Executive Summary: Zillow Assessed Tax Property Value Prediction

Project Objective:

Our team was tasked to build a data-driven tool that estimates the assessed value of residential properties using the provided dataset. The goal of this project was to improve Zillow's pricing accuracy, support data-driven decision making, and help provide more reliable home value estimates to customers.

Data Analysis:

These data included features such as square footage, number of bedrooms and bathrooms, lot size, location, pool and garage features and more. The data needed significant pre-processing due to many empty fields (some with over 95% incomplete values), removal of internal identifiers lacking predictive value, and handling of extreme outliers. After analyzing the shape and content of the data, we were left with a reliable and well-structured dataset ready for supervised modeling. Exploratory data analysis revealed that larger homes are indicative of higher property value, bathroom count is a strong price signal, and the target property values were highly right skewed. Specifically, a relatively small number of higher-end properties were pulling the average upward significantly.

Model Performance:

We tested three model approaches; a more straightforward Lasso Regression (a linear method), a rule-based Decision Tree Regression, and Gradient Boosting, a model that builds on its own errors to improve predictions. The best performing model was Gradient Boosting, which predicted property values with an average error of ~$196,000 on property data it had not seen before. Despite this large value, this error represents strong performance given the context that the property values in this dataset span a wide distribution. The linear model was consistently underperforming, despite attempting to tune parameters as we went, and the Decision Tree model seemed to be memorizing training data and did not perform well on unseen data.

Recommendation:

We recommend using a Gradient Boosting model as the foundation for an automated valuation tool. It has proven to reliably capture complex, non-linear relationships between home features and price, and strikes the best balance between accuracy and consistency, which may not be fully captured by simpler models. The next priority should be enriching the input data with neighborhood-specific features, such as proximity to schools, walkability, and comparable sales, which will strengthen the predictive value beyond the current available features.