Team:
Leshma Gudupu
Margi Hingrajia

Project Report: House Price Prediction

1. Project Overview

The objective of this project is to predict the sale price of houses based on various features using machine learning models. The dataset consists of multiple features such as the size of the house, the number of rooms, the year of construction, and the neighborhood, among others. The goal is to create a model that can accurately predict house prices based on these features, which can be useful for real estate investors, property buyers, and other stakeholders in the housing market.

2. Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for model training. The dataset contains both numerical and categorical features, with missing values that need to be handled before model training. Below are the preprocessing steps applied:

Handling Missing Values:

For categorical columns, missing values were filled with the string "None".

For numerical columns, missing values were filled with the median of the respective column.

One-Hot Encoding:

Categorical variables were transformed into a numerical format using one-hot encoding, creating binary columns for each category. This helps the models interpret the data in a format suitable for machine learning algorithms.

Log Transformation:

Since the target variable (SalePrice) is highly skewed, a log transformation was applied to it. This helps to stabilize variance and make the model more effective.

Train-Test Split:

The dataset was split into training and validation sets (80% training, 20% validation) to evaluate model performance.

3. Exploratory Data Analysis (EDA)

EDA is an essential part of any data science project as it helps understand the underlying structure of the data and identify trends, patterns, and potential issues. Here are the key insights derived from the EDA:

Target Distribution:

The distribution of house prices was skewed, indicating the need for a transformation to normalize the data.

Correlation Analysis:

Strong positive correlations were observed between certain numerical features such as "GrLivArea" (above-ground living area) and "SalePrice". Features like "OverallQual" (Overall quality) also showed strong correlations with the target variable.

Outlier Detection:

Certain outliers were identified, such as properties with extremely high sale prices or unusual square footage. These outliers could impact the performance of the model, so they were handled during preprocessing.

Categorical Feature Analysis:

The analysis of categorical features like "Neighborhood" showed that location is an important factor influencing house prices, with certain neighborhoods consistently having higher prices.

4. Model Building and Evaluation

Several machine learning models were tested for predicting house prices. The models used are:

Linear Regression:

A simple linear regression model was trained using the features to predict the house prices. It served as the baseline model.

Ridge Regression:

Ridge regression was used to handle multicollinearity by adding a regularization term. It is especially useful when there are many features with correlated variables.

Lasso Regression:

Lasso regression is another regularization technique that helps perform feature selection by shrinking some coefficients to zero.

Random Forest Regressor:

A Random Forest model was used, which works well with both numerical and categorical features and is capable of capturing non-linear relationships in the data.

XGBoost:

XGBoost is an advanced boosting algorithm that has been proven to perform well in regression tasks. It works by combining several weak models to create a strong model.

5. Comparison of Models

Each model was trained on the same training data, and the performance was evaluated using Root Mean Squared Error (RMSE) on the validation set. The results are summarized below:

Model	RMSE
Linear Regression	0.13215
Ridge Regression	0.13645
Lasso Regression	0.13796
Random Forest	0.14572
XGBoost	0.13201

XGBoost performed the best with the lowest RMSE value of 0.13201, followed by Linear Regression with 0.13215.

The Random Forest model performed the worst with an RMSE of 0.14572, which could be due to overfitting or inadequate tuning of hyperparameters.

6. Feature Importance & Insights

Feature importance refers to how much each feature contributes to the prediction of the target variable. In this project, XGBoost provided insights into the most important features based on their contribution to the model's prediction. Some of the most important features included:

OverallQual (Overall quality): The quality of the property is a major predictor of its price.

GrLivArea (Above ground living area): The larger the living area, the higher the price.

Year Built (Year of construction): Newer homes tend to have higher prices.

TotRmsAbvGrd (Total rooms above ground): More rooms generally correspond to higher prices.

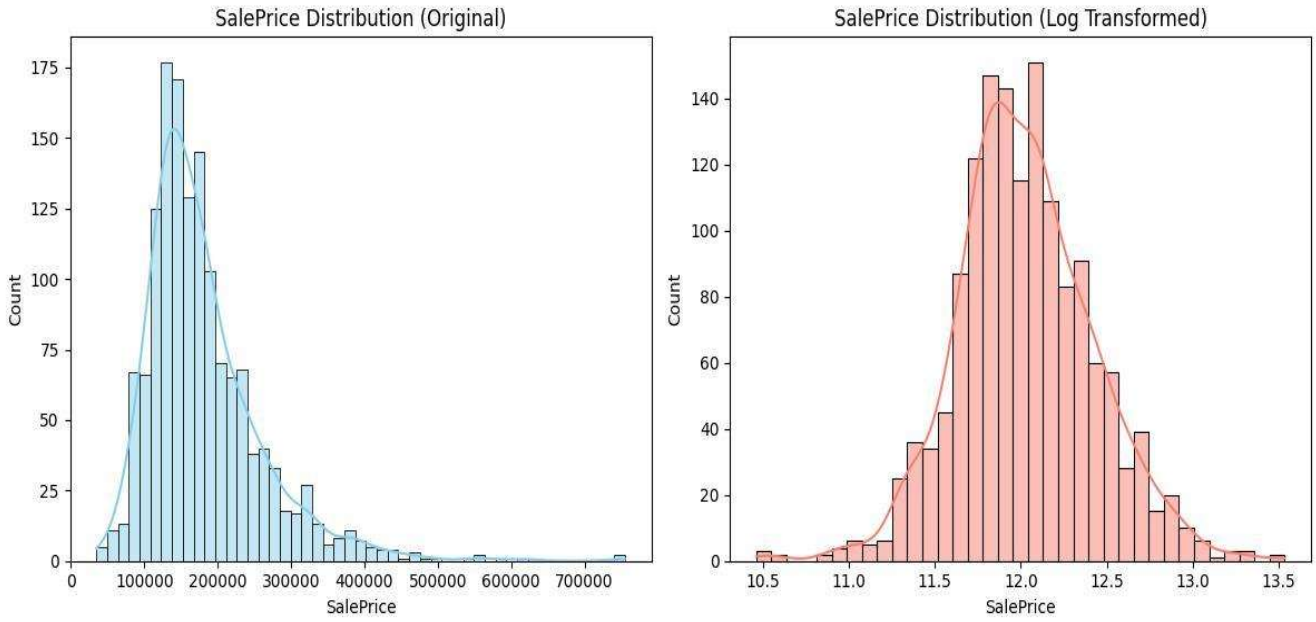
By analyzing feature importance, we gain valuable insights into what drives house prices. This can inform both real estate investment strategies and decisions for property buyers.

7. Visualizations

Several visualizations were created to better understand the data and model performance:

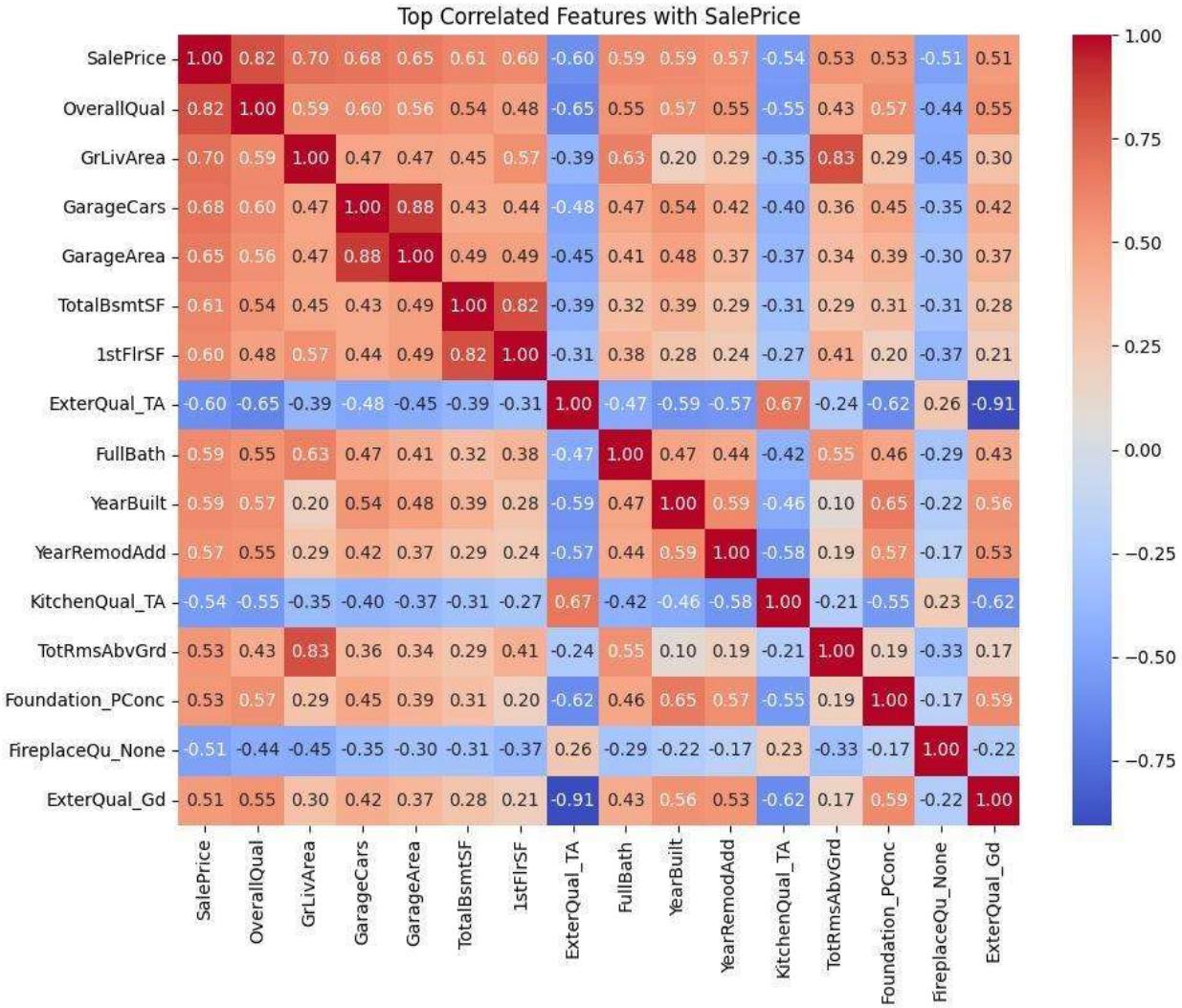
Target Distribution:

A histogram of the target variable, SalePrice, was shown before and after applying a log transformation.



Correlation Heatmap:

A heatmap of correlations was plotted to identify features with high correlations with the target variable.

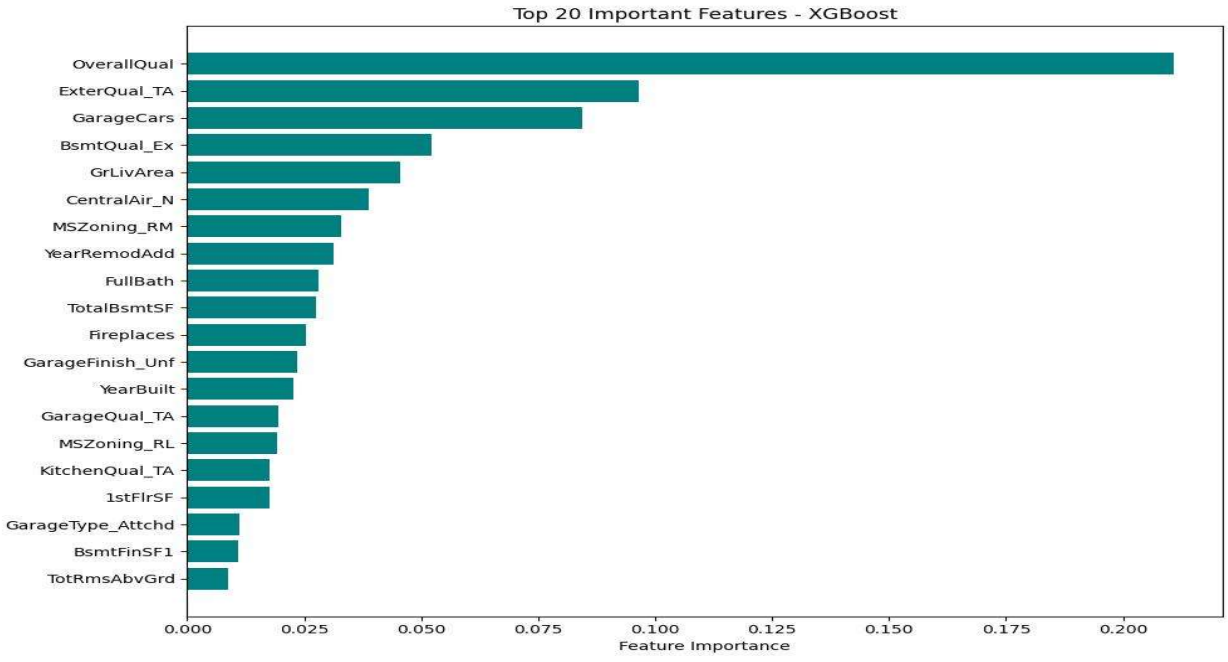


Model Performance Comparison:

A bar plot was created to compare the RMSE values of different models. This helps visualize which models performed better.

Feature Importance Plot:

For the final XGBoost model, a bar chart of feature importance was displayed, showing how much each feature contributed to the final predictions.



8. Conclusion & Recommendations

The following conclusions were drawn from the project:

Best Performing Model: The XGBoost model achieved the lowest RMSE, indicating it was the best model for this problem. It outperformed other models such as Linear Regression and Random Forest.

Data Preprocessing Importance: Proper handling of missing values, encoding categorical variables, and transforming the target variable significantly improved model performance.

Feature Selection: Key features such as "OverallQual" and "GrLivArea" were highly important for predicting house prices. Understanding feature importance can help focus on the most relevant aspects when analyzing properties.

Further Improvements: Hyperparameter tuning, using advanced techniques like cross-validation, and testing additional models such as gradient boosting machines or deep learning could potentially improve performance further.

Recommendations:

For Real Estate Investors: This model can assist in identifying properties likely to have higher sale prices based on their features, location, and overall quality.

For Homebuyers: Buyers can use this model to estimate the fair price of homes in different neighborhoods and assess the impact of specific features on property prices.

Links:

Code:
https://colab.research.google.com/drive/1wsmJhKm76mKxQl7LomDOZblP2KTUEZoN?usp=sharing

Video presentation:
https://1drv.ms/v/c/714c7d964557e947/EaYpbtldbc1MjzfX1JofuSoBI7WU-DCvhVdN1HsFWs1nVQ?e=7EDug6

Dataset:
https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data?select=sample_submission.csv