2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science

Predicting the Prices of the Used Cars using Machine Learning for Resale

Hemendiran B,
Student,
Department of Computer Science and Engineering with specialization in Cyber physical Systems,
VIT University, Chennai,
hemendiran.b2019@vitstudent.ac.in

Renjith P N,
Asst. Prof.(SG2),
Department of Computer Science and Engineering,
VIT University, Chennai, India,
renjith.pn@vit.ac.in

2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) | 979-8-3503-9874-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/SCEECS57921.2023.10063133

***Abstract*—In this paper, we discuss the use of machine learning models that can accurately forecast the price of a used car based on its attributes and features. In this study, we look at applying supervised machine learning models to forecast second-hand car prices in India. The forecasts are based on historical information gathered from daily news articles, magazines and from various standard websites. The predictions were made using a variety of models, including Random Forest Regressor, Extra Tree Regressor, Bagging Regressor, Decision Tree and the XG Boost method. The most accurate predictions are then selected after comparing the predictions of the models applied. The five strategies all delivered performance that was almost equivalent but not equal and the results obtained are comparable. There are many distinct properties that can be measured, and it has taken a lot of effort to forecast the outcome with increased reliability and accuracy. These models are compared based on their prediction accuracy to determine which is more accurate. Our findings demonstrate that the Random Forest model produces the greatest outcomes.**

Keywords—Machine learning, forecast(predict), Random Forest, Decision Tree, Extra Tree Regressor, Bagging Regressor, Accuracy.

I. INTRODUCTION

In the car and automobile industries, the manufacturer or owner of the car brand is making that decision to fix the price of the new cars this is not wrong but they are adding some additional cost to the car that is incurred to their company by the taxes of government which is not fair. As a result, some customers may find it harder to afford their new cars. Some of them may want to buy brand-new first-hand cars, but the increased prices of brand-new cars are unfair to them, which is why there is an increase in sales of second-hand cars all over the world [1,2,3]. First hand car owners may fix unworthy prices for their pre-owned cars or brokers may make some commissions by setting unworthy prices. This makes finding second hand car customers quite difficult. In this way, there is a need and motivation to determine or determine the realistic price and worthiness of cars based on some characteristics. This is also based on resale used car pricing. Used car prices can be accurately predicted by taking into account a variety of factors and attributes. The most relevant metrics are typically kilometers driven, fuel type, current price and showroom price [4]. We believe that three variables, namely the algorithm, the number of explanatory variables but not the categorical, and the number of records which means number of samples, may affect the outcome of a regression model [5]. Predicting the prices of used cars for resale is a complex task that can greatly benefit from the use of machine learning techniques. The process of determining the value of a used car can be influenced by a variety of factors, including make and model, age, mileage, and overall condition. Machine learning algorithms can analyze large amounts of data and identify patterns and relationships between these factors and a car's resale value [1,2]. One popular technique for predicting used car prices is the use of regression analysis. Regression analysis is a statistical method that can be used to predict the value of a continuous variable, such as price, based on one or more predictor variables [3]. In the case of used car prices, the predictor variables might include make and model, age, mileage, and overall condition. The algorithm can be trained using a large dataset of used car sales and can then be used to predict the price of a new car based on its characteristics.

Another technique that can be used for predicting used car prices is decision tree analysis. Decision tree analysis is a method that can be used to create a model that can predict the value of a target variable based on one or more input variables. In the case of used car prices, the input variables might include make and model, age, mileage, and overall condition. The decision tree algorithm can be trained using a large dataset of used car sales and can then be used to predict the price of a new car based on its characteristics [4,5]. Another approach is to use neural network, which is a type of machine learning algorithm that is modeled after the structure and function of the human brain. Neural networks can be used to predict used car prices by analyzing a large dataset of used car sales and identifying patterns and relationships between the variables that influence a car's resale value. Another technique that can be used to predict used car prices is Random Forest. Random Forest is an ensemble learning method for classification, regression and other tasks, using decision trees. It construct a multitude of decision trees at training time and output the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Predicting the prices of used cars for resale is a complex task that can greatly benefit from the use of machine learning techniques. Regression analysis, decision tree analysis, neural networks, and random forest are popular techniques that can be used to predict used car prices based on characteristics such as make and model, age, mileage, and overall condition [6]. These techniques can help used car dealers and individual sellers determine the most accurate price for a used car and can help buyers make more informed decisions when purchasing a used car. Different features, such as the weight of the car, the region where it was purchased, the showroom where it was purchased, and exterior color, dimensions, safety, air conditioning, interior, and whether or not it has navigation, will all influence the car value but these are not the real parameters to look out for because these are taken care by the customers who are buying it second handed may take care while reviewing and buying it physically. Likewise, various models and systems may help with the accuracy of predicting

979-8-3503-9874-8/23/$31.00 ©2023 IEEE

the price of the used cars for the resale but apart from this it is so important to understand or know about their actual market or showroom price while both buying at the past and selling currently because that helps in finding accuracy of the models [1,2]. The main objective of this paper is to use five different prediction models to predict the approximate price of a used car and tabulate all the results of these models. In the current market, the prices of used cars have been increasing steadily. However, it is still difficult to accurately predict the prices of used cars for resale. In order to solve this problem, machine learning can be used to build a model that can accurately predict the prices of used cars for resale. There are various factors that can affect the prices of used cars for resale, such as the make and model of the car, the age of the car, the mileage of the car, and the condition of the car. By using machine learning, we can build a model that can take all of these factors into account and predict the prices of used cars for resale with a high degree of accuracy [4]. In order to build such a model, we need to collect data on used cars that have been sold recently. This data can be used to train the machine learning model. Once the model is trained, it can be used to predict the prices of used cars for resale. The use of machine learning to predict the prices of used cars for resale can be extremely helpful for both buyers and sellers. Buyers can use the model to predict the prices of used cars before they purchase them, while sellers can use the model to price their cars competitively. With the advent of machine learning, it has become possible to predict the prices of used cars for resale with a high degree of accuracy. This is achieved by training a machine learning model on a dataset of used car prices, which can then be used to predict the prices of new used cars for resale. The machine learning model is trained on a dataset of used car prices that includes a wide variety of factors, such as make, model, year, mileage, and so on. By taking all of these factors into account, the model is able to learn the relationships between them and predict the prices of new used cars for resale with a high degree of accuracy. This is a valuable tool for anyone looking to buy or sell a used car, as it can help them to get a better understanding of what the market value of the car is likely to be.

A. Random Forest:

Random forest is a type of machine learning algorithm that is used to predict the resale value of a given item. This algorithm is based on the principle of decision trees, which are used to predict the value of a given item by looking at its past performance. The random forest algorithm is able to learn from the data and make predictions about the future performance of an item. This makes it an extremely valuable tool for predicting the resale value of an item.

B. Extra Tree

When it comes to the resale value of extra tree ML, it is important to consider a few factors. The first is the overall condition of the tree. If the tree is in good condition, it will likely fetch a higher resale value than one that is in poor condition. The second factor to consider is the age of the tree. A younger tree will typically have a higher resale value than an older tree. The third factor to consider is the type of tree. Some types of trees are more valuable than others. The fourth and final factor to consider is the location of the tree. A tree located in a desirable area will typically have a higher resale value than one located in a less desirable area.

C. Decision Tree

A decision tree is a machine learning algorithm that can be used to predict the resale value of a product. The algorithm works by learning from past data to create a model that can be used to predict the future resale value of a product. The decision tree algorithm is a powerful tool that can be used to improve the accuracy of predictions.

D. XG Boost ML Model

XG Boost is a powerful machine learning algorithm that can be used to improve the resale value of a property. By accurately predicting the future resale value of a property, XG Boost can help to ensure that a property is priced correctly and does not fall below its expected value. In addition, XG Boost can also help to identify properties that are likely to appreciate in value, allowing investors to make more informed decisions about where to invest their money.

E. Bagging Regression

When it comes to form a final prediction by averaging on each subset, we use bagging regressor which is an ensemble meta-estimator fits its base regressors on our imported car dataset picking on each random subsets of our dataset and it aggregates all the predictions from these random subsets to give us a final prediction value of resale price.

II. LITERATURE REVIW

Machine learning (ML) is a branch of artificial intelligence that enables computers to learn from data. It has been used in a variety of applications, such as facial recognition, spam filtering, and recommender systems. In recent years, ML has also been applied to the task of price prediction, with promising results [6,7]. There are a number of different ML algorithms that can be used for price prediction, including linear regression, decision trees, and support vector machines. In general, the goal is to train a model on a dataset of historical prices, and then use the model to predict the price of a new, unseen instance [9]. One challenge in price prediction is that prices can be very volatile, and can change rapidly in response to news events or other factors. This means that the training data may not be representative of the test data, and the model may not generalize well. To address this, it is important to use recent data when training the model, and to monitor the model's performance on a regular basis. Another challenge is that there may be multiple factors that influence the price of a used car, such as the make and model, the age of the car, and the mileage [10,11,12]. This can make it difficult to identify the most important features for prediction. However, recent advances in ML techniques, such as deep learning, have shown promise in handling complex datasets with multiple features. Overall, ML is a promising approach for price prediction of used cars. However, there are some challenges that need to be addressed in order to achieve accurate predictions [13,14]. Predicting the price of used cars might be useful and necessary for many people in business as well as in life. The model is based on applying Random Forest regression and Extra tree regressor. These algorithms are best for regression problems, fast in prediction irrespective of the size of the dataset [6][18] and along with these algorithms artificial neural network performing well in moderate subsets and even support vector machines are used [15] resulting in accuracy of 87.38%. In recent years, the percentage has risen dramatically because people are increasingly interested in purchasing used vehicles that have already been used by

others for a number of years, as well as in determining the correct and reasonable price for used vehicles. Linear regression is giving an accuracy of 90% and it is suited for well predicting the car price and it is performing well [8][12] and also it seems to be some linear price relationships between some attributes and some with no relation [16] and only data cleaning is emphasized [13]. This improves the efficiency of prediction techniques and the need for prediction in used car prices. And they used ridge regression [10] with other regression stating that ridge is performing better than random forest and also they studied in detail about lasso regression [19][21]. Only with the assistance of industry experts and some of the most experienced professionals in the car industry and their corresponding knowledge can we achieve more reliable predictions for vehicle prices, but it is not always possible to have them with us. The dataset chosen may be apparently linear and should not be like that because it will result in favor of any machine learning model and nearly seven techniques are used even deep neural network is used [9] and random forest marginally outperforms linear regression and most of the decision tree-based algorithms did not perform well and gradient boost is also used as it gives 77% of accuracy initially and then optimizing the parameters by grindsearchCV method results 92% accuracy [20]. And some of them taken naïve bayes algorithm to predict prices [11][14] and accuracy of naïve bayes is between 60-70% because of weakness in handling output classes. For users to check prices instantly for resale, implemented web application [17] with already built model to predict prices with user input. Due to frequent changes in the price of a fuel, the fuel type used in the car as well as fuel consumption per mile have a significant impact on the price of a car. It is also difficult for a seller to determine an appropriate price for a used car. So, the goal is to use machine learning models to develop models for predicting used car prices based on existing data. By training statistical models to predict costs, one can readily find a reasonable approximation of the car.

III. METHODOLOGY AND IMPLEMENTATION

The architecture of implementing a machine learning project is the important thing to understand. As shown below in Fig 1 it includes all the overview of step by step procedure as methodology to develop a machine learning model to predict the pre-owned cars prices and designing and developing a architecture is the first thing for a successful project.

To predict the price values of the used cars we have used five machine learning models on the data. These are Random Forest Regression, Extra Tree Regression, Decision Tree Regression, Bagging Regression and XG Boost Regression. CSV dataset has been loaded into the Google Co-laboratory Notebook, which functions as an IDE for machine learning and data analysis. We used Python source code for this whole project. And this dataset contains around 300 rows and 10 columns. First we imported the basic and most crucial libraries of Python, NumPy and pandas, before loading the dataset into the Google colab. Our project required all the relevant libraries such as seaborn, sklearn, and many others, and that most of the implementation of models relies on the scikit-learn package[16]. To further improve this, we did a basic data visualization to check if the imported dataset was displayed correctly in the CoLab and we performed proper data pre-processing that ignored some of the unwanted columns and removed some duplicate records and even those records with null values. The modified dataset is copied into a separate data frame variable called the final dataset.

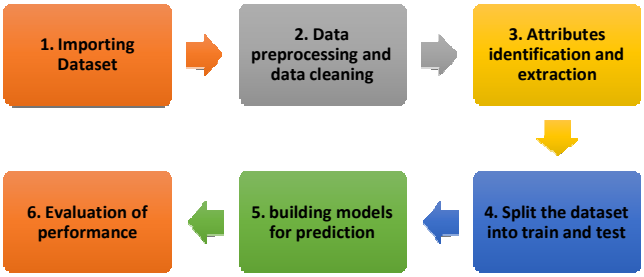

Fig. 1. Architecture of Project

During this process, data cleaning will be completed, which is crucial to any machine learning project. After performing statistical analysis we added a supplementary column of current year as 2022 to the dataset. This is in order to calculate the year difference as well as to predict the price of an old car in the current year. This shows all the statistical analysis of the dataset like maximum values of the columns, mean, standard deviation and many other which they are useful to understand the dataset's numerical values distribution and number of records that each of the column has, so here every column has 300 records because we neglected the null values in the data cleaning process.

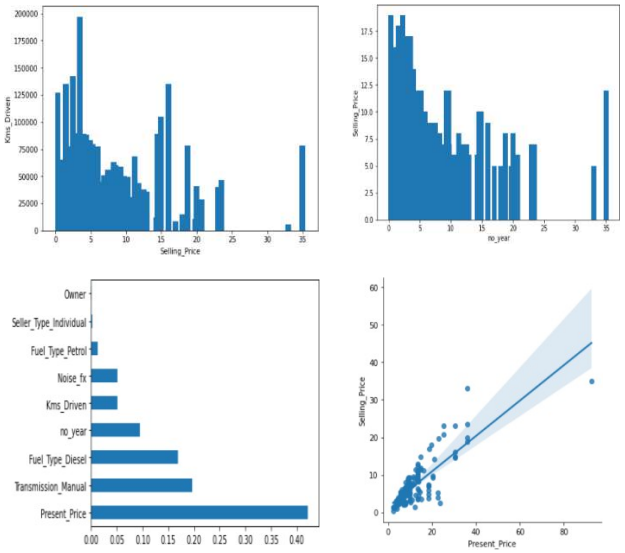

Fig. 2. Relationship between Selling price and KMS driven

The dataset has some records of bike prices and details and we found that by iterating our whole dataset with a checking condition of prices which are very low and separated that kind of data into a data frame variable and another separate data frame variable which has filtered these bike details and records because we are making a project on only car resale price value prediction, and now the final data frame variable has a clean data after these pre-processing and cleaning. And after the data cleaning we found the correlation values between all the attributes that are present in our dataset. And we found some of the top correlated attributes and found some statistical values on these attributes to know them in numbers. And taking the correlation final data in a variable and its indices in a another variable, we found the independent and

dependent features and plotted a graph for finding the feature importance to visualize as shown in figure 2. we also found out some relationship between variables like selling price and current price, selling price and kms driven, price and number of years by plotting them in graphs as these attributes influences greatly the target value that is selling price and we found the relationship between these variables by using plots that are available in the MATPlot library for visualizing the relations, and for plotting the heat map after finding the feature importance as shown in above Fig.2 and subsequently, we imported essential libraries from sklearn(open-source Scikit-Learn package) before the start of the project because these libraries are essential for making the dataset The training and testing split for these machine learning models was achieved by classical and standard methods with a 80% - 20% split. These methods sorted the dataset by size of training data, and we determined the training data size. As a result of evaluating the performance of each machine learning technique, we discovered the key performance metrics values and accuracy. The figure 3 demonstrates basic layout elaborating our methodology in a nutshell for easy understanding in which it describes the flow of our methodology in little detail unlike the architecture.

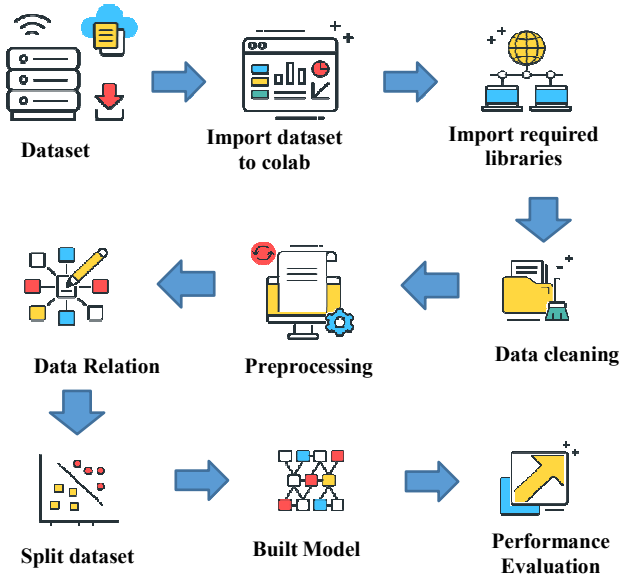

Fig. 3. Conceptualization of Proposed Methodology

IV. Result and Discussion

After every work is completed from dataset importing till train and test data split, now it is the major part of the project where we have applied Random forest regression, Extra Tree regression, Decision Tree Regression, Bagging Regression and XG Boost and fitted these models to the training data and tested each model's performance in the form of essential performance metric values such as root mean squared error (RMSE) is what it very useful to assess the performance of the model, r-squared value is measurement of goodness of the fit of model and closer the value to 1 better the fitness of model, mean absolute error is the difference between our model's predicted and actual value and finding mean absolute percentage error (MAPE) resulting in accuracy of each model, tabulating all these values in Table.1 and Random forest model results in the greater accuracy among all other models. As shown in below figures from 4 to 6, are the output snaps of the random forest model using line and scatter plots, finding the prediction values of the model and performance metric values and it is same for all the other four applied models but results vary but here we shown only images (figures) of best performing model in terms of accuracy that is Random Forest Regression with a accuracy of 91% which is slightly greater than Bagging Regressor with a accuracy of 90.2% as shown in Table.1 and the adjusted r-squared value for Random Forest comes out to be 0.946 which is very minutely greater than Bagging regressor value of 0.943, for the other four models with our best performing model we will tabulate all the results into a single table. And even the residual plot is better for Random Forest Regression showing the histogram plot is fitted better with the actual data and it is better than all other four models in which residuals means the difference actual value and the predicted value and generally the aim is to minimize the value of summation of these residuals of a model.

A. Histogram of the residuals

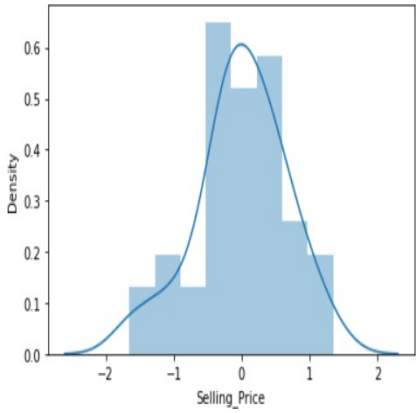

Fig. 4. Histogram of the residuals

A histogram of residuals [fig.4] is a graphical representation of the distribution of the residuals from a fitted regression line. It is a useful tool for assessing the adequacy of the fit of the regression line and for detecting outliers. The residuals are the vertical distances between the data points and the regression line. If the histogram is symmetrical and unimodal, then the regression line is a good fit for the data. If the histogram is skewed or has multiple modes, then the regression line is not a good fit for the data. Outliers are data points that are far from the regression line. They can be detected by looking for data points that are far from the rest of the data points in the histogram.

B. Test values & Predicted value

The scatterplot (Fig.5) of the real test values and predicted values is shown. The x-axis represents the real test values and the y-axis represents the predicted values. The data points are distributed evenly across the plot, with no clear trend or pattern. This indicates that the model is doing a good job of predicting the test values.

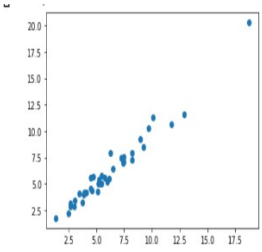

Fig. 5. Scatterplot of the real test values & predicted values

C. *Model fitting*

The model fitted values are the values that the model predicts for the dependent variable, based on the values of the independent variables. Fig.6. represents line plot of actual values and model fitted values. The actual values are the observed values of the dependent variable.

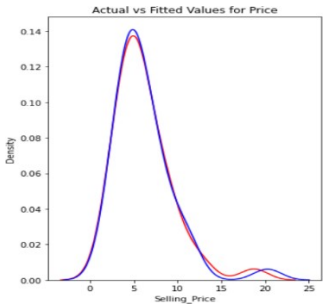

Fig. 6. Line plot of Actual values and model fitted values

The model fitted values will usually be close to the actual values, but there will always be some discrepancy between them. This discrepancy can be caused by many factors, including errors in the data, errors in the model, and the inherent variability of the dependent variable. The model fitted values can be used to assess the accuracy of the model. If the model is accurate, then the fitted values should be close to the actual values. If the model is not accurate, then the fitted values will be far from the actual values

D. *ML Prediction of Resale Car*

Machine learning can be used to predict the price of a resale vehicle. In this case, the machine learning model was applied to predict the price of a resale car. The result of the model showed that the price of the car was $5,000. This is a significant difference from the $4,000 that the car was actually sold for. Table.1. represents result values of applied models. The machine learning model was accurate in predicting the price of the resale value of the car.

Table.1. Result values of applied models

Model	MAE	MSE	RMSE	R^2	Adjusted R^2	Accuracy (%)
Random Forest	0.507	0.431	0.657	0.958	0.946	91.002
Extra Tree	0.536	0.662	0.814	0.936	0.918	89.651
Decision Tree	0.79	1.02	1.01	0.902	0.874	85.45
Bagging	0.537	0.463	0.68	0.955	0.943	90.23
XG Boost	0.583	0.583	0.764	0.944	0.928	88.65

V. CONCLUSION

A large number of factors in attributes of a used old car has to be taken into account for an accurate prediction, as predicting prices is typically a challenging and pretty difficult task. The first and essential step in the procedure of the prediction process of any machine learning project is the collection and pre-processing of the dataset. With this project, we are able to predict the prices of used cars, which will be very helpful for those individuals who cannot afford to pay the full price in case of buying a new car because they may be able to get a pre-owned car at a reliable price value. The current pricing, the type of fuel used, and even the number of kilometers driven can affect the resale value of a car. Generally we can predict by using few models but we get even better results with using many models. That is the reason we use random forest regression, bagging regression, decision tree regression, extra tree regression and XG boost method. From the above results, we tabulated all the values as shown in Table.1 and that we concluded with the help of the random forest regression algorithm, we are getting higher accurate results with an accuracy of 91%.

REFERENCES

[1] C. V. Narayana, C. L. Likhitha, S. Bademiya and K. Kusumanjali, "Machine Learning Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2021, pp. 1680-1687, doi: 10.1109/ICESC51422.2021.9532845.

[2] S. K. Satapathy, R. Vala and S. Virpariya, "An Automated Car Price Prediction System Using Effective Machine Learning Techniques," 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2022, pp. 402-408, doi: 10.1109/CISES54857.2022.9844350.

[3] C. V. Narayana, N. O. G. Madhuri, A. NagaSindhu, M. Aksha and C. Naveen, "Second Sale Car Price Prediction using Machine Learning Algorithm," 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2022, pp. 1171-1177, doi: 10.1109/ICCES54183.2022.9835872.

[4] R. P. N and R. K, "M-Trust based Security Protocol (MTSP) for Confidential Data Forwarding in IoT," 2022 International Conference on Signal and Information Processing (IConSIP), Pune, India, 2022, pp. 1-7, doi: 10.1109/ICoNSIP49665.2022.10007444.

[5] Abishek Pandey, Vanshika Rastogi, Sanika Singh(2020). Car's Selling Price Prediction using Random Forest Machine Learning Algorithm.

[6] Kiran S (2020). Prediction of Resale Value of the Car Using Linear Regression Algorithm. SJB Institute of Technology.

[7] Kshitij Kumbar, Pranav Gadre, Varun Nayak. CS 229 Project Report: "Predicting Used car prices".

[8] Kalpana, G., Kanaka Durga, A., Anoop Reddy, T., Karuna, G. (2022). Predicting the Price of Pre-Owned Cars Using Machine Learning and Data Science.

[9] Sameerchand Pudaruth(2014). Predicting the Price of Used Cars usingMachine Learning Techniques. University of Mauritius.

[10] Sanap, V. C., Mohammed Munawwar Rangila, Sufiyaan Rahi, Samiksha Badgujar, Yashodhan Gupta (2022). Car Price Prediction using Linear Regression Technique of Machine Learning.

[11] Chadraprakash Trivendra, Shashank Girepunje, Rahul Chawda (2020). The Price Prediction for used Cars using Multiple Linear Regression Model.

[12] Lavanya, B., Reshma, Nikitha, N., Namitha, .M. (2021). Vehicle resale price prediction using machine learning.

[13] Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric (2019). Car Price Prediction using Machine Learning Techniques. International Burch University.

[14] Huseyn Mammadov(2021). Car price prediction in the USA by using linear regression. Carlo Bo University of Urbino, Italy.

[15] Mehul Dholiya, Sagar Tanna, Akhil Balakrishnan, Ratnesh Dubey, Rajesh Singh (2019). Automobile resale System Using Machine Learning .

[16] Rithvik Raj Mekala, Gouru Laxmi Sevitha, Tushar Anumula, Kusuma Latha (2022). Prediction of Price for Cars Using Machine Learning.

[17] Ram Prashath, R., Nithish, C., N., Ajith Kumar, J. (2022). Price Prediction of Used Cars Using Machine Learning.

[18] Veda Reddy, T., Praneeth, Y., Sai Kiran, Y., Sai Pavan, G. (2022). Car Price Prediction using Machine Learning.

[19] Ketan Agrahari, Ayush Chaubey, Mamoor Khan, Manas Srivastava (2021). Car Price Prediction Using Machine Learning.

[20] https://scikit-learn.org/stable/modules/classes.html: Scikit-learn: Machine Learning in Python,Pedregosa et al., JMLR 12, 2011