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) | 978-1-6654-2867-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICESC51422.2021.9532845

Chejarla Venkat Narayana
Computer Science and Engineering
Lakireddy Bali Reddy college of Engineering
Mylavaram,521230, Andhra Pradesh, India
cvnreddy.chejarla@gmail.com

Chinta Lakshmi Likhitha
Computer Science and Engineering
Lakireddy Bali Reddy college of Engineering
Mylavaram,521230, Andhra Pradesh, India
ch.likhithachinta@gmail.com

Syed Bademiya
Computer Science and Engineering
Lakireddy Bali Reddy college of Engineering
Mylavaram,521230, Andhra Pradesh, India
sdbademiya99@gmail.com

Karre Kusumanjali
Computer Science and Engineering
Lakireddy Bali Reddy college of Engineering
Mylavaram,521230, Andhra Pradesh, India
kusumakarre444@gmail.com

Abstract:

It is generally known that, taking wise and challenging decisions is really a crucial task in every business. Taking improper decisions can cause huge loss and even lead to shutdown of business. To propose a novel solution for this challenge, this research work majorly focuses on one of the retail businesses i.e., used car sales business. The proposed research work shows that, the predictive analytical models will be a great add-on to business mainly for assisting the decision making process. Predictive Analytics is a process, where the businesses use statistical methods and technologies to analyze their historical data for delivering new insights and plan the future accordingly. The major objective of our paper is to build a prediction model i.e., a fair price mechanism to predict the cars selling price based on their features like the car model, the number of years that a car is old, the type of fuel it uses, the type of seller, the type of transmission and the number of kilometers that the car has driven so far. This paper will help to get an approximation about selling price of a used car based on its features and reduces the seller and consumer risk in business. The proposed model utilizes the machine learning algorithms and regression techniques of statistics like linear, decision tree and random forest regressions to achieve this task.

Keywords: Machine Learning, Car Price Prediction, Predictive Analytics, Regression Techniques, Linear Regression, Decision Tree Regression, Random Forest Regression, Regression Analysis, Prediction Models

I. INTRODUCTION

Based on various reports and statistics of Indian Used Car Market, the market of used cars has reached to twice the market of new car sales. The organized sales of used cars have seen a tremendous growth over the past 3 years, and it is expected to raise up to 15% or more during the forecast period 2020-2025 [19]. There are lot of factors creating the impact on this growth, such as customer economy standards, assuming the used cars as best transport media, providing a positive firsthand knowledge about price and low financing cost for used vehicles and many other. Some of the most prominent organizations taking control over the current market of used cars sales are OLX, Cars24, Car Dekho, Droom and many other. Organizations need to come up with various prediction models to plan their future events, promote their brand and extend their business network.

The used car market has got the central attention in India's automotive industry. The ratio of new cars to old cars was 1:1.2 a few years ago, but it is now 1: 2.2 i.e., when 10 new cars are sold, there are 22 used cars available in market. The expected sales of used cars in 2008-2009 was 37 lakhs. For the financial year 2018-2019, the sales were expected as 62 lakhs, valued Rs. 1.62 lakh crores [20] and it is growing day by day.

Predicting the selling price of used cars is not a simple task. It needs good domain knowledge and estimates about features of car. Brand model, age of car, color of car, kilometers that car has driven, owner and seller type and fuel type of car can be some of the major factors we need to consider while judging the price of used cars. The list of features is not limited, we can consider lot of available features that can add value to our predictions.

In this paper, we are going to explore some regression techniques like Linear, Decision Tree and Random Forest regression and machine learning algorithms to forecast the selling price of used car based on its features. We will apply evaluation metrics on our models and finally, we will choose the model with good performance and minimum error rate.

A. Regression

In statistical modelling, regression can be defined as a process to estimate the relationships between a dependent variable and a group of independent features. Here, the target variable refers to the variable whose value need to be predicted and set of independent variables refers to the features that affect the value of target variable [14]. Mathematically, we can represent the regression as,

$$Y_i = f\,(Xi, parameters) + E \qquad (1)$$

Here, Y_i is target variable, X_i is set of input features and E is Error rate. To predict the target variable, we need f(Xi) called as model that closely fits to our data and various parameters can be involved in the model based on the algorithm we have chosen. Regression analysis is widely used for forecasting purpose. We can observe lot of regression techniques, and these are all based on type of data they work on and shape of the model that can be derived from data.

B. Machine Learning

Machine Learning is a field of research which employs data, algorithms for computers to learn and refine over time. Artificial Intelligence relies heavily on it. Machine Learning algorithms helps in creating the models based on training data to make predictions and decisions without having to be specifically instructed [15].

There are mainly three types of machine learnings possible such as supervised, unsupervised, and reinforced learning. Supervised learning deals with labelled data and its key task is to map the input features to the output label. Unsupervised learning deals with unlabeled data and its key task is to find the patterns hidden inside the data. Reinforced learning involves trial and error mechanism in interacting with environment i.e., agent will get positive reward when it takes the correct step towards goal otherwise negative rewards and feedbacks to get improved with the situation.

Regression and Classification belongs to supervised learning techniques, Clustering is an example of unsupervised learning technique and game playing and hidden Markov models are the use cases of reinforced learning techniques. In this paper, we are going to use the concept of supervised machine learning technique known as Regression.

C. Predictive Analytics

Analytics uses a combination of statistical methods such as predictive modelling and machine learning to analyze historical and current data to make forecasts for future events. Predictive models aid in detection of trends in data, allowing threats and possibilities to be identified [16].

Predictive Analytics playing a crucial role in various domains like retail industry, banking sector, healthcare, IT industry, entertainment, sports, IoT, social media and many other for various purposes like customer targeting, sales forecasting, inventory and resource managing, market analysis, risk and feedback assessment and future planning of firms.

This paper is majorly focusing on building a car price prediction model that could serve as a fair price mechanism and can create impact in retail business i.e., automotive industry in our case. These type of prediction models can create a great impact in businesses by forecasting the future and helps in firm planning.

II. LITERATURE REVIEW

We can find lot of research works related to predicting the price of products based on their features. Enis Gegic and their batch discussed about some regression models that were designed to predict the predict the price of a car using its features in their paper [1]. One of the major drawbacks of this research is the data they considered. They had taken 1105 samples of limited features only. Since data gathered using web scraper, so there are several samples with few attribute values.

Listiani has addressed a regression model that was developed using Support Vector Machines (SVM 's) to predict the price of leased cars with good precision than multivariate or simple multiple regression in their paper [2]. The SVM performs well with multidimensional datasets and is less vulnerable to over and under fittings. SVM's advantage over simple regression is not quantified in terms of metrics. It can be a drawback of this paper.

Sameerchand Pudaruth has studied some regression methods for car price prediction in Mauritius [3]. The limitation of this study is the data they considered. The dataset consists of very few records and features.

Noor and Jan studied about multiple linear regression methods for car price prediction in their paper [4]. The dataset was developed over a period of 2 months and consists of various features. Out of all the features the authors considered only engine type, car present price and car model as input features for their predictions.

Wu et al. used a neuro-fuzzy knowledge-based methods for car price prediction in their paper [5]. They considered the features like car brand, car engine type and year of car production. The findings of this model are close to those of simple regression model. This system has a proven track record of good performance against 2 million vehicles [6].

Kshitij Kumbara and their batch used a variety of regression techniques for car price prediction in their paper [7]. They considered the features like Mileage, Make, VIN, Model, Year and Location. The data used is huge but missing lot of important features that decides the resale value of cars.

Maduvanthi and their batch used a variety of regression techniques for car price prediction in their paper [8]. They considered the features like Mileage, Model, Fuel type, Horsepower and Price. They have followed Analytic Hierarchy process which results in good and accurate results. One of the major drawbacks of this paper is the data they considered. The data is lacking various important features like transmission type of car, seller type and many more that could impact the prediction of price for used cars.

Praful Rane and their batch discussed about some regression techniques in their paper [9]. The dataset details, performance metrics of models and comparative study of models is not mentioned in their paper. Proper visualizations or execution pseudocodes are not provided in this paper. We can state this as another flaw of paper.

Ashutosh Datt Sharma and Vibhor Sharma has discussed about Linear Regression model for predicting the price of used cars [10]. The dataset contains important features and the results also clear. This paper gives a brief explanation about data preprocessing techniques also, but it missed the feature importance score calculation and presentation of coefficients of linear model.

Chuancan chen and their batch had done a comparative analysis on evaluation of various models that predicts the used car price in their paper [11]. They have taken the data from several used car markets in shangai i.e., more than 1 lakh sales transactions were collected and analyzed. This paper helps in evaluating the models at various circumstances and assumptions.

Nabarun Pal and their batch has studied about some regression techniques to know the worth of a used car in their paper [12]. The methodology of this paper is very clear. Preprocessing and feature engineering tasks is not presented in the paper. This can be stated as a flaw of this paper. Moreover, this paper gives a good overview about machine learning process.

V. Suma has studied about prediction of demand for Refurbished Electronics in her paper [13]. The methodology of paper is clear, but it used very limited data that is collected from various ecommerce sites using the web crawlers. Moreover, this paper gives a good overview about predictive modelling and business analytics.

Table 1: *Outline of Related works*

S.NO	Paper and Authors	Limitation
1	Enis Gegic and their batch discussed about some regression models to predict the predict the price of a car using its features in their paper [1].	Considered 1105 samples of limited feature. Since data is gathered using web scraper, so there are several samples with few attribute values
2	Listiani has addressed a regression model that was developed using SVM's to predict the price of leased cars within their paper [2].	SVM's advantage over simple regression is not quantified in terms of metrics.
3	Sameerchand Pudaruth has studied some regression methods for car price prediction in Mauritius [3].	The dataset consists of very few records and features.
4	Noor and Jan studied about multiple linear regression methods for car price prediction in their paper [4]	The dataset was developed over a period of 2 months only and consists of various features.
5	Wu et al. used a neuro-fuzzy knowledge-based methods, adaptive neural networks and expert systems for car price prediction in their paper [5]	In this paper, for price forecasting, the effective criteria are simply assumed to be the car's model, year of manufacture, and engine style. Furthermore, the car's equipment is the thought to improve price performance.
6	Kshitij Kumbara and their batch used a variety of regression techniques for car price prediction in their paper [7]. They considered the features like Mileage, Make, VIN, Model, Year and Location	The data used is huge, but it is missing lot of important fields that decides the resale value of cars.
7	Maduvanthi and their batch used a variety of regression techniques for car price prediction in their paper [8]. They have followed Analytic Hierarchy process which has given good and accurate results.	The dataset is lacking various important features like transmission type of car, seller type and many more that could impact the prediction of price for used cars.
8	Praful Rane and their batch discussed about some regression techniques in their paper [9].	The dataset details, performance metrics of models and comparative study of models is not mentioned in their paper.
9	Ashutosh Datt Sharma and Vibhor Sharma has discussed about Linear Regression model for predicting the price of used cars [10].	it missed the features score calculation and presentation of model coefficients in linear regression.
10	Nabarun Pal and their batch has studied about some regression techniques to know the worth of a used car in their paper [12].	Preprocessing and feature engineering tasks is not presented in the paper. This can be stated as a flaw of this paper.
11	V. Suma has studied about prediction of demand for Refurbished Electronics in her paper [13].	The methodology of paper is clear, but it used very limited data that is collected from various ecommerce sites using the web crawlers.

III. PROPOSED WORK AND METHODOLOGY

We are going to follow the approach specified in Fig. 1. To build a car price prediction model. Firstly, we are going to collect the data and we will do exploratory analysis on it to get a summarized view about data i.e., what are the columns involved, how many records are there, what types of data is involved in dataset, is there any missing or null values present in dataset and many other observations and then we will go for preprocessing tasks like handling of missing data, categorical variables, and scaling of features. Later we will do feature engineering tasks like handling the outliers, dividing the dataset into train and test splits, and analyzing the importance of features in building the model. At last, we will generate various prediction models using machine learning algorithms and we will evaluate them.

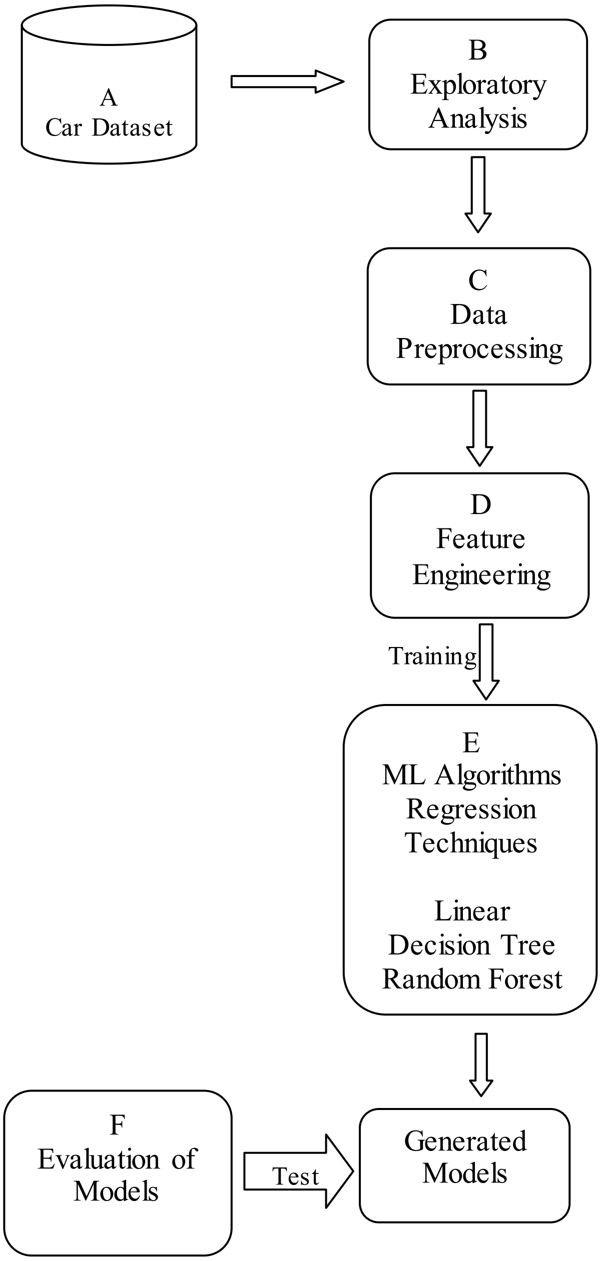
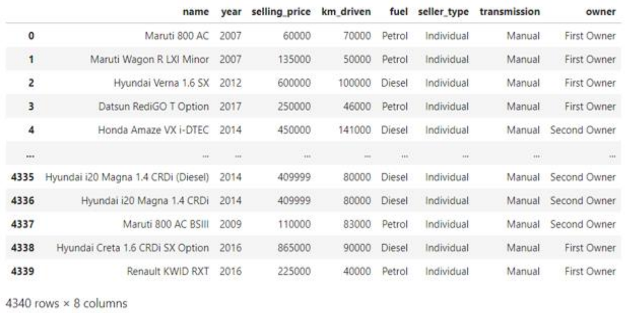


__Fig. 1.__ Block Diagram of Proposed Work

A. *Dataset Description*

We collected the data from Kaggle i.e., gathered by Nehal Birla and other two collaborators, which is an open source and community-maintained website. You can get similar datasets from Kaggle. The dataset [17] consists of more than 4000 records and various input features like model of car, year of model, kilometers driven so far, fuel type, seller type, transmission type and the owner type as specified in Fig. 2. The dataset consists of both numerical and categorical data i.e., data present in selling price, and kilometers driven columns are of numeric type and the remaining input fields are holding the categorical data.



	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
0	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner
1	Maruti Wagon R LXI Minor	2007	135000	50000	Petrol	Individual	Manual	First Owner
2	Hyundai Verna 1.6 SX	2012	600000	100000	Diesel	Individual	Manual	First Owner
3	Datsun RediGO T Option	2017	250000	46000	Petrol	Individual	Manual	First Owner
4	Honda Amaze VX i-DTEC	2014	450000	141000	Diesel	Individual	Manual	Second Owner
...	...	...	...	...	...	...	...	...
4335	Hyundai i20 Magna 1.4 CRDi (Diesel)	2014	409999	80000	Diesel	Individual	Manual	Second Owner
4336	Hyundai i20 Magna 1.4 CRDi	2014	409999	80000	Diesel	Individual	Manual	Second Owner
4337	Maruti 800 AC BSIII	2009	110000	83000	Petrol	Individual	Manual	Second Owner
4338	Hyundai Creta 1.6 CRDi SX Option	2016	865000	90000	Diesel	Individual	Manual	First Owner
4339	Renault KWID RXT	2016	225000	40000	Petrol	Individual	Manual	First Owner

4340 rows × 8 columns

__Fig. 2.__ Car Sales Dataset

a) ***Name:*** It is one of the data fields in our dataset. The dataset contains more than 1400 unique car names or models from the well-known 29 car brands like Maruti, Hyundai, Dastun, Tata, Chevrolet, Toyota, Jaguar, Mercedes-Benz, Audi, Skoda, Jeep, BMW, Mahindra, Ford, Nissan, Renault, Fiat, Volkswagen, Volvo, MG, Force, Isuzu, Opel Corsa, Ambassador, Kia and many more.

b) ***Year:*** It is one of the data fields in our dataset. It represents the year in which the car is bought. Our dataset covers various cars that are bought in the duration of 1992 to 2019.

c) ***Kilometers Driven:*** It is one of the data fields in our dataset. It represents the number of kilometers that a car has driven so far. Our dataset covers various cars which are driven in the range of 1 to 800000 kilometers.

d) ***Fuel Type:*** It is one of the data fields in our dataset. It represents the type of fuel used in car. Our dataset covers various fuel types like Petrol, Diesel, CNG, LPG and Electric.

e) ***Seller Type:*** It is one of the data fields in our dataset. It represents the type of seller. Our dataset covers various types of sellers like Individual, Dealers and Trustmark Dealers.

f) ***Transmission Type:*** It is one of the data fields in our dataset. It represents the type of car. Our dataset covers various types of cars like Manual and Automatic.

*g) **Owner Type:*** It is one of the data fields in our dataset. Our dataset covers various types of owners like Firsthand, Second hand, Third hand, Fourth & above owners, and Test-Driven cars also.

*h) **Selling Price:*** It is one of the data fields in our dataset. It represents the selling price of various used cars. It is the field to be predicted. Our dataset covers various cars whose selling price is in the range of 200000 to 9000000 Rupees.

B. Exploratory Analysis

Exploratory Data Analysis (EDA) is used to understand the data, to gather all possible insights, perform initial investigations, discover anomalies and patterns, test the hypothesis by using various statistics and visualization techniques. Some of our EDA results are shown as follows.

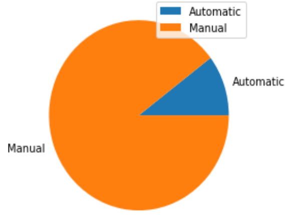


Fig. 3. *Transmission Vs Cars Count*

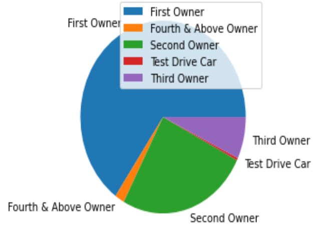


Fig. 4. *Owner Vs Cars Count*

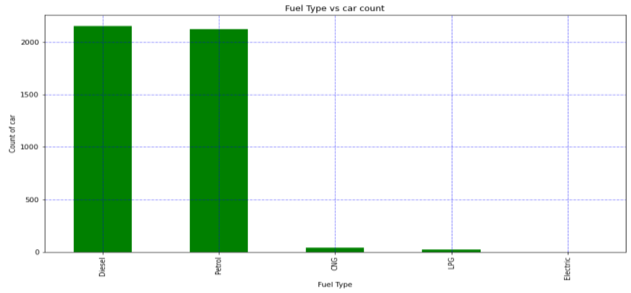


Fig. 5. *Fuel Type Vs Cars Count*

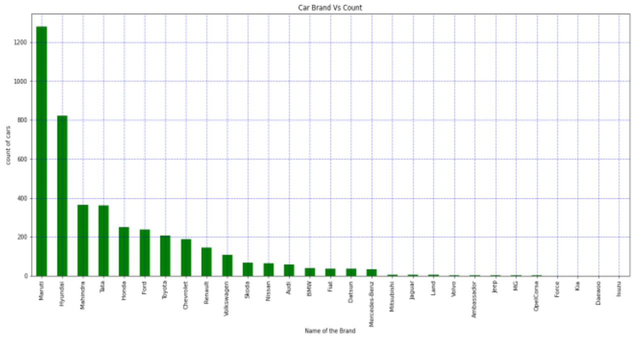


Fig. 6. *Brand Vs Cars Count*

The visualizations clearly states that most of the sold cars are having manual as transmission type, Diesel and Petrol as fuel type, kilometers driven as less than 3 lakhs, sold by first owners and of brands Maruti and Hyundai.

C. Data Preprocessing

It is the first and most important step in the process of developing predictive models. Generally, real world data may contain noises, missing values and outliers which cannot be employed directly in machine learning models. Preprocessing of data is a method of preparing the raw data for use in machine learning algorithms.

*a) **Checking Missing and Null Values:*** The dataset may contain missing or null values. Improper handling of missing and null values can result in false predictions or models. It is a very prior task before building the model. We can use various techniques such as imputation methods, predictions methods, replace nan values with mean, median and mode strategies to handle the missing and null values. Our dataset is not having any null values which is specified in heat map as Fig. 7.

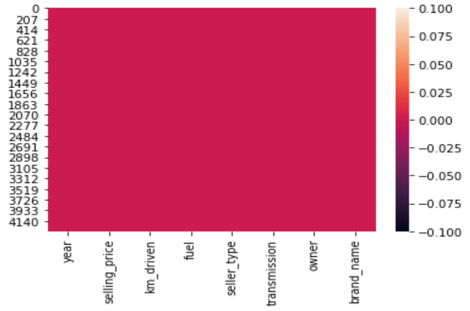
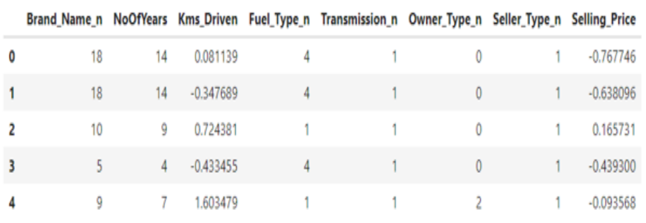


Fig. 7. *Heat Map of Null Values*

*b) **Encoding Categorical Variables:*** We cannot apply the categorical fields directly into our algorithms or models. We to encode the categorical variables like name of car, fuel type, owner type, transmission type and seller type to use them in our algorithms. Generally, we can follow label encodings i.e., convert every value into numeric value which ranges between 0 to number of categories-1.

*c) **Features Scaling:*** Generally, fields with huge data values can dominate the fields of small data values and may hide valuable insights. To avoid this issue, we will scale or normalize the data present in the fields so that every field can contribute equally to build the model. There are lot of normalization techniques available such as scaling to a range, clipping, z-score, min max and many other.



	Brand_Name_n	NoOfYears	Kms_Driven	Fuel_Type_n	Transmission_n	Owner_Type_n	Seller_Type_n	Selling_Price
0	18	14	0.081139	4	1	0	1	-0.767746
1	18	14	-0.347689	4	1	0	1	-0.638096
2	10	9	0.724381	1	1	0	1	0.165731
3	5	4	-0.433455	4	1	0	1	-0.439300
4	9	7	1.603479	1	1	2	1	-0.093568

Fig. 8. *Encoded and Scaled Features*

$$z = (x - \mu) / \sigma \qquad (2)$$

We can observe the scaling of kilometers driven and selling price columns and encoding of all categorical fields in Fig.8.

D. Feature Engineering

Feature Engineering refers to preparation of proper input dataset, compatible with the machine learning algorithms. It involves outlier analysis, feature splitting as input and target variables and splitting of dataset into train and test sets.

a) ***Outlier Analysis:*** The outlier is a datapoint which deviates from the existed data points. Outliers may occur in data due to human or machine errors while collecting the data. We can detect and remove the outliers using multiple methods like visualization of data, z-score normalization, and box plot.

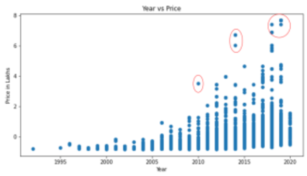


Fig. 9. *Outlier detection and removal*

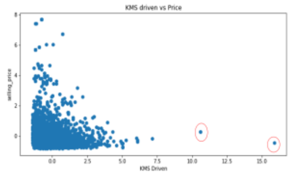


Fig. 10. *Outlier detection and removal*

Outliers can affect the model performance. Removal of outliers will help us to get good accuracy in predicting the resale value of car. Some of the outliers detected and removed from dataset are shown in Fig. 9. and Fig. 10. The scatter plots have shown that there are very few cars with more than 7 lakh kilometers driven value are sold. We can consider them as outliers, and we removed them from our dataset for improving the model performance.

b) ***Feature Importance****:* It is a fundamental concept in machine learning that has a significant impact on model's performance. Model performance can be harmed by irrelevant or partially relevant attributes. It reduces overfitting of algorithm, time of training and improves the performance of model.

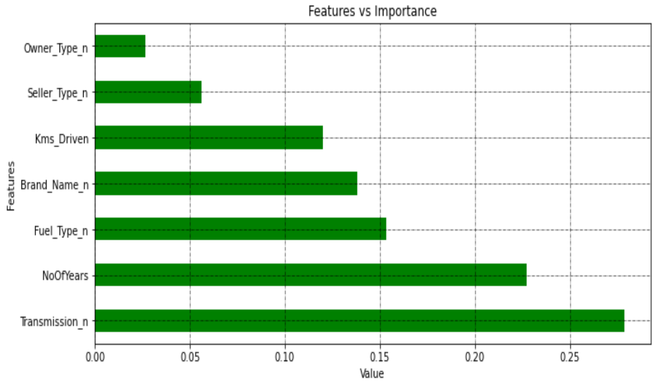


Fig. 11. *Features Importance*

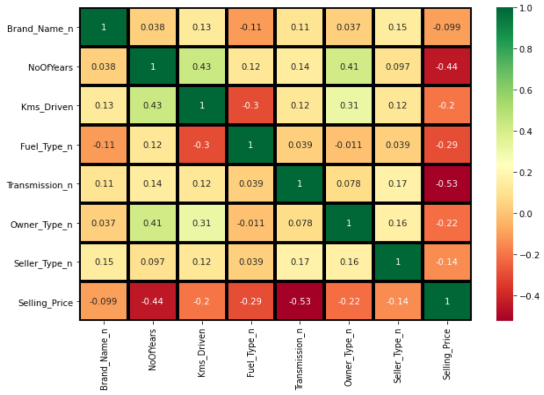
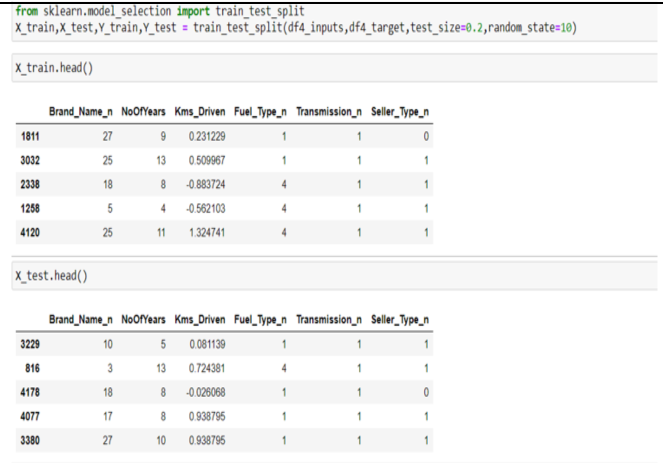


Fig. 12. *Correlation Matrix*

Feature Importance assigns a value to each of our data field. The greater the score, the more essential or relevant is to our target variable. In this paper, we are going to use Extra Tree Regressor class, feature importance property of model and correlation matrix to get the top relevant features of dataset as shown in Fig. 11 and Fig. 12.

We can observe that the fields like transmission type of car, fuel type of car, brand name of car, kilometers driven value, age of car and seller type are having impact on predicting the price of used car.

c) ***Train and Test Splits:*** We split the data into train and test splits such that 80% of data used for training and 20% of data is used for testing the model as shown in Fig. 13.



```
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test = train_test_split(df4_inputs,df4_target,test_size=0.2,random_state=10)
```

```
X_train.head()
```

```
      Brand_Name_n  NoOfYears  Kms_Driven  Fuel_Type_n  Transmission_n  Seller_Type_n
1811            27          9    0.231229            1               1              0
3032            25         13    0.509967            1               1              1
2338            18          8   -0.883724            4               1              1
1258             5          4   -0.562103            4               1              1
4120            25         11    1.324741            4               1              1
```

```
X_test.head()
```

```
      Brand_Name_n  NoOfYears  Kms_Driven  Fuel_Type_n  Transmission_n  Seller_Type_n
3229            10          5    0.081139            1               1              1
 816             3         13    0.724381            4               1              1
4178            18          8   -0.026068            1               1              0
4077            17          8    0.938795            1               1              1
3380            27         10    0.938795            1               1              1
```

Fig. 13. *Train and Test Splits*

E. ML Regression Techniques

*a) **Linear Regression:*** Linear regression aids in predicting the value of target variable based on given collection of independent features. In this technique, we can observe a linear relationship between a target variable 'y' and group of input features like 'x_1, x_2, x_3, …, x_n'.

The relationship between predictor variables and target variable is described by regression coefficients. The direction of association between a predictor variable and target variable is shown by the sign of each coefficient. A positive sign means that as the predictor variable rises, the target variable also rises. A negative sign means that the target variable drops as the predictor variable rises.

The coefficients between input features and target variables for our dataset is shown in Fig. 14. By using these coefficients, we can form a linear model to predict the price of used cars.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + E \qquad (3)$$

Here Y represents the target variable and x_1, x_2 ...., x_n represnets the input feature values and β_1, β_2, β_3, ..., β_n represents the coeffiecients of features and β_0 represents the intercept value.

Pseudocode:

```
1. Train the data with linear regression algorithm to get the coefficients and intercept value.
2. represent the linear model as
   $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ..... + \beta_n x_n + E.$
3. Here Y represents the target variable and $x_1$, $x_2$ ...., $x_n$ represnets the input feature values and $\beta_1$, $\beta_2$, $\beta_3$, ...., $\beta_n$ represents the coeffiecients of features and $\beta_0$ represents the intercept value.
```

*b) **Decision Tree Regression:*** This technique uses a decision tree to go from observations to final decisions. In these tree structures, leaf nodes represent the target class labels or the values to be predicted and the intermediate nodes represents combination of features that lead to final decisions or values.

Attribute selection and decision making is based on some properties like Entropy, Information gain and gini index.

Pseudocode:

```
1. Place the dataset's best attribute at the top of tree.
2. Split the training set into subsets. Subsets should be created so that each subset has data with the same attribute value.
3. Repeat steps 1 and 2 on each subset until you have found leaf nodes in all of the tree's branches.
```

*c) **Random Forest Regression:*** In Random decision forests are an ensemble learning technique, that does a classification, regression or other tasks by creating a large number of decision trees during training and outputs the values or decisions by considering all those trees.

Model Parameters: Number of Estimators that indicates the number of trees to be constructed in the forest and Maximum number of features that represents the number of features to be trained over each decision tree.

Pseudocode:

```
1. Select 'k' features at random from a total of 'n' features, where k<n.
2. Calculate the node 'nd' using the optimal split among the 'k' features.
3. Using the optimal split, divide the nodes into child nodes.
4. Repeat 1 to 3 steps until we reach to leaf node i.e., actual prediction.
5. To make a 'm' number of trees, repeat steps 1 to 4 for a 'm' number of times.
```

F. Model Evaluation and Results

On training the machine learning algorithm with the data, we will get different prediction models with different parameters and accuracies. Some of the regression models and their accuracies were shown below. We can use Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to evaluate the regression models [18].

MAE: It is defined as average of magnitude difference between predicted value and true value of all observations.

MSE: It is defined as average of squared difference between predicted and true value for all observations.

RMSE: It is defined as square root of MSE.

$$MAE = (1 / n) \,.\, \sum(y^{\wedge} - y) \qquad (4)$$

$$MSE = (1 / n) \,.\, \sum(y^{\wedge} - y)^2 \qquad (5)$$

$$RMSE = \surd\, (1 / n) \,.\, \sum(y^{\wedge} - y)^2 \qquad (6)$$

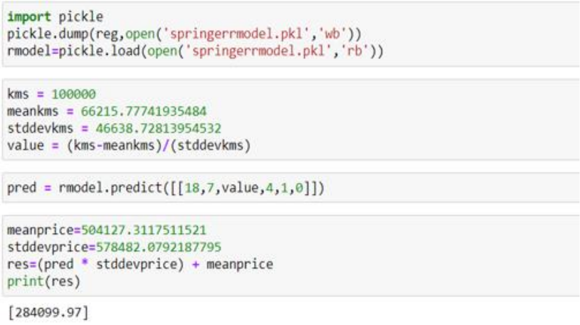
Here $y^{\wedge}$ is predicted value and y is true value.

Table 2: *Summary of Models*

Model	MAE	MSE	RMSE
Linear Regression	0.37998	0.36805	0.60667
Decision Tree	0.23565	0.21704	0.46587
Random Forest	0.19780	0.10122	0.31816

According to rule of thumb, RMSE values between 0.2 and 0.5 indicate that model can reasonably predict the data accurately. From the results observed, Random Forest Model is giving good accuracy of 85 percentage and minimum error rate. So, we are going to use that model for predicting the price of used car. We can give inputs via console or web user interface and can observe the results.

Giving Inputs from conosole and Web Interface:



```
import pickle
pickle.dump(reg,open('springerrmodel.pkl','wb'))
rmodel=pickle.load(open('springerrmodel.pkl','rb'))
```

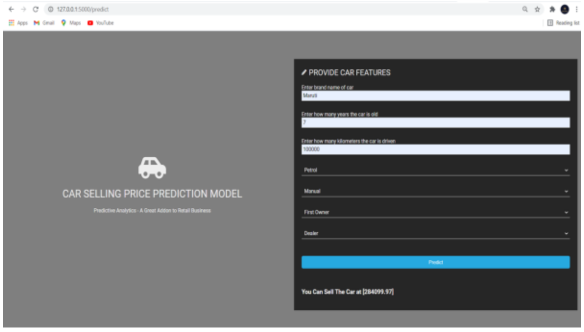
```
kms = 100000
meankms = 66215.77741935484
stddevkms = 46638.72813954532
value = (kms-meankms)/(stddevkms)
```

```
pred = rmodel.predict([[18,7,value,4,1,0]])
```

```
meanprice=504127.3117511521
stddevprice=578482.0792187795
res=(pred * stddevprice) + meanprice
print(res)
```

```
[284099.97]
```

Fig. 14. *Inputs from Python Console*



127.0.0.1:5000/predict
Apps
Gmail
Maps
YouTube
Reading list
CAR SELLING PRICE PREDICTION MODEL
Predictive Analytics - A Great Addon to Retail Business

PROVIDE CAR FEATURES

Enter brand name of car: Maruti

Enter how many years the car is old: 7

Enter how many kilometers the car is driven: 100000

Petrol

Manual

First Owner

Dealer

You Can Sell The Car at [284099.97]

Fig. 15. *Inputs from web user interface*

Here, 18 represents encoded brand name as Maruti, 7 represents the age of car, value represents the normalized kilometers driven, 4 represents the fuel type as petrol, 1 represents the transmission type as manual and 0 represents the seller as dealer type. On giving this input, we can be able to see the predicted car selling price as 2,84,099.97 Rupees.

IV. CONCLUSION

We know that building prediction models is a challenging task. It involves domain knowledge, feature analysis, machine learning techniques and many other complexities. In this paper we followed a clear methodology to build a fair price mechanism i.e., car price prediction models. In this paper, we did exploratory analysis on data to know a brief understanding about the dataset. Later, we applied Data preprocessing techniques like encoding categorical variables, scaling of features and we did feature engineering tasks like removing the outliers and knowing the importance of features to make our data compatible for models. Out of all the models, random forest model fits well to our data and giving a good accuracy of 85 percentage. So, we considered that model to predict the price of used car. We also provided a web interface to make this project more interactive.

V. FUTURE SCOPE

Although, this system has achieved a good performance in car price prediction problem. We can further extend by adding new sales data and other input features like safety index of car, number of doors and seats, color of car, steering type of car, weight of car and many other to know any hidden insights. As the businesses changes every day, keeping an eye on data and models is essential to get good insights.

References

[1] Car Price Preditcions using Machine Learning Techniques by Enis Geic TEM Journal Febraury 2019 International Burch University https://www.temjournal.com/content/81/TEMJournalFebruary2019_113_118.pdf

[2] Listiani Support Vector regression for car leasing prediction 2009 (Master thesis TU Hamburg-Harburg)

[3] Predicting the price of used cars by sameerchand pudaruth 2014 at International Journal of Information and Computation Technology http://ripublication.com/irph/ijict_spl/ijictv4n7spl_17.pdf

[4] Noor & Jan Machine Learning Techniques for Vehicle Price Prediction on june 2017 at International Journal of Computer Applications.

[5] Wu, J.D., HSU, C. C., & Chen, H.C Adaptive Genetic Algorithms and Expert systems for price prediction

[6] Du, J., Xie, L., & Schroeder, S. Price forecasting and Elasticity Estimation, Optimal distribution of auction vehicles system.

[7] http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26612934.pdf

[8] https://www.researchgate.net/publication/332072545_Car_Sales_Prediction_Using_Machine_Learning_Algorithmns

[9] International Research Journal of Engineering and Technology https://www.irjet.net/archives/V8/i4/IRJET-V8I4278.pdf

[10] International Research Journal of Modernization in Engineering Technology and science by Ashutosh Datt Sharma and Vibhor Sharma https://irjmets.com/rootaccess/forms/uploads/IRJMETS462275.pdf

[11] Comparative analysis of models evaluation by Chuancan chen and batch https://aip.scitation.org/doi/abs/10.1063/1.4982530

[12] Random forest technique to know the worth of car by Nabarun Pal and batch, https://arxiv.org/ftp/arxiv/papers/1711/1711.06970.pdf

[13] Suma, V., Data Mining based prediction of demand in Indian market for Refurbished Electronics, Journal of Soft Computing Paradigm.

[14] Regression Analysis Wikipedia

[15] Machine Learning Wikipedia

[16] Predictive Analytics Wikipedia

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

[18] Mean-squared-error, mean-absolute-error, root-mean-squared-error study to night.

[19] Indian Used Car Market Growth Trends and Forecast – reportlinker.com

[20] https://www.cars24.com/blog/used-car-market-india-untold-story/