Zillow Prize: Zillow's Home Value Prediction

Group Members:

1. Varshini Bonagiri
2. Muvva Sreeja
3. Nikhitha Tubati
4. Nandigam Sarath Chandra Pavan Ram

Description:

This project is to enhance the Zestimate home valuation algorithm. The goal was to build predictive models that accurately estimate home prices based on historical real estate data.

Dataset has a full list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) data in 2016. Training data has all the transactions before October 15, 2016, plus some of the transactions after October 15, 2016. The test data in the public leaderboard has the rest of the transactions between October 15 and December 31, 2016. The rest of the test data, which is used for calculating the private leaderboard, is all the properties in October 15, 2017, to December 15, 2017.

We must predict the log-error between their Zestimate and the actual sale price, given all the features of a home. The log error is defined as

$$\textbf{logerror} = \log(\textbf{Zestimate}) - \log(\textbf{SalePrice})$$

Forecast home values at six time points for all properties:

- October 2016 (201610), November 2016 (201611), December 2016 (201612)
- October 2017 (201710), November 2017 (201711), and December 2017 (201712)

We are planning to handle missing values using median imputation for numerical data and mode imputation for categorical variables. Feature engineering will involve creating attributes like property age and price per square foot, while categorical data will be encoded using label encoding. We will apply MinMax scaling for numerical

standardization and implement outlier detection techniques to identify and correct erroneous transaction values, ensuring high-quality data for model training.

We plan to research and analyze various models before finalizing the most accurate one. Our approach involves experimenting with multiple algorithms and selecting the best-performing model. Based on our research so far, we will begin with baseline models like Linear Regression and Decision Trees, then explore advanced techniques such as Gradient Boosting (XGBoost, LightGBM, CatBoost) for improved accuracy. Additionally, we may consider neural networks for deep learning exploration. To handle the time series component, we will analyze trends in sales price changes. Model performance will be evaluated using Root Mean Squared Log Error (RMSLE), with cross-validation strategies based on time-based splits for robust evaluation.

Responsibilities:

Member 1:

1. Handle missing values using appropriate techniques.
2. Experiment with advanced models like XGBoost, LightGBM, and CatBoost.
3. Implement time-based cross-validation to ensure robust evaluation.
4. Create visualizations to support data insights and model performance.

Member 2:

1. Identify and remove or adjust outliers in the dataset.
2. Research and implement models (Linear Regression, Decision Trees).
3. Calculate performance metrics such as RMSLE for model comparison.
4. Summarize model results and prepare a structured report.

Member 3:

1. Normalize and scale numerical features for consistency.
2. Optimize hyperparameters and fine-tune model performance.
3. Analyze trends and patterns in sales price changes over time.

4. Develop presentation materials for final project submission.

Member 4:

1. Engineer new features such as property age and price per square foot.
2. Train models to predict property values at six specific time points.
3. Assess the impact of time-dependent features on predictions.
4. Maintain detailed documentation of all steps, including challenges and findings.