Multinomial Regression to Predict House Price

Shubham Kumar
AITS
7th August, 2019
Delhi, India
skagnihotri1@gmail.com

***Abstract*— The objective of this report is to predict the house price using two multinomial regression algorithms. Linear Regression and Support Vector Machine are used to predict the prices. The dataset is taken from Kaggle. Dataset consist of different aspects which affect house price. Hyper tuning and Regularization are used to make model more robust and powerful.**

Keywords—Linear Regression, Support Vector Machine (SVM), Regularization, Encoding, Grid Search

I. Introduction

Regression is the technique to predict the continuous data. It consists of the predicting function known as Hypothesis, which is a function of features (dependent variables). For accuracy to hypothesis, Cost(loss) function which calculates the root mean square distance between the true and predicted values. Gradient Descent is used to minimize the cost function and make hypothesis more powerful. The algorithms used are Linear Regression with l1 and l2 regularization and Support Vector Machines. Grid Search and Cross Val Score are used to tune the hyper parameters and make Bias and Variance small and converge the algorithm to global minima.

II. Machine Learning

Machine learning (ML) is a class of algorithm that allows software applications to become more accurate in forecast outcomes without being precisely programmed. The basic assertion of machine learning is to construct algorithms that accepts input data and use statistical analysis to forecast an output while updating outputs as new data becomes available. The actions involved in machine learning are similar to that of data mining and predictive modeling. Both require searching through data to seek patterns and adjusting program actions accordingly. Many people are accustomed with machine learning from shopping on the internet and being served ads related to their purchase. This happens because recommendation engines use machine learning to customize online ad deliveries in almost real time. Beyond customized marketing, other prevalent machine learning cases include fraud detection, spam filtering, network security threat detection, predictive maintenance and building news feeds. Machine learning algorithms are often classed as supervised or unsupervised. Supervised algorithms require a data scientist or data analyst with machine learning skills to administer both input and desired output, in addition to feedback about the accuracy of predictions during algorithm training. Data scientists determine which variables, or features, the model should analyze and to develop predictions. When training is complete, the algorithm will apply what was learned to new data. Unsupervised algorithms do not require to be trained with desired outcome data. Instead, they use a progressive approach called deep learning to review data and arrive at conclusions. Unsupervised learning algorithms – also known as neural networks– are used for more complex processing tasks than supervised learning systems, including image recognition, speech-to-text and natural language generation.

III. Preprocessing And Multinomial Regression Methods

A. Preprocessing

Preprocessing means cleaning of data i.e. removal of "nan" values, one hot encoding, filling of "nan" values, etc.

"nan" values are if present in bulk in a certain column or in some rows then the rows or columns are just dropped. But if, they are very few we replace them with mean, median or mode values according to the need.

Categorical columns are the columns which only have string value, so to use them they are one hot encoded and dropped and the new generated column are added. It is important to drop one column from every one hot encoded column to save the algorithm from dummy variable trap.

TABLE 1. Column Type

	Number
Categorical Variables	40
Non-Categorical Variable	12

As the preprocessing steps are completed, we train out model on the training set.

There are many regression algorithms but most popular regression algorithms are Linear Regression and SVM. First the model is trained on the training set and then accuracy is checked on the validation set.

After training is done model accuracy was analyzed on the validation data. The cross_val_score of the model were plotted.

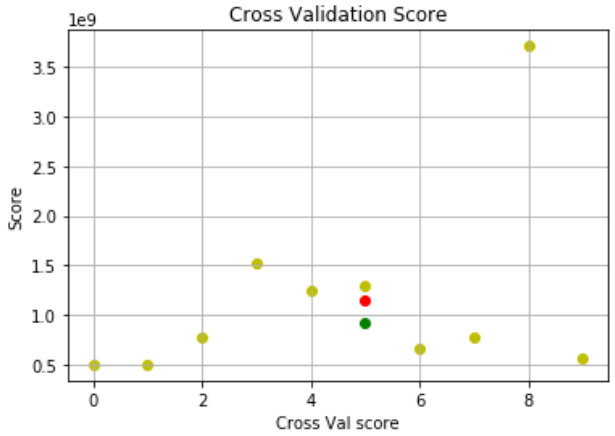

Fig 1. Cross Validation score (R2 score) of Linear Regression with L1 regularization.

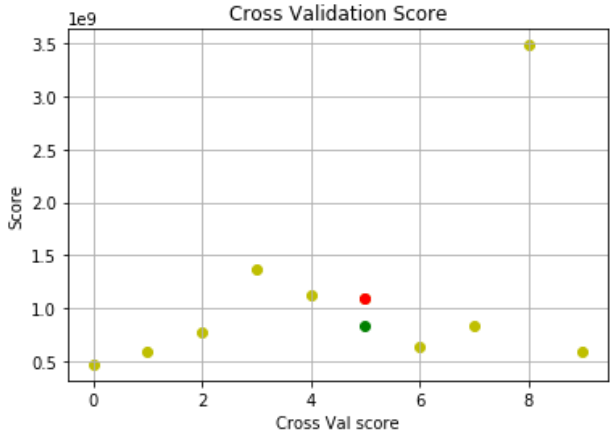

Fig 2. Cross Validation score (R2 score) of Linear Regression with L2 regularization.

And for SVM Grid Search was used to hyper tune the values of parameter "C".

B. *Linear Regression*

Linear Regression is the method if fitting a line or curve on a dataset, which is used to predict the result. It gives continuous prediction. The predicting function is known as hypothesis, which consist of independent variables(features) and dependent variables(target).

Hypothesis equation: -

$$f(x,y,z)=w_1x+w_2y+w_3z$$

Cost Function is used to calculate the error between the predicted and true price. Generally, mean square distance is calculated but sometimes when a large number od outliers are there in data absolute distance is then used.

Cost Function: -

$$MSE=1N\sum_{i=1n}(y_i-(mx_i+b))2$$

Gradient descent is used to minimize the cost function and fit the model to the dataset more accurately.

C. *Support Vector Machine*

SVM or Support Vector Machine is a linear algorithm for classification and regression tasks. It can solve linear (using 'linear' kernel) and nonlinear (using kernels such as 'rbf', 'polynomial' and 'sigmoid') task and work great for many daily life problems. The concept of SVM is easy: The algorithm creates a separation line and support vector lines which separates the data into different classes.

IV. Result And Conclusion

A. *Result*

Linear Regression with l1 and l2 regularization is used and the following graphs between predicted values and true values were plotted.

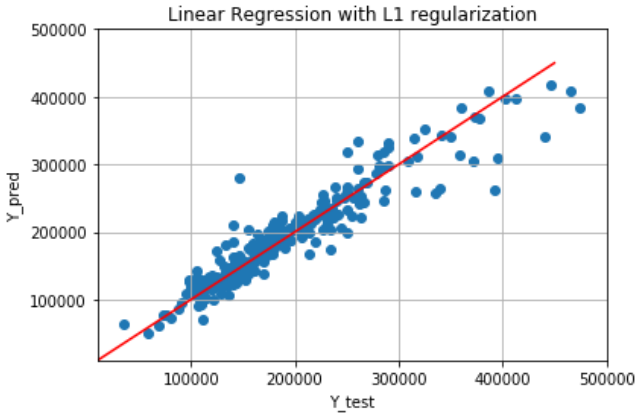

Fig. 1. Linear Regression with absolute mean Regularization.

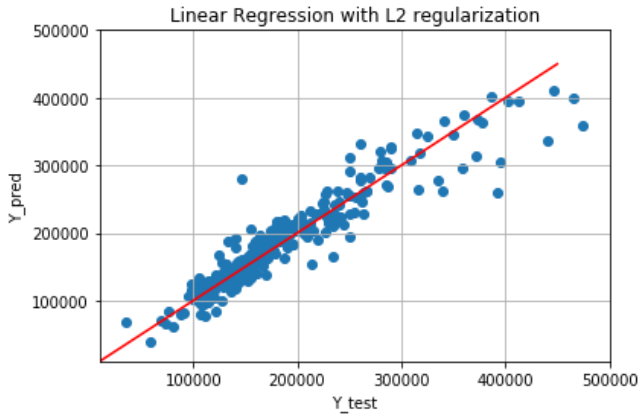

Fig. 2. Linear Regression with mean square Regularizations.

Support Vector Machine with Grid search hyper parameter tuning is used and a graph between predicated and true values is plotted.

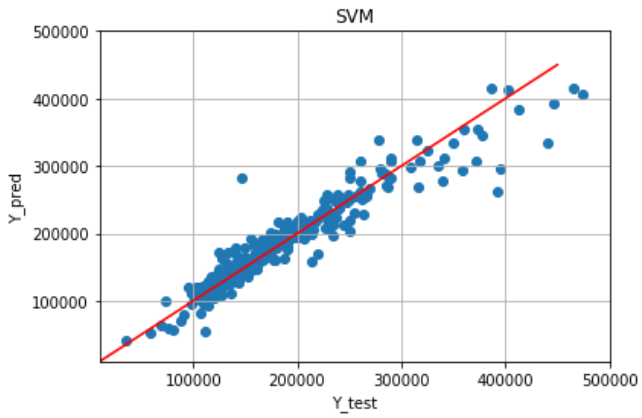

Fig. 3. Support Vector Machine with parameter tuned with k-fold cross validation and grid search.

B. Conclusion

Both are really powerful model and are generally used. The models perfumed really good on the data. Linear Regression giving 89%(approx.) accuracy i.e. R2 score where as SVM giving 85%(approx.) accuracy i.e. R2 score. Linear Regression performing better then support vector machine. Linear Regression time cost is also less than that of SVM.