Housing Prices Prediction

Project Write up

Understanding and forecasting house prices is an important sign of economic stability and a component of personal financial planning. The difficulty stems from the complexities of real estate markets, which are influenced by a variety of factors ranging from the macroeconomic climate to the characteristics of local communities. This research aims to give significant tools for real estate buyers, sellers, and analysts by addressing this complicated problem using machine learning approaches to generate predictive insights from historical housing data.

Data Overview

The dataset used in this investigation is obtained from a large database of housing sales records. It includes a wide range of features that characterize the physical characteristics of the property, the quality and condition of its numerous elements, and the characteristics of its surroundings. The dataset contains approximately 1,000 records, each representing a unique instance of a property's potential market worth. This dataset will serve as the foundation for our predictive models.

Data Preprocessing

The initial phase of the project was focused on data preprocessing, a critical step to ensure the quality of the analysis. This stage of the project was clearly documented, with a focus on transparency and reproducibility. The data preprocessing involved several steps outlined in the project's codebase:

Handling Missing Values: An approach was adopted to manage missing data, as seen in the code snippet below. Numerical attributes with missing values were imputed with the mean of the respective feature, while categorical attributes with missing entries were populated with a 'None' placeholder to indicate the data's absence. This strategy was carefully chosen to maintain a consistent aspect of the dataset while preparing it for the modeling process.

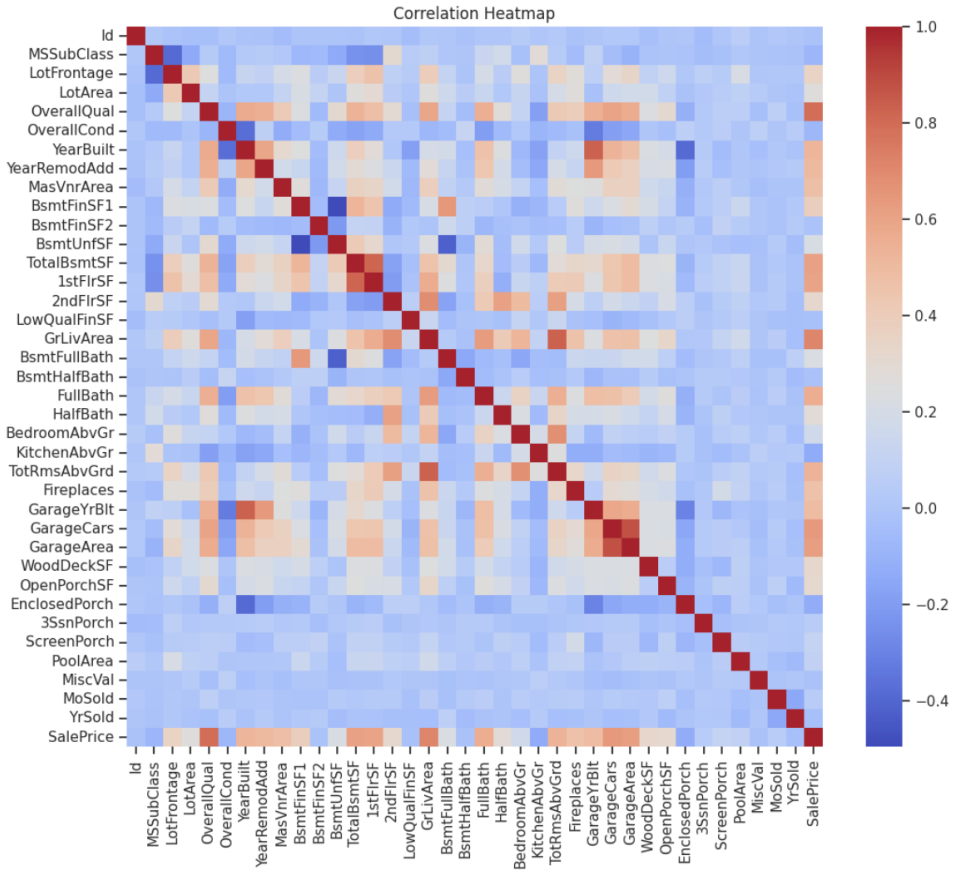
```
for column in ['LotFrontage', 'GarageYrBlt', 'MasVnrArea']:
    numerical_attributes[column].fillna(numerical_attributes[column].mean(), inplace=True)

columns_to_delete_categorical = ['PoolQC', 'MiscFeature', 'Alley', 'Fence', 'MasVnrType', 'FireplaceQu']
categorical_columns = categorical_columns.drop(columns_to_delete_categorical, axis=1)
categorical_columns = categorical_columns.fillna('None')
```

Pruning Features: The decision to retain or exclude certain features was based on the proportion of missing data. Features with a substantial number of missing values were excluded, ensuring the reliability of the dataset for the model.

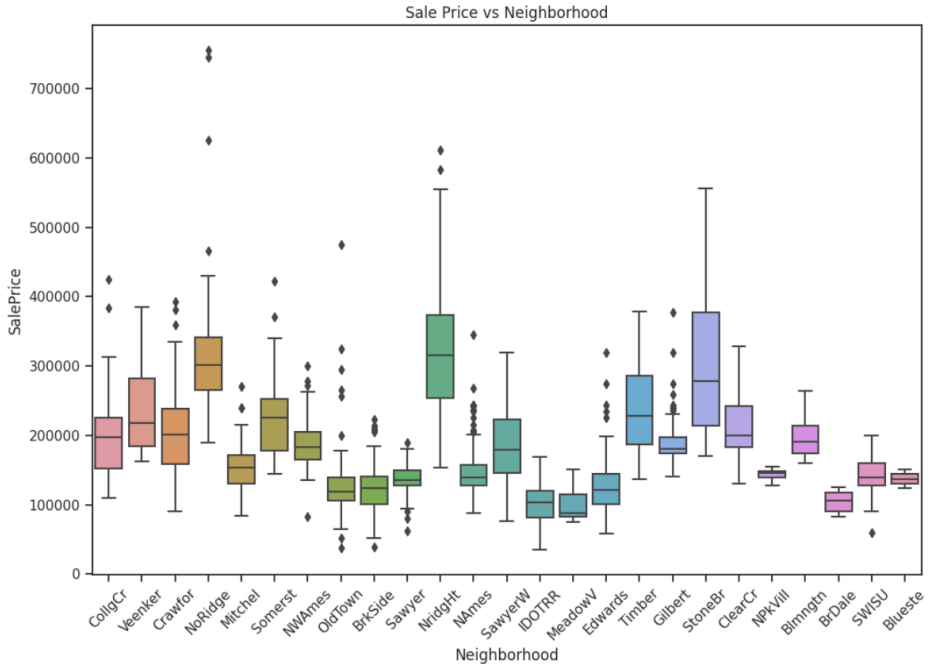
These preprocessing steps were critical for the dataset, setting a solid base for the subsequent stages of feature engineering and model development.The subsequent visualizations created after the preprocessing steps provided empirical insights into the dataset's structure and the relationships between variables:

Visualization 1 showcases the distribution of 'SalePrice', offering a glimpse into the range and skewness of property values within the dataset.

Visualization 2 presents the correlation heatmap, revealing the strength and direction of relationships between various numerical features.



Visualization 3 offers box plots that compare 'SalePrice' across different categories of the 'Neighborhood' feature, illustrating the impact of location on housing prices.



Visualization 4 provides scatter plots for 'GrLivArea' against 'SalePrice', underscoring the relationship between living area size and property value.

The insights from these visualizations were important in informing the feature engineering and modeling phases, ultimately leading to the development of a predictive model aimed at estimating housing prices with a significant degree of accuracy.

Feature Engineering

The project's feature engineering phase was designed to enhance the dataset's predictive capacity by introducing new features and transforming existing ones:

Interaction Term:
Recognizing the compounded effect of a home's living area and basement size, an interaction term 'GrLivArea_TotalBsmtSF' was engineered. This new feature aimed to capture the synergistic value these two aspects might have on the overall price.

Logarithmic Transformation:
A logarithmic transformation was applied to the target variable 'SalePrice'. This transformation was used when dealing with right-skewed distributions, as it helps stabilize variance and normalize the data. The transformed 'SalePrice' was then used as the target variable for model training.

Model Development

For the model development, the RandomForest Regressor was selected for its proficiency in managing datasets with a multitude of features and its protection against overfitting. This method constructs a multitude of decision trees during training and outputs the average prediction of the individual trees, leading to a more accurate and stable prediction.

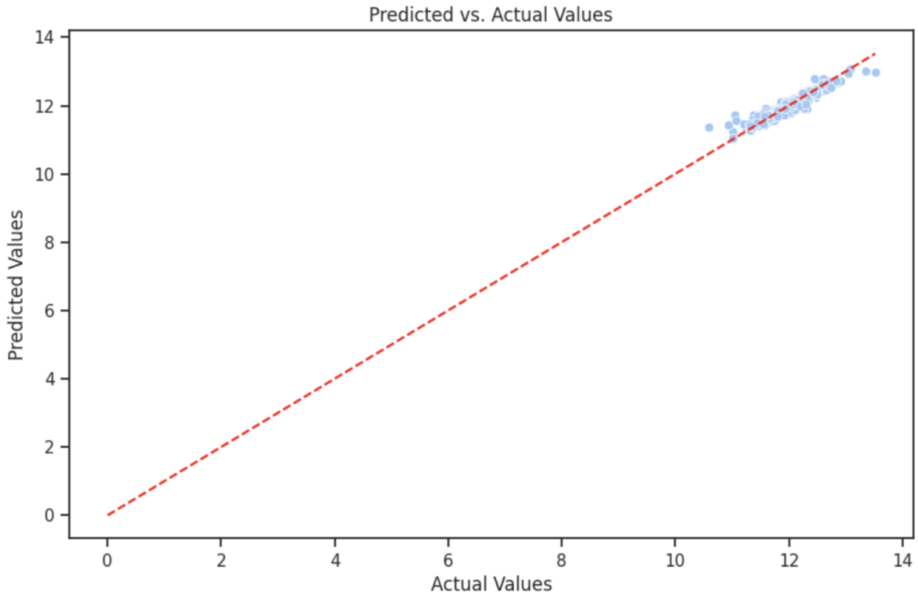
To further refine the model, hyperparameter tuning was conducted using GridSearchCV. This process involved iterating through predefined parameter grids for 'n_estimators' and 'max_depth' to identify the combination that produced the best results based on the negative mean squared error scoring method. The best parameters from GridSearchCV were then used to train the final model.

Model Evaluation

The model's effectiveness was quantified using the Mean Absolute Error (MAE) metric on the validation set, providing a clear measure of the average magnitude of the errors in the predictions:

```
Validation Mean Absolute Error: 0.09625614620733751
```

In addition to the quantitative evaluation, qualitative analysis was performed through visual means:



Prediction vs. Actual Scatter Plot:

This scatter plot was essential for visualizing the model's predictions in relation to the actual sale prices. Ideally, the points should align closely with the identity line, indicating precise predictions. Deviations from this line were indicative of prediction inaccuracies and their patterns could suggest potential areas for model improvement.

Conclusion:

In conclusion, our project tackled housing prices, combining data analysis and machine learning to uncover what drives a home's value while also showing how tech can help us understand and predict real estate trends in practical, everyday terms.