Housing Prediction

Ames Housing Price Prediction Model

Developing a Predictive Model for Real Estate Valuation

Ames Housing Price Prediction Pipeline

Overview of data processing stages for modeling

Data Cleaning

The process of identifying and correcting errors in the dataset to enhance data quality and integrity, ensuring reliable analysis.

Feature Engineering

Creating new features or modifying existing ones to improve model performance, often through transformations or combinations of raw data.

Model Development

The phase where predictive models are built and tested using algorithms, incorporating validated features to forecast housing prices.

Key Steps in Data Cleaning

Essential procedures for preparing housing data for analysis

01 **Outlier Detection Process**

Removed 24 outlier properties to ensure data quality and accuracy.

02 **Handling Missing Values**

Categorical and numerical features were addressed to maintain dataset integrity.

03 **Feature Engineering Techniques**

Created new features such as houseage and totalsf to enhance model performance.

04 **Reducing Feature Set**

Dropped low-importance and highly correlated features for better analysis.

Insights from EDA Analysis

Exploring correlations and transformations in housing prices

0.85

High Correlation

Overall Quality showed a strong positive correlation with SalePrice, indicating its importance in pricing.

1.2x

Log Transformation

Applying a **log transformation** to SalePrice helped stabilize variance and improve model performance.

4

Key Zones

The **zoning classification** had a notable impact on property values, with certain zones fetching higher prices.

Regression Models Overview

Key Techniques for Price Prediction

Base Models in Regression

Various **base models** are utilized for predicting housing prices, including **Linear Regression**, **Random Forest**, and **XGBoost**. Each model offers distinct advantages and can be chosen based on the dataset characteristics.

Advanced Ensemble Methods

To enhance predictive accuracy, **ensemble methods** such as **Voting Regressor** and **Stacking Regressor** are employed. These techniques combine multiple models to improve performance and robustness.

Training Data Strategies

The **training approach** involves an **80-20 train-test split** along with **cross-validation**. This strategy ensures reliable evaluation of model performance and reduces overfitting risks.

Comparison of Model Performance

Evaluation of various regression models for housing price predictions

Best performance by Stacking Regressor

The Stacking Regressor demonstrated superior accuracy and predictive power compared to other models evaluated in the study.

Voting Regressor ranks second

Following closely, the Voting Regressor also showed commendable results, significantly contributing to the overall model ensemble.

Final model advantages noted

The final model's key strength lies in its ability to combine the **strengths** of multiple regression techniques, enhancing overall effectiveness.