Samantha Feinberg | Alina Hota | Abhiroop Kumar
Connor Therrien | Janani Vakkanti

PREDICTION CONTEST (GROUP 2)

Objective

Predict Austin house prices on the 'austinhouses_holdout.csv' using the 'austinhouses.csv' using the *'latestPrice'* variable. The analysis compared different modeling approaches on expanded predictors.

Data Preparation and Feature Engineering

1. X/Y Graphs: Correlation analysis of the Austin housing dataset revealed how different features relate to house price. The correlation matrix provided a numerical measure of these relationships, highlighting linear trends, outliers, and the strength of the connections. Together, these tools identified key features for predicting house prices.
2. Data preparation: This process involved excluding unique identifiers and using a selection of predictors, including *streetAddress*, *homeType*, and *avgSchoolSize*. Missing values were imputed by omitting the *"NA"* values for categorical data. Additionally, some features were transformed into factors to suit specific modeling techniques.
3. Feature Engineering: Custom columns were created from existing features to improve the model performance, to give more accurate price predictions and a lower RMSE. *age_when_sold* to capture property lifecycle, *area_ratio* to measure house density on its lot, *total_rooms* as a metric for size, *school_quality_score* to represent the combination of school quality and proximity, and *log_latestPrice* as the log of the latestPrice variable.

Models

A total of 9 models were run on the target variable, *log_latestPrice* - Linear Regression (0.26) was a baseline model, while Ridge (0.31) and Lasso (0.31) used regularization to prevent overfitting. Stepwise, achieved a lower RMSE of 0.31. Other methods like Bagging (0.25), Random Forest (0.28), and Boosting (0.25) performed better. But the *Unpruned Regression Tree* model performed the best, outperforming even its pruned counterpart (0.24), achieving the lowest RMSE at *0.18*.

Final Prediction & Conclusions

Based on the lowest RMSE of 0.18, the unpruned regression tree was selected as the final model. This model's predictions on the holdout dataset demonstrated strong performance, suggesting that the predicted prices exhibit similar skewness to the original prices, indicating robust performance.

latestPrice	Original Price	Predicted Price
Min.	5.8	92.59
1st Qu.	310	314.25
Median	400	399.28
Mean	486.6	479.53
3rd Qu.	550	554.43
Max.	6250	3558.04