CS-20131 Maaz Jawaid

Detailed Report on House Price Prediction

1. Introduction

The objective of this project is to predict house prices using various features related to properties. The dataset includes both numerical and categorical features that describe different aspects of the houses, including their characteristics, conditions, and historical details.

1.1 Data Collection

The dataset was obtained from Kaggle, a platform known for hosting machine learning competitions and providing datasets for practice and real-world applications. The specific competition where this dataset was sourced aimed to predict house prices based on various explanatory variables.

2. Data Preprocessing

2.1 Loading and Understanding the Data

The dataset is loaded into a DataFrame for initial inspection to understand its structure and content. This includes:

- Displaying the first few rows to get an overview of the data.
- Checking the data types of each column.
- Summarizing the numerical and categorical features.

2.2 Handling Missing Values

- Missing values are identified using `isnull().sum()` and visualized using a heatmap.
- Columns with a significant number of missing values are dropped.
- Missing values in numerical columns are filled with the median, and in categorical columns with the mode.

3. Feature Engineering

3.1 Encoding Categorical Variables

Categorical variables are converted into numerical form using one-hot encoding. Specific columns that are numerical but represent categories (e.g., `MSSubClass`, `YearBuilt`, `YearRemodAdd`, `GarageYrBlt`, `YrSold`) are converted to string type for proper one-hot encoding.

3.2 Handling Temporal Variables

The `MoSold` column (month of sale) is converted to its corresponding month abbreviation.

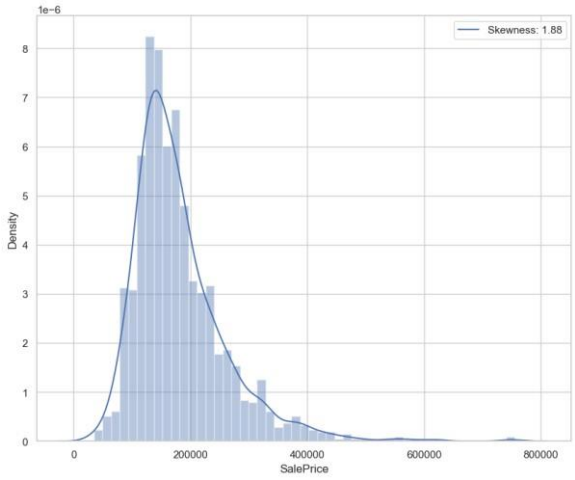
3.3 Log Transformation

Log transformation is applied to skewed numerical features to reduce skewness.

4. Exploratory Data Analysis (EDA)

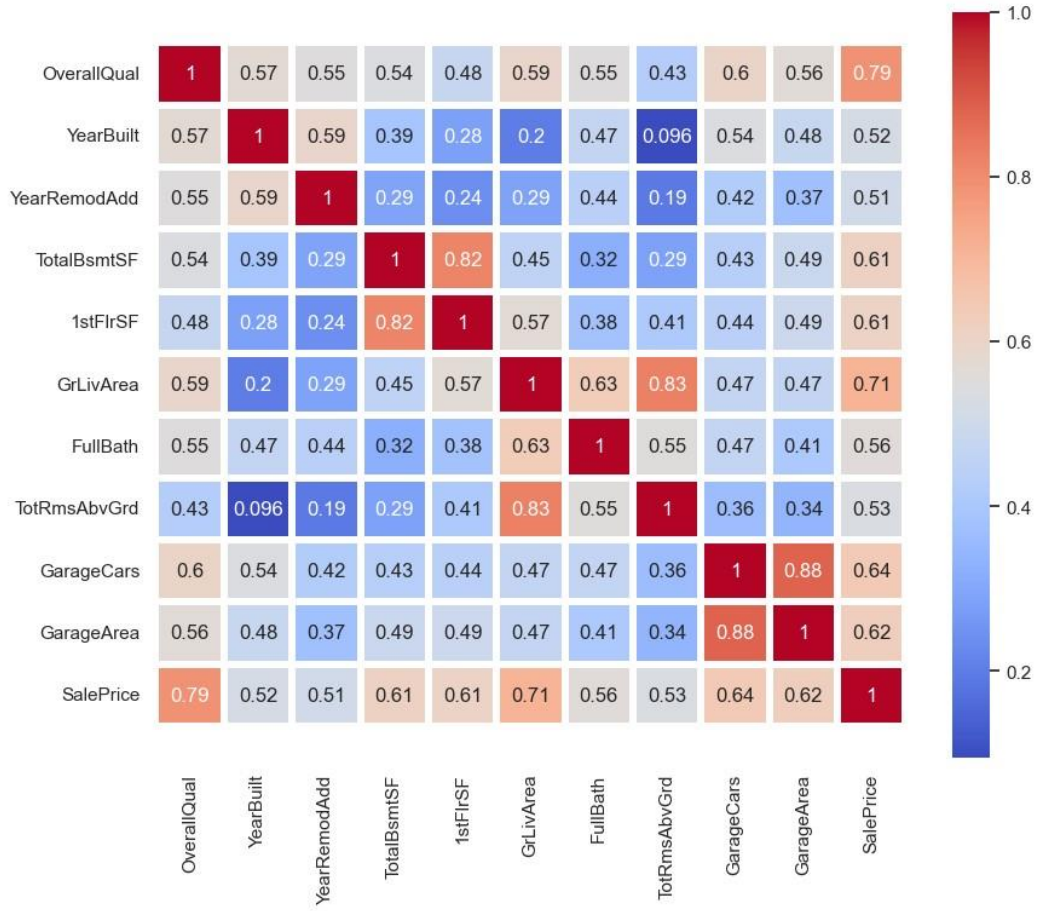
4.1 Target Variable Distribution

The distribution of the target variable, `SalePrice`, is plotted to understand its skewness.



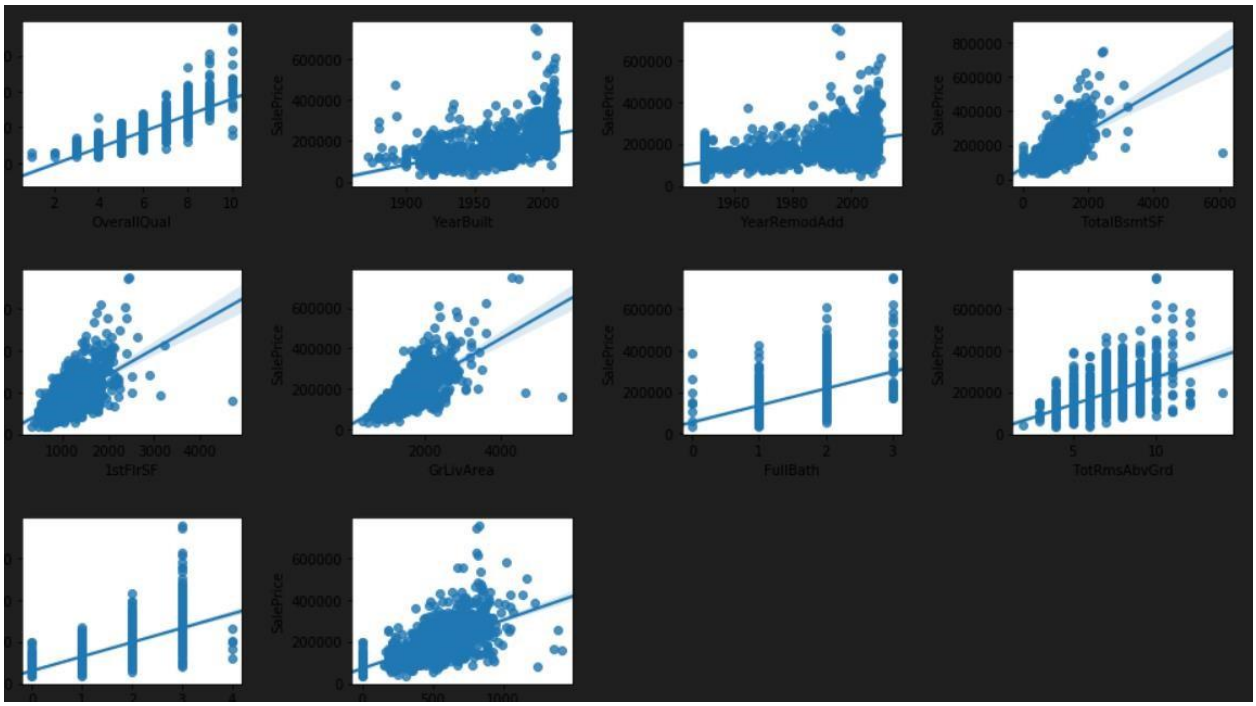
4.2 Correlation Analysis

A heatmap is created to show the correlation between different features and `SalePrice`.



4.3 Feature Relationships

Scatter plots or box plots are used to visualize the relationship between the most significant features and the target variable.



5. Model Building

5.1 Splitting the Data

The data is split into training and testing sets based on the original training and test data sizes.

5.2 Scaling the Data

Features are scaled using a `RobustScaler` to handle outliers effectively.

5.3 Model Selection

Various models are trained and evaluated using cross-validation, including:

- **Linear Regression**
- **Ridge Regression**
- **Lasso Regression**
- **Support Vector Regression (SVR)**
- **Decision Tree Regressor**
- **Random Forest Regressor**
- **Bagging Regressor**
- **Gradient Boosting Regressor**
- **XGBoost Regressor**

5.4 Model Tuning

Hyperparameter tuning is performed for the SVR model using `RandomizedSearchCV`.

Code Snippet:

5.5 Model Evaluation

The best-performing model, which is the SVR model after hyperparameter tuning, is selected. Predictions are made on the test set, and results are saved in a CSV file for submission.

Conclusion

The report summarizes the data preprocessing steps, exploratory data analysis, feature engineering, and model building process. The SVR model with tuned hyperparameters achieved the highest crossvalidation score. The predictions made by this model are saved and ready for submission.