House price prediction using linear regression

Shalini Chaurasia

Machine learning intern
AI tech systems
shalini.s.chaurasia@gmail.com

Kanpur,India

www.ai-techsystems.com

Abstract - House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. There are three factors that influence the price of a house which include physical conditions, concept and location.The goal of this project is to create a regression model that are able to 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. This can lead to people making poorly informed decisions about whether to buy or sell their houses and which prices are reasonable. Because houses are long term investments, it is imperative that people make their decisions with the most accurate information possible.

Keywords - exploratory data analysis, data visualisation, handling missing data, co relation matrix, model training, linear regression, gradient boosting.

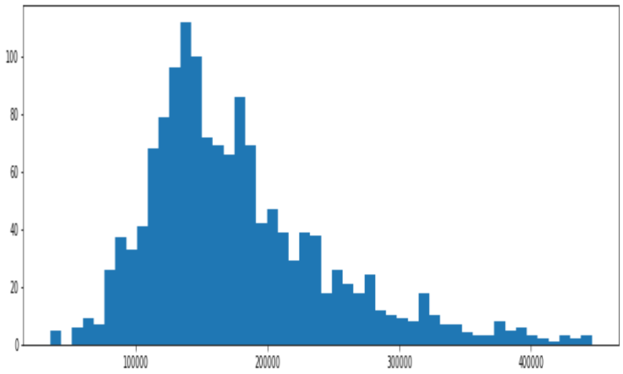
I. INTRODUCTION

Development of civilization is the foundation of prediction of the house prices. 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. Many researchers have already worked to unravel the mysteries of the

We all know that house price is a number from some defined assortment, so obviously prediction of prices of houses is a regression task. 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. All these indicate that house price prediction is an emerging research area of regression which requires the knowledge of machine learning. This has motivated to work in this domain.The goal of this project is to create a linear regression model that is able to accurately estimate the price of the house given the features.

II. METHODOLOGY

A. Dataset

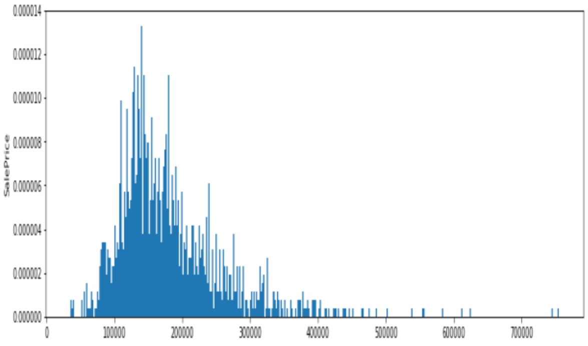


The dataset used in this project is an open-source dataset taken from Kaggle.com. It consists of 3000 records that have the possibility of affecting the property prices. The house prices data set has 81 features and the objective is to predict SalePrice. Some of the parameters are Area in square meters, 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 have to 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 function because our model can only train on numerical columns.

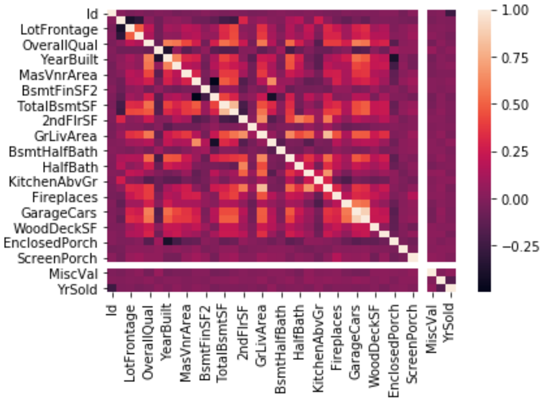
B. Exploratory Data Analysis

In this phase our main aim is to 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 correlation towards SalePrice..I performed some bivariate analysis on the data to get a better overview of the data and to find outliers in our dataset. Outliers can occur due to some kind of errors while collecting the data and need to be removed so that it don't affect the performance of our model.

I applied the XGBoost machine learning technique and to the data. As for feature engineering, I replaced missing data with the most frequent value/category in each column, removed outliers (any prices above 450k) and did one-hot-encoding on the categorical values.



C. Data Preprocessing



We can notice that some variables are strongly correlated with SalePrice. Specifically, these six features : OverallQual, GrlivArea, TotalBsmtSF, 1stFLrSF, GarageCars, and GrageArea. Moreover, some variables are strongly correlated with each other which means that we might have a multicollinearity. 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 in the same time with the construction of the houses.Therefore,we can consider that Year built and GarageYrBlt as the same variable. The correlation matrix shows only the value of the correlation but it doesn't reveal the nature of the correlation. 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 make this data with different format usable for our algorithms, categorical data was converted into separated indicator data, which expands the number of features in this dataset. 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.

D. Model

I am using cross-validation, the scaling has to be done independently for the training and the testing sets. 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. In order to further improve our models, I also performed gradient boosting regressor.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.

n_estimators = 1000, max_depth
= 3, min_samples_split = 2,
learning_rate = 0.05,
base_score=0.5, booster=gbtree

III. CONCLUSION

Data preprocessing have been proven to be a crucial part of our work, for instance addressing the non-linearity problem with log transformation improved the performance dramatically. Moreover, removing the outliers also yield better results. Encoding the features according to their type: nominal, ordinal, and numerical is also critical to our work.

One way of improving our results is creating an ordinal version of the location, because, as we know, location is quite important factor in most housing prices. We can also improve our model doing more feature engineering,

IV. REFERENCE

[1]D. X. Zhu and K. L. Wei, —The Land Prices and Housing Prices ——
Empirical Research Based on Panel Data of 11 Provinces and
Municipalities in Eastern China,‖ Int. Conf. Manag. Sci. Eng., no. 2009,
pp. 2118–2123, 2013

[2] De Cook, Dean. “Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project.” Journal of Statistics Education, vol. 19, no. 3, 2011.

[3]F. S. Gharehchopogh, T. H. Bonab, and S. R. Khaze, —A Linear
Regression Approach to Prediction of Stock Market Trading Volume: A
Case Study,‖ Int. J. Manag. Value Supply Chain., vol. 4, no. 3, pp. 25–
31, 2013.