PROJECT REPORT

USED CAR PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES

submitted in partial fulfillment of the criteria for award
of
POST GRADUATE PROGRAM
in
DATA SCIENCE AND ENGINEERING

by

PRASSANTH E	**(VX3WJ2477S)**
PRATIK ASARKAR	**(45BA1HZ8S1)**
PRIYANKA PARLIKAR	**(YG5GCUACL6)**
RUDRA PRATAP SEN	**(G53ZF1NIFJ)**
ANN MARIA JOHN	**(MYDEW11NAD)**

Under the supervision of
Mr. Srikar Muppidi

GREAT LAKES INSTITUTE OF MANAGEMENT
Bangalore – 560 102, INDIA

JULY 2020

GREAT LAKES
INSTITUTE OF MANAGEMENT, CHENNAI

greatlearning
Learning for Life

ABSTRACT

Summary:

A car price prediction has been a high interest research area, as it requires noticeable effort and knowledge of the field expert. Considerable number of distinct attributes are examined for the reliable and accurate prediction. To build a model for predicting the price of used cars in Germany, machine learning techniques like multiple linear regression, ridge regression, lasso regression, elastic net regression, decision tree regressor, random forest regressor, KNN regressor, gradient boosting and artificial neural networks were applied. The predictions are based on data scraped from used car listing on Ebay-Kleinanzeigen (German), retrieved from Kaggle (https://www.kaggle.com/orgesleka/used-cars-database#autos.csv). Respective performances of different algorithms were then compared to find one that best suits the available data set.

Conclusion:

KNN Bagging Regressor proved its capability in generating a good prediction model with a better trade-off between bias and variance error. With hyper parameter tuning using grid search CV, it showed a better accuracy than the standard solution, multiple linear regression. Therefore, KNN Bagging Regressor was chosen as the final model.

Keywords: used car price prediction, machine learning, multiple linear regression, ridge regression, lasso regression, elastic net regression, decision tree regressor, random forest regressor, KNN regressor, gradient boosting and artificial neural networks, grid search CV.

ACKNOWLEDGEMENT

First and foremost, we praise and thank almighty God for showering his perennial blessing on us and for giving us the courage all through the execution of our work.

At the outset, we are indebted to our Mentor Mr. Srikar Muppidi for his time, valuable inputs and guidance. His experience, support and structured thought process guided us to be on the right track towards completion of this project.

We are extremely gifted and fortunate to have Ms. Neha N. as our Academic counsellor. Her in-depth knowledge coupled with her passion in delivering the subjects in a lucid manner has helped us a lot. We are thankful to her for her guidance towards entire coursework.

We also thank all the course faculty of the DSE program for providing us a strong foundation in various concepts of analytics & machine learning.

Last but not the least, we would like to sincerely thank our respective families for giving us the necessary support, space and time to complete this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Prassanth E

Pratik Asarkar

Priyanka Parlikar

Rudra Pratap Sen

Ann Maria John

Date: 28th Jul, 2020

Place: Bangalore

CERTIFICATE OF COMPLETION

I hereby certify that the project titled "Used Car Price Prediction Using Machine Learning Techniques" was undertaken and completed under my guidance and supervision Prassanth E, Pratik Asarkar, Priyanka Parlikar, Rudra Pratap Sen and Ann Maria John, students of the December 2019 batch of the Post Graduate Program in Data Science & Engineering, Bangalore.

Mr. Srikar Muppidi

Date: 28th Jul, 2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

S. No.	***Abbreviation***	***Detail***
1.	*KNN*	*k-nearest neighbors*
2.	*ANN*	*Artificial Neural Networks*
3.	*ANOVA*	*Analysis of variance*
4.	*VIF*	*Variance Inflation Factor*
5.	*Q-Q Plot*	*Quantile-Quantile Plot*
6.	*RMSE*	*Root Mean Squared Error*
7.	*NB*	*Naive Bayes*
8.	*DT*	*Decision Tree*
9.	*RF*	*Random Forest*
10.	*GB*	*Gradient Boosting*
11.	*OHE*	*One Hot Encoding*
12.	*CV*	*Cross Validation*
13.	*ACF*	*Auto-Correlation Function*

EXECUTIVE SUMMARY

Background and need for study: Due to the increased price of new cars and the incapability of customers to buy new cars due to the lack of funds, used cars sales are on a global increase. But, the challenging part for used car dealers is to predict the price of used cars with accuracy. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. The increase in e-commerce usage over the past few years has created potential in the used car market, enabling the used car dealers to reach out to a huge customer base. The target business is online used car dealers and the target customers are buyers who would like to purchase used cars via online portal.

Scope and objectives of the study: This project aims to analyze how features of a used-car influence its market price and to predict the price based on the car features in the given data. The final product of the project are machine learning models that can predict market value estimation of a used car given its features. The organized sector is dominating the German used car market due to higher network chains with a record of satisfactory relationship with their customers. This gives the organized dealerships an edge over the unorganized dealers in terms of enhanced quality of documentation process, certified inspection and many other factors. Additionally, the multi-brand dealerships recorded a larger market share in the organized used car market. To predict the price of used cars, the business would require pointers about various features pertaining to the used car. The study will aid in price prediction based on the trained model on the used car dataset and give insights on how the price varies depending on the features.

Approach and Methodology: The data is scraped from used car listing on Ebay-Kleinanzeigen (German), retrieved from Kaggle. After processing the dataset and cleaning the inconsistencies, the numerical and categorical features used in the purchasing intention prediction model is generated. Various Regression algorithms are used to predict used car price based on set of independent variables like brand, model, kilometer, powerPS, FuelType, VehicleType, etc. The predictive models are also used to identify the variables that strongly influence the price using variable importance and probabilistic approaches. The models are evaluated using relevant model performance measures to arrive at the most robust models for prediction.

Key Learnings: The data obtained from Kaggle convey important information about the features of a used car, and how they influence its market price.

Recommendations & actionable insights: The high-level recommendations for the project are developed by predicting used car prices. These are then linked to the model findings to recommend actionable insights.

CHAPTER 1 – PROJECT OVERVIEW

1.1.Problem Statement

The challenging part for used car dealers is to predict the price of used cars with accuracy. The prices of new cars in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But due to the increased price of new cars and the incapability of customers to buy new cars due to the lack of funds, used cars sales are on a global increase. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a used car's actual market value. It is important to know their actual market value while both buying and selling.

1.2.Need of Study

The increase in e-commerce usage over the past few years has created potential in the used car market, enabling the used car dealers to reach out to a huge customer base. To be able to predict used cars market value can help both buyers and dealers (sellers in the organized used car market).

1.2.1. Target Business: Online User-Car Dealers

They are one of the biggest target group that can be interested in results of this study. If used car dealers better understand what makes a car desirable, what the important features are for a used car, then they may consider this knowledge and market products with better price. E-commerce is being preferred over other businesses as the proportion of revenue investment, which is apparently better than other options and requires less investment.

1.2.2. Target Customers

Buyers who would like to purchase used cars via online portal, wherein, it's a big corner to pay too much or sell less then it's market value.

Hence, we can see that estimating the price of used cars is of very high commercial importance.

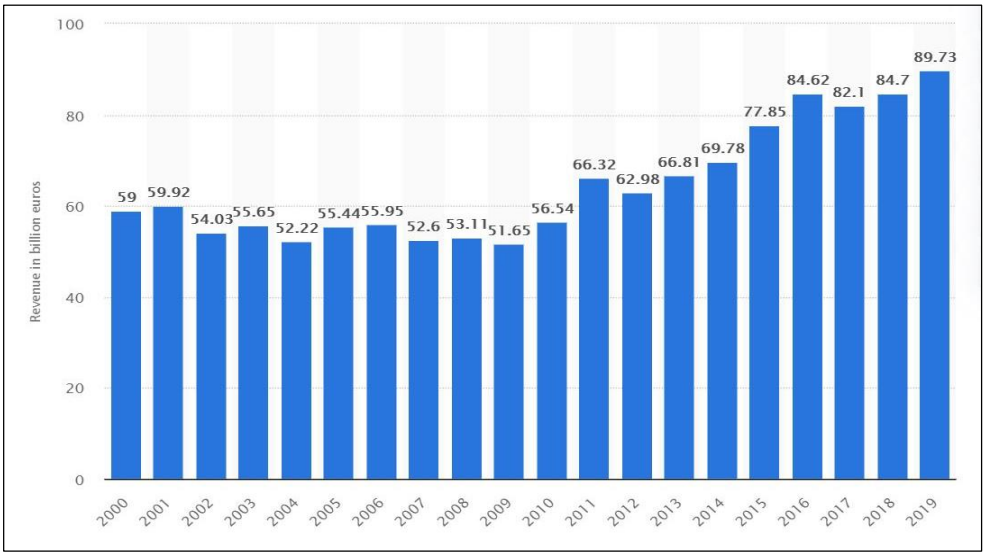

Fig. 1: Revenue of the used car market in Germany between 2000 and 2019

1.3.Scope of Study

The organized sector is dominating the German used car market due to higher network chains with a record of satisfactory relationship with their customers. This gives the organized dealerships an edge over the unorganized dealers in terms of enhanced quality of documentation process, certified inspection and many other factors. Additionally, the multi-brand dealerships recorded a larger market share in the organized used car market as multi-brand used car dealers have various brands and models available with them and the consumers have the choice of comparing and then purchasing the used cars. Furthermore, the organized used car dealers have a larger geographical presence in the country.

To predict the price of used cars, the business would require pointers about various features pertaining to the used car. The study will aid in price prediction based on the trained model on the used car dataset and give insights on how the price varies depending on the features.

The objectives of the study are:

i. Perform data cleaning and visualization, in order to reach an elementary understanding of each car feature and its influence on the market price.

ii. Build and evaluate models using machine learning algorithms for price prediction in order to provide a real-time used car evaluation service.

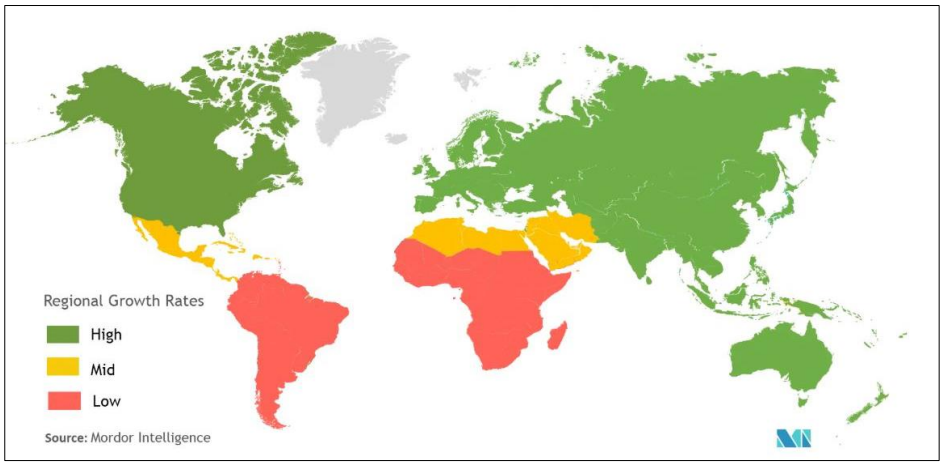

Fig. 2: Used Car Market – Growth Rate by Region (2020 – 2025)

1.4.Complexity Involved

The dataset has some complexity which needs to be resolved in order to get better results of the predicted price. The complexity in the dataset includes:

- The dataset has many outliers and missing values that need to be treated.
- High number of categories for features like model was difficult to handle as data would be spread over a large area, and so encoding techniques were done.

1.5.Data Sources

In order to predict used car prices, the data was retrieved from Kaggle (https://www.kaggle.com/orgesleka/used-cars-database#autos.csv). It contains 20 features with 371528 raw observations, scraped from used car listing on Ebay-Kleinanzeigen (German). Each record belongs to a unique used car.

1.6.Dataset Description

The dataset consists of both numerical and categorical variables.

The below table consists of numerical features and their description.

Feature Name	Feature Description	Min.	Max.	Std. Deviation
Price	The price on the ad to sell the car	0	2147484000	3587954
Year of Registration	Year in which the car was first registered	1000	9999	92.866598

Power (PS)	Power of the car in PS	0	20000	192.139578
Kilometer	How many kilometers the car has driven	5000	150000	40112.337051
Month of Registration	At which month the car was first registered	0	12	3.712412
No. of pictures	Number of pictures in the ad	0	0	0
Postal code	Postal code of the city where the car is available.	1067	99998	25799.08247

Table 1: Numerical features used in the price prediction model

Table 1 shows the seven numerical features along with their statistical parameters. Among these, Price is monetary values in Euros. PowerPS represents the power of the car in PS (Pferdestärke, German) equivalent of horsepower or Torque (1 pferdestarke = 0.9863200706195 horsepower).

The below table consists of categorical features and their description.

Feature Name	**Feature Description**	**Number of Categorical Values**
Name	Name of the car	233531
Seller	Private or dealer	2
Offer type	Offer or Application	2
A/B Test	Test or Control	2
Vehicle type	Describes the type of Vehicle	8
Gearbox	Whether the car has manual or automatic gearbox	2
Model	The model of a car is the name used by a manufacturer to market a range of similar cars	251
Fuel type	The Fuel system on which the car runs	7
Brand	The Brand which the car belongs to	40
Not repaired damage	If the car has a damage which is not repaired yet	2

Table 2: Categorical features used in the price prediction model

Table 2 shows the thirteen categorical features along with the number of categorical values for each of them. Further we can divide above features into ordinal and nominal categorical features. The features which have inherent order in it are called ordinal features while features which do not have inherent order are called nominal features. In the context of this dataset, all the features are nominal categorical features.

The below table consists of date-time features and their description.

Feature Name	Feature Description	Unique Values	First Value	Last Value
Date crawled	When this ad was first crawled, all field-values are taken from this date	280500	2016-03-05	2016-04-07
Date created	The date for which the ad at EBay was created	114	2014-03-10	2016-04-07
Last seen online	When the crawler saw this ad last online	182806	2016-03-05	2016-04-07

Table 3: Date-time type features used in the price prediction model

Table 3 shows the three date-time features along with their statistical parameters. These features represent the time duration in the form of date.

1.7.Data Preparation

In the following section, we are going to experiment with the dataset by data cleaning, feature selection, feature extraction, and feature engineering.

1.7.1. Data Cleaning

1.7.1.1. Outlier Treatment

Outliers are data points that are far from other data points. In other words, they're unusual values in a dataset. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results. Let’s get a sense of how we handled outliers in our dataset.

i. Price

Looking into the real used car market in Germany, the prices of a vehicle could range from anywhere between 100 euros to 80,000 euros. But we did have our records which had price

range starting from 0 euros to as high as 2.147484e+09 euros. These values seemed practically impossible for any used car in the market and seemed to be some sort of data collection error. As these values above practical range (80,000 euros) were very low in numbers, we capped our price range between 100 euros to 1,00,000 euros thus getting rid of major chunk of outliers. We used isolation forest anomaly detection technique and could still see some outliers in our price range which we handled using log transformation technique to normalize the range.

ii. powerPS

The power of a German car could range from anywhere between 40 to 800 PS. But we did have our records which had power range starting from 0 PS to as high as 20,000 PS. These values seemed practically impossible for any used car in the market and seemed to be some sort of data collection error. As the values above practical range (800 PS) were very low in numbers, we capped our power range between 40 PS to 800 PS thus getting rid of major chunk of outliers. We used isolation forest anomaly detection technique for outlier detection.

1.7.1.2. Missing Value Imputation

A common problem when dealing with real-world data is missing values. These can arise for many reasons and have to be either filled in or removed before we train a machine learning model. First, let's get a sense of how many missing values are in each column.

Following are the features and their corresponding count of missing values in the dataset after price range was capped:

	count
notRepairedDamage	65197
vehicleType	32947
fuelType	28773
model	17703
gearbox	16621
state	183

Fig. 3: Count of missing values

From the above features, we decided to drop the missing values for the features 'notRepairedDamage', 'fuelType', 'model', 'gearbox' and 'state'. The reason for NOT imputing these values was the fact that all these variables were quite independent of any other features in the dataset, so imputing these values depending on other features would actually result in increase in multicollinearity between independent features. Secondly, as we have a

huge dataset at our disposal, we could afford losing some data as it would not cost us much, as far as information loss is concerned.

For the feature 'vehicleType', we did KNN imputation using model, brand and price as these feature come very close to predicting the type of vehicle.

1.7.2. Feature Selection

A major chunk of our feature selection was done based on the statistical analysis of the features. All of the features passed the significance test except for abtest for which we got a p-value greater than alpha (0.05).

Apart from that, features like seller and offerType were highly imbalanced and had no significant influence on price prediction due to the extreme imbalance so we kept it out from the final dataset.

Features dateCrawled, noOfpictures practically served no purpose in predicting the price of a used car. So these features were also kept out from the final dataset.

1.7.3. Feature Extraction

i. ageOfVehicle

Considering two categorical columns (yearOfRegistration and monthOfRegistration) seemed not a feasible solution to find out how old the car is. So we extracted a new feature from these two columns called "AgeOfVehicle" which severed as an additional continuous variable which described how old the car is in years. Additionally, we got rid of two variables from the dataset namely, yearOfRegistration and monthOfRegistration.

ii. No_of_days_online

This is one variable which could be quite tricky to understand. Our idea behind introducing this feature was that, as the number of days of a post/advertisement of a certain used car being online increases, there could be chance for an interested buyer to negotiate the price of the car he/she is interested in. So, to let the buyer know exactly the valuation of a certain car after being online for certain number of days, would be a helpful added entity. So using the columns dateCreated and lastSeen, we came up with the new feature No_of_days_online, thus getting rid of the former two features.

iii. State

This feature was extracted from the postalCode feature. It gave an additional support to price prediction depending on in which state the vehicle is available.

iv. CountryOfManufacture

This feature was extracted from the feature brand. Similar to the feature State, CountryOfManufacture also provided an additional support to the price prediction wherein it was estimated if the price of the used car varies depending on the country of manufacture

1.7.4. Feature Engineering

There were many categorical features in our dataset, so finding the best encoding technique for these variables was certainly a challenging task. We decided to divide the categorical features into high and low cardinality features. For low cardinality features, we used One Hot Encoding technique (OHE) as it did not add too many columns to our dataset. On the other hand, we had high cardinality features, which we could not deal with just using OHE as it could add high dimensionality to our data. So we went through the rigorous process for combining OHE for low cardinality features with some other different techniques for handling the high cardinality features.

Depending on the best RMSE scores, we decided to stick with a combination of OHE and k-fold target encoding technique for our final dataset.

CHAPTER 2 – EXPLORATORY DATA ANALYSIS

In the following section, we are going to experiment with visualization, in order to reach an elementary understanding of each car feature and its influence on the used car market price. The purpose of exploratory data analysis is two-fold:

- To understand the data in terms of price across various independent variables/features.
- Get insights on various features.

2.1.Univariate Analysis

i. Price

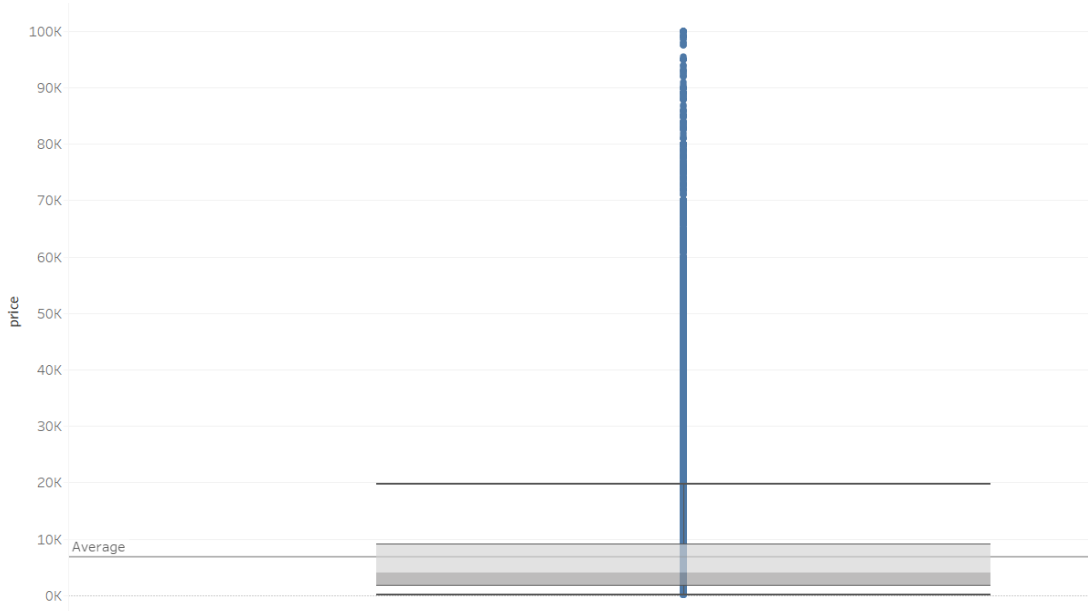

Fig. 4: Box-Whisker Plot: Price

ii. powerPS

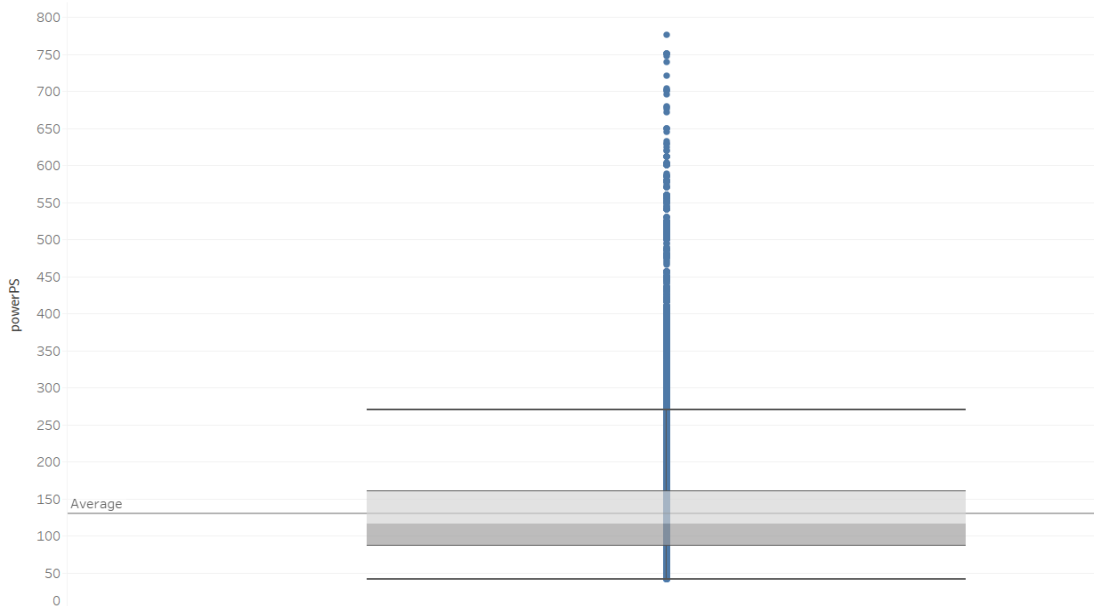

Fig. 5: Box-Whisker Plot: powerPS

Statistical Information	Price	powerPS
25% (lower quartile)	100	41
50% (Median – middle quartile)	4000	116
75% (upper quartile)	19,800	270
Mean	6847	129

Table 4: Statistical Information for Price and powerPS

iii. ageOfVehicle

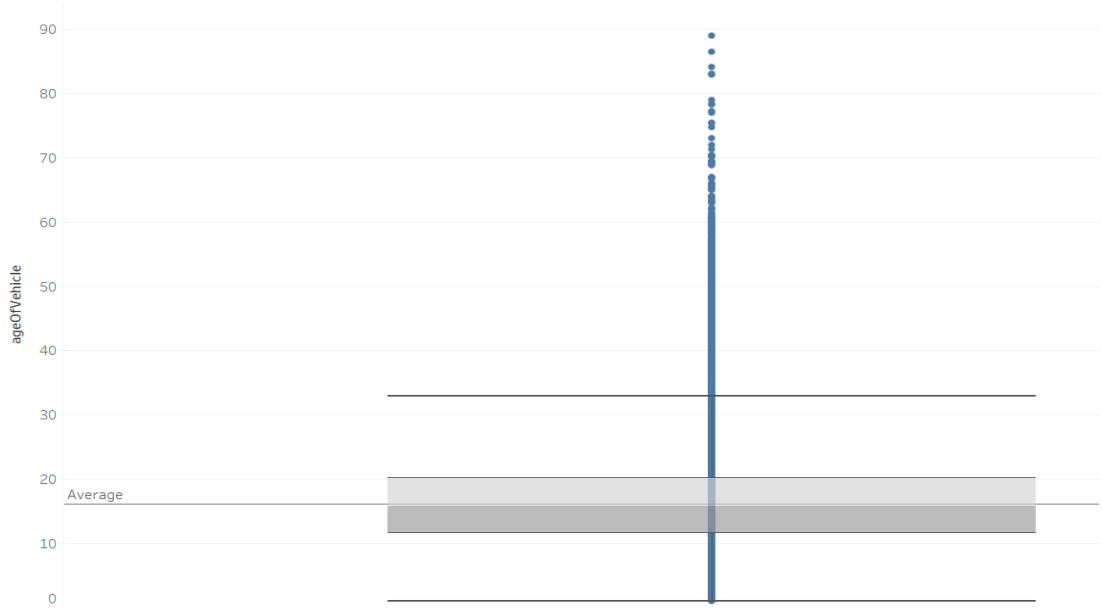

Fig. 6: Box-Whisker Plot: ageOfVehicle

iv. No_of_days_online

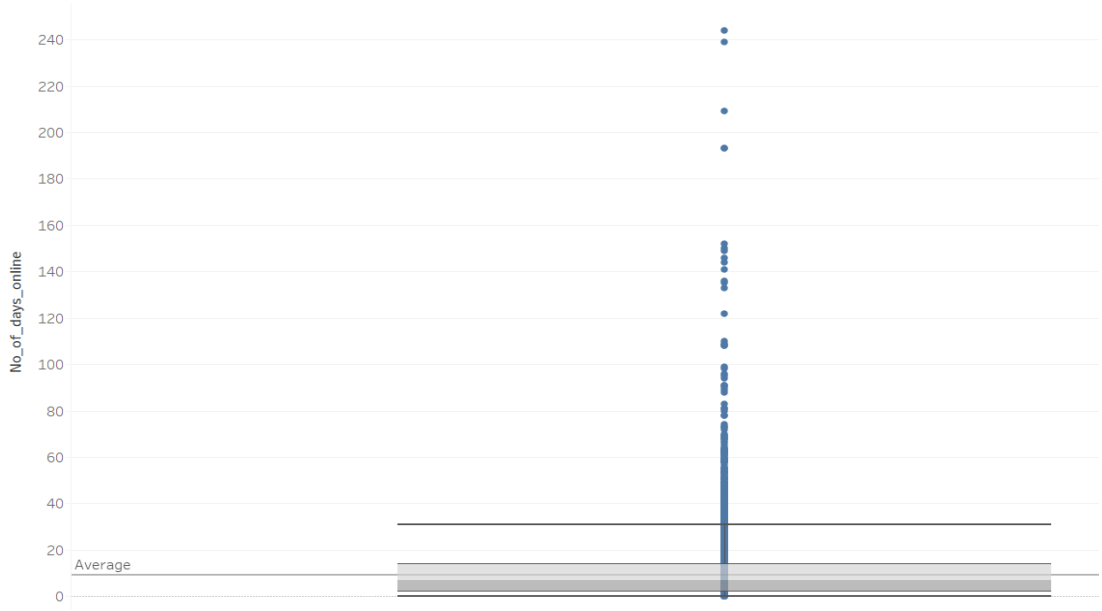

Fig. 7:Box-Whisker Plot: No_of_days_online

Statistical Information	ageOfVehicle	No_of_days_online
25% (lower quartile)	1.08	0
50% (Median – middle quartile)	15.75	7
75% (upper quartile)	33	31
Mean	16.10	9.24

Table 5: Statistical Information for ageOfVehicle and No_of_days_online

Inference: Transformation technique was applied to reduce the effect of outliers only for the target variable price.

v. Count of brand

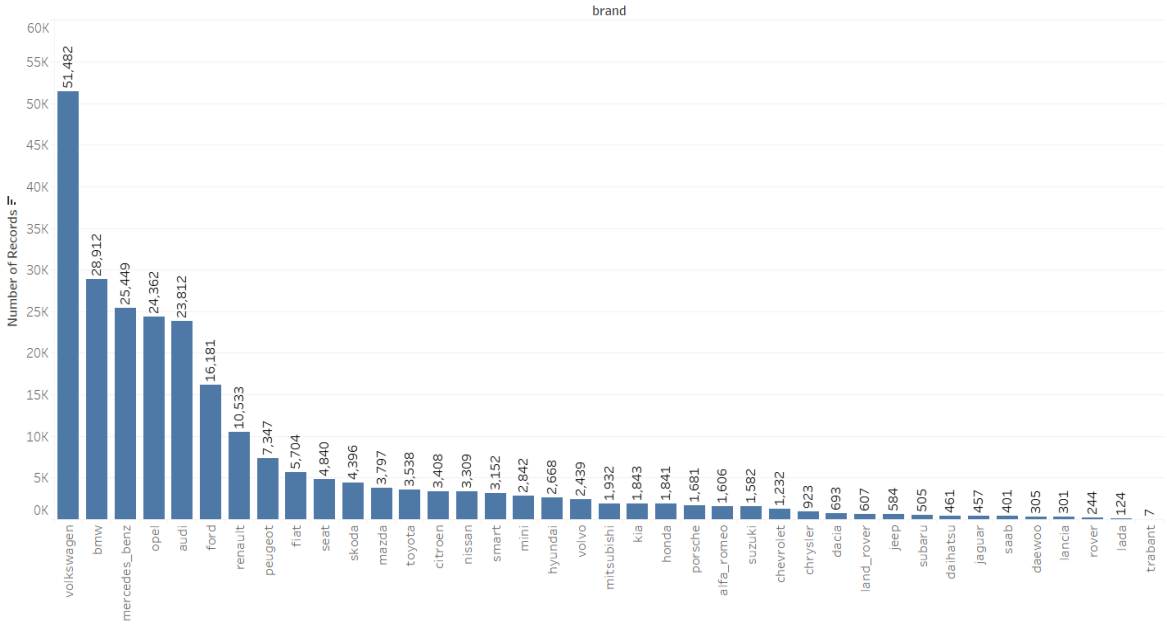

Fig. 8: Count of brand

Inference: Volkswagen, BMW, Mercedes Benz, Opel and Audi are the top most brands available in used car segment.

vi. Count of vehicleType

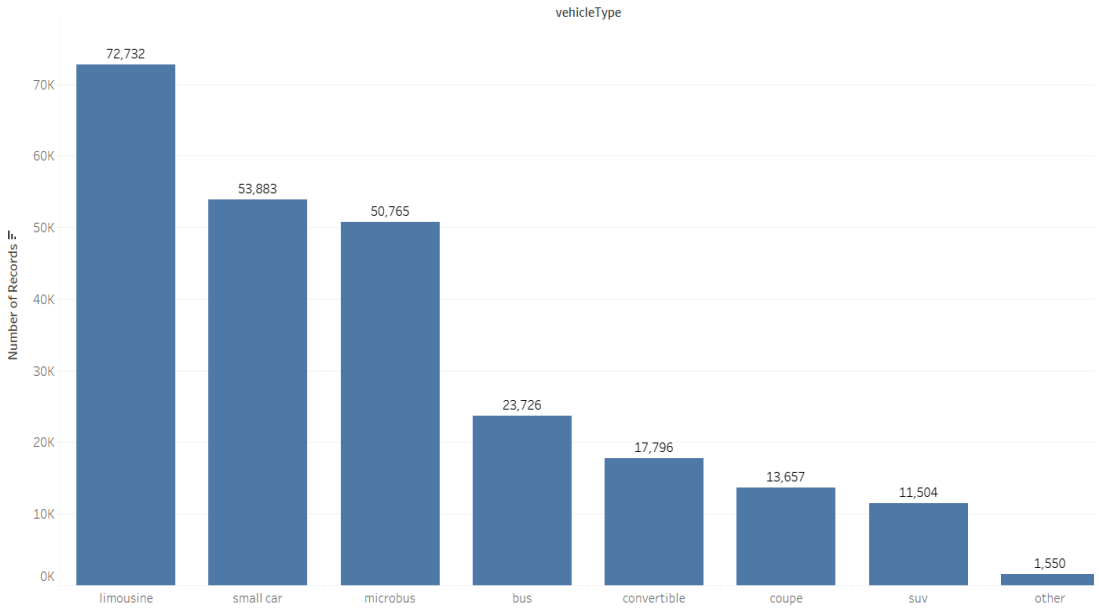

Fig. 9: Count of vehicleType

Inference: Limousine in Germany means Sedan vehicle type, from the count plot we can see that Limousine, small car and microbus are the most used vehicle type. Brands like BMW and Mercedes Benz sedan cars are used as taxi in Germany. Hence, Limousine are most in demand.

vii. Count of notRepairedDamage

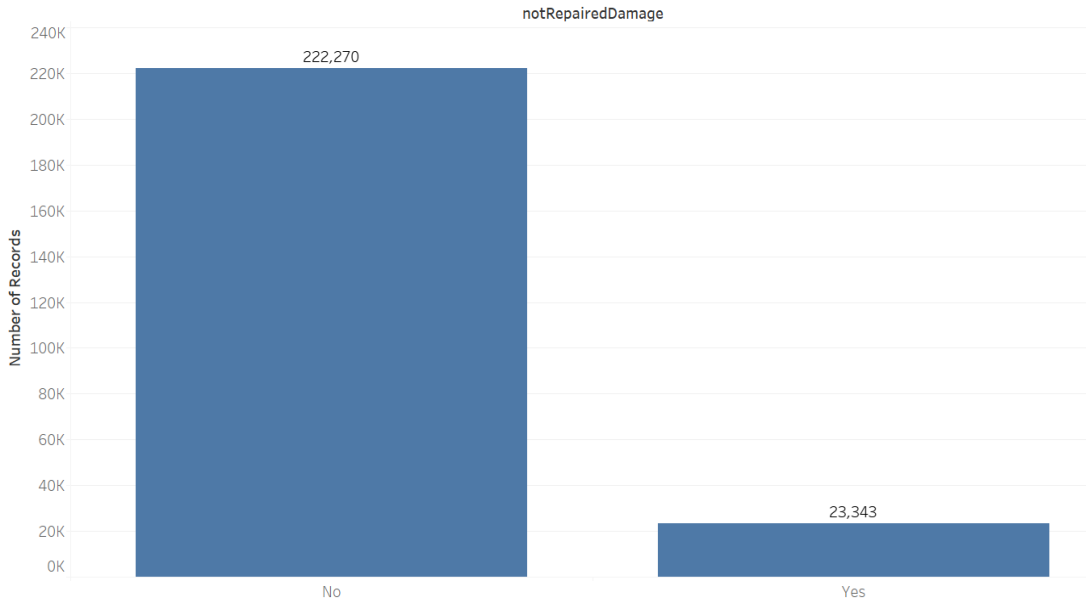

Fig. 10: Count of notRepairedDamage

viii. Count of fuelType

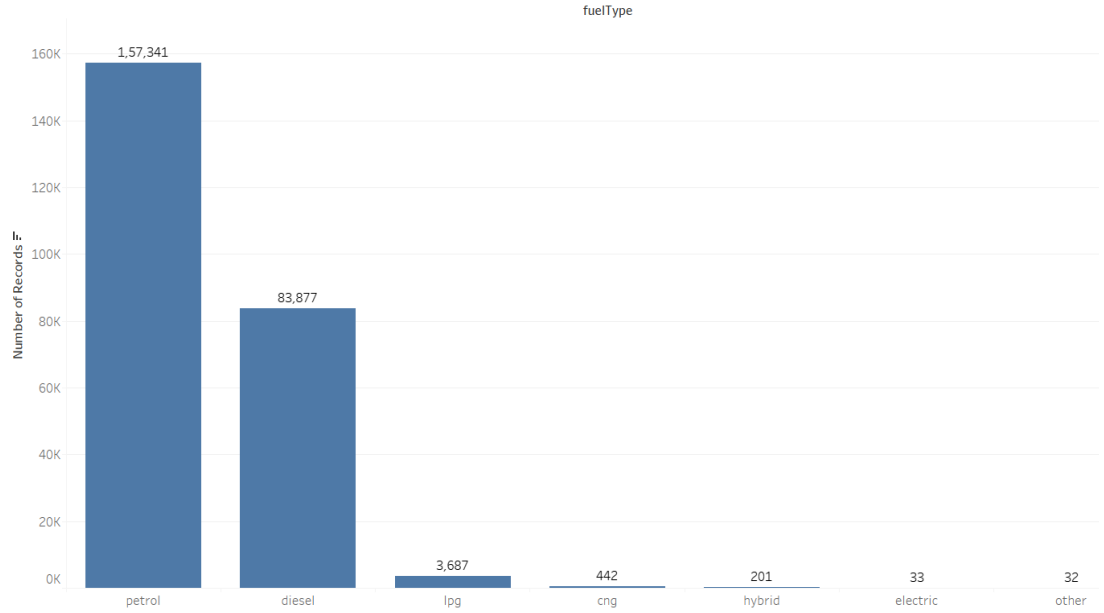

Fig. 11: Count of fuelType

Inference: Petrol and Diesel cars are most in demand and as the maintenance cost for petrol cars is less and more reliable compared to diesel cars, customers tend to buy petrol cars often.

2.2.Bivariate Analysis

i. Price v/s fuelType

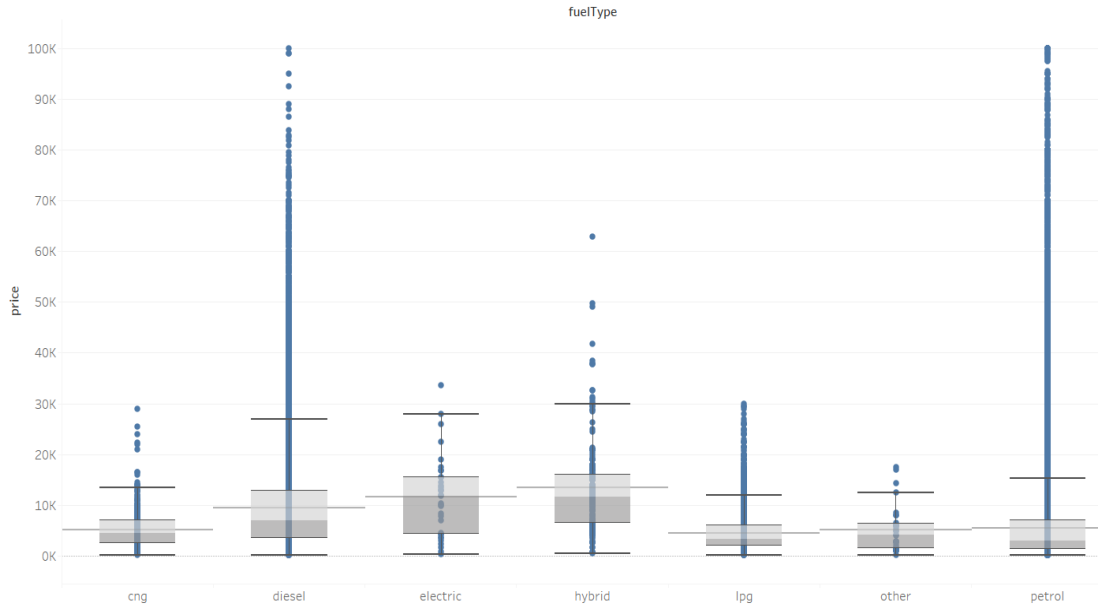

Fig. 12: Price v/s fuelType

Inference: Cars with different fuel type have different prices, from the plot we can see that cars of electric and hybrid fuel type are expensive as compared to other fuel types. Diesel and petrol fuel type cars belong to moderate range of price. CNG and LPG cars are the least expensive. The horizontal grey line shows the mean price of each fuel type cars, also p-value of fuel type is 0 when One-way ANOVA test was performed, hence fuel type of car becomes significant to predict price.

ii. Price Vs gearbox

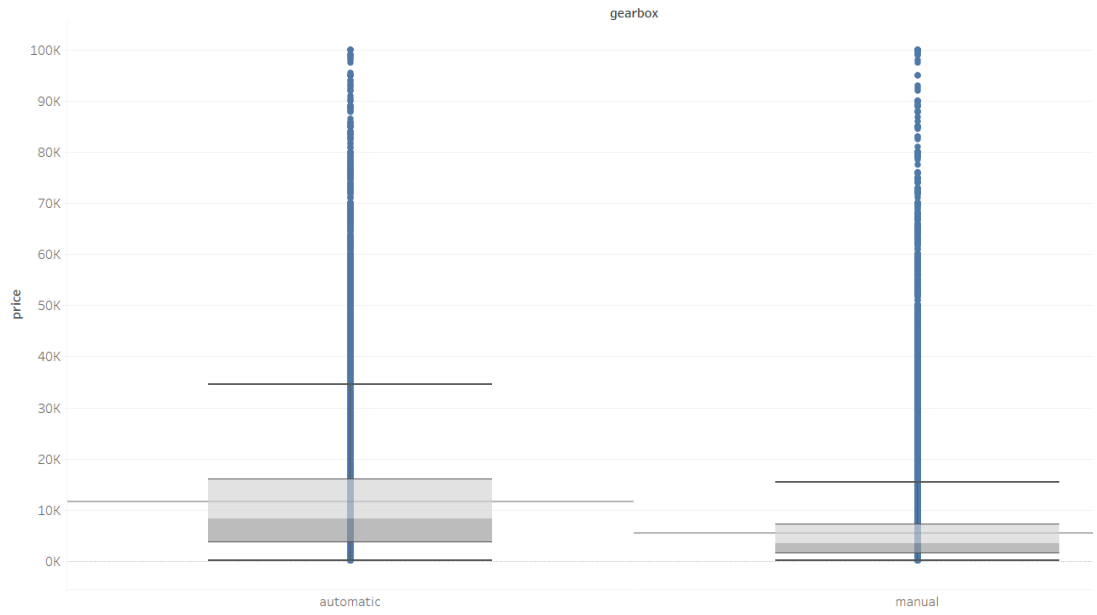

Fig. 13: Price Vs gearbox

Inference: Cars with automatic gearbox are of higher prices as compared to cars with manual gearbox. Mean price of cars with automatic and manual gearbox are different, p-value of gearbox is 0 which indicates that it is a significant feature in price prediction.

iii. Price v/s vehicleType

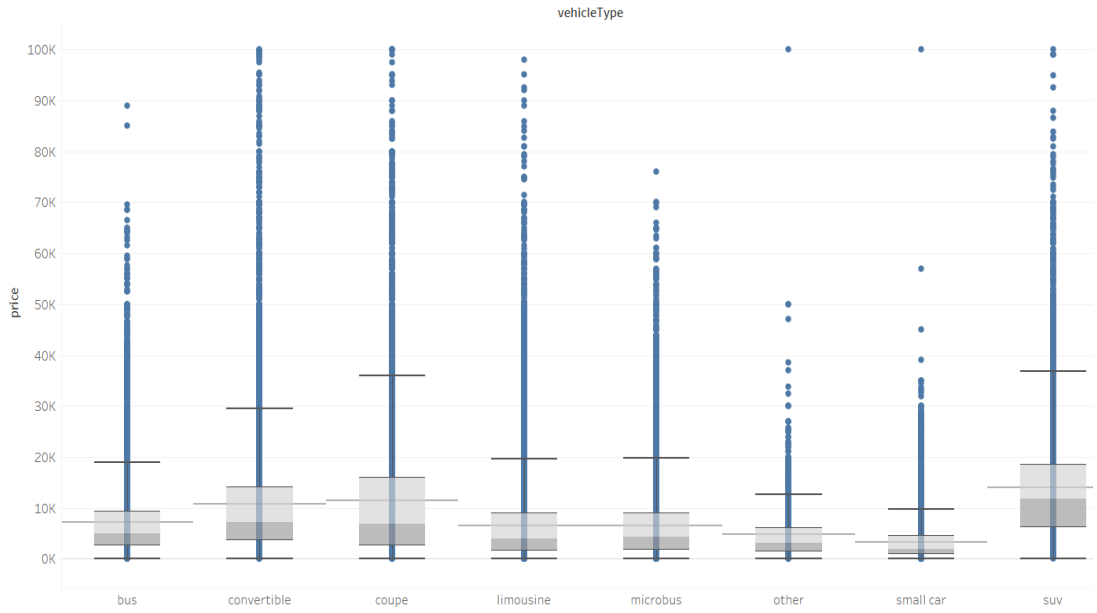

Fig. 14: Price v/s vehicleType

Inference: SUV, coupe and convertible vehicle types comes under higher price ranges. The mean price difference between convertible and coupe; limousine and microbus is less hence for price prediction one may consider only one vehicle type among them, though here we have considered each vehicle type distinctly.

iv. Price v/s notRepairedDamage

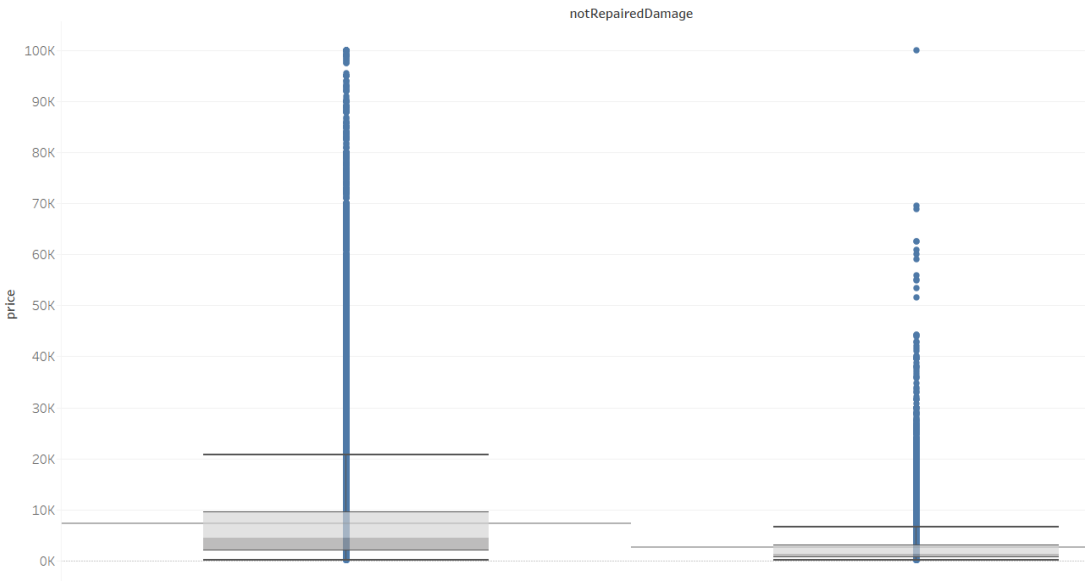

Fig. 15: Price v/s notRepairedDamage

Inference: Cars that do not have unrepaired damages have higher prices compared to the ones that have not repaired damages. Customers generally prefer cars in good condition over the ones that would bring about additional maintenance cost.

v. Average price of brands

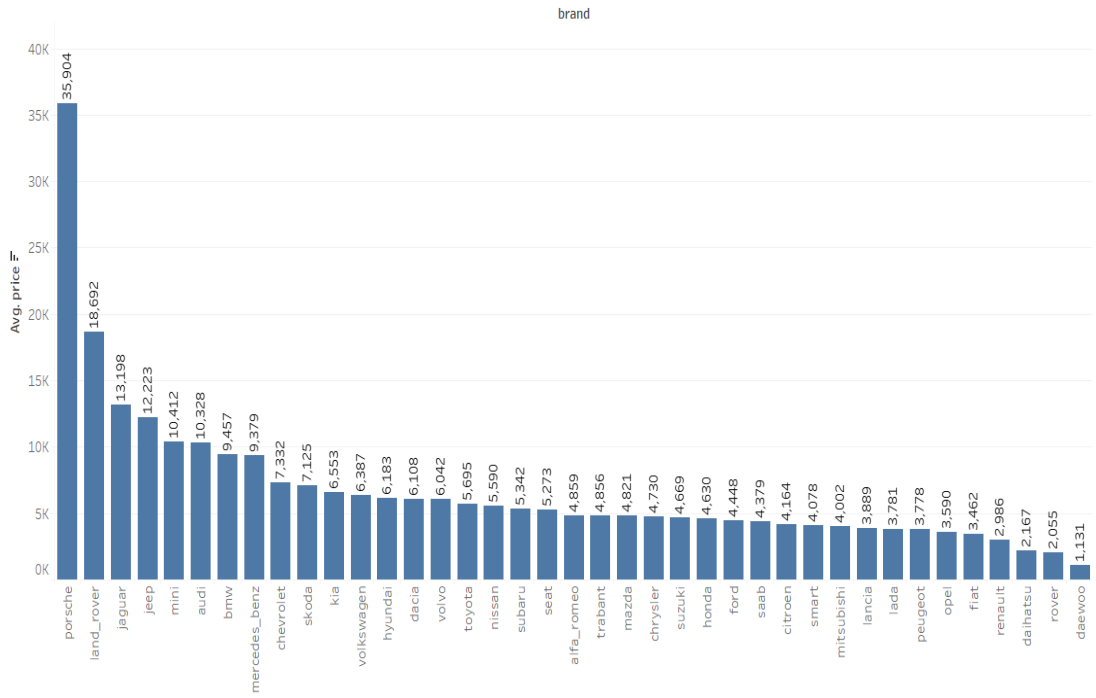

Fig. 16: Average price of brands

Inference: Porsche, land rover and jaguar top the brand list being the most expensive used cars.

vi. Average price v/s kilometer

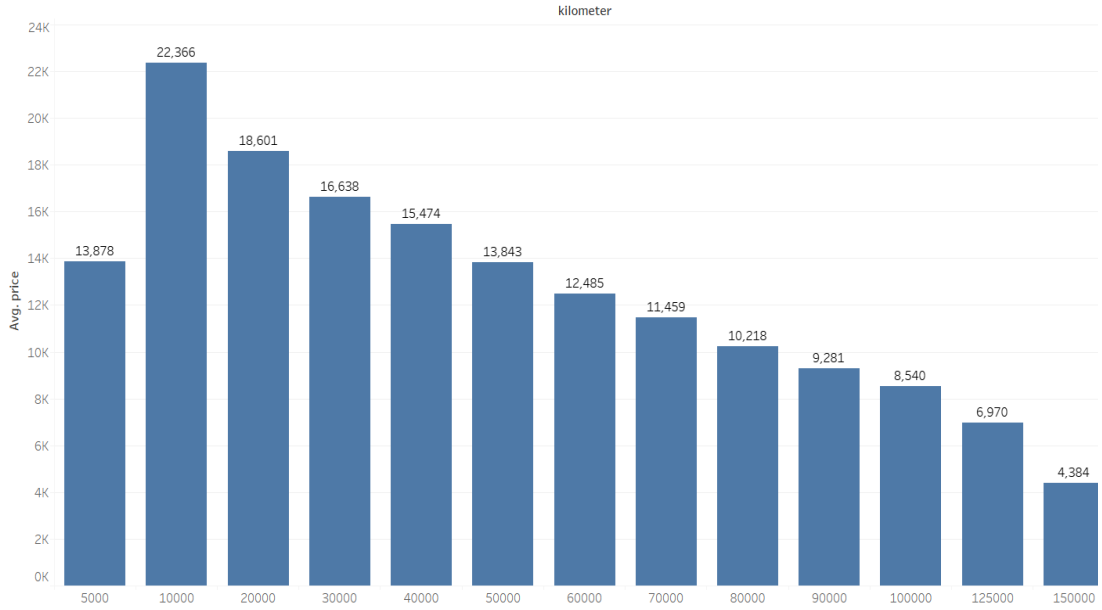

Fig. 17: Average price v/s kilometer

Inference: Here, we have considered kilometer as a categorical feature against average price. As the kilometers driven increases, average price decreases, this could probably mean that customers are keen on buying used cars that are less driven and are relatively in good condition. Also, we can see the average price for 5000 km is less as compared to 10000 km, there could be three possible reasons:

1. The data does not comprise much information about used cars that are driven 5000 kilometers.
2. The cars might have some not be repaired damages; that would result in lower prices.
3. Fuel type and vehicle type could rank higher when it comes to used car price determination.

vii. Price v/s ageOfVehicle

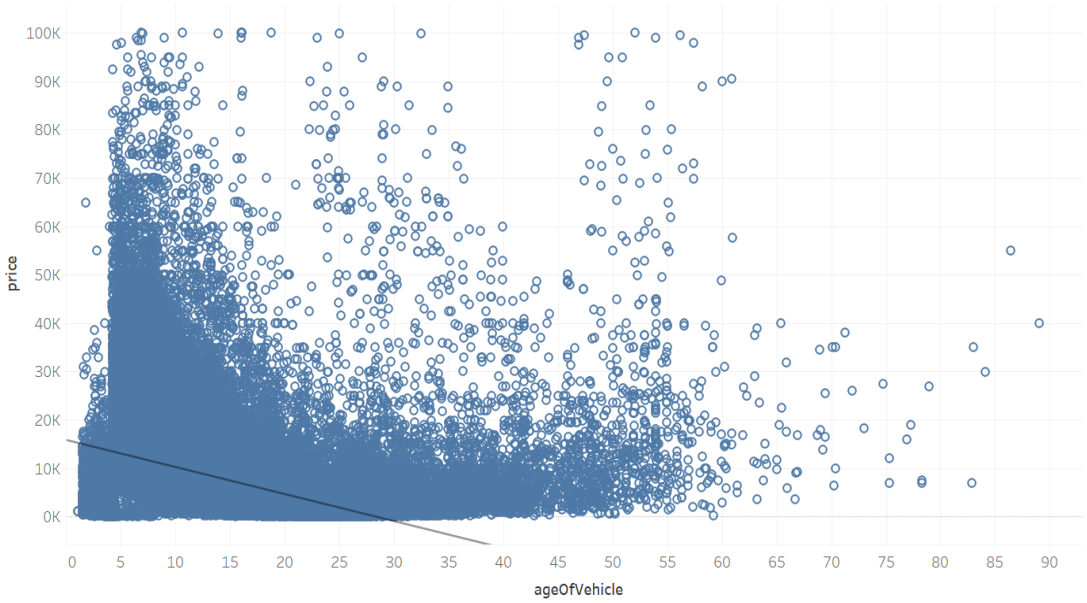

Fig. 18: Price v/s ageOfVehicle

Inference: The trend line in the above scatter plot indicates a strong negative correlation between price and age of vehicle. Exceptions include Porsche, Land Rover, Mercedes Benz and BMW that belong to luxury cars, wherein even though the age of vehicle is greater than 20 years, price would be higher shows the reliability of the German cars.

viii. Price v/s PowerPS

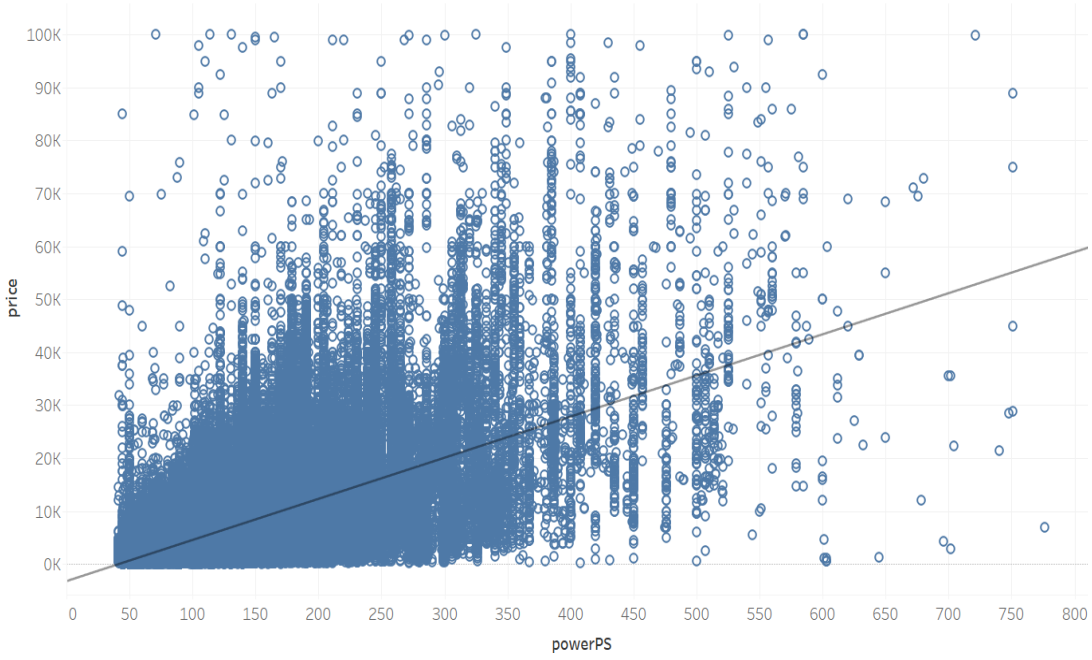

Fig. 19: Price v/s PowerPS

Inference: The above trend line shows a positive correlation between price and powerPS. Basic vehicle type like small cars and bus have less powerPS as compared to sport cars that has powerPS more than 300. Such high powerPS cars belongs to brands like Mercedes Benz, Porsche, BMW, Audi etc., that are well known for manufacturing sports cars.

ix. Average powerPS v/s brand

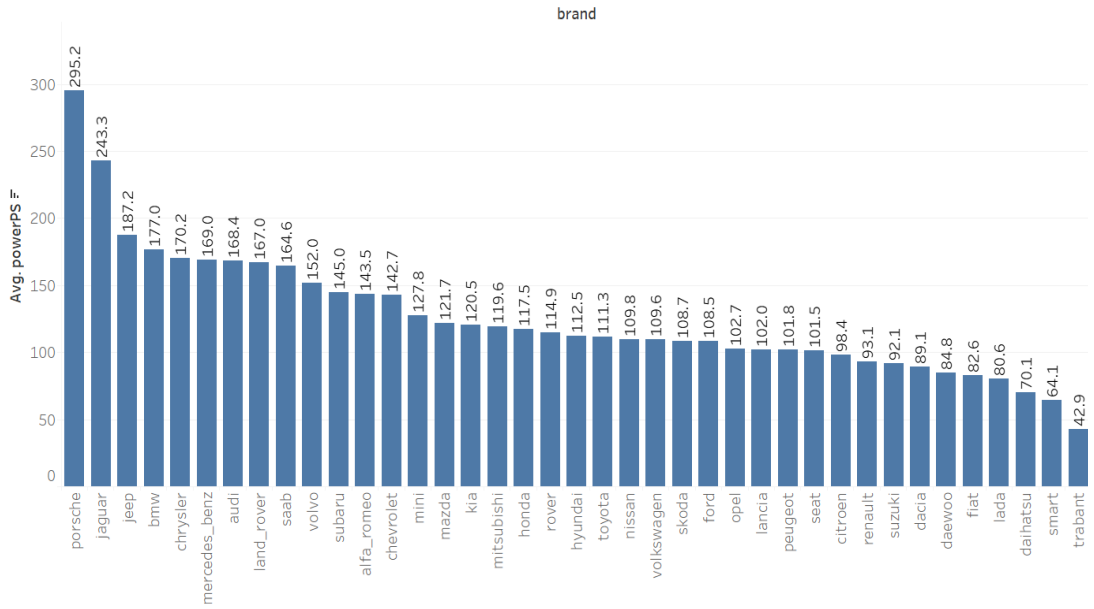

Fig. 20: Average powerPS v/s brand

Inference: As we have seen in earlier plot between average price and brand; Porsche, Land Rover, Jaguar, Jeep, BMW are one of those brands whose average price is higher as compared to other brands. The bar plot between average price and powerPS also states same brands justifying the positive correlation between price and powerPS.

x. Average powerPS v/s fuelType

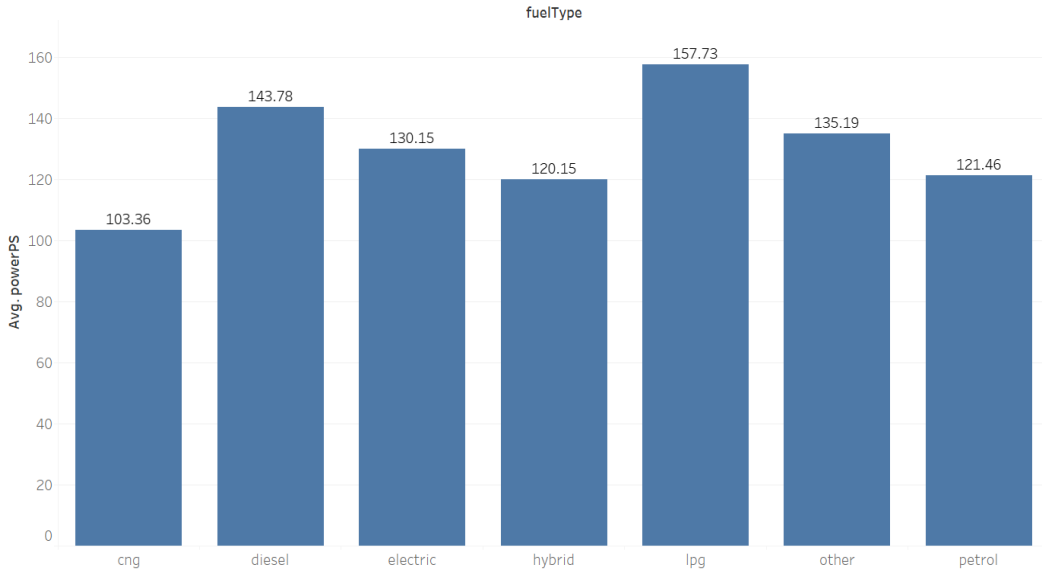

Fig. 21: Average powerPS v/s fuelType

Inference: As Germany is more likely to ban diesel and petrol cars in the country by 2030, LPG and electric cars seems to be an efficient alternative in terms of powerPS performance, also LPG cars are the cheapest cars, this could be one of the main reason as customers are selling diesel and petrol cars, hence available in huge amount in dealer's inventory. Diesel cars with automatic gearbox gives 2nd best average powerPS hugely comes within higher price ranges could be mainly sport cars.

xi. Average powerPS v/s gearbox

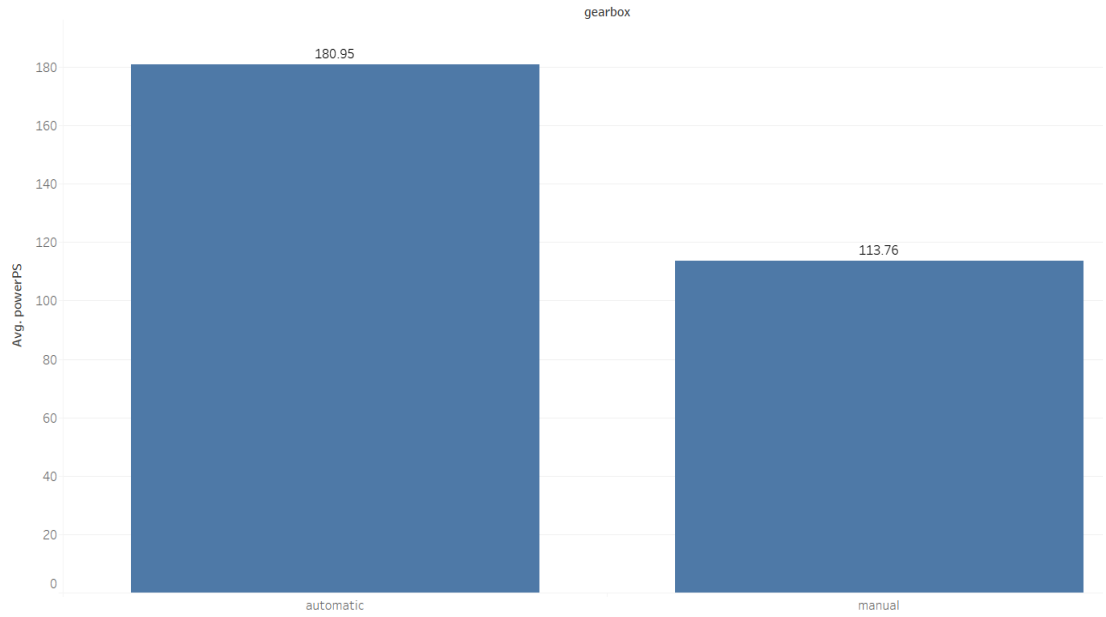

Fig. 22: Average powerPS v/s gearbox

Inference: Cars with low power generally have manual gearbox, whereas, mid-range cars will have an equal number of manual and automatic gearboxes. Nowadays, in many high range cars, manufacturers prefer automatic gearboxes because only then they can achieve their said 0-100 speeds. Additionally, launch control and gear response time will be better in cars with automatic gearboxes.

xii. Count of fuelType in terms of gearbox

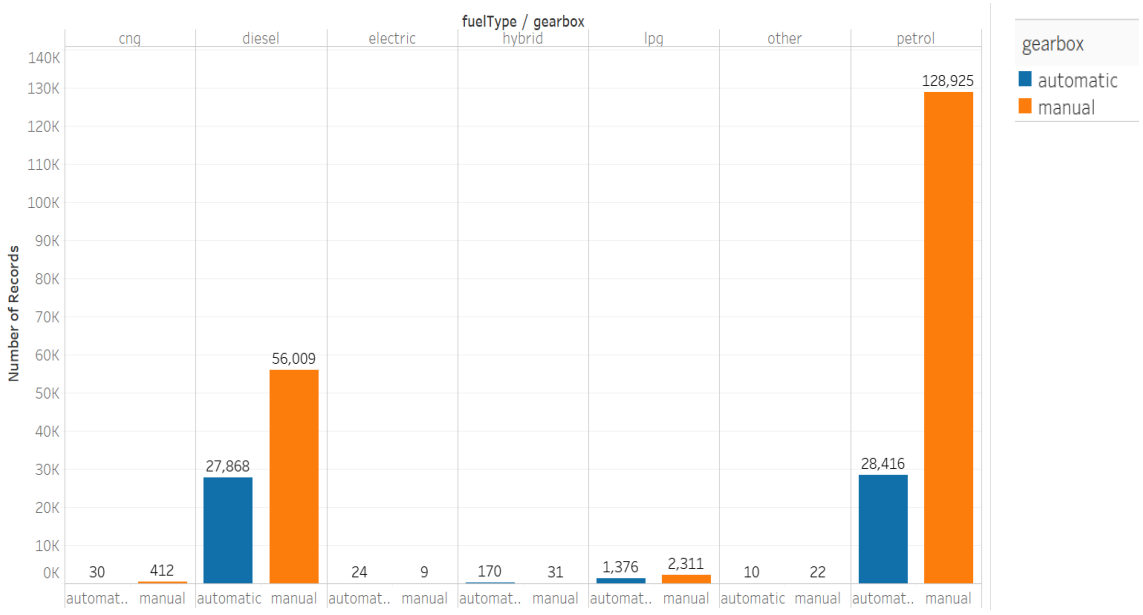

Fig. 23: Count of fuelType in terms of gearbox

xiii. Count of vehicleType in terms of fuelType

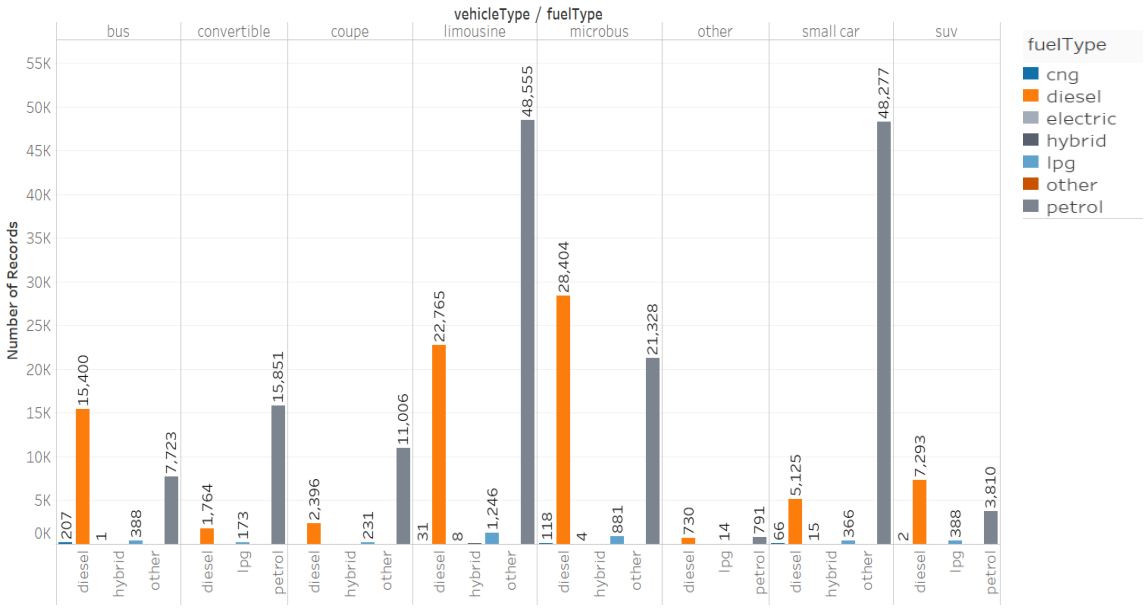

Fig. 24: Count of vehicleType in terms of fuelType

xiv. Count of vehicleType in terms of gearbox

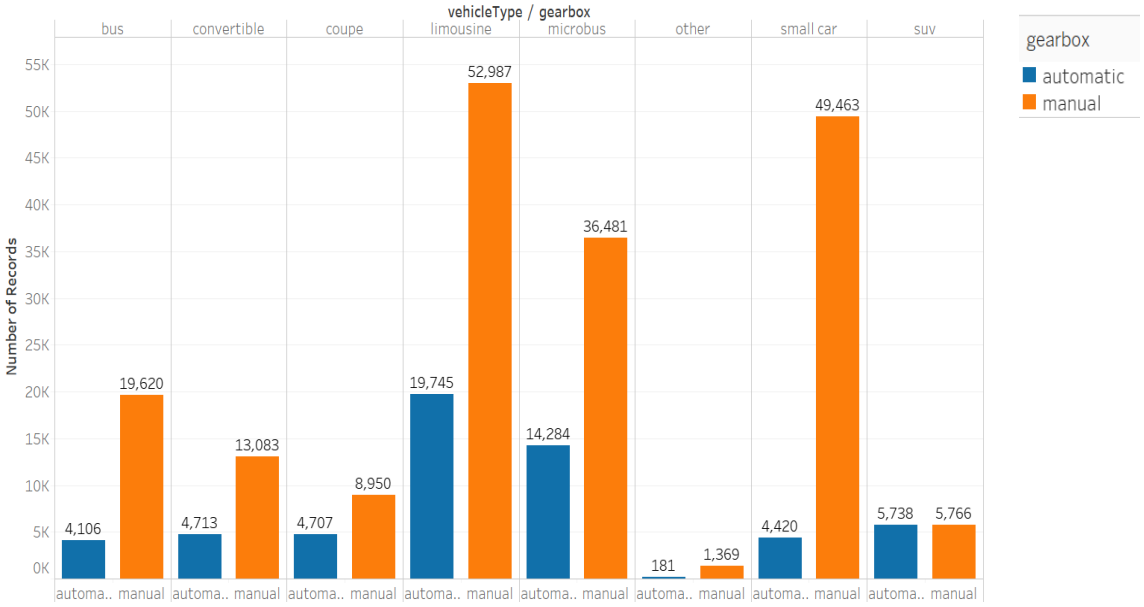

Fig. 25: Count of vehicleType in terms of gearbox

xv. Count of vehicleType in terms of notRepairedDamage

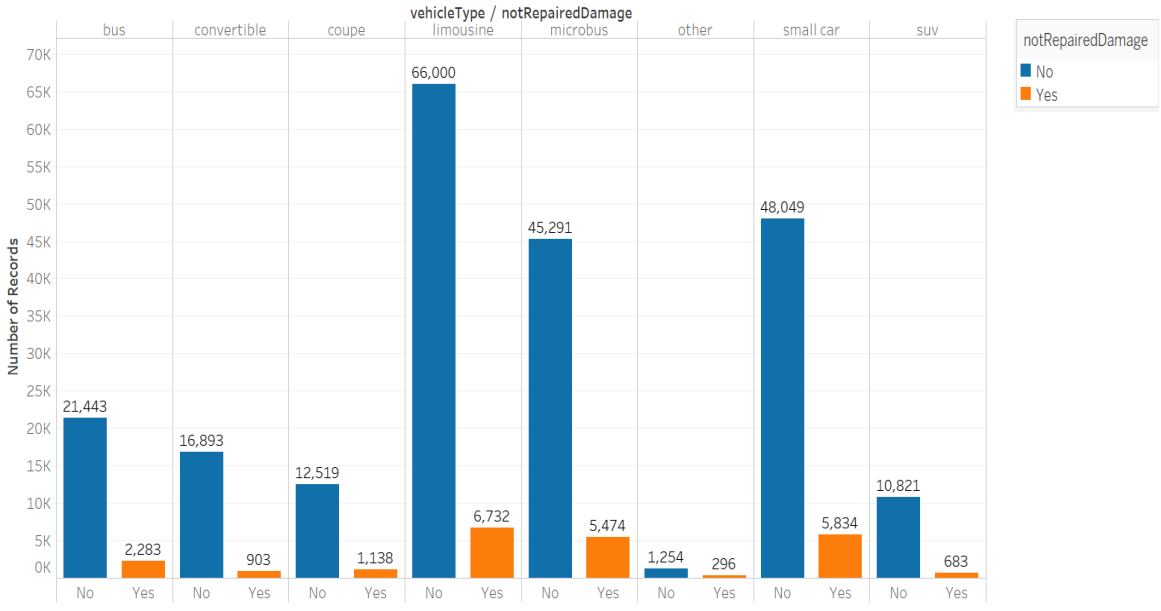

Fig. 26: Count of vehicleType in terms of notRepairedDamage

2.3.Multivariate Analysis

i. Price, vehicleType and gearbox

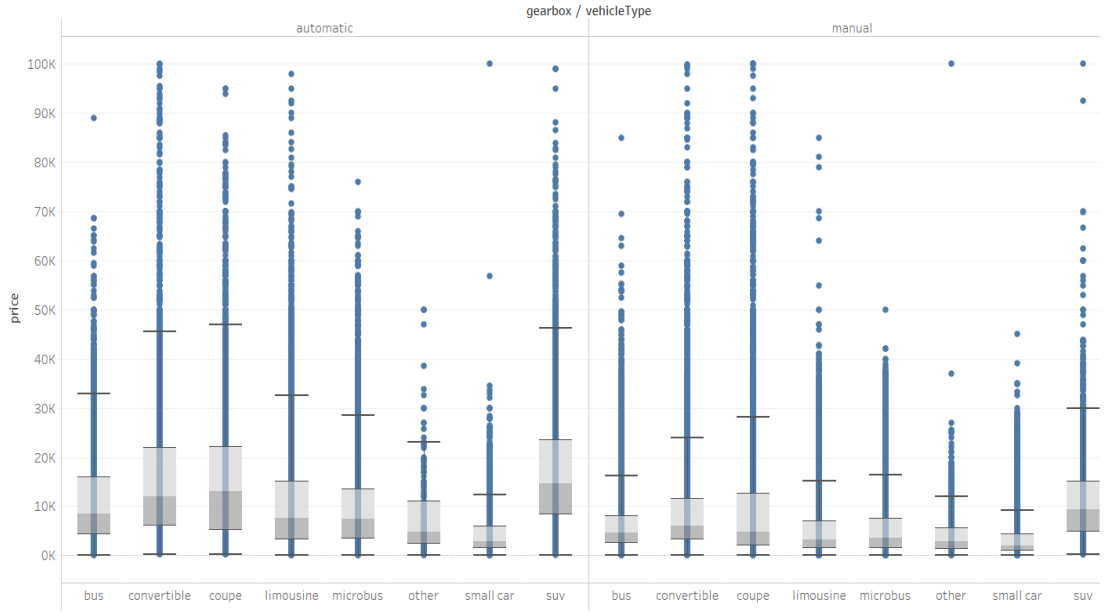

Fig. 27: Price, vehicleType and gearbox

Inference: As we have seen earlier in the plot between price and vehicle type; SUV, convertible and coupe are of higher prices as compared to other vehicle types; when compared in terms of gearbox we get the same results in both automatic and manual gearboxes. Thus in general, SUV, convertible and coupe are of higher prices irrespective of gearbox.

ii. Price, fuelType and gearbox

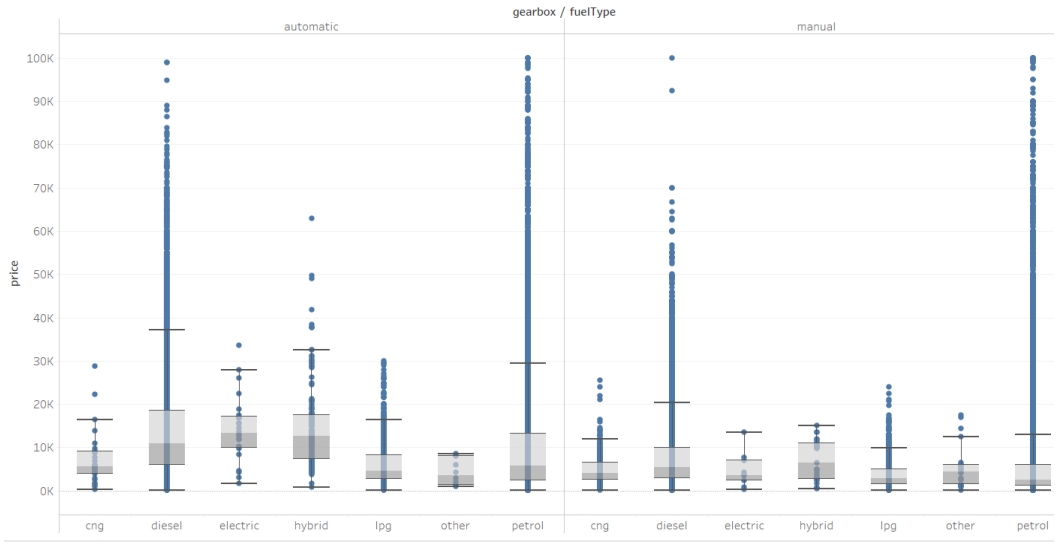

Fig. 28: Price, fuelType and gearbox

Inference: Box plot between price and fuel type stated that the electric and hybrid fuel type cars belong to higher price ranges. But when compared in terms of gearbox, diesel cars are as expensive as electric and hybrid for automatic gearbox; while for manual gearbox, electric cars are of lesser price as compared to diesel and hybrid cars. Hence, gearbox plays a major role for selecting cars in terms of fuel type.

iii. Correlation Matrix

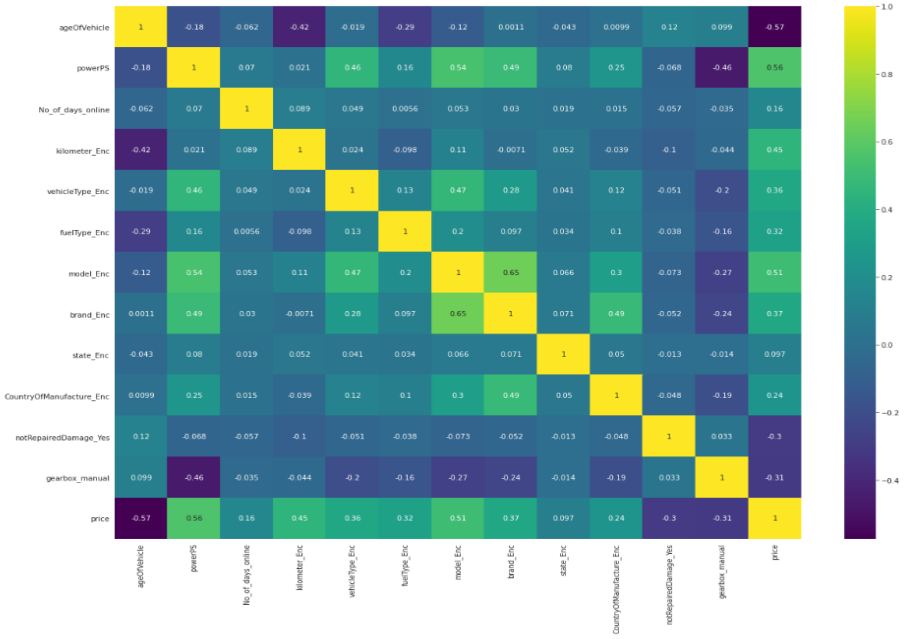

Fig. 29: Correlation Matrix

CHAPTER 3 – MODEL BUILDING AND EVALUATION

3.1.Base Model

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.730
Model:	OLS	**Adj. R-squared:**	0.730
Method:	Least Squares	**F-statistic:**	4.416e+04
Date:	Tue, 30 Jun 2020	**Prob (F-statistic):**	0.00
Time:	07:48:44	**Log-Likelihood:**	-1.7343e+05
No. Observations:	196400	**AIC:**	3.469e+05
Df Residuals:	196387	**BIC:**	3.470e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.1289	0.016	374.992	0.000	6.097	6.161
ageOfVehicle	-0.0541	0.000	-216.670	0.000	-0.055	-0.054
powerPS	0.0054	3.02e-05	179.431	0.000	0.005	0.005
No_of_days_online	0.0084	0.000	55.053	0.000	0.008	0.009
kilometer_Enc	8.436e-05	4.01e-07	210.457	0.000	8.36e-05	8.51e-05
vehicleType_Enc	3.884e-05	5.74e-07	67.652	0.000	3.77e-05	4e-05
fuelType_Enc	9.304e-05	7.89e-07	117.883	0.000	9.15e-05	9.46e-05
model_Enc	3.224e-05	4.18e-07	77.064	0.000	3.14e-05	3.31e-05
brand_Enc	1.203e-05	5.6e-07	21.472	0.000	1.09e-05	1.31e-05
state_Enc	2.429e-05	1.71e-06	14.177	0.000	2.09e-05	2.76e-05
CountryOfManufacture_Enc	5.374e-05	8.91e-07	60.283	0.000	5.2e-05	5.55e-05
notRepairedDamage_Yes	-0.6978	0.005	-153.160	0.000	-0.707	-0.689
gearbox_manual	-0.0307	0.004	-8.632	0.000	-0.038	-0.024

Omnibus:	24814.913	**Durbin-Watson:**	2.004
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	279572.261
Skew:	-0.145	**Prob(JB):**	0.00
Kurtosis:	8.838	**Cond. No.**	2.31e+05

Key Observations:

i. R-squared: 0.73

ii. No. Observations: n = 196400

iii. Df Residuals: n-p-1 = 196387

iv. Prob (F-statistic): the 12 features are helpful to predict the price as pvalue < alpha i.e., reject H0.

iv. Durbin-Watson: 2.004: no auto correlation.

v. Prob(JB): 0.00; pvalue > alpha; fail to reject H0: residuals are normally distributed.

3.2.Assumptions of Linear Regression

3.2.1. No Auto Correlation

For Durbin-Watson test we got a value of 2.004 which lies in the acceptance range of 1.5 - 2.5, hence, no autocorrelation. Moreover, the ACF plot also clearly confirms the same, that is no auto-correlation between the residuals, as clearly shown below.

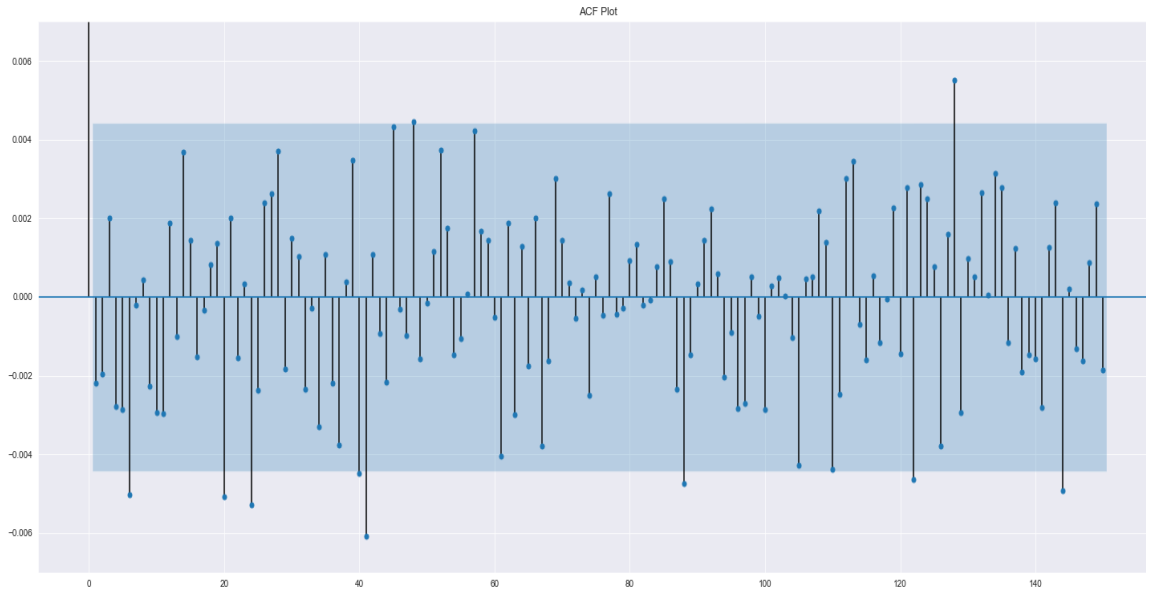

Fig. 30: ACF plot

3.2.2. Normality of residuals

The Jarque Bera test resulted in a p-value of 0.0 along with a test statistic value of 279629.60 which is greater than the t-critical value of 5.99. Moreover, our residuals deviated from normality towards the extreme which we can clearly see from the Q-Q - plot below. So we rejected the null hypothesis and concluded that residuals are not normal.

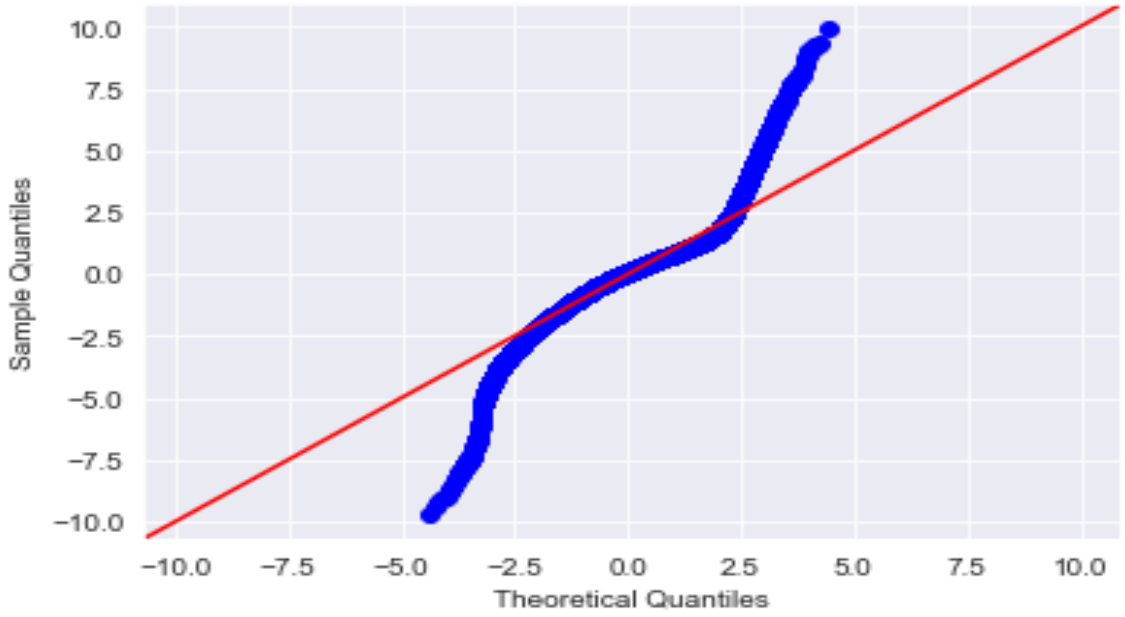

Fig. 31: Q–Q plot

3.2.3. Linearity of Residuals

We get a p-value of 0.399 for Linear Rainbow test which is higher than 0.05. Moreover, from the below scatter plot the residuals are symmetrically distributed in the former one and around horizontal line in the latter one. In both cases linearity is observed.

Fig. 32: Scatter Plot of Residuals

3.2.4. Homoscedasticity

If the variance of the residuals is symmetrically distributed across the regression line, then the data is said to be homoscedastic. The Goldfeld-Quandt test gives a p-value of 0.3132 which is higher than 0.05. Moreover, we can visually see that Homoscedasticity is present as shown below.

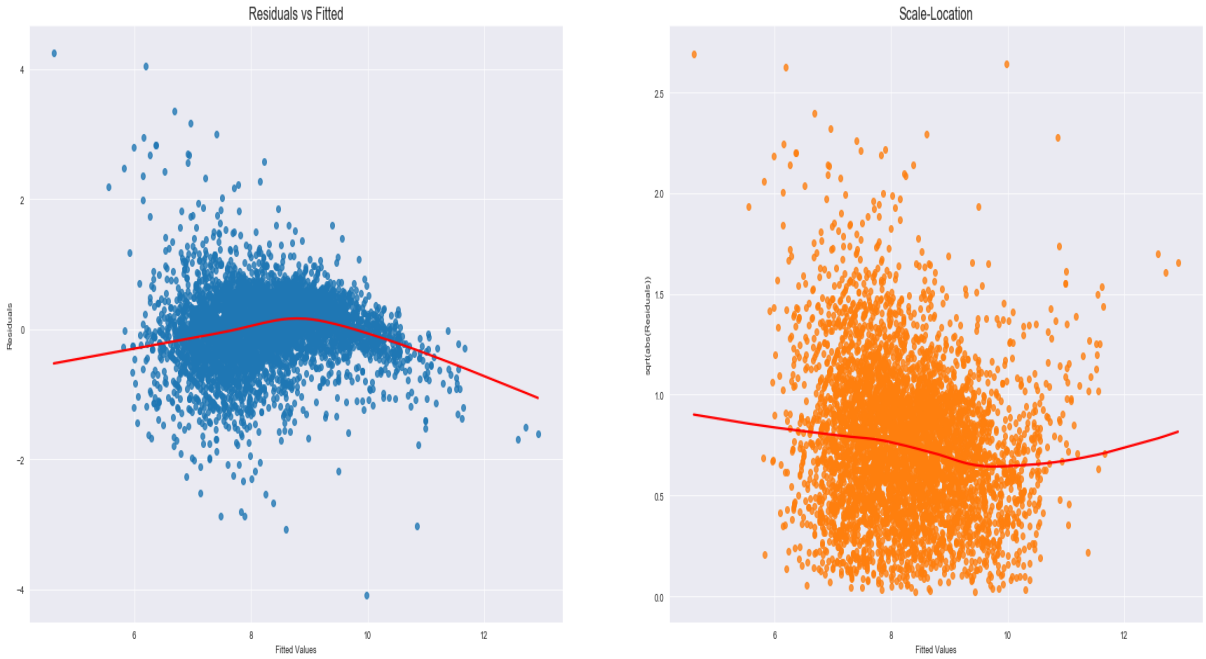

Fig. 33: Scatter plot of residuals v/s fitted

3.2.5. No Multicollinearity

The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high.

Following are the results for variance inflation factor test:

Feature	VIF
ageOfVehicle	1.482666
powerPS	1.975102
No_of_days_online	1.015263
vehicleType	1.421693
kilometer	1.356916
fuelType	1.236748
model	2.265722
brand	2.214775
state	1.012671
CountryOfManufacture	1.333841
notRepairedDamage	1.025145
gearbox	1.304639

Table 6: VIF values for the features

The VIF for all the variables are very low as we can see from the above values. So we can conclude that there is no multicollinearity present in our data.

3.3.Model Performance Measures Used for Evaluating Models

The various models built, must be evaluated based on certain model performance measures to identify the most robust models. The choice of the right model performance measures is highly critical hence, Root Mean Squared Error (RMSE) was also considered, in addition to R-squared.

3.3.1. Root Mean Squared Errors (RMSE)

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction.

3.3.2. R-Squared

R-squared is the coefficient of determination; it ranges from zero to one, with zero indicating that the proposed model does not improve prediction over the mean model, and one indicating perfect prediction. Improvement in the regression model results in proportional increases in R-squared.

Machine learning algorithms are classified as two distinct groups: parametric and non-parametric. Herein, parametricness is related to a pair of model complexity and the number of rows in the train set. We can classify algorithms as non-parametric when model becomes more complex if number of samples in the training set increases. Vice versa, a model would be parametric if model becomes stable when number of examples in the training set increases.

3.4.Parametric Models

Models that simplify the function to a known form are called parametric machine learning models. Low variance algorithms tend to be less complex, with a simple or rigid underlying structure. Examples include regression, Naive Bayes (NB), linear algorithms and parametric algorithms. Like, a regression can be regularized to further reduce complexity.

Under parametric modelling, Linear Regression, Ridge Regression, Lasso Regression and Elastic Net Regression were considered. Before model building log transformation was applied.

```
y_par_mod = pd.DataFrame(np.log1p(y['price']).values,columns=['log_price'])
```

```
lr = LinearRegression()
ridge = Ridge(alpha = 0.1,normalize = True) # Scaling is mandatory for all distance based calculations
lasso = Lasso(alpha = 0.0001,normalize = True)
elasticnet = ElasticNet(l1_ratio=0.01,alpha = 0.0001,normalize = True)
```

3.4.1. Coefficients for parametric models

Linear Regression Model

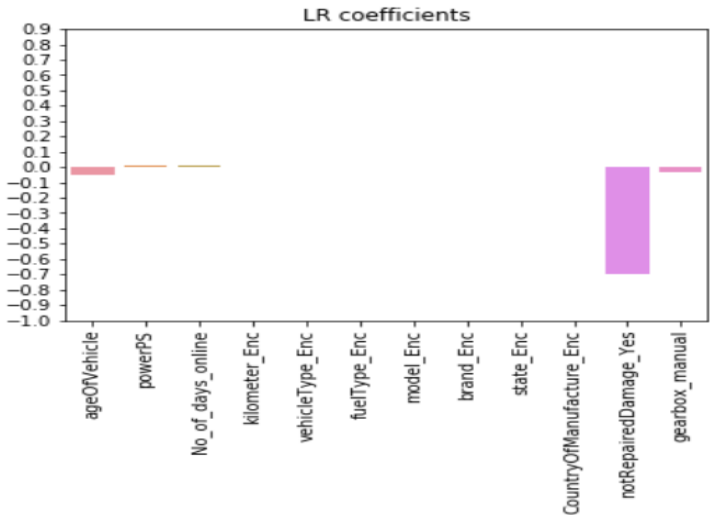

Fig. 34: Model coefficients for linear regression

From above plot, we can see the most significant features to predict price of used cars are ageOfVehicle, powerPS, No_of_days_online, notRepairedDamage_Yes and gearbox_manual.

Ridge Regression Model

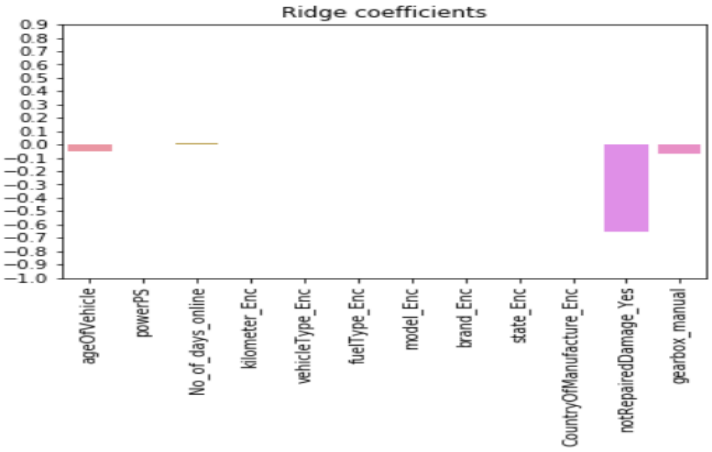

Fig. 35: Model coefficients for ridge regression

Ridge Regression is one of the regularization techniques that is used to reduce the overfit nature of the model. In this model, we can see that coefficient of powerPS has been reduced as it approaches zero indicating it's insignificance nature in predicting target variable.

Lasso Regression Model

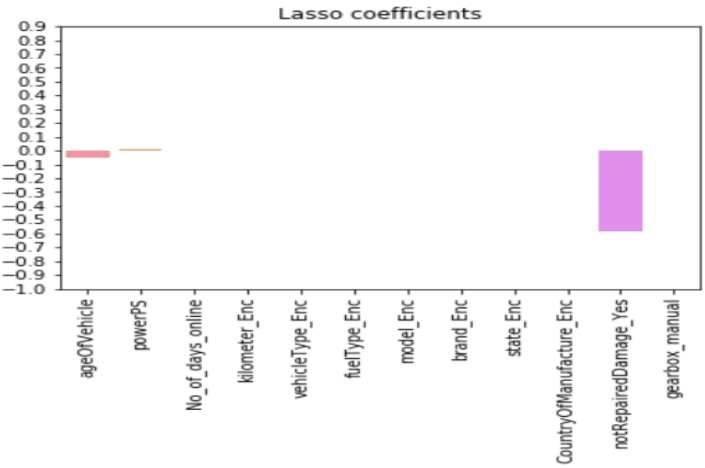

Fig. 36: Model coefficients for elastic regression

Lasso Regression is the Regularization technique with high power penalty that reduces the overfit nature of model in such a way that coefficients of insignificant features become zero. Here, we can see that coefficient of No_of_days_online and gerabox_manual is zero as compared to Linear Regression model.

ElasticNet Regression Model

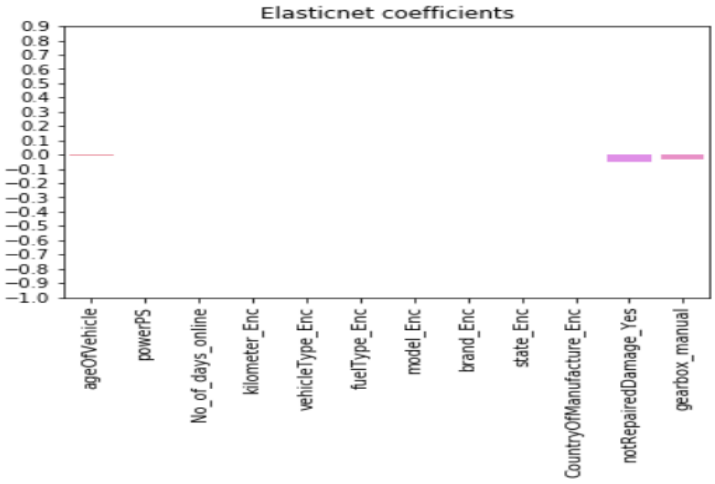

Fig. 37: Model coefficients for elastic net regression

ElasticNet is the combination of both Ridge and Lasso Regression, where it applies both the penalties in given ratio. From above plot we can see, ageOfVehicle, notRepairedDamage_Yes and gearbox_manual are the most significnat features in predicting price of used cars.

Used k-fold cross validation technique to get the RMSE values with number of splits equal to 10.

Bias and Variance errors for various parametric models.

```
RMSE scores : 0.586 (+/- 0.00652309) [Linear_Regression]

RMSE scores : 0.588 (+/- 0.00621691) [Ridge]

RMSE scores : 0.596 (+/- 0.00613996) [LASSO]

RMSE scores : 1.047 (+/- 0.00700431) [ElasticNet]
```

Fig. 38: Bias and Variance errors for various parametric models

Machine Learning Method	Bias Error	Variance Error
Linear Regression	0.586	0.0065
Ridge Regression	0.588	0.0062
Lasso Regression	0.596	0.0061
Elastic Net Regression	1.047	0.0070

Table 7: Bias and Variance errors for various parametric models

Linear Regression model gave the best results out of all the parametric models. Ridge model increased the bias error by about 0.34% but yielded an improvement in variance error by approximately 4.69%, whereas, Lasso model increased the bias error by about 1.67% but yielded an improvement in variance error by approximately 5.87%.

From the parametric models, linear regression model gave the best results. But, as per assumptions of linear regression, the residuals do not have a normal distribution. Hence, non-parametric models were considered.

3.5.Non-Parametric Models

Low bias algorithms tend to more complex, with a flexible underlying structure. Examples include decision tree (DT), nearest neighbors, non-linear algorithms, non-parametric algorithms. Decision trees can be pruned to reduce complexity. Algorithms that are too complex produce overfit models that memorize the noise instead of the signal.

Models that were considered under non-parametric category include:

1. DT Regressor
2. DT Bagging Regressor
3. DT Adaboost Regressor
4. RF Regressor
5. RF Adaboost Regressor
6. KNN Regressor
7. KNN Bagging Regressor
8. Gradient Boost Regressor

Standard Scalar function was used to scale the data with a mean of the transformed data equal to 0 and the variance equal to 1 before building the model.

GridSearchCV was used for hyper parameter tuning of DT, Random Forest (RF) and KNN Regressor. For DT Regressor, hyper parameter tuning was applied to get optimal maximum depth of tree and criterion to calculate errors. For RF Regressor, hyper parameter tuning was applied to get optimal number of trees and its maximum depth. For KNN Regressor, hyper parameter tuning applied to get optimal number of nearest data points and criterion of whether to consider distance for the calculations.

A bagging regressor is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a

final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it. Gradient Boosting (GB) for regression builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function. For Bagging and Boosting Regressor, k-fold cross validation technique (with 5 folds) was used to get the hyper parameters that could give maximum accuracy with a trade-off between bias and variance error.

Scoring method used to find bias and variance errors is 'Negative Mean Squared Error' and to find the R-squared value is 'r2'.

Bias and variance errors for various non-parametric models.

```
DTree  :  3166.656666292124  --  20.688297393962202
DTree Bagged  :  3048.6105212780576  --  22.834231612754063
DTree Boosted  :  2920.243202345049  --  84.49880036135615
KNN  :  2902.0643393016458  --  21.264488151889246
KNN Bagged  :  2968.214229098139  --  12.085551507327446
RF  :  2874.4468242214257  --  44.693950897618045
RF Boosted  :  2810.063420003883  --  40.17073280040146
Gradient Boost  :  2728.8733868780087  --  32.40123373222216
```

Fig. 39: Bias and variance errors for various non-parametric models

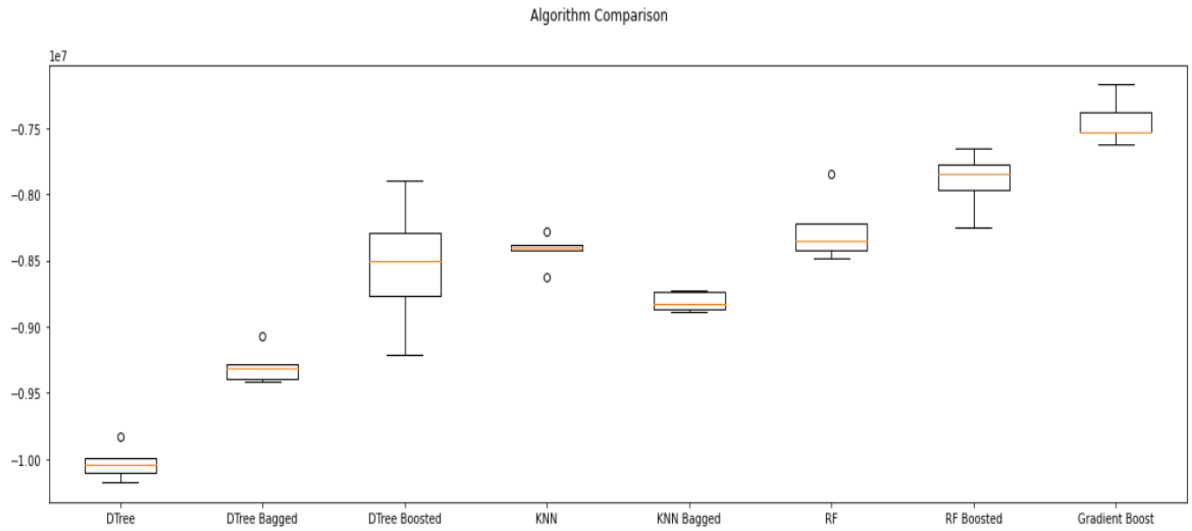

Fig. 40: Box-plot showing bias and variance errors for various non-parametric models

R-squared value (coefficient of determination) for various non-parametric models.

```
DTree  :  0.8449657928744093
DTree Bagged  :  0.8563115350229993
DTree Boosted  :  0.8681219220876164
KNN  :  0.8698168777240948
KNN Bagged  :  0.8638136745699165
RF  :  0.8721904277046238
RF Boosted  :  0.8778941513969005
Gradient Boost  :  0.8848697399665386
```

Fig. 41: R-squared value (coefficient of determination) for various non-parametric models

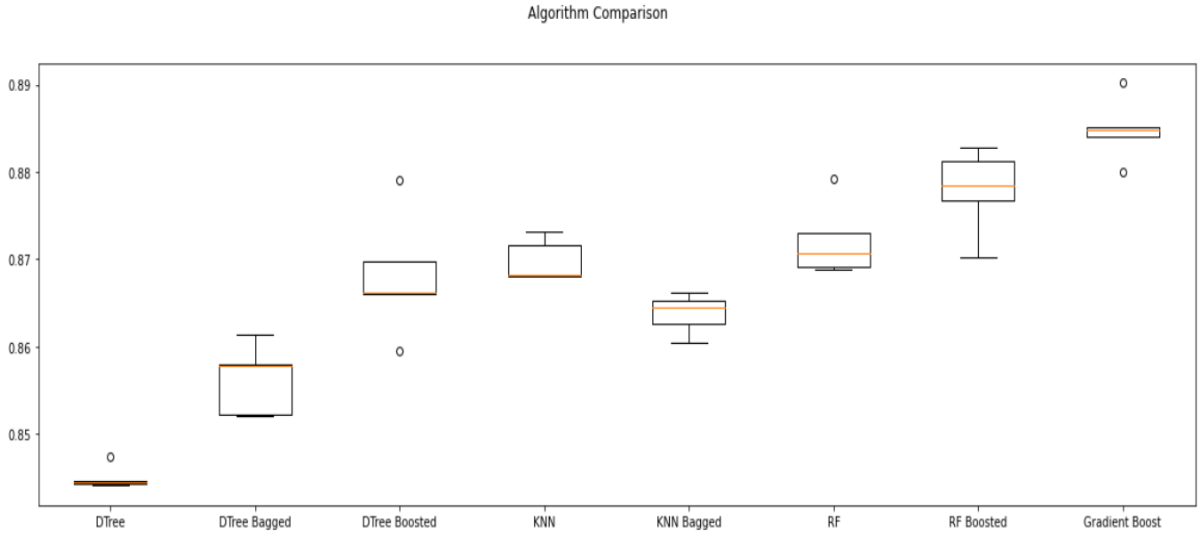

Fig. 42: Box-plot showing coefficient of determination for various non-parametric models

Bias and variance errors and R-squared (coefficient of determination) for various non-parametric models.

Machine Learning Method	Bias Error	Variance Error	Coefficient of determination
DT	3166.6566	20.6882	0.8449
DT Bagged	3048.6105	22.8342	0.8563
DT Boosted	2920.2432	84.4988	0.8681
KNN	2902.0643	21.2644	0.8698
KNN Bagged	2968.2142	12.0855	0.8638
RF	2874.4468	44.6939	0.8721
RF Boosted	2810.0634	40.1707	0.8778
Gradient Boost	2728.8733	32.4012	0.8848

Table 8: Bias - variance errors, coefficient of determination for various non-parametric models

Gradient Boost Regressor model seems to give the best results. But if KNN Bagging Regressor model is used, bias error tends to increase by about 8% but that would yield an improvement in variance error by approximately 62%. R squared value for Gradient Boost is 88.48% while for KNN Bagging Regressor is 86.38%, difference in R squared value is approximately 2%.

3.6.Artificial Neural Network

Artificial Neural Networks (ANN) are comprised of simple elements, called neurons, each of which can make simple mathematical decisions. Together, the neurons can analyse complex problems, emulate almost any function including very complex ones, and provide accurate answers. A shallow neural network has three layers of neurons: an input layer, a hidden layer, and an output layer. A Deep Neural Network (DNN) has more than one hidden layer, which increases the complexity of the model and can significantly improve prediction power. A neural network is a flexible model that adapts itself to the shape of the data. So it can pick best type of regression, and more hidden layer can be added to make the model more complex and increase its prediction capability.

A shallow neural network with a single hidden layer of 16 neurons could able to get coefficient of determination of around 0.84, which can be further increased by increasing the complexity of the model. After fine tuning the model with all the activation function RELU (Rectified Linear Units) gave the best result.

Next number of hidden layers and the number of neurons in each layer has to be decided. It was found out by using randomized search. 3 hidden layers with 256 neurons each gave us the best result with the batch size of 64. In order to prevent our model from overfitting, L2 regularization was added. Further, experiments were carried out with regularizations and by adding dropout layer. Since regularization was not added, doing normalization at each layer (Batch Normalization) doesn't improve the score. Early stopping was done in order to avoid overtraining.

After careful fine tuning, the best model was chosen with 3 hidden layers 16,256 and 256 neurons in each layer with a batch size of 64. Coefficient of determination for this model is around 0.88.

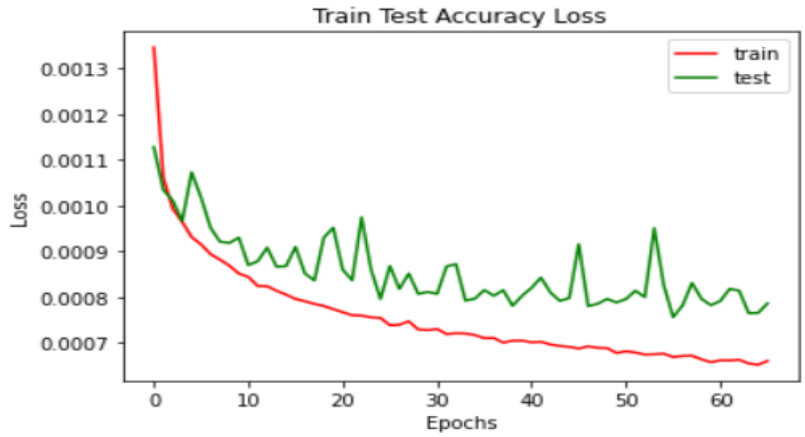

Fig. 43: Train test accuracy loss at each epoch for ANN

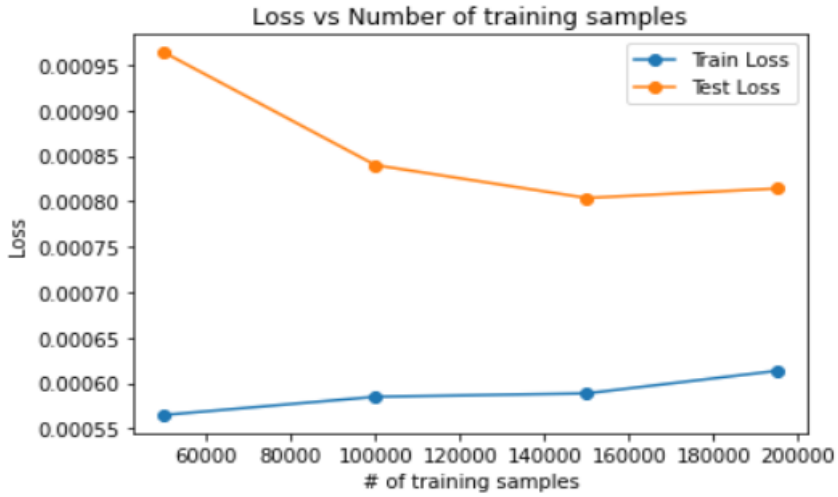

Fig. 44: Train test loss Vs number of training samples

From this learning curve it is evident that number of samples needs to be further increased for neural network to learn better.

Mean RMSE score for neural network model is around 2839.18 with the deviation of 45.8 euros.

```
RMSE scores : 2839.182 (+/- 45.82048079) [Neural Network]
```

CHAPTER 4 – CONCLUSION

Car price prediction can be a challenging task due to the high number of attributes that should be considered for the accurate prediction. The major step in the prediction process is collection and preprocessing of the data. In the following section, the experiment process, results and analysis will be summarized to conclude the aforementioned tasks.

We used data retrieved from Kaggle, scraped from used car listing on Ebay-Kleinanzeigen (German). As stated in the project review chapter, the two objectives carried out in this project were:

i. Perform data cleaning and visualization, in order to reach an elementary understanding of each car feature and its influence on the market price.

We used isolation forest anomaly detection technique for outlier detection and outliers for features 'price' and 'powerPS' were capped, as the outlier values seemed practically impossible. Missing values for 'notRepairedDamage', 'fuelType', 'model', 'gearbox' and 'state' were dropped as all these variables were quite independent of any other features in the dataset, and imputing these would result in an increase in multicollinearity between independent features. Secondly, as we have a huge dataset at our disposal, we could afford losing some data as it would not cost us much, as far as information loss is concerned. For the feature 'vehicleType', we did KNN imputation using model, brand and price as these feature come very close to predicting the type of vehicle. A major chunk of our feature selection was done based on the statistical analysis of the features. All of the features passed the significance test except for 'abtest' (p-value greater than alpha (0.05)). Apart from that, features like 'seller' and 'offerType' were highly imbalanced and had no significant influence on price prediction, hence, due to the extreme imbalance these were removed from the final dataset. Features 'dateCrawled', 'noOfpictures' practically served no purpose in predicting the price of a used car. So these features were also kept out from the final dataset. Four features namely, 'ageOfVehicle', 'No_of_days_online', 'State' and 'CountryOfManufacture' were extracted from existing features. Categorical features were divided into high and low cardinality features and depending on the best RMSE scores, we decided to stick with a combination of OHE and k-fold target encoding technique for our final dataset. Exploratory data analysis was performed to understand the data in terms of price across various independent variables/features and also to get insights on various features.

ii. Build and evaluate models using machine learning algorithms for price prediction in order to provide a real-time used car evaluation service.

This project examined the different machine learning techniques for predicting the price of the used car. KNN Bagging Regressor has successfully proved its capability in generating a good prediction model. With

the hyper parameter setting, it achieves a better accuracy than the standard solution, multiple linear regression.

Parametric models like Linear Regression, Ridge Regression, Lasso Regression and Elastic Net Regression were considered initially. Before model building log transformation was applied. Linear Regression model gave the best results out of all the parametric models. Ridge model increased the bias error by about 0.34% but yielded an improvement in variance error by approximately 4.69%, whereas, Lasso model increased the bias error by about 1.67% but yielded an improvement in variance error by approximately 5.87%. But, as per assumptions of linear regression, the residuals do not have a normal distribution. Hence, non-parametric models were considered.

Among non - parametric models, gradient boost regressor model gave the best results. But if KNN bagging regressor model is used, bias error tends to increase by about 8% but that would yield an improvement in variance error by approximately 62%. All the models seemed to provide comparable accuracies. R-squared value for Gradient Boost is 88.48% while for KNN Bagging Regressor is 86.38%, difference in R-squared value is approximately 2%.

Based on results, KNN Bagging Regressor proved its capability in generating a good prediction model with a better trade-off between bias and variance error. With hyper parameter tuning using grid search CV, it showed a better accuracy than the standard solution, multiple linear regression. Therefore, KNN Bagging Regressor was chosen as the final model.

- Price range in worst case scenario: [Actual value - Error (RMSE + Standard Deviation)] to [Actual value + Error (RMSE + Standard Deviation)]

- Price range in best case scenario: [Actual value - Error (RMSE - Standard Deviation)] to [Actual value + Error (RMSE - Standard Deviation)]

From the analysis, the most significant features to predict price of used cars are ageofVehicle, powerPS, notRepairedDamage_Yes, and gearbox_manual, although, the other independent features have a p-value less than significance level (0.05), they provide unique information regarding the attributes of a used car and how they influence used cars market price.

We created a user defined function where input would be all independent features and output would be predicted price of that particular car model. The '.pkl' file can be used to deploy our model using any platform (like Flask).

```python
def pred(age,powerPS, No_of_days_online, kilometer, vehicleType, fuelType, model, brand, state, mfg, damage,gearbox):
    map_dict = np.load('map_dict.npy',allow_pickle = 'TRUE').item()
#     gbr_model = joblib.load('gbr_model.pkl')
    knn_bagged_model = joblib.load('knn_bagged_model.pkl')
    yn = {'Yes' : 1, 'No' : 0}
    km = map_dict['kilometer'][kilometer]
    vt = map_dict['vehicleType'][vehicleType]
    ft = map_dict['fuelType'][fuelType]
    mod = map_dict['model'][model]
    br = map_dict['brand'][brand]
    st = map_dict['state'][state]
    com = map_dict['CountryOfManufacture'][mfg]
    d = yn[damage]
    gb = yn[gearbox]
    arr = np.array([age,powerPS, No_of_days_online, km, vt, ft, mod, br, st, com, d,gb])
    ss = joblib.load('ss.pkl')
    arr_t = ss.transform(arr.reshape(1,-1))
#     pred_val = gbr_model.predict(arr_t)
    pred_val = knn_bagged_model.predict(arr_t)
    return pred_val.tolist()

```

```python
pred(age = 19, powerPS = 75, No_of_days_online = 0, kilometer = 150000, vehicleType = 'small car',fuelType = 'petrol', \
     model = 'golf', brand = 'volkswagen', state = 'Bayern', mfg = 'Germany',damage = 'No', gearbox = 'Yes')
```

```
[1608.231753692143]
```

CHAPTER 5 – RECOMMENDATIONS AND ACTIONABLE INSIGHTS

This project aimed to analyze how features of a used-car influence its market price and to predict the price based on the car features in the given data. The final product of the project is machine learning model that can predict market value estimation of a used car given its features in order to provide a real time used car evaluation service. Such services are helpful to both organized and unorganized market.

5.1. Revenue Generation recommendations for the target business – online used car dealers:

- Lead management solution: Online used car dealers will be able to increase their customer base, as customers would have a handy approach to search about their desired used cars. This will be an effective source for lead generation.
- Advertisement Revenue: Ad revenue based on high user traffic and target marketing based on user profile.
- Subscription plans: Another way to earn revenue by charging a subscription /membership fee from the buyer or seller to access/avail/buy/sell the services or products available on portal.
- Trust Mark Certification: Dealers can start with Trust Mark certification and warranty program that endorses the quality of used cars. The certification is aimed at providing a used car buyer with the assurance that the used car they are buying is free from any major defects that the papers are genuine and the car has not been involved in an accident.
- Commission on Car Loan and Insurance: Dealers can now tie up with car loan and insurance providers where they work on a commission model.

5.2. Actionable insights

- The main limitation of this study was the low number of independent predictors that have been used. As future work, we intend to collect more data and to use them again on artificial neural networks, and other advanced techniques like fuzzy logic and genetic algorithms to predict car prices.
- Capping on values of price and powerPS features during our analysis can be overcome by accurate data collection.

CHAPTER 6 – REFERENCES

1. Amjadh Ifthikar and Kaneeka Vidanage (2018) "Valuation of Used Vehicles: A Computational Intelligence Approach".

2. Dr. Shiva Shankar. K.C (2016) "A Study on Consumer Behavior Towards PreOwned Cars in India", Indian Journal of Research, 5(11).

3. Edmunds Forecasts (2020) "New Vehicle Sales Drop in March to Close a Down First Quarter in 2020". Available from: https://www.edmunds.com/industry/press/new-vehicle-sales-drop-in-march-to-close-a-down-first-quarter-in-2020-edmunds-forecasts.html [Accessed 09 June, 2020].

4. Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric (2019) "Car Price Prediction using Machine Learning Techniques", TEM Journal, 8 (1), 113-118.

5. Evgeniya Koptyug (2019) "Leading import countries for motor vehicles from Germany based on export value 2018". Available from: https://www.statista.com/statistics/587701/leading-import-countries-german-motor-vehicles-by-export-value/ [Accessed 09 June, 2020].

6. Evgeniya Koptyug (2020) "Revenue of the market for second-hand cars in Germany from 2000 to 2019 (in billion euros)". Available from: https://www.statista.com/statistics/589610/revenue-used-cars-germany/ [Accessed 19 June, 2020].

7. Graham Rapier (2020) "Used car may get even cheaper than in the last recession as the coronavirus forces dealerships to offer unprecedented deals". Available from: https://www.businessinsider.in/business/news/used-cars-may-get-even-cheaper-than-in-the-last-recession-as-the-coronavirus-forces-dealerships-to-offer-unprecedented-deals/articleshow/75072952.cms [Accessed on 10 June, 2020].

8. Greg Gardner (2018) "Auto Sales Are Down. Here's Why They'll Continue To Fall (Forbes)". Available from: https://www.forbes.com/sites/greggardner/2018/03/12/auto-sales-are-down-heres-why-theyll-continue-to-fall/#1612ab622dcb [Accessed 09 June, 2020].

9. Harikrushna Vanpariya (2018) "Using Different Machine Learning Techniques for Predicting the Price of Used Cars", International Journal for Scientific Research & Development, 6(10).

10. Ken Research (2020) "Germany Used Car Market is expected to reach about EUR 105 Billion in Gross Transaction Value (GTV) by 2023: Ken Research". Available from: https://www.kenresearch.com/blog/2020/02/germany-used-car-market-value/ [Accessed 25 June, 2020].

11. Mariana Listiani (2009) "Support Vector Regression Analysis for Price Prediction in a Car Leasing Application", Master Thesis.

12. Monika Singh (2019) "Germany Used Car Market Outlook to 2023 - Surge in Multi-Brand Dealerships Coupled with Improved Quality and Inspection of Used Cars to boost Used Cars Market". Available from: https://www.kenresearch.com/automotive-transportation-and-warehousing/automotive-and-

automotive-components/germany-used-car-market-outlook/277739-100.html [Accessed 18 June, 2020].

13. Mordor Intelligence "Used car market– growth, trends, and forecast (2020 - 2025)". Available from: https://www.mordorintelligence.com/industry-reports/global-used-car-market-growth-trends-and-forecast-2019-2024 [Accessed 10 June, 2020].
14. Sameerchand Pudaruth (2014) "Predicting the Price of Used Cars using Machine Learning Techniques", International Journal of Information & Computation Technology, 4(7), 753-764.
15. Statista Research Department (2020) "Worldwide car sales 2010 – 2020". Available from: https://www.statista.com/statistics/200002/international-car-sales-since-1990/ [Accessed 09 June, 2020].
16. Stephen Edelstein (2020) "COVID-19 may already be causing new car sales to fall". Available from: https://www.digitaltrends.com/cars/coronavirus-set-to-torpedo-new-car-sales-for-march-2020-and-beyond-report-says/ [Accessed 09 June, 2020].