Final Project: Housing Price Prediction

Tao Sun

Business Framing

Background:

In the real estate market, it is very important for stakeholders to accurately predict the sale price of real estate. Accurate forecasting helps to make informed decisions, optimize investment returns and understand market trends. A company called "RealLink" specializes in real estate analysis. Their services include providing market insights, property valuations and investment advice to real estate investor clients.

Opportunity:

By building more accurate home price forecasting models, Realink can provide accurate investment recommendations, thereby gaining a competitive edge in the market and improving customer satisfaction.

Problem Statement

Objective:

The main goal is to develop a machine learning model that can predict home sales prices with high accuracy. The model will utilize various attribute characteristics, including size, location, condition, material, etc.

Challenges:

- Models must handle a wide range of features, some of which may have missing or erroneous data. Moreover, we need to analyze the relationship between features to select an appropriate model.
- Properties vary greatly in type, size, condition and location. We need to find a model that generalizes well across different property profiles.

Data:

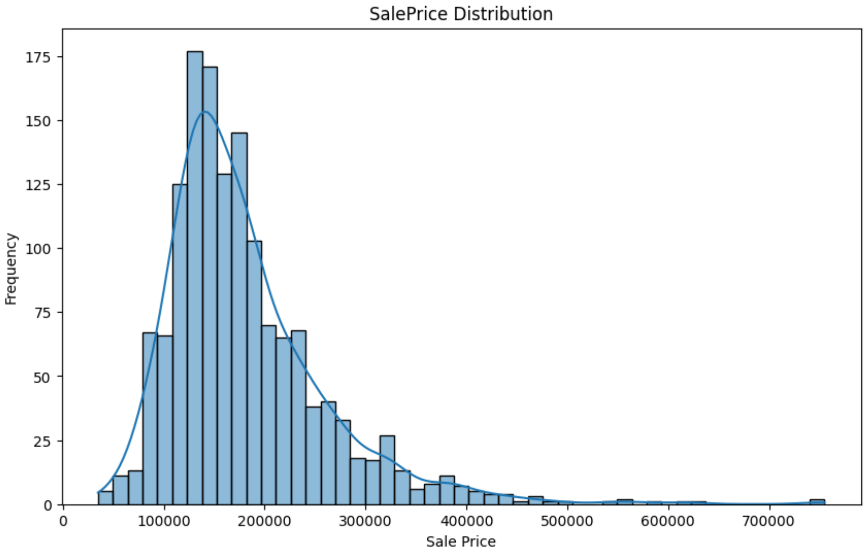
The model will use the provided dataset, which includes historical sales prices and property characteristics. The dataset includes variables such as building category, zoning classification, lot size, number of bedrooms, and more.

Evaluation Metric:

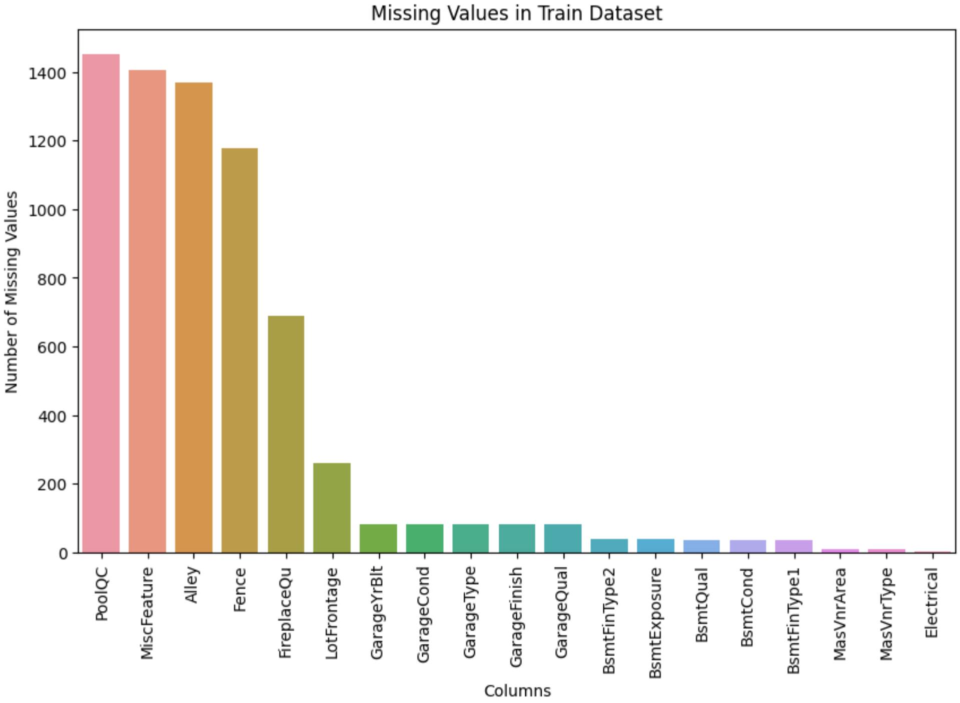
The performance of the model will be evaluated based on the root mean square error (RMSE) between the predicted price and the logarithm of the actual sales price.

EDA

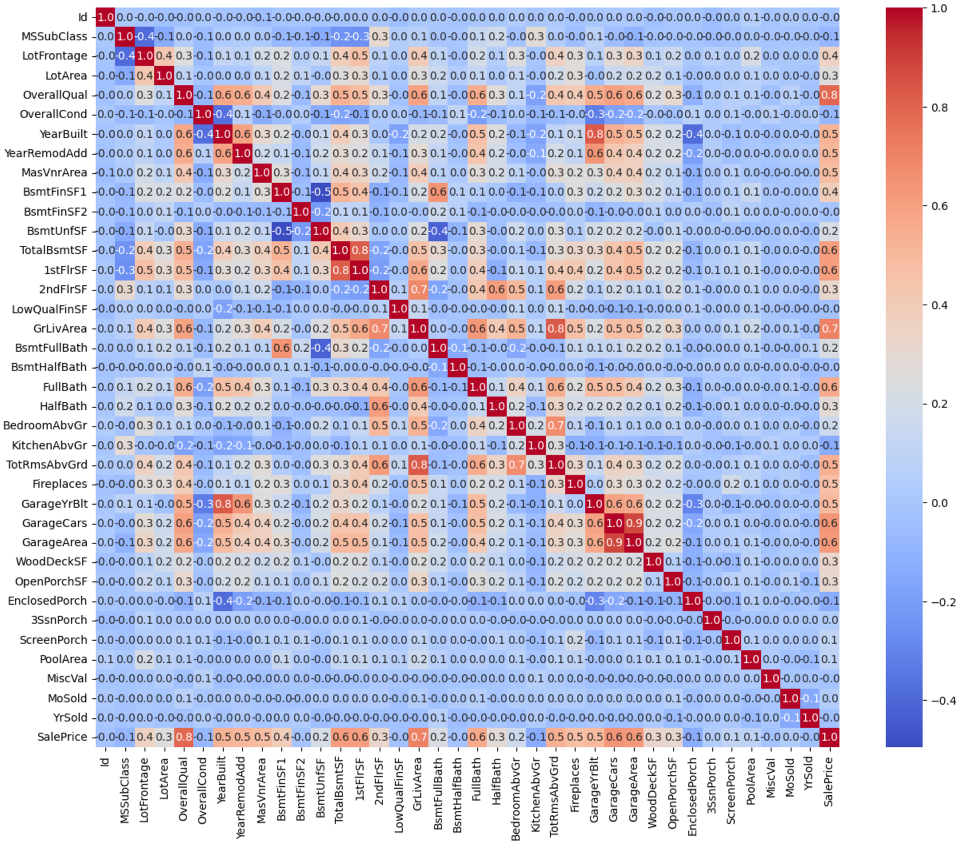
The training data contains 81 columns and 1460 rows. SalePrice is the column we need to predict. The mean of the SalePrice in training data is 180921.195890. And after checking the distribution of the SalePrice, SalePrice follows normal distribution.



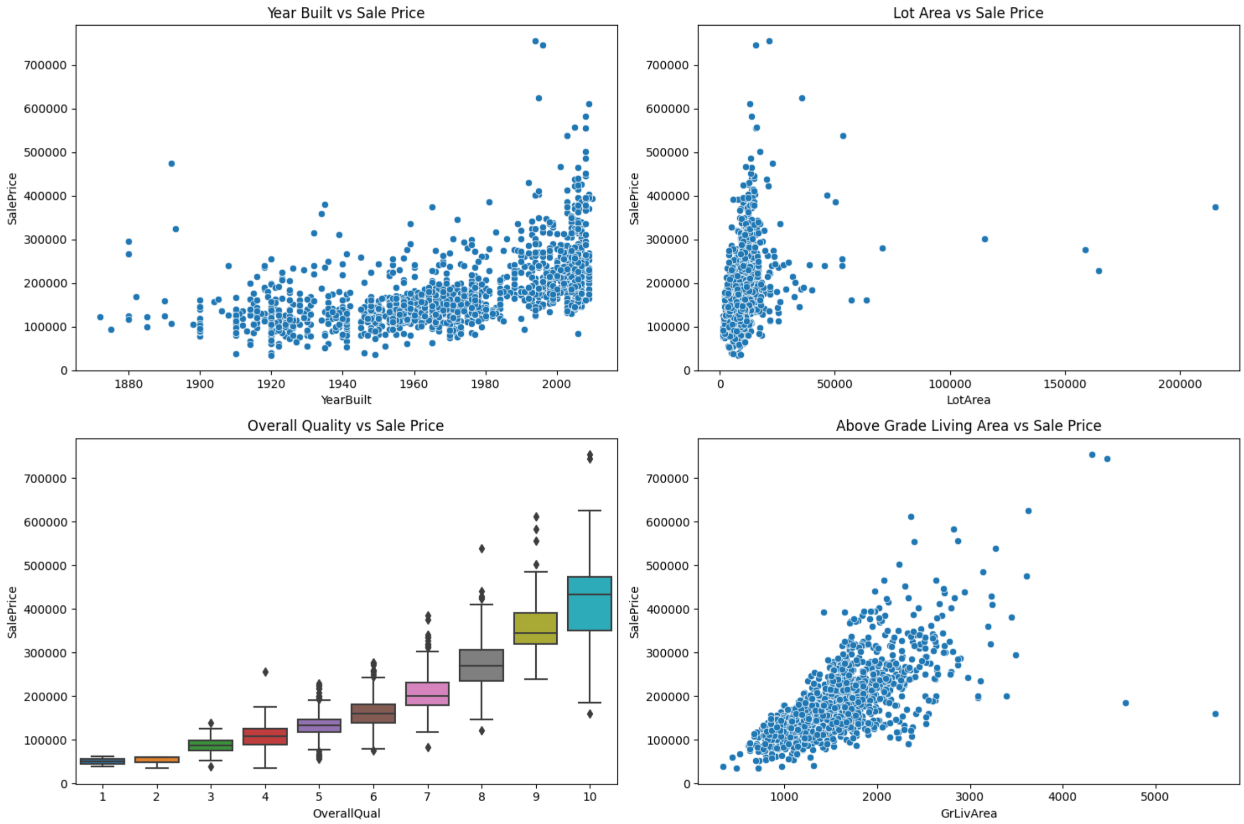
The training data set contains some Null values in some columns. In PoolQC, MiscFeature, Alley, Fence, FireplaceQu columns, they contains many Null values and we need to make decisions to choose important features for the model. Including some columns with many null values may lower accuracy of the model.



According to the heatmap, some numeric columns, such as OverallQual and GrLiveAre, has high positive correlation with the SalePrice.



So, I plot some scatterplots and boxplot for Sale Price with some important numeric columns. YearBuilt, Overall Quality, and Above Grade Living Area show linear relationships with Sale Price. But some features don’t describe linear relationships, such as Lot area feature.



Feature Selection

Based on the EDA, some columns with missing values need to be dropped.

Modeling

According to the EDA, Tree-Based Models can be a good decision because:

- Tree-based models are particularly good at capturing the complex and nonlinear relationships unique to real estate data. By using tree-based models, we eliminate the need for manual feature engineering. There are nearly 80 features in the training set. This definitely reduces our time for feature engineering.
- The data set contains a mixture of numerical and categorical variables. Tree-based models can handle this diversity naturally and do not require extensive preprocessing.
- Many features contain outliers. Tree-Based Models are more tolerant of outliers.
- Tree-Based Models are easier to understand for the market. This helps real estate market stakeholders save more time on model understanding.

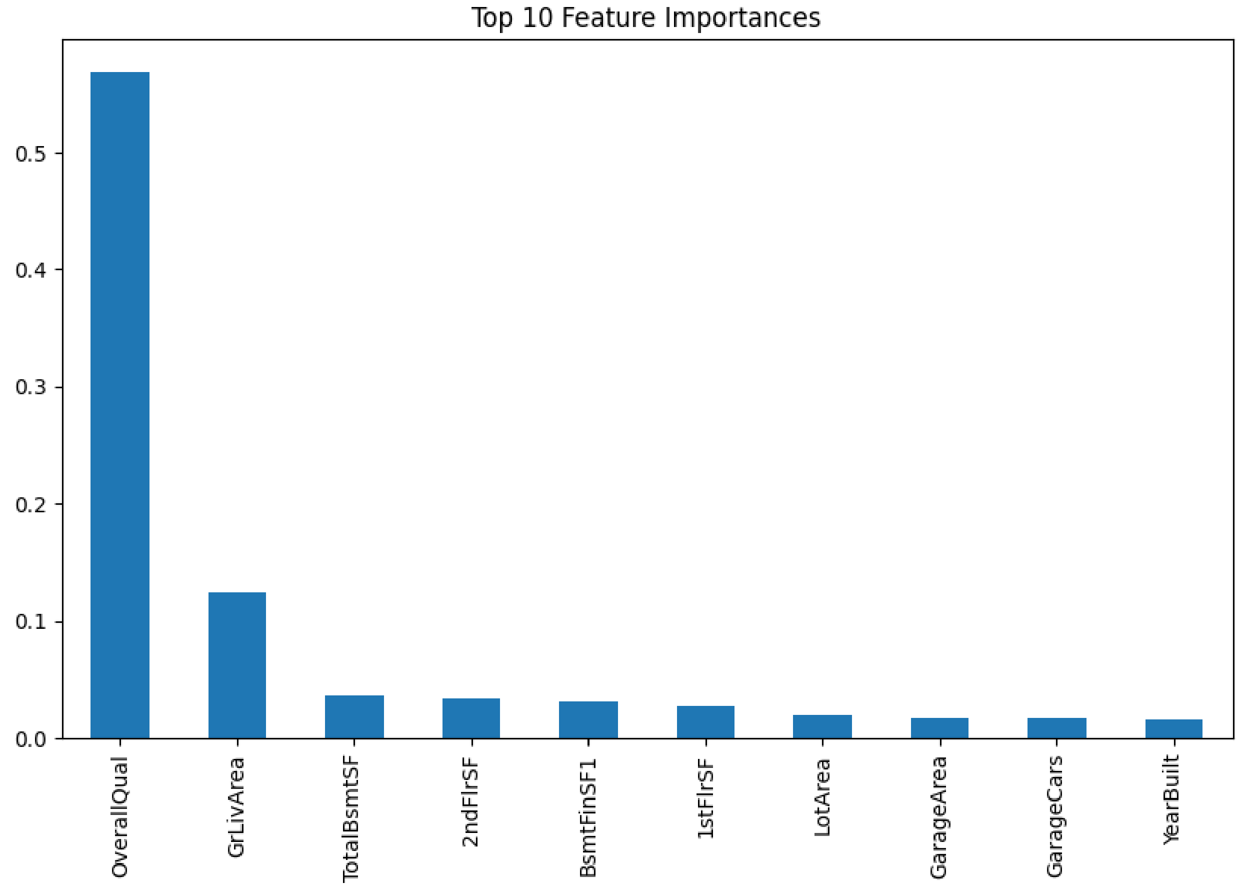
Therefore, I plan to first use a random forest model to predict the sales price.

Result

The RMSE for the random forest model with default parameters is 28833.398297058924.

And then, I use random search with cross-validation to figure out a better model. After fitting 100 models, the RMSE of the best model is 28554.91886991951.

The top 10 important features in the model show in the graph below:



Discussion

Overall Quality has the highest importance, which indicates that the overall material and finish of the house are the most predictive of the sale price. This means that a buyer often value the general condition and quality of the house.

The improvement in RMSE from the default model to the best model obtained by random search is relatively small. This suggests that while random search may have found better hyperparameter configurations, there may be limits to how much the model can be improved given the current characteristics.