HOUSE PRICE PREDICTION USING ADVANCED REGRESSION TECHNIQUES

-Surya Prabha V P

Abstract

This project focuses on predicting house prices using a dataset from a Kaggle competition. I followed a complete data science workflow, including **exploratory data analysis**, **feature engineering**, and **model building**. I implemented and compared several advanced regression techniques, ultimately finding that **Lasso Regression** provided the most accurate predictions. This project demonstrates my proficiency in using Python for data analysis and machine learning to solve a real-world regression problem.

Introduction

House price prediction is a classic machine learning regression problem with significant real-world applications for real estate agents, buyers, and sellers. My goal was to predict a continuous value (SalePrice) based on various features of a house. This report details the steps I took to build a predictive model, starting from data preprocessing and concluding with a comparison of advanced regression techniques.

Dataset Description & EDA

The dataset, sourced from a Kaggle competition, contains 79 explanatory variables and a single target variable, SalePrice, for 1,460 houses.

Data Cleaning

- **Missing Values:** I handled missing data by filling it with 'None' for categorical features and 0 for numerical features where a missing value indicated the absence of a feature (e.g., PoolQC for a house with no pool). For the LotFrontage feature, which is truly missing, I filled the values with the mean.
- **Outliers:** During my bivariate analysis, I identified and removed two significant outliers in the GrLivArea feature, as they were houses with

exceptionally large living areas but unusually low prices, which would have negatively impacted my model.

EDA Findings

- **Distribution of SalePrice:** The target variable, SalePrice, was heavily right-skewed. To meet the assumptions of many linear models and improve performance, I applied a **log transformation** to normalize the distribution.
- **Key Feature Insights:** A correlation heatmap revealed that **OverallQual** (Overall Material and Finish Quality) and **GrLivArea** (Above-Ground Living Area) were the features most correlated with the SalePrice. A **T-test** was performed on OverallQual, and the p-value of 0.0000 proved that high-quality houses have a statistically significant difference in price from low-quality houses. The Neighborhood feature was also found to be a powerful predictor of price.

Methodology (Model Building)

After data preprocessing and EDA, I prepared the data for modeling by applying **one-hot encoding** to all categorical variables. This converted 79 original features into **259 numerical features** that the models could understand. I then split the data into a training set and a testing set.

I applied and evaluated four different regression models:

- **Linear Regression:** Served as a baseline to establish an initial performance score.
- **Lasso Regression:** A regularized model that shrinks the coefficients of less important features, effectively performing feature selection.
- **XGBoost:** A powerful gradient boosting algorithm known for its high performance on structured data.
- **Random Forest:** An ensemble model that uses multiple decision trees to improve accuracy.

Results and Discussion

I evaluated each model's performance using three key metrics: **Root Mean Squared Error (RMSE)**, **R^2 score**, and **Mean Absolute Error (MAE)**. The results are summarized in the table below:

Model	RMSE (lower is better)	R^2 (closer to 1 is better)	MAE (lower is better)
Lasso Regression	**0.1245**	**0.9081**	**0.0862**
XGBoost	0.1411	0.8819	0.0932
Linear Regression	0.1449	0.8883	0.0935
Random Forest	0.1468	0.8722	0.0978

- **Best-Performing Model:** Lasso Regression was the top performer, achieving the lowest RMSE and MAE, as well as the highest R^2 score.
- **Why Lasso Excelled:** Its superior performance is likely due to its regularization technique. With 259 features after one-hot encoding, Lasso helped prevent overfitting by automatically selecting the most important features and shrinking the coefficients of irrelevant ones to zero.
- **Model Comparison:** While XGBoost and Linear Regression also performed well, the subtle improvements from Lasso's regularization were enough to give it the edge. Random Forest, an ensemble model, also provided solid performance.

Conclusion

- **Project Goal Achieved:** I successfully built and evaluated several models for house price prediction, fulfilling the project's objective.
- **Key Findings:** The analysis confirmed that OverallQual and GrLivArea are the most influential features in determining a house's value.
- **Best Model Recommendation:** Lasso Regression is the best model for this dataset due to its high accuracy and ability to perform automatic feature selection, which is ideal for a high-dimensional dataset.