Data Analysis Report on House Prices Prediction using TFDF

Mohammad Hossein Basouli
Shahid Beheshti University of Tehran

June 6, 2025

Abstract

This report presents an analysis of a House Price Prediction work, done for a Kaggle competetion, in which we try to predict sale price of a house given many features. Also we would try to extract some of the most important features that have significant determinationation on the house sale price.

I Data

We have used the house **Ames Housing dataset** which was compiled by **Dean De Cock** [1]. The data is represented as follows: each row is a record of a house (a total of 1460 training examples), along with 81 columns (80 features and the label **SalePrice**). The dataset includes both continous data and categorical data. To get a feel of how the data look like, we present Figures 1 and Figure 2, representing how the label **SalePrice** and numerical features are being distributed respectively.

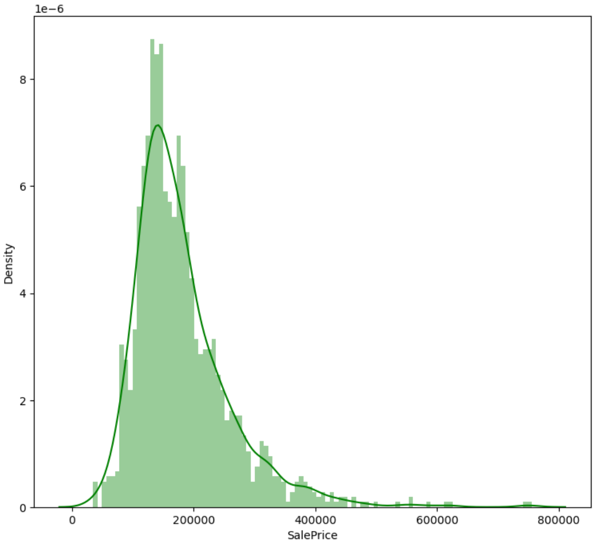

Figure 1: Distribution of SalePrice

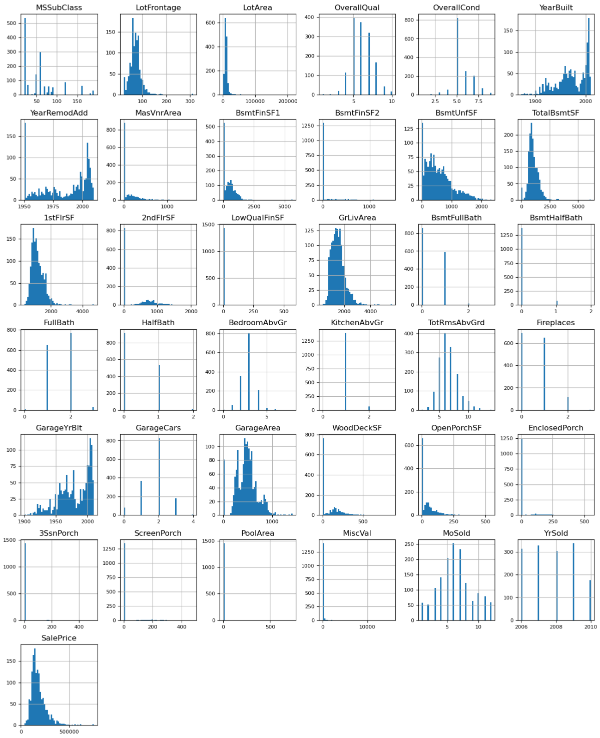

Figure 2: Distribution of numerical features

II Features

As mentioned earlier, the data includes both continous and categorical features. Some of the continous features are: lot area in squared feet, lot frontage, year sold and misc value. In addition, categorical features include street that the house is nearby to, alley that it's located in and shape of the lot.

III Methodology: Random Forests

III.I Strategy

We have used Random Forests in order to learn insights from the data and predict the final **SalePrice** since, they work very well in situations where we have a mix of numeric, categorical and missing features and there is no need for any preprocessing. Firstly we split the dataset into two pieces; train dataset and validation dataset, each having a size of 30% and 70% of the original dataset respectively and train our model.

III.II Result

Our resutls of evaluation of the model can be seen in Figure 3 and Table 2 at the top of the following page as well as important features and their measure of importance based on number of times that they have been selected as the root in the trees in Figure 4 shown at the bottom of the following page.

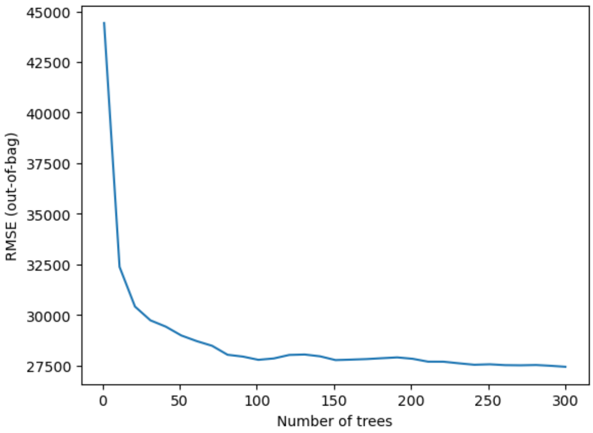

Figure 3: Evaluation on the Out of bag (OOB) data

Table 1: RMSE for Out-of-Bag (OOB) and Testing Data

	OOB data RMSE	**Testing data RMSE**
Random Forests	27438.6758	33064.0629

Note: All results are based on dataset size of 470

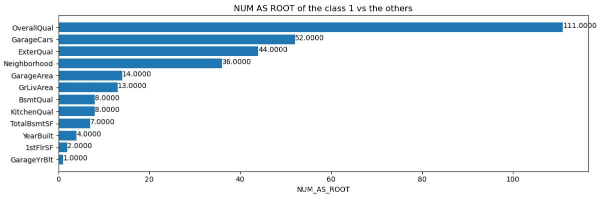

Figure 4: Number of times as root for each of the important features

III.III Analysis

Looking at Figure 1, we can see that our model makes a pretty good prediction based on observable statistics about **SalePrice**. Also, we can see that four of the features seem to be really important in the prediction of the **SalePrice**: **Overall Quality, Number of Cars that could be put in the Garage, Exteriror Quality of the house and Neighborhood**

IV Conclusion

We can conclude that core, well-known components of a house (such as those mentioned in the **Analysis** section) play a crucial role in determining the final **SalePrice** of the house.

References

[1] D. De Cock, “Ames housing dataset.” `https://www.kaggle.com/datasets/prevek18/ames-housing-dataset`, 2023. Accessed: 2025-06-06.