Atlanta Housing Price Prediction

Chamonia Johnson cjohnson435@student.gsu.edu

Tyler Thomas tthomas149@student.gsu.edu

**PROBLEM STATEMENT:**

Our goal is to create a predictive price modeling tool where users can input listing, price, amenities, and variables to our Machine Learning Model to predict the output price of a home in the Atlanta area.

**DATA SOURCES:**

<https://www.kaggle.com/datasets/yellowj4acket/real-estate-georgia/data>

**DATA PREPARATION:**

Data Loading: imported our data into Google Colab using the pd.read function. We also imported pandas, numpy, sklearn as well.

Data Cleaning: dropped Real Esate listings that weren’t in Atlanta. Then we dropped all of the variables that weren’t going to be used in our analysis. The final step in our data cleaning process is to replace 0 values with NAN values and to drop those values.

Data Preprocess: defined our numerical values (price, pricePerSquareFoot, livingArea, bathrooms, bedrooms) and categorical values (city, state, streetAddress, homeType, county, levels). We used pipeline and one hot encoder to transform the numerical and categorical values.

Data Visualization: we scaled our numeric features and visualized Lot Area against Selling Price.

Modeling- Linear Regression: Used the train test split to have training and testing data. Imported Linear Regression package and trained the model and evaluated the training and testing score. We used R squared and MSE as evaluating metrics.

Modeling- KNN Regression: Used the train test split to have training and testing data. Imported KNN Regression package and trained the model and evaluated the training and testing score. R squared and MSE were used as evaluating metrics

Modeling- Random Forest: Used the train test split to have training and testing data. We imported the Random Forest package and trained the model and evaluated the training and testing score. R squared and MSE were the evaluating metrics

For all of these models, the inputs were country, city, state, yearBuilt, streetAddress, zipcode, livingArea, bedrooms, bathrooms, levels, homeType, and county.

**EVALUATION:**

Using Mean Squared Error and R squared to determine the best model, we created a bar chart to display the results. The lower the MSE, the better the model seems to be; while the higher the R Squared, the better the model seems to be performing. Linear regression had the highest R squared and the lowest Mean Squared Error compared to KNN and Random Forest.

**RECOMMENDATIONS:**

Based our evaluation, we chose to use the Linear Regression Model when predicting the Price of homes in the Atlanta area.