Machine Learning Coursework Report

Bhanu Prakash Bandi - 001183470 - bb5505q

Abstract

Buying their own house is the dream for many, keeping that in mind, project is done to predict the prices of the house based on the different factors(variables) using various Machine learning algorithms. The goal is to provide the best housing price as a deciding factor based on the people's interest. The set of algorithms that are examined is effective since the data set is quantitative.Hence, most efficient algorithms based on the model performance is used for predicting the best house price.

1. Introduction

"Aim to make our evaluations based on every basic parameter that is considered while determining the price" [2]. The topic that we choose for the project is to predict the house price based on the different characteristics of the house. The source of the data is taken from Immowelt, Immoscout24 and some of it from web scrapping. Data set comprises of quantitative attributes namely 13 independent features and 1 dependent feature(specifically the variable that is to predicted). The characteristics of the residential house structure such as number of rooms, garden area, living area square units etc. are included which is particularly used to analyse the house price based on the people's interest.

1.1. Pre-processing

Initially, the general pre-processing is done where each independent is analysed with dependent variable to understand the data more specifically and also, exploratory data analysis (different plots) is used to figure out how data points is distributed over dependent variable with each independent variable. Here, we are analysing the data and deriving insights which intern helps to carry out further analysis for predicting the price and also it gives an idea as to which may be the deciding factors that can be used for the algorithm. Log transformation, outliers and correlation plot is used to figure out the data distribution as the part of pre-processing.

1.2. To predict the target variable

Once all the necessary analysis is derived then the data set is passed into the different Machine learning regressor models since the data is quantitative. Different Machine learning algorithms namely, Random forest regressor model, Ridge regression, Gradient Boost regressor, KNeighbours regressor are used to check which algorithm provides the best regressor fit for the price prediction. Algorithms are experimenting on hyperparameter tuning to identify at which point(out of n estimators) the model estimates the best fit and gives the highest accuracy to predict the house prices.

The Machine learning models performance is evaluated using r2 score which is determined to select the best algorithm that is used to predict the house price and gives most efficient results(outcomes).

2. Methods

The dataset is quantitative for which general preprocessing is done to analyse each of the independent variable with the target variable. The derived insights from the pre-processing are further used to carry out the analysis. The target variable(Price) is later transformed by dividing the living space square area since the dependent feature is right skewed. Later, the log transformation is performed on living space variable since it is highly associated with the target variable.

Once the data is completely pre-processed, the machine learning algorithms are performed on the dataset to predict the price which is explained below in detail.

2.1. Ridge regression

Ridge regression is the type of regularised linear regression. Ridge regression uses the least-squares criterian but it also adds the penalty for large variations in weight parameters. The addition of a parameters as penalty is called regularization. Regularization helps in preventing overfitting by reducing complexity of the data. Ridge regression model is one of the best algorithm which is used to analyse the data that undergoes multicollinearity.

"In especial, Lasso regression could model cases with a million features. In order to avoid overfitting, Ridge regression performs L2 regularization and Lasso regression performs L1 regularization. In the following part, we investigate the Ridge regression algorithm" [3].

Ridge regression uses the $L2$ regularization and minimizes sum of squares of weight entries. The influence of the regularization term is controlled by the alpha parameter. Inter-

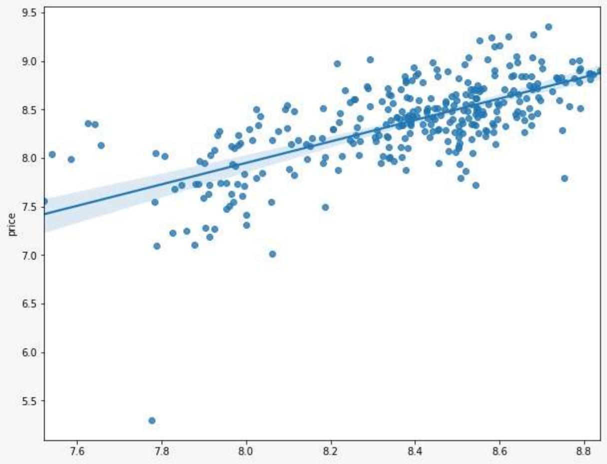

Figure 1: Ridge Regression Regplot
The figure 1 shows the ridge regression plot where most of the data is not lying on the hypothesis line which indicates that the model is underfitting.

cept and Coefficient values are calculated by the regression model where

”intercept” is a parameter which calculates the bias value for the ridge regression model and

”coef” is a parameter which calculates the coefficients of ridge regression model.

Ridge regression algorithm minimizes the least squared error while regularizing the norm of the weights and the equation of the ridge regression model is given below:

$$RSS_{ridge}(w,b) = \sum_{i=1}^{N}(y_i - (wx_i + b))^2 + \alpha \sum_{j=1}^{P} w_j^2 \quad (1)$$

where,

RSS is the sum of squares of residuals,

α is the regularization parameter,

y_i is the true value,

$\alpha \sum_{j=1}^{P} w_j^2$ is the regularization term which penalizes the weights of the linear model.

From the model, we obtain the calculated R^2 score i.e., **0.449** for test data and **0.424** for training data which is the least score to predict the house price.

2.2. Random Forest Regressor

RandomForestRegressor is one of the sofisticated models which works based on multiple decision trees. This algorithm uses the ensemble learning method. Ensemble learning method is one of the most prominently used technique which combines the predictions with multiple algorithms to make more accurate predictions than the single model.

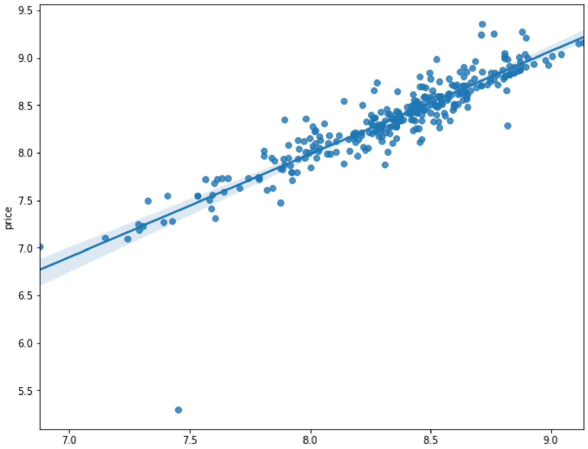

Figure 2: Random Forest Regressor Regplot
The figure 2 graph shows that most of the data points are lying on the hypothesis line which seems to be good.

Random forest regressor adopts ensemble learning method since the multiple decision trees are combined to make more accurate predictions than a single decision tree algorithm. Iterating through different estimators gives us the intuition of number of estimators at which we get the best score.

The two hyperparameters used in this model are n_estimators and max_features.

- The parameter n_estimators is building number of trees before choosing the majority votes or average on predictions.
- The parameter max_features is set to ’sqrt’, this is to select number of features for our model.

The **Random Forest Regressor** algorithm results the highest r^2 score of **8**2.28% after tuning with n_estimators (hyperparameter) at 1400 estimator.

2.3. Gradient Boost Regressor

GradientBoosting is a type of machine learning model. It is based on the idea of the probability calculated for the data GradientBoosting relies on the principle of ensemble technique of decision trees to predict the dependent(target) variable. This algorithm works on the basis of loss function, weak learners and additive model where,

- loss function is used to optimise the model $[L(y, \theta(x))]$
- weak learners(Decision trees are used as weak learners) to make predictions.
- Additive model helps to add weak learners to minimise the loss function.

All the weak learners(points) are putting up together to make the strong fit of the model.this algorithm relies on the gen-

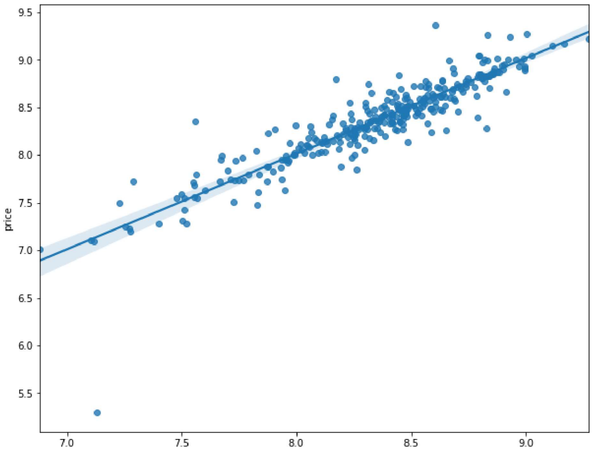

Figure 3: Gradient Boosting Regressor Regplot
The figure 3 shows that most of the data points lies on the hypothesis line which indicates that it is the best fit and learning the data more efficiently.

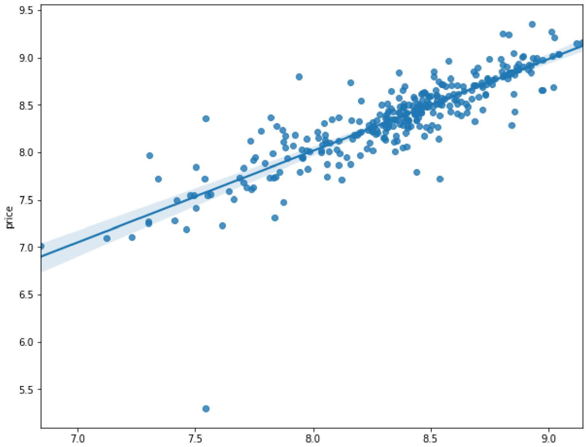

Figure 4: KNeighbors Regressor Regplot
The figure 4 infers that the not all data points lies on the hypothesis line but only some of the data points are near to the objective function.

eral intuition that the best possible next model. Gradient Boosting algorithm can be explained by AdaBoost algorithm. The AdaBoost algorithm trains each observation of the decision tree by assigning the equal weight to it.

Gradient Boost Algorithm uses two hyperparameters namely max_depth and n_estimators.

- max_depth is a hyperparameter which takes an integer input, states that at these many steps the model will converge.
- loss parameter takes 'ls' which is least-squares while calculating the loss.

$$R^2 = 1 - \frac{RSS}{TSS} \tag{2}$$

where,

R^2 is the co-efficient of determination.

RSS is the sum of squares of residuals.

TSS is the total sum of squares.

For the fitted **Gradient Boost regressor** model, At 2100 estimators, we get highest r^2 score of **82.98%**.

Gradient boosting algorithm has high accuracy value when compared to all the other algorithms regarding house price predictions [4].

2.4. KNeighbors Regressor

kNeighbors algorithm adopts the **feature similarity** technique to predict the values of any new data points. A nearest neighbor algorithm needs demands 4 things to be specified:

- A distance metric, typically Euclidian.
- How many 'nearest' neighbors to look at?
- Optional weighting function on the neighbor points.
- Method for aggregating the classes of neighbor points, that is Simple Majority Vote.

KNeighborsRegressor works based on averaging the distances of trained data from the new data point. Here 'K' defines the number of data points to consider for predicting the new value. Based on feature similarity, the new data point is assigned a value on the basis of how closely the points of the training dataset resembles.

For the default parameters of the KNeighbors, the training score is nearly **80%** and test score is nearly **74%**.

Table 1. Hyperparameters for Ridge regression, Random forest, Gradient Boost and KNeighbours Regressor that are used and values given are mentioned in the table

Parameter name	Value
Regularization parameter(alpha)	20.0
n_estimators in RFR	1400
n_estimators in GBR	2100
max_depth in GBR	5
learning rate in GBR	0.1
n_neighbours in KNN	1

3. Experiments

The housing dataset is quantitative and also, the dependent variable(target variable) is distributed over right-skewed(positively skewed) which states that the data points are low relative. Experiments are carried out with respect to dependent variable in contrast with each of the independent

Table 2. The price prediction of housing data is evaluated by R^2 score. The digits that are bold indicates the best R^2 score value.

Model	R^2 score
Ridge Regression	0.449
Random Forest Regressor	0.822
Gradient Boost Regressor	**0.829**
KNeighbours Regressor	0.754

variable to understand and analyse the data in more specific manner which interns helps to get the best fit model whilst fitting different algorithms.

3.1. Experimental settings

The major task is to predict the price most accurately with the given characteristics of the house for which several methods(models) are used. Unlike this, the challenge was to tune the hyperparameters in most effective way.

In Random forest regressor and KNeighbours Regressor, the parameter n estimators is considered as ’i’ to understand at which point the model is estimating the dataset efficiently. In the same way, the hyperparameters such as max depth, alpha value and learning rate are choosed of all the algorithms are shown in Table 1.

3.2. Evaluation criteria

The criteria that is used to evaluate the performance of the model is r2 score, mean square error and mean absolute error. Also, Regplot is used as a metric to plot the data for estimating the regression model.

3.3. Results

The report is presented with four different models that is RidgeRegression, RandomForestRegressor, GradientBoostingRegressor and KNeighborsRegressor.

- RidgeRegression is explaining the 42 percent variance of the generated values for the training data but testing data is explaining nearly 45 percent variance of the model, this model is suffering from underfitting because most of the data is deviated from the hypothesis line.
- RandomForestRegressor went through different number of estimators, that is hyperparameter tuning, the best results are giving out for the 1400 n_estimators parameter. The $r2$ score for the training data is 0.96 and for the testing data it is 0.82.
- Similarly for the GradientBoostingRegressor, running iterations on number of estimators with 1900 the training $r2$_score is 0.97 and 0.83 for the testing data, it means 83 percent of variance can be explained by the data.
- KNeighborsRegressor tested with default parameters such as n_neighbors, the $r2$_score for the training data is 0.79 and $r2$_score for the testing data is 0.73, seems the data is under fitting because the training data failed to explain the maximum variance of the data.

4. Conclusion

Among the Ridge Regression, Random Forest Regressor, Gradient Boosting Regressor and KNeighbors Regressor. ”The Random Forest was found to consistently perform better than the kNN algorithm” [1]. The Random Forest Regressor and Gradient Boosting Regressor are giving best results, these 2 models are nearly the same but Gradient Boosting performing 1 percent more than Random Forest Regressor. Therefore, [4] ”Gradient Boosting Regressor is the best model to work on dataset which fulfills the objective effectively.”

References

[1] Isak Engström and Alan Ihre. Predicting house prices with machine learning methods, 2019.

[2] Visit Limsombunchai. House price prediction: hedonic price model vs. artificial neural network. In *New Zealand agricultural and resource economics society conference*, pages 25–26, 2004.

[3] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, and Rick Siow Mong Goh. A hybrid regression technique for house prices prediction. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 319–323. IEEE, 2017.

[4] CH Raga Madhuri, G Anuradha, and M Vani Pujitha. House price prediction using regression techniques: a comparative study. In *2019 International Conference on Smart Structures and Systems (ICSSS)*, pages 1–5. IEEE, 2019.