Vehicle Price Prediction System using Machine Learning Techniques

Kanwal Noor
Computer Software Engineering department,
UET Peshawar (Mardan campus).
KP Pakistan.

Sadaqat Jan
Computer Software Engineering department,
UET Peshawar (Mardan campus),
KP Pakistan.

ABSTRACT

This paper presents a vehicle price prediction system by using the supervised machine learning technique. The research uses multiple linear regression as the machine learning prediction method which offered 98% prediction precision. Using multiple linear regression, there are multiple independent variables but one and only one dependent variable whose actual and predicted values are compared to find precision of results. This paper proposes a system where price is dependent variable which is predicted, and this price is derived from factors like vehicle's model, make, city, version, color, mileage, alloy rims and power steering.

General Terms

Machine Learning.

Keywords

Multiple Linear regression, Car Price, Regression model.

1. INTRODUCTION

Vehicle price prediction especially when the vehicle is used and not coming direct from the factory, is both a critical and important task. With increase in demand for used cars and upto 8 percent decrease in demand for the new cars in 2013, more and more vehicle buyers are finding alternatives of buying new cars outright. People prefer to buy cars through lease which is a legal contract between buyer and seller. The seller category includes direct seller or third party, business entity or insurance company. Under lease contract, the buyers pay regular installments of the item purchased for a pre-defined period of time. These lease installments are dependent upon the estimated price of the vehicle and thus, sellers are interested to know about fair estimated price of their vehicles. It is found through studies that finding fair estimated price of a used car is important as well as challenging. So, there is a need of accurate price prediction mechanism for the used cars. Prediction techniques of machine learning can be helpful in this regard. Machine learning uses two techniques, i.e., inductive and deductive. The deductive learning is based on the usage of existing facts and knowledge to deduce new knowledge and facts while in inductive machine learning new computer programs are created by finding patterns and rules in the new data sets which were never explored before.

We use deductive approach of multiple linear regression since it creates new values based on existing values. In this technique, there is single dependent variable Y and there can be multiple independent variables X. The relationship among variables is direct or linear. This paper has the following goals:

i. Design: The research includes the design of a system explaining linear relationship between X and Y which are price and other factors like model and make of the car.

ii. Predict: The research predicts the price of the vehicle using linear regression model which identifies different patterns and projects and predicts the value of the vehicle.

iii. Confirm: The research finds out which variable associated with the vehicle is the best predictor of its price.

There are many types of linear regressions and this research uses multiple linear regression where there are more than one independent variables. The data associated with the investigation was very large because there are thousands of used cars and each car's data comprises of values of many features. Both data gathering and analysis are complex. In the beginning, two thousand records of used cars were recorded and the data was obtained from pakwheels which is a well-known online company for reselling used and new cars in Pakistan. Research used only those cars that contained price details so that the results could be verified. Features like car's model, make, version, city, color, mileage, engine capacity, alloy rims, power steering, engine type and price were included.

The organization of this paper is such that Section II is on related work, Section III discusses the methodology and Section IV presents the results and in depth discussion. Section V concludes the paper and offers directions for future research.

2. RELATED WORK

Researchers more often predict prices of products using some previous data and so did Pudaruth [1] who predicted prices of cars in Mauritius and these cars were not new rather second hand. He used multiple linear regression, k-nearest neighbors, naïve Bayes and decision trees algorithm in order to predict the prices. The comparison of prediction results from these techniques showed that the prices from these methods are closely comparable. However, it was found that decision tree algorithm and naïve bayes method were unable to classify and predict numeric values. Pudaruth's research also concluded that limited number of instances in data set do not offer high prediction accuracies [1].

Multivariate regression model helps in classifying and predicting values of numeric format. Kuiper [2] used this model to predict price of 2005 General Motor (GM) cars. The price prediction of cars does not require any special knowledge so the data available online is enough to predict prices like the data available on www.pakwheels.com. Kuiper [2] did the same i.e. car price prediction and introduced variable selection techniques which helped in finding which variables are more relevant for inclusion in model. He encouraged students to use different models and find how checking model assumptions work. Another similar research by Listiani [3] uses Support Vector Machines (SVM) to predict the prices of leased cars. This research showed that SVM is far more accurate in predicting prices as compared to the multiple linear regression when a very large dataset is available. SVM also handles high dimensional data better and avoids both the under-fitting and over-fitting issues. Genetic algorithm is used by Listiani [3] to find important features for SVM. However, the technique does not show in terms of variance and mean standard deviation why SVM is better than simple multiple regression.

Limsombunchai [4] concluded in his research that neural networks are better in estimating price of a house. His method offered higher prediction accuracy as compared to hedonic method. Although neural network (NN) operates like hedonic price theory because it defines presence of attributes associated with the house and help in prediction, yet NN operates such that the model is trained first and then tested for prediction. Using both the methods, it was found that NN gives higher R-sq and smaller root mean square error (RMSE) while the hedonic method offered lower values. The limitation of this research was that it offered strong evidence of prediction superiority but did not talk of forecasting capability between the two methods used. Also, actual house prices are missing in the research and only estimated prices were used avoiding difficulties of data collection [4].

In another research, Bourassa et al. [5] represented how price of a house is related to the prices of adjacent properties. They used four models to check this relation. Different models were considered including 2 Ordinary Least Squares (OLS), 4 geo-statistical and 2 lattice models which showed that when submarket variables are included in the models, they give higher predictions accuracies and R-sq value than those obtained without adding submarket variables. Research also stated that OLS with submarket variables has high accuracy as compared to other models. Percentage raised from 39.8% to 46.8% when submarket variables were added to OLS model. Nau [6] has outlined that the following assumptions apply to simple regression model:

1. There exists a linear relationship, Y=βiXi, between the variable Xi and the expected value of Y where the value of β is constant.
2. Unexplained variations in Y exist due to independent variables which are random in nature.
3. Variance in all these variables is almost the same.
4. They follow normal distribution.

According to Nau [6], the following requirements must be satisfied by an effective regression model:

a) The model should gather useful data and the source and method of gathering of this data must be known.
b) The model must have the ability of performing descriptive analysis on data where general patterns could be understood well.
c) There should be the capability of comparison across different models.
d) There should be a capability of validation to see whether a given model accomplishes the stated assumption. In case of no compliance against the stated assumption, the requirement of an alternative model could be identified.
e) Based on a given accuracy level, it should have the capability to choose the appropriate model based.
f) There should be the capability to derive useful insight from the whole process [6].

3. METHODOLOGY

This research aims to develop a good regression model to offer accurate prediction of car price. In order to do this, we need some previous data of used cars for which we use price and some other standard attributes. Car price is considered as the dependent variable while other attributes as the independent variables. Let X be the input and Y be the output, the linear regression correlation can be expressed as:

$$Y = \beta_0 + \beta_1 X \qquad (1)$$

In the above equation, $\beta_{0,}$ $\beta_1$shows the regression coefficients, Y is the output or required variable and X shows the input. The above equation represents relation in case of single input. In case of multiple inputs, equation will be:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots \beta_n X_n \qquad (2)$$

In the above equation, X_1, $X2, \ldots\ldots\ldots X_n$ represent multiple inputs. For n inputs there are n+1 regression coefficients.

Used cars have very high dimension input space. There are various attributes and features that have impact on car price which naturally generates large set of data leading to complexity in analyzing it [7]. The focus of this research is to build such a model which has the capabilities of dealing with high complexity and gives accurate results irrespective of the magnitude of data set. The input data is gathered from pakwheels in a month or two. In the beginning, 2000 records of used cars were recorded. The collected data included variable values for price, engine capacity, color, advertisement date, number of views, mileage in kilometer, power steering, alloy rims, transmission, type of engine, registered city, city, version, model, make and model year. Once the data collection was over, we processed data using multiple linear regression technique for price prediction. In this research, statistical software Minitab was used in which we input the data and analyze the results via linear regression application. Initially, all attributes were considered, but later we applied the variable selection techniques on our input data and found the most significant variables and skipped all other insignificant variables. Table 1 summarizes sample of our input data fed into regression model for price prediction.

Table1: Input Data

SERIAL NO	MODEL YEAR	MODEL	ENGINE TYPE	PRICE (RS)
1	2011	Vitz	petrol	1560000
2	2012	Alto	CNG	840000
3	2006	city	Petrol	995000
4	2011	Swift	Petrol	950000
5	2007	Passo	Petrol	900000
6	2003	corolla	CNG	1090000
7	2007	Alto	CNG	560000
8	2008	Alto	CNG	520000

9	2010	city	Petrol	1250000
10	2012	city	Petrol	1495000
11	2011	city	Petrol	1395000
12	2010	civic	Petrol	1450000
13	2004	Smart	Petrol	850000
14	2011	Corolla Axio	Petrol	1000000
15	2010	Corolla Axio	Petrol	907750

As mentioned earlier, 2000 records were collected initially and later only 1699 records were left once pre-processing was applied on data. Preprocessed data is then further processed using multiple linear regression. Since the data (e.g., engine type and model as shown in Table 1) is textual in nature, therefore, these attributes have been coded and converted into category codes, i.e., 1,0, using Minitab because regression works on numeric data. This data is then fed to regression model for estimating price.

4. RESULTS

Least square method has been used for model estimation with the following results obtained from Minitab:

4.1 Result of R-sq value

Frost [8] discusses that, a linear model explains the percentage response variation in variable called R-square (R-sq). This means that a high R-square value is an indication of better fitness of the model to the data resulting in more accurate results.

Table 1: R-square and adjusted R-square

S	R-sq	R-sq (adj)	R-sq (pred)
92622.8	98.61%	98.50%	*

Standard deviation of the error term is shown in the Table 2 where the 98% variation is represented by 98.61% of R-sq value. The modified version of R-sq called the adjusted R-sq value (R-sq (adj)) is adjusted for predictors in the model. This shows that a high percentage (98.50%) in the dependent/response variable is due to engine type, model year and car's model. To see how well the model predicts the response for new input records, the value of R-sq (pred) is used [9].

The natural variability which exists in each data set cannot be explained with the current model. Since, too many predictors can be added to the proposed model due to high R-sq value, therefore only R-sq is not enough to explain all such occurrences. Due to this reason, we have used the adjusted R-sq for checking accuracy of the model [9].

4.2 Results of the predicted response

Using Minitab, we get the price being predicted in the additional column "FIT". Apart from it, the residual value being the difference between actual and predicted response variable is also calculated. Results' samples of actual and predicted price are shown in Table 3 indicating the number of observations, FIT, Resid and standard Resid.

Table 2: Actual and Predicted Price

Obs	Price	Fit	Resid	Std Resid
1	1560000	1372547	187453	2.05 R
6	1090000	1277393	187393	2.03 R
13	850000	850000	0	*
14	1000000	984842	15158	0.20
15	907750	938065	-30315	-0.40
17	875000	1065347	-190347	-2.06 R
22	800000	800000	0	*
25	2850000	2850000	0	*
26	1750000	1750000	0	*
28	5200000	5200000	-0	*
29	2650000	2575000	75000	0.99
32	925000	925000	0	*
34	2150000	2150000	0	*
35	2675000	2575000	100000	1.32
36	2450000	2450000	0	*
37	1800000	1800000	0	*
38	3650000	3650000	0	*
39	2400000	2575000	-175000	-2.30 R
42	2000000	2000000	0	*
44	1000000	984842	15158	0.20

Price column in Table 3 shows the actual price collected from pakwheels while "Fit" column shows the price which is being predicted as calculated by the Minitab. As explained earlier, the difference of actual and predicted price is represented by the *Resid* while the *StdResid* shows the value of the standardized residual which is obtained by dividing the value of residual with its standard deviation estimate. The last column also shows R with the numeric values in some cases which highlights an observation with large standardized residual. Moreover, it can also signify the representation of outliers in the proposed model. Normally, a value greater than 2 for standardized residuals is considered large. Since raw residual have no capability to detect outlier, therefore, Standardized residuals are considered a better measure for such detection [10].

4.3 Results of Prediction Interval and Confidence Interval

Price prediction can be done by giving input values of the required information, *i.e.*, engine type, model and model year in the proposed regression model. When required information is entered, we obtained the following Predicted Interval (PI) and Confidence Interval (CI) values (shown in Table 4).

Table 3: PC and CI

Variable	Setting		
Model Year	2013		
Engine Type	Petrol		
Model	Corolla		
Fit	SE Fit	95% CI	95% PI
1792980	9318.04	(1774703, 1811257)	(1610386, 1975573)

Fit column In Table 4 provides value 1,792,980 which is the predicted price for newly made observation. The standard error value of the Fit is SE Fit which indicates variation in the estimated mean response for a given setting of variables [11]. In the Minitab, two other values are generated which are PI and CI for checking prediction accuracy. The PI range shows that the new predicted value will fall in that range [12]. The table also shows CI range which tells range in which the mean value will fall. It can be seen that the PI range is wider than CI range and it is always so because of the uncertainty which is involved in predicting single value rather than the mean value [12].

4.4 Results of Regression plots

Plot of regression shows if residual values are normally distributed or not. Plots shown in Figure 1 and Figure 2 were obtained after applying regression model on the data. In the results, the normal probability plot of residuals versus percent are drawn in upward slop line.

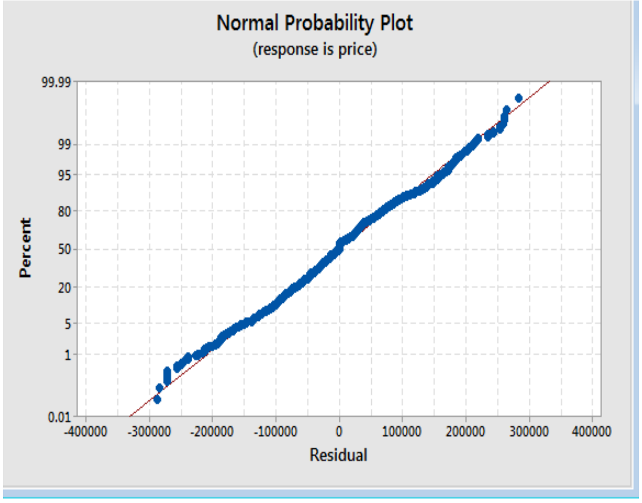

Figure 1: Normal Probability Plot

The Figure 1 shows that the blue points are clustered around the red line which is almost a straight ascending line. The balanced distribution of points on line helps in concluding that the residuals are normally distributed and hence the assumption is true that normality is valid [10]. The values in Figure 1 are plotted as histogram in Figure 2 showing normal distribution. If a curve is drawn over the bars in Figure 2, it will exhibit an inverted bell showing how price mechanism works at highest and lowest prices (low occurrences) and mean/average price (higher occurrences thus higher frequencies).

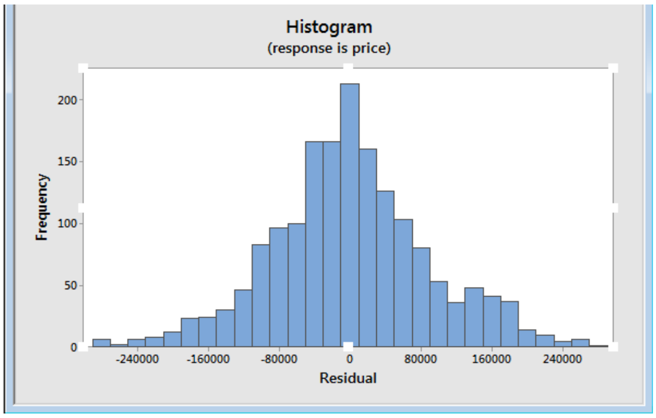

Figure 2: Histogram

5. CONCLUSION

The data set used in this paper can be very valuable in conducting similar research using different prediction techniques. The prices of vehicles can be predicted using this data set on same or different prediction software as well. The data obtained under this research facilitated in prediction of prices of used cars through linear regression method. Many assumptions were made on the basis of the data set. The proposed system evaluated variables and selected the most relevant variables out of the dataset and reduced the complexity of model by eliminating unrelated variables during processing and analysis phase. The future price prediction of used cars with the help of same data set will comprise of using fuzzy logic, KNN and genetic algorithm.

6. REFERENCES

[1] Pudaruth,S. 2014. “Predicting the Price of Used Cars Using Machine Learning Techniques”, International Journal of information & Computation Technology,4(7), p.753-764.

[2] Kuiper, S. 2008. “Introduction to Multiple Regression: How Much Is Your Car Worth?”, Journal of Statistics Education, 16(3).

[3] Listiani M. 2009. Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. Master Thesis. Hamburg University of Technology.

[4] Limsombunchai, V. 2004. House price prediction: Hedonic price model vs. artificial neural network. In New Zealand Agricultural and Resource Economics Society Conference, New Zealand, pp. 25-26.

[5] Bourassa, S.C., Cantoni, E. and Hoesli, M. 2007. “Spatial dependence, housing submarkets, and house price prediction”, The Journal of Real Estate Finance and Economics, 35(2), p.143-160.

[6] Nau, R. 2014. Notes on linear regression analysis, Lecture handouts, Duke University, Furqa School of Business, 26 nov 2014.

[7] Singh, Y., Bhatia, P. K., & Sangwan, O. 2007. “A review of studies on machine learning techniques”, International Journal of Computer Science and Security, 1(1), 70-84.

[8] Frost, J. 2013. Regression analysis: How do I interpret R-squared and assess the goodness-of-fit. The Minitab Blog, 30. Available online from: http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit (Last accesed: 29-11-206).

[9] Frost,J. 2013. Multiple Regression Analysis: Use Adjusted R-squared and Predicted R-squared to Include the Correct Number of Variables. Available online from:http://blog.minitab.com/blog/adventures-in-statistics/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables (Last accessed: 29-11-2016).

[10] Minitab Express Support. Interpret all statistics and graphs for Multiple Regression.[Online] Available from: http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/multiple-regression/interpret-the-results/all-statistics-and-graphs/

[11] Minitab Express Support. Interpret all statistics for Predict.[Online] Available from: http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/predict/interpret-the-results/all-statistics/

[12] Frost,J. 2012. How to Predict with Minitab:Using BMI to Predict the Body Fat Percentage,Part 2.[Online] Feb 23 2012. Available from: http://blog.minitab.com/blog/adventures-in-statistics/how-to-predict-with-minitab-using-bmi-to-predict-the-body-fat-percentage-part-2

[13] https://www.pakwheels.com/ (Last accessed on 29-11-2016)