e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:11/November -2020 Impact Factor- 5.354 www.irjmets.com

USED CAR PRICE PREDICTION USING LINEAR REGRESSION MODEL

Ashutosh Datt Sharma*1, Vibhor Sharma*2

*1Student, Department of Information Technology, Maharaja Agrasen Institute of Technology, Rohini, Delhi, India.

*2Assistant Professor, Department of Information Technology, Maharaja Agrasen Institute of Technology, Rohini, Delhi, India.

ABSTRACT

The production rates of cars have been rising progressively during the past decade, with almost 92 million cars being produced in the year 2019. This has provided the used car market with a big rise which has now come into picture as a well-growing industry. The recent arrival of various online portals and websites has provided with the need of the customers, clients, dealers and the sellers to be updated with the current scenario and trends to know the actual value of any used car in the current market. While there are numerous applications of machine learning in real life but one of the most pronounced application is it's use in solving the prediction problems. Again, there is an end number of topics on which the prediction can be done. This project is very much focused and based upon one of such application. Making the use of a Machine Learning Algorithm such as Linear Regression, we will try to predict the price of a used car and build a statistical model based on provided data with a given set of attributes.

Keywords: Cars, Price, Model, Predict, Features, Python, Module, Dataset, Plot.

I. INTRODUCTION

The manufacture fixes the costs of recent cars within the industry together with some additional costs that are incurred by the govt. majorly within the type of various taxes. So, customers that buy a replacement car remain assured of the money that they invest to be righteous. But because of such increase in prices of the new cars and therefore the inability of the many customers to shop for a replacement car thanks to the dearth of sufficient funds, they like used cars which has resulted into a world increase within the sales of used cars. Therefore, there's a necessity to possess a second hand car price prediction model to accurately determine the worthiness of the car considering a range of features. Although there are various online portals and websites which offers these services, their prediction method might not be necessarily the most effective. Besides, a special model and system can contribute in predicting power for a second user car's actual market price. it's mandatory to understand the particular market price of a car based upon it's features while buying or selling it.

Predicting the resale value of a car isn't a simple task. It requires knowledge about the varied number of things which are most significant in determining the worth of the used car. the foremost prominent attributes are usually the quantity of years that the car is employed, its build (and model), the car origin (the original country of the manufacturer), its mileage (the number of kilometers it's run) and its horsepower. Since fuel prices are rising, fuel economy is additionally important. Unfortunately, most of the people generally don't know exact amount of fuel their car consumes for every kilometers driven. Other factors like the kind of fuel, acceleration, the inside style, the quantity of its cylinders (measured in cc or cubic centimeters), the braking system, its size, safety index, number of doors, weight of the car, paint color, prestigious awards won by the automaker, customer reviews, whether it's a sports car, its physical state, whether it's automatic or manual transmission, whether it's controller, whether it belonged to a private or a corporation and other options like system, cooling, cosmic wheels, steering mechanism, GPS navigator all may influence the worth furthermore. another factors that buyers sometimes consider important are the knowledge like locality of previous owners and if the car has been involved in any reasonably accidents and had repairs. The physical appearance of car also majorly influences the value of the car. Therefore, we will easily conclude that the worth of car depends on numerous factors. But fairly often, information about all the factors mentioned above isn't available and therefore the customer needs to make his/her decision to get at a selected price supported few factors only. during this work, only a tiny and low subset of the factors

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:11/November -2020 Impact Factor- 5.354 www.irjmets.com

that are mentioned above is taken into account for the prediction model.

To be able to predict used cars value can help both buyers and sellers.

Used car sellers (dealers): They'll be benefitted from this model and thus they'll have an interest in results obtained from this study. If used car sellers better understand the important features of a second hand car and what makes a car desirable then they will consider this information and offer a far better price.

Online pricing services: There are many online portals moreover as websites that estimates the worth of a second hand car. they'll have an honest prediction model but having another model could help them to get better ends up in prediction and thus provides a better prediction of price to their users. Hence, the model developed during this study might be helpful for online web services which predicts the value of used cars.

Individuals: Plenty of people are there who have an interest within the used car market at some points in their life because they need to sell their car or buy a second hand car. Therefore, this might give them a platform to estimate the worth of any used car they need to sell or buy and are come up with an accurate value of their car in accordance to it's condition.

II. METHODOLOGY

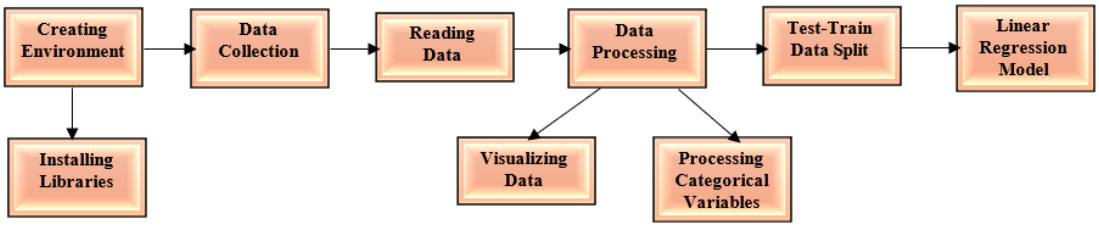

Figure-1: Workflow of Study.

Data was collected from the web portal of Kaggle.com where vehicle dataset from cardekho is provided for selling and buying cars. The following attributes were captured for each car:

Car Name, Year, Selling Price, Present or the Current Price, Kilometers driven, Fuel Type: Petrol, Diesel or CNG (Compressed Natural Gas) ,Seller Type: Dealer or Individual, Transmission : Automatic or Manual , Owner (No. of previous owners) .

Only those cars for which their price were mentioned in the dataset were listed and taken under consideration for this study. Further modification of the data was done as the data with null entries were removed to maintain homogeneity of the dataset as this could have affected our prediction model. Since it is very hard to find enough data records for Car name and model, the column Car_Name(Model) was also dropped from the dataset. Additionally, the Car Name is not of much help in this study. Also data provided for some of the attributes was sparse but considering their importance they were taken into consideration for this study.

	A	B	C	D	E	F	G	H	I
1	Car_Name	Year	Selling_Pri	Present_Pr	Kms_Drive	Fuel_Type	Seller_Typ	Transmissi	Owner
2	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
3	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
4	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
5	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
6	swift	2014	4.6	6.87	42450	Diesel	Dealer	Manual	0
7	vitara brez	2018	9.25	9.83	2071	Diesel	Dealer	Manual	0
8	ciaz	2015	6.75	8.12	18796	Petrol	Dealer	Manual	0
9	s cross	2015	6.5	8.61	33429	Diesel	Dealer	Manual	0
10	ciaz	2016	8.75	8.89	20273	Diesel	Dealer	Manual	0
11	ciaz	2015	7.45	8.92	42367	Diesel	Dealer	Manual	0
12	alto 800	2017	2.85	3.6	2135	Petrol	Dealer	Manual	0
13	ciaz	2015	6.85	10.38	51000	Diesel	Dealer	Manual	0
14	ciaz	2015	7.5	9.94	15000	Petrol	Dealer	Automatic	0
15	ertiga	2015	6.1	7.71	26000	Petrol	Dealer	Manual	0
16	dzire	2009	2.25	7.21	77427	Petrol	Dealer	Manual	0

Figure 2: The sample of collected Data.

After raw data was collected the further processing of data was done.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:11/November -2020 Impact Factor- 5.354 www.irjmets.com

NumPy: NumPy is acronym for 'Numerical Python' or 'Numeric Python'. Being an open source module of Python it is known for providing quick mathematical calculations on arrays and matrices. NumPy in conjunction with Machine Learning modules such as Scikit-learn, Pandas, Matplotlib, TensorFlow, etc. complete the Python Machine Learning Ecosystem as arrays and matrices are also an important part of the Machine Learning Environment. The essential multi-dimensional array-oriented computing functionalities that are designed for high-level mathematical functions and scientific computation are provided by NumPy. The homogeneous multidimensional array is the main object of NumPy module. It is a table with similar kind of components, i.e. integers or string or characters (homogeneous) and generally integers. In NumPy, dimensions are referred to as axes. The total count of axes is named rank.

Pandas: Like NumPy, Pandas is one of those python libraries which are most widely used in data science. It supports structures and data analysis tools which are easy to use and also provides high performance. In contrast to NumPy library that provides objects for multi-dimensional arrays, Pandas provides in-memory 2-d(2-dimensional) table object referred to as data frame. It is sort of a spreadsheet with row labels and column names. Therefore, with 2-d tables, pandas is capable of providing several other additional functions like building pivot tables, computing columns based on other columns and plotting graphs. To create a Pandas Series object pd.Series function is used. Every row is given an index and by default is assigned numerical value beginning from 0. Similar to NumPy, Pandas additionally offer the simple mathematical functions such as addition, subtraction, etc. It also provides conditional operations and broadcasting along with basic mathematical functionalities. A spreadsheet with cell values, column names, and row index labels is represented by Pandas data frame object. To visualize a data frame, we can consider dictionaries of Series. Rows and columns of a data frame are easy and intuitive to access. Pandas additionally offer SQL-like functionality to filter and sort rows according to the given conditions.

Matplotlib: Matplotlib is especially deployed for basic plotting. Bars, pies, lines, scatter plots and so on are part of visualization using matplotlib. Multiple figures of this module can be opened, however have to be closed explicitly. Only the current figure is closed by plt.close() while plt.close('all') would shut them all. For data visualization in Python, Matplotlib is a graphics package well integrated with NumPy and Pandas. The MATLAB plotting commands are closely mirrored by the pyplot module. Therefore, the MATLAB users could simply transit to plotting with Python. Matplotlib has different stateful APIs for plotting and works with data frames and arrays. The object represents the figures and aces and therefore plot() like calls without parameters suffices, avoiding any need to manage parameters. Matplotlib is extremely customizable and powerful. Pandas uses Matplotlib and it is also a neat wrapper around Matplotlib.

Seaborn: Seaborn provides various visualization patterns. It has easy and interesting default themes and uses fewer syntax. Statistics visualization is the speciality of seaborn and it is employed while summarizing data in visuals and additionally depict the data distribution. Seaborn creates multiple figures which typically results in OOM (out of memory) issues. Seaborn is additionally integrated for functioning with Pandas data frames. It extends the Matplotlib library for making ideal graphics with Python employing simple and easy methods. Seaborn is much more intuitive than Matplotlib and works with an entire dataset. In Seaborn, replot() is the API used with 'kind' parameter which specifies the type of plot that can be line, bar, or many of the other types. Since, Seaborn is not stateful, it is necessary for plot() to pass the object. Seaborn avoids plenty of boilerplate by providing commonly used default themes. Seaborn is employed for use cases that are more specific and also, under the hood it is Matplotlib. Statistical plotting is what it is especially meant for.

Scikit-learn: A variety of supervised and unsupervised learning algorithms via a homogeneous interface in Python are provided by this module. It is authorised underneath a permissive simplified BSD license and is distributed underneath several Linux distributions, promoting education as well as the business purposes. SciPy (Scientific Python) must be installed before one can use scikit-learn because the library is built upon SciPy. Scikits is conventionally the name used for the extensions or modules of SciPy. The learning algorithms are provided by this module and it is called scikit-learn. A level of robustness and support needed for use in production systems is the vision for this library. This basically implies a major focus on considerations namely ease in use, quality of code, collaboration, documentation and performance. The main focus of the library is on modelling the data. It does not target on loading, manipulating and summarizing the data.

Processing Categorical Variables: The above modules and libraries were then employed to read and visualize data with help of various plots and tables for a better understanding of our dataset. After that to deal with the categorical variables One Hot Encoding technique was employed which basically helps in adding details easily for better prediction and analysis. In the dataset Fuel had three sub categories: Petrol, Diesel and CNG while Transmission and Seller Type had 2 categories, using this function it simplifies the processing of data in

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:11/November -2020 Impact Factor- 5.354 www.irjmets.com

analysis. A binary column for each category is created using One Hot Encoding technique and additionally a sparse matrix or dense array is returned depending on the sparse parameter.

Train-Test Split: The dataset was then split into training and testing data. The train data is used for training our model while the test data was employed for testing and making predictions using our model. For this study we used 80% data to train our model while 20% data was then employed as test data for predictions. A Linear Regression model was employed for this study.

Linear Regression:
Linear regression is employed to find out the extent up to which there is a linear relationship between some dependent and one or more independent variables. Linear regression is generally of two types, simple linear regression and multiple linear regression. In this project there are more than one independent variable therefore, a "Multiple Linear Regression model was used.
The linear regression model can be represented by the following equation :-

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

Figure 3: Equation of the Model.

- Y is the predicted value
- θ_0 is the bias term.
- $\theta_1,...,\theta_n$ are the model parameters
- $x_1, x_2,...,x_n$ are the feature values.

III. MODELING AND ANALYSIS

Reading and Understanding Data: The dataset is imported and read for understanding.

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

Figure-4: Understanding the Data.

```
Car_Name         0
Year             0
Selling_Price    0
Present_Price    0
Kms_Driven       0
Fuel_Type        0
Seller_Type      0
Transmission     0
Owner            0
dtype: int64
```

Figure-5: Checking Null Values.

The Figure 3 above is an overview of our dataset that simply describes that what exactly does our dataset looks like. It simply displays all the attributes which are: Car Name, Year, Selling Price, Present Price, Kms Driven, Fuel Type, Seller Type, Transmission, Owner (Number of previous owners). Figure 4 covers the fact that the dataset does not contains any Null entries. Null entries are basically any kind of missing values in the dataset. It is necessary to know about the missing values or null entries because a null entry would affect the homogeneity of the dataset or the continuity of our data and this could create problem while data modelling and building the model. So, to avoid any such kind of problems we have to make sure that our dataset does not have any missing values or entries and in order to do that we would have to remove those data points which have any missing value in them from the whole dataset.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:11/November -2020 Impact Factor- 5.354 www.irjmets.com

Visualizing Data with Target Variable :

Figure-5: Selling Price vs Seller Type.

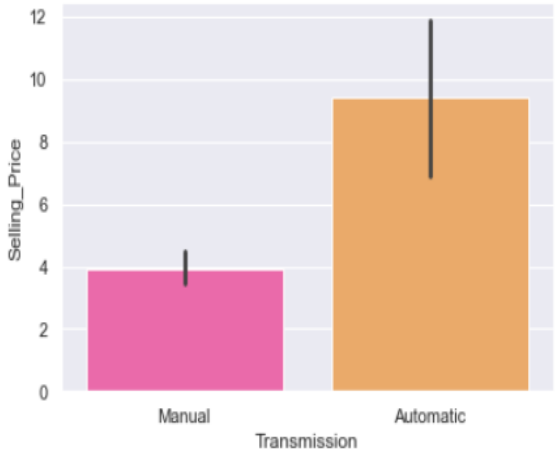

Figure-5: Selling Price vs Transmission.

From Figure 5 we can come to the conclusion that used cars have a higher selling price when sold by dealers in comparison to being sold by individuals. Similarly, Figure 6 tells us the fact that selling price of the cars with manual transmission is lower than those cars which are having automatic transmission.

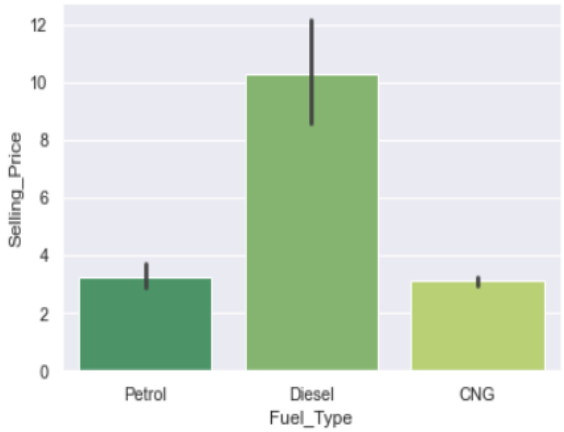

Figure-7: Selling Price vs Fuel Type.

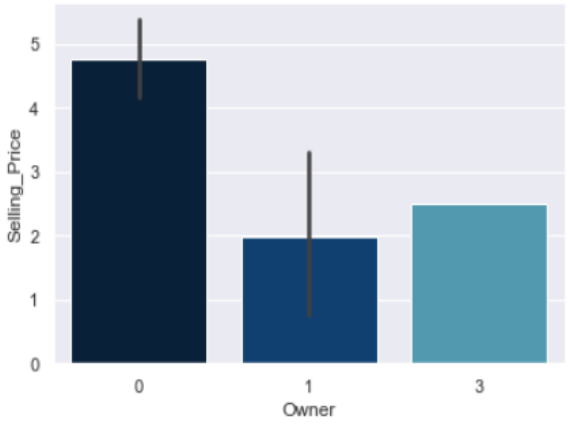

Figure-8: Selling Price vs Owner.

From Figure 7 it can be concluded that used cars with Diesel as fuel type have higher selling price as compared to those which have Petrol and CNG as fuel type. Additionally, Figure 8 clarifies that the selling price of cars with no previous owners is higher than rest of the cars.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:11/November -2020 Impact Factor- 5.354 www.irjmets.com

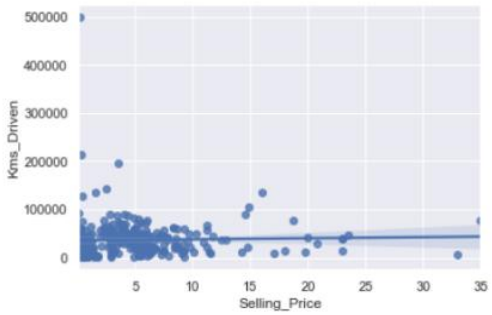

Figure 9: Kms Driven vs Selling Price.

Figure 10: Present Price vs Selling Price.

The Figure 9 above concludes that lesser number of kilometers driven would increase the selling price of the used cars. Apart from that, Figure 10 depicts that a greater present price of the car would also result in a greater selling price of the used car.

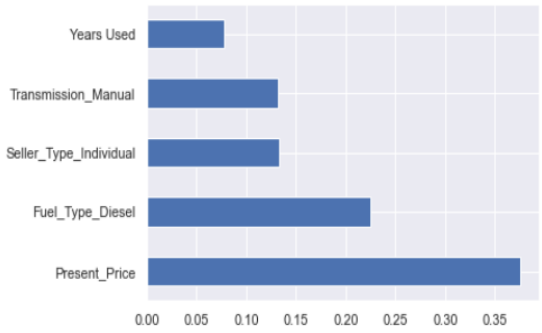

Figure 11: Selling Price vs Years Used.

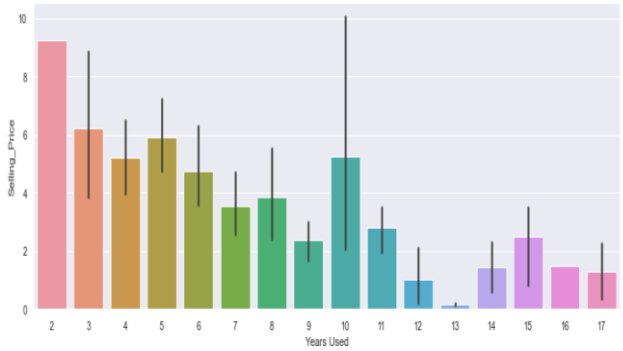

Figure 12: Correlation Features.

From Fig-11, it can be concluded that the cars used for lesser number of years generally have a higher selling price as compared to the cars used for greater number of years. Fig-12 is a plot showing correlation features and their importance ant it depicts that Present Price is the most important correlation feature followed by Fuel Type (One Hot Encoded), Seller Type (One Hot Encoded), Transmission (One Hot Encoded) and Years Used.

	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner	Years Used
0	3.35	5.59	27000	Petrol	Dealer	Manual	0	6
1	4.75	9.54	43000	Diesel	Dealer	Manual	0	7
2	7.25	9.85	6900	Petrol	Dealer	Manual	0	3
3	2.85	4.15	5200	Petrol	Dealer	Manual	0	9
4	4.60	6.87	42450	Diesel	Dealer	Manual	0	6

Figure 13: Final Dataset.

IV. RESULTS AND DISCUSSION

The r_2 score of the Multiple Linear Regression Model obtained is **0.86** and a plot of Originals vs Predictions in which red curve represents original selling price of cars and blue curve represents the predicted selling price of cars shows that they both are very close to each other. Fig-14 shows both the r_2 score as well as the Originals vs Predictions Plot.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:11/November -2020 Impact Factor- 5.354 www.irjmets.com

```
LinearRegression()

r_2 score : 0.8625260513315252
```

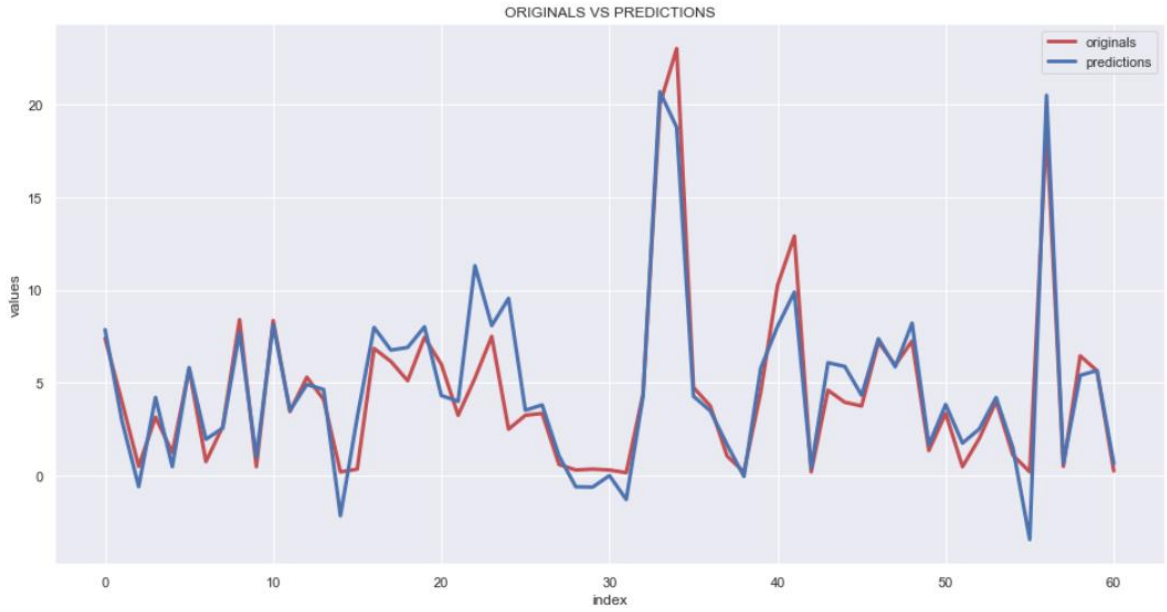

Figure 14: Model Accuracy and Originals vs Predictions Curve.

V. CONCLUSION

In this study, a Linear Regression Model was successfully implemented employing various prominent algorithms from the python libraries and modules. After the collection of data was done, further processing of data was done. The null entries and missing datapoints were removed from the dataset and the categorical variables were also processed using One Hot Encoding technique. The results showed that there is a positive correlation between Selling Price and Present Price while a negative correlation between Selling Price and Kms Driven, Years Used and Owner (Number of Previous Owners). Positive correlation can be referred to as Direct proportion while Negative correlation can be referred to as Inverse Proportion. Also, it was concluded that Selling Price of cars was higher when sold by dealers when compared to individuals. Similarly, the Selling Price was higher for cars that were automatic in transmission. It was also observed that Selling Price of cars with Fuel Type Diesel was higher than those having Petrol and CNG as Fuel Type. The **r2 score** of Linear Regression was **0.86** which is good and predictions were quite close to the original selling prices.

VI. REFERENCES

[1] Sameerchand Pudaruth, Computer Science and Engineering Department, University of Mauritius, Reduit, MAURITIUS. Predicting the Price of Used Cars using Machine Learning Techniques. International Journal of Information & Computation Technology, 2014.

[2] Saamiyah Peerun, Nushrah Henna Chummun and Sameerchand Pudaruth, University of Mauritius, Reduit, Mauritius. Predicting the Price of Second-hand Cars using Artificial Neural Networks. Proceedings of the Second International Conference on Data Mining, Internet Computing, and Big Data, Reduit, Mauritius 2015.

[3] Nabarun Pal(Department of Metallurgical and Materials Engineering, Indian Institute of Technology Roorkee, Roorkee, India), Priya Arora(Department of Computer Science, Texas A & M University Texas, United States), Sai Sumanth Palakurthy(Department of Computer Science and Engineering, IIT (ISM) Dhanbad, Dhanbad, India), Dhanasekar Sundararaman (Department of Information Technology, SSN College of Engineering, Chennai, India), Puneet Kohli (Department of Computer Science, Texas A & M University, Texas, United States). How much is my car worth? A methodology for predicting used cars prices using Random Forest. Future of Information and Communications Conference (FICC) 2018.

[4] Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric, International Burch University, Sarajevo, Bosnia and Herzegovina. Car Price Prediction using Machine Learning Techniques. TEM Journal, February 2019.

[5] Ashish Chandak , Prajwal Ganorkar , Shyam Sharma , Ayushi Bagmar, Soumya Tiwari, Information Technology, Shri Ramdeobaba College of Engineering, Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur. Car Price Prediction Using Machine Learning. India International Journal of Computer Sciences and Engineering, May 2019.

[6] Pattabiraman Venkatasubbu, Mukkesh Ganesh. Used Cars Price Prediction using Supervised Learning Techniques. International Journal of Engineering and Advanced Technology (IJEAT), December 2019.

[7] Laveena D'Costa, Ashoka Wilson D'Souza, Abhijith K, Deepthi Maria Varghese. Predicting True Value of Used Car using Multiple Linear Regression Model. International Journal of Recent Technology and Engineering (IJRTE). January 2020.

[8] S.E.Viswapriya, Durbaka Sai Sandeep Sharma, Gandavarapu Sathya Kiran. Vehicle Price Prediction using SVM Techniques. International Journal of Innovative Technology and Exploring Engineering (IJITEE), June 2020.