**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 economic 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 predict the value of homes in given region with an impressive 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 a new set of parameters can be fed into the model to predict an unknown outcome of interest.

Real estate data are a prime candidate for analysis 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**

**References**

*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>