PROJECT REPORT

Pricing Ground Inc.

Abstract

The goal of this project is to investigate how house pricing fluctuates depending on different factors as well as predict the house price given those factors. Some of these factors can be the different physical characteristics of the house, the year it was built or its location. The relationship between some of these factors and the sales price is shown through graphs. It is interesting to see how much some components can influence the house price more than others, and some you didn't expect them to. We will then be using Linear Regression, a Machine Learning algorithm. Housing prices are affected by multiple factors so we will be using Multiple linear regression analysis. Our approach will be very similar to those currently used for multiple linear aggression in that we will start by data analysis and continue to refine our results (to obtain accurate weights for the correlation coefficients between different features and the house pricing) with a gradient descent optimizer.

Introduction

Housing prices today have reached an all time high due to external factors other than those that typically define a house's "price tag". This is a great opportunity to study data that showcase how real estate prices outside this present crisis. Do factors that we personally think correlate with the sale price of a house the most actually reflect the results? Most likely not. We was pleasantly surprised to realize that characteristics of a house that we thought were detrimental to a selling price, come secondary by popular demand.

Related Work

Most of the published work related to Housing Pricing Prediction uses a similar approach in presenting their algorithm to ours. We started with an extensive data analysis in order to get a better understanding of the data we are working with and be able to recognize data that needed processing or conversion. Then this data will be used to run the training model and generate the results. However, a big difference between the other publications and this report is the use of certain libraries. Our solution using JAX, a Python library designed for high-performance LM research.

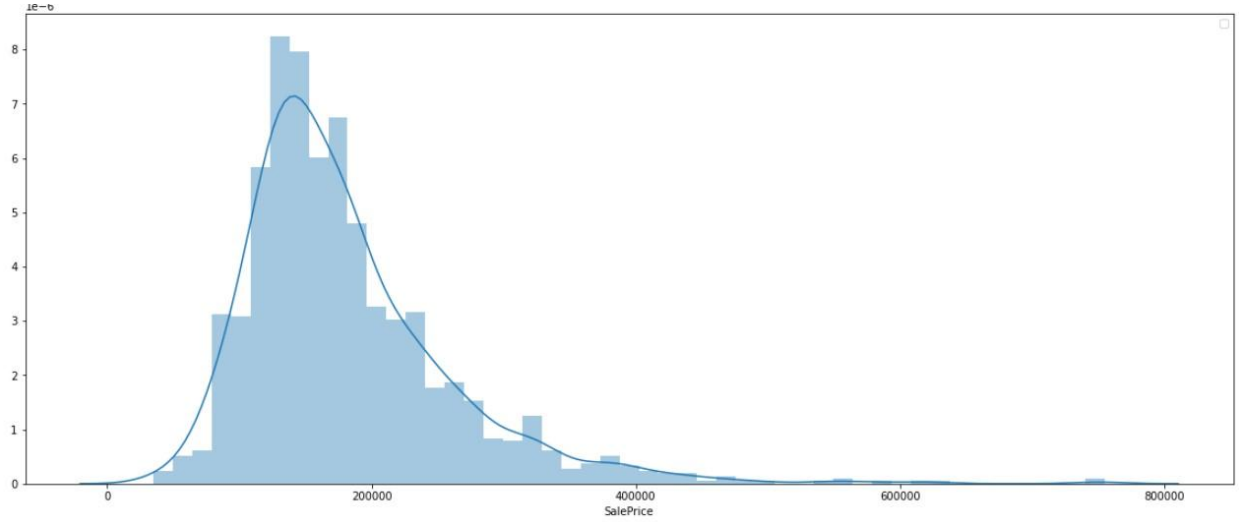
Data

The data is taken from the Kaggle competition called “House Prices - Advanced Regression Techniques”. There are 1460 rows which are the training data points and 80 columns which are the different features.

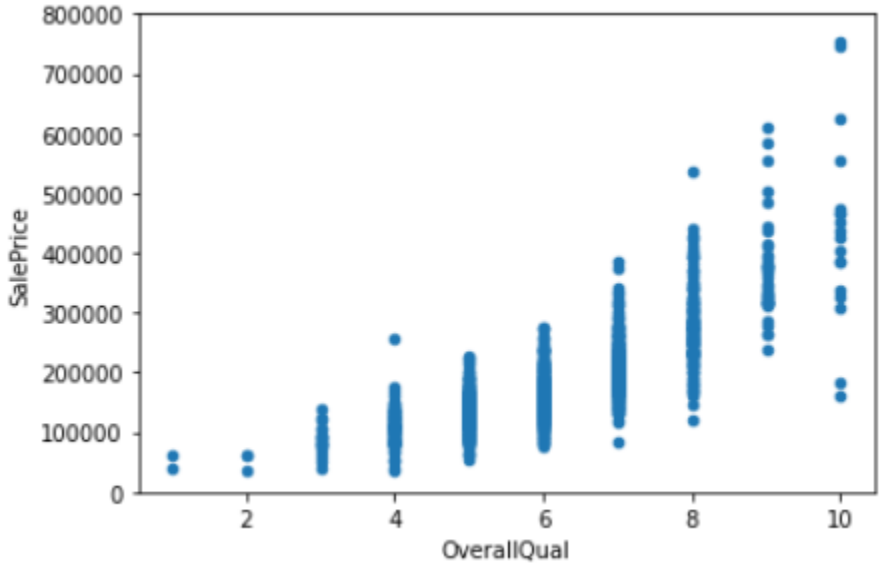
The analysis of the SalePrice data points gave us some interesting results. The mean sale price was calculated at $180,921, the maximum was priced at $755,000 and the minimum was found to be $34,900. As we processed the data we noticed that there were many missing values , NA, which had to be dropped or replaced. The way this was done is described in the Methods section and the reason we decided to drop these data points is mostly null and dropping them will not affect the training model.

Methods

The first step we took was familiarizing ourselves with the data that we were given from the Kaggle competition. The house sales prices were plotted in a graph that followed Gaussian distribution. The graph also shows that most house sale prices are placed between $150k-$250k, which was expected as the mean value was found around $180k.



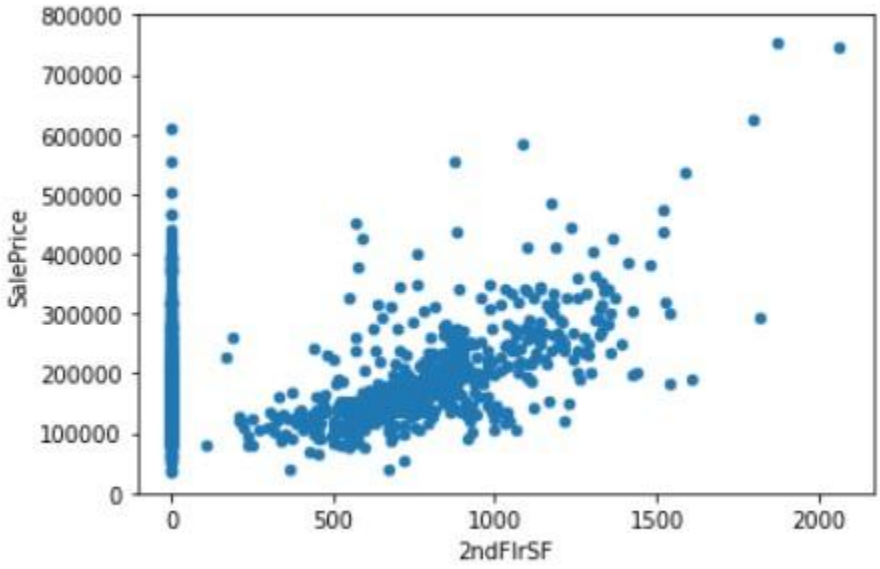
Next we decided to analyze our numerical data and find the 10 most correlated values with SalePrice. Upon results the most influential feature was found to be ‘OverallQual’ with a correlation coefficient 0.79. The OverallQual, which is the overall quality of the house, was given as a rating from 1-10 for each house and the higher the rating the higher the house sale price.



Second most correlated numerical feature with the house sale price was found to be the size of the living room area, followed by the amount of cars that fit in the garage, the garage

area, the total basement area, the first floor square footage, the size of the full bath, the total rooms of the house, the year it was built and the year it was remodeled.

Next we found the least correlated features with the sale price of the house. The three features that were found to have the smallest correlation coefficient, 0.32, were the OpemPorchSF which is the square footage of of the open porch, the 2ndFlrSF which is the square footage of the second floor and the WoodDeckSF which is the square footage of of the wood deck.



This graph that shows the correlation between SalePrice and 2ndFlrSF is very interesting, as it shows that having a two floor house doesn't necessarily mean that the house will be more expensive.

Next we decided to process the data and deal with the inconsistencies that we previously noticed. We ran a test to check which are the features that contained the most missing,NA, data. Then we decided to drop the top five numerical features that had the most missing data. These were PoolQC and MiscFeature with over 1400 missing data, Alley, Fence and FireplaceQu. The ‘ID’ and ‘SalePrice’ columns are also being dropped because it

Experiments

We run a few experiments to make sure that the data we are working with is ready to be used in our training model. This is something that most related already published work also did. It is important to clear the data of any NA, values that might confuse the model while training. Housing Price Prediction can be achieved with many different models. Our purpose is to present the one that we believe will be the most effective one. After trying out models like Gradient Boosting Regressor, Random Forest Regressor, Light Gradient Boosting Regressor, we decided that the best model is Multivariable Linear Regression.

Conclusion

Predicting House Prices is not a simple task and it requires a lot of understanding of the data you are working with as well as being able to understand the best possible algorithm for the training model. However at the same time it is a very interesting process. We were able to understand in a small scale what is considered the most influential characteristic of the house, which was the living room area and one of the features that correlated the least with the price was the second floor square footage which is something that surprised us.

References

Data - https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Google Collab Viewer - https://colab.research.google.com/drive/1BFPnm-NvxO09dh5Xp011nAZe5Qha3ggy

Sources

https://towardsdatascience.com/predicting-house-prices-with-linear-regression-machine-learning-from-scratch-part-ii-47a0238aeac1

https://www.kaggle.com/code/ahmedmsoliman/house-prices

https://www.kaggle.com/code/ashydv/housing-price-prediction-linear-regression/notebook

https://www.kaggle.com/code/adibouayjan/house-price-step-by-step-modeling

https://www.kaggle.com/code/masatoshikato/houseprice-my-solution/notebook