CAR PRICE PREDICTION USING MULTIPLE REGRESSION

SUBJECT: DOMAIN APPLICATION OF PREDICTIVE ANALYTICS

BOJAVARDHAN R CHEERAPUREDDY

School of Computing

NATIONAL COLLEGE OF IRELAND

Dublin, Ireland

x19191421@student.ncirl.ie

ABSTRACT

People in many developing countries like Europe, USA, Canada adopted leasing culture because of heavy prices on new cars. The demand for used cars has become a booming business now a days because of steadily increase in car manufacturing. Taking advantage of this situation third party sellers used to sell the cars with unrealistic prices compared to the original cost of the car. Because of heavy price on the new cars the usage of used cars has been increasing. Predicting a used car price was a highly promising area of research because it needed significant effort and expertise. For predicting the price of a used car, we are examining various attributes in-order to give accurate price prediction. This paper concentrates mostly on the factors that are showing major impact on the price of a used car. As the number of dependent variables are more, we are using multiple regression for better and accurate price value.

Keywords – Price Prediction; Used Cars; Multiple Regression; Feature Selection; Prices.

I. INTRODUCTION

In the present market situation, the sales for used car have been increasing due to heavy prices on new cars. Manufactures are fixing the price of new cars with some extra expenses acquired by the government as charges. Because of heavy prices people are unable to afford buying a new car. In Many developing countries selling of used cars became a booming business in the market industry. The prediction of price for a used car becomes a major challenge for the companies. Because they want to satisfy the customer and should make them feel that their investment on the used car is worthy. So, it becomes a much-needed issue for the companies to predict the correct price for the car. Predicting the price of a used vehicle is a fascinating and important issue that need to be discussed. Because of unfavorable price estimation of cars and the nomadic behavior of people in many developed nations, the used vehicles are generally purchased on the rental basis, where there is an understanding between the purchaser and vender. Foreseeing the resale estimation of a vehicle is definitely not an easy task the price of these used cars are relies on various components. The most significant factor is typically the age of vehicle, manufacturing model, mileage, horse power, no of kilometers driven etc.

So, the car sales depend on all these parameters. For proving that the sales of cars are increasing enormously I have considered a dataset from the KAGGLE website where the whole data refers to a Chinese automobile company where they are in plan to start a branch in US for opening a branch in US the company is interested in knowing all the

manufacturing details of how the US citizens were interested in purchasing cars. This whole dataset is having different number of entities related to the car such as id's of the car, brand name of the car, what kind of fuel type is used for that car, body of the car and the locality of engine all these have been mentioned in the dataset. My study will even help the company to understand how the value of the car price can vary in different locations according to their customers response in different countries. For calculating the price values, I have used multiple regression model in machine learning where this model is performed when there are multiple independent variables and when all these variables helps in predicting the car price model. So, by calculating the price of the car the Chinese company will get benefited in identifying the people's interest in US and they can revamp the sales by establishing a branch in US as well. So, if once all the prices of the cars are identified based on different entities the company identifies different possible solutions in improving their business strategies based on the engine, brand, and different factors. Because individuals have their own interest in choosing cars based on brand, mileage, driving wheel and body of the car. If once we identify the car price based on these parameters, we can verify in which area of the people were showing major in interest in their respective entities. So that the company can get possible solutions in improving the sales by providing the features in which the people in different areas were interested. So, this whole project will help Chinese company in identifying the possible solutions to improve their sales using machine learning algorithms.

II. RELATED WORK

Several associated works were performed formerly almost about used car rate prediction. Robert Tibshirani [1] suggests a new technique called Lasso, which keeps down the remaining sum of squares. It gives a sub-set of attributes that are used to be include in the system to get minimum error rate. Likewise, decision trees affected from over-fitting if they're now no longer pruned/shrunk.

Listiani M [2], suggest a model which was established by using SVM (Support Vector Machines) can estimate the rate of car this has been rented with higher precision than the simple multivariate regression. SVM is superior in dealing the dataset's with extra dimensions and it's less prone on both over-fitting and under-fitting. The only disadvantage of this thesis is to modify the simple regression to SVM was not visible in the basic index like mean, standard deviation or variance.

Noor and Sadaqat Jan [3] compile a prototype for car rate prediction by making use of linear regression. They performed varying selection of dataset that was created before the period of eliminating the features only following are considered model number, model name, engine type, price, mileage as the features. Through the given data the result of achieving the accuracy rate of 98%.

Pudaruth [4] predicts the used cars price by using various machine learning algorithms, likely: Naive Bayes, multiple linear regression, decision tree and k-nearest neighbors. The dataset collected from everyday newspapers in the local area Mauritius, as time passes it will impact the car price. Classifying the price based on the analysis of the work that has done by collecting the data, in addition the classification algorithms for finding the data given to predict the numeric values gives good results.

The report by Gonggi S [5] build a model for predicting the used car price by using Artificial Neural Networks (ANN).He taken attributes like brand, miles travel, car life into consideration. The model that built could deal the nonlinear relations of the data which was not there in previous models to the simple linear techniques. This non-linear can predict the price of a car with better accuracy compared to other models.

Wu et al. [6] carried out car rate prediction study, through the use of neuro-fuzzy information-based system. The author took the subsequent attributes into consideration: time of production, brand and sort

of engine. Their prediction version produced comparable results because of the easy regression standard.

Additionally, they made an professional system named ODAV (Optimal Distribution of Auction Vehicles) as there may be an excessive value for selling the vehicles on the termination of the rental year by vehicle dealers. This device offers insights to the great value for cars, in addition to the place in which the best value may be gained. Regression version primarily supported on k-nearest neighboring machine studying algorithm that is used for predicting the rate of a vehicle. This device has a bent to be enormously successful due to the fact that greater than 2 million cars have been exchanged via this system [5].

Another method was given with the aid of using Richardson work in his paper [7]. His principle became that the vehicle producers produce greater durable vehicles. This has roots in environmental issues approximately the weather and it offers better gasoline efficiency. Richardson implemented more than one regression evaluation and verified that hybrid automobiles maintain their cost for longer period than traditional vehicles.

Sun et al. [8] suggested the software of on-line used vehicle price evaluation version by the usage of the optimized BP neural networking algorithm. They brought up a brand new optimization technique known as Like Block-Monte Carlo Method (LB-MCM) to optimize secret neurons. The end result proven that the optimised version yielded better accuracy rate compared to the non-optimized standard based at the preceding associated works, we found out that no one have applied gradient boosting approach for knowing the rate of used vehicle yet. By using gradient boosted regression trees, we determined to create a model for price prediction system for used cars.

Bharambe and Dharmadhikari (2015)[9] proposed a system by using artificial neural networks (ANN) that analyse the market price and predict the behavior of present market.There are so many factors that effect the stock market by dialy news like dividents, earnings, change of board and profits. Neural network market has a derive meaning in losse of data remarkably, this has been increased the forcasting the stock marketing. The system that has claimed acurate rate more than the existing ones by 25%.

Rose (2003)[10] predict the cars that are specifically manufacturing from specfic companyes by using NN. the neural network(NN) were able to estimate the ones who are searching for used cars through the retention rate. By using this system the sales of the product is estiated with the accuracy. There work, is to apply neural networks in a brand new application, i.e., that of predicting the rate of second-hand cars.

Jassbi et al. (2011)[11] used two methods to predict the color of the car that differentate the price was found that the thickness of the color was 2/99 for neral networks and 17/86 for regression.Ahangar et al.(2010)[12] uses neural network with linear regression to predict the stock market. They descoverd NN palys major role in speed and accuracy while compared with linear regression.

Regression method is one of the statistical analysis techniques used to describe the relationship between one response variable with one or more explanatory variables [13]. Linear regression is divided into 2 namely simple linear regression and multiple linear regression. prediting the SP of the used vechiles by takinng the additional attributes into consideration like diatance, color of cars, manufacture year and cities of Semarang, Jakarta, and Bandung those type of cars have the trust value of 63.20%. Though the value of car can be vary by the concluding the calculation is correct or not. The data used to find linear equations have independent variables and dependent variables.

There are mainly five steps to decide like, information, alternatives,product,purchase evalution, and problem recognition. Because most of the second-hand cars are damaged and needed repair.

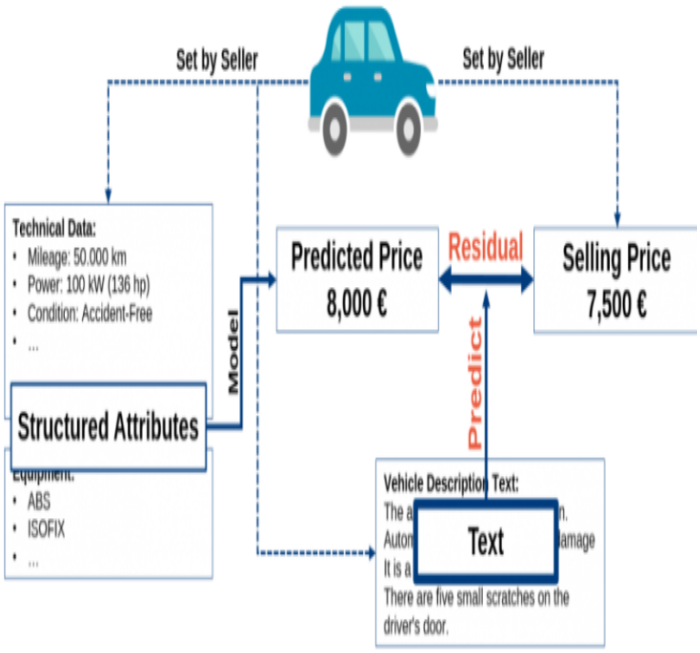

Fig (1) Process Flow

Using Wikipedia, genre, source, internet search and twitter they've anticipated the counts of tweets and page views. The proposed proactive smart decision guide System for the recognition of the web article. Prediction the popularity accuracy using the algorithm that is by taking the rolling home windows assessment strategy. Most of the researches are primarily based on studying the early person feedback and additionally, they consider the sooner potential information content material and domains. In the other prediction they anticipated the recognition of the thing not only based on their attraction however additionally with many different research articles with which it's far competing with many other different journals. Prediction models like SVM, Ranking SVM's, Naive Bayes, Artificial Neural Networks (ANN), and investigated and extra advanced methods like random forest, precision can be boomed by adaptive boosting.

III. METHODOLOGY

Multiple linear regression deals with the relation between the multiple independent variables with one dependent variable. When the predictive variables are more than 2 then it is called a multiple linear regression. In this car price prediction project, the dependent variable is prising of the car whereas the dependent variables are car mileage, year, kilometers driven and horse power are the dependent variables. The dependent variable car price is depended on the all independent variables as they help in prediction the price of the car with dependency to one another. The result is influenced by the other independent variables. Change in one independent variable may affect in the performance of the entire model as everything is inter related to each other. Thus, this model describes how well the response variable linearly depend on the other independent variables.

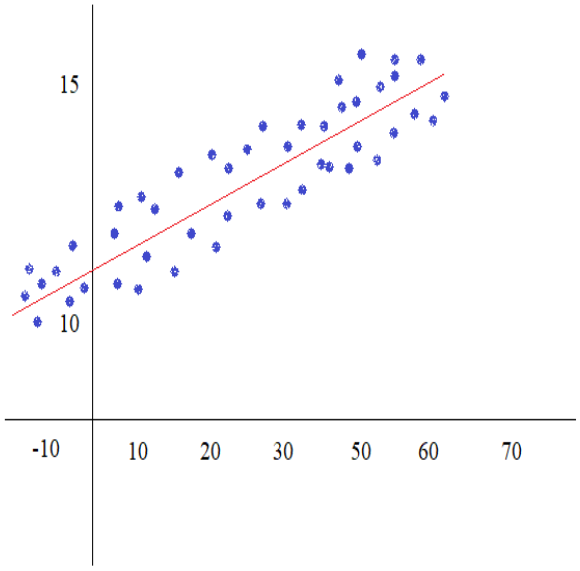

Fig (2) Scatter Plot

In this project, the main objective is to predict the price of the car. This can be predicted with the help of the variables like mileage, kilometres driven, year car bought and horse power. Here the resulting variable or the dependent variable be the car price and the other variables, horse power, kilometres driven, year car bought and mileage will be the independent variables. Thus, the car price is linearly dependent on these

independent variables. The red line in the figure: shows the multicollinearity among the dependent and independent variables. The multiple regression undergoes various steps in performing the model on the dataset. This model in the project follows CRISP-DM framework.

This Multi linear regression methodology undergoes 5 major steps:

1. Selecting the variables
2. Testing the model
3. Refining the model
4. Addressing the issues
5. Validating the model

Once we uploaded the data we check for null or missing values in the dataset because they show a major impact on the prediction values. Once all the null values are find we will replace them with duplicate values or omit the null values once the pre processing is done we then split the data as training set and test set. After splitting we will fit the regression algorithm for the dataset and then will do the evaluation metrices and check whether the model used for the dataset is fitted perfectly and giving good predictions.

IV. EXPERIMENTS AND RESULTS

The Q-Q plot is majorly used for the regression models. As I have performed Multiple Regression model, I have divided the data into train and test sets to predict the car prices. So, once after dividing the data the Q-Q plot plays an indispensable role in informing the model whether the data is divided equally for both train and test sets.

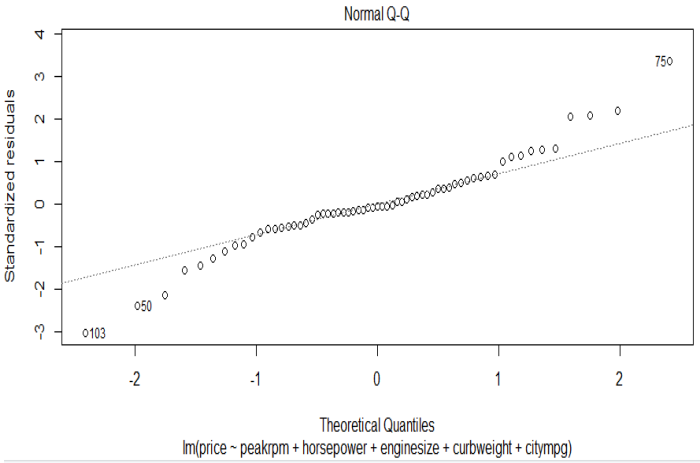

Fig (3) Normal Q-Q Plot

From the above-mentioned values, we can understand that how well the model is performing. If we consider, F-Stat value it is around 103.5 which is immensely high compared to 1. Based on this value we can conclude that the predicted and response variables are highly connected. In the same way, the whole model is predicting the car price values with an accuracy of 90 % which can be known based on Multiple R-Squared value and it is very close to 1 which means we can say that the model values predicted by the multiple regression model are perfect and is a good fit model. In addition to these values we can analyze the value of p-value which is less than 0.001.

```
Residual standard error: 2988 on 57 degrees of freedom
Multiple R-squared:  0.9008,    Adjusted R-squared:  0.8921
F-statistic: 103.5 on 5 and 57 DF,  p-value: < 2.2e-16
```

Fig (4) Residual Values

V. CONCLUSION

Predicting the price of car has become a challenging task because there are more numbers of independent variables which will be used for predicting the price accurately. In this paper, using various variables we are predicting the price of the car accurately by using multiple regression. The dataset was collected from Kaggle. This research will help most of people who are seeking to buy used car with correct price. This paper concentrates mostly on the factors that are showing major impact on the price of a used car. As the number of dependent variables are more, we are using multiple regression for better and accurate price value

REFERENCES

[1] Robert T. (1996) Regression Shrinkage and Selection Via the Lasso. In: Journal of the Royal Statistical Society: Series B (Methodological) Volume 58, Issue 1

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

[3] N. Kanwal and J. Sadaqat, "Vehicle Price Prediction System using Machine Learning Techniques," International Jounal of Computer Ap-plications, vol. 167, no. 9, pp. 27–31, 2017.

[4] S. Pudaruth, "Predicting the Price of Used Cars using Machine Learning Techniques," International Journal of Information & Computation Technology, vol. 4, no. 7, pp. 753–764, 2014.

[5] Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on (Vol. 2, pp. 682-685). IEEE.

[6] Wu, J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications, 36(4), 7809-7817.

[7] Richardson, M. S. (2009). Determinants of used car resale value. Retrieved from: https://digitalcc.coloradocollege.edu/islandora/object /coccc%3A1346 [accessed: August 1, 2018.]

[8] N. Sun, H. Bai, Y. Geng, and H. Shi, "Price evaluation model in second-hand car system based on BP neural network theory," in 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), jun 2017, pp. 431–436.

[9] Bharambe, M. M. P., and Dharmadhikari, S. C. (2015) "Stock Market Analysis Based on Artificial Neural Network with Big data". Fourth Post Graduate Conference, 24-25th March 2015, Pune, India.

[10] Rose, D. (2003) "Predicting Car Production using a Neural Network Technical Paper- Vetronics (Inhouse)". Thesis, U.S. Army Tank Automotive Research, Development and Engineering Center (TARDEC)

[11] Jassibi, J., Alborzi, M. and Ghoreshi, F. (2011) "Car Paint Thickness Control using Artificial Neural Network and Regression Method". Journal of Industrial Engineering International, Vol. 7, No. 14, pp. 1-6, November 2010

[12] Ahangar, R. G., Mahmood and Y., Hassen P.M. (2010) "The Comparison of Methods, Artificial Neural Network with Linear Regression using Specific Variables for Prediction Stock Prices in Tehran Stock Exchange". International Journal of Computer Science and Information Security, Vol.7, No. 2, pp. 38-46.