■ House Prices: Advanced Regression Techniques

Overview

This project predicts home sale prices using the Kaggle 'House Prices: Advanced Regression Techniques' dataset. It includes preprocessing, visualization, and machine learning regression modeling to predict SalePrice.

Goal

Predict the SalePrice for each house in the test dataset. The competition evaluates predictions using Root Mean Squared Error (RMSE) between the logarithm of predicted and actual prices.

Dataset

Files used: - train.csv: 1460 rows × 81 columns (features + SalePrice target) - test.csv: 1459 rows × 80 columns (no SalePrice) - data_description.txt: Column descriptions - competition.ipynb: Main notebook containing analysis and model training.

Data Preprocessing

• Handled missing numeric values with median • Handled missing categorical values with most frequent • Encoded categorical columns using One-Hot Encoding • Log-transformed SalePrice for stability • Split data into training and validation folds

Model

A Random Forest Regressor was used inside a scikit-learn Pipeline: - SimpleImputer (median/mode) - OneHotEncoder - RandomForestRegressor (n_estimators=500, random_state=42) Metric: Cross-Validation Log RMSE (Root Mean Squared Error on log-transformed SalePrice).

Visualization

Key plot: Living Area vs Log(SalePrice) This visualization shows a strong positive correlation between above-ground living area (GrLivArea) and log(SalePrice). It also highlights outliers and demonstrates why log-transforming SalePrice is effective for modeling.

Results

Baseline RandomForestRegressor achieved approximately: • Cross-Validation Log RMSE ≈ 0.135

Submission Format

Submission file example: Id,SalePrice 1461,169000.1 1462,187724.12 1463,175221.0

Future Improvements

• Add feature engineering (TotalSF, OverallScore) • Try Gradient Boosting / XGBoost / LightGBM • Perform hyperparameter tuning • Remove outliers to improve model performance

Author

Your Name Data Science Student | Kaggle Competitor Created: November 2025