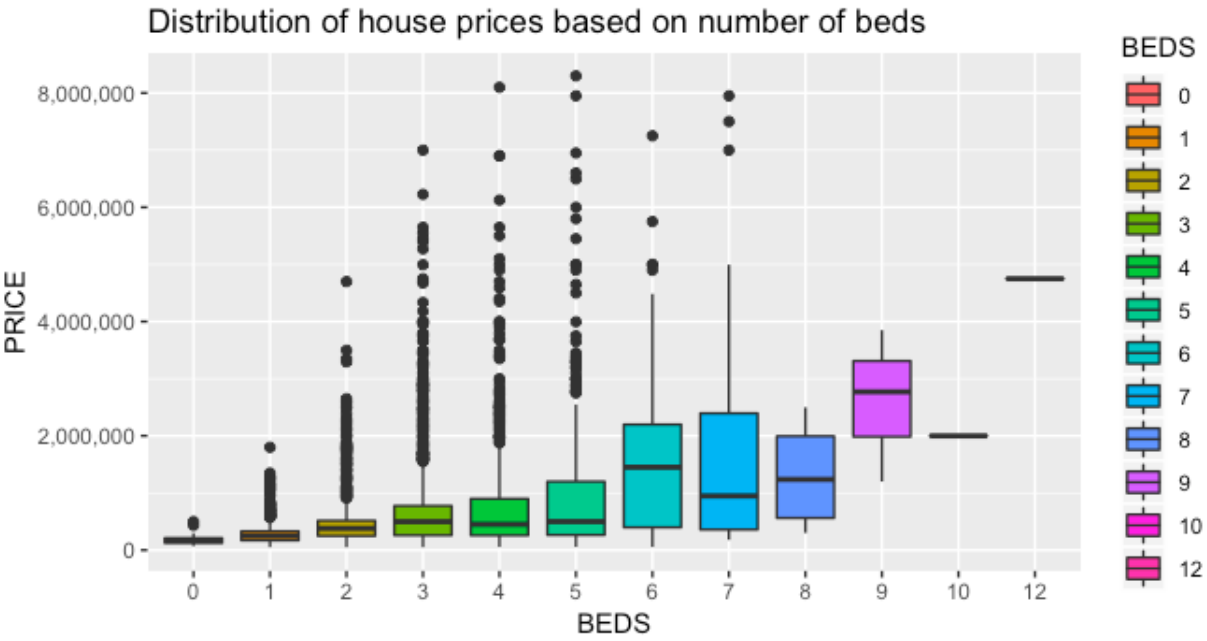

Figure 4: Distribution of house prices based on number of beds.

From the distribution of housing prices based on number of *beds*, one may conclude that with an increase in number of beds in the house, variation of price also rises [fig 4]. There is an upward trend between beds and price. However, there are a few observations for large number of beds with low variation. Therefore, number of beds in larger groups can be binned together, especially for number of beds exceeding

seven. Additionally, hypothesis tests were conducted to determine whether there was a significant difference between house mean prices and adjacent number of beds.

The variation in housing prices and number of baths are positively correlated [fig 5]. Similar to treating the *beds* variable, the *baths* variable was also binned into different groups. Hypothesis tests were conducted to determine whether there was a significant difference between house mean prices and adjacent number of baths.

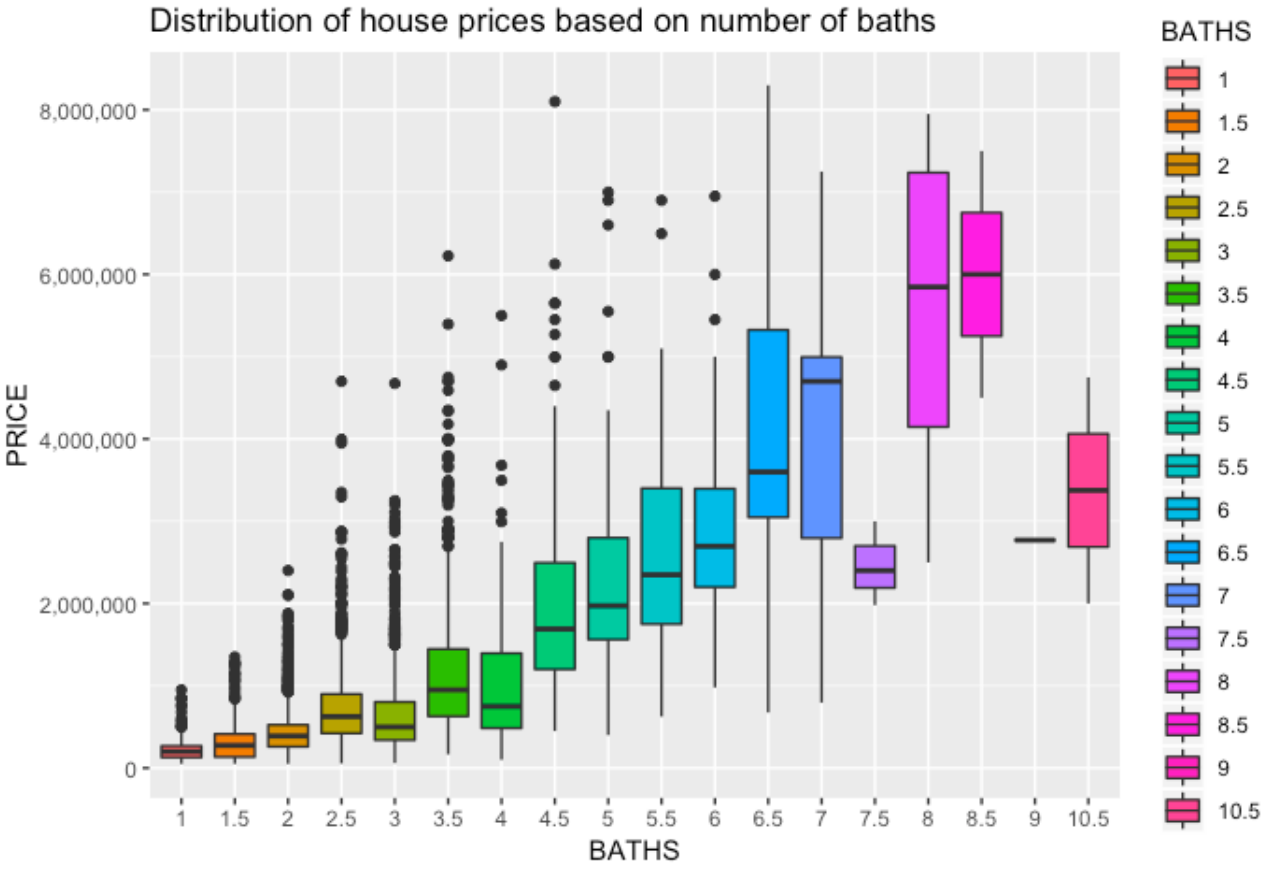

Figure 5: Distribution of house prices based on number of baths.

To further investigate the data, the interactive map was created [fig 6]. It provides a variety of summary metrics about the houses for sale per community. Insights obtained from map could be summarized in the following way:

- The most expensive community with the highest mean price of $1240K and highest median price of $800K is Lincoln Park.
- On average, the most expensive houses are sold to the North of Loop, while on the South of Chicago, average price would not exceed $200K for majority of the

communities.

- It can be seen on map that the most expensive communities have low median house age. We can conclude that either 'old' houses are not on sale or those communities are more recently built.
- The rate of level 1 schools is considerably larger for the northern communities, compared to the southern ones, though we can't say that communities with the higher house prices have higher rate of level 1 schools.
- As for the crime rate, Loop has the highest, mainly because its popularity with the tourists. This creates a favorable environment for pick-pocketing. In general, it can be seen that for the northern communities with the higher housing prices, crime rate is substantially lower.

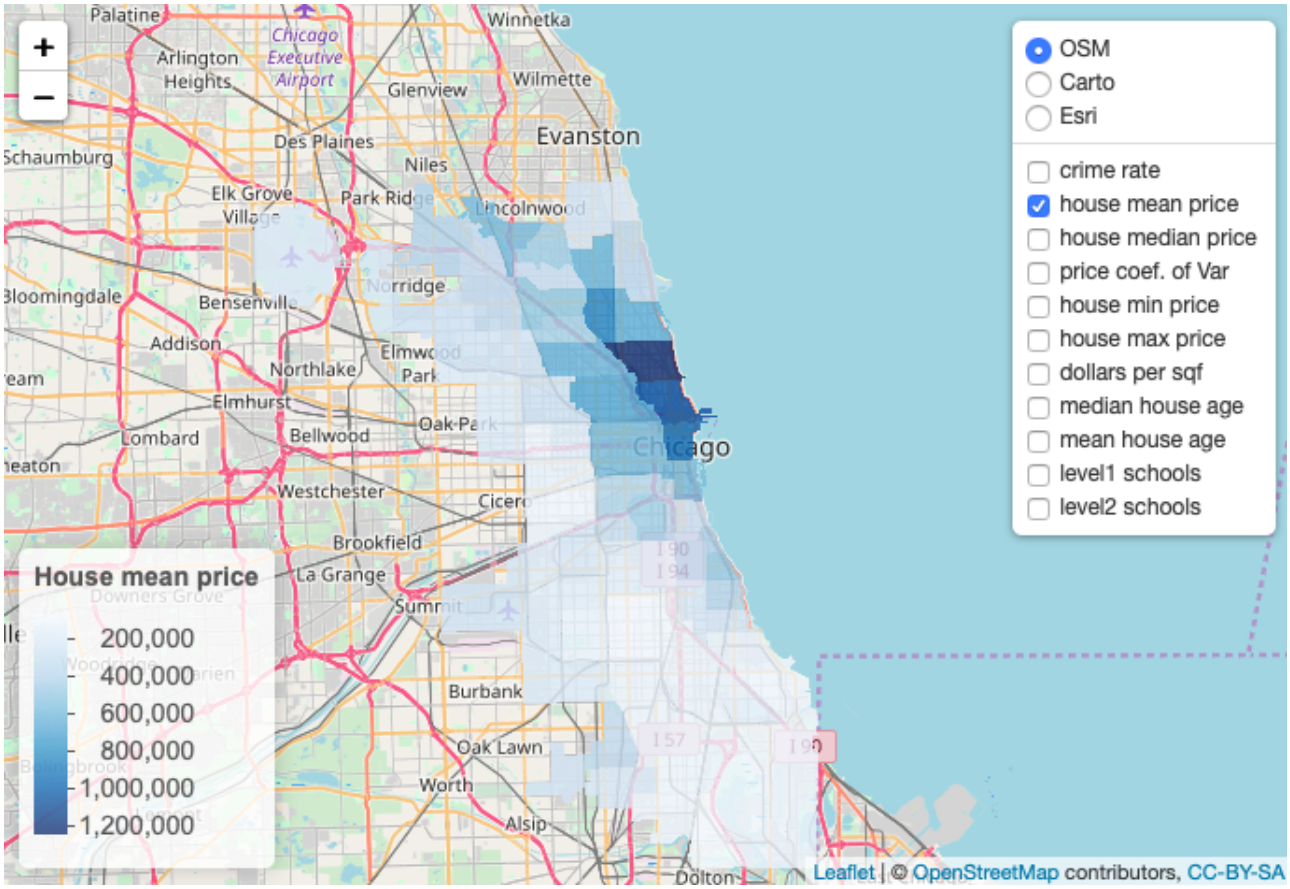

Figure 6: House's mean price by communities.

Interactive map provided in the project codes, please run "houseMap.html".

Modeling

1. Linear Methods

1.1 Methodology

The multiple linear regression (MLR) model assumed that the target variable is linearly dependant on multiple predictors with assumptions below.

- Linear relationship between target variable and predictors;
- Normal distribution of residuals;
- No multicollinearity. Predictor variables should not be highly correlated;
- Homoscedasticity. Variance around regression line should be the same for all predictors.

1.2 Results

Based on exploratory data analysis, log-transformed target variable was used for all models. The following results were obtained in the analysis.

Model	RMSE on log(Y)	Model description
OLS 1	0.3671265	Fit using all variables
OLS 2	0.3322985	Fit using all variables (22), re-binned Beds and Baths
OLS 3	0.3724802	Simpler model (13 variables) with multicollinearity removed.

Table 1: Results of Linear Models

For model 1, MLR with all independent variables was used to train the model. Variables *Beds* and *Baths* were treated as numerical variables. In this simple model, a substantial RMSE on log(Y) = 0.3671265 was obtained.

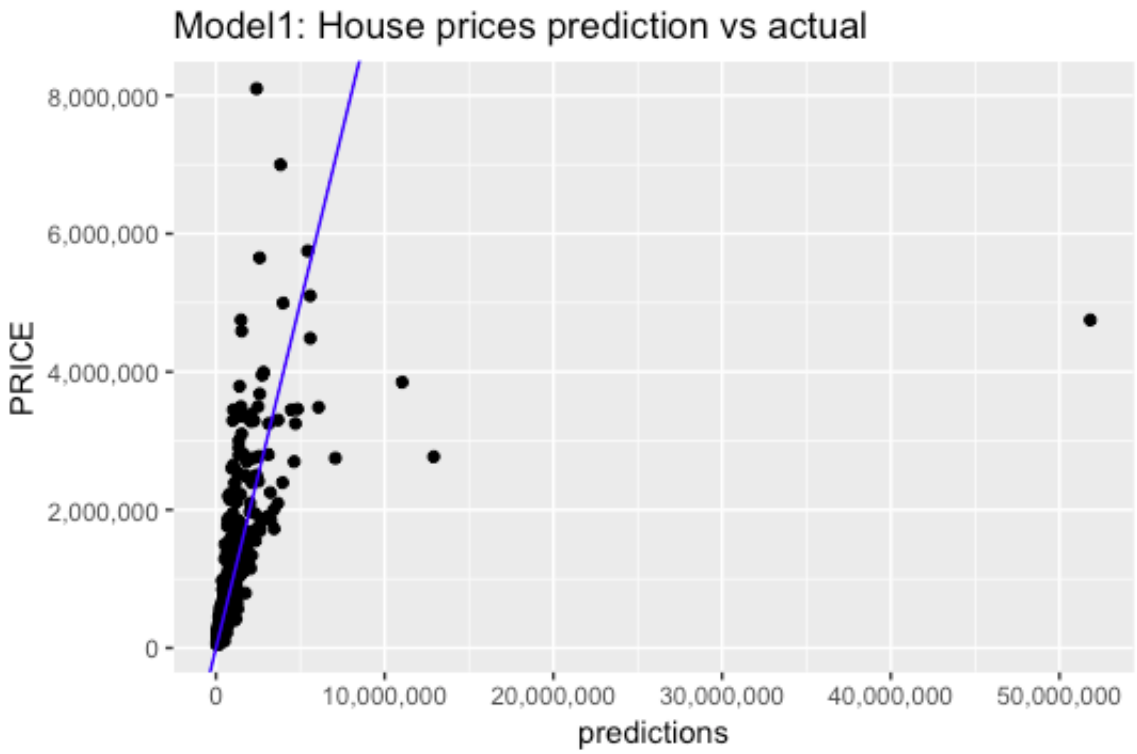

Figure 7: Model1. Predicted vs Actual.

From the plot [fig 7], *Price* is observed to be over-predicted in some special cases, and it was the case for observations with high number of beds and baths. To overcome that, the outlying number of beds and baths, were binned together into a separate category.

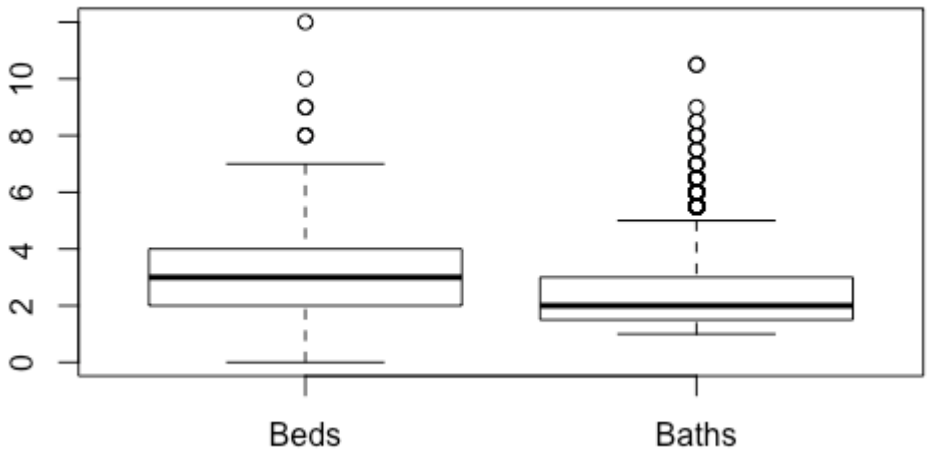

Figure 8: Boxplots for 'Baths' and 'Beds' variables.

Based on the boxplots [fig 8], *baths* levels 5 and higher were labeled as '5+' category, and *beds* level 7 and higher were labeled as '7+' category. With this changes, model 2 was fitted giving a lower RMSE on log(Y) = 0.3322985 with more accurate predictions [fig 9].

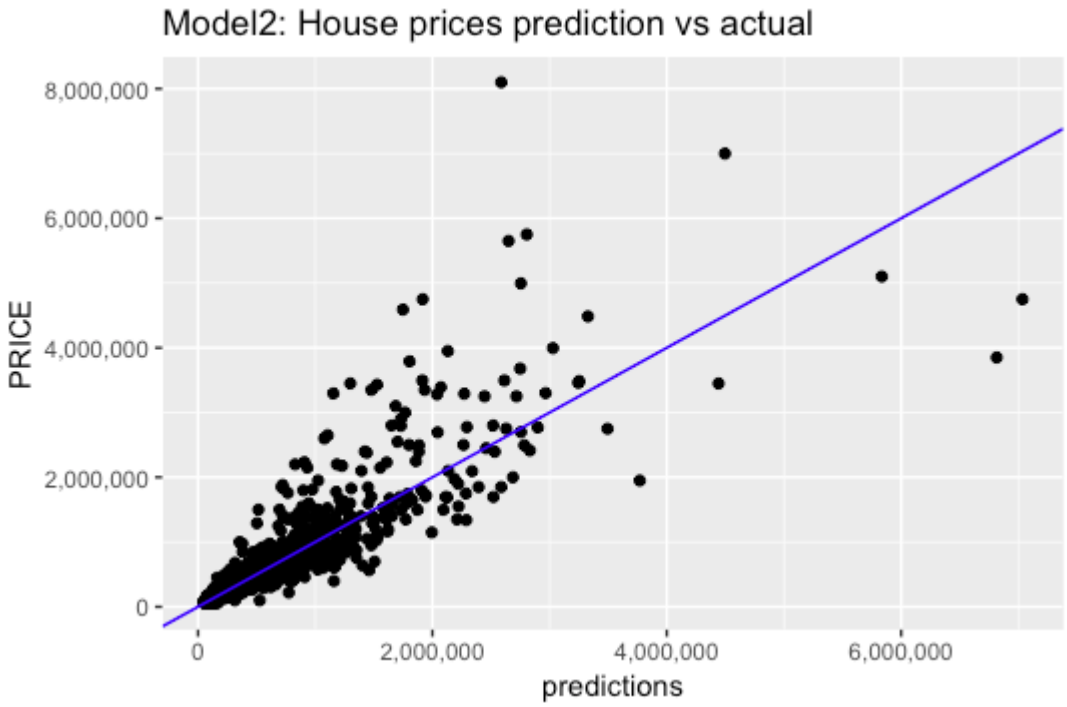

Figure 9: Model2. Predicted vs Actual.

Given the improvement in predicting power of the model and decreasing RMSE on log(Y), further relabeling of Beds and Baths was investigated.

Two-sample t-tests with alpha level .05 were performed comparing whether there is a difference in mean price for houses with 1.5 baths and 2 baths, 2 baths and 2.5 baths and so on. Same tests were performed for beds. As a result of those tests, some of the Beds and Baths levels were relabeled, as there were no difference in mean price between them [fig 10].

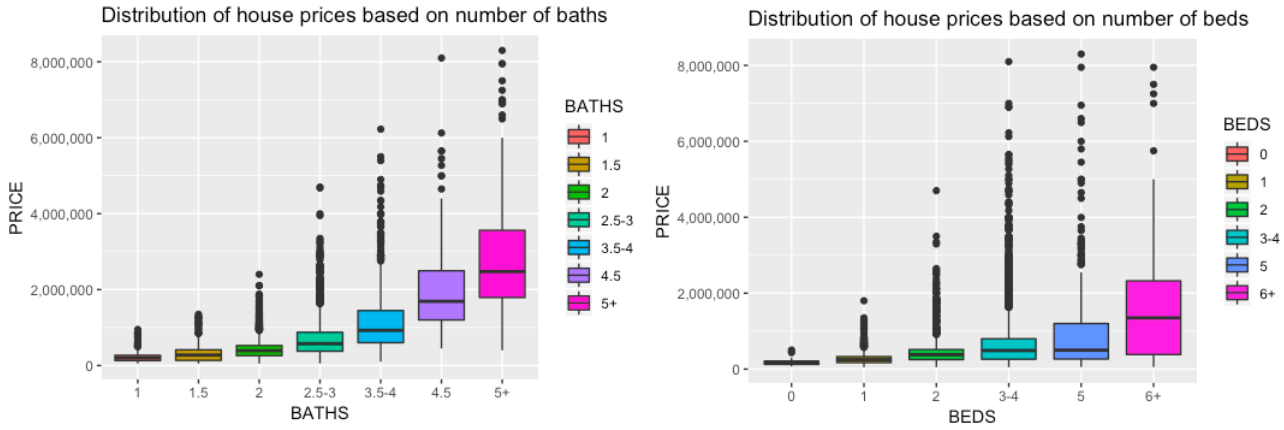

Figure 10: Side-by-side boxplots for Beds and Baths after relabeling.

Further relabeling of *beds* and *baths* had a tiny effect on RMSE on log(Y) with 0.0002 decrease.

In the next step, last fitted model, was checked for multicollinearity, by analyzing the Variance Inflation Factor (VIF), which quantifies the severity of multicollinearity. All variables with VIF larger than 9 were removed, though some of the levels of BEDS also had high VIF, but this variable was kept, as it is one of the main criteria when buying the house. Moreover, *hardship_index* variable was also kept, as it is a score incorporating main socioeconomic indicators and can help explain the variance in mean price.

Finally, the diagnostic plots for the final model were accessed:

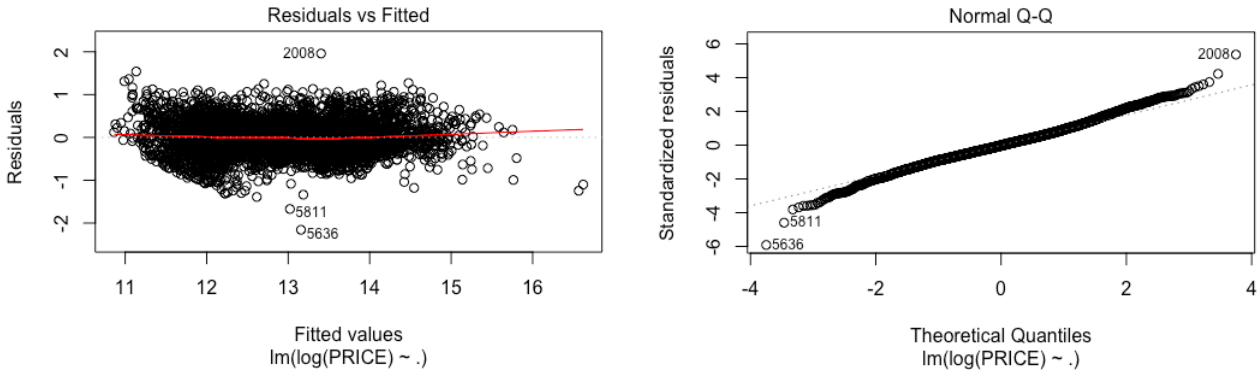

Figure 11: Diagnostic plots for final model.

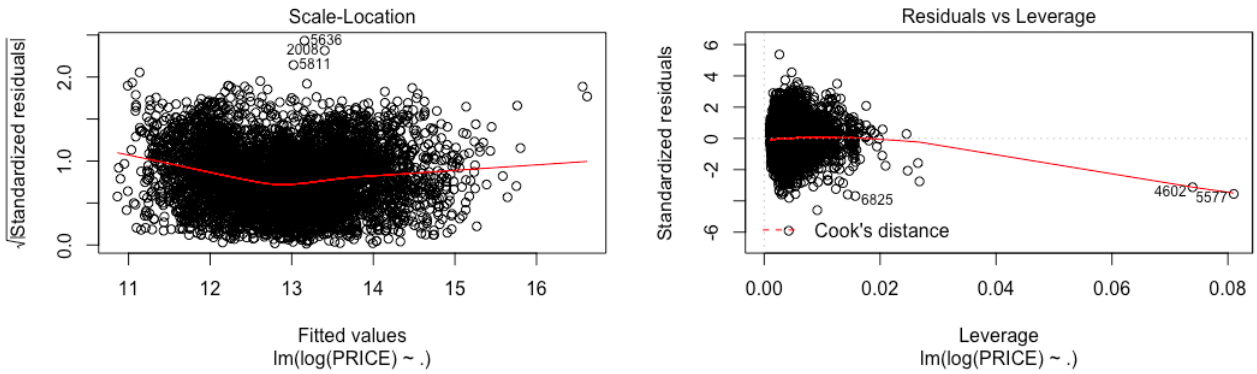

Figure 12: Diagnostic plots for final model.

We conclude that residuals are equally spread along horizontal line, so there is no non-linear patterns [fig 11]. Residuals are normally distributed based on QQ-plot [fig 11], which confirms one of the assumptions of linear regression. Homoscedasticity assumption was checked with 'Scale-Location' plot [fig 12]. The equal variance assumption was also true, as residuals are spread equally along the ranges of predictors. Finally, 'Residuals vs Leverage' plot doesn't show outliers [fig 12]. As a result, we have a good model.

2. Tree-based Methods

2.1 Methodology

Tree-based methods have been favorite techniques in many industries with proven successful cases for prediction. These methods are considered non-parametric, making no assumption on the distribution of data and the structure of the true model. They require less data cleaning and are not influenced by outliers and multicollinearity to some fair extend. The simplest method is decision trees, which can be used for both regression and classification problems and provide an useful and simple tool for interpretation. However, a simple model like Decision Tree tends to not have good predicting power. More complicated methods, such as random forests and boosting, usually yield better results, though there is a trade-off between interpretability and prediction accuracy. As the purpose of this project is accurately predicting housing prices, random forests and XGBoost algorithms were chosen.

Random forests involve building many decision trees on a bootstrapped training set. When building decision trees, a random selection of m predictors among the full

set of p predictors is chosen in each considered split. The split can only use one of those m selected predictors. This method of bagging is great to overcome the problem of choosing only a few strong candidates in the splits. It decorrelates the trees, making the results of averaging trees more robust.[12]

XGBoost stands for eXtreme Gradient Boosting, implementing the gradient boosting techniques but being optimized for speed and performance. The boosting idea involves growing trees sequentially, meaning that each tree is built based on the information from previously grown trees.[12] XGBoost, however, applies more penalties in the boosting equation when updating trees and residuals compared to the traditional boosting method, while leveraging the structure of hardware to speed up computing time.[11] Therefore, it has been considered the best model among all tree-based models.

2.2 Results

After running different models with tuning parameters, a model for random forest and 2 models for XGBoost are chosen to compare the predicting power with the test data set. *Beds* and *Baths* variables were kept as numeric in these models because they helped improve the performance of the models. The random forest model built 500 trees and 8 predictors were considered at each split. This model was fitted using all described variables in our training set (24 variables in total). All the XGBoost models were trained with cross validation of 5 folds, using the RMSE metric to evaluate the model. The first XGBoost model considered all available predictors, while the second model excluded some variables that were highly correlated to each other. The results of these models are displayed in the table 2 below, and figure 13 shows some important features when considered for a split in each type of model. The 'Gain' statistics evaluates the improvement in accuracy brought by a feature to the branches.

Model	RMSE on log(Y)	Model description
Random Forest	0.24364	Fit using all variables
XGBoost 1	0.23894	Fit using all variables
XGBoost 2	0.239424	Drop *perc_under*18_*over*64 + *hardship_index*

Table 2: Results of Tree-based Models

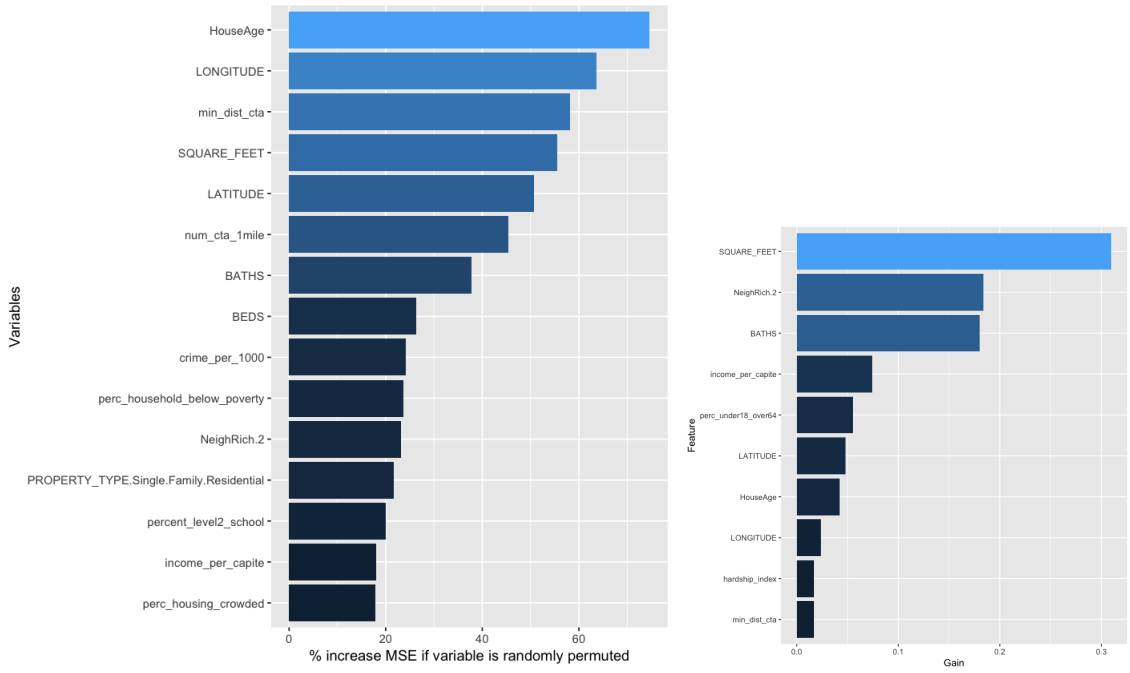

Figure 13: Important Features Based on Random Forests (left) and XGBoost1 (right)

Conclusions

The linear regression models achieve good performance in explaining our target variable. However, due to the multicollinearity of a few variables, model "OLS3" where the collinear terms are dropped is chosen in order to satisfy the linear regression assumptions. This model, in fact, has a slightly lower performance as compared to other linear models. The model that could accurately predict the price of the houses is XGBoost using all the predictor variables. The key insights from the analysis is that the factors that most influence the price of the house include the square footage of the house, the number of baths in the house and the socioeconomic condition of the communities. As for future improvement, principal component analysis (PCA) can be performed on the collinear terms, among the predictor variables, this can be then used in the linear regression model to boost its performance. Additional future work could involve transforming the entire process of the project from cleaning the data to building models and making predictions to a production level application where a user can select a region of interest and the output being an approximate price of the property.

Appendix

Outcome of the Random Forest model

```
Call:
 randomForest(formula = PRICE ~ ., data = trainData, ntree = 500,
importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 8

          Mean of squared residuals: 0.05893583
                    % Var explained: 92.38
```

References

[1] Real Estate, Homes for Sale, MLS Listings, Agents | Redfin. (n.d.). Retrieved from https://www.redfin.com/

[2] 2010 Census Data Summarized to Chicago Community Areas - Spreadsheet: 2010 Census Data Summarized to Chicago Community Areas - CMAP Data Hub. (n.d.). Retrieved from https://datahub.cmap.illinois.gov/dataset/2010-census-data-summarized-to-chicago-community-areas/resource/b30b47bf-bb0d-46b6-853b-47270fb7f626

[3] Census Data - Selected socioeconomic indicators in Chicago, 2008 – 2012 | City of Chicago | Data Portal. (n.d.). Retrieved from https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

[4] Public Health Statistics- Life Expectancy By Community Area | City of Chicago | Data Portal. (n.d.). Retrieved from https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Life-Expectancy-By-Commun/qjr3-bm53

[5] Public Health Statistics- Selected public health indicators by Chicago community area | City of Chicago | Data Portal. (n.d.). Retrieved from https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu

[6] CTA - System Information - List of 'L' Stops | City of Chicago | Data Portal. (n.d.). Retrieved from https://data.cityofchicago.org/Transportation/CTA-System-Information-List-of-L-Stops/8pix-ypme

[7] Crimes - 2018 | City of Chicago | Data Portal. (n.d.). Retrieved from https://data.cityofchicago.org/Public-Safety/Crimes-2018/3i3m-jwuy

[8] Chicago Public Schools - Progress Report Cards (2011-2012) | City of Chicago | Data Portal. (n.d.). Retrieved from https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t

[9] American Mathematical Society. (n.d.). Retrieved from http://www.ams.org/publicoutreach/feature-column/fcarc-taxi

[10] Distance on a sphere: The Haversine Formula | GeoNet. (2017, October 5). Retrieved from https://community.esri.com/groups/coordinate-reference-systems/blog/2017/10/05/haversine-formula

[11] Introduction to Boosted Trees — xgboost 0.81 documentation. (n.d.). Retrieved from https://xgboost.readthedocs.io/en/latest/tutorials/model.html

[12] James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). An introduction to statistical learning with applications in R. New York: Springer.