Capstone III Project Proposal: Zestimate Prediction

Scott McCracken, 2023

The Problem

In 2017, Zillow released “Zillow Prize,” a competition challenging participants to beat Zillow’s *Zestimate*, a predictor of home future sale prices. The competition data is still open to the public although it finished some time ago.

Homes are typically the most expensive purchases any individual will make, and at the time of the competition, Zillow was actively predicting the sale price of over 100 million homes in the United States. Zillow started this competition in hopes of gaining insights to improve its Zestimate algorithm and reduce the log error of actual sale price compared to the Zestimate. An accurate algorithm would help it compete with other real estate marketplaces like Redfin.

The Data

The competition uses data from homes in LA, Orange and Ventura counties in 2016 and 2017. The data is organized in pairs of tables, a training set for the 2016 year and a testing set for the 2017 year. Each set has one table with over 50 features of homes, such as square footage, the number of rooms, the existence of a pool, etc, and a separate table that contains all homes that sold with the transaction and the log error of its price from the Zestimate. The tables can be joined via a Parcel ID.

The goal of the competition is to predict the log error of each home, which will help reveal which types of homes the Zestimate is relatively good or bad at predicting.

There are nearly three million records in the training data set, and many of the features have null values. There are also several categorical features, including some that may appear as numerical at first, such as land use codes. The data is accompanied by a dictionary that helps explain the column names.

The Approach

This project may have a long Exploratory Data Analysis process, as it will be important to see if any features are correlated with log error and if there are unexpected patterns in any of the features. It may be helpful to categorize the log error values into tiers of ‘good’ versus ‘bad’ predictions and plot those tiers against different features.

The data includes latitude and longitude coordinates, and it might make sense to use a Tableau dashboard to visualize log error or certain features geographically.

This is a regression problem, so a Gradient Boosting or Support Vector Machines model may be effective. I will need to decide how to deal with missing values and scaling before training these models.

The Deliverables

The competition calls for log error predictions for six different time points: Oct, Nov, and Dec in 2016 and 2017. Models are judged on their mean absolute error between their prediction and the actual log error.

It will also be important to deliver visualizations that describe which types of properties lead to errors for the Zestimate. This may be done in a slide deck or a dashboard.