Model Reporting for listing_mortgage_info Dataset

Dataset Overview

The listing_mortgage_info dataset contains 98,502 entries with the following six columns:

- zpid: Unique identifier for each entry (non-null, integer).
- bucketType: Category type for the listing (non-null, object).
- rate: Mortgage rate (may contain missing values, float).
- rateSource: Source of the rate information (non-null, object).
- lastUpdatedTimestamp: Timestamp of the last update (non-null, object).
- lastUpdatedDate: Date of the last update (non-null, object).

Target Variable

The target variable for our analyses is rate, which represents the mortgage rate. This column contains 4,835 missing values out of 98,502 total entries.

Models and Analytical Approaches

We analyzed the dataset using four machine learning models:

- Linear Regression: A fundamental regression model to identify linear relationships between predictors and the target variable.
- Decision Tree: A non-linear model that splits data into decision nodes for prediction.
- Random Forest: An ensemble model based on decision trees, improving prediction robustness and accuracy.
- K-Nearest Neighbors (KNN): A distance-based algorithm that predicts the target value by averaging the values of the nearest neighbors.

Each model was evaluated on both:

- ✓ Unscaled data: Original data without any standardization or normalization.
- ✓ Scaled data: Data transformed to ensure features have zero mean and unit variance.

Comparative Analysis: Scaled vs. Unscaled Data

The models were compared on two key performance metrics:

- Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values (lower is better).
- R^2 Score: Represents the proportion of variance in the target variable explained by the model (higher is better).

Performance improvements after scaling were particularly significant for the KNN model, as it is sensitive to feature magnitudes.

The analysis highlights the importance of data preprocessing and scaling for models like KNN. Scaling not only improved performance metrics but also enhanced the interpretability and stability of predictions across all models.

Further details and visualizations of the model results are provided in the subsequent sections.