Machine Learning based Predicting House Prices using Regression Techniques

Tarek Ali Berbesh
ID: 202200639
Faculty of Graduate Studies for Statistical Research
Cairo University.

***Abstract*——— Predictive models for determining the sale price of houses in cities is still remaining as more challenging and tricky task. The sale price of properties in cities depends on a number of interdependent factors. Key factors that might affect the price include area of the property, location of the property and its amenities. In this research work, an analytical study has been carried out by considering the data set that remains open to the public by illustrating the available housing properties . The data set has eighty one features. In this study, an attempt has been made to construct a predictive model for evaluating the price based on the factors that affect the price. Modelling explorations apply some regression techniques such as multiple linear regression (Least Squares), Lasso and Ridge regression models, support vector regression, and boosting algorithms such as Extreme Gradient Boost Regression (XG Boost). Such models are used to build a predictive model, and to pick the best performing model by performing a comparative analysis on the predictive errors obtained between these models. Here, the attempt is to construct a predictive model for evaluating the price based on factors that affects the price.**

1. Introduction

Modeling uses machine learning algorithms, where machine learns from the data and uses them to predict a new data. The most frequently used model for predictive analysis is regression. As we know, the proposed model for accurately predicting future outcomes has applications in economics, business, banking sector, healthcare industry, e-commerce, entertainment, sports etc. One such method used to forecast house prices are based on multiple factors . In metropolitan cities , the prospective home buyer considers several factors such as location, size of the land, proximity to parks, schools, hospitals, power generation facilities, and most

importantly the house price. Multiple linear regression is one of the statistical techniques for assessing the relationship between the (dependent) target variable and several independent variables. Regression techniques are widely used to build a model based on several factors to predict price. In this study, we have made an attempt to build house price prediction regression model for data set . We have considered five prediction models, they are ordinary least squares model, Lasso and Ridge regression models, SVR model, and XGBoost regression model. A comparative study was carried out with evaluation metrics as well. Once we get a good fit, we can use the model to forecast monetary value of that particular housing property . The paper is divided into the following sections: Section 2 addresses previous related work, Section 3 explains the description of the data set used, pre-processing of data and exploratory analysis of data before regression model is built. Section 4 Problem Formulation Section 5 presents a summary of the regression models developed in the comparison study and the evaluation metrics is used. Section 5 sums up the models and conclusion and results.

2. Related Work

Pow, Nissan, Emil Janulewicz, and L. Liu [1] used four regression techniques namely Linear Regression, Support Vector Machine, K-Nearest Neighbors (KNN) and Random Forest Regression and an ensemble approach by combining KNN and Random Forest Technique for predicting the property's price value. The ensemble approach predicted the prices with least error of 0.0985 and applying PCA didn't improve the prediction error. Several studies have also focused on the collection of features and extraction procedures. Wu, Jiao Yang [2] has compared various feature selection and feature extraction algorithms combined with Support Vector Regression. Some researchers have developed neural network models to predict house prices. Limsombunchai, compared hedonic pricing structure with artificial neural network model to predict the house prices .The R-Squared value obtained for Neural Network model was greater when compared to hedonic model and the RMSE value of Neural Network model was relatively lower. Hence they concluded that Artificial Neural Network performs better when compared with Hedonic model. Cebula applies the hedonic price model to predict housing prices in the City [3]of Savannah, Georgia. The log price of houses has been shown to be positively and substantially associated with the number of bathrooms, bedrooms, fireplaces, garage spaces, stories and the house's total square feet [4]. Jirong, Mingcang and Liuguangyan apply support vector machine

(SVM) regression to predict China's housing prices from 1993 to 2002. They have applied the genetic algorithm to tune the hyper-parameters in the SVM regression model. The error scores obtained for the SVM regression model was less than 4Tay and Ho compared the pricing prediction between regression analysis and artificial neural network in predicting apartment's prices. It was concluded that that the neural network model performs better than regression analysis model with a mean absolute error of 3.9

3. Data

The two data sets-train set and test data considered in the paper. It consists of features that describe house-property . There are 81 features in both the data sets. The features can be explained as follows:

3.1. Description

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass : The building class
- MSZoning : The general zoning classification
- LotFrontage : Linear feet of street connected to property
- LotArea : Lot size in square feet
- Street : Type of road access
- Alley : Type of alley access
- LotShape : General shape of property
- LandContour : Flatness of the property
- Utilities : Type of utilities available
- LotConfig : Lot configuration
- LandSlope : Slope of property
- Neighborhood : Physical locations within Ames city limits
- Condition1 : Proximity to main road or railroad
- Condition2 : Proximity to main road or railroad (if a second is present)
- BldgType : Type of dwelling
- HouseStyle : Style of dwelling
- OverallQual : Overall material and finish quality
- OverallCond : Overall condition rating
- YearBuilt : Original construction date
- YearRemodAdd : Remodel date
- RoofStyle : Type of roof
- RoofMatl : Roof material
- Exterior1st : Exterior covering on house
- Exterior2nd : Exterior covering on house (if more than one material)
- MasVnrType : Masonry veneer type
- MasVnrArea : Masonry veneer area in square feet
- ExterQual : Exterior material quality
- ExterCond : Present condition of the material on the exterior
- Foundation : Type of foundation
- BsmtQual : Height of the basement
- BsmtCond : General condition of the basement
- BsmtExposure : Walkout or garden level basement walls
- BsmtFinType1 : Quality of basement finished area
- BsmtFinSF1 : Type 1 finished square feet
- BsmtFinType2 : Quality of second finished area (if present)
- BsmtFinSF2 : Type 2 finished square feet
- BsmtUnfSF : Unfinished square feet of basement area
- TotalBsmtSF : Total square feet of basement area
- Heating : Type of heating
- HeatingQC : Heating quality and condition
- CentralAir : Central air conditioning
- Electrical : Electrical system
- 1stFlrSF : First Floor square feet
- 2ndFlrSF : Second floor square feet
- LowQualFinSF : Low quality finished square feet (all floors)
- GrLivArea : Above grade (ground) living area square feet
- BsmtFullBath : Basement full bathrooms
- BsmtHalfBath : Basement half bathrooms
- FullBath : Full bathrooms above grade
- HalfBath : Half baths above grade
- Bedroom : Number of bedrooms above basement level
- Kitchen : Number of kitchens
- KitchenQual : Kitchen quality
- TotRmsAbvGrd : Total rooms above grade (does not include bathrooms)
- Functional : Home functionality rating
- Fireplaces : Number of fireplaces
- FireplaceQu : Fireplace quality
- GarageType : Garage location
- GarageYrBlt : Year garage was built
- GarageFinish : Interior finish of the garage
- GarageCars : Size of garage in car capacity
- GarageArea : Size of garage in square feet
- GarageQual : Garage quality
- GarageCond : Garage condition
- PavedDrive : Paved driveway
- WoodDeckSF : Wood deck area in square feet
- OpenPorchSF : Open porch area in square feet
- EnclosedPorch : Enclosed porch area in square feet
- 3SsnPorch : Three season porch area in square feet
- ScreenPorch : Screen porch area in square feet
- PoolArea : Pool area in square feet
- PoolQC : Pool quality
- Fence : Fence quality
- MiscFeature : Miscellaneous feature not covered in other categories
- MiscVal : Value of miscellaneous feature
- MoSold : Month Sold
- YrSold : Year Sold
- SaleType : Type of sale
- SaleCondition : Condition of sale

3.2. Data Understanding

The purpose is to create a model that can estimate housing prices. We divide the set of data into functions and target variable. In this section, we will try to understand overview of original data set, with its original features and then we will make an exploratory analysis of the data set and attempt to get useful observations. The data set consists of 1460 records with 81 explanatory variables.. While building regression models we are often required to convert the categorical i.e. text features to its numeric representation. The two most common ways to do this is to use label encoder or one hot encoder. Label encoding in python can be achieved by using sklearn library.

3.3. Data Prepossessing

The general steps in data pre-processing are:

- Converting categorical features into numerical variables in order to fit linear regression model.
- Imputing null records with appropriate values.
- Scaling of data
- Split into train –test sets.
- Visualisation.

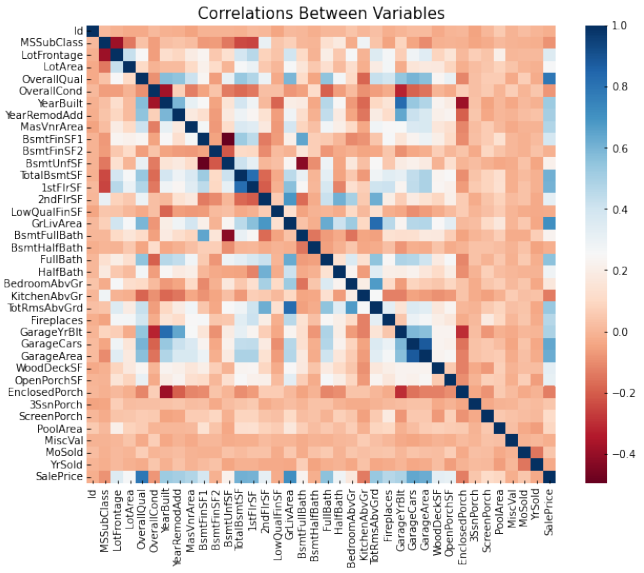

Figure 1: Research Framework

4. Problem Formulation

Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. In this project, house prices will be predicted given explanatory variables that cover many aspects of residential houses. 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.

5. Model

Linear regression is one of the most well- known algorithms in statistics and machine learning. The objective of a linear regression model is to find a relationship between one or more features (independent/explanatory/predictor variables) and a continuous target variable (dependent/response) variable. If there is only one feature, the model is simple linear regression and if there are multiple features, the model is multiple linear regression [5].

- Basic Linear Model: The formulation for multiple regression model is

$$Y_i = \alpha_0 + \alpha_1 X_i + \epsilon_i \tag{1}$$

  assumptions in the model are:

  - The error terms are normally distributed.
  - The error terms have constant variance.
  - The model carries out a linear relationship between the target variable and the functions.

  Here the multiple regression models are developed by the least square approach (Ordinary Least Squares / OLS). The accuracy of the designed model is difficult to measure without evaluating its output on both train and test data sets.
- Ridge Regression
  Ridge Regression Ridge regression model is a regularization model, where an extra variable (tuning parameter) is added and optimized to address the effect of multiple variables in linear regression which is usually referred as noise in statistical context. In mathematical form, the model can be expressed as

$$y = xb + e \tag{2}$$

  Here, y is the dependent variable x refers to features in matrix form and b refers to regression coefficients and e represents residuals.
- Lasso Regression LASSO means least absolute shrinkage, and the selection operator is an LR technique that also regularizes functionality. It is identical to ridge regression, except that it varies in the values of regularisation. The absolute values of the sum of regression coefficients are taken into consideration. It even sets the coefficients to zero so it completely reduces the errors. So selection of features are resulted by lasso regression.
- SVR (Support Vector Regression) whereas in SVR we try to fit he error within a certain threshold. It is a regression algorithm and uses a similar Support Vector Machines (SVM) methodology for regression Analysis
- XGBoost Regression Model XGBoost stands for extreme gradient boosting which is most efficient technique for either regression or classification problem. It is decision tree based algorithm that make use of gradient boosting framework. It provides the features that greatly have impact on performance of

model. This technique helps in developing a model that have less variance and more stability.

6. Results

[H]

MAE	MSE	RMSE	R2 Score	RMSE (Cross-Validation)
XGBRegressor	1.770609e+04	7.596317e+08	2.756142e+04	9.009649e-01
SVR	1.784316e+04	1.132136e+09	3.364723e+04	8.524005e-01
RandomForestRegressor	1.811501e+04	9.829418e+08	3.135190e+04	8.718514e-01
Ridge	2.343550e+04	1.404264e+09	3.747351e+04	8.169225e-01
Lasso	2.356046e+04	1.414338e+09	3.760768e+04	8.156092e-01
LinearRegression	2.356789e+04	1.414931e+09	3.761557e+04	8.155318e-01
Polynomial Regression (degree=2)	1.494080e+15	5.962634e+31	7.721810e+15	-7.773639e+21
ElasticNet	2.379274e+04	1.718446e+09	4.145414e+04	7.759618e-01

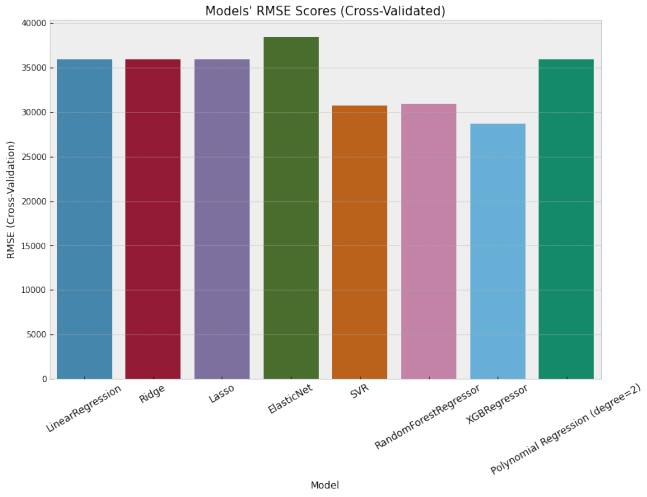

Figure 2: Results

References

[1] N. Pow, E. Janulewicz, and L. Liu, "Applied machine learning project 4 prediction of real estate property prices in montréal," *Course project, COMP-598, Fall/2014, McGill University*, 2014.

[2] J. Y. Wu, "Housing price prediction using support vector regression," 2017.

[3] V. Limsombunchao, "House price prediction: hedonic price model vs. artificial neural network," 2004.

[4] R. J. Cebula, "The hedonic pricing model applied to the housing market of the city of savannah and its savannah historic landmark district," *Review of Regional Studies*, vol. 39, no. 1, pp. 9–22, 2009.

[5] J. S. Raj, J. V. Ananthi, *et al.*, "Recurrent neural networks and nonlinear prediction in support vector machines," *Journal of Soft Computing Paradigm (JSCP)*, vol. 1, no. 01, pp. 33–40, 2019.