CAR RESALE VALUE PREDICTION

Name of the Team leader: Rakshana.H

Team ID: PNT2022TMID00778

Roll number: 2019PECCS169

Registration Number: 211419104213

Name of the Team Member 1: Swarnamalya.N

Roll number: 2019PECCS201

Registration Number: 211419104279

Name of the Team Member 2: Swathi.G

Roll number: 2019PECCS280

Registration Number: 211419104280

Name of the Team Member 3: Shakthi Arumugaraj

Roll number: 2019PECCS174

Registration Number: 211419104227

ABSTRACT

Predicting used car prices is one of the important and intriguing research fields. The market for used cars has seen an upsurge in demand, which has impacted both purchasers and sellers' businesses increased. Expertise is needed for dependable and accurate prediction. knowledge of the subject because the cost of cars is a factor on a number of crucial criteria. In this study, a supervised Regression using the KNN (K Nearest Neighbor) machine learning algorithm method to evaluate used-car prices. Through this investigation, A variety of trained to test ratios were used to analyse the data. As as a consequence, the suggested model is fitted with an accuracy of about 85% like the improved model. The predictions are then evaluated and compared in order to find those which provide the best performances. A seemingly easy problem turned out to be indeed very difficult to resolve with high accuracy. All the four methods provided comparable performance. In the future, we intend to use more sophisticated algorithms to make the predictions . Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle’s price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. On a dataset made up of the selling prices of various makes and models across American cities, we put several learning techniques into practise and evaluate their effectiveness. Our findings demonstrate that, while computationally intensive, the Random Forest model and K-Means clustering with linear regression produce the best outcomes. Traditional linear regression also produced good results, with the benefit of requiring much less training time than the approaches outlined above.

LITERATURE SURVEY

[1] Used automobile prices have been predicted using a variety of studies and related works utilising various methodology and approaches, with accuracy ranging from 50% to 90%. In (Pudaruth, 2014), the researcher suggested to forecast used automobile prices in Mauritius. He employed a variety of machine learning approaches to obtain his results, including decision trees, K-nearest neighbours, multiple regression, and naive bayes algorithms.

Achieved results ranged from accuracy of 60-70 percent, the author suggested using more sophisticated models and algorithms to make the evaluation, with the main weakness off the decision tree and naïve Bayes that it is required to discretize the price and classify it which accrue to more inaccuracies. Moreover, he suggested a larger set of data of data to train the models hence the data gathered was not sufficient.

[2] (2018) (Monburinon et al.) Data from a German e-commerce site, totaling 304,133 rows and 11 characteristics, was collected in order to estimate the pricing of used cars using several methods. The results were measured using Mean Absolute Error (MEA) in order to compare the findings. Each model received the same training and testing datasets. The gradient boosted regression tree produced the best results, with mean absolute errors of 0.28, 0.35 for multiple linear regression, and 0.55 for mean absolute errors. The authors recommended changing the settings in next studies to produce better outcomes as well as switching from label encoding to one hot encoding for more accurate categorical data interpretations.

[3] Three alternative machine learning methods were employed by (Gegic, Isakovic, Keco, Masetic, & Kevric, 2019) from the International Burch University in Sarajevo to forecast used automobile values. After preprocessing, the used automobile data scraped from a local Bosnian website totaled 797 car samples. The techniques suggested were

Support Vector Machine, Random Forest, and Artificial Neural Network. Results showed that employing a single machine learning algorithm alone produced results that were less than 50% accurate, however combining the algorithms with pre-calculation of pricing using Random Forest produced results that were up to 87.38% accurate.

[4] After pre-processing 1699 records from a used car website in Pakistan called Pak Wheels, (Noor & Jan, 2017) were able to predict the price of the cars with a high level of accuracy using multiple linear regression models. They were able to achieve an accuracy of 98%. This was accomplished by reducing the total number of attributes using variable selection technique to include only significant attributes and to lessen the complexity of the model.

[5] 2020 (K. Samruddhi & Kumar) In order to predict used car prices from a data set obtained from Kaggle that contained 14 different attributes, it was suggested using a supervised machine learning model using K-Nearest Neighbor. Using this method, accuracy reached up to 85% after varying the value of K as well as changing the percentage of training data to testing data; as would be expected, when increasing the percentage of data that is tested, better accuracy results are obtained. Using the K fold approach, the model was additionally cross-validated with 5 and 10 folds.

[6] (Gongqi, Yansong, & Qiang, 2011) proposed using Artificial Neural Network (ANN) through a combined method of BP neural network and nonlinear curve fit and have achieved accurate value prediction with a feasible model.

[7] Support Vector Machines (SVM) are significantly more accurate in big datasets with high dimensional data than multiple linear regression, according to research from (Listiani, 2009) who used SVM to evaluate lease automobile costs. While the SVM might take up to a day to compute the results, multiple linear regression computations

can take few minutes. Although multiple linear regression is straightforward, SVM is far more precise.The use of multiple linear regression may be more appropriate in our instance because the study provides Samples with up to 178 characteristics, which is significantly more than the suggested variable in our study.

[8] (Kuiper, 2008) Collected data from General Motor of cars that are produced in 2005, where he as well used variable selection technique to include the most relevant attributes in his model to reduce the complexity of the data. He proposed used Multivariate regression model that would be more suitable for values with numeric format. In order to predict the price of used cars, researchers.

[9](Nabarun Pal, 2018) used a supervised learning method known as Random Forest. Kaggle's dataset was used as a basis for predicting used car prices. In order to determine the price impact of each feature, careful exploratory data analysis was performed. 500 Decision Trees were trained with Random Forests. It is most commonly used for classification, but they turned it into a regression model by transforming the problem into an equivalent regression problem. Using experimental results, it was found that training accuracy was 95.82%, and testing accuracy was 83.63%. By selecting the most correlated features, the model can accurately predict the car price. In light of the number of works that have been done in this field, another group of researchers.

[10] (Jian Da Wu, 2017) conducted research on this topic and tried to develop a system that consists of three components: a data acquisition system, a price forecasting algorithm, and a performance analysis. Due to its adaptive learning capability, a conventional artificial neural network (ANN) with a back-propagation network is compared to the proposed ANFIS. In the ANFIS, qualitative fuzzy logic approximation as well as adaptive neural network capabilities are included. Using ANFIS as an expert system in predicting used car prices showed better results in the experiment. Using GUI, the consumer can

get accurate and convenient information about used cars' purchasing prices, and experiments proved that the proposed system could provide accurate and convenient price forecasting. Hence, from all literature review it is concluded that used cars price prediction is an important topic which is the area of many researchers nowadays. So far, the best achieved accuracy is 83.63% on kaggle’s dataset using random forest technique. The researchers have tested multiple regressors and final model is regression model using linear regression

Method :

This kind of subject may be evaluated using mathematical models created from quantitative data. By evaluating the contribution of each independent variable to the determination of the dependent variable, a multiple variable regression may be used to interpret the data (in this case, resale value).

Additionally, significance may be determined by looking at the p-values for each variable. The implementation of a statistical model will help support this assertion and uncover some of the key factors influencing the resale value of cars.

Data Collection :

Quantitative information will be employed in this regression. The sources of the data are what one would anticipate for information on secondhand cars. Kelly Blue Book, Edmunds, a government fuel efficiency database, and Car and Driver are the four sources that were used. Both Kelly Blue Book and Edmunds will be used as data sources, and each will provide a unique perspective on the independent variables that will be employed. These sources will cooperate to provide information on the price of a car, both new and old, as well as information on the age, mileage, manufacture, condition, miles per gallon, safety ratings, and hybrid technology. With the aid of these variables, a regression can be performed and an equation may be calculated.

Expected Outcomes :

Before I can make predictions regarding the influence each variable will have on resale value, a review of prior research and

literature is appropriate. This will allow me to make a more confident prediction as well as confirm which variables are needed to produce a strong equation that explains much of the variations in vehicle depreciation.

An expected equation could look like this: Resale Value (DV) = Intercept- B3(Age) - B4(Mileage) + Bl(Make) + B2(MPG) + B5(Hybrid Tech)

REFERENCES

[1] Pudaruth, Sameerchand. "Predicting the price of used cars using machine learning techniques." Int. J. Inf. Comput. Technol 4, no. 7 (2014): 753-764.

[2] Monburinon, Nitis, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, and Pitchayakit Boonpou. "Prediction of prices for used car by using regression models." In 2018 5th International Conference on Business and Industrial Research (ICBIR), pp. 115-119. IEEE, 2018.

[3] Gegic, Enis, BecirIsakovic, Dino Keco, Zerina Masetic, and Jasmin Kevric. "Car price prediction using machine learning techniques." TEM Journal 8, no. 1 (2019): 113.

[4] Noor, Kanwal, and Sadaqat Jan. "Vehicle price prediction system using machine learning techniques." International Journal of Computer Applications 167, no. 9 (2017): 27-31.

[5] https://ieeexplore.ieee.org/Xplore/home.jsp

[6]https://www.analyticsvidhya.com/blog/2018/08/k nearestneighbor-introduction-regression-python/

[7] https://machinelearningmastery.com/k-fold-cross-validation