RUTGERS

MSDS 596
Regression Analysis

"CAR PRICE ANALYSIS AND MODELING"

Final Project Report

Under the guidance of: Dr. Koulik Khamaru

Jayanth Dasamantharao

CONTENTS

ABSTRACT

Over the past few years, the sale of used cars has become a lucrative business in several markets worldwide. Well-maintained used cars are becoming more popular than new ones today. It shouldn't come as a surprise that used cars have as many fans as the new ones. Predicting used car prices is therefore a highly interesting subject. Apart from being affordable for buyers, they also allow sellers to move around without keeping a car. Through this project, we aim to estimate used car prices by using attributes that correlate closely with the price such as production year, car model, mileage, engine size, number of airbags, levy and so on. To accomplish this, linear regression techniques have been employed. In order to develop a model based on the extracted raw data, a pre-processing step where data cleaning will be followed by the application of linear regression concept- model selection and then train the model and test it to check its accuracy. To predict used car prices, we will build a linear regression model in accordance with the accuracy of the results.

1. INTRODUCTION

The primary goal of this project is to predict the accuracy of a linear regression model based on the sale price of used cars. This model can help both buyers and sellers to make informed decisions based on the value of a car. Linear regression is a common method for predicting numerical values. In the context of used car prices, linear regression models can be used to predict the price of a used car based on its target variables. By doing Linear Regression Analysis on a large dataset, we developed a model that is accurate in predicting the sale price of a used car based on its characteristics and features.

Considering a large dataset such as this, outliers are required to predict the accuracy for a model. They can have a significant impact on the results of analysis, as they can skew the mean and other measures of central tendency. To identify these outliers in the dataset, visual methods such as box plots and joint plots were used. Once the outliers were identified, interquartile range (IQR), variance inflation factor (VIF) and jackknife residual method were implemented to handle them.

To make predictions about the sale price of a used car, we can use a Linear Regression model that is trained on a large dataset of historical car sales data. To get the accuracy of a model using Linear Regression, goodness of fit is the key. The process of model selection was done using various methods which includes forward selection, backward selection, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC). The model was selected based on lower AIC & BIC values. The selected model has higher forward selection and backward elimination size which predicts better performances. All the predictors are significantly nonzero at the level α. Furthermore, correlation heat maps were used to identify the strength and direction of relationships between different variables in the dataset.

Comparison between two probability distributions was done by plotting their quantities against each other. Quantile-quantile (QQ) plots were used to check whether the sample comes from the obtained distribution. These help us to provide a visual representation of the similarity or difference between two distributions in the model. Further, the model was evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE). The dataset which is used for the model shows a minimum RMSE value which shows accurate predictions. This can be used to compare the performance of different estimators and acts as an important metric for evaluating the performance of regression algorithms.

As a result of various analyses and observations, this model can learn the relationship between the various characteristics and features of a car, for calculating the sale price, and can use this information to calculate the accuracy of the model.

2. PROBLEM FORMULATION

The primary objective is to develop a linear regression model that can predict the accuracy of the car prices from the dataset. We will also train the model and test it on the dataset to assess its ability. Also, analysis of various aspects and factors that influence the prices of cars in the United States that lead to the actual used car price valuation will be done. Considering this is an interesting topic in the research community, and in continuing their footsteps, we hope to achieve significant results using more advanced methods.

3. DATA SET SOURCE

The project deals with used cars in the United States. The dataset was obtained from Kaggle.com and scraped in order to build the effective intelligent model. The dataset consists of 19237 rows and 18 different columns. Below are the description and sample of the dataset.

S.No.	Column	Description
1	ID	The ID of each car
2	Price	Price of each car in USD
3	Levy	Interest rate on the price in USD
4	Manufacturer	Manufacturer of the car. E.g.: Toyota, Ford, etc.
5	Model	Model of the car. E.g.: Elantra, Camry, Sonata, etc.
6	Production Year	Manufactured year of the car
7	Category	Type of the car. E.g.: Sedan, Hatchback, Jeep, etc.
8	Leather Interior	Whether it has leather interior or not
9	Fuel Type	Type of Fuel. E.g.: Hybrid, Petrol, Diesel, etc.
10	Engine Volume	Volume of the Engine
11	Mileage	Total distance traveled in kilometers
12	Cylinders	Number of cylinders
13	Gear Box Type	Type of Gear Box. E.g.: Automatic, Manual, etc.
14	Drive Wheels	Wheel of a motor vehicle that transmits force. E.g.: Rear, Front, etc.
15	Doors	Number of doors
16	Wheel	Hand drive. E.g.: Left, Right, etc.
17	Color	Color of the car
18	Airbags	Number of Airbags in the car

ID	Price	Levy	Manufacturer	Model	Prod. year	Category	Leather interior	Fuel type	Engine volume	Mileage	Cylinders	Gear box type	Drive wheels	Doors	Wheel	Color	Airbags
45654403	13328	1399	LEXUS	RX 450	2010	Jeep	Yes	Hybrid	3.5	186005 km	6	Automatic	4x4	04-May	Left wheel	Silver	12
44731507	16621	1018	CHEVROLET	Equinox	2011	Jeep	No	Petrol	3	192000 km	6	Tiptronic	4x4	04-May	Left wheel	Black	8
45774419	8467	-	HONDA	FIT	2006	Hatchback	No	Petrol	1.3	200000 km	4	Variator	Front	04-May	Right-hand drive	Black	2
45769185	3607	862	FORD	Escape	2011	Jeep	Yes	Hybrid	2.5	168966 km	4	Automatic	4x4	04-May	Left wheel	White	0
45809263	11726	446	HONDA	FIT	2014	Hatchback	Yes	Petrol	1.3	91901 km	4	Automatic	Front	04-May	Left wheel	Silver	4
45802912	39493	891	HYUNDAI	Santa FE	2016	Jeep	Yes	Diesel	2	160931 km	4	Automatic	Front	04-May	Left wheel	White	4
45656768	1803	761	TOYOTA	Prius	2010	Hatchback	Yes	Hybrid	1.8	258909 km	4	Automatic	Front	04-May	Left wheel	White	12
45816158	549	751	HYUNDAI	Sonata	2013	Sedan	Yes	Petrol	2.4	216118 km	4	Automatic	Front	04-May	Left wheel	Grey	12
45641395	1098	394	TOYOTA	Camry	2014	Sedan	Yes	Hybrid	2.5	398069 km	4	Automatic	Front	04-May	Left wheel	Black	12
45756839	26657	-	LEXUS	RX 350	2007	Jeep	Yes	Petrol	3.5	128500 km	6	Automatic	4x4	04-May	Left wheel	Silver	12
45621750	941	1053	MERCEDES-BEN	E 350	2014	Sedan	Yes	Diesel	3.5	184467 km	6	Automatic	Rear	04-May	Left wheel	White	12
45814819	8781	-	FORD	Transit	1999	Microbus	No	CNG	4	0 km	8	Manual	Rear	02-Mar	Left wheel	Blue	0
45815568	3000	-	OPEL	Vectra	1997	Goods wagon	No	CNG	1.6	350000 km	4	Manual	Front	04-May	Left wheel	White	4
45661288	1019	1055	LEXUS	RX 450	2013	Jeep	Yes	Hybrid	3.5	138038 km	6	Automatic	Front	04-May	Left wheel	White	12
45732604	59464	891	HYUNDAI	Santa FE	2016	Jeep	Yes	Diesel	2	76000 km	4	Automatic	Front	04-May	Left wheel	White	4
45465200	549	1079	TOYOTA	CHR	2018	Jeep	Yes	Petrol	2	74146 km	4	Automatic	Front	04-May	Left wheel	White	12
45772281	7683	810	HYUNDAI	Elantra	2016	Sedan	Yes	Petrol	1.8	121840 km	4	Automatic	Front	04-May	Left wheel	Blue	12
45797221	28382	810	HYUNDAI	Elantra	2016	Sedan	Yes	Petrol	1.8	54317 km	4	Automatic	Front	04-May	Left wheel	White	4
45772104	549	2386	HYUNDAI	Sonata	2006	Sedan	Yes	Petrol	3.3	295059 km	6	Automatic	Rear	04-May	Left wheel	Blue	12
45653306	941	1850	LEXUS	RX 400	2008	Jeep	Yes	Hybrid	3.5	364523 km	6	Automatic	4x4	04-May	Left wheel	Black	12
45801686	18826	531	HYUNDAI	Elantra	2012	Sedan	Yes	Petrol	1.6	112645 km	4	Automatic	Front	04-May	Left wheel	Silver	4

4. METHODOLOGY

The methodology involves the following data engineering steps and regression techniques:

4.1 Pre-processing

Pre-processing is the process of cleaning and preparing data for analysis. This is an important step in data analysis because raw data is often noisy and unstructured and may contain missing or incorrect values. Pre-processing data is a critical step in data analysis, as it helps to ensure that the data is of high quality and is in a form that can be effectively analyzed. By carefully pre-processing the data, one can improve the accuracy and reliability of the analysis, and gain insights that might not be possible with raw, unprocessed data.

There are several steps involved in preprocessing data for analysis, including:

4.1.1 Data cleaning: This involves identifying and correcting errors and inconsistencies in the data. The missing values were filled, outliers were removed, and data formats were standardized.

4.1.2 Data transformation: This involves converting the data into a form that is more suitable for analysis. Data scaling was done along with application of mathematical transformations and combining multiple datasets in order to get the accuracy.

4.1.3 Data reduction: This involves reducing the amount of data by selecting a subset of the data or aggregating the data. This can help to improve the efficiency of the analysis and avoid the risk of overfitting the model to the data.

4.2 Exploratory Data Analysis (EDA)

Exploratory data analysis was used to analyze data that focuses on discovering patterns and relationships in the data, rather than testing hypotheses or making predictions. Through this, the dataset was explored and understood to visualize and summarize the main characteristics and to identify any trends or outliers.

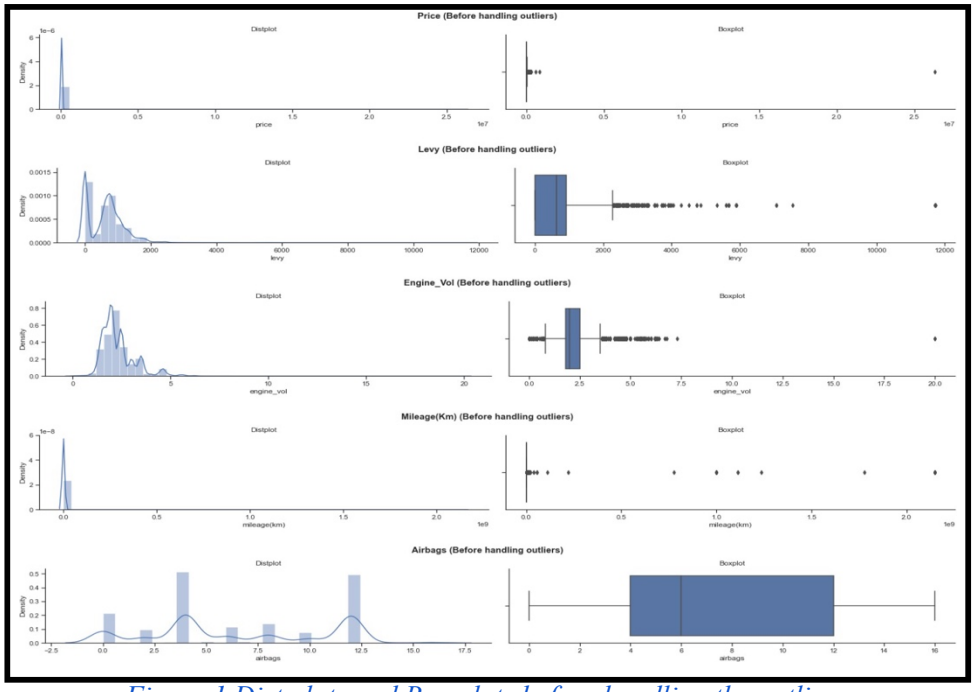

Figure 1 Dist plots and Box plots before handling the outliers

Fig-1 shows the distribution plots and box plots of Price, Levy, Engine Volume, Mileage and Airbags before handling the outliers. The total outliers in price are: 1055 - 5.57%, in levy are: 160 - 0.85%, in engine_vol are: 1358 - 7.18%, in mileage(km) are: 635 - 3.36%, and in airbags are: 0 - 0.0%.

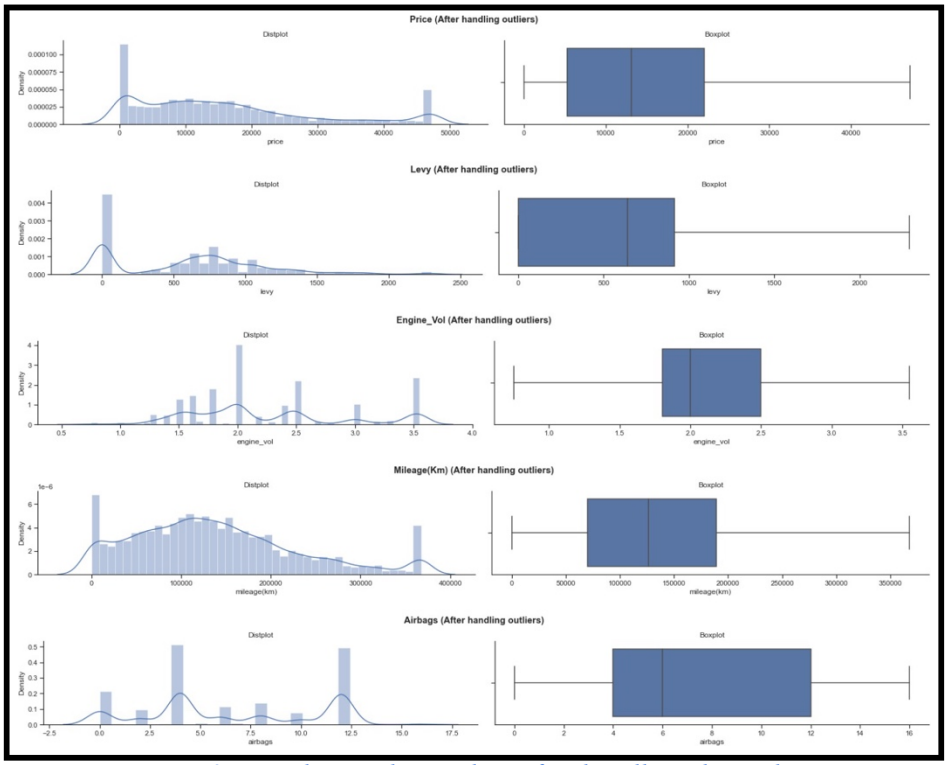

Figure 2 Dist plots and Box plots after handling the outliers

Fig-2 shows the distribution plots and box plots of Price, Levy, Engine Volume, Mileage and Airbags after handling the outliers. Mileage, levy, and the price are right skewed. The variety of cars for values of airbags are very high. The variety of cars for particular values of engine volume are very high.

4.3 Data Visualization

The data was visualized using various plots like heat map, box, distant, and others. From data visualization, potential important features like distributions of data, local patterns, gaps, missing values, outliers and correlation between target variables and so on were identified.

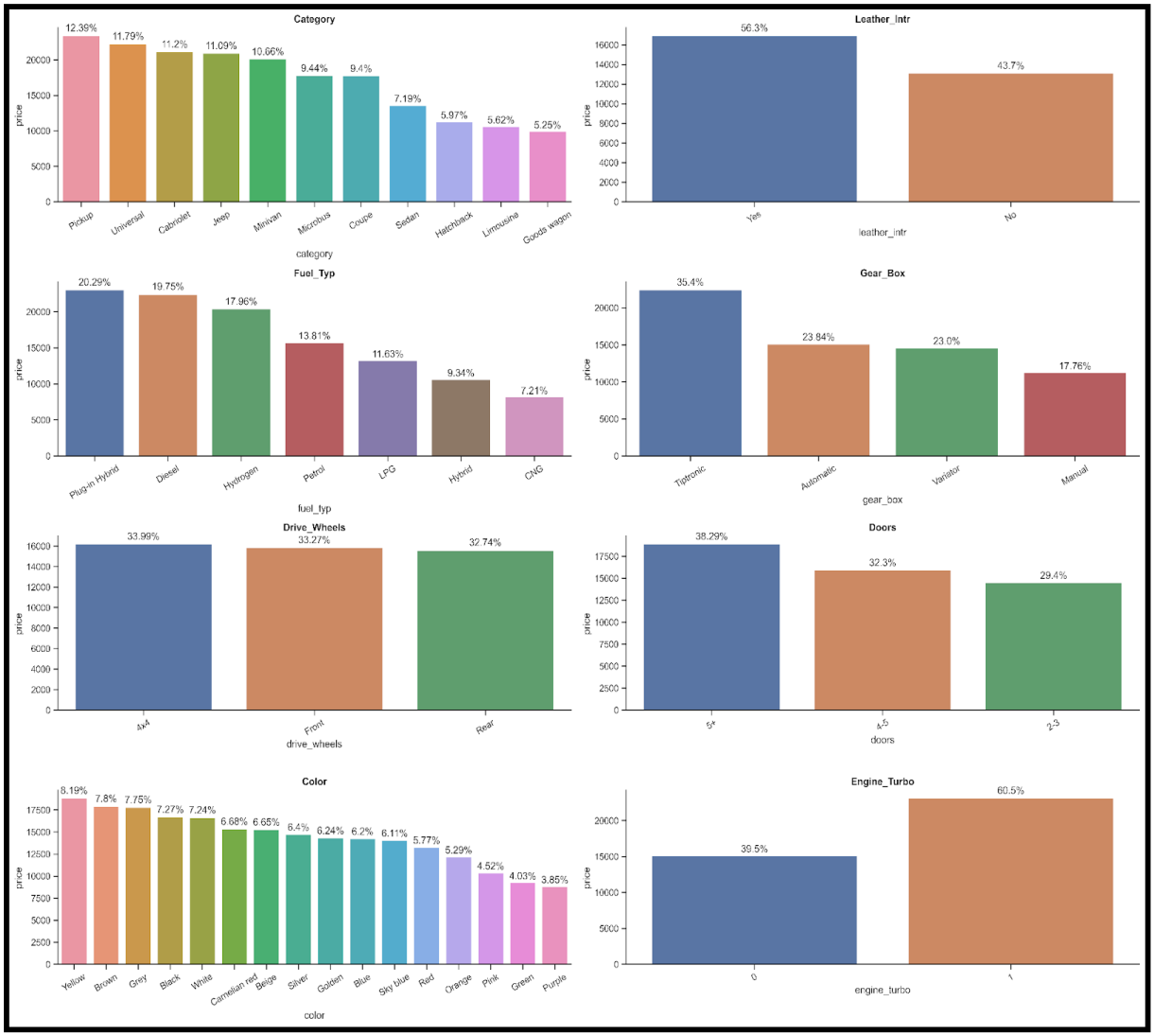

Figure 3 Mean price wise analysis of each feature

Fig-3 shows the mean price wise analysis of each feature**.** Plug-in Hybrids have the highest average price, while CNGs have the lowest. In comparison, the average price of a Tiptronic gearbox is the highest and the lowest for a manual gearbox. There is almost no difference in prices between 4x4, front, and rear drive wheels. In general, 4x4, front-drive, and rear-drive wheels cost about the same. The average price of 5+ doors is maximum and for the 2-3 doors the average price is the lowest. Purple color

averages the lowest price, and yellow color is at its highest average price. Cars with turbo engines are more expensive on average.

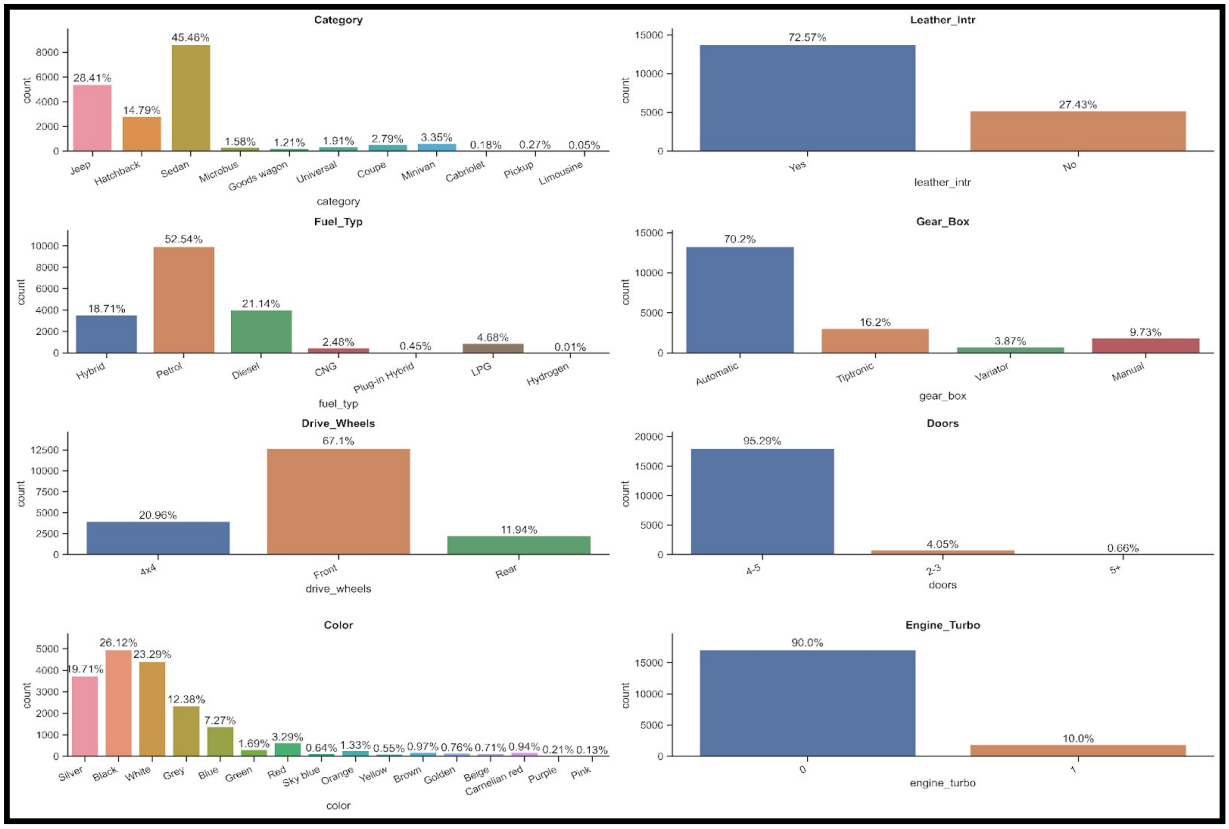

Figure 4 Bar pot for the variety of cars

Fig-4 shows a bar plot of a variety of cars. Sedans make up about 45.44 % of the market, Jeeps are 28.42 %, Hatchbacks are 14.79%, and all other categories have market shares of less than 5%. The fuel type of 52.55% of cars is petrol, 21.14% is diesel, and 18.7% is hybrid, while other fuel types such as CNG, LPG, and hydrogen, and plug-in hybrids are very rare. About 67% of car variants have front drive wheels. Most car variants come with leather interiors. In about 70% of car variants, the transmission is automatic. About 95% of car variants have 4-5 doors. About 90% of car variants have a turbo engine. Most car variants are black (26.13%), white (23.29%) and (19.71%).

4.4 Model Selection

Model selection was used to best describe the relationship between the variables in the data set. The process involved identifying the model that strikes the right balance between simplicity and accuracy which was an important step in the data modeling process.

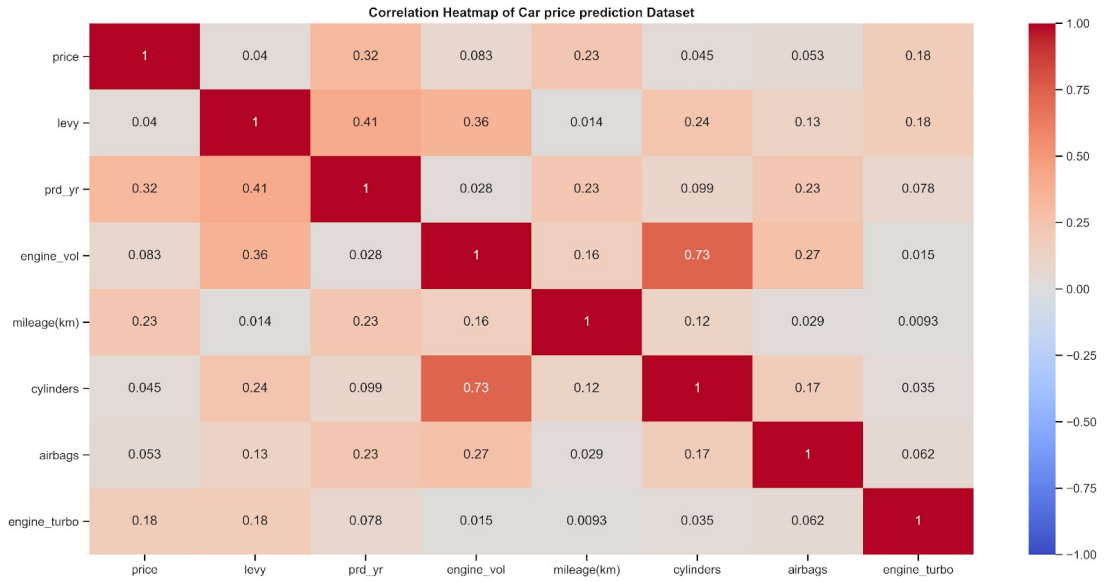

Figure 5 Correlation Heatmap of all the target variables

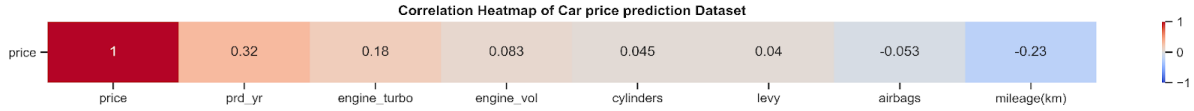

Figure 6 Correlation Heatmap of continuous variables

Fig-5 and Fig-6 are the correlation heatmaps. The price has a high positive correlation with production year, and engine turbo but has a high negative correlation with mileage. It also has a very little correlation with engine volume, cylinder, levy, and airbags. The production year and engine turbo have a high positive correlation to the price followed by engine volume, cylinders and levy. The airbags and mileage have negative correlation to the price.

After visualization, feature extraction was performed. This compresses the amount of data into manageable quantities for algorithms to process. This process was performed using Variance Inflation Factor (VIF) to check for multicollinearity and dropped columns which showed high VIF for the column Cylinder which was dropped.

```
VIF without removing anything:
                    VIF
Variables
levy           2.931734
prd_yr        17.751825
engine_vol    28.882986
mileage(km)    3.415502
cylinders     33.471335
airbags        3.632481
engine_turbo   1.164048
=========================================||
VIF after removing 'cylinders'
                    VIF
Variables
levy           2.930144
prd_yr        13.409287
engine_vol    14.852612
mileage(km)    3.415468
airbags        3.626076
engine_turbo   1.159125
```

Further, an appropriate model to represent the relationship between the dependent and independent variables was selected to check the goodness of the fit.

Forward Selection:

```
##
## Residual standard error: 9941 on 18861 degrees of freedom
## Multiple R-squared:  0.4531, Adjusted R-squared:  0.4515
## F-statistic: 274.2 on 57 and 18861 DF,  p-value: < 2.2e-16
```

Backward Elimination:

```
## Residual standard error: 9936 on 18857 degrees of freedom
## Multiple R-squared:  0.4538, Adjusted R-squared:  0.452
## F-statistic: 256.8 on 61 and 18857 DF,  p-value: < 2.2e-16
```

Even after Forward Selection and Backward Elimination, the R squared value of the model hasn't changed much.

4.5 Model Evaluation

The regression model can be evaluated on following parameters:

1. Mean Square Error (MSE): MSE is the single value that provides information about the goodness of regression line. Smaller the MSE value, better the fit because smaller value implies smaller magnitude of errors.

2. Root Mean Square Error (RMSE): RMSE is the quadratic scoring rule that also measures the average magnitude of the error. It is the square root of the average squared difference between prediction and actual observation.

3. Mean Absolute Error (MAE): This measure represents the average absolute difference between the actual and predicted values in the dataset. It represents the average residual from the dataset.

QQ plot:

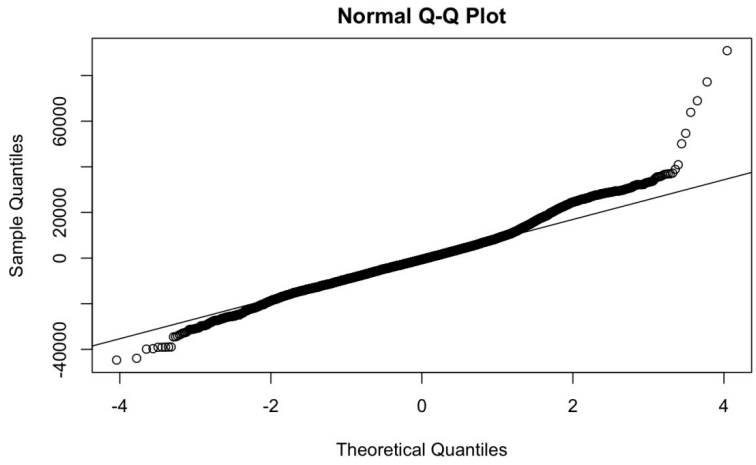

Figure 9 Normal Q-Q Plot

Fig-9 is the Normal Q-Q Plot that is Light-tailed, which means that the plotted dataset has fewer extreme values than the other dataset and is less likely to have outliers.

CONCLUSION

Using Linear Regression techniques, this project proposed a scalable framework for US based used car model analysis. An efficient linear regression model was developed by training, testing and evaluating the dataset. As a result, the Mean Squared Error (MSE) is 96803636.48, Mean Absolute Error (MAE) is 7443.95, Root Mean Squared Error (RMSE) is 9838.88, Trained R Squared value is 0.45, Tested R Squared value is 0.46 and Adjusted R Squared value is 0.45. The linear regression model gave a decent R Squared value, but the model can be further improved by using other Machine Learning Algorithms.

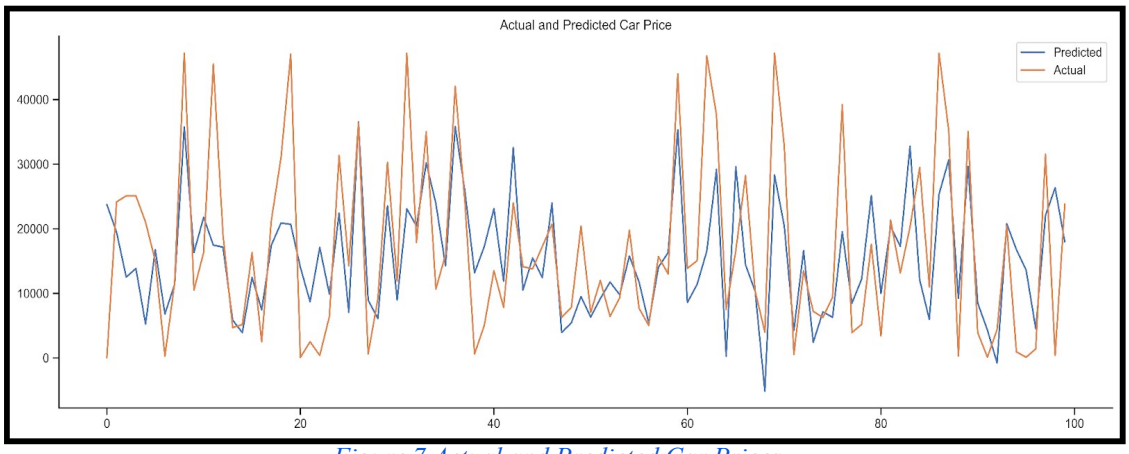

Figure 7 Actual and Predicted Car Prices

Fig-7 shows the comparison of actual and predicted car prices

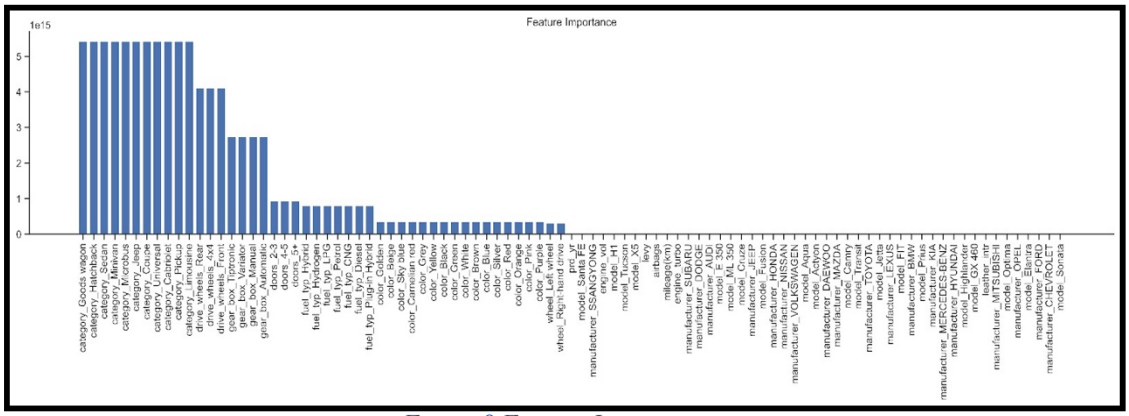

Figure 8 Feature Importance

Fig-8 shows the features of the car that played an important role in determining the prices

REFERENCES

1. kaggle.com/dataset.
2. Linear Regression in R by Julian Faraway.
3. https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/
4. https://stackoverflow.com/questions/66443371/standardize-or-normalize-categorical-values.
5. N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, 2018, pp. 115-11