2022 7th International Conference on Communication and Electronics Systems (ICCES) | 978-1-6654-9634-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCES54183.2022.9835872

Second Sale Car Price Prediction using Machine Learning Algorithm

Chejarla Venkata Narayana , Nukathoti Ooha Gnana Madhuri , Atmakuri NagaSindhu,
Mulupuri Aksha , Chalavadi Naveen
Department of Computer Science and Engineering,
Lakireddy Bali Reddy college of Engineering,
Mylavaram, 521230, Andhra Pradesh, India
Email: cvnreddy.chejarla@gmail.com , oohagnanamadhuri@gmail.com , sindhuatmakuri56@gmail.com
mulupuriakshaa@gmail.com , naveenchalavadi00@gmail.com

Abstract -- **Every business firm recognizes the need of making sound and challenging decisions. Poor decisions can lead to substantial losses and even the demise of a firm. This paper is focused on one of the retail enterprises, which deals with the used car sales. The major goal is to develop a prediction model that can estimate the selling price of used cars based on key factors. Machine learning techniques such as Random Forest Regression, Feature engineering technique such as Extra Trees Regression are employed to accomplish the goal as Random Forest Regression is modeled for prediction analysis and Extra Trees Regression fits the number of decision trees. The results are so encouraging with our approach.**

Keywords: Hyperparameter Tuning, Categorical data, RandomisedSearchCV, Prediction Model

I. INTRODUCTION

Nowadays the Indian Used Car Market become so exuberant, the sales figure of secondhand cars has grown to be double when comparing with the size of the new purchased cars. Automatically used car sales have become more organized in recent years. Over the last three years, it was witnessed amazing progress, and the prices are expected to rise by up to 17% or even more. To effectively determine the car's worthiness, a prudent mechanism is essential. In this paper, a prediction model is built so that it estimates the selling price of already used cars based on their features. This in turn reduces the burden and risk from the seller, consumer and also provides a positive firsthand knowledge about price and low financing cost for used cars.

OLX, Cars24, Car Dekho, and many other organizations are majorly handling the present market for used car sales. Some of the major features to consider while determining the price of used cars include the brand name, year, kilometers driven, owner and seller type, fuel type, transmission type, number of seats and steering type. The list of attributes was not at all restricted; these can be extended widely and automatically can upgrade the reliability of the model. In this paper, the approach adopted is based on renew Machine Learning algorithms such as random forest regression, to predict the selling price of a used car based on key features. Evaluation metrics are applied on the proposed method and compared the results with the existing method. Finally, it is concluded that the proposed method provides better performance with less error percentage.

II. LITERATURE SURVEY

As numerous authors worked on this study, they discovered some flaws in their research over the existing one. Let's take a look at each of the limitations one by one. In their study [1], Enis Gegic and his colleagues explored certain regression models that were aimed to predict the price of a car based on its features. One of the most significant flaws in this study is that they used more samples with limited features.

In their study [2], Syed Bademiya and their batch discussed about used car sale prices using Linear Regression algorithms but they the drawback is they used limited features and the values of the metrics were too high which impacts the accuracy.

In her Master's thesis article [3], Listian discussed the regression model constructed by them using the Support Vector Machines (SVM). In estimating the price of a leased car, these are more accurate than multivariate regression or basic multiple regression. The flaw of SVM over simple regression in terms of metrics is not quantifiable.

In Mauritius, Sameer Chand Pudaruth investigated different regression approaches for predicting car prices [4]. But there are only a few records and features in the collection. This is considered as the fault.

To forecast the expected value of private used cars Gongii [5], suggested a new methodology which uses the artificial neural networks. The model was adapted to assist the illogical relationships that are hard to estimate by using traditional linear regression methods.

In their study Wu et al [6], used fuzzy method to estimate car prices. They included the features such as mileage, brand name, year, and location. The data used is substantial, yet it lacks several key features that influence car resale value.

In their paper [7] A. Gelman and J. Hill used number of regression techniques to determine the used car price. They employed the Analytical Hierarchy method, which obtained the good and faultless observations. The drawback is that the data lacks a number of major key parameters such as transmission type, seller type, and others that could impact the used car price.

In their study [8], Nabarun Paul and his colleagues investigated Random forest regression algorithms for determining the value of a second hand car. The paper does not discuss the feature engineering tasks. This can be considered as a weakness of the paper.

In their paper [9], V. Suma used the enhanced neural network algorithm to estimate the price for each type of vehicle (car). They included the features such as the car's brand, engine type, and manufacture year. But its training time is too long to be considered as the drawback.

In their study [10], Madhuvanthi and their batch employed the linear regression technique for predicting the price of used cars. This publication includes a brief overview of data procedures. However it omits the feature of calculating significance scores and displaying linear model coefficients.

Ashutosh Datt and Vibhor Sharma [11] studied the prediction of used cars using Linear Regression. But the drawback is that this paper missed performing the feature and calculating the model coefficients.

In their study [12], Ning Sun, Hongxi Bai, Yuxia Geng, Huizhu Shi discussed the prediction of the cars based on BP Neural Network Theory. The drawback is that it requires vast data to examine the used car price.

III. PROPOSED SYSTEM

The proposed model which is shown in below Figure 1 is a combination of the two machine learning algorithms i.e. Random Forest Algorithm and Extra Trees Regression algorithm.

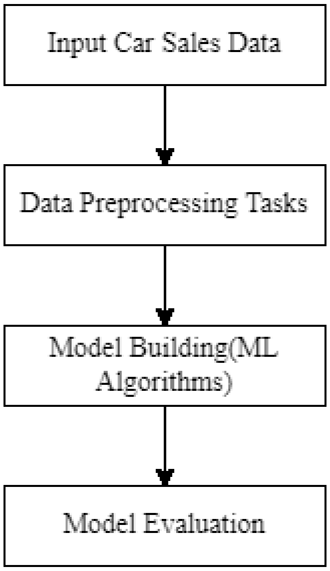


Fig 1: Over View of Proposed system

At first the dataset is loaded and it was collected from Kaggle.com. And then the dataset is divided into two segments, one is for training and the other is for testing. Then applying the preprocessing techniques that includes the identifying missing or null values and encrypting the Categorical Values. For obtaining the optimal performance, Hyperparameters are used. The two machine learning techniques namely Random Forest and Extra Trees Regression Algorithms are used, and the Hyperparameters are also effectively tuned using RandomisedSearchCV. The prediction model is tested and calculated its accuracy after the model predicts a result.

A. BLOCK DIAGRAM

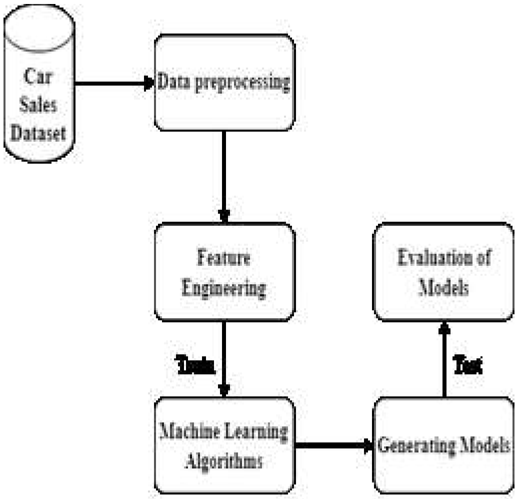


Fig 2: Block Diagram of Proposed system

- After collecting the data set preprocessing chores are done including handling missing data, encoding categorical variables.
- Next feature engineering is employed which includes removing the outliers and separating the dataset into train and test phases.
- Finally, the prediction model is developed using Machine Learning Algorithm.

B.DATASET ILLUSTRATION

The dataset was collected through Kaggle [13], since it is an open platform to download any kind of datasets. This set includes more than 4250 records with various fields that are discussed below.
The required features are:
Name: It is used to describe the car's brand name or model name.
Year: It denotes the year in which the car was purchased. Our data set includes a variety of vehicles purchased between 1995 and 2020.
Kilometers Driven: It reflects the total number of kilometers driven by a vehicle.
Fuel Type: It denotes the type of fuel utilized in a vehicle. Our data includes a variety of fuel types, including gasoline, diesel, compressed natural gas, liquid propane, and electricity.
Seller Type: It denotes the seller's personality. Individuals and Dealers are among the sellers represented in our database.
Transmission Type: It denotes the vehicle's make and model. Our database includes both manual and automatic vehicles.
Owner Type: It represents first-hand, second-hand, third-hand, fourth-hand, and above-fourth-hand automobile owners, as well as test-driven automobiles.
Selling Price: It represents the amount of money a seller expects on the car.
No. of Seats: The number of seats in the car is represented by this number.
Steering Type: It represents the car's type of steering. Our dataset consists tilted, adjustable, and telescopic steering types.

C.DATA PREPROCESSING

It is the first and most important step in the process of developing predictive models.
Pre-processing of data is a process of arranging the raw data that suits for machine learning algorithms.

a) HANDLING MISSING VALUES
If there are any null values or missing values they should be ignored or replace null values with mean, median and mode strategies to handle the missing and null values.

b) ENCODING CATEGORICAL DATA
Since the prediction is done using Random Forest Regression, the non-uniform data fields are converted to integers ranging from 0 to 1 for accepting the parameters.

c) SCALING OF THE FEATURES
In general the fields with high values will dominate the low values, so to overcome this issue, a lot of normalization techniques are used such as scaling to a range, clipping, z-score, minmax and many others. As $z = (x - \mu) / \sigma$ *(2).*

D. FEATURE ENGINEERING

The process of selecting, changing, and transforming original data into feature set which is used to develop a model using machine learning techniques is known as feature engineering. It includes dealing with outliers and depicting the correlation matrix.

a) OUTLIERS ANALYSIS
Outlier analysis is a process of data scanning that includes the identification of unusual entries in the dataset. It is a data analysis process that involves identifying abnormal observations in a dataset.

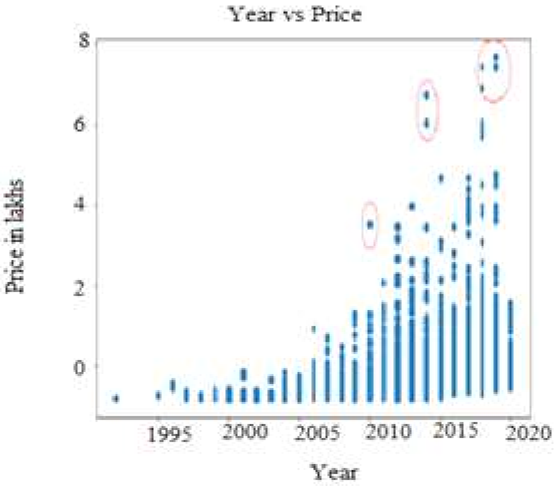


Fig 3: Identifying Outliers

Outliers in data can emerge as a result of human or computer errors during data collection as shown in Fig 3 and 4.These are identified using scatter plot. To remove these, the values are replaced with null.

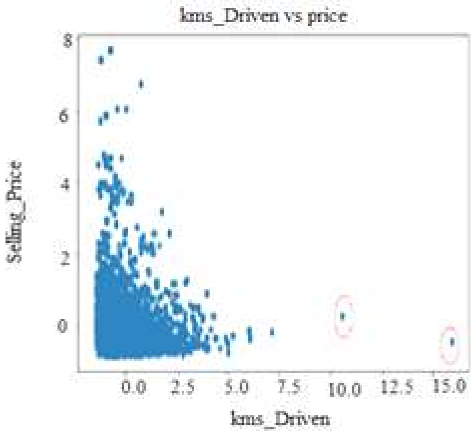


Fig4: Outlier Analysis

b) CORRELATION MATRIX

A correlation matrix is a table which represents the relation between the features. It gives the summarized view of the feature set. The below Figure 5 describes the correlation matrix of the features which is represented on the Heat Map. The dark green cells in the graph indicate that the features are highly positively correlated, while the red cells signify that they are negatively correlated.

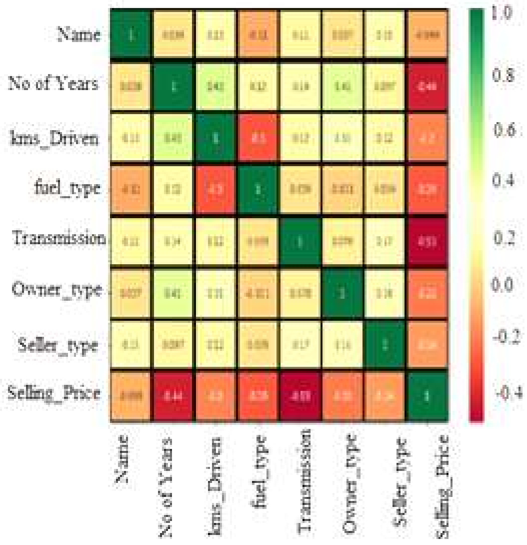


Fig 5: Correlation Matrix on Heat Map

c) TRAIN AND TEST SPLITS

As shown in the below Figure 6, the dataset is dissected into two phases one is for training which is of 80% and the other is for testing which is of 20%. This is the convention that in any Machine Learning Algorithm the dataset is divided into two parts. Training phase describes the creating the model and the testing phase determines the bringing accuracy from the created model.

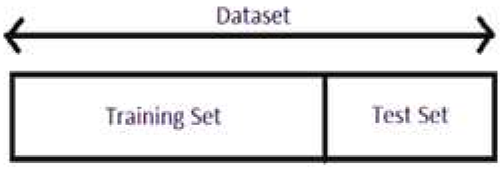


Fig 6: Train and Test Splits

E.FEATURE IMPORTANCE

Feature selection is a crucial machine learning approach with a significant impact on model performance. It minimizes over fitting, training time, and also boosts model performance. Each of our data items is given a value by Feature Importance. The higher the standard, the more important and relevant the variable is to our objective. To retrieve the top relevant features of the dataset Extra Trees Regressor class is used which is shown in below Figure 7.
Extra Trees Regressor Classifier is an ensemble technique which is used to aggregate the outputs of multiple decision trees that are collected in a "forest" for it's result.

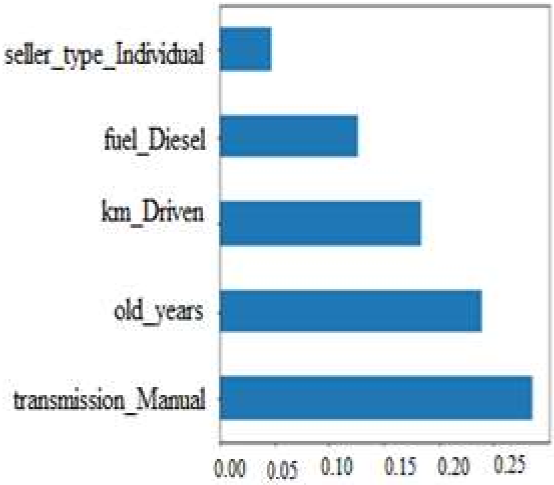


Fig 7: Relevancy Scores of Data Fields

a) Distribution between the Model Prediction and Test Dataset

The normal distribution of the model with the test dataset can be seen in the histplot shown below figure 8. As a result it says that the prediction of this model is highly accurate.

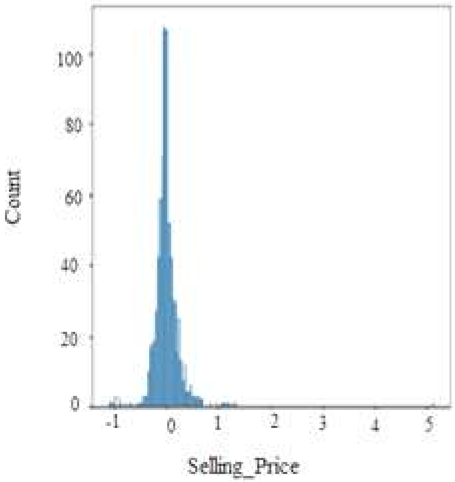


Fig 8: Histplot showing Predictions

b) Scatter Plot for Price Distribution

The scatter plot in Figure 9 shows a straight distribution, illustrates that this model is faultless. As a result, we can wind up that the forecasting of selling price using the current dataset is perfect.

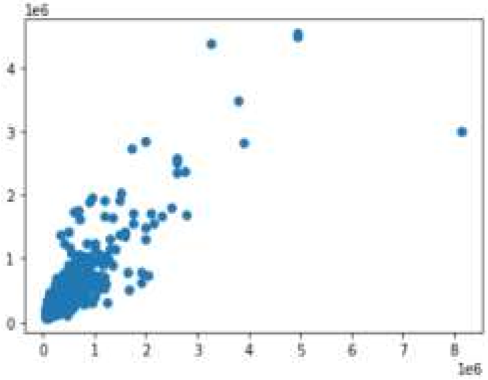


Fig 9: Scatter-plot for Price Distribution

F.ALGORITHMS

MACHINE LEARNING

Machine learning [14] is a subpart of Artificial Intelligence. Machine Learning algorithms aid in the creating the models based on historical data such that predictions and decisions can be done without involving the explicit instructions. Using these techniques such as supervised learning, unsupervised learning, and reinforcement learning, the models are developed and maintained.

Unsupervised Learning is a type of machine learning technique which uses the unlabeled data to find hidden patterns.
Reinforcement Learning is also a machine learning technique that works based on the feedback-back. If the agents works properly and correctly according to the given instructions then it will receives the reward otherwise it will be punished.

Random Forest Regression: The Random Forest algorithm is used to construct decision trees at the training phase. This method is also known as an Ensemble-Bagging method[15]. Random Forest is a mixture of both regression and classification techniques. It also works better for large datasets.
The term Ensemble is derived from the word "assemble", which refers to the process of combining several elements into a single working model or product. Bagging is a type of Ensemble Technique where the result is obtained by using the voting-win method .

Pseudo Code:

```
1. Randomly select ‘x’ features from a total of ‘n’
   features, where x<n.
2. Now calculate the node ‘n1’ by using the
   optimal split among the ‘x’ features.
3. The nodes are divided into child nodes by using
   the optimal split.
4. Steps 1 to 3 are repeated until the leaf node is
   reached. This is known as the actual prediction.
5. To get the ‘m’ number of trees, steps 1 to 4 are
   repeated for obtaining ‘m’ number of times.
```

Hyperparameter Tuning: The hyperparameters [16] are used to get optimized performance in terms of free of faults and speed. The arguments available in Scikit-learn library are:

- **n_estimators**: It describes the number of decision trees to be occurred in the random forest.
- **max_features:** It describes the maximum number of features considered for splitting a node.
- **min_sample_split**: The minimum number of leaves required to distinguish an internal node in the decision tree.
- **max_depth:** It describes the number of levels should the decision tree contains.
- **min_sample_leaf:** It describes the number of minimum leaf nodes in the tree.

Before the training starts we set the hyperparameters and the algorithm uses them to train the parameters. Behind the training phase, these parameters are uninterruptedly being modified and the updated ones which are formed at the end of the training phase set up the model. By this way, these hyper parameters tune to optimize the performance.

RandomisedSearchCV: It is a scikit-learn hyper-parameter tuning method in Python. It's utilized to implement a "fit" and "scoring" approach, as well as to forecast the best model.

IV. RESULTS EVALUATION

The regression models can be evaluated using the metrics [17] such as Mean Sqaured Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These are the metrics that describes the accuracy of the performance between Existing [2] and Proposed Systems.

MAE: It is the average difference between the expected and actual values for all observations.

MSE: It is the sum of all squared differences between expected and actual values.

RMSE: It is defined as the square root of sum of squared differences between the predicted and actual values.

$$MAE = (1 / n). \sum(\hat{x} - x)$$

$$MSE = (1 / n). \sum(\hat{x} - x)^2$$

$$RMSE = \sqrt{(1 / n)}. \sum(\hat{x} - x)^2$$

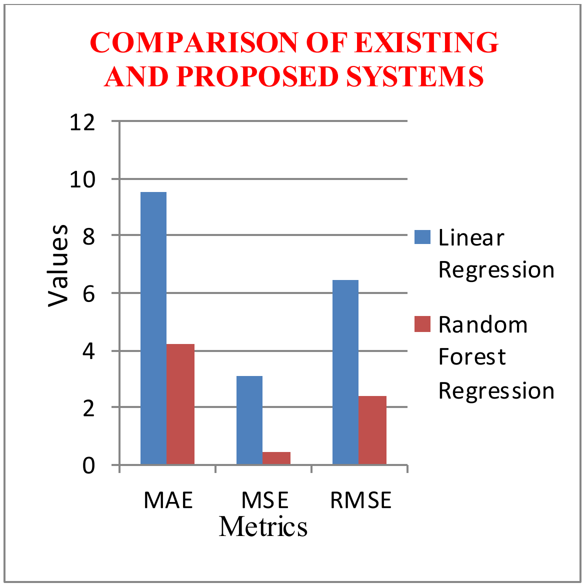
Here $\hat{x}$ is predicted value and x is true value.

A. Comparison between the Linear Regression and Random Forest Regression Algorithms

ALGORITHM	MAE	MSE	RMSE
Linear Regression	9.5351	3.1199	6.4762
Random Forest Regression	4.1918	0.4069	2.4082

Table 1: Summary of Models

The above table is obtained by passing the predictions into the metrics i.e., MAE, MSE and RMSE. Based on the values we obtained, the lesser values signify better performance in the error functions and improved the accuracy.



From the above bar graphs it is concluded that the lesser the value of the metrics the more accurate it gives from the model. So the Random Forest Regression gives the best accuracy when compared with the Linear Regression Model [2].

B. Prediction Output from Web Page

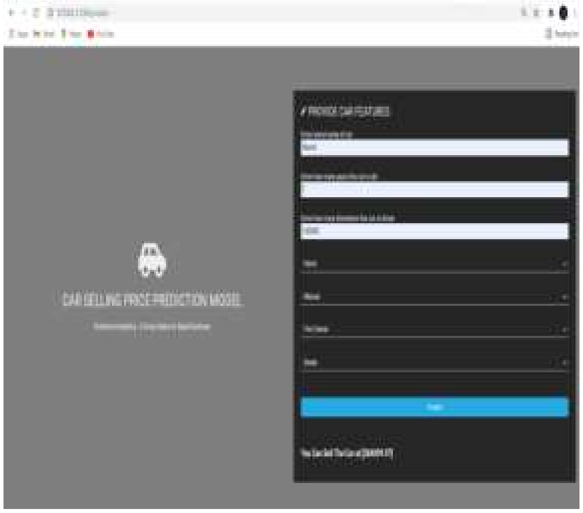


Fig 10: Inputs from Web User Interface

V. CONCLUSION

Building prediction models is a difficult undertaking. The data techniques were used such as encoding categorical variables, and performing feature engineering activities. Among many Machine Learning algorithms, the Random Forest model, adjusts the dataset well and shows the accuracy about 90%. And a web interface was also built to help end users with this task. Finally to conclude that Extra Trees Regression gives good result in feature engineering and also Random Forest Regression algorithm is best suited for forecasting the price of the second cars with much accuracy.

REFERENCES

[1] Car Price Predictions using Machine Learning Techniques by Enis GeicTEM International Burch University TEM Journal. Volume 8, Issue 1, Pages 113-118, ISSN 2217-8309, DOI: 10.18421/TEM81-16, February 2019.

[2] 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), 2021, pp. 1680-1687, doi: 10.1109/ICESC51422.2021.9532845.

[3] Mariana Listiani Support Vector regression for car leasing prediction 2009 Matriculation Number: 33750 Information and Media Technology Hamburg University of Technology

[4] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753-764 © International Research Publications House http://www. irphouse.com

[5] GONGGI, 2011. New model for residual value prediction of used cars based on BP neural network and non-linear curve fit. In: Proceedings of the 3 rd IEEE International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Vol 2. pp. 682-685, IEEE Computer Society, Washington DC, USA.

[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] Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Analytical Methods for Social Research). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511790942

[8] @article{Pal2018HowMI, title={How much is my car worth? A methodology for predicting used cars prices using Random Forest}, author={Nabarun Pal and Priya Arora and Dhanasekar Sundararaman and Puneet Kohli and Sai Sumanth Palakurthy}, journal={ArXiv}, year={2018}, volume={abs/1711.06970}}

[9] @article{Suma2020DataMB, title={Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics}, author={Dr. V. Suma}, journal={Journal of Soft Computing Paradigm}, year={2020}}

[10] TY - JOUR, AU - Madhuvanthi, K., AU - Kailasanathan, Nallakaruppan, AU - N C, Senthilkumar, AU - Somayaji, Siva, PY 2019/03/29, SP - T1 - Car Sales Prediction Using Machine Learning Algorithmns VL - 8 JO - International Journal of Innovative Technology and Exploring Engineering

[11]International Research Journal of Modernization in Engineering Technology and science by Ashutosh Datt Sharma and Vibhor Sharma Volume:03/Issue:06/June-2021 Impact Factor- 5.354 available from: https://irjmets.com/rootaccess/forms/uploads/IRJMETS462275.pdf

[12] N. Sun, H. Bai, Y. Geng and H. Shi, "Price evaluation model in second-hand car system based on BP neural network theory," 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2017, pp. 431-436, doi: 10.1109/SNPD.2017.8022758.

[13] Car Dataset Available from: https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho

[14] [Online] Machine Learning – Wikipedia https://en.wikipedia.org/wiki/Machine_learning

[15] [Online] Available from: https://www.geeksforgeeks.org/random-forest-regression-in-python/

[16] [Online] Available from: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

[17] Mean-squared-error, mean-absolute-error, root-mean-squared-error.