Modeling Prices of Houses in the Real Estate Industry

Introduction

The goal of this project is to train a model that can predict house prices accurately with minimal errors for house buyers and house sellers. The value of a house is usually determined by several factors which include location, square footage, number of bedrooms, number of baths and so on. Due to the recent increase in the price of houses, many seekers of houses are now interested in knowing the very important aspects that give a house its value. In this project, we are interested in modeling the key variables that determine the value of a house and come out with a predictive model that will help home seeker to determine in advance the value of a house.

Problem statement

House buyers in the real estate industry always longs to find a reasonable price for the property they wish to buy. Many buyers are also completely at sea about what factors determine the cost of a given house. This has caused many to believing that prices of houses in recent times are over-priced. Sellers of houses on the other hand sometimes find it extremely difficult to get a fair price for their property. Many sellers do not even know what factors are to be considered before pricing their property. Since the aim of both buyers and sellers in the real estate industry is to get a fair price for the property they are buying and selling respectively, the goal of this project is to provide a predictive model that can adequately determine the price of a house with minimum margin of error. This will help sellers and buyers to know the fair price of a particular house in advance. This will also help to eliminate the idea of bargaining which sometimes leads to cheating on the other party.

Methodology

In this project, we are going to train a machine learning model that can predict the price of a house using real estate data in the following link: https://www.kaggle.com/amitabhajoy/bengaluru-house-price-data/activity. The data is made up of 13320 observations or houses sold in India. The response variable in this study is the price of a house. Since the price of a house is a quantitative measured, this is a regression problem and we will train a regression model to predict the prices of homes. The explanatory variable considered in this study are; area type, house availability, house location, house size, society, total square feet, number of bathrooms, and balcony. Variables such as area_type, society, balcony, and availability are droped from the study since they do not contribute much in determining the price of a house.

Three different models will be fitted to the given data. The best model will be selected based on higher predicted accruacy among the three models. The two models considered in this case are linear regression model, the lasso regression model and the decision trees regression model.

1. Linear Regression

Linear regression is a very simple model that models the linear effects of covariates on the response variable. It generally assumes that the relationship between the response variable and the set of predictors space is linear. The linear regression model is defined as: $y = B0 + B1x1 + B2x2 +...+ Bp*xp$, where B0, B1, ..., Bp are the regression parameters, y is the response variable and x1, ..., xp are the set of predictors space. In this project, y is the price of house, and x will represent the set of predictors spaces.

2. Lasso Regression

The lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model. Lasso was originally formulated for linear regression models. This simple case reveals a substantial amount about the estimator. These include its relationship to ridge regression and best subset selection and the connections between lasso coefficient estimates and so-called soft thresholding. It also reveals that (like standard linear regression) the coefficient estimates do not need to be unique if covariates are collinear. reference: https://en.wikipedia.org/wiki/Lasso_(statistics)

3. Decision Trees

A decision tree is a flowchart-like structure which is made up of internal nodes and terminal nodels. The termonal nodes are also called the leaf nodes or decision nodes. For regression problems, the final decision node is the average of all observations in the node, whiles for classification problems we choose the class with majority. We are interested in seeing the performances of decision trees in this data because decision can capture both linear and non-linear covariates on the response variable.