**Predicting Zillow Home Value Estimates Using XGBoost Regression**

Ben Karabinus

Department of Computer Science, University of Denver

COMP 4442: Probability and Statistics 2

Dr. Wendy Christensen

November 14, 2021

**Abstract**

The United States (U.S.) real estate market can be described as a complex blend of diverse geographies and regional price parities constantly in flux as home values rise and fall with the ebb and flow of the economy. In recent years extreme gradient boosting, a popular method for analysis used in the field of machine learning, has been a top choice for predicting changes in real estate markets across the globe. This document summarizes an analysis of home values in regions across the U.S. as related to socioeconomic factors in said region. The goal of the analysis was to design a gradient boosted machine learning model for use in predicting future home values, specifically an XGBoost regression model. Results show that XGBoost regression is robust to complex data collected on the economic state of different regions in the U.S. and can be used to predict the value of homes in said region with a relative degree of accuracy.

# **Predicting Zillow Home Value Estimates Using XGBoost Regression**

XGBoost regression is a popular supervised machine learning algorithm which falls into the category of tree-based methods. Tree-based methods seek to create predictions for both continuous numerical outcomes and group membership (categorical predictions). Models are trained using past observations of an outcome of interest (dependent variable) and a set of parameters (independent variable(s)) related to the outcome. Once a model has been trained new records can be fed into the model and used to create predictions.

Real estate data are a prime candidate for analysis and prediction using tree-based methods. The analysis explained in this document covers how to use XGBoost regression to predict Zillow Home Value Index (ZHVI) estimates for home price. Key areas of discussion include data sources and variable definitions, exploratory data analysis, a brief explanation of the primary statistical method used, a summary of method application and the outcome of the analysis.

**Data Sources and Variable Definitions**

**Data sources**

Data for the analysis were collected from a combination of both government and private sources. The outcome of interest, the ZHVI home value estimate (Zestimate), was collected from Zillow.com. The ZHVI is a smoothed, seasonally adjusted measure of typical home values in the U.S. (Zillow Home Value Index, 2021). ZHVI estimates account for market changes and differences in housing type across regions. Regions correspond to typical measures used by U.S. Government agencies when collecting data for statistical research (See Appendix A). The single category for region type used in this research was the metropolitan statistical area (MSA). All ZHVI Zestimates used were collected for records from year 2019. For a brief definition of how the ZHVI Zestimate was used in this analysis please reference supporting content (see Appendix A). The variable “Size” is a factor with six levels condo, one, two, three, four and five. Each level represents the size of a home with five representing homes with five or more bedrooms. The size variable was created by recoding the “ZHVI Indicator Type” variable included in the ZHVI data.

Population (POP) and Per Capita Personal Income (Income) were collected from the Bureau of Economic Analysis (BEA) website bea.gov. Both variables were collected for each region in the data set and only for year 2019. Per capita personal income is calculated dividing the total personal income in a region by its population.

Regional price parity (RPP) measures the differences in price levels for consumer goods across MSA’s for a given year and is expressed as a percentage of the overall national price level. For example, if consumer goods in Denver cost 10% more than the U.S. average then the regional price parity for Denver would be 110%. All regional price parities used in the data set were collected for year 2019.

Unemployment rates used as predictors in the were collected from the U.S. Census Bureau. Rates were collected for each region in the data set and were only collected for year 2019.

# **Exploratory Data Analysis**

## **Summary statistics**

Prior to implementing XGBoost a brief exploratory data analysis was performed. Exploratory analysis showed that the combined data set is complex with high dimensionality of factor levels in categorical variables (Region = 368 levels) and the presence of extreme outliers in both continuous predictors and the dependent variable Value. The mean of POP (population) was 739,774 with a standard deviation of 1,677,914 (See Appendix B, Figure B.1). Though not as extreme as Value, each of the continuous predictors produced a left skewed distribution, again showing the presence of extreme outliers (See Appendix B, Figure B.2). Box plots were created to show the spread of Value for each home size in the data (condo, one, two, three, four, five). The box plots showed the distribution of the value of homes for each size was left skewed indicating the presence of extreme outliers (See Appendix B, Figure B.4).

After performing exploratory data analysis normalization of the data seemed to be warranted. That being said, no data preparation steps were performed. A key aspect of the analysis was to assess the robustness of the algorithm to complex data sets.

# **Application of XGBoost Regression**

## **Gradient Descent and Gradient Boosted Machines**

Gradient descent is a powerful algorithm used to optimize mathematical models by efficiently minimizing a specified cost function. The cost function can be any differentiable expression making gradient descent not only powerful, but extremely versatile. Gradient boosted machines (GBMs) are a class of supervised machine learning models which builds upon the principles of gradient descent. GBMs are a sequential tree method. Sequential tree methods prescribe building a series of models (weak learners) which improve upon the errors of their predecessors (See Appendix B, Figure B.5).

## **XGBoost**

XGBoost for regression is a popular flavor of gradient boosted machine learning models. Like other boosted methods, XGBoost prescribes building a sequence of models each slightly improving upon its predecessor. What’s makes XGBoost different is that the algorithm allows for a great deal of flexibility when specifying parameters for model tuning. Creating an XGBoost regression model begins with making an initial prediction. Residual error is then calculated in the same fashion as regular least squares regression by taking the difference of the observed and predicted values. The residuals are used to create a root node or “stump” for the first tree in the sequential tree model. Next a similarity score is calculated for the root node. The similarity for a tree node is the sum of the squared residuals in a node divided by the number of residuals plus a regularization parameter “lambda” (See Appendix B, Figure B.6). Lambda is used to reduce similarity scores sensitivity to individual observations helping to avoid overfitting (Rao, 2021). After calculating the similarity score for a root node, the leaves of a tree are created by grouping the residuals based upon model parameters or “predictors.” Following the split, similarity scores are calculated for each new leaf. The information gain for any split is the sum of similarity scores for each leaf minus the similarity score of the parent or root node (See Appendix B, Figure B.7). When creating a split each parameter in the model is tested independent of the others. The split resulting in the greatest gain is chosen. This process is repeated until a specified benchmark is achieved. XGBoost regression allows for the option to specify a plethora of hyperparameters for model tuning during tree creation. Hyperparameter tuning makes XGBoost regression robust to overfitting and gives the algorithm ability to perform well on large, complicated data sets.

# **Applying XGBoost Regression to ZHVI Data**

## **Tidymodels and XGBoost in R**

XGBoost regression is designed for use with large, complicated data sets and is built to be robust to over-fitting. The robustness of the algorithm comes both from both process design and the ability to specify different hyperparameters for model tuning. Even with its sturdy design and ability to handle all types of data, implementing XGBoost regression using the R programming language requires some specific steps. Mainly those related to data preparation and model tuning. XGBoost prescribes that all categorical variables are one hot encoded prior to being fed into the model. One hot encoding is the process of transforming categorical values into a machine-readable coding schema. To summarize this process categories can be thought of as columns and membership in each category is specified by a one for true or a zero for false in each column (DelSole, 2018). Enter the Tidymodels library. Tidymodels is a framework for implementing modeling and machine learning processes in the R programming language (Tidymodels, 2021). Creation of an XGBoost model for predicting Zillow Home Value Estimates (Zestimates) was done using the Tidymodels framework. The first step in creating an XGBoost model using Tidymodels is to split the data into training and test sets and the then specify a recipe for data preparation (See Appendix B, Figure B.8). The recipe maps out the necessary steps for preparing data to be used with XGBoost and creates a repeatable process for preparing data. When training XGBoost regression models it’s common place to use cross fold validation during model training to avoid overfitting. The Tidymodels package is designed so the recipe can be applied when creating folds. That is, data processing and the creation of folds for cross-validation is done in one step. After preparing the data the Tidymodels “Parsnip” package is used to specify the XGBoost model and specific parameters for tuning. A grid space for testing model performance with different parameter combinations is created to train the model. For this analysis a grid space of size one hundred was created with five folds for cross-fold validation, giving five hundred unique parameter combinations to be tested during the training process. The XGBoost model is then tuned, and the best parameters are selected. (See Appendix B, Figure B.10). Root mean square error was used as the gauge for model performance in this analysis. A visual assessment of the training process shows a steady decline in root mean square error until around the 1500th tree built during training (See Appendix B, Figure B.11). The final tree count used during the training process was 1598. Regional price parity (RPP) was determined as the predictor which contributed the most gain to the final model (See Appendix B, Figure B.12).

## **Results**

After determining the best parameters, the final fit was created using the training data. Predictions were made on both the training and test sets to check for the presence of overfitting and assess the performance of the model on new data. Results showed a root mean square error of approximately 30,434.72 on the training set and 48,939.28 on the test set (See Appendix B, Figure B.13). Given the mean of Value was 228,186 with a standard deviation of 161,634 the model performed considerably well. The results show that XGBoost regression is a suitable method for predicting home prices in regions across the U.S. using socioeconomic factors in the region and that the algorithm is robust to large, complicated data sets.

**References**

DelSole, M (2018, April 24). *What is One Hot Encoding and How to Do It.* Medium. <https://medium.com/@michaeldelsole/what-is-one-hot-encoding-and-how-to-do-it-f0ae272f1179>

*Gradient Boosting Machines.* UC Business Analytics R Programming Guide. Retrieved October 28, 2021, from <https://uc-r.github.io/gbm_regression#learn>

Rao, S. (2021, August 22). *XGBoost Regression: Explain it to me like I’m 10.* Towards Data Science. <https://towardsdatascience.com/xgboost-regression-explain-it-to-me-like-im-10-2cf324b0bbdb>

*Tidymodels.* Homepage. Retrieved November 12, 2021, from <https://www.tidymodels.org/>

*Zillow Home Value Index (ZHVI).* Housing Data. Retrieved November 3, 2021, from <https://www.zillow.com/research/data/>

# **Appendix A**

Data Dictionary

Income – The per Capita personal income for a given region in the data set.

POP – The population for a given region in the data set.

Region – Region corresponds to metropolitan statistical areas in the U.S. MSA’s are defined as a standardized county or equivalent-based area having at least one urbanized area of 50,000 or more population, plus adjacent territory that has a high degree of social and economic integration with the core, as measured by commuting ties.

RPP - RPPs measure the differences in price levels across states and metropolitan areas for a given year and are expressed as a percentage of the overall national price level.

Size – A factor with 6 levels condo, one, two, three, four and five. Each level represents the size of a home with five representing homes with 5 or more bedrooms.

UnempRate – The unemployment rate for a given metropolitan statistical area. Unemployment rates were collected for fiscal year 2019 from the U.S. Census Bureau.

Value – The typical home value for a given region. This value is calculated using Zillow’s proprietary “Zestimate.” Value is a seasonally adjusted and smoothed measure. For supplemental information on how the Zestimate is calculated please see [ZHVI User Guide](https://www.zillow.com/research/zhvi-user-guide/).

**Appendix B**

Supporting Statistics and Visualizations

Figure B.1. Summary Statistics

Text

Description automatically generated

Figure B.2. Distribution of Continuous Variables

Chart, line chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart

Description automatically generated

Figure B.3 Scatter Plot of Continuous Predictors

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Figure B.4 Boxplot of Value by Home Size

Chart, box and whisker chart

Description automatically generated with medium confidence

Figure B.5 Gradient Boosted Machines

Diagram

Description automatically generated

Figure B.6 XGBoost Regression, Root Node

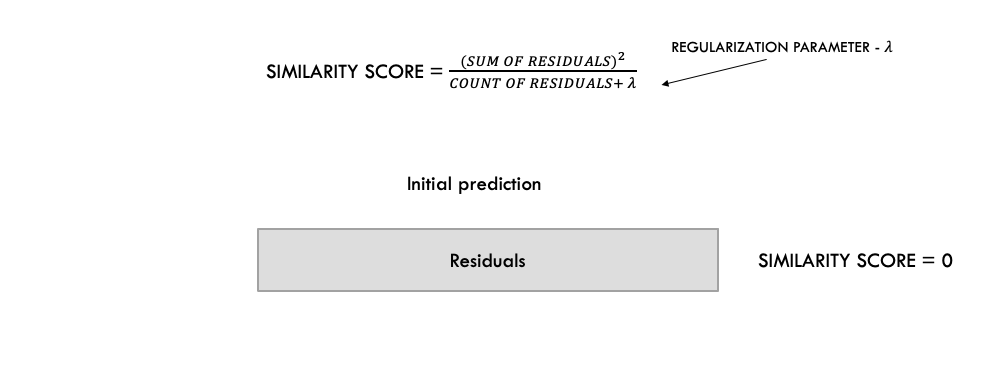


Figure B.7 XGBoost Regression, Split

Diagram

Description automatically generated

Figure B.8 Specifying a Recipe for XGBoost

Text

Description automatically generated

Figure B.9 TIdymodels, Model Specification

Text

Description automatically generated

Figure B.10 Parameter Selection

Graphical user interface, text, application

Description automatically generated

Figure B.11 Root Mean Square for Model Training

Chart, line chart

Description automatically generated

Figure B.12 Variable Importance

Chart, funnel chart

Description automatically generated

Figure B.13 Results

Graphical user interface, text

Description automatically generated