Prediction of House Price’s Using Logistic Regression

Chethan Yajman Krupakar
State University of New York, Albany
Department of Computer Science
cy695277
cyajmankrupakar@albany.edu

***Abstract*—The house prices are the vital impression of the economy of a location. Unlike the stock prices the house prices can be predicted with some degree of accuracy. Accurate prediction of the house prices is very vital for the prospective house owners, investors, developer, appraiser, tax assessors and other real estate market participants. For many people, buying a house is one of the most valuable investment in their life. The location of the house with other attributes of the house affect the price of the property. The prediction models used earlier were based on the chronological sequence influenced by many factors, which made it difficult to predict the house prices correctly.**

In this paper, we have used logical scientific prediction model -Logistic regression to predict the house prices. The logistic regression was first initialized with the data set of the recently sold houses of a specific location. Once it is initialized, it can be further used to predict the prices of the house in the real-time market. Our results show that the logistic regression approach to the existing problem has been successful, and it performed exceptionally compared to the existing models.

I. Introduction

Houses are mainly used as a shelter to fulfil the most basic need of a person. If a prospective buyer wants to know about the housing trends in various other places, he would have very less knowledge about it. It is very important for the people to invest in a property worth their money based on the specifications they require rather than be charged exorbitant prices by the Real Estate Brokers. As per the Land Registry there will be over millions of transactions, which doesn’t make sense to the public. When we plot the previous sales data, we can see the variation in the prices of the houses during a fixed period. When we can see the average prices at which the houses are sold, the median price and the number of houses sold in the respective period. We collected the data of the average asking price for the houses in the specific period and plotted it on a graph. We can see the differences in the two prices, due to the lack of knowledge of the purchaser. We aim at providing the simple solution to the problem by providing the estimated prices based on different attributes which help in bridging the gap in the market. It also helps the buyer by keeping him well informed about the house and the attributes he needs and the price associated with it.

In this project, we are dealing with the data set of the Ames Housing, IOWA. We use almost all the important attributes to make the best possible prediction of the prices of the houses. This is a very challenging problem as it helps us learn about the real-life housing market problems and help make calculated decisions. The dataset represents residential properties in Ames, Iowa.

Also, I wanted to study about Advanced regression techniques and how they could be utilized to achieve the result. The goal of this project is to create a regression model and a classification model that can accurately estimate the price of the house given the features.

II. Related Work

There are many approaches that has been done to the project. The LASSO, which is known as Least Absolute Shrinkage and Selection Operator is a type of regression model that does various regularization and selection. The LASSO model allows the variables coefficients to 0, which in turn reduces the dimensionality of the model. Reducing a number of variables increases prediction accuracy and interpretability. [1]

The other method which was used was the GBM Gradient Boosting Models, which is the popular algorithm used on kaggle. A variant of the Gradient Boosting Models known as the XGBoost is the best fit for the prediction. This algorithm is like the random forest, in which multiple decision trees are used to optimize over the cost function. [1]

The neural nets have also been used for the deep learning approach. The neural network resembles the functions of that of a human brain. It is a process of steps, in which the input to one layer computes and generates the output, this output becomes the input for the further layers. This method is useful as it can handle complex databases and multiple variable type datasets. [1]

The random forest method uses multiple decision trees and gives one mean prediction tree. The approach is not very clear as lack of coefficients which is lacking the output from regression model. The random forest can be quite robust against outliers and do not require any assumptions of normality.

The other model that is used for the prediction is the Ensemble model or the model Averaging. In this method, many other models are used to roughly create an average prediction. This is a very efficient model as it takes the values by calculating the average of other models, hence it can get the leverage of all other models. Hence this model can have the highest prediction accuracy. [1]

III. Proposed Approaches

We have two different data sets obtained from Kaggle.com, namely the training data-set and the test data set. The training data set contains 1460 entries based the houses sold. We used the training data set to train the model. We are provided with the 1459 test set for which we must predict the sale prices. The test data set contains many values out of which few of them do not have any values and are populated with “n/a” in the table. So, before starting the process, cleaning of the data set is most important. If we train our model based of the attributes which are not present in the test data set, then there are chances that process can seriously affect the accuracy of the model.

To begin with we can examine the range of house prices that we are dealing with by using the box plot which help us better visualize the data we are dealing with.

By looking at the Box plot we can get to know that the maximum price at which the house is sold is around $500K. As we can see there are very few houses which are sold above the price of $500K. So, if we include even those entries it might average out the values of the houses with lower prices. Hence, to make the model more evenly distributed we can plot the histogram to see the distribution.

Looking at the histogram we can confirm that there are very less houses that has sold price above $500K. This helps us make out results more efficient.

In some of the attributes most of them had values except the few where it was important and cannot be neglected as it would affect the price of the property. So, we appended those attributes with 0 or with mean values as compared to the other others.

We are using the logistic regression, here to analyze the data. We are taking the training set of 1460 entries to train the model. There are 76 different attributes present in the dataset, out of which we have taken 36 attributes, for more accuracy using the best parameter function.

IV. IMPLEMENTATION

Building a model is a very important and time consuming process. We should first analyze the attributes and include only independent variables. So, in the initial stages, we looked for the attributes which had more “n/a” values. To overcome this problem, we selected the important attributes which was populated with the data, so that we can get a linear output. It is very important to select the correct attributes, If there training set is not used to model properly then there will be discrepancy in the resulting output. The steps followed are:

- Descriptive Analysis
- Univariable Analysis
- Testing of Collinearity
- Multivariable Analysis
- Model Diagnostics

V. Algorithm and techniques used

Regression analysis is a predictive modelling technique which helps in investigating the relationship between an independent variable and a dependent variable. Regression is an very important tool used for modelling and analyzing the data.

The Logistic Regression predicts the probability of an outcome that can have only two values. We can predict the probability of event success or failure using logistic regression. We use many numerical and categorical outliers to predict the outcome. The logistic regression is not appropriate for predicting the values of binary variable.

The logistic regression output a logistic curve, which can have values between 0 and 1.

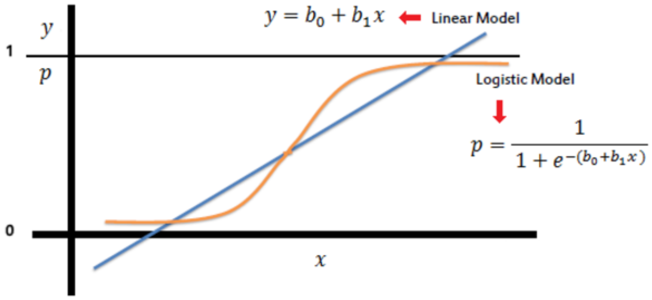


Figure 3.1: Logistic curve [2]

In the logistic regression, the constant (b0) moves left and right and the slope (b1) defines the steepness of the curve. The logistic regression can be transformed into the equation. [2]

$$\frac{p}{1-p} = \exp\left(b_0 + b_1 x\right)$$

[2]

When we take natural log on both sides,

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

[2]

Logistic regression can handle any number of numerical/categorical variables.

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}}$$ [2]

Maximum-likelihood estimation is a common learning algorithm used to make assumptions about the distribution of the data. The best coefficients values obtained would result in a model that can predict values 1 or 0.

Hence when we provide the input to the logistic regression it should be of a specific type.

The dataset must a contain a unique identifier for each record as it is used to identify that row values.

The dataset must contain fixed number of columns for the training set as well as the testing set. Based on the training set the values are calculates, hence if the testing set has any additional columns then it might increase the time of calculation and can affect the accuracy of the output.

The model must contain atleast one column that has numeric value used to predict the output by taking it as an input.

The logistic regression is a linear algorithm. It assumes that the input as well as the output variable has a linear relationship. [3]

It can overfit multiple highly correlated inputs. Calculating the pairwise correlation between all inputs and removing highly correlated inputs. [3]

In building the model we selected the most important attributes from the file "train_after_extracting_Number_Columns.csv" and create a linear model for the price. The process of building any model involves starting with an initial model. The models gets better by several iterations. The process model may do well on the training data but may perform badly for the set of test data commonly known as outfitting. To determine the outfitting we can compare the in-sample and out-sample datasets.

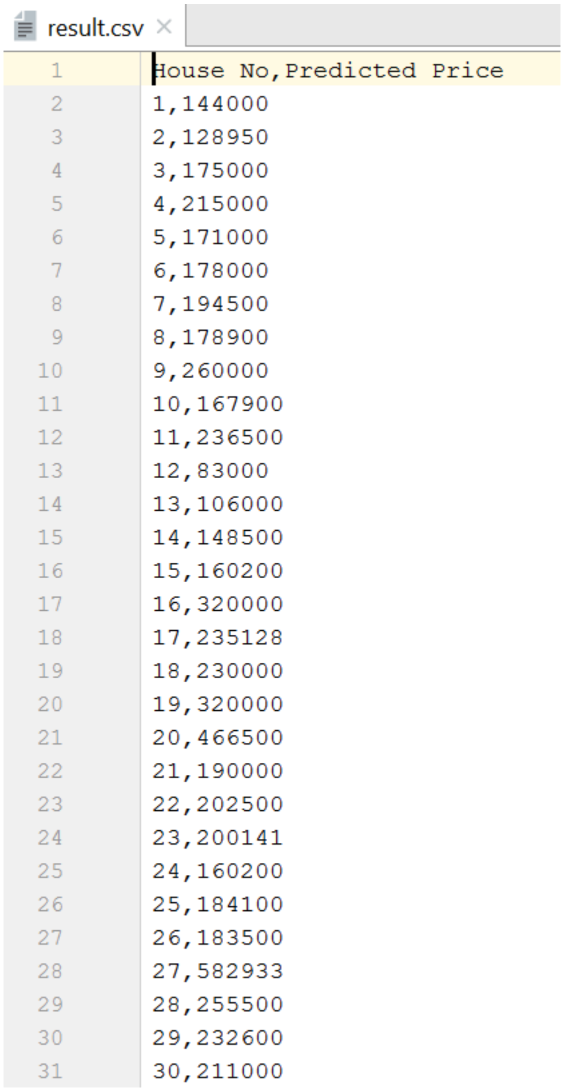
In the project we are using the input file "train_after_extracting_Number_Columns.csv" and reading it line by line. Next we make the training set X and Y vectors by taking the record with the input columns and then storing it by splitting by using a delimiter "," for all the 36 attributes.

After training X and Y vectors, we train our logistic regression model by using the X and Y vectors. Now we take the input test file and read all the rows and store it into another variable. Now we are creating another testing vector called XX but just initializing the vector YY.

Now we predict the house prices based on the trained data set. After writing the data set we then check the accuracy of our accuracy by applying it back on the training set. So if our model is trained correctly then we will get the similar house prices for the houses that are sold and the information provided in the training dataset.

The results are par with that of the training set making it accurate.

The output for the program is in a .csv file format, it is as shown in the Figure 5.1.



```
House No,Predicted Price
1,144000
2,128950
3,175000
4,215000
5,171000
6,178000
7,194500
8,178900
9,260000
10,167900
11,236500
12,83000
13,106000
14,148500
15,160200
16,320000
17,235128
18,230000
19,320000
20,466500
21,190000
22,202500
23,200141
24,160200
25,184100
26,183500
27,582933
28,255500
29,232600
30,211000
```

Figure 5.1: Output file

VI. Conclusion

This study was aimed to predict the House prices in the Ames region of IOWA using a effective prediction tool, Logistic regression. In the beginning the training was done based on the training set that was selected. The training set with the attributes are the input to the logistic regression model. It produces the output for the house number and the predicted price for it.

While in the process of building the model there were few places where very important, like in Albany the AC is not much used as the climate is cold in most of the time. But in Ames, Iowa the weather is very hot and Humid, so the presence of an AC is very important. The presence of an Heater becomes an important attribute in the house selection process in Albany.

Another important factor for deciding the house prices was the presence of an addition bathroom. If the house had an additional bathroom instead of the larger garage space, the people are ready to pay a higher price for it.

Our model consists of 36 variables, which all are significant t 0.05 level. The data set has unlimited potential that can be utilized in lower level logistic classes. We conclude out project with the prediction model having 54.53% accuracy due to change in input values

We were able to achieve a accuracy of about 69.63% by removing the extraneous attributes.

In conclusion, the results show that the Logistic regression model can generate high level of accuracy.

VII. Future scope

In the future scope of the project, we can use various other advanced regression techniques. The various regression models based on Machine learning would be able to provide a more accurate result upto 90%.

To obtain an lower test error, we might need to cross-validate both the training set as well as test dataset together, though this would be more computationally intensive.

References

[1] http://kevinfw.com/post/predicting-ames-house-prices/

[2] http://www.saedsayad.com/logistic_regression.htm

[3] http://machinelearningmastery.com/logistic-regression-for-machine-learning/

[4] http://vicken.me/ames.html

[5] https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/microsoft-logistic-regression-algorithm

[6] http://www.saedsayad.com/logistic_regression.htm

[7] http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7603227

[8] https://www.analyticsvidhya.com/blog/2015/10/basics-logistic-regression/