SMART ASSISTANT

Used Vehicle Price Prediction System for Sri Lanka

NAME : R.D.SILVA

INDEX NO : 18/ENG/105

REG NO : EN91445

Content

Table of Figures

1. INTRODUCTION

Sri Lanka is a country where vehicle manufacturing is not done inside. Thus, when importing brand new vehicles from other countries an extra amount in other words, a tax has to pay to the manufacturer as the buyer is not living in the same country according to the government regulations. This tax amount is not the same value for all the vehicles as it changes from vehicle to vehicle. Thus, the tax depends on the vehicle that import. In addition to the tax if the raw materials that use to design a brand vehicle gets increase or the inflation rates in the manufacturing country gets increase, the price of a brand-new vehicle also gets increased. As a result of that the price of a brand-new vehicles also gets increased heavily when importing them. This is a serious problem that Sri Lanka have been facing during past few years. Thus, the worst case is the decrement of the rupee during the past decade in Sri Lanka. This has been a crisis for the country and it mainly affects to the goods which are importing from other countries. Brand-new vehicles are the one of the most popular and most imported good in the past several years and the price of those climbed as a mountain with the increased tax rates by the government as a result of the crisis mentioned above. This will clearly target mainly two part of people in the country who are, middle class level people and lower-class level people. Thus, with the crisis mentioned above, the dream of buying a brand-new vehicle will be cracked up for middle-class and lower-class level people in Sri Lanka as they cannot afford such amount for a brand-new vehicle. Therefore, the only option they left is to go with a used vehicle which they can afford if they need a vehicle. It is the best and mostly suitable option in Sri Lanka for nowadays.

Still there exists many problems when buying a used vehicle. The biggest and most heard problem is the cheating people when buying a used vehicle. In social media these kinds of news are trending in past few years and the reason for that is fixing unrealistic prices for the used cars. When a person selling a used vehicle, the seller of the vehicle can either be the person who used the vehicle or a third party. The price of a vehicle varying upon different conditions and situations of the vehicle. Most of the time, those sellers trap customers when buying a used vehicle buying listing higher price for a vehicle which not much worth. Unfortunately, many people have fallen into those traps and pay for a used vehicle which is not worth. For instance, a used Suzuki Wagon R which has latest technology and many features was opted to sell at a price more than a brand-new Toyota Vitz car. Those car sellers or the third-party people provokes buyers to buy their vehicle rather than a vehicle for a fair price to just to trap the buyer and get more profit. Sometimes they

hid some problems of the vehicle and anyhow provoke buyers to buy their vehicle. This a serious issue which needs to consider when it comes to buying a used vehicle. Furthermore, when a user tries to sell his/her used vehicle, most of the time they contact an agent or a vehicle dealer who is the third party mentioned above to get the work done easier. At the end of the day when the vehicle is sold, vehicle owner has to pay an extra amount to that guy for the job done by him. Vehicle owners do not need to do that if he/she knows what the best and suitable price is for selling the vehicle, but the problem is most of them doesn't have some knowledge to predict the price without any proofs. Sometimes that agent or the dealer can be culprit who traps buyers to list unrealistic price for a used vehicle as mentioned above. Thus, if a vehicle owner doesn't contact an agent for selling his/her vehicle and assume he/she has some knowledge about vehicles. Still in the owners has doubts when selling their vehicle as the price of selling can be overestimated or underestimated. If the selling price is underestimated, then the seller will suffer a loss and if it is overestimated, then the seller won't be able sell the vehicle as buyers won't buy a vehicle which is not much worth.

'Smart Assistant' is the best solution for all the problems discussed in above. It is a web application which predicts the price of a used vehicle in Sri Lanka more accurately. The web application accepts some essential features of a vehicle as inputs such as model, brand name, mileage driven, fuel type, transmission type, capacity, year and provide the predicted the price as the output to the user. The base foundation for this system will be designed using supervised machine learning techniques such as Linear Regression, Random Forest, Decision Tree and Naive Bayes. Then by creating several models best one which has highest accuracy will be exported to a flask server and then through the web application, user can view the predicted price of the vehicle. The data set for creating the models will be taken from ikman.lk which is the best and most popular web site in Sri Lanka for selling ad buying used vehicles. Thus, for this project a data set that has been uploaded to Kaggle web site will be taken which has around 30,500 data. In addition to that web scrapping will be done using some techniques such as Microsoft Power Automate to extract data directly from ikman.lk rather than waiting for a data set because since this is a price prediction system, it needs to be changed at least once a week as the price is a factor that changes from day to day. Thus, for that purpose web scrapping can be done to modify the model with up-to date data set and for this system both the data set and web scrapping will be used to create the best model.

2. BACKGROUND

Price prediction is a vast research topic that has been studied by many around the world. Unfortunately, in Sri Lanka there has not been a better system for this issue. Researchers more often predict prices of products using some previous data. Pudaruth predicted prices of cars in Mauritius and these cars were not new rather second hand [1]. He used multiple linear regression, k-nearest neighbours, Naïve Bayes and decision trees algorithms 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 the decision tree algorithm and Naive Bayes method were unable to classify and predict numeric values. Pudaruth's research also concluded that the limited number of instances in the does not offer high prediction accuracies [1].

The multivariate regression model helps in classifying and predicting values of numeric format. In [2], it shows how to use this multivariate regression model to predict the 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. The author of the article [2] did the same car price prediction and introduced variable selection techniques which helped in finding which variables are more relevant for inclusion in the model.

In 2019, Pal et al [3] discovered a methodology for predicting used cars prices using Random Forest. The paper evaluated used car price prediction using Kaggle data set which gave an accuracy of 83.62% for test data and 95% for train-data. The most relevant features used for this prediction were price, kilometer, brand, and vehicle type and identified by filtering out outliers and irrelevant features of the data set. Being a sophisticated model, Random Forest provided good accuracy in comparison to prior work using these data sets. In [4] it shows that, to build a model for predicting the price of used cars in Bosnia and Herzegovina. They have applied three machine learning techniques namely Artificial Neural Network, Support Vector Machine and Random Forest. However, the mentioned techniques were applied to work as an ensemble. The data used for the prediction was collected from the web portal autopijaca.ba using web scraper that was written in PHP programming language. Respective performances of different algorithms were then compared to find one that best suits the available. The final prediction model was integrated into a Java application. Furthermore, the model was evaluated using test data and the accuracy of 87.38% was obtained.

In 2019, Dholiya et al have presented an automobile resales system using Machine Learning [5]. The fair idea of what the vehicle could cost them. The system is a web application which could also provide the user with a list of choices of different types of cars based on the details of the car the user is looking for. It helps give the buyer/seller with substantial information based on which they can make the decision. This system uses Multiple Linear Regression as the algorithm to make predictions and this model has been trained using historical data that was gathered over a long period of time. Based on the KDD (Knowledge Discovery in Databases) process, the raw data was first collected. It was then preprocessed and was cleaned in order to find useful patterns useful in order to make some meaning out of those patterns. This data was then used to train the model using Multiple Linear Regression in Java as well as in Python.

Gonggi proposed a new model based on artificial neural networks to forecast the residual value of private used cars [6]. The main features used in this study were mileage, manufacturer and estimated useful life. The model was optimized to handle nonlinear relationships which cannot be done with simple linear regression methods. It was found that this model was reasonably accurate in predicting the residual value of used cars.

Listiani presented another similar research that uses Support Vector Machines (SVM) to predict the prices of leased cars [7]. This research showed that SVM is far more accurate in predicting prices as compared to the multiple linear regression when a very large data set is available. SVM also handles high dimensional data better and avoids both the under-fitting and over-fitting issues. Genetic algorithm is used 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.

Thus, there exist some drawbacks of existing systems mentioned in above. Some of them are, Most of researches haven’t used a recent data set, Almost all the systems just showed the predicted the price at the end and not some other related pr recommended vehicles which has a similar price with the predicted price, Almost all the researches used several models to build the system and selected the best one which has the highest accuracy but with lesser number of data which is an issue when it comes to accuracy and some researches just used a single model with recent and larger data set but still accuracy issue comes as using one model it can’t be concluded that is the best model for the system.

3. AIMS & OBJECTIVES

'Smart Assistant' answer for all the problems and drawbacks discussed in above. There are mainly 4 aims and its objectives can be presented in 'Smart Assistant'.

3.1 To give a fair idea for the user of what factors could affect the price when buying a used vehicle

Once user entered the inputs to the web application which are some essential features of a vehicle such as model, brand name, mileage driven, fuel type, transmission type, capacity, year, output will be displayed. Then user can change those values and see what are the factors that change the price of a used vehicle and can get some knowledge about it when buying one. Thus, to achieve the aim as the objective, multiple linear regression machine learning model with other supervised learning techniques will be built to predict the price of a used vehicle for given inputs more accurately when buying or selling.

3.2 To enable user to search vehicle prices without any skepticism

As mentioned in previous sections user can still have some doubts when buying or selling a used vehicle. When selling used vehicle, price can be overestimated or overestimated and when buying a used vehicle user can think whether the price of the vehicle is too much for the features it contains. The accuracy is the best answer for the doubts that user has and to achieve this, several models will be built using supervised machine learning techniques and the best one will be selected which has the highest accuracy.

3.3 To find available vehicle prices accurately easily at anytime

For this purpose, web application will be designed as the end product which user can access freely at anytime from anywhere.

3.4 To give an idea for the user about alternative vehicles that is similar to the predicted price

As a main drawback of the recent studies is that they only showed the predicted price as the output to the user. It will better if user can get to know what are the other alternative or related vehicles that they can go with the predicted price. For that purpose, this web application will display some alternative or recommended vehicles that has a similar price with the predicted price in some range when buying one so that user can see what are the other vehicles that they can buy with that predicted price. It is a great advantage for the user as they can even change their mind by viewing those and can go for a better vehicle than predicted one.

4. IMPLEMENTATION PLAN

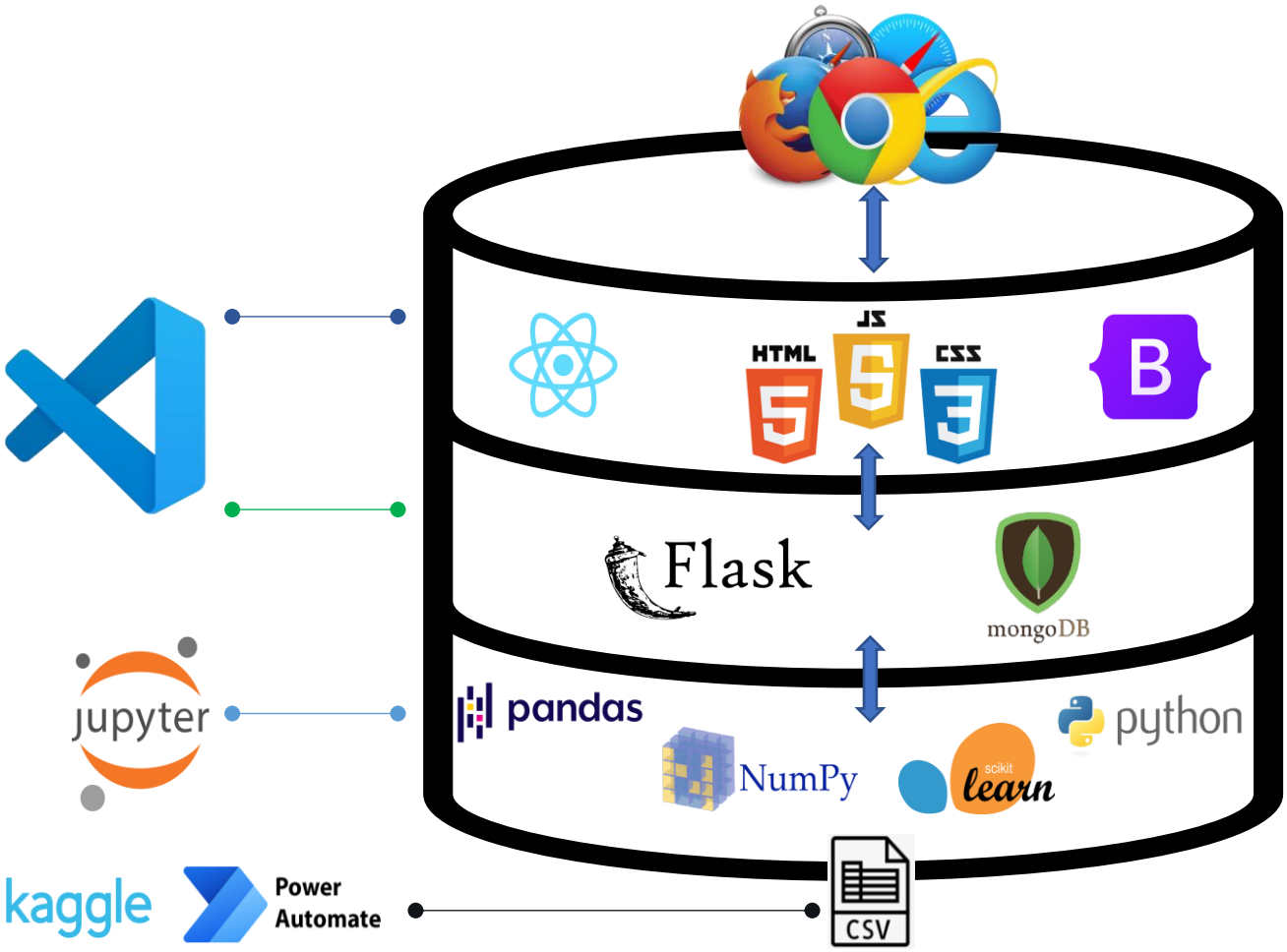

Figure 1: Technology stack

The above figure illustrates the technologies and tool that will be used in the implementation. As the first step data set is required and there exist several methods get a data set. For this project a data set which was recently uploaded to Kaggle web site will be used with some web scrapping with the favor with Microsoft Power Automate tool. The main reason for using that tool is as mentioned above this is price prediction system and the price of a vehicle in Sri Lanka changes from day to day due to several reasons. if the data set used for building the model is not an up-to date one in other words a recent data set, then there is no use of building a model because the predicted price won't be accurate. Thus, web scrapping method is the best option to get a recent data set and to update the model as the developer I have to be active in ikman.lk website or the best option is I need to contact a person who knows when will the prices of vehicles will get changed and how they change. In addition to that, historical data of vehicle prices also can be

viewed from ikman.lk website to get more accurate prediction. Once the data set was selected, it will be feed to the Jupyter notebook and with the help of Python and its libraries such as pandas, NumPy and scikit learn displayed in the above figure, several models will be built. The best model among built will be chosen by considering the accuracy of each model and it will be exported to the Python Flask server. Then after the web application is designed with React JS, Bootstrap, HTML, CSS and JS those two parts can be merged. Once it is done user can use the web application to predict the price of a used vehicle in buying or selling. In addition to that the inputs that the user entering will be stored in a database to future developments such as when modifying or rebuilding the model with recent data set, those stored data can be analyzed and even more they can be used to rebuild the model as those data doesn't need to preprocess again or transform and it they are accurate. Thus, for that purpose MongoDB will be used.

5. MILESTONES & TIMELINE

5.1 Milestones

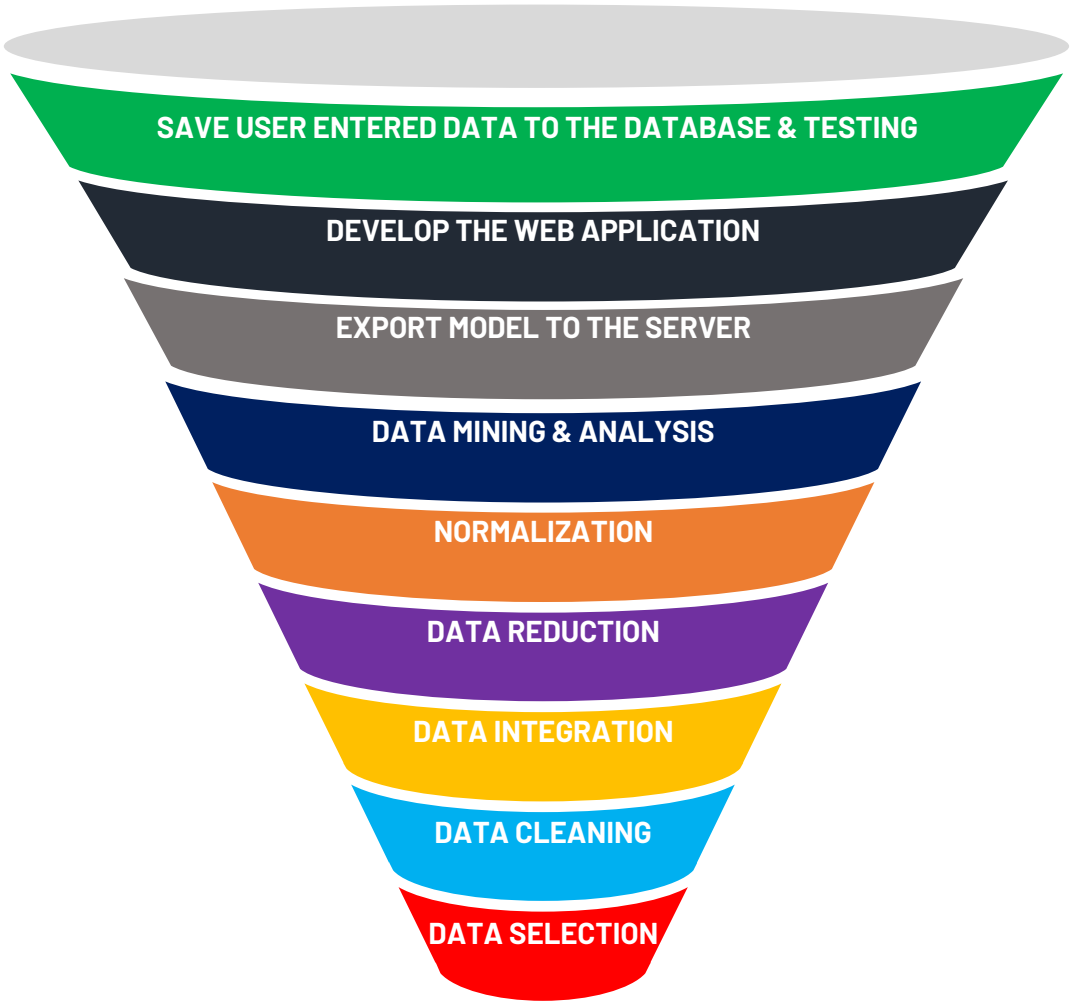

Figure 2: Milestones

Milestone of each stage is as below.

- Data Selection: A suitable data set needs to be selected from a reliable source. For that purpose, ikman.lk will be used as the website as it is the most popular and most used website in Sri Lanka for selling and buying used vehicles. For this project a data set was taken from Kaggle website which uploaded a month ago and it has 30,500 data. In addition to that web scrapping will be used as this model needs to be modify at least once a month.
- Data Cleaning: As the first step of data preprocessing, data cleaning will be done as the data set contains null values for some attributes. They need to be neglected or replaced by values using some algorithms.

- Data Integration: Some attributes have values with different data formats. They will be converted to single format.
- Data Reduction: As the final step of data preprocessing, data reduction will be done. The original data set contains 20 attributes and most of them are not needed to build a model. Thus, they will be neglected when building the model.
- Normalization: In data transformation, normalization will be done because in the data set, the price of vehicles has a large range. To increase the accuracy and to decrease the time consume for building the model, the price attribute will be scaled between 0 to 1 using algorithms.
- Data Mining & Analysis: Once the above steps are done, several models will be built using supervised machine learning techniques and the best one will be selected by comparing the accuracy.
- Export Model to the Server: Once the best model is selected, it will be exported to the created python flask server.
- Develop the Web Application: After the above step, a web application will be developed using the technologies mentioned above with all the features (mainly has two parts buying and selling). When designing this web application, admin login part will be included to view the user entered values when buying and selling vehicles.
- Database Configuration and Testing: As the final step, database configuration will be designed to store the user entering values and finally overall testing of the system.

5.2 Timeline

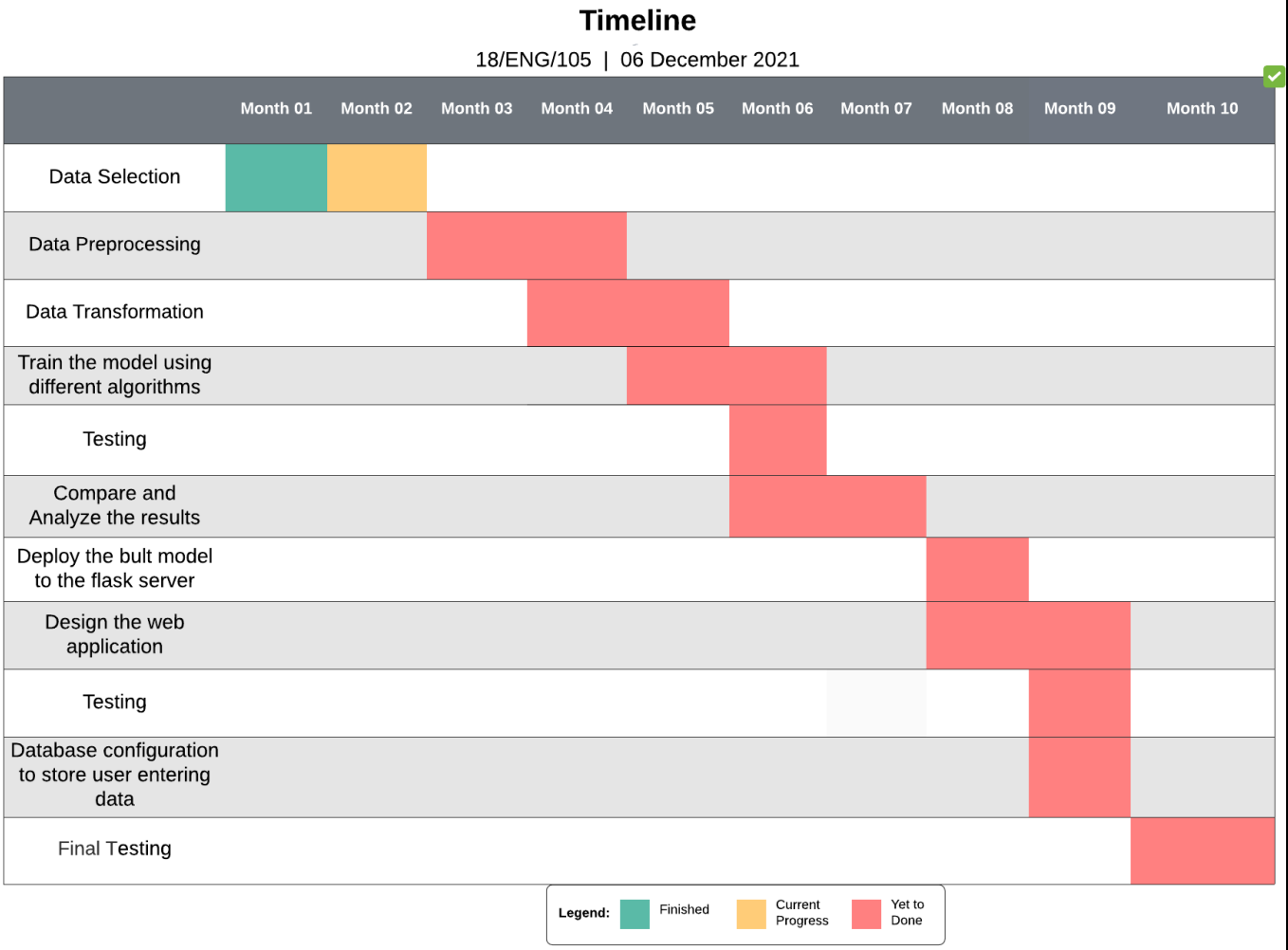

Figure 3: Gantt Chart

6. REFERENCES

[1]Pudaruth, S. (2014) 'Predicting the Price of Used Cars using Machine Learning Techniques', *International Journal of Information & Computation Technology*, 4(7), pp. 753–764. Available at: http://www.irphouse.com.

[2] Kuiper, S. (2008) 'Introduction to Multiple Regression: How Much Is Your Car Worth?', *Journal of Statistics Education*, 16(3). doi: 10.1080/10691898.2008.11889579.

[3] Pal, N. *et al.* (2019) 'How Much is my car worth? A methodology for predicting used cars' prices using random forest', *Advances in Intelligent Systems and Computing*, 886, pp. 413–422. doi: 10.1007/978-3- 030-03402-3_28.

[4] Gegic, E. *et al.* (2019) 'Car price prediction using machine learning techniques', *TEM Journal*, 8(1), pp. 113–118. doi: 10.18421/TEM81-16.

[5] Dholiya, M. *et al.* (2019) 'Automobile Resale System Using Machine Learning', *International Research Journal of Engineering and Technology(IRJET)*, 6(4), pp. 3122–3125.

[6] Gonggi, S. (2011) 'New model for residual value prediction of used cars based on BP neural network and non-linear curve fit', *Proceedings of the 3rd IEEE International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, 2(1), pp. 682–685.

[7] Listiani, M. (2009) *Support Vector Regression Analysis for Price Prediction in a Car Leasing Application*, *Technology*. Hamburg University of Technology.