Prediction Of House Price Using Regression

Mayank Goyal
Machine Learning Intern
www.ai-techsystems.com
Jaipur, India
mayankgoyal1993@gmail.com

Abstract-

House prices increase every year, so there is a need for a system to predict house prices in the future. The goal of this project is to create a regression model that can accurately estimate the price of the house given the features. There are many factors that influence the potential price of a house, making it more complicated for an individual to decide how much a house is worth on their own without external help. Houses are long term investments, it is imperative that people make their decisions.

Keywords-

Machine Learning, Regression

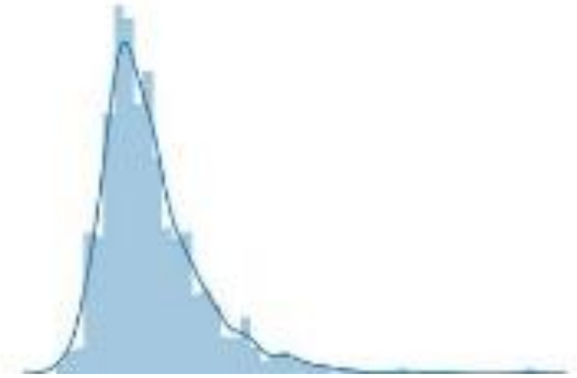
I. INTRODUCTION

It becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy. Increase of demand of houses day by day. Accurate prediction of house prices has been always a fascination for the buyers, sellers and for the bankers also. To forecast house price one person usually tries to locate similar properties at his or her neighbourhood and based on collected data that person will try to predict the house price.

II. METHODOLOGY

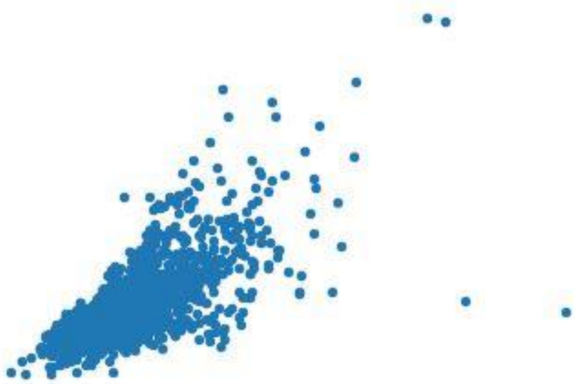
A. DATASET

The dataset used in this project is an open-source dataset taken from Kaggle.com. It consists of 3000 records that have the possibility of affecting the property prices. The house prices data set has 81 features and the objective is to predict SalePrice. Some of the parameters are Area in square meters, Overall quality which rates the overall condition and finishing of the house, Location, Year in which house was built, Numbers of Bedrooms and bathrooms, Garage area and number of cars that can fit in garage, swimming pool area, selling year of the house and Price at which house is sold. The SalePrice is the label which we must predict through regression techniques. Some parameters had numerical values while some had categorical values. We converted categorical columns to numerical columns using pandas get dummies



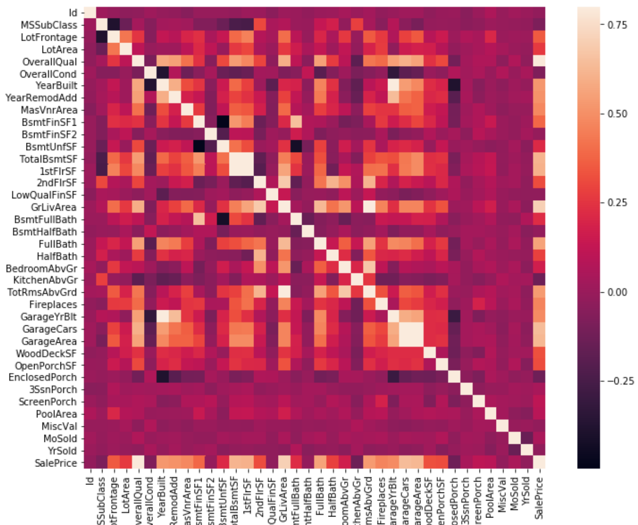
B. DATA ANALYSIS

In this phase our main aim is to analyse the data and have a better understanding of the features involved in our data. It might be possible that some are left behind but I will be focusing on the features that have the highest dependency towards SalePrice of the house. I performed some bivariate analysis on the data to get a better overview of the data and to find outliers in our dataset. I applied the XGBoost machine learning technique so that the data can be fastly boosted. As for feature engineering, I replaced missing data with the most frequent value/category in each column and encode the data so that the data can be used to train the model. And made an instance of gradient boosting regressor and fitted it with our data. For best results and more optimized Model we changed the parameters of gradient boosting regressor when training the data.



C. DATA PREPROCESSING

We must pre-process the data so that the data can be encoded and proceed for training phase. Moreover, some variables are strongly correlated with each other means the encoding of those data is not easy. Subsequently, we need to take them into consideration when selecting and preparing the features to use in our modelling. For example, there is a strong correlation between Yearbuilt and GarageYrBlt which means that most Garages are built at the time when the houses were built. Therefore, we can consider that Yearbuilt and GarageYrBlt as the same variable. On the other hand, scatter or some other charts can show the nature of the correlation whether it is linear or has another shape**.** In order to convert the data in encoded form, we need to perform different operations and changes in data to encode it in the numeric form. Besides, there were some features that had values of N/A; we replaced them with the mean/median/0 of their columns so that they don’t influence the distribution. We also need to assume other values to encode the data at the training phase. Uniformity in dataset is also maintained at the encoded phase.



D. MODEL

The model is separately train and test with train.csv and test.csv dataset. House price prediction is predicted and checked with sample_sumission.csv dataset. I present the evaluation of different scikit-learn modelling algorithms. We aim to measure the performance of each model and compare it with the other models. I fitted our model with linear regression algorithm on train data and predicted house prices on the test data. The accuracy of the model is 96.8%. In order to further improve our models, I also performed gradient boosting.

III. CONCLUSION

Encoding of dataset have been proven to be the important part to train the model, for instance addressing the non- linearity problem with log transformation improved the performance dramatically. Moreover, removing the outliers also yield better results. Data pre-processing is also the important part. One way of improving our results is giving the dependency of house prices on location, because, as we know, location is very important factor in most housing prices. If the house is located in the good locality which is quite and clean then the price of house should be high in that locality. Repeatedly, feature engineering can also increase the accuracy of the model.

IV. REFERENCE

[1] www.pandas.pydata.org
[2] www.geeksforgeels.com
[3] encyclopedia
[4] www.stackoverflow.com