House Price Prediction Project

05/18/2024

Table of Contents

Project Objective

Objective:

- Accurately predict house prices
- Analyze the impact of different features on house prices using machine learning models

Dataset and Features

Source:

- Kaggle House Prices: Advanced Regression Techniques

Features:

- 79 explanatory variables
- Target variable: SalePrice (house sale price)

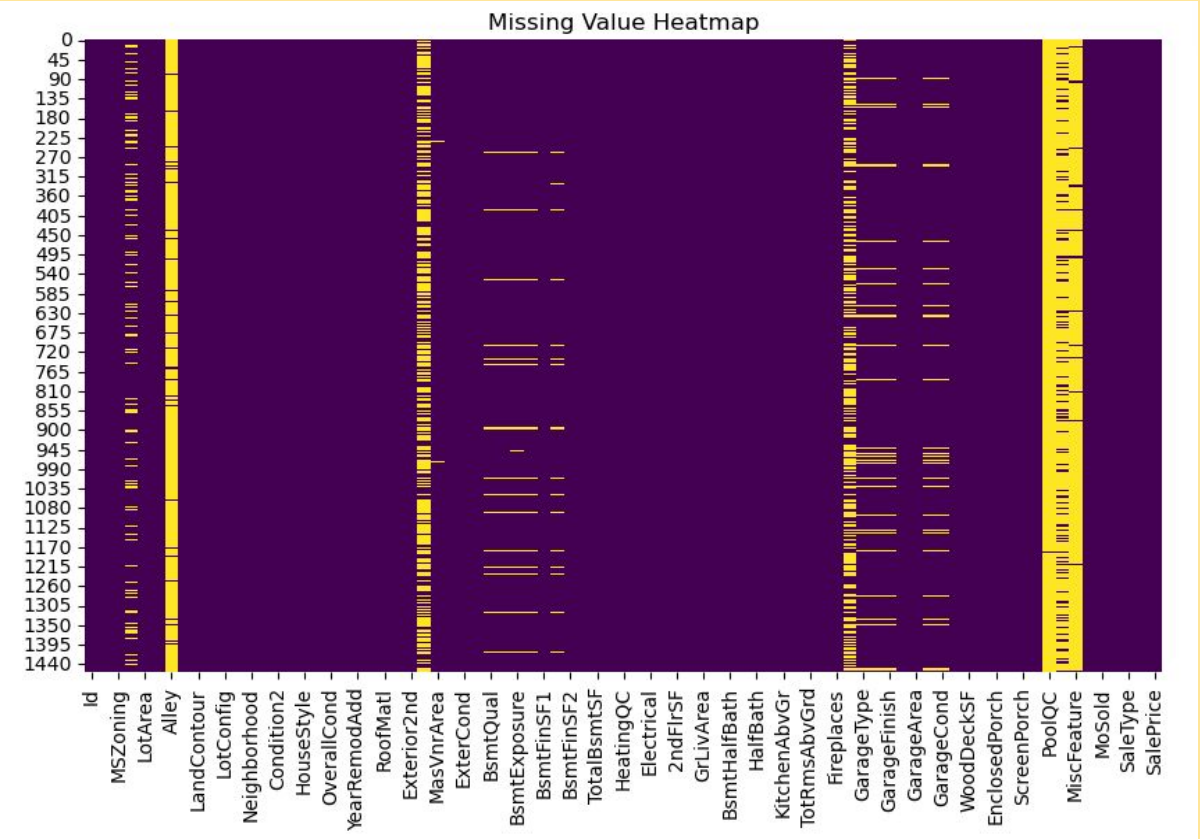
Data Processing

Steps:

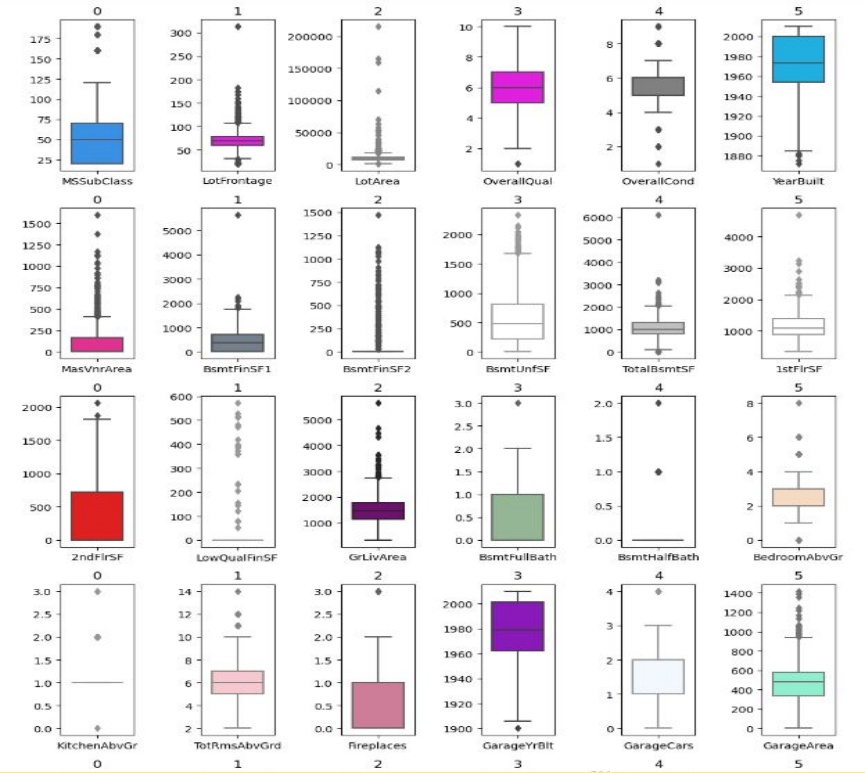
- Handling Missing Values
- Detecting and Handling Outliers
- Feature Engineering
- Feature Transformations (Label Encoding, One-Hot Encoding, Scaling)

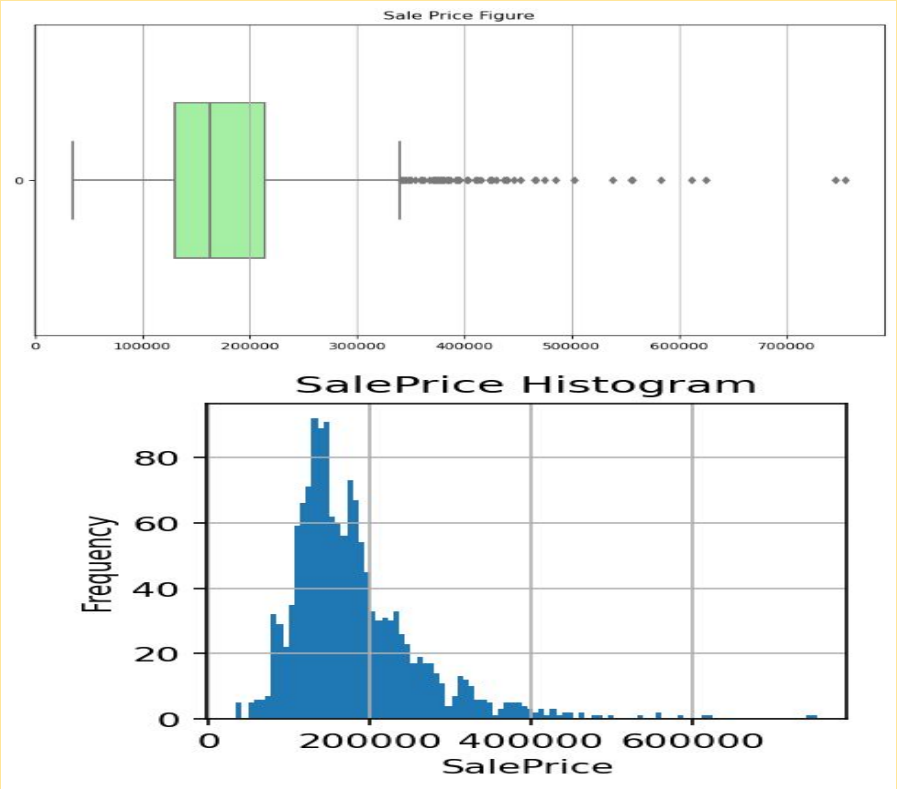
Handling Missing Values

- Identify Missing Data
- Imputation



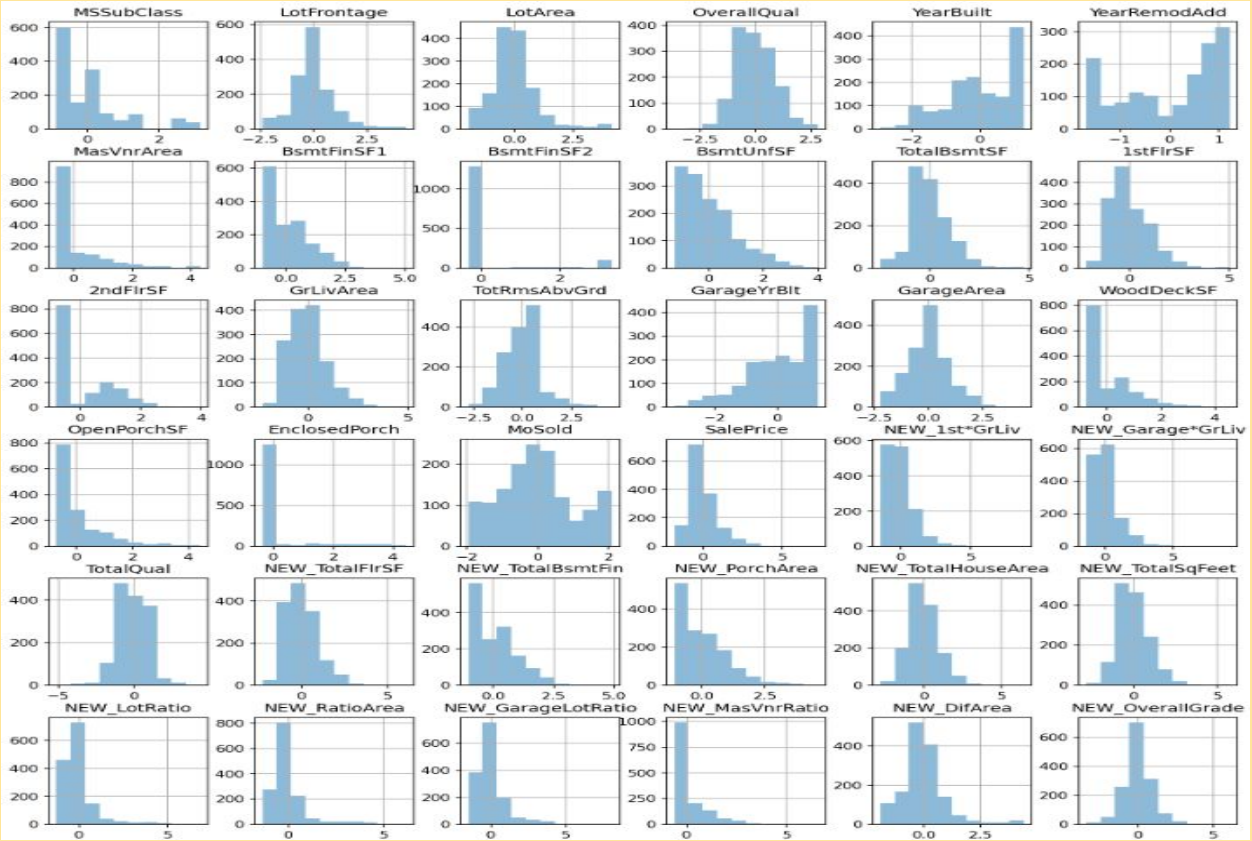
Detecting and Handling Outliers





Feature Engineering

- Creation of New Features
- Feature Interaction



Feature Transformations (Label Encoding, One-Hot Encoding, Scaling)

- **Label Encoding:** Categorical variables are converted into numerical values using label encoding.
- **One-Hot Encoding:** For nominal categorical variables, one-hot encoding is used to avoid any ordinal assumptions.
- **Scaling:** Numerical features are scaled using standardization or normalization to ensure they have a standard scale, which is crucial for algorithms sensitive to feature scaling.

Modeling

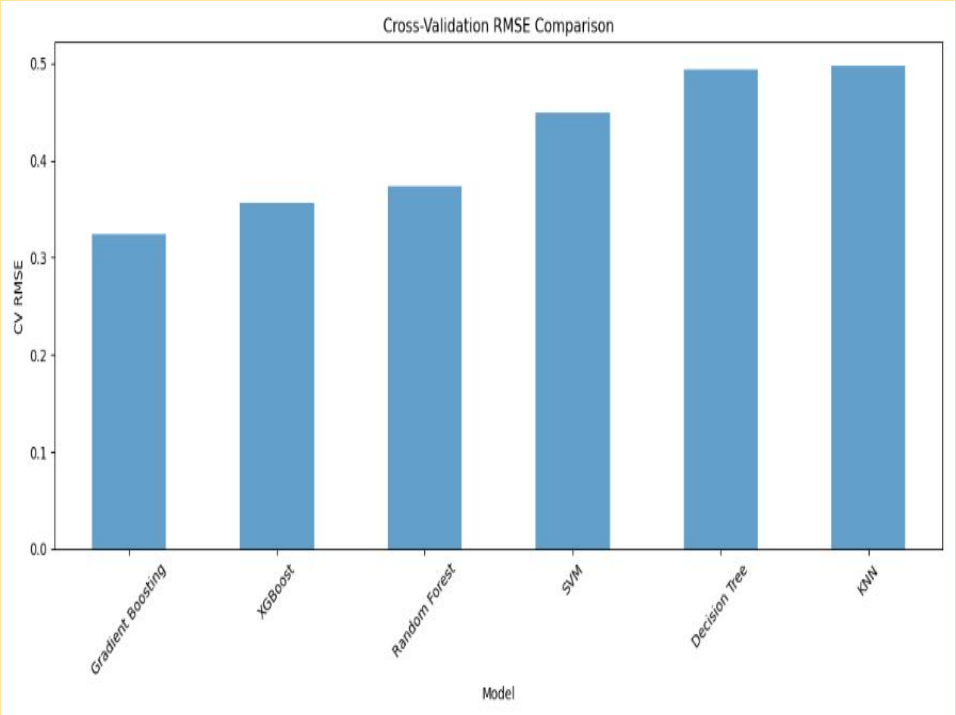
Models Tried:

- Linear Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- XGBoost

Model Performance Evaluation

XGBoost:

- - Best performance
- - Low error rate
- - High R2 score
- - Cross-Validation RMSE: Lowest among tested models



Results and Evaluation

- XGBoost model showed the best performance for predicting house prices.
- Gradient Boosting also performed well.
- Linear Regression was not suitable for this dataset.

Recommendations and Future Work

Recommendations:

- Use the XGBoost model for predicting house prices.
- Monitor and retrain the model regularly with updated data.
- Further optimize and fine-tune hyperparameters to enhance performance.

Future Work:

- Enrich the model with additional data sources.
- Explore other algorithms to improve generalization.

Thank You

- Questions?