Faculty of Science and Technology

Computer Engineering & Informatics

DISSERTATION REPORT

Real Time Prediction of Used Car Value

CST4090

Supervisor:

Dr.Krishnadas Nanath

Submitted By:

ANSU VELLARA ANISH

MSc Data Science

M00734048

02nd October 2020

ABSTRACT

Abstract Research is about the prediction of the car value for sale in UAE. The Research domain is the automobile industry, where the user can get the value of the car predicted depending on the features of the car given. The data for the prediction will be web scraped lively, depending on the input given. The input used for scraping includes the car details such as Brand name and Model name. A similar Model will be scraped, which is then used as the data for training the model. Research is trying to predict using machine learning algorithms such as Linear Regression, Decision Tree, XGBoost, and Random Forest. The Expected end product of this Research is a product that gives the car's value when the user inputs the data of the car they want to sell. The outcome expected for the Research is to the prediction of the car value, as mentioned. This Research would be beneficial for the Online Dealers, Any Online Pricing services, or individual who wish to sell the car which could help them to get a fair price so that neither their sale will not get delayed due to a high price that they ask nor get any loss due to loss price that they demand due to lack of knowledge of the market price.

TABLE OF CONTENTS

ABSTRACT 1
List of Figures: 3
ACKNOWLEDGEMENT 6
1: INTRODUCTION 7
1.1 BACKGROUND: 8
1.2.1 Terminology Used: 8
1.2 PROBLEM STATEMENT 10
1.3 AIMS & OBJECTIVES: 10
1.4 DISSERTATION ORGANISATION: 11
2: LITERATURE REVIEW 12
2.1 INTRODUCTION: 12
2.2 RESEARCH TIMELINE: 12
2.3 METHODOLOGY 13
2.4 REVIEW 16
2.4.1 Level 1: 16
2.4.2 Level 2 : 17
2.4.3 Level 3 : 19
2.4.4 Level 4 : 20
2.5 OUTPUT OF THE REVIEW 22
3: METHODOLOGY 24
3.1 DATA EXTRACTION : 24
3.1.1 DATA PLATFORM 25
3.2 DATA : 25
3.3 DATA PREPARATION 26
3.3.1 Data Cleaning: 26
3.3.2 Data Pre-processing: 27
3.4 DATA ANALYTICS PIPELINE 27
3.4.1 VISUALIZATION 28
3.4.2 MODELLING 34
3.4.3 Hyper Parameter Tuning of Model: 38
3.5 PROTOTYPE OF THE END PRODUCT: 38
4: RESULTS & FINDINGS 40
4.1 PRE-PROCESSING: 40
4.2 MODELING: 41

4.2.1 Linear Regression 41
4.2.2 Gradient Boosting: 45
4.2.3 RANDOM FOREST 48
4.3 HYPER PARAMETER TUNING: 49
4.3.1 Random Forest: 49
4.3.2 Gradient Boosting Model: 50
4.4 RESULT OBTAINED FROM RAPIDMINER: 52
4.5 FINAL PRODUCT 53
5: DISCUSSIONS 56
5.1 DATA 56
5.2 MODELING 57
5.2.1 Linear Regression 57
5.2.2 Gradient Boosting Model: 58
5.2.3 Random Forest: 59
5.3 SUMMARISING COMPARISON OF MODELS: 59
5.4 HYPER PARAMETER TUNING 60
5.4 OPTIMIZATION: 60
5.5 RESEARCH PRODUCT: 60
6: CONCLUSION 61
6.1 SUMMARY OF THE THESIS MODULES: 61
6.1.1 Data Extraction: 61
6.1.2 Predictive Model Selection: 61
6.1.3 Optimization: 62
6.2 MANAGERIAL IMPLICATION 62
6.3 THEORETICAL IMPLICATIONS 63
6.4 LIMITATIONS: 63
6.5 FUTURE WORK: 65
6.5.1 Feature Level Work: 65
6.5.2 Model Level Work: 65
7 : BIBLIOGRAPHY 66
8.APPENDICES 70

List of Figures:

Figure 1:Used Car Vs. New Car Sales 7
Figure 2:Used Car value index UAE 7

Figure 3:Dissertation Organization 11
Figure 4-Chapter Format 12
Figure 5-Number of Papers reviewed 13
Figure 6-Predictive analytics status (a) 2006 & (b) 2020 13
Figure 7-Final Set of Publication 14
Figure 8-Classification of the Papers 15
Figure 9-Hierarchical Paradigm for Review 15
Figure 10-Classification of the papers based on levels 15
Figure 11-Distribution of data 16
Figure 12-Business Value of Predictive Analytics 17
Figure 13-Output of review 22
Figure 14:PIPELINE 28
Figure 15: Data Structures 28
Figure 16:Correlation Plot 29
Figure 17:Missingness 29
Figure 18:Percentil 95 30
Figure 19:Sum(Price) vs. Variables 32
Figure 20:Median(Price) vs. Variables 33
Figure 21:SUM(Price) Vs. SUM(Mileage) 33
Figure 22:Median(Price)vs Median(Mileage) 34
Figure 23:Correlation value 34
Figure 24:Regression Curve of Price vs. Year 35
Figure 25:Regression Plot Year and Mileage Vs. Price 35
Figure 26:Regression Plot Year and Mileage Vs. log10(Price) 35
Figure 27:Inc MSE and Inc Node Purity Plot 36
Figure 28:Gradient Boost Model 37
Figure 29:Removing Outlier 40
Figure 30:Null Check 40
Figure 31:Price value Distribution graph 41
Figure 32:Structure of Linear Model 42
Figure 33:Regression Prediction Plot 42
Figure 34:LR with Normal Split and Original Price Value 43
Figure 35:Structure of LR with log10(Price) 43
Figure 36:LR With normal split and log10(Price) 44
Figure 37:Random Splitting of data 44
Figure 38:LR with Random split and Price 44
Figure 39:LR with Random Split and log10(Prices) 45
Figure 40:LR with CV 45
Figure 41:Relative Influence Graph 45
Figure 42:GBM Normal Split 46
Figure 43:GBM Normal split with relative influenced variable 46
Figure 44:GBM with log10(Price) and relative influenced values 46
Figure 45:GBM with Random Split 47
Figure 46:Plotting Optimum Tree graph 47
Figure 47:GBM with optimum tree value result 47
Figure 48:Random Forest Normal Split 48
Figure 49:Random Forest with log10(Price) 48
Figure 50:Random Forest with Random Split 48

Figure 51:Comparison of Models 49
Figure 52:Random Forest with Normal split Hyperparameter Tuning 49
Figure 53:Random Forest with Log split Hyperparameter Tuning 50
Figure 54:Random Forest with Normal split and Log split Hyperparameter Tuning Graph 50
Figure 55:Hyper Parameter Tuning of GBM 51
Figure 56:Application of Hyper Parameter Tuned Result 51
Figure 57: Variable Statistics 52
Figure 58:Ensemble tree in GBM 52
Figure 59:(a)SVM (b)Decision Tree 52
Figure 60:Deep Learning and Linear Regression 52
Figure 61:GBM and Random Forest 53
Figure 62:Comparison of Result based on RapidMiner 53
Figure 63:UI of the product 54
Figure 64:The UI send data request in backend and scraping 55
Figure 65:Output from the prediction 55
Figure 66:Final Output 55

ACKNOWLEDGEMENT

In the successful accomplishment and completion of this project, many thanks to the almighty God. I would like to extend my gratitude to many people who have best owned upon me their blessings and heart pledged support and to everyone who has directly or indirectly been part of this journey of completion of the project.

Firstly, I would like to extend my sincere thanks of gratitude to my Professor and Mentor, Dr. Krishnadas Nanath who has guided and supported me throughout the course to complete this project on this amazing topic, “Real-Time Prediction of Used Car Value”. This topic's choice resulted from one of the activities initiated by Dr. Krishnadas, where we were asked to choose a topic and find the relevant datasets from the open sources and related hashtags. I was keen to explore this topic further and decided to choose this to complete my thesis.

I would also like to extend my gratitude to my parents, peers and friends who have helped me and encouraged me throughout the year of my study at Middlesex University. Special mention to my parents who supported me and encouraged me to believe in myself. I am grateful to the faculties and professors who are always ready to help me and share their knowledge and experience. I have learned so much from my professors and classmates based on their experience and knowledge. I am lucky to have the best mentor and advisor which resulted in successfully completing the project, without whom I could not have done this any better.

1: INTRODUCTION

Data Analytics deals with analyzing the data and making any conclusion that can be derived that can bring business values. Machine Learning is the automated process to find any relationship with the variables to derive the outcome. The machine learning in the analysis of the problems. Research is intended to predict and outcome for Ecommerce customers to the customer Industry and the automotive industry. Proposed research deals with the prediction of used car values.

As per the study conducted by the FOX business, It was shown that people having a credit rating with prime and super-prime has a priority of buying a used car rather than a new model as they come in less value and the easiness to change frequently. The sale of the used car compared to the new has been increasing in the market since times.

Graphical figure 1 depicts the sales of used cars vs. the new car in different countries(Dark lines as Used car and dotted lines New car)-it is clear that the sales of the used car are much higher than the new car. The importance of the proposed research lies in this zone. This research helps in predicting the value of the used car that the user is selling. It would help the customers value their car, resulting in a sale that does not incur a loss and avoids the delay due to the high price of the car without the market knowledge.

The research scope is high that it incurs a high demand in many sectors as the car price valuation is very important in many sectors such as banks for approval of loans, Motor insurers for the premium calculation.

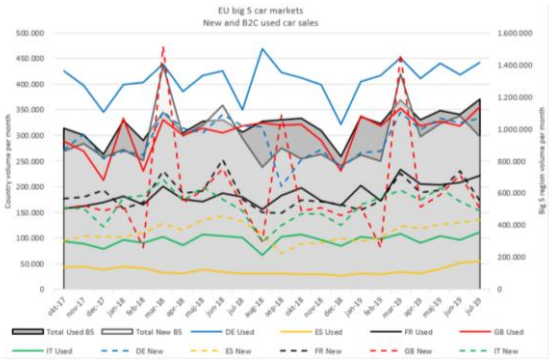


Figure 1:Used Car Vs. New Car Sales

The research is based on the realtime data from the dubizzle used for the prediction depending on the users input.The real time data for the prediction make the research more live as the importance of such project is very high

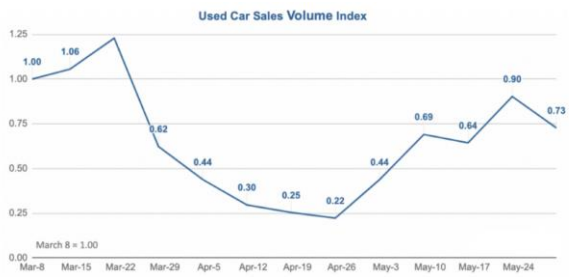


Figure 2:Used Car value index UAE

The used car value index of UAE shows the sale of the used car hence proving the importance of this research.

1.1 BACKGROUND:

The proposed research aims at the prediction of the used car values for sales. The research is intended to help the customers get around the value of the car they are going to sell. The sales market price has gone down as the number of sales has increased massively due to the loss of a job or any other financial situation that has currently affected the market. A massive sale that has incurred a considerable loss due to lack of knowledge of the market price resulted in the delay of the sale to high price demand, which in turn resulted in a sale of minimal price for the fast sales due to urgency. The intended research looks at the current situation where the research helps the user predict the car's value, depending on the current market values. It would help the user in making the sale without any delay or incurring any loss.

The automobile market behavior is of considerable interest due to many reasons. The current situations of COVID have a major impact on public transports. The public transport timings are not flexible nowadays due to the lockdown, and they are accommodating only a limited number of people ensuring the safety of the travelers and commuters. It will lead to people buying more of used cars for commuting purposes as this will be an affordable option for people going for work to ensure they have flexibility in the time they commute.

The used cars are mostly bought online via commercial websites based on the information provided by the owner or car dealer, which include at least a name, type, year of make, condition of the car, the value of the car which owner is offering the car for and the image of the car itself. This information is not just enough for a buyer to determine if this car is priced at the right value. The buyer needs to ensure if the car is worth the value, the car needs to be checked with a detailed inspection. For the same reason, the owner or the car dealer should know the current market value of the car they are selling. It requires proper research and awareness of the current market value trend and forecasts the price accordingly, which is the main purpose of this project, and this will be a helping aid for the sellers/ resellers. The research is highly applicable to the e-commerce industry in customer to customer deals and the automotive industry. It is very helpful for any e-commerce automobile sales to predict the car's value for sales.

1.2.1 Terminology Used:

- Machine Learning:
  It is defined as programming a system to optimize the performance using and past experience data by building any model.

- Car sale Value:
  The Car sale value is defined as the price at which the car can be sold.

- Customer 2 Customer Sales:
  It is defined as the sale that happens between two individuals.

- Predictive Analysis:

"Predictive analytics is a set of business intelligence (BI) technologies that uncover relationships and patterns within large volumes of data that can be used to predict behavior and events."

- Regression Analysis:
  It is the method of finding the variable that impacts the value to be predicted and uses those variables to predict the outcome.

- Ensemble Model
  "Ensemble modeling is the process of running two or more related but different analytical models and then synthesizing the results into a single score or spread in order to improve the accuracy of predictive analytics and data mining applications."

- Boosting
  Boosting is an algorithm that makes the high-quality prediction by training sequence of multiple weak models, which compensates its previous model's weakness.

- Tree Model
  "Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each tree in the random forest spits out a class prediction, and the class with the most votes becomes our model's prediction."

- Real-Time Prediction:
  "Real-time predictive analytics is the process of extracting useful information from data sets in real-time."

The research works for sales in the UAE. The market statistics show that the used car sales in UAE have continuously grown in recent years. The competition in the sector has widely increased that there are multiple players. UAE sales have been divided into seven emirates(Dubai, Sharjah, Abu Dhabi, Umm Al Quwain, Ras Al Khaimah, Fujairah, and Ajman), which Dubai marks the highest sales. The sales have been divided into three

1.Dealer sale
2.Online Portal sale
3.Customer to Customer sales

The Dealer sales mark the highest sales compared to the online portal sale and customer to customer sales. There are around 400 dealerships and 40 classified portals for used car sales. The customer to customer sales is very less compared to both sales as the individual expects the price variance and expects any fraudulent activities leading to a loss for them. When doing business with the customer, the dealers buy the car at a low price and refurbish it and sell it at a higher price. It affects the customers who do not know about the market value get easily cheated by them who demand that the car expects to get that value only.

The market is expected to raise its sales in the next five years; hence the demand for price prediction increases, which is the intended outcome of this research. Multiple research has

shown that the Used car sales have been widely increased over the new car. The study says that nine out of the ten customers are interested in the used car, but the customer to customer sales share, which is very less.

1.2 PROBLEM STATEMENT

This research's background studies prove that the sale of the used car is higher in UAE, but customers' sale to customer share is very less. The dealer's sale share is very much high compared to others due to the offers they apply to the customers, Which means the dealers buy the car from the user at a lower price and resell it at a higher price showing the discount. The sales of the user, in turn, result in loss due to the poor knowledge of the market value. The dealers also charge an amount of commission from the user; hence is it essential to promote the customer to customer sale which is very less due to the user fear of fraudulent activities of loss they might incur from an unknown person which might be a higher value than compared to the registered dealers. There can be a chance that user sales get delayed due to the higher demand of the sales value.

The intended research helps predict the value of the car that the user wants to sell, which would promote the customer to customer sale, which in turn helps the user to sell the car value by demanding the market price without causing any delay in the sales. This research would have a high demand in this time as the car sales have increased widely through the internet sources, which could help users fall prey for fraudulent acts of the dealers. This research would speed up the process of sales, helping the seller and buyer get the value predicted and hence to decide the final sale price so that it would end up in a loss for both.

The research opportunity lies in the time of higher sales time where the market of the sales is very less, and the urgency of the sale to happen is very high.

1.3 AIMS & OBJECTIVES:

The aims and objectives of this research would help in building an effective solution for the problem statement.

AIM

> **Identify and predict the car price with the variables such as mileage, year of production, color, cylinder number, and other given specification.**

The research aims to predict the value of the used car depending on the given specification of the car that they want to sell.

OBJECTIVE:

The objective of this research would help in fulfilling the aim that is mentioned.

- **O1: Interface for the user to input the description of the car to be sold is to be Generated**

  O1 deals with the generation of the interface for the user to enter the specification of the car that they want to sell. It acts like a webpage for the user.

- **O2: The Data set will be lively scraped from the internet using the input from the user**

The second objective is the real-time value collection of the data. This is done by scraping the data from dubbizle.com by taking the input from the user. The scraped data will have the data of all the cars of the same model in sale for the current time.

- **O3: The Data analysis of the scraped data using multiple algorithms for the prediction**

  This objective is the analysis of the data that is scraped for the analysis. This uses multiple algorithms for the predive analysis for the prediction of the final value.

- **O4: The selection of the model that gave the accurate output for building the product**

  The research uses multiple algorithms for the prediction to find the algorithm that produces the best accuracy with less error rate—the real deals with the selection of the final model.

- **O5: Optimization of the predicted value**

  The predicted value may not be optimized. It is to be optimized to give the final result to the user depending on the sale time demand of the customer.

1.4 DISSERTATION ORGANISATION:

The proposed research lies on the importance of the prediction of used car value. The dissertation consist of study and development of a product based on the multiple analysis of the model conducted. The structure of the dissertation are as follows

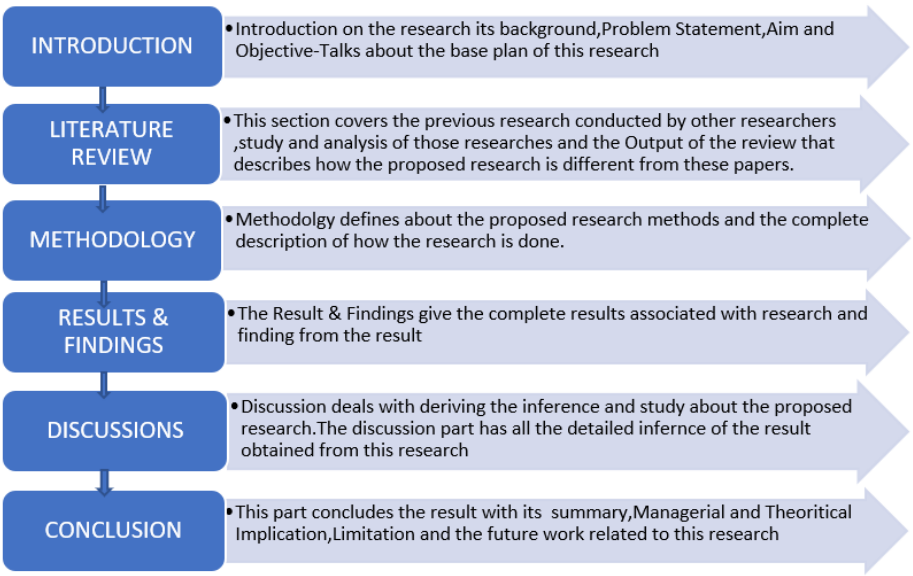


Figure 3:Dissertation Organization

2: LITERATURE REVIEW

2.1 INTRODUCTION:

Machine learning and predictive analytics have been in playing the market for a decade. It has been used widely for the prediction of many business values. There has been tremendous growth in the insights derived through predictive analytics—the number of research and study conducted in this area. The study related to this field is so massive that it proves the power of the prediction, which would strengthen any research conducted in the field.

The structure of the chapter follows the below structure.

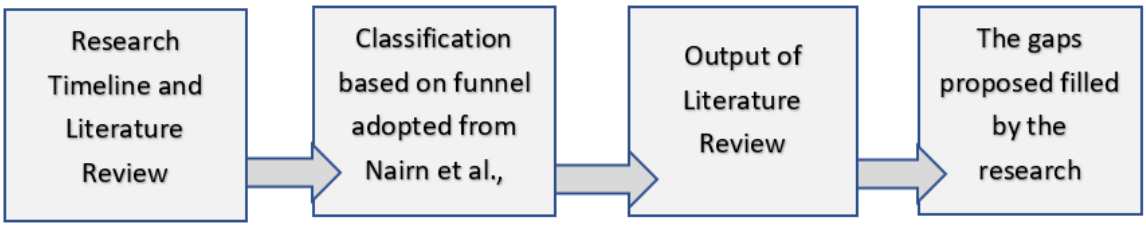


Figure 4-Chapter Format

The chapter begins with the research time of the literature review, which defines how machine learning has shaped up over the years. It specifies how machine learning has been changing over the years, and similar researches related to this proposed research has been conducted in these years. The next section covers the methodology of the research. It covers how this review has been structured and how different researches have been collected to study the same. It is followed by the classification of our research based on the architecture of Nairn et al. It defines a funnel structure with four levels, which helps in differentiating the heterogenous papers used in this review.

The output of the literature review gives us the idea of the algorithms used in different researches. This would help the intended research to understand the working of different algorithms in similar researches and select multiple algorithms for the proposed research and choosing the best algorithm with the best accuracy and less error rate. This is followed by the mentioning of the gaps in the reviewed researches and studies with the intended research. This section describes the change that is being made in the proposed research from other researches that would help in solving the problem statement much efficiently.

2.2 RESEARCH TIMELINE:

Machine learning came into existence from 1952, where the computer was able to learn using neural networks and perceptron. The kind of research at the time was pattern and shape recognition. The times till now, it has marked importance and the need for machine learning. With certain practices, machine learning did not have any progress since 1970 due to the argument of Neumann architecture being more accessible than a neural network.1982 again marked rise in interest for this neural network. It then got developed with backpropagation, which is a critical content in a neural network. Then again, the study went into a black box.

The graph marks a timely research journal article that was published related to the proposed research. From the graph, it is observed that there is a sudden surge in the number of studies

conducted between 2010 and 2020. It is because the 21st century marked a significant difference in machine learning. There have been several algorithms developed in machine learning for prediction and analysis. 2009 marked the beginning of the new deep learning. It leads machine learning to be used widely in services and applications; hence many studies and projects were conducted after 2010, which lead to a peak in the number of publications in Figure 2. The 21st century led to many achievements in machine learning, such as ImageNet, Googles AphaGo. The preceding time of machine learning had its discovery and application.

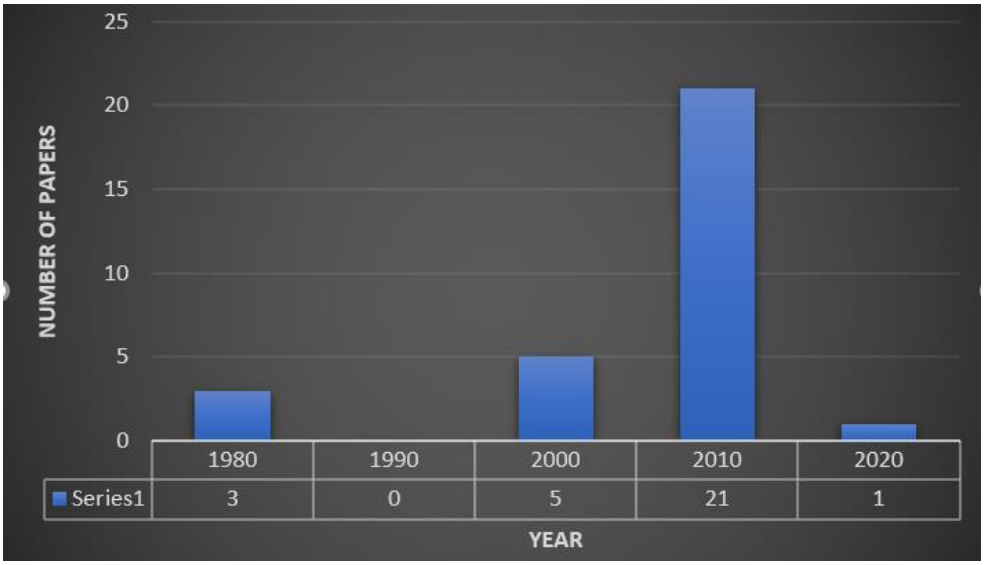


Figure 5-Number of Papers reviewed

Predictive Analytics and machine learning growth have tremendously increased. Figure 3a shows the status of the predictive analytics in 2006 were the predictive analytics was in 45 % exploration stage. The Figure 3b shows the growth of business market value in billions. These two figures widely explain the growth of predictive analytics and machine learning in the market.

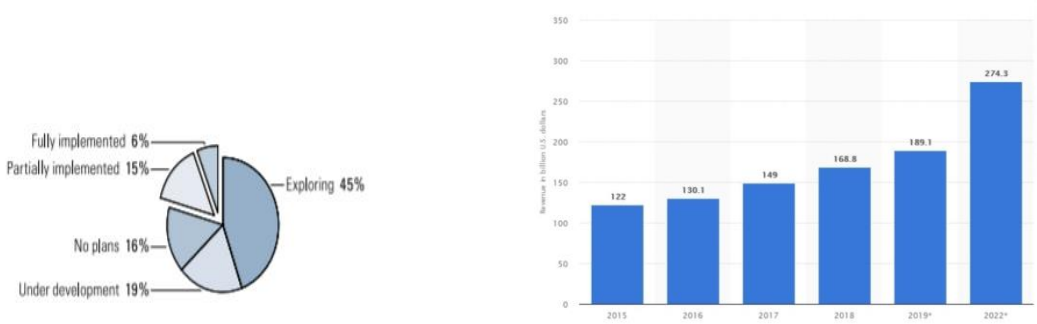


Figure 6-Predictive analytics status (a) 2006 & (b) 2020

Multiple numbers of people studied the proposed research on car value prediction. Multiple numbers of similar researches were conduction for the similar topic of prediction since 2011.

2.3 METHODOLOGY

The search for the articles and publications started with the databases: Google Scholar and Unihub Library Search. The publication article and journals were taken for the proceedings. There were multiple keywords used for the research of the papers. Data Science, Predictive

Analytics, Used car sales in the automotive industry, Used car sales value prediction, Machine learning for Prediction, Algorithms in machine learning for continuous value prediction. Out of this research keyword the Used car value prediction, Machine learning for continuous value prediction, Predictive analytics gave the maximum number of papers. There were more than 50 results were obtained, and the final 30 papers were selected by excluding those that had the least significance.

Sl No	Types Of Publication	No of Articles
Journal		
1.	National Bureau of Economic Research	1
2.	American Economic Review	1
3.	Rand Journal of Economics	1
4.	Bioinfoematic,Oxford Academy	1
5.	IEEE Explorer	7
6.	Sage Journals	1
7.	Stanford Education	3
8.	Data Science Journals	1
9.	American Journal of Applied Science	1
10.	European Journal of Science&Technology	1
11.	TEM Journal	1
12.	Journal of Physics	1
13.	International Journal of Information& Computation Technology	1
14.	Institute of Information Systems	1
Total		**22**
Articles		
1.	Springer	4
2.	Libert Inc Publisher	1
3.	Communications of ACM	1
Total		**6**
Conferences		
1.	Winter Simulation Conference	1
2.	Cogent Business & Management	1
Total		**2**
----------		**---------**
Grand Total		**30**

Figure 7-Final Set of Publication

The 30 papers selected has 22 papers in journal six as articles and two from conferences.

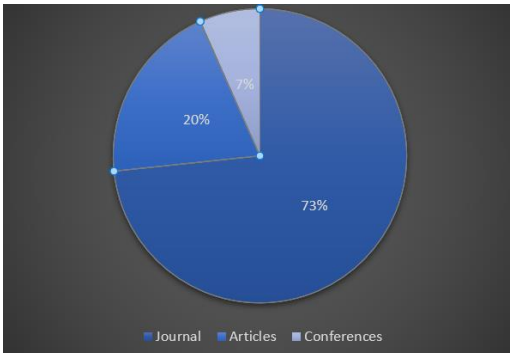


Figure 8-Classification of the Papers

The Hierarchical paradigm was chosen for the research review.

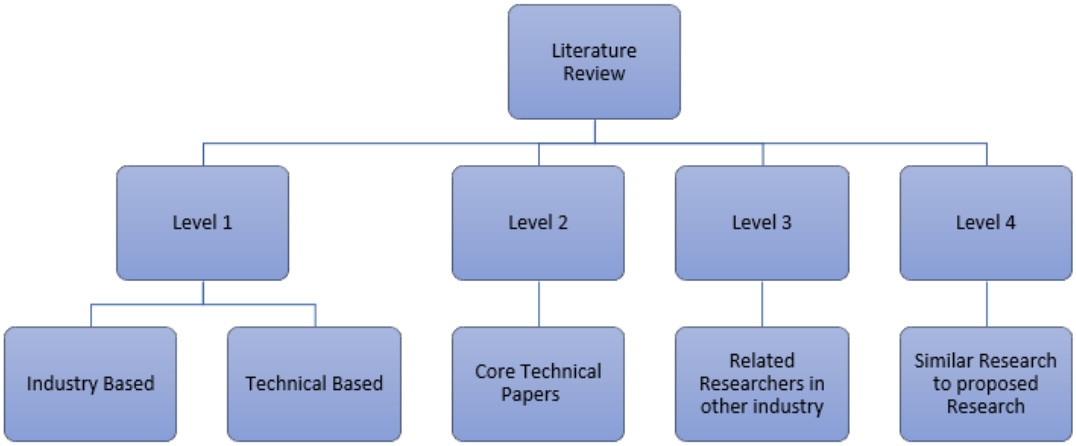


Figure 9-Hierarchical Paradigm for Review

The first level deals with the papers that are related to industry and the general technical idea about machine learning—the Level 2 deals with the study of predictive analytics. Level 3 speaks of the papers of similar researches in continuous value prediction, and level 4 dealt with all the similar works conducted.

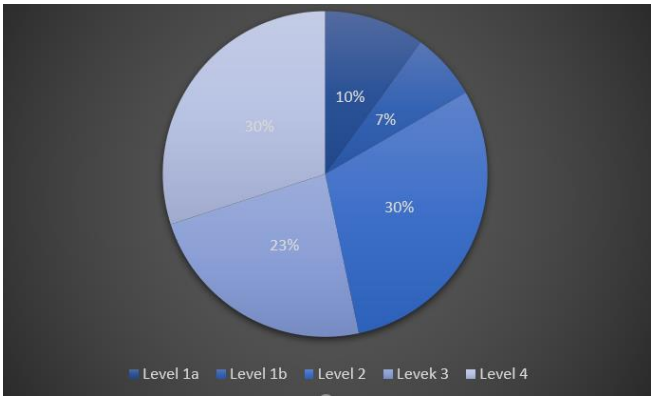


Figure 10-Classification of the papers based on levels

2.4 REVIEW

2.4.1 Level 1:

This level deal with the papers, researches, article, and books related to the industry and the Machine learning and data science. Most of the work in level 1is based on gaining the general idea in the automotive industry and the subject study of Machine learning and data science.

A total of 5 papers at this level constitute the general study. Out of the five papers, the 3 constitutes the journal papers, and two are contributed articles published. This level has the least number of papers associated.

This level dealt with the Level 1a industry study and the Level 1b technical study of the machine learning and data science. The primary issue that speaks in the industry study is the issue of equilibrium in the market in sales of the primary and secondary(Berkovec 1985) and any heuristic thinking in the market(Lacetera et al. 2012). The observation of the study also says about the differences between the past decade and the current trend.

The importance of the data used for the study is being described by Dhar(2013). There was a huge increase in the amount of data generated in recent times. Figure 8 shows the surge in data generation. The past decade study methods did not have any subtle method for modeling and did not have any assumptions about any features of the data such linearity, monotonicity, which lead to a downside in picking up the noise in the data. The distinguishing of signal and noise will be less if the features of the data are not considered. The observation of research led to an idea of the importance of feature management and the biased nature of the industry.

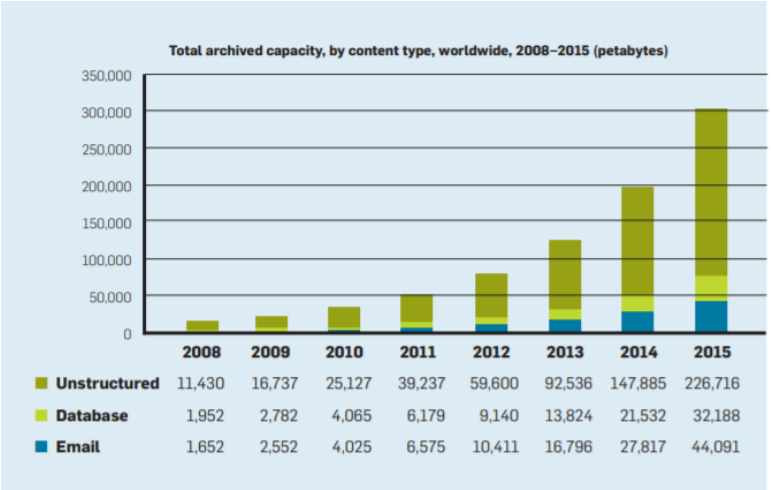


Figure 11-Distribution of data

The observation from the research of Lacetera et al. (2012) is that the value of the car is highly dependent on the value of odometer, out of which the left four-digit value of it profoundly affects the sales.

Papers	Description
Foster and Tom(2013)	The paper talks about how data science helps in making data-driven decisions and the fundamental concept of data science.
Dhar(2013)	The implication of data science from the research and business stand point. General analysis of the data science of predictive analysis with earlier years and current situation.
Gillingham et. al(2019)	This describes about the trade between the used car and the new car and modelling the equilibrium in the automobile market.
Lacetera et.al(2012)	Aim to study how the left four digit bias of the odometer value affect the affects the sale.
Berkovec(1985)	Presents the model of demand, supply , scrappage and hence to combine it to an equilibrium to understand the future trend in the market.

2.4.2 Level 2 :

This is an essential section that talks about the technical description of the predictive analytics and models that are being used in multiple types of research. This method constitutes 30% of papers of the research. This method deals with the idea of predicting its applications in areas. The core area of this level is the accuracy, sampling methods, error rate that can be used in predictive analytics and different models in predictive analytics.

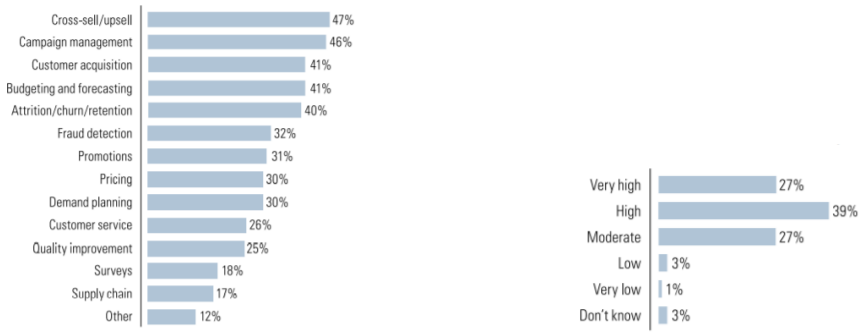


Figure 12-Business Value of Predictive Analytics

The papers of this level are highly speaking about the issue that is faced during the predictive analysis methods such as the sampling, accuracy improvement, and decreasing the error rates. The major part of predictive analytics is dependent on the mentioned methods to derive the best outcome.(Hastie.et.al,2003) states that misspecification of the model, randomness in the model even after the model is specified accurately, and sample size that is being used constitutes the bias measure.

It is observed that feature selection plays a very crucial step while developing a predictor. As the noise in the variable increases, the accuracy decreases, and it may lead to overfitting. It is challenging to select an accurate model from numerous available models. It is also essential to assess the error rate of the chosen model(Molinaro,2005). It is also observed in the paper that the ratio of the sample size and noise ratio profoundly influences

the resampling methods. The sample size and difference among the resampling methods are inversely proportional.

Knofczynski & Mundfrom(2008) papers dealt with the issue of the sample size. It tried to figure out the minimum sample size that is required for the model with multiple scenarios by varying the correlation between the predictor variable and the criterion variable. It is observed that there is a relationship that exists between the size of the sample and the squared multiple correlation coefficient.

Papers	Description
Boonsiritomachai.et.al (2016)	Paper describes the predictive analytics. Its definition, applications, and business values trends and the process of predictive analytics.
Hastie.et.al(2003)	The description of the predictive analysis of how accuracy can be increased, and an error rate can be decreased.
Molinaro(2005)	The study was on the feature selection; the model is selected and its predictive assessment. It also estimates the error rate when a feature is selected.
Ljung(2002)	Paper aims at the study of fundamental properties of Prediction Error Methods, which is highly applicable for the wide variety of structuring a model.
Knofczynski & Mundfrom(2008)	Describe the relationships that exist between sample size and the correlation coefficient.
Rodrı´guez.et.al(2010)	This aims to look at the variance of how the K cross-validation is sensitive to the training set.
Adra(2016)	Extensibility and reusability of the model being used when the predictive analysis is being done with real-time data
Burman(1989)	Study of the concept of the v-fold cross-validation and the repeated learning and the testing method.
Breiman(1984)	It is the analysis of the classification and regression models and their accuracy.

The error rate analysis is vital for model selection. Rodrı´guez.et.al(2010) dealt with analysis in bias, variance, and statistical properties of the K-fold Cross-validation. It is observed that it would be good to use a less biased error estimator for measuring the least error rate. Breiman(1984) talks about the issue of pruning a tree and getting an accurate estimation of the misclassifications. Cross-validation is very expensive in practical problems(Burman,1989), from his study, it is observed that the v-fold cross-validation reduces the cost issue of cross-validation. It is observed that the bias and variance value get decreased when the value of 'v' increases; the bias of the model depends on the ratio of the test set and the learning set.

It is essential to understand the issue that comes with real-time data. The real-time data is taken such that it should not affect the validity of the model that is used(Adra,2016). It has been observed from his research that real-time data will maximize the effectiveness of the operation and even improves the delivery time of the organization.

The observations of this level include how the accuracy of the model is to be increased, and the error rate is to be decreased with the model on a selection of the methods.

2.4.3 Level 3 :

This section deals with papers that have researched a prediction of discrete or continuous values. This section has the least number of papers associated with it. The level generally looked at the issue of the prediction of the value of an item for sale. The eight papers referred to different types of predictions and multiple ways to approach the problem statement they mentioned.

Geng.et.al (2011) tried to address the car sale prediction with the idea of Particle Swarm Optimization. The method proposed helped in finding the optimal values for the parameters in SVM. It is observed that the model solved the problem of the overfitting and gave the least Mean Absolute percentage error.

The e-commerce norm had a high effect on the image of the sales(Chen.et.al,2018).The prediction of the sale using the product image of the study performed. The feature visualization, a seller, can determine which of the feature is correlated to the higher price.

Limsombunchai(2004)researched on the prediction of the house prices with the Neural network and hedonic price models. The comparison of the model proved that the Neural network gave a more accurate result over the other model. It is observed from the study that the lack of real-time data was an issue of the result. The real-time data will perform differently compared to the pre-recorded data.

Internet Word of mouth is a factor that severely affects the sales in e-commerce(Chiu and Shu,2017). It gave the idea that sales are also affected by eWOM. Kejela.et.al(2014) studied with real-time data; it is observed that the gradient boost model(GBM) outperformed the general linear model(GLM). The error distribution in GBM is very less compared to GLM.

Papers	Description
Geng.et.al (2011)	The paper describes the car sale prediction with the particle swarm optimization and how it helps in prediction.
chen et.al(2018)	The paper describes the prediction with the help of product Images and how it affects sales.
Limsombunchai(2004)	The research aimed to predict the house prices with multiple algorithms and to compare the results of the multiple models
Chiu and Shu(2017)	This shows the customer behavior in e-commerce sales with the internet word of mouth and how it affects sales.
Kejela.et.al(2014)	The study shows how a machine learning algorithm would give an accurate prediction for both continuous and categorical variables with the help of real-time data.
Markou.et.al(2018)	The demand prediction of taxis with multiple models with the data from the web at regular traffic hours.
Huang.et.al(2015)	The study done with the trigger model over a single model in ecommerce. It focusses on forecasting a sales with high effectiveness and high accuracy.

2.4.4 Level 4 :

The level contains the papers of similar research conducted by other researchers. All the paper contributes to the details of the researches conducted with multiple models. This level gives an in-depth idea about the prediction of the used car values. The primary issue that researchers dealt with at this level was data availability. The researchers used data from multiple online sources and the paper which was already existing.

The objective of the research conducted by each contributes to the automotive industry for predicting the nominal value of the used car, which would help the customers. Voß and Lessmann (2013) in the research discussed the increasing competition in the market, which led to a decrease in sales. They proposed a well-versed prediction of sale value with the car functions and other values. There were 18 models used by them for the prediction and got a result with the linear model that has less significance than a random forest and ES model. The study conducted by Ozer and Omer(2019) made predictions with normality test, analysis on correlation, and linear regression. The data was taken from a different period, and it is observed that the value change in the variable had a significant effect on the sale value. Thus we can conclude from his observation that the variables had a significant effect on the models.

Khitij.et.al(2019) used six algorithms for the prediction. The tree model chosen for the prediction did not perform well due to the linearity in the dataset. The concluded with the result of overfitting in tree and outperformance of the ensemble model used. Puteri and Safitri(2020) made a prediction with the linear regression. They made changes in the values of the price and tried different predictions. There were significant changes observed when the value of the car was predicted. It can be observed that the logarithmic value or any other processed values can give a better accurate result than the normal price value in the data set.

Papers	Description
Voß and Lessmann (2013)	Study based on the competition in the market that reduced the marginal car sale. Used 18 models for the prediction.
Pudaruth(2014)	Prediction of the used car after the lease period. The methods used were Multiple linear regression, KNN, Decision Tree, and Naïve Bayes. Data was collected from a daily newspaper.
Ozer and Omer(2019)	The prediction of the used car was made with linear regression on ikinciyeni.com data.
Gegic, E .et.al(2019)	Prediction with Neural network, support vector machine, and random forest model. The data was taken from autopijaca.ba
Puteri and Safitri(2020)	Prediction of used car value in Indonesia. Uses Multiple linear regression—obtained data from www.mobil123.com.
Khitij.et.al(2019)	The used car worthiness depending on factors such as make, model, year. The study used Linear Regression, Gradient Boost, Random Forest, XGBoost, KMeans + Linear Regression Ensemble, Light GBM. Data obtained from Kaggle.
Ho and Wo(2012).	The aim was to make a prediction based on the price as a categorical value. Used Support Vector Machine and logistic regression for prediction.Obtained data from Kaggle.
Pal(2018)	The prediction with a random forest tree was made on the data from the Kaggle.
Gongqi.et.al(2011)	The research aimed to predict the residual value of var with various conditions using BP Neural network and non-linear curve

There exist a trend of selling the car after the lease period. Pudaruth(2014) tried to predict the values of the cars after the lease period. The research was done with Naïve Bayes, Decision tree, KNN, and multiple linear regression. He derives that the correlation between the price and mileage is high. The high correlation between price *year and a very weak positive correlation between the cylinder value * price.The price as categorical value prediction was done by (Albert Ho and Alice Wo,2012). The result obtained by their study was in contrast to the machine learning output; the balanced data set did not perform well. It is observed from his prediction that the data normalization, using of F1 score, Auc, and boosting was beneficial for prediction. Gegic E.et.al(2019)used Artificial Neural Network, SVM, and Random Forest, model. The research applied on three categories of data as cheap, moderate, and expensive data. The study draws a conclusion of the consumption of more computational resource for the method used than a single algorithm in Machine Learning.

Gongqi.et.al(2011) researched the residual value prediction with the BP neural network and non-linear curve. The result of the research was obtained as a curve. The proposed model by the researcher could obtain an accurate and convenient model with a residual value of the used car in various conditions. The prediction with the random forest was used by Pal.et.al(). The research could produce an outcome with good accuracy compared to

the related works that used the same data sets. The common problem of overfitting in the tree was overcome in this study by averaging the predictions of each tree and hence reduced the variance and product consistency.

2.5 OUTPUT OF THE REVIEW

The used car value prediction is a complex process as the value of the car may not be fixed as any other product. It may vary from a country to another country, a region to another region. The result obtained may depend only on the limited data that is used during the prediction time. The use of the different variables, methods, and periods mat result in differences in the analysis result. The outcome of the review includes challenges in the Prediction, Business values of the predictive analysis.

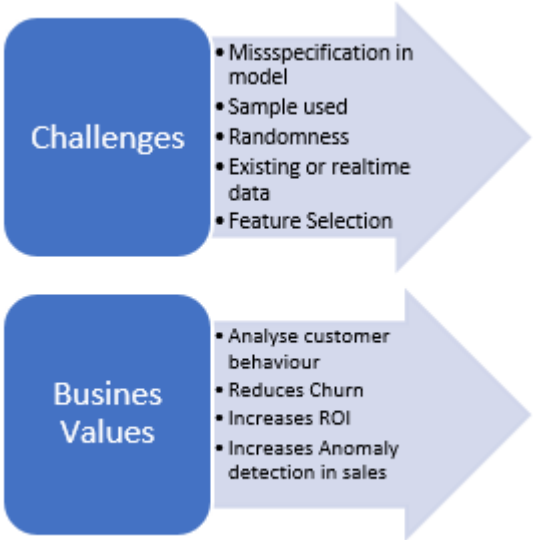


Figure 13-Output of review

The challenges in the prediction include the misspecification in the model due to any bias. The sample size is another criterion that profoundly affects the prediction. As the size of the sample is smaller, the bias will be higher. Randomness in the model can affect the error in the Prediction (Hastie,2003). The data selection for the prediction profoundly affects the prediction. The existing data prediction gives the prediction based on the data that is pre-recorded the real-time data has the current market data, which is more tend to give the accurate result, but there are chances that real-time data can disturb the validity of the model(Adra,2016). Feature selection is a crucial step in the development of class prediction. It should be chosen such that the variables are not noisy, affecting the accuracy and leading to overfitting(Ransohoff, 2004).

The business values produced by predictive analytics using machine learning is very high. The specific case includes customer behavior prediction. This would help to understand the behavior and improve the sale. The reduction in the churn rate is highly improved by the prediction. The ROI of the company with prediction has a higher ROI. It also helps to increase the prediction of the sales anomaly rate helps in the improvement of the sales.

Multiple types of research were conducted in the case of the prediction of used car values. The reviewed researches used the pre-existing data from multiple sources. The price of the car is

highly dependent on the time and the variables used for the prediction. The pre-existing data that is available for the prediction may not give good results for the research as they are pre-recorded and may not be related to the current situation in the market. It would not give the best result and affect the computation. The proposed study and the research of the real-time prediction of the used car values use the current data that is web scraped from dubizzle, which has the current value of a car in the market. The data is chosen such that it takes the brand name and model and model name and scrapes the data depending on the data given by the user. The research would bring in good results in the pricing of the car for sale to the end-user.

3: METHODOLOGY

The Proposed Research of prediction of used car values mainly focuses on predicting the values based on the predictive algorithm. The research intends to focus on the aim and objectives described. The aim primarily intends to predict the value, and the objective revolves around the successive application to complete the aim. The objective of the research raises the research question of how data dependency is associated with the outcome. The core of the research is based on predictive analytics with multiple algorithms to model. The model is hence compared to choose one that has less error and best accurate to build the product.

The research started with data extraction using R codes. The research uses multiple tools such as rapid miner, Jupiter Notebook, Tableau. It was followed by an understanding of researches conducted by other researchers. A deep understanding of the methodology that other researches had used was studied, and the methods were selected. Post selection of the method followed the study regarding data. Comparing different papers regarding the data used for the modeling was taken, and could see that all of the researchers conducted the studies based on the pre-existing data. The proposed research was not based on heuristic thinking of pre-existing data instead considered it the black space between the existing research and proposed research. The proposed research is based on the live data depending on the input. Data extraction was followed by the data understanding using Tableau and then modeled using R codes and packages. The rapid miner tool was used to make an auto model for multiple models to see the error rates and accuracy developed by multiple models.

3.1 DATA EXTRACTION :

Data, a critical element of Machine Learning, plays a vital role in the modeling. The data selection and extraction is very crucial. The dependency of each variable and its correlation is vital to the predicted value price; hence the proposed research uses live data rather than previously recorded data.

The data extraction is done on live scrapping based on user input. The data was scraped from dubizzle.com. The data scraping was done using the RVEST package in R. The rvest package helps extract the text data from the HTML page using the HTML tag and HTML nodes. The HTML page consists of multiple pages with multiple brand and model car details. In order to extract data from multiple pages, we use the package called PURRR. For an accurate prediction, the data set should be scrapped so that scraped data only has the data of the same brand and model that the user has input. The steps followed for the scraping include:

1. Getting the generalized URL for the page with the same modeled car
   The data is scraped from the particular page with multiple cars.

2. Get the HTML for the webpage that is needed to be scrapped.
   The scraped landing webpage has limited availability of the data. The central data is available on the landing page; hence to get the detailed data of the particular car, it is needed to redirect to the detailed page of each car and get the details. Hence from the landing page, all the reference links are scraped using the HTML NODE class name and stored in a data frame for the mentioned number of pages.

3. The extracted links of all cars stored in the data frame are taken one by one and select the HTML tag and HTML Node to be scraped and stored back in the data frame.

The 'mapdf' function of the PURRR package helps in parsing the multiple pages and hence scraping the data from multiple pages.

3.1.1 DATA PLATFORM

The proposed research chooses dubizzle.com as the data platform. The scraping mentioned is done in the dubizzle data, and it is stored for the modeling. There are multiple other data platforms available such as Yalla motors, sellanycar.com.The data can also be extracted from that site.As the data platform changes, the data format changes; Hence only one source can be used at a time. The proposed research has chosen the mentioned data platform because, in comparison with multiple other sources, the data available of this source was high, and the meaningful data for the prediction with more data variables were available here.

The HTML Extraction code attached in Annexure 2

3.2 DATA :

The scraped data consist of 14 variables Description, data types, and possible value of the data as mentioned below:

VARIABLE NAME	DESCRIPTION	POSSIBLE VALUE	TYPE
Lr	The web scraped links from the model car's landing pages, and the link is used for scraping the details.		
year	Model Year of Car	1990+	int
mileage	Kilometers Driven by the car	0+	int
color	Colour of the car	Any color	Char
transmission	Transmission Type of the car (Manual, Automatic)	Automatic or Petrol	Char
spec	The country of Manufacture	GCC, European, Japanese, American, Others	Char
bodycond	The interior and external Body condition of the car	Perfect inside and out, Normal wear & tear a few issues, No accidents, very few faults, A bit of wear & tear, all repaired, Lots of wear & tear to the body	Char
mechcond	The mechanical condition of the car. Mention if any mechanical issue associated with the car	Perfect inside and out, Minor faults, all fixed, Major faults, all fixed, Major faults fixed, small remain, Ongoing minor & major faults	Char
body	The type of car	Coupe, Crossover, Hard Top Convertible, Hatchback, Sedan, Soft Top Convertible, Sports	Char

		Car, SUV, Van, Wagon, Other	
warranty	If there exists any warranty for the car	Yes, No, Does not Apply	Char
trim	Mentions the number of the door	2,3,4,5	int
cylinder	Number of Cylinders	3,4,5,6,8,10,12,Unknown	Char
fuel	Fuel type of car	Gasoline, Petrol, Hybrid, Electric	Char
seller type	Type of user who posted the advertisement if an owner or a dealer	Owner, Dealer, Dealership/Certified Pre-Owned, Unknown	Char
price	Price of the car	0+	int

The data variables type and the possible value taken by them are shown. All the possible values expected are mentioned. The values are given a list of drop-down for the users. The list of data scraped from the dubbizle has mentioned has only these values.

3.3 DATA PREPARATION

3.3.1 Data Cleaning:

The data cleaning is essential because it identifies and removes the errors and duplicate in the data to get reliable data. The cleaning method would improve the quality of the data to be trained and enable us to get accuracy in decision making.

The data extracted shown above listed. The user is expected to input all the variables as user input. The user selects the value from the drop-down list. Since the user is given the option to select from the drop-down, the possibility of the data entered by the user to be wrong does not exist. Since the scraped data has only given values available rate of error possible to happen is significantly less.

The extraction of data was not post stored and processed since the data extraction and modeling takes place simultaneously. Hence the cleaning was done simultaneously along with the extraction and then stored to the data frame. Since the data extracted only has a fixed value, the possibilities that can happen is the error in the text extraction of the HTML. Possible data cleaning ways are as follows:

1.The error in the scraped data during text extraction

VARIABLE NAME	Scraped Value	Cleaned Value
mileage	“\n \n \n 66,000 \n”	66000
colour	“\n \n Black \n”	Black
bodycond	“\n \n No accidents”	No accidents
mechcond	“\n \n Minor faults\n”	Minor faults
trim	“\n \n 4+ door \n”	5
cylinder	“\n \n \n 2 \n”	2
fuel	“\n \n Gasoline \n ”	Gasoline
seller type	“\n \n Owner \n”	Owner
price	“\n \n AED 99,000\n”	99000

2.There can be chances that certain values in the html page can be unfilled.

If the html accessed is null it can lead to break of scraping due to the empty value. It is cleaned by forced adding 'NA' Value to empty column.

3.The 'NA' accessed from page.

It can be removed by removing those rows with empty rows using

```
na.omit(yy)
```

4.subseting the data set to remove lr from the data set, which is of no use.

3.3.2 Data Pre-processing:

Data needs to be pre-processed to get accurate type of data. The extracted data, in default, have the data type depending on the type it was extracted. The default data type may not give the exact meaning and value to it as it is intended; hence the variable has to be changed to the exact data type it is intended to be. The exact data type improves the accuracy of prediction.

VARIABLE NAME	Old Type	New Type
year	Int	Numeric
mileage	Int	Numeric
color	Char	Factor
transmission	Char	Factor
spec	Char	Factor
bodycond	Char	Factor
mechcond	Char	Factor
body	Char	Factor
warranty	Char	Factor
trim	Int	Factor
cylinder	Int	Factor
fuel	Char	Factor
seller type	Char	Factor
price	Int	Numeric

3.4 DATA ANALYTICS PIPELINE

The pipeline will automate the workflow.

Advantages:

Flexibility: Computation is made easy by each module. Change to one module can be made without affecting the other module

Extensibility: Easy to make new functionality

Scalability: If any part occurs in any issue, it can be possible to scale them separately.

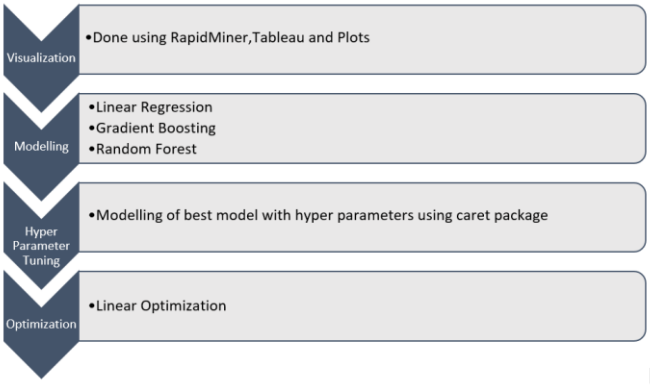


Figure 14:PIPELINE

3.4.1 VISUALIZATION

The research models with three algorithms. The algorithms are chosen depending on the performance marked by the others in the literature review conducted. The research is based on the prediction of used car value. Hence the dependent variable is the price, and independent variables are year, mileage, color, specs, warranty, body, bodycond, mechcond, fuel, trim, cylinder, seller_type, and transmission. Hence there are one dependent and 13 independent variables. The study for the variable is based on data input given FORD EXPLORER. The data structure is as follows.

3.4.1.1 Data Visualization:

year	mileage	colour	price	warranty	specs	transmission	bodycond	body	mechcond	fuel	trim	cylinder	seller_type
2016	93314	White	66000	Yes	GCC	Automatic	SUV	SUV	Perfect inside and out	Gasoline	5	6	Dealer
2018	35	Black	149000	Yes	GCC	Automatic	Perfect inside and out	SUV	Perfect inside and out	Gasoline	4	6	Dealer
2019	2647	White	155000	Yes	American	Automatic	Perfect inside and out	SUV	Perfect inside and out	Gasoline	4	6	Dealer
2020	1500	White	179000	Yes	American	Automatic	Perfect inside and out	Truck	Perfect inside and out	Gasoline	4	6	Dealer
2018	100	White	149000	Yes	GCC	Automatic	Perfect inside and out	SUV	Perfect inside and out	Gasoline	4	6	Dealer

Figure 15: Data Structures

Correlation factors help us to analyze the dependency relation of variables with each other. The variables are converted to numeric to graph the relationship with other using fit_transform in python.

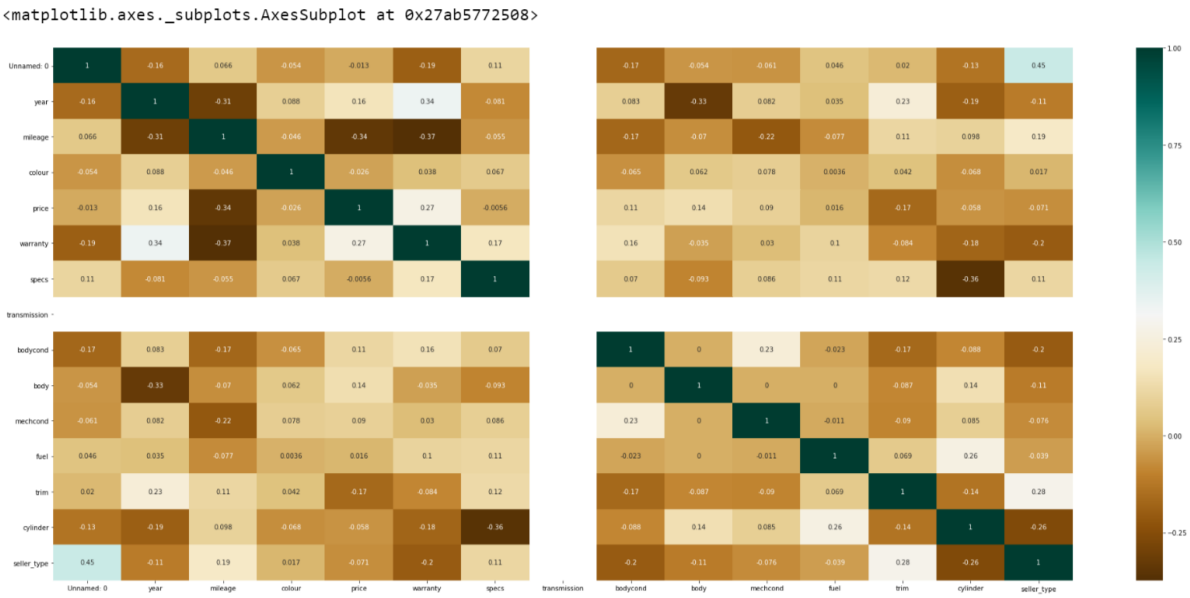


Figure 16:Correlation Plot

From the plot, we can gain information such as

1. Variables are not very highly correlated; hence it is suitable for using the modeling.
2. Highest correlation variable with the price is mileage(-0.34).There exists a negative correlation with mileage; hence, as the mileage increases, price decreases.
3. There exist a small negative correlation(-0.37) between warranty and mileage.
4.Correlation of 0.34 exist between warranty and year. It implies that the older the car, the warranty does not apply.
5.Warranty and price have a positive correlation. The cars with a warranty have a higher price.
6.Year and mileage are positively correlated. Newer the car Mileage is less .There can be a case of vice versa; hence the correlation value is less.

The below image is extracted using the missingno package in python and helps locate any missing values in data, and hence it is clear that the data set does not have any null values. This helps us to confirm that the pre-processing step has removed any null values if it existed.

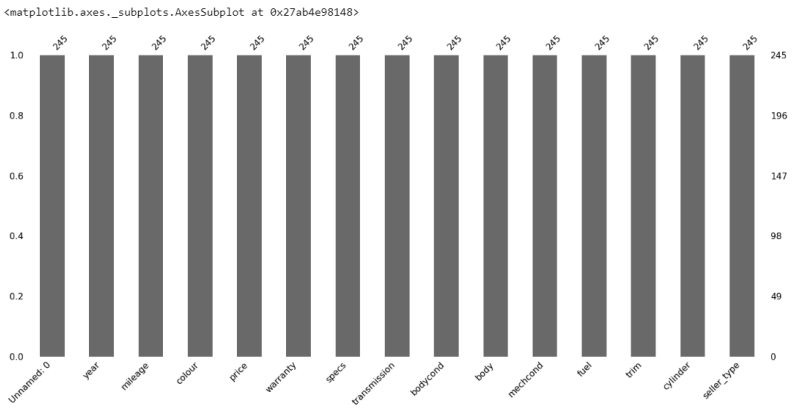
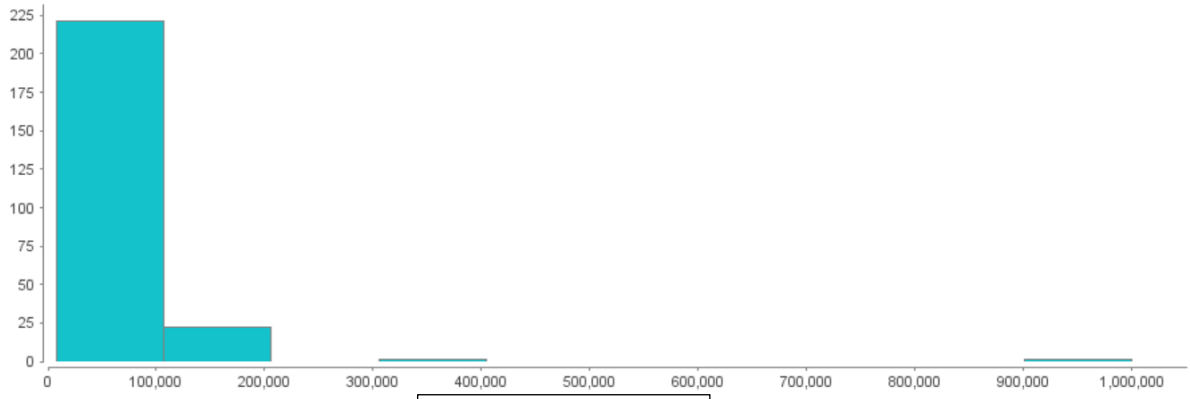


Figure 17:Missingness

Let us look at the variables to see how the values are distributed in the data.

Dependent Variable:



Price

The dependent variable price is distributed, as shown. We can see that there exists 1 data value of price 1,000,000 AED, and it cannot be an outlier. The data input value of 1,000,000 to be bid for ford vehicles is very rare, and hence it can be considered false data. The value range of the particular car lies between 0-200000 AED. The data of 3,00000 can also be false data.

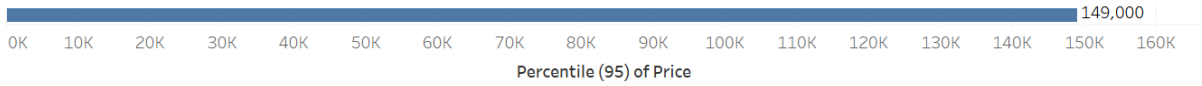
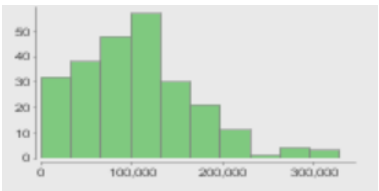


Figure 18:Percentil 95

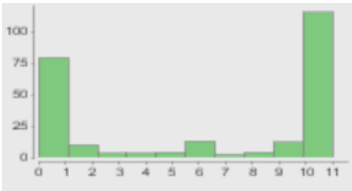
The percentile graph of the price variable proves the mentioned data are false data. The 95 percentile value of the price is 149000. There can be a possible chance of rare higher prices; hence any prices above percentile 95 + 25000 AED can be considered false data and can be removed.

Independent Variables:

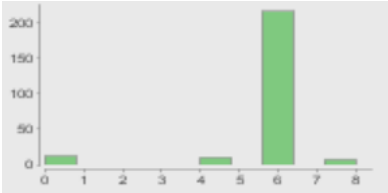
The independent variable distribution graph helps us understand how the independent values are done by transforming the categorical data into categorical values.



Mileage

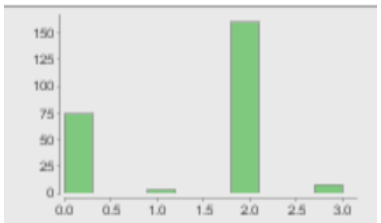


Colour

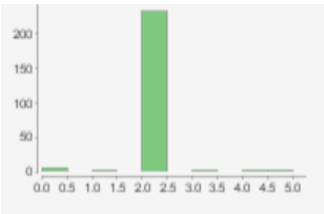


Cylinder

1.Mileage distribution shows that most of the selling cars are in the range of 100000.
2.Colour distribution shows that specific colors are prominent in cars
3.Ford explorer mostly comes in 6 cylinders. Similarly the distribution will be different brands.



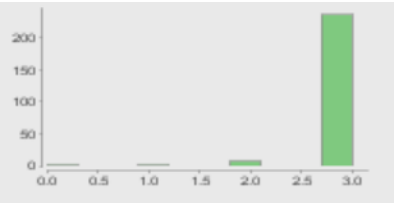
Seller Type



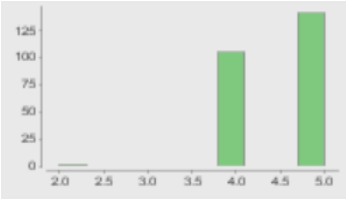
Body

Fuel

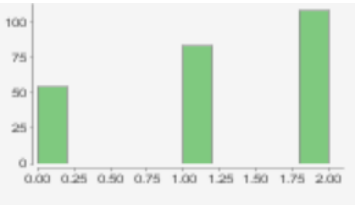
4.The owner and dealer based sell more.
5.Most of the Ford explorer comes in SUV type.
6.The Fuel is Gasoline



Mech Condition

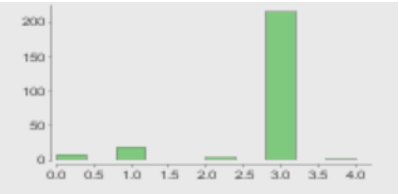


Trim

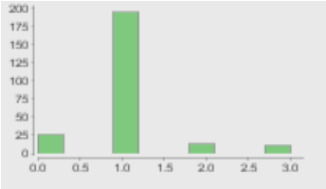


Warranty

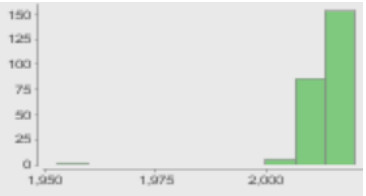
7.Most of the cars are of good mechanical condition
8.Most of the Ford explorers are 5 or 4 doors.
9.There can be cars with warranty and without equally available in the market



Body Condition



Spec



Mech Condition

10.The perfect body condition cars are more in the sales for the market. However, mechanical condition distribution shows that there can be a variable kind of mechanical condition. It implies that the car comes to sale with an excellent outer appearance but may have some mechanical issues.
11.GCC vehicles are more in the market.

The relation between Dependent and Independent Variables:

The below images depicts the relation between the

(a)certain variables with sum(price)

Figure 19:Sum(Price) vs. Variables

1.The visualization between variables and sum (price) .This helps us understand which variables are more prominent in sales of the car.
2.The color variable shows that White, Black followed by the silver colors are the color of the car that are mostly available sales. It proves one of the assumptions of the car sale that we have.
3.Perfect mechanical and body condition cars are the highest in the sales market.
4The sale type of car usually comes by owner and dealers. A high percentage of cars have a warranty

(b)certain variables with Median(price)

Let us now draw the inference on the price with regards to a few variables. For drawing inference, we use the median of the price so that we can see how variable types are affecting the price

1.The sales market has more cars of color White, black, and silver, but the price of rare colors such as burgundy has more median (price).This is because the prominent colors have the price value spread on the range whereas rare colors come with more price and few numbers.

2. Price concerning the body and mechanical condition is evident that vehicle with no-fault and minor faults has a higher price; hence the assumptions come true.

3.Price of a car with a warranty is more.

4.Seller type with owner and dealer has higher sales, whereas Dealership/Certified was less compared, but the price of Dealership and Certified is more. It can be because of the varied price range in a car similar to the color, and the dealership charges the seller more cost for the profit and charges more. Hence the sale price is higher.

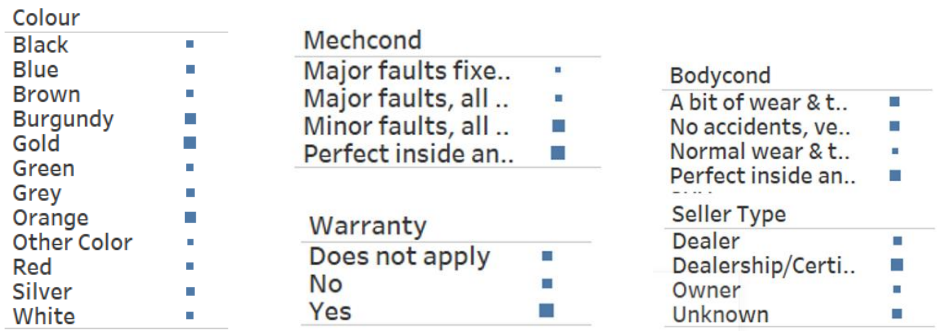


Figure 20:Median(Price) vs. Variables

(c)The relation with sum(mileage) and sum(price)

Most of the graph and assumption talks about the correlation between mileage and the price. The correlation index of the negative value indicated an inverse relation between Mileage and Price.

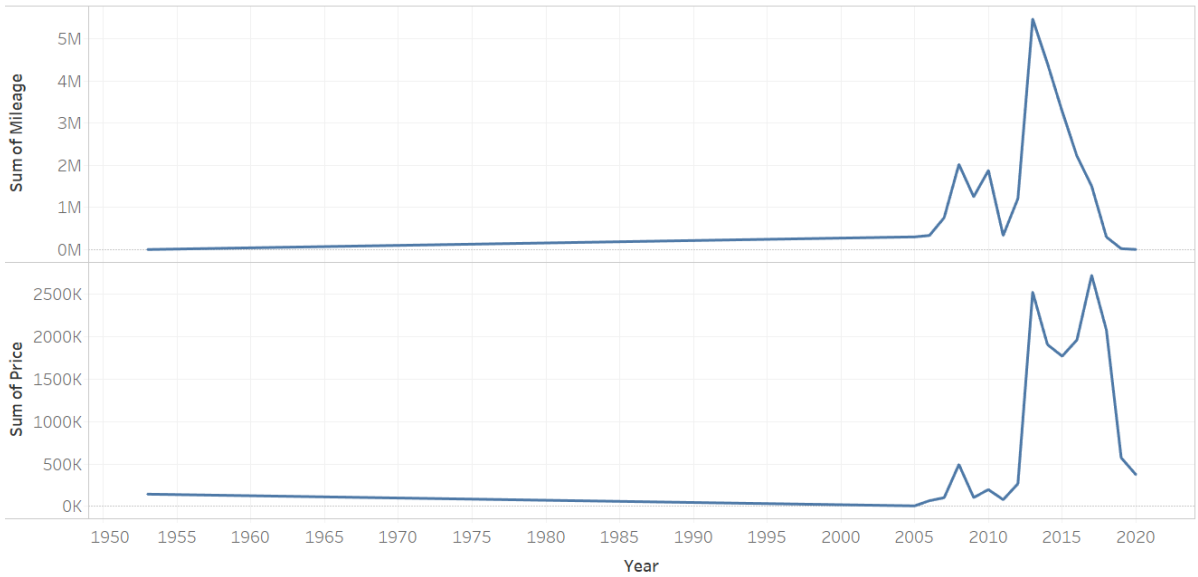


Figure 21:SUM(Price) Vs. SUM(Mileage)

The sum of the mileage and price graph shos the same rice pattern as it is the calculated sum. There was a fall in price post-2014 due to oil price reduction.It severely affected the market, which led to a drop in the graph of mileage from 2014. If we consider the median price and median mileage, we can see that the inverse relationship between the mileage and price is apparent, and they inversely draw a similar pattern. Hence it is visible the relationship of price to mileage.

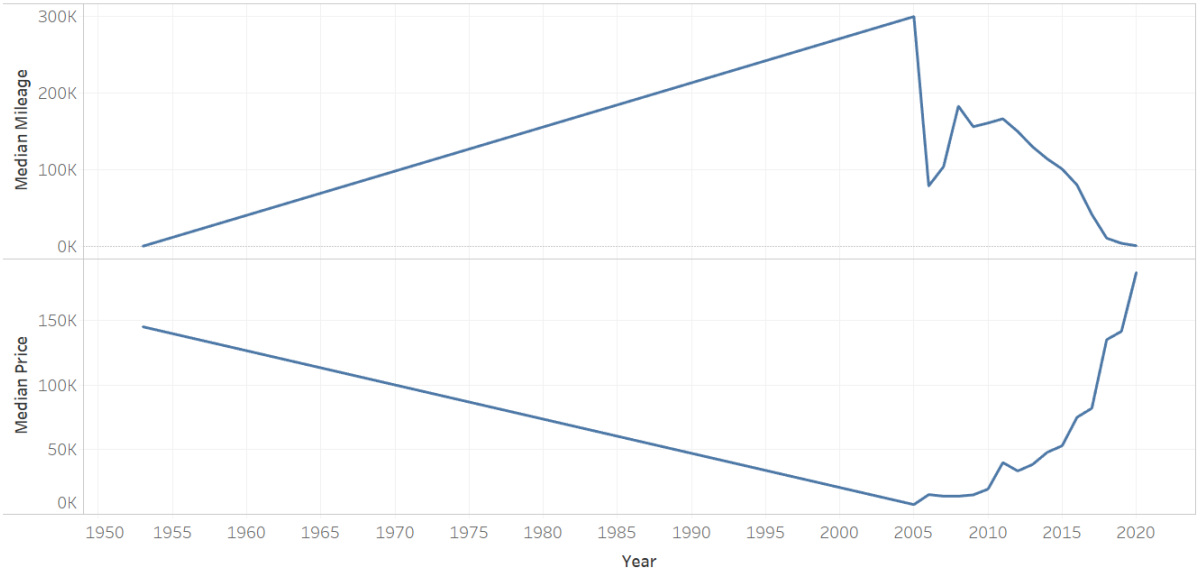


Figure 22:Median(Price)vs Median(Mileage)

Let us look at different models expected to perform well-chosen through many studies conducted.

3.4.2 MODELLING

3.4.2.1 Linear Regression:

Linear Regression(LR) is one of the most accessible models that takes the least time to execute the models. It can be used as a baseline model for comparison with the other models. Linear regression is models with all the data variables to obtain the summary to see which variables fit best for the model. The linear regression's primary aim is to estimate the values of the parameter based on the data.

LR is applied to the data set in three types of data split for test and train.

The data found as false data was removed in the data cleaning, and it and other pre-processing parts are done so that it is good enough to be used for modeling.

The initial modeling started with inserting all 13 independent variables and price as the dependent variable. The correlation coefficient computed has shown that the variables do not have a strong relationship with the price, and mileage and year are inversely related.

	Unnamed: 0	year	mileage	price
Unnamed: 0	1.000000	-0.162032	0.065531	-0.012948
year	-0.162032	1.000000	-0.308251	0.159740
mileage	0.065531	-0.308251	1.000000	-0.336487
price	-0.012948	0.159740	-0.336487	1.000000

Figure 23:Correlation value

The below graphs show the data distribution of price vs. year—this graphical representation has false data.

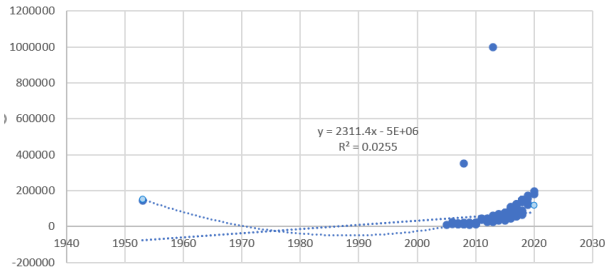


Figure 24:Regression Curve of Price vs. Year

The below graph depicts the relation between the independent variable mileage and year with the dependent variable price. The plotted graph shows that they do not have an entirely linear relationship. The false data in graph1 was removed and obtained a new graph to obtain consistency and more linearity. We can see that the R2 value changed from 0.02 to 0.66.

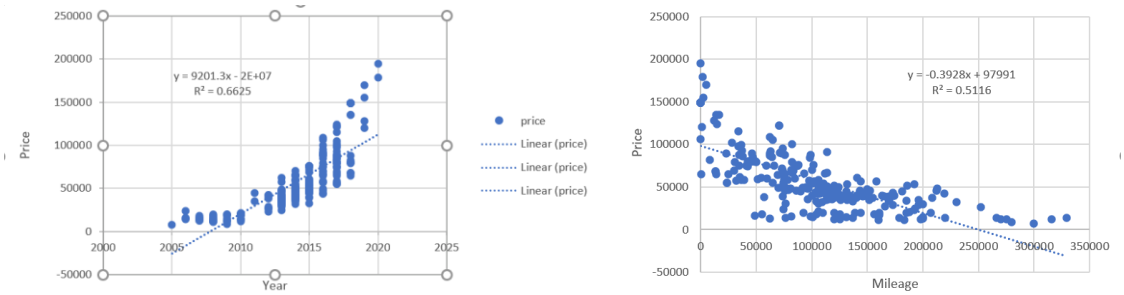


Figure 25:Regression Plot Year and Mileage Vs. Price

The logarithmic value of the price is considered to increase the linearity and R2 value. The result was studied.

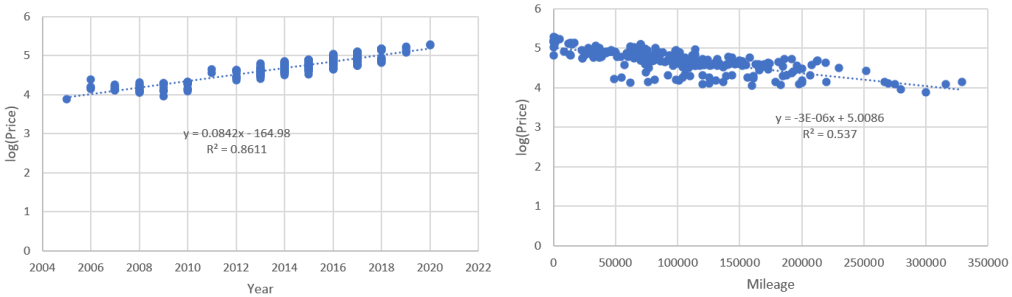


Figure 26:Regression Plot Year and Mileage Vs. log10(Price)

There was a massive changed result obtained. The year vs. price graph turned to be linear, with an increase of r2 value from 0.66 to 0.86. The mileage vs. price graph had an increase in the value of r2 from 0.53 from 0.51. The distribution got narrowed and got more linear.

After removing an outlier, the model can be considered a base model; the logarithmic value of the price is taken to be the dependent variable. The model is built on three variants of data set to see the performance

Let us now remove the particular variable with the least importance and do the modeling to see the modeling's characteristics. The variables such as mileage, year, trim, warranty, mechcond, bodycond, body, and seller_type are used for modeling.

The result is obtained based on the actual and predicted Mix Max Accuracy, MAPE(Mean Average Percentage Error), and AIC(Akaike's Information Criterion), which define the fitness of our model from the data it was generated. A comparison of actuals and predicted are made based on the standard deviation of predicted and actuals. MAPE is expected to be the least possible, and AIC should lie between 2K – 2(log-likelihood); hence value should be as least to 2k.

The result obtained was compared with the baseline model and the model with logarithmic price values. The result of all the three models has been analyzed and studies. The results are now stored to compare with the other models.

3.4.2.2Random Forest:

The random forest model, also called a Random decision forest, belongs to an ensemble method. It can be used for both regressions as well as classification problems. This method was developed for the problem of overfitting. This model was applied with data as in linear regression to test the accuracy of the prediction. The random forest model is one of the meta estimators that can fit multiple decision trees on the samples and uses the averaging to control overfitting and improve the accuracy. The random forest model has the main advantage of improving the bagging as it can correlate the tree by introducing the splitting based on a random subset of the variables.

The variable's importance is one of the significant factors in random forest modeling. Let us look at the variable importance of random forest modeling. Hence compare the feature importance to see the index value of feature importance with node purity and MSE.

The node purity, which is calculated base on the Gini index impurity, is a measure of the probability that a variable wrongly classified when chosen. The higher the node purity value, it decreases the Gini score, and the higher would be the variable importance. Hence we can see the essential features that can be used. The features that can be used in primary are mileage, warranty, color, trim, fuel, specs, body, cylinder.

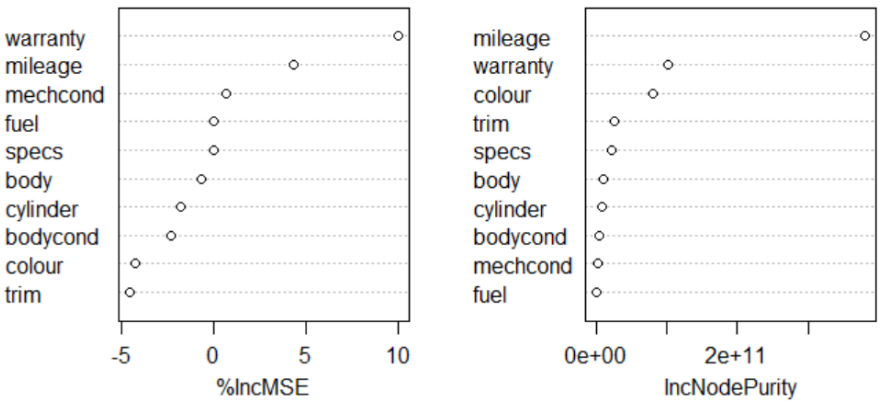


Figure 27:Inc MSE and Inc Node Purity Plot

The features are now given the model based on the three splits as similar in linear regression. Unlike the linear regression, the random forest differs in the way of creating model and prediction. The random forest runs multiple trees that are independent of each other tree. The random forest is run with tuning based on the three splits mentioned and tries to get the best

possible 'mtry' parameter value possible. It is the number of the variables that are available for splitting at the node. The fitting tris to run for multiple trees and tries to get the best mtry parameter available, and hence it is fed to run the prediction model.

The evaluation factors for the random forest tells if the value of the model fit the prediction. The model given is evaluated using the MAPE, AIC score, Min-max accuracy, comparison of the actuals and predicted, and the RMSE value. The random forest is run on the R with the package called 'RandomForest.'

3.4.2.3 Gradient Boosting Model:

Gradient Boost model(gbm) is being used in the proposed research to gradually and sequentially train the data.The customization depending on the needs of the application, can be done with the loss function. This model will continuously work on the error fitting and allow the user to experiment with multiple models. The implementation comparing to the other model is easy, and the training time is less for gbm.

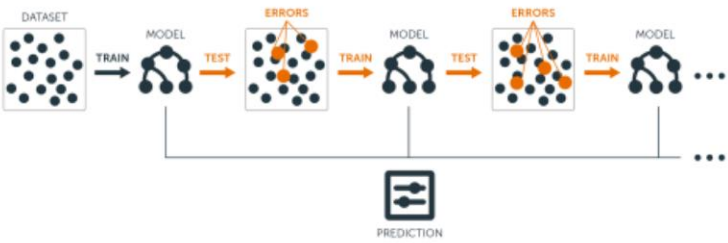


Figure 28:Gradient Boost Model

The modeling is done based on the data set with three splits as similar in other models. The package required for the gradient boosting model to run is gbm.T he gbm model feds the parameters such as

n.trees: the number of trees to be fit
interaction. Depth: Control the complexity in the boosted ensemble.
Shrinkage: reduces the learning speed to reduce the impact of any unstable regression coefficients.
distribution: distribution of the variable

The gbm model requires the gbm library. Initially, the formula which specifies the dependent and independent variables are given. The distribution for the response variable is given. Here we consider the Gaussian model to get the squared error. The data specification then follows it, and other parameters such as n.tree, default value considered is 100. ntree value is tested for various values to find fitting of each tree and best is selected. We also provide the interaction depth to specify the depth of tree and shrinkage for the learning rate as it can reduce any impact of additionally fitted learners. It reduces the size of incremental steps, and hence penalizing the vitality of every consecutive iteration.cv folds criterion is added to remove the early stopping in gbm. Method argument specifies the technique that is used for deciding the optimum number of trees. The method used here is cv.

The gbm model is evaluated using the error rate. The gbm package, which has default gbm.perf(), is used in evaluation in finding the optimum numbers in the iterations. The

accuracy measure, error rate(MAPE, MAE, RMSE), and actual and predicted comparison are also used to evaluate the model.

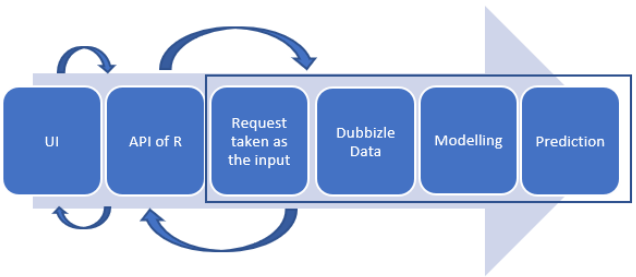
3.4.3 Hyper Parameter Tuning of Model:

The research now ran by changing multiple parameters with different datasets and different models. With the evaluation criteria, the models are compared with the baseline model and other models to see each's performance. Out of all the models, we can see that gbm with regular split gave the best accuracy and least error and gave good values for the other criterion. Hence, we perform the hyperparameter tuning on gbm to see the performance and choose the best possible values for which our model is entirely fitting.

The hyperparameter started with the tuning of the n_estimators and the learning rate. The n_Estimators were capturing the tree. It is tuned to reduce the computing expense and to increase accuracy. The R uses the trainControl function of the caret package for tuning the parameters. The train control output is given to the model as the argument for training—the hyperparameter tuning. The fitcontrol function has parameters such as method, which takes the method for tuning the optimum number of trees. The method used is repeatedcv as it can repeat the random sampling validation. The method argument gives the number of resampling iterations to be done. Repeats are used, as our method is repeatedcv. It says the number of the complete set of the folds. The classProbs is a boolean mentioning if the class probability is computed with predicted value in every sample. The traincontrol output is fed to train (), which does the hyperparameter tuning as the parameter to trControl.The other parameters for the train function are a method that describes which model is to be done. The gbm is our expected model, and we feed gbm to the method, and tune length is set for the number of levels for trying different parameter combinations. The higher the tunelength, the more combination is applied. We apply1, 5, and 10 as to compare the performance of tuned parameters.

3.5 PROTOTYPE OF THE END PRODUCT:

The research product is a web page that can be filled with the data as per the requirement. The user entry fields are given appropriate value for the selection depending on the data similar in the data plat form.When the user submits the data it is send as a request to the backend (Rcode).The backend code that we have generated is exposed as an API using plumber package which is then consumed by the UI and hence getting the output of prediction as a response.The response will be shown as the output in UI.



1.The user inputs the data sent to the backend (R Code) as a request through the API Exposed.

2.The request from the UI is taken as the input and fed the code the brand and model for scraping from dubizzle.

3.The scraped data is given for modeling using gbm model.

4.The prediction is done and result is send back to the UI as a response to the UI through the API exposed which will be displayed as the output.

4: RESULTS & FINDINGS

The predicted research was modeled on multiple algorithms to find the best model that can fit the data. The research looks at a good value for the results that are to be obtained from the models. Let us now look at the results obtained from different models.

4.1 PRE-PROCESSING:

One of the significant parts is that data processing is to find if any false data or outlier can happen in data. The real-time data can be very challenging. When considering data from a data source for car sales, it can be twice the error rate than other data as customers tend to bid any value for sale. Here we use the quantile function, which gives the percentile value of the cars. 95% of the price gives the value that the model can bid. However, there can be a chance that a particular buffer value more than the maximum value can happen. Hence we add 50000 along with the percentile value to make the range as the maximum. We can find two outlier values of 1000000 and 350000, which is a cost even higher than the new owner car.so we remove it as false data.

```{r}
print(quantile(yyy$price,seq(0.95,1))+50000)
```

```
    95%
 199000
```

```{r}
yy<- subset(yyy,yyy$price>quantile(yyy$price,seq(0.95,1))+50000)
yy
```

year <int>	mileage <int>	colour <chr>	price <int>	warranty <chr>	specs <chr>	transmission <chr>	bodycond <chr>
2013	140000	Black	1000000	Yes	GCC	Automatic	Perfect inside and out
2008	50000	White	350000	No	GCC	Automatic	Perfect inside and out

Figure 29:Removing Outlier

The Null values of the data are tested with complete.cases(). Our data proves to be complete without any null values.

```{r}
complete.cases(carDf)
```

```
  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [14] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [27] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [40] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [53] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [66] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [79] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 [92] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[105] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[118] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[131] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[144] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[157] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[170] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[183] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[196] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[209] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[222] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

Figure 30:Null Check

The price value is distributed along with a long-range. The distribution of the variable is normalized. The logarithmic value is applied to the price variable. The regression graph in the above methodology firmly proves that price logarithmic value gives a more accurate prediction. The distribution of the price is shown below.

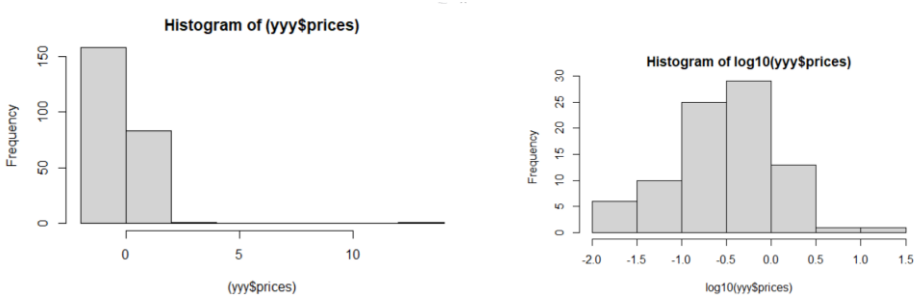


Figure 31:Price value Distribution graph

4.2 MODELING:

4.2.1 Linear Regression

The linear regression is the most accessible model in terms of runtime. Let us consider the normal data split value as the baseline model.

```
Call:
lm(formula = prices ~ mileages + year + colour + warranty + bodycond +
    body + mechcond + trim, data = traindatas)

Residuals:
   Min     1Q Median     3Q    Max
-64466 -17498  -3365   9726 912445
```

```
(Intercept)                                                                  .
mileages                                                                     *
year                                                                         .
colour\n                                        Blue\n
colour\n                                        Brown\n
colour\n                                        Burgundy\n
colour\n                                        Gold\n
colour\n                                        Green\n
colour\n                                        Grey\n
colour\n                                        Orange\n
colour\n                                        Other Color\n
colour\n                                        Red\n
colour\n                                        Silver\n
colour\n                                        White\n
warrantyNo
warrantyYes                                                                  .
bodycondNo accidents, very few faults
bodycondNormal wear & tear, a few issues
bodycondPerfect inside and out
bodycondSUV
bodyOther
bodySedan
bodySports
bodySUV
bodyTruck                                                                    *
mechcondMajor faults, all fixed
mechcondMinor faults, all fixed
mechcondPerfect inside and out
trim
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69590 on 196 degrees of freedom
Multiple R-squared:  0.1943,    Adjusted R-squared:  0.07923
F-statistic: 1.688 on 28 and 196 DF,  p-value: 0.02168
```

Figure 32:Structure of Linear Model

The linear regression baseline model ran with the data such as mileages, year, color, warranty, bodycond, mechcond, body, and trim. The model ran with all possible data variables. The variable transmission and seller_types are omitted as the transmission had only one value and seller_type had the value dealership/certified in very few numbers, resulting in an error of new testing variables. The residual error is the average distance that the value can have from the regression line. The value is 69590, which is very high. It is to keep to low as possible on 196 degrees of freedom. i.e, on 196 varying independent values, the residual error is 69590. The multiple R2 shows the correlation of the dependent variable with the independent variable. The value is expected to be more than 0.50. Fstatistic also compares the correlation of price with other independent variables.

The regression-based on the graph is shown below the year and mileage, two significant variables for the prediction. The above pictures depict the linear regression result over normal price values, and below depicts the variation over logarithmic. The year variation of the regression graph performs well with the log value of the price.The variation is still high

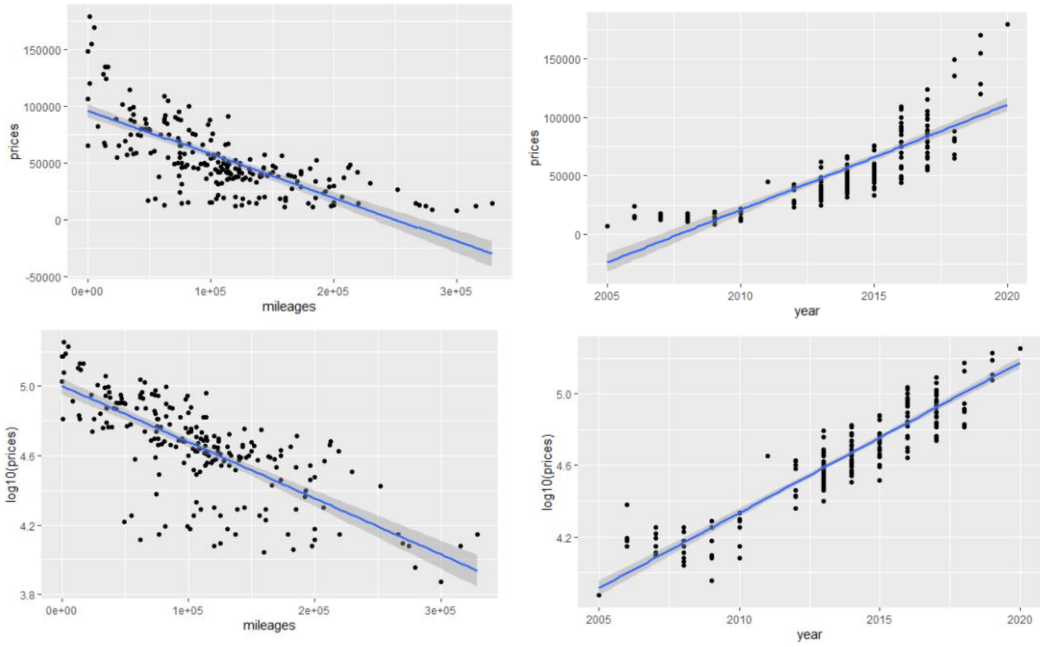


Figure 33:Regression Prediction Plot

The prediction with normal value varies as follows.

actuals <dbl>	predicteds <dbl>
40000	41278.85
79000	103162.32
60500	62837.38
76000	80922.34
70000	54389.90
40000	27145.02
39000	41126.67
19999	14100.14
53000	46164.21
59900	71950.21

```
[1] "Min_Max Train Accuracy"
[1] 0.5292624
[1] "Min_Max Test Accuracy"
[1] 0.8180206
[1] "MAPE"
[1] 0.1931012
[1] "MAE"
[1] 12388.27
[1] "RMSE"
[1] 15500.25
```

Figure 34:LR with Normal Split and Original Price Value

The values on the modeling of the linear regression are shown. It is seen that with normal values of prices, the test accuracy is much higher than train accuracy. When the price is standardized using the log values, the results vary from the previous result. The significance of the variable changes very much when the value of the price is changed. The accuracy value increases, and train test accuracy variation and the Average error rates get lowered. The results with the log values are shown below.

```
(Intercept)                                                                  ***
mileages                                                                     ***
year                                                                         ***
colour\n                                    Blue\n
colour\n                                    Brown\n
colour\n                                    Burgundy\n
colour\n                                    Gold\n
colour\n                                    Green\n
colour\n                                    Grey\n                           .
colour\n                                    Orange\n
colour\n                                    Other Color\n
colour\n                                    Red\n
colour\n                                    Silver\n
colour\n                                    White\n
warrantyNo
warrantyYes                                                                  *
bodycondNo accidents, very few faults
bodycondNormal wear & tear, a few issues
bodycondPerfect inside and out
bodycondSUV
bodyOther
bodySedan
bodySports
bodySUV
bodyTruck

Residual standard error: 0.08871 on 196 degrees of freedom
Multiple R-squared:  0.9155,    Adjusted R-squared:  0.9034
F-statistic: 75.84 on 28 and 196 DF,  p-value: < 2.2e-16
```

Figure 35:Structure of LR with log10(Price)

actuals <dbl>	predicteds <dbl>
4.602060	4.595772
4.897627	5.050251
4.781755	4.738075
4.880814	4.809574
4.845098	4.673654
4.602060	4.479918
4.591065	4.501472
4.301008	4.242464
4.724276	4.631783
4.777427	4.736939

```
[1] "Min_Max Train Accuracy"
[1] 0.9335942
[1] "Min_Max Test Accuracy"
[1] 0.9809354
[1] "MAPE"
[1] 0.01914043
[1] "MAE"
[1] 0.09107148
[1] "RMSE"
[1] 0.1034186
```

Figure 36:LR With normal split and log10(Price)

Random Split

```{r}
trains = sample(1:nrow(carDf),floor(nrow(carDf)*0.80))
tests = -trains

trainRandom=carDf[trains,]
dim(trainRandom)

testRandom=carDf[tests,]
dim((testRandom))
```

```
[1] 191  20
[1] 48 20
```

Figure 37:Random Splitting of data

actuals <dbl>	predicteds <dbl>
37000	41857.097
66666	43285.113
42000	41876.953
170000	99176.477
149000	100646.759
66800	76228.172
13999	36499.794
37000	39178.445
75000	47379.201
135000	96497.825

```
[1] "Min_Max Train Accuracy"
[1] 0.5696168
[1] "Min_Max Test Accuracy"
[1] 0.7057327
[1] "MAPE"
[1] 0.3882259
[1] "MAE"
[1] 16596.26
[1] "RMSE"
[1] 22217.29
```

Figure 38:LR with Random split and Price

The values of the random split are very less compared to a normal split. The logarithmic value of the price is added to the variable pricelog.

```
Call:
lm(formula = priceslog ~ mileages + warranty + trim, data = trainRandom)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.75259 -0.07877  0.02667  0.13311  0.38485 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.860e+00  1.383e-01  35.135  < 2e-16 ***
mileages    -2.582e-06  2.360e-07 -10.938  < 2e-16 ***
warrantyNo  -1.269e-02  3.785e-02  -0.335    0.738    
warrantyYes  1.581e-01  3.949e-02   4.003 9.01e-05 ***
trim         1.711e-03  2.837e-02   0.060    0.952    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1919 on 186 degrees of freedom
Multiple R-squared:  0.5744,	Adjusted R-squared:  0.5653 
F-statistic: 62.76 on 4 and 186 DF,  p-value: < 2.2e-16
```

s

actuals <dbl>	predicteds <dbl>
4.819544	4.785683
5.173186	5.024641
4.568202	4.561284
4.579784	4.541379
4.875061	4.631217
4.574031	4.663433
4.602049	4.725399
4.681241	4.549407
4.681241	4.780191
4.897627	4.644126

```
[1] "Min_Max Train Accuracy"
[1] 0.9442847
[1] "Min_Max Test Accuracy"
[1] 0.9806085
[1] "MAPE"
[1] 0.01967497
[1] "MAE"
[1] 0.09226795
[1] "RMSE"
[1] 0.1278523
```

Figure 39:LR with Random Split and log10(Prices)

The performance of the random split is slightly better than the normal split. The cross-validation result is shown below, and the cross-validation results are not that promising.

```
[1] "MAE"
[1] 0.195369
[1] "RMSE"
[1] 0.2579713
```

Figure 40:LR with CV

4.2.2 Gradient Boosting:

Gradient Boosting is the other model used for prediction.GB is the easiest and fast learning model. Let us look at the results of the model.

	var <chr>	rel.inf <dbl>
year	year	54.87077627
mileages	mileages	36.18520766
warranty	warranty	4.94918688
specs	specs	2.74171429
trim	trim	1.04630284
bodycond	bodycond	0.18866562
body	body	0.01814644
mechcond	mechcond	0.00000000
fuel	fuel	0.00000000

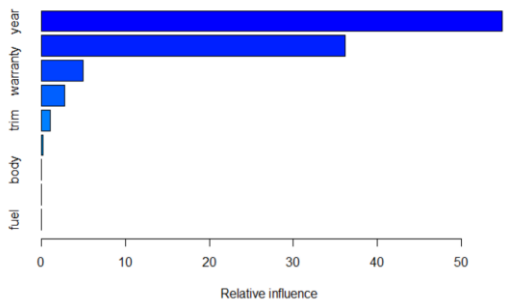


Figure 41:Relative Influence Graph

The relative influence index in the gbm model fitting summary gives the variable influence of the data to the dependent variable in the gbm model fitting. The model is run for all the 3 data splits. The normal fit does not perform as expected. The min-max accuracy of train and test had a large variation. The error rate for the gbm with the normal split is very less compared to linear regression. In the Relative influence graph, the variable with the least influence and 0 value can be removed to see how the accuracy and error value gets affected for the normal value of the price.

actuals <dbl>	predicteds <dbl>
48000	41418.853
86000	91766.685
39500	187005.513
24000	20114.740
12500	35249.506
76000	55461.609
35000	32068.071
59000	57040.819
34000	74967.479
65000	99463.235

```
[1] "Min_Max Train Accuracy"
[1] 0.4920667
[1] "Min_Max Test Accuracy"
[1] 0.6939355
[1] "MAPE"
[1] 0.4849163
[1] "MAE"
[1] 31330.27
[1] "RMSE"
[1] 67463.96
```

Figure 42:GBM Normal Split

The value with the least relative index was removed, and the other variables were used for the valuation. The accuracy and error rates were better compared to the previous model. Let us now look at how the performance will be when the value of the price is standardized to the log value. The log valued performed very well with high accuracy of 0.999. The variation from the training data is also no that large. The high accuracy can be overfitting, but the MAPE, MAE, and RMSE values are very least show that the model performance is good.

actuals <dbl>	predicteds <dbl>
55000	54964.53
10000	9367.27
12500	12657.72
68500	67736.13
19000	18821.99
40845	41038.95
39000	40522.01
11500	11594.56
33000	32898.70
89990	87990.14

```
[1] "Min_Max Train Accuracy"
[1] 0.511236
[1] "Min_Max Test Accuracy"
[1] 0.9888082
[1] "MAPE"
[1] 0.01135177
[1] "MAE"
[1] 478.7291
[1] "RMSE"
[1] 856.3509
```

Figure 43:GBM Normal split with relative influenced variable

actuals <dbl>	predicteds <dbl>
4.740363	4.742821
4.000000	3.991617
4.096910	4.099071
4.835691	4.829370
4.278754	4.276978
4.611139	4.612946
4.591065	4.602656
4.060698	4.063198
4.518514	4.515290
4.954194	4.948750

```
[1] "Min_Max Train Accuracy"
[1] 0.9259743
[1] "Min_Max Test Accuracy"
[1] 0.9990221
[1] "MAPE"
[1] 0.0009786184
[1] "MAE"
[1] 0.004492721
[1] "RMSE"
[1] 0.005798623
```

Figure 44:GBM with log10(Price) and relative influenced values

The Raandom split value's performance sheet with least variables also performs well, but the error rate compared to the normal split is higher. Hence, the gbm with a normal split of log value of price gives the best result.

```
[1] "Min_Max Train Accuracy"
[1] 0.9337829
[1] "Min_Max Test Accuracy"
[1] 0.9909684
[1] "MAPE"
[1] 0.009471482
[1] "MAE"
[1] 0.04142401
[1] "RMSE"
[1] 0.0956043
```

Figure 45:GBM with Random Split

Let us select the optimum number of the tree using the perf() of gbm and another plot to indicate the tree's optimum number on this technique.

```
[1] "Optimum Tree in oob"
[1] 361
attr(,"smoother")
Call:
loess(formula = object$oobag.improve ~ x, enp.target = min(max(4,
    length(x)/10), 50))

Number of Observations: 2000
Equivalent Number of Parameters: 39.99
Residual Standard Error: 1.294e-05
[1] "Optimum Tree in cv"
[1] 1982
```

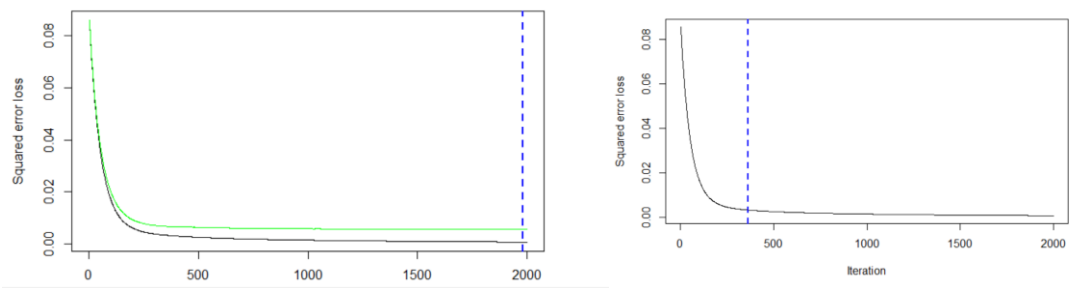


Figure 46:Plotting Optimum Tree graph

The graph indicates the test's error marked in green line and the error in train marked in black line. The blue line indicates the optimum number of iterations. Beyond that value, the train test accuracy may increase; hence the training stops at the optimum number. The result with optimum value gave a good result and slightly more train accuracy than normal split and less deviation from train and test accuracy.

```
[1] "Min_Max Train Accuracy"
[1] 0.9261077
[1] "Min_Max Test Accuracy"
[1] 0.997973
[1] "MAPE"
[1] 0.002030889
[1] "MAE"
[1] 0.009505661
[1] "RMSE"
[1] 0.01301764
```

Figure 47:GBM with optimum tree value result

4.2.3 RANDOM FOREST

The random forest is yet another model that was used for the valuation of the car. The random forest result depends on the variable and creates several decision trees to fit the model. The data split is similar to another model with another performance-related modeling.

The random forest model started modeling with a normal value of the price. Like other models, the random forest also outperformed.It has a high value of error rate. Where as the modeling was done on the logarithmic value performed well and produced a result of good accuracy and less error rate. The result of random forest compared to gbm, the random forest did not perform well. Similary, the modeling was done the random splitting with the log value of the price. The model performed well. The result of accuracy was slightly higher than the normal split with the least error.

actuals <dbl>	predicteds <dbl>
13000	33433.03
32000	32596.92
45000	29024.42
17000	38021.64
65000	112233.85
68000	86606.16
68000	86606.16
12000	40030.97
115000	76850.22
42500	43627.22

```
[1] "Min_Max Train Accuracy"
[1] 0.5747172
[1] "Min_Max Test Accuracy"
[1] 0.7066647
[1] "MAPE"
[1] 0.5197348
[1] "MAE"
[1] 24590.38
[1] "RMSE"
[1] 50502.54
```

Figure 48:Random Forest Normal Split

actuals <dbl>	predicteds <dbl>
4.113943	4.454551
4.505150	4.406021
4.653213	4.416418
4.230449	4.527605
4.812913	4.895722
4.832509	4.792671
4.832509	4.792671
4.079181	4.592998
5.060698	4.865204
4.628389	4.623057

```
[1] "Min_Max Train Accuracy"
[1] 0.9418089
[1] "Min_Max Test Accuracy"
[1] 0.9648828
[1] "MAPE"
[1] 0.03643048
[1] "MAE"
[1] 0.167871
[1] "RMSE"
[1] 0.2356366
```

Figure 49:Random Forest with log10(Price)

```
[1] "Min_Max Train Accuracy"
[1] 0.9403843
[1] "Min_Max Test Accuracy"
[1] 0.9727599
[1] "MAPE"
[1] 0.02807019
[1] "MAE"
[1] 0.1288174
[1] "RMSE"
[1] 0.1670161
```

Figure 50:Random Forest with Random Split

Let us compare the models with the evaluation matrix of min-max accuracy of train and test data, Mean average percentage error(MAPE), Mean Average Error(MAE), Root Mean Square Error(RMSE).

Linear Regression -LR,	Gradient Boosting-GBM,	Random Forest-RF,
Cross-Validation -CV		

Split	Model	Price	Train Accuracy	Test Accuracy	MAPE	MAE	RMSE
Normal	LR	Normal	.53	0.81	0.19	12388.27	15500.25
	LR	Log10	0.934	0.980	0.191	0.0910	0.1034
	GBM	Normal	0.4920	0.6939	0.484	31330.27	67463.96
	GBM	Value basedin relative influence	0.511	0.989	0.0113	478.7291	856.3509
	GBM	Log10	0.9259	0.99902	0.00097	0.004492	0.005798
	GBM	Log10 with optimum tree	0.9261	0.997973	0.002	0.01	0.013
	RF	Normal	0.5747	0.766647	0.5197348	24590.38	50502.54
	RF	Log10	0.9418	0.96488	0.03643	0.167871	0.23563
RANDOM	LR	Normal	0.57	0.706	0.389	16596.26	22217.29
	LR	Log10	0.9442	0.9806	0.1967	0.0922	0.1278
	GBM	Log 10	0.9337	0.9909	0.009471	0.041424	0.956043
	RF	Log10	0.94038	0.9727	0.02807	0.1288174	0.1670161
CV	LR					0.195	0.2578

Figure 51:Comparison of Models

The three models performed better than others; let us Hyperparameter these models to get the best value for the parameters.

4.3 HYPER PARAMETER TUNING:

4.3.1 Random Forest:

Let us compare the tuning results of the random forest with normal price value and log10 Price Value.

```
Call:
 randomForest(formula = prices ~ mileages + colour + specs + warranty +      body +
bodycond + fuel + mechcond + trim, data = traindatas,      importance = TRUE, ntree =
500)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 5602010422
                    % Var explained: 3.13
             %IncMSE IncNodePurity
mileages  3.9176544393  373716416490
colour   -4.6906524144   71368129584
specs     1.1382309415   25258559476
warranty  6.3613907143   95386068286
body     -0.0008832017    7866527671
bodycond -1.6190277628    4252544261
fuel      0.0000000000      66391958
mechcond -0.7825561236    1831984518
trim     -3.7534679440   22888978127
```

```
Random Forest

200 samples
  9 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 180, 180, 180, 180, 181, 180, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
   1    50425.06  0.4923322  30562.05
   3    45255.51  0.4805479  25007.79
   4    45566.08  0.4753444  24317.85
   5    46462.62  0.4769606  23878.79
   7    48231.33  0.4719294  23602.36
   8    48907.62  0.4815925  23422.03
  12    51853.78  0.4855477  23425.94
  15    52128.83  0.4842549  23174.81
  17    53359.53  0.4832446  23396.10
  18    52765.06  0.4961983  23154.35
  23    56323.93  0.4931496  23842.85
  27    58450.94  0.5090338  24105.51

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 3.
```

Figure 52:Random Forest with Normal split Hyperparameter Tuning

The log10 value of the price gave a better output. The normal value of price gave an optimal model at mtry three, where the RMSE value is 45255.51, R2 value of 0.4805479, and MAE is 25007.79.The log10 value of price gave an optimum value with mtry 14 with RMSE = 0.1843,R2 value = 0.6336,MAE=0.1291.The RF hyperparameter tuned model does not give that good value.

```
Call:
 randomForest(formula = priceslog ~ mileages + colour + specs +      warranty + body +
bodycond + fuel + mechcond + trim, data = traindatas,      importance = TRUE, ntree =
500)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 0.03909896
                    % Var explained: 57.37
           %IncMSE IncNodePurity
mileages 43.6920904   7.842366231
colour   -1.5619242   1.351962976
specs    25.9489507   1.619640190
warranty 17.2301126   2.740016943
body      0.7991976   0.186637798
bodycond  3.8390222   0.631769583
fuel      0.0000000   0.007689107
mechcond -4.1010318   0.365915467
trim     -1.8615820   0.336745710
```

```
Random Forest

200 samples
  9 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 180, 181, 180, 181, 179, 179, ...
Resampling results across tuning parameters:

  mtry  RMSE       Rsquared   MAE
   1    0.2687362  0.4939273  0.2003478
   3    0.2140866  0.5798714  0.1607306
   7    0.1903310  0.6192795  0.1393535
   8    0.1886192  0.6228890  0.1365654
  12    0.1852784  0.6296910  0.1304783
  13    0.1852305  0.6308315  0.1299616
  14    0.1843724  0.6336012  0.1291067
  15    0.1847978  0.6328041  0.1287766
  16    0.1850279  0.6312092  0.1289108
  18    0.1850902  0.6314032  0.1288778
  21    0.1849116  0.6329392  0.1287075
  23    0.1858481  0.6305722  0.1287591
  26    0.1853965  0.6318078  0.1286789

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 14.
```

Figure 53:Random Forest with Log split Hyperparameter Tuning

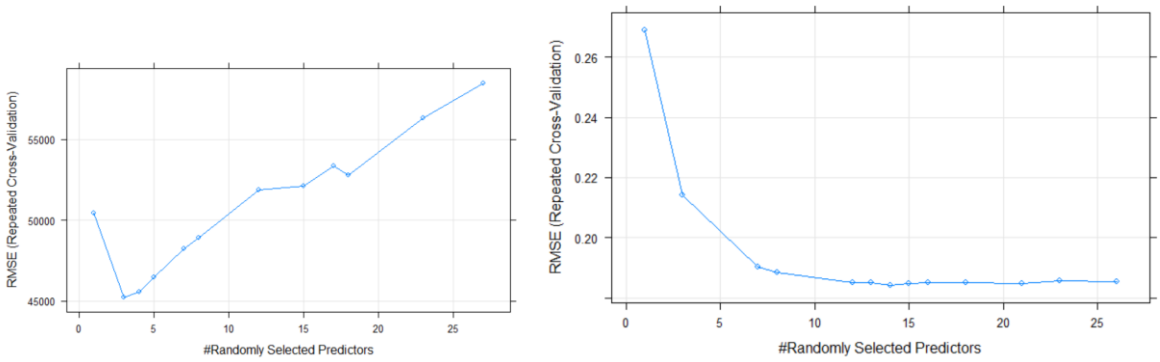


Figure 54:Random Forest with Normal split and Log split Hyperparameter Tuning Graph

4.3.2 Gradient Boosting Model:

Gradient Boosting hyperparameter tuning was done using trainControl, which is fed to train function with a tunelength of 50 on the cv method. The result of the train function.

```
Stochastic Gradient Boosting

275 samples
  4 predictor

No pre-processing
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 183, 184, 183
Resampling results across tuning parameters:

  interaction.depth  n.trees  RMSE       Rsquared   MAE
  1                   50      0.08913966 0.9157788  0.06769140
  1                  100      0.08914724 0.9158902  0.06750856
  1                  150      0.08927434 0.9172742  0.06744010
  1                  200      0.08989837 0.9164524  0.06785243
  1                  250      0.08789882 0.9202478  0.06558191
  1                  300      0.08770760 0.9205651  0.06523465
  1                  350      0.08703323 0.9215609  0.06434749
  1                  400      0.08581306 0.9233150  0.06274598
  1                  450      0.08499229 0.9247906  0.06156022
  1                  500      0.08444044 0.9262304  0.06082869
  1                  550      0.08470576 0.9257302  0.06056166
  1                  600      0.08350994 0.9271652  0.05921835
  1                  650      0.08288992 0.9279880  0.05811288
  1                  700      0.08337316 0.9274405  0.05826268
  1                  750      0.08210673 0.9293794  0.05666824
  1                  800      0.08202259 0.9296998  0.05649678
  1                  850      0.08290737 0.9288998  0.05650724
  1                  900      0.08097480 0.9310919  0.05473918
  1                  950      0.08154758 0.9308966  0.05480989
  29                  50      0.08232904 0.9281511  0.06214459
  29                 100      0.07738931 0.9361201  0.05463183
  29                 150      0.07479716 0.9398655  0.05038558
  29                 200      0.07442463 0.9401524  0.04818253
  29                 250      0.07478984 0.9397408  0.04626297
  29                 300      0.07514792 0.9388874  0.04560606
  29                 350      0.07543276 0.9379835  0.04473887
  29                 400      0.07603827 0.9372542  0.04443392
  29                 450      0.07642211 0.9363641  0.04421064
  29                 500      0.07682313 0.9355275  0.04370435
  29                 550      0.07738136 0.9343664  0.04356734
  29                 600      0.07854266 0.9325522  0.04413757
  29                 650      0.07908693 0.9313569  0.04420516
  29                 700      0.07961490 0.9303823  0.04435433
  29                 750      0.07937487 0.9304958  0.04403240
  29                 800      0.08011224 0.9291039  0.04413157
  29                 850      0.08012178 0.9290111  0.04401664
  29                 900      0.08035279 0.9286237  0.04377487
  29                 950      0.08081199 0.9277769  0.04382186
  29                1000      0.08098251 0.9274179  0.04390962
  29                1050      0.08100049 0.9271974  0.04383621
  29                1100      0.08126355 0.9267915  0.04346477
  29                1150      0.08145021 0.9264473  0.04342148
  29                1200      0.08174232 0.9260318  0.04348396
  29                1250      0.08163172 0.9261631  0.04330876
```

29	1150	0.08145021	0.9264473	0.04342148
29	1200	0.08174232	0.9260318	0.04348396
29	1250	0.08163172	0.9261631	0.04330876
29	1300	0.08157404	0.9262301	0.04369312
29	1350	0.08188791	0.9256872	0.04348077
29	1400	0.08190245	0.9256013	0.04350354
29	1450	0.08220754	0.9250652	0.04348428
29	1500	0.08212706	0.9251667	0.04345440
29	1550	0.08222228	0.9249463	0.04347113
29	1600	0.08235253	0.9246746	0.04364481
29	1650	0.08237855	0.9247165	0.04345619
29	1700	0.08263057	0.9242840	0.04362027
29	1750	0.08272491	0.9240522	0.04370916
29	1800	0.08280253	0.9237811	0.04365689
29	1850	0.08306700	0.9233676	0.04374623
29	1900	0.08307920	0.9233594	0.04379812
29	1950	0.08284644	0.9235559	0.04365313
29	2000	0.08313798	0.9231980	0.04379788
29	2050	0.08342575	0.9227453	0.04383133
29	2100	0.08348456	0.9225482	0.04383061
29	2150	0.08331799	0.9228780	0.04383220
29	2200	0.08331587	0.9229654	0.04382944
29	2250	0.08330445	0.9229313	0.04368784
29	2300	0.08352648	0.9224179	0.04393121
29	2350	0.08342065	0.9226136	0.04388618
29	2400	0.08356291	0.9224298	0.04375652
29	2450	0.08354872	0.9225079	0.04378610
29	2500	0.08352898	0.9224192	0.04383574
30	50	0.08244503	0.9270961	0.06099731
30	100	0.07863034	0.9343082	0.05513639
30	150	0.07579251	0.9381966	0.05024753

49	1150	0.08015558	0.9295694	0.04319599
49	1200	0.08026621	0.9292960	0.04316925
49	1250	0.08059335	0.9286290	0.04325858
49	1300	0.08076192	0.9283112	0.04311760
49	1350	0.08125245	0.9274766	0.04334731
49	1400	0.08115697	0.9277743	0.04319860
49	1450	0.08134834	0.9273350	0.04315506
49	1500	0.08136124	0.9273693	0.04317906
49	1550	0.08170480	0.9268228	0.04331911
49	1600	0.08178756	0.9265266	0.04315866
49	1650	0.08176898	0.9265731	0.04305515
49	1700	0.08169434	0.9265964	0.04317500
49	1750	0.08204999	0.9259873	0.04324040
49	1800	0.08206227	0.9260899	0.04310558
49	1850	0.08220177	0.9257434	0.04315409
49	1900	0.08211091	0.9258682	0.04303361
49	1950	0.08243187	0.9252606	0.04322599
49	2000	0.08251099	0.9250935	0.04332664
49	2050	0.08248954	0.9251102	0.04327669
49	2100	0.08252564	0.9250501	0.04334685
49	2150	0.08249765	0.9252459	0.04314656
49	2200	0.08259384	0.9249931	0.04316244
49	2250	0.08284473	0.9245366	0.04336907
49	2300	0.08252410	0.9249961	0.04307602
49	2350	0.08270938	0.9246923	0.04339809
49	2400	0.08271713	0.9246170	0.04327747
49	2450	0.08302049	0.9241040	0.04344503
49	2500	0.08276971	0.9244581	0.04331250

```
Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning
 parameter 'n.minobsinnode' was held constant at a value of 10
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were n.trees = 200, interaction.depth = 29,
 shrinkage = 0.1 and n.minobsinnode = 10.
```

Figure 55:Hyper Parameter Tuning of GBM

The model obtains the best fitting at interaction.depth = 29, shrinkage =0.1, n.trees=200 with the RMSE = 0.07442, R2 Accuracy of 0.94015, MAE =0.0481825

actuals <dbl>	predicteds <dbl>
4.740363	4.740091
4.000000	3.988145
4.096910	4.099708
4.835691	4.831186
4.278754	4.275571
4.611139	4.614586
4.591065	4.604342
4.060698	4.062392
4.518514	4.515535
4.954194	4.947599

```
[1] "Min_Max Train Accuracy"
[1] 0.9259443
[1] "Min_Max Test Accuracy"
[1] 0.9987715
[1] "MAPE"
[1] 0.001229666
[1] "MAE"
[1] 0.005658223
[1] "RMSE"
[1] 0.007556397
```

Figure 56:Application of Hyper Parameter Tuned Result

The result obtained from the train() is fed to the gbm model and saw the accuracy and accuracy of good values of accuracy and the least error.

Comparing all the models, we can see that gbm is the best model fit for the final product. The gbm model will be given to the final product with parameter values as shown

interaction.depth = 29	shrinkage =0.1	n.trees=200

4.4 RESULT OBTAINED FROM RAPIDMINER:

Name	Correlation	ID-ness	Stabillity
year	2.55%	6.94%	16.33%
mileage	11.32%	76.73%	2.45%
trim	2.77%	1.22%	57.14%

Figure 57: Variable Statistics

The data ran on the rapid miner shows the major three significant variables that we ontained from plots and modelling results.

The gbm ensemble tree produced by the performance in the rapid miner is shown below.

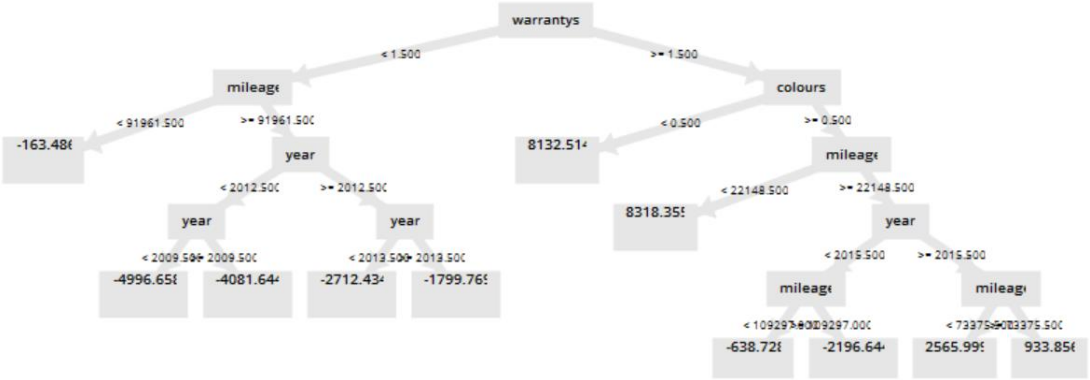


Figure 58:Ensemble tree in GBM

The data was modeled on others to see the performance. The SVM results prove that it does not provide a good result as expected.

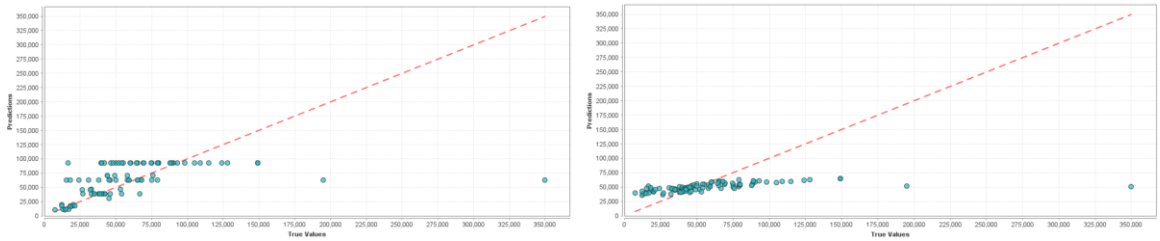


Figure 59:(a)SVM (b)Decision Tree

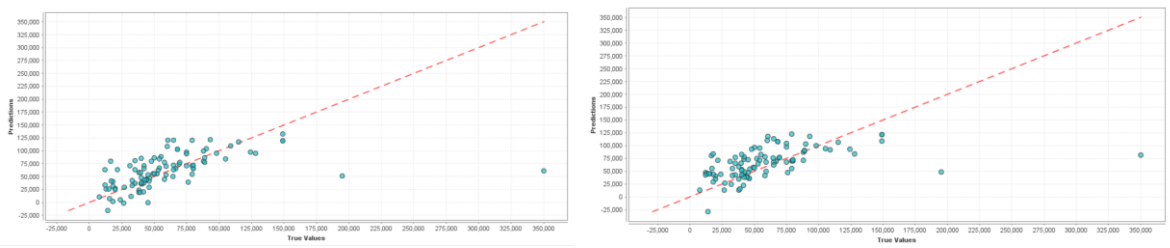


Figure 60:Deep Learning and Linear Regression

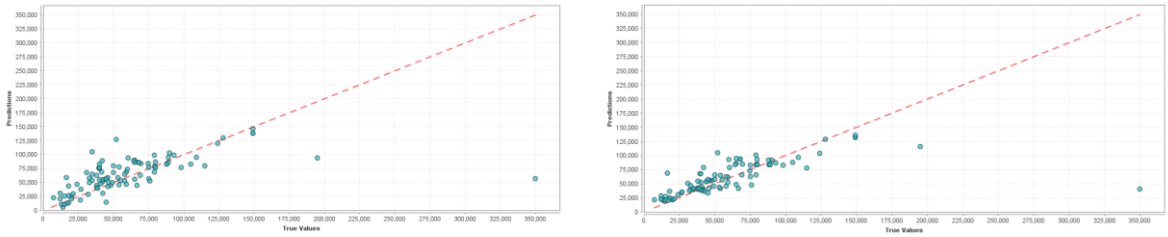


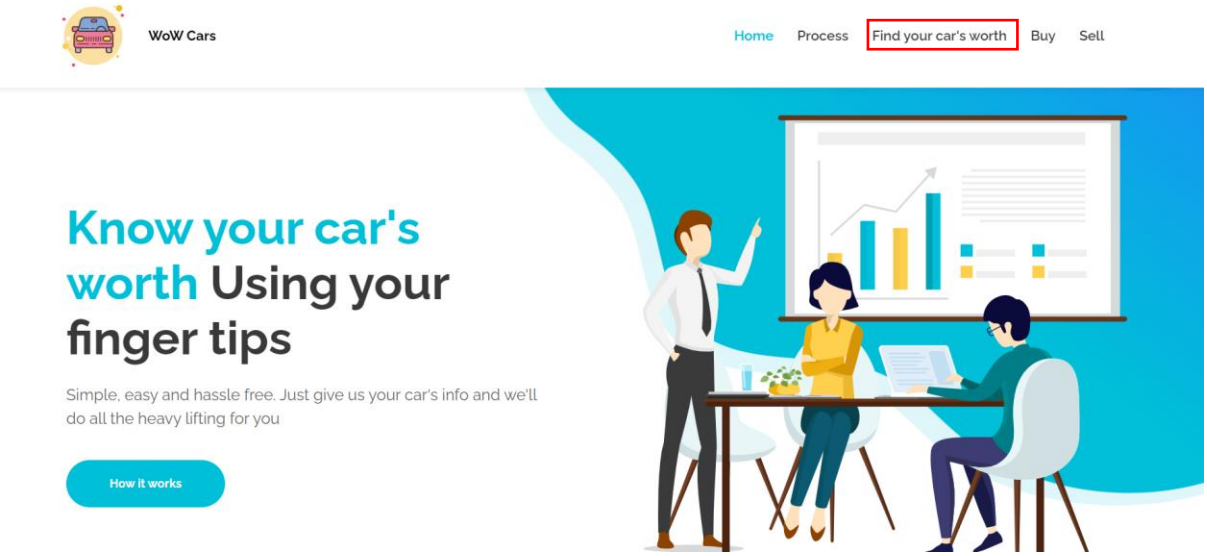
Figure 61:GBM and Random Forest

Model	Relative Error	Standard Deviation	Total Time	Training Time (1,000 Rows)	Scoring Time (1,000 Rows)
Generalized Linear Model	0.3	0.0	2232.0	522.4	132.7
Deep Learning	0.3	0.0	2290.0	1689.8	122.4
Decision Tree	0.2	0.1	377.0	24.5	61.2
Random Forest	0.2	0.0	9044.0	4604.1	154030.6
Gradient Boosted Trees	0.2	0.0	18271.0	273.5	1000.0
Support Vector Machine	0.3	0.0	1175.0	57.1	71.4

Figure 62:Comparison of Result based on RapidMiner

On comparing the model results in the rapid miner with the proposed research model, the RF and GBM perform the best. RF has the least variation from the regression line but the training time taken for the RF is 20 times higher than gbm. Similarly, the Deep learning model also performed well compared to other models, but the relative error is higher than other models.

4.5 FINAL PRODUCT



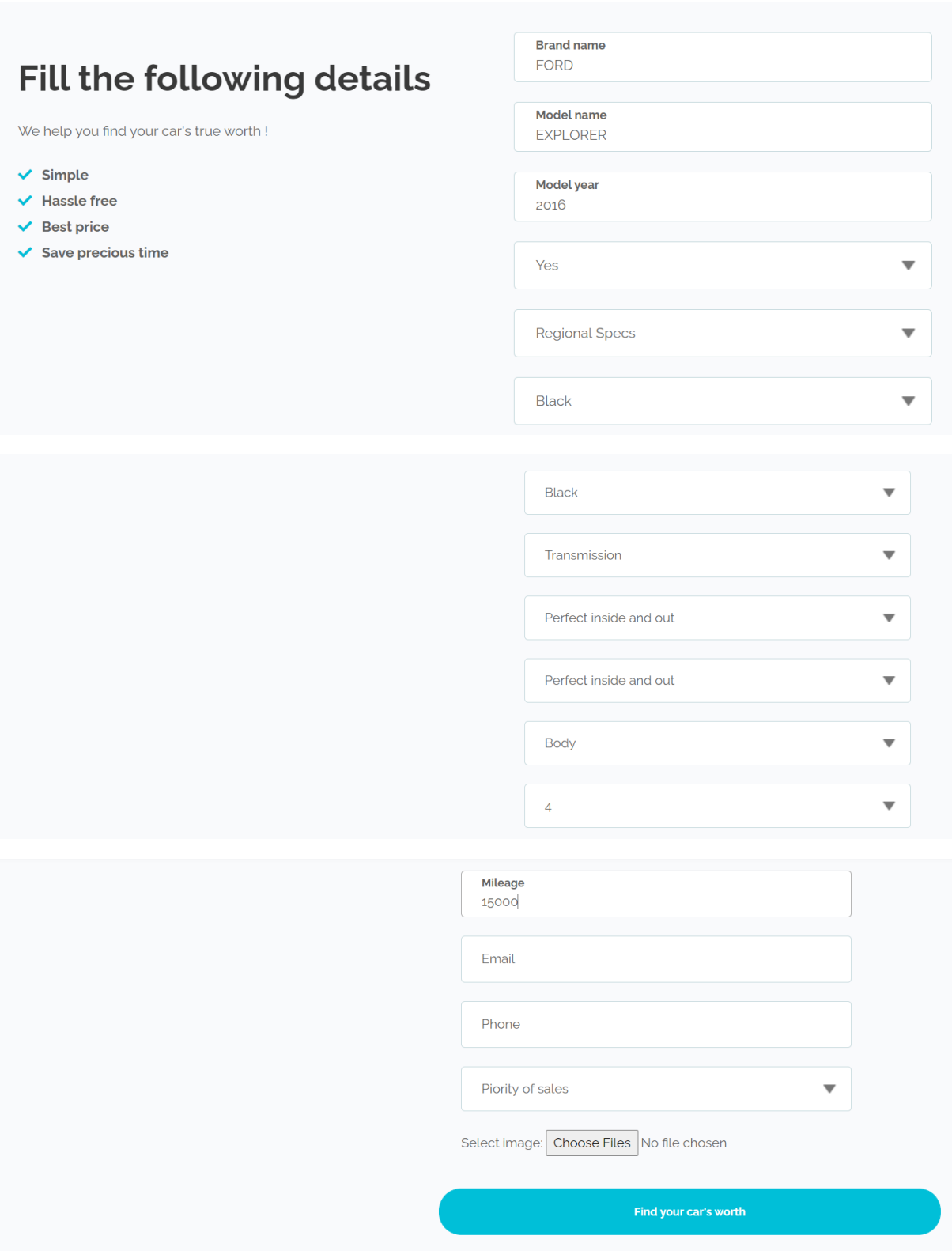


Figure 63:UI of the product

The UI has the request the and once the find your car's worth button is clicked. The API is started as shown in the below figure, and scraping starts followed by the prediction, and the result is sent back to the UI, as shown in the figures below.

```
> plumber::plumb("D:/MSc/Study/@5Project/@3Work/@3shiny/ProjectFinalAPI.R")$run(port = 8000)
Running plumber API at http://127.0.0.1:8000
Running swagger Docs at http://127.0.0.1:8000/__docs__/
[1] "Hello"
{html_document}
<html xmlns:fb="https://www.facebook.com/2008/fbml" lang="en" dir="ltr">
[1] <head>\n<meta charset="utf-8">\n<meta name="viewport" content="width=device-width,initial-scale=1.0">\n<title>Buy &amp; sell any car onlin ...
[2] <body id="classified-listing-body" style="\n      \n      \n    " class="\n      classified-listing-p\n      country-uae\n      site-dubai ...
 [1] "148,679" "120,000" "8,000"   "163,000" "118,665" "84,000"  "108,000" "93,000"  "128,000" "26,000"  "79,000"  "48,499"  "250,000" "111,602"
[15] "124,000" "24,905"  "144,118" "82,000"  "85,000"  "250,000" "75,378"  "78,000"  "65,000"  "152,430" "82,104"
[1] "mileage is"
 [1] "148679" "120000" "8000"   "163000" "118665" "84000"  "108000" "93000"  "128000" "26000"  "79000"  "48499"  "250000" "111602" "124000"
[16] "24905"  "144118" "82000"  "85000"  "250000" "75378"  "78000"  "65000"  "152430" "82104"
 [1] "59500"  "56000"  "145000" "25000"  "47500"  "38000"  "14500"  "68000"  "36000"  "120000" "46500"  "62000"  "10000"  "48000"  "71000"
[16] "128500" "34500"  "99950"  "79950"  "39950"  "55000"  "52000"  "55000"  "43500"  "64800"
 [1] "White"  "Red"    "Silver" "White"  "Brown"  "White"  "White"  "White"  "Grey"   "Blue"   "White"  "Silver" "Grey"   "White"  "Black"
[16] "Black"  "Black"  "Silver" "White"  "Brown"  "White"  "Silver" "Black"  "White"  "White"
 [1] "4" "5" "4" "5" "5" "4" "5" "4" "5" "5" "5" "4" "4" "5" "5" "4" "4" "5" "5" "5" "5" "5" "4" "4" "4"
{html_document}
<html xmlns:fb="https://www.facebook.com/2008/fbml" lang="en" dir="ltr">
[1] <head>\n<meta charset="utf-8">\n<meta name="viewport" content="width=device-width,initial-scale=1.0">\n<title>Buy &amp; sell any car onlin ...
[2] <body id="classified-listing-body" style="\n      \n      \n    " class="\n      classified-listing-p\n      country-uae\n      site-dubai ...
 [1] "91,234"  "100,000" "150,000" "93,252"  "59,691"  "98,928"  "50,139"  "237,000" "130,000" "34,000"  "148,000" "90,845"  "150,000" "92,000"
[15] "121,000" "35,000"  "17,000"  "235,000" "35,000"  "96,000"  "162,000" "80,000"  "56,000"  "106,000" "62,000"
[1] "mileage is"
 [1] "91234"  "100000" "150000" "93252"  "59691"  "98928"  "50139"  "237000" "130000" "34000"  "148000" "90845"  "150000" "92000"  "121000"
[16] "35000"  "17000"  "235000" "35000"  "96000"  "162000" "80000"  "56000"  "106000" "62000"
 [1] "58500"  "49500"  "46000"  "58500"  "75990"  "89995"  "79995"  "11500"  "19000"  "92500"  "46000"  "74000"  "47500"  "42000"  "47000"
[16] "78000"  "145000" "23000"  "89000"  "13500"  "42500"  "20000"  "88000"  "36000"  "62000"
```

Figure 64:The UI send data request in backend and scraping

```
[1] "The Specification of Car for which Price is to be Predicted"
  year mileage trim colour
1 2016   15000    4  Black
[1] "Predicted Value of Car is "
[1] 102984
```

Figure 65:Output from the prediction

Figure 66:Final Output

5: DISCUSSIONS

Proposed research depends on the output obtained from different models—the research of car value done modeling with three algorithms.The research was based on real-time data. The evaluation of the classification tasks can be done using multiple evaluation factors such as recall, sensitivity,f1 score. The regression evaluation can be done on RMSE, MAE, MAPE, and Accuracy score. The proposed research depending factor is the price; hence we use supervised learning. The best model selection done by modeling three algorithms was made a data scraped out of a brand. The model used is Linear Regression, Gradient Boost Model, and Random Forest supervised learning. The dependent variable of the data is the price, which is a continuous data. Hence the evaluation factors that can be used are RMSE, Min-Max Accuracy, MAE & MAPE. The result obtained from algorithms and its comparisons have been shown in Figure 51

5.1 DATA

The result obtained from different models says that the prediction accuracy varies in different ranges. The data we obtained through scraping is very dynamic that price range and mileage may not be of a fixed range. It can vary from 0 to any value. There can be possible chances that data entered in the site to be inaccurate by the customer. There can be chances to bid the higher prices knowingly or unknowingly. Hence it can highly affect the data modeling. The data consist of 1 dependent 13 independent variables. The multiple correlation graphs and models are shown in Figure 27 show that the highest correlated variables are mileage ,year, warranty, and color. If we look at each variable one by one, let us see how the variables' relation is dependent on the dependent variable.

Mileage - Figure 22 shows that the mileage is inversely related to price. As the car's mileage increases, the value of car-driven increases, and the price decreases.

Year - Year is one of the other factors that is highly affecting the value of the car. It is also inversely related to the price. It implies that as the car gets older, the price gets lowered.

Colour - Colour is a factor depending on the price, but it is not relative to the year and mileage. The relation of color over price is shown in Figure 19 & 20. It is clear that certain color is more in markets, and similarly, those colors that are least in the market has more price.

Warranty - Warranty relation with price can be seen it the Gini index impurity graph of the RF Model. It means that the probability of the impurity in the variable. The warranty has the least impurity and more variable importance. Those cars that come in the market with a warranty will cost a higher price than those that are not.

Trim & body- Trim value, which tells about the number of doors and body, has less influence than the factors mentioned above.

bodycond & mechcond - The body condition and mech condition have less influence on price than the other factors. They are based on the market price. If a customer applies to get the car for lower prices, the heuristic thinking that profit on the price can replace the body and mech condition issue as per the buyer's wish. Hence they have the least relation with the price.

Seller_type, Cylinder, Specs-These variables influence the price and are shown in Figures 19 & 20, but in this use case, since the data contains the least number of factor levels, they cannot be used to train the model. Hence they are the least significant.

Transmission & fuel – Transmission & fuel are a variable affecting the sale, but in the particular use case, the data set is related to a particular brand hence the variation of transmission would be less due to which it is not used in the modeling.

Value of Price:

The modeling was done over two types of value of prices. Normal Value in the data extracted and log10 value of the price. It was done as the value of the price was distributed along with the range. The performance of the log value price was more evident than the normal value. As the price's value is applied with log10, the variable range decreases, and values become close. The variation of the value in the prediction gets lowered, hence the error rate.

Data Platform:

The proposed research was done with the live data scraped from the dubbizle.com. The research is highly dependent on data that is being scraped; hence, it is very important to get the data from a reliable source with full data content. The dubizzle.com website is a good source and a market player that can be relied on. Multiple other platforms are available for the scraping. It is important is to understand how the data source platform can change the data available for scraping, which is further used in the modeling.

The different platform provides different data available. Certain data that are available from one source may not be available in another. The basic data for sales may be available in all. We can see that warranty and door were one of the factors that were shown to be influencing. Compared with other data sources such as sellanycar.com and yallamotors, these details were not available; hence, the data prediction is affected by it. There can be chances that prediction may vary highly. It is important that scraping and data training are to be changed widely to apply the data source. A few variables can also be added, which can create value for the data. Hence the data source is one important factor.

Advantage:
The advantage of using log value is that error rates get decreased, and accuracy increases.
Disadvantage:
The reverse conversion leads to the normal value, and prediction accuracy decreases.

5.2 MODELING

The modeling of this research is done on three supervised learning algorithms.

5.2.1 Linear Regression

The linear regression is the easiest model for training and least tun time. The variables such as mileages, year, color, warranty, bodycond, mechcond, warranty,bodycond, body, mechcond, trim are used. Only these variables are considered for the modeling as other variables are least significant to the modeling and did not have multiple levels for training. The significant variables are mileages, year, body, and bodycond. The F-statistic in Figure 32 is to be as low as possible from 0. The p-value tells about the relation of the model with the hypothesis. The 0.02 displays that there is a small evident relation with the other variables. However, it is no that closely related to the variable relations that are not high.

The modeling was done on both the value of prices as mentioned and studied the result. Figure 33 shows the linearity of the model prediction after the price was converted to the log values. The log value was applied to the mileages, but the mileage variation was not that promising similar to price; hence, we keep the normal value for the prediction. We can see that the price's normal value gave a bad performance with less accuracy rate and higher error rates in Figure 34. The value in the actual and predicted column is an evident result variation of the result.

The values of the logged value prediction show that the model is a good fit for the prediction. The MAPE, MAE, and RMSE are less as the price's log value has the least variation, but there is a slight variation in train accuracy and test accuracy. This can be because of the chance of overfitting the model, but the possibility of linear regression to overfit is least.

The test result of the random split was done. It was done to see how random splitting effects modeling. The random splitting helps in the shuffling of the data. The result produced from the random split is less compared to the normal split. It has less accuracy and more error rate. The less result of the random split can be of the shuffling result where the variation in train data was higher compared to normal split. The cross-validation method also produced a result that was not much promising. The cross-validation method expects to perform well in all models as they are trying for all data combinations, but this user case outperformed as the variation was increasing in the different split. The normal split expects the data scraped of similar data, and variation is least.

5.2.2 Gradient Boosting Model:

The gradient boosting model is the model that is easy to implement and fast learning model. The Gradient Boosting Model. The relative influence graph (Figure 41) shows the more relative variables to the dependent variable used in the prediction. The graph shows that the year has the strongest influence on modeling. It proves the heuristic assumption of the car sale depending on the year, followed by the mileage. The modeling was done on the normal and random split with the original value and the price's log10 value. The normal split accuracy of the gbm was very less compared to the linear baseline model and higher error rate, but the variation of the train and test accuracy is less compared to the linear model. It proves the tend of this model to fit better than the linear model. The gbm model produced the evident good result with a log price. The test accuracy was around 0.9990. The value of accuracy can lead to thinking of the overfitting, and the variation of test and train accuracy variation also says a slight chance of overfitting, but the error rates of the modeling can be considered for thinking the model to be a good fit.

The modeling was on the random split similar to the linear regression with log values. The similarity was similar to the normal split value. The train and test accuracy variation were slightly less as the model tend to fit all the possible values. However, the error rate is slightly higher than the normal split, as it can be caused by the variation in data due to the shuffling.

The model's performance was analyzed using gbm.perf()to see the optimum number of the needed tree. It was done using both the cv method and OOB (out of the bag). These optimum values produced were 361 and 1982, respectively. The value helps to understand the variation of test train accuracy. The value is such that any value beyond produces a variation in test and train accuracy; hence the model training stops at that optimum number. The performance graph value of optimum cv value is given to find the model fitting with it and produce a more positive

result than other models. The test train variation was the least, and error rates were less compared to other models.

The rapid miner study on the gbm also proves to have a higher performance of gbm over other models. The log value selection creates and ambiguity on the model's selection though the value of performance was best in the log value of price, the use case request to reverse the value back to normal, and hence the normal value tends to have the variation of old accuracy. Though the variation is expected in the ambiguity, the log value is chosen to produce less variation than the normal value.

5.2.3 Random Forest:

The random forest model is another model that was used for our prediction. The random forest model build tree and its advantage of tree generation with the least error are very evident. Hence this model is highly dependent on the variables. The Figure 27 with the IncMSE and IncNodePurity prove the variable importance. The variable's Gini score is to be kept as low as possible to get a pure variable. The least value of Inc Node purity shows that the fuel, body cond, and mechcond are least significant to the modeling.

The random forest was done on a normal split with normal values of the price and log10 value of price and the random split on log10m value of the price. The Normal value gave an accuracy similar to gbm. The overlook of the result was done for the log10 price of the normal split model. The result was much promising, with the least train test variation. The random forest tends to build a tree with the best fir; hence, this model produced the least variation and training and test results. The error rates were least that, but we are not that promising compared to the gbm model. The random split was done on the log10 value prediction. This time prediction was very promising, with the least variation in test train accuracy and the least error rate compared to the normal split.

5.3 SUMMARISING COMPARISON OF MODELS:

- On comparison of all models with the accuracy and the error rates, as shown in Figure 51. When we compare the accuracy, we can see that all models ran with the original price's original value outperformed. They gave the least accuracy and higher error rates. It was because of the variation in the price. The price value varied in the higher range and hence giving the least accuracy and higher error rates.
- Those models with the log value of high price prediction as the price variation got reduced in the log value and decreased the error.
- The linear model performs least compared to other models as gbm, and the random forest is build to fit the most with trees.
- Random model performance was least compared to the Normal split as the data variation increases when the data get shuffled.
- The gbm and Rf model of log10 price performed well compared to all other models as the models tend to fit the maximum using decision trees and ensemble method hence producing better results.
- The gbm has less error rate, and RF has the least train and test accuracy variation. Their characteristic of the RF model tends to fit multiple decision trees and reduce overfitting, and the gbm ensemble method continuous and sequential work for error fitting.

- The chosen model shown in Figure 51 marked is the best, and it is tuned to get the best parameter producing accurate results.

5.4 HYPER PARAMETER TUNING

The hyperparameter tuning is done on the random forest model. With the hyperparameter tuning, we compare the normal value of price and log10 value of the price. The hyperparameter tuning was done on traincontrol() with repeatedcv and repeat. The value from the traincontrol is fed to the train(), producing results in multiple mtry to the best result produced. The same method was applied to the log10 values. The optimized result will be produced on the fitcontrol, but the results produced are not promising.

The tuning was done on GBM. The fitcontrol function was applied with the cv method of fold value -3. The output fed to the train produced the best result at the interaction.depth value=29 and tree =200.It produced good accuracy, and the result is fed to model to the best fit model. The gbm model produced the best accurate result with error compared to all.

The Gradient Boost model (price-log) at interaction.Depth = 29, shrinkage =0.1, n.trees=200 is chosen to be the final model as it produces the best result compared to all other.

5.4 OPTIMIZATION:

The optimization works to optimize the result, depending on the sales' user priority, but the optimization result was not obtained for the proposed research as the data have the least number of a continuous variable. The data had only mileage as the best continuous variable; hence, running an optimization for that regression equation depends on other categorical variables. So, this is not a good candidate for our linear optimization

5.5 RESEARCH PRODUCT:

The research product is fed with the mentioned gradient boosting model with the mentioned feature values. The product produces a dynamic value for each input given. The variations of each scraped data can vary depending on the input. The model was chosen based on the results of the selected data. It may outperform in certain cases. The UI's input is passed to the code's API to scrap and model and return the UI results.

6: CONCLUSION

The Research, Real-time Prediction of used car values are based on the user's input, and the prediction based on the data extracted depending on the user input. This research showed that predictive analytics is an excellent tool in delivering the prediction value for the problem statement—the research intended to predict the live data extraction. The data is the central part that plays a role in the prediction, positively influencing the outcome. The chapter summarizes the approach, Managerial and Theoretical Implications, Limitations of the Research, and The future applied to it. The research aimed to get multiple perspectives of the model and eventually compare the performance of the models.

The research aimed to Predict the value of the used car with real-time data. The research primarily focussed on modules such as data extraction, prediction. The research's predictive analytics was initiated with real data collection, which was done by the scraping method. The scraped data is trained with multiple models to see how different prediction works for the aim and how it results with the best accuracy and least error rate. This research helps to understand how a prediction works effectively with real-time data rather than pre-existing data recorded a while back.

The Prediction algorithms that have been used differed each other in the characteristics with the prediction power and accuracy. The result of a different model that has been used for the model selected given out a result that has a considerable variation in the result from good accuracy- the least error rate to feasible model accuracy and large error rate. The limitations in the model limited the research from bringing out a much better and accurate result. The scraped data has the least error rate as the data from the website was specific, and the pre-processing part was worked to find the empty content in the scraped data, and finally, the variable type was changed to numeric and factors accordingly. This final data was used for modeling.

6.1 SUMMARY OF THE THESIS MODULES:

6.1.1 Data Extraction:

Data extraction played a vital role in the research. The live data extraction depends on the user input, which has to scrape when the user enters his data. The data was scrapped based on the brand and model of the car entered by the user. The pre-existing data, which consists of the car and its details, will intend to have all types of a car with less data, which can critically affect the prediction. The research has shown that the data extraction part can get any data extracted depending on the user input, which is good enough for the prediction part to work, bringing out the best accurate result than training on a pre-existing data. The dependency of the data in the model is highly visible and marked this research. The module played an essential role in the intended research to achieve the best result to attain the aim.

6.1.2 Predictive Model Selection:

This module intended to select a model from multiple models depending on the user Input. The scraped data used for the modeling shows how machine learning is influencing the prediction. The model selection done after multiple model training on the data helped understand how each model is working on training data. Another result drawn from the multiple models for selection is how they positively influence the feature selection in understanding how certain variables are affected in the modeling. Hyperparameter tuning helped to find the model with high accuracy and the least error rate. The data and its type play an essential role in modeling.

Modeling based on tuning the parameter helped find a parameter that could bring out the least error rate.

6.1.3 Optimization:

The optimization was focusing on the sales priority, but as the data has the least number of continuous variables, running optimization is not feasible, and the result was not that meaningful.

6.2 MANAGERIAL IMPLICATION

The outcome of the research to take place is a managerial Implication of the research. The Car value is the price that gets dropped from the moment the car is bought. As each year passes on, the depreciation of the car gets continued. The features of the cars are very much affecting factors for the valuation of the car. Multiple levels of sectors can use this research. The proposed research aimed to help the users in predicting the value of their used car. The research mainly focuses on users such as customers to draw a market idea about the price for them to bid for the sale. Though it is intended mainly for the individual customer, the research is highly useful for Cooperate users who are in the automobile market for the sales and buy out of the used car. The automotive industry and user car sales are one of the growing sectors. The sales of the used cars have been increased from day today. The research product is the car's valuation, which plays a vital role in the sales market. It can also be used by banks, NBFCs, and even by the motor insurers.

The Dealers, Organizations, and web portals that deal with car sales can use this product for the valuation. This product plays a key to customer interaction for any organization and deals with them in clearing the car's value. This research product can also work with buying users who intend to a used car. Users can use this product for valuing the car that they are planning to buy. The car valuation of the used car helps the user in the hassle-free assessment of the car they will sell or buy. It would help the organization reduce the time for sales with proven value for the seller rather than procrastinate the sale on the means of sales value.

The growing market of used car sales calls for the importance of such products, which could improve the market growth and sales without much loss in the sales for both ends. The usage of the product would improve the credibility of the value that the dealer or organization bid. A negative impact raised in users' minds is valuing the car by a seller is instead a figure raised by them rather than based on the market value. It can indeed be overruled with the product's usage, which would make the customer understand how the market value can be applied to their car. Rather than an automotive industry, it is useful for the NBFCs and banks for valuing the car the approval of the loan amount for the user. It can also be used as the motor insurers, which would help them calculate the motor insurances' Premiums.

Research product plays a major role in all sectors and organizations that deal with the used car sales. The product helps calculate the loan amount to be approved, the Sale amount that would not incur a loss, and premium calculation of the insurance. If someone underprices the value, it highly affects sales. In the buyer case, if a motor insurer underprices the value, It will affect the customer when they are trying for the next sale for the car. It can cause a lower value claim payout in the policy, which would lead to a loss when any theft happens. Hence it very important to value the car where the product plays a significant role, which would help in an accurate value of the car depending on the Market value that would affect neither the customer

nor the seller. It could improve the credibility of the user organization's valuation of the car between the customers overwriting all the pre-existing negative impacts on users.

6.3 THEORETICAL IMPLICATIONS

All the research intends to bring out the best accuracy and least error rate prediction. The proposed research would bring out the best outcome for the users and the intended customers. The integration of multiple researches for a worthy product is highly appreciated for market benefit. It brings an outcome to improve the area of study and to improvise the scope of research. The theoretical implication of the research always lies in the understanding of the theory of the research. It involves the application of the research beyond any set of conditions that the research initially built.

The research key role in the day to day life is highly important. At the same phase, the intended research could be used by much other research in the work for the development of much other valuable work that could benefit the users in a much better and best way. The automotive industry is one of the major industries where several studies and computer vision projects have been carried for better outcomes and predictions. A broad set of projects are carried with the automotive industry, which could use this intended research, and to integrate it, bring out the best outcome for society.

The intended research value is based on the features of the cars. Some projects have been carried with image processing on the identification of the damages in the car. The researchers used an image data set to train the model and evaluate the user input car-based as the test to predict the car's damages. It would help identify the dents, scratches, repainting, or any kind of changes that were made. Other researches are carried out with any kind of loan prediction and insurance premium predictor.

The proposed research can be integrated with the mentioned researches in bringing out a much better model. The research can integrate to make project outcomes to get the user's data input, such as the brand model year and other features along with multiple image input of the car. The model will run the algorithm to identify the severity of the damages in the car. It then evaluates an index value of the damage severity and then uses that index value as a variable for the model that predicts the car's value. The integrated research would bring out an outcome that could be developed as a product that would be beneficial for multiple users and customers.

The proposed research is an outcome that could bring out a base project, which, in turn, integrated with many other modules research could develop research of greater scope giving an outcome to the society. Another possibility of integrating the proposed research includes the proposed research to predict Loan Value for a car, Prediction of Motor Insurance. The Loan value prediction, motor insurance premium value prediction is sensitive such that the error rate would highly influence the outcome. The mentioned sensitive predictions are highly correlated to the variable values that are used for the modeling. The better the data higher would be the accuracy and least error rates. Hence the proposed research would positively contribute and can be integrated into such projects for best accuracy prediction.

6.4 LIMITATIONS:

The research outcome expects a model that predicts the value of the used car with real-time data with the user input—each research focus to provide the best outcome with the least error rate. Though the research is diverged to produces the best out, multiple limitations limited the

study that could be applied to prediction. Limitations are crucial factors that influence the thesis. A few Major limitations to the predicted researches are mentioned below

1.Availability of the data:

The data required for the research is used by scrapping from the internet rather than pre-existing data. Hence the availability of the data is least. As mentioned in the previous section of the research. The data is crucial for prediction, and the number of data of similar car models was significantly less (¬300). As the number of data increases, the variability of the factors can be increased in the prediction, bringing out much better results than available. The more the data number more robust, the prediction to be.

2.The Scraping Issues:

The proposed research uses web scraping for extracting the data depending on the user input. There were multiple issues associated with the scraping, which limited the research. Few of them are :

1.The unavailability of whole data in the vehicle mentioned.

2. Any changes in the structure of the source from where we scrape the data would highly limit the computation

3.Difficulty in the structure of the textual data from scrape.

4.The saving of data to the data frame gets distracted with any error that occurs. The scraping was a multiple number of od issues, and limitation is considered to improve the research.

3.Feature Availability:

The features of the car are significant factors that are critical in modeling. The more the number of features available more good would be the prediction. The features with good correlation were significantly less. There was more number of features that are least correlated to price than highly correlated ones. The features such as service availability, damage rating, and expected mileage would add more value to the prediction.

4.Data cleaning:

Data cleaning is another area that is to be done rigorously. The 'NAs' were handly removed those rows filled with NAs, which reduced the data availability further. The data cleaning lies with the modeling part; hence, the scraped beauty depends on the source. The Unwanted formatting of the text and inclusion of text to the data when scraping could cause an error in the prediction, which must be handled carefully.

5.Data Preprocessing:

Data Preprocessing and data cleaning would result in making the data more rigorous for use and prediction. The data format change depending on the source, and changing the type of data from numeric to factor is to be done carefully, which highly affects the accuracy.

6.Computation time:

One of the significant factors that each research would look at is the computation time required for the product's performance. The research computation time is highly dependent on the number of variables and scraping. There was a recorded time of 3min required for computation and prediction for a model with five variables, and it then increased to 25 minutes for the computation of 14 variables.

The limitations are another milestone for future work. Every limitation is expected to be overwritten and improved in future work.

6.5 FUTURE WORK:

The future work for this research would aim to improvise at the modeling level and the feature level. The future work majorly aims at the development of the project, along with image processing. Feature level looks at the user's basic features of prediction depending on the user. The model-level looks at the research's improvisation, technically overcoming the limitations of the current product and adding more functionality and hence improving the prediction.

6.5.1 Feature Level Work:

1.The proposed research majorly focuses on the seller level. Future work expects to cover the buyer level aspect. Its focus on prediction and optimization difference depending on the user type of seller and buyer.

2.The future work expects to attach images to the variable. It can be used for computer vision for detecting the severity of the damages and hence calculate the damage severity index. The damage severity index calculation with computer vision studies would highly benefit the prediction in giving the best outcome. The index predicted by the computer vision algorithms can be used as another variable for predicting the values.

6.5.2 Model Level Work:

1.A better data platform is to find that the source's data is very accurate, depending on the market value.

2.The currently proposed research uses the method such as linear regression, GBM, Random Forest, and Light GBM for the modeling and the future work plan to use Neural Network to see how the performance accuracy and error rate are changed produces the best model.

3.Addition of the computer vision and image processing for the damage severity index prediction.

4.A new optimization method to improve the prediction and final result.

7 : BIBLIOGRAPHY

Alpaydin,Ethem(2014),*Introduction to Machine learning*.United States:MIT Press, 2014

Bercovec,James(1985),' New car sales and used car stocks: a model of the automobile market',*The RAND Journal of Economics*, Vol. 16, No. 2 (Summer, 1985), pp. 195-214

Boonsiritomachai, W., McGrath, G. M., & Burgess, S. (2016). Exploring business intelligence and its depth of maturity in Thai SMEs. *Cogent Business & Management*, *3*(1), 1220663. https://doi.org/10.1080/23311975.2016.1220663

Breiman,L. and Spector,P. (1992) Submodel selection and evaluation in regression. The X-random case. Int. Stat. Rev., 60, 291–319.

Breiman,L., Friedman,J.H., Olshen,R.A. and Stone,C.J. (1984) Classification and Regression Trees. Wadsworth and Brooks/Cole, Monterey, CA.

Burman,P. (1989) A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika, 76, 503–514.

C. Chiu and C. Shu, "Monthly car sales prediction using Internet Word-of-Mouth (eWOM)," (*2017),IEEE International Conference on Innovations in Intelligent SysTems and Applications (INISTA)*, Gdynia, 2017, pp. 345-348, doi: 10.1109/INISTA.2017.8001183.

Chen, S., Chou, E., & Yang, R.R. (2018). The Price is Right: Predicting Prices with Product Images. *ArXiv, abs/1803.11227*.

Dhar,Vasant(2013),' Data Science and Prediction',DOI:10.1145/2500499

G. Kejela, R. M. Esteves and C. Rong, "Predictive Analytics of Sensor Data Using Distributed Machine Learning Techniques," (*2014) IEEE 6th International Conference on Cloud Computing Technology and Science*, Singapore, 2014, pp. 626-631, doi: 10.1109/CloudCom.2014.44.

Gallo,Amy (2015), 'A Refresher on Regression Analysis',Hardvard Business Review,04 November.Available at: https://hbr.org/2015/11/a-refresher-on-regression-analysis (Accessed on 13 July 2020)

Gegic, E. et al. (2019) 'Car Price Prediction using Machine Learning Techniques', TEM Journal, 8(1), p. 113-118,DOI:10.18421/TEM81-16.

Gillingham.et.al(2019),' Equilibrium Trade in Automobile Markets',The national bureau of economic research, (DOI): 10.3386/w25840

Haine, Alice. (2019) 'Nine in 10 UAE car buyers will choose second-hand over new in 2019', *The National (Business)*18 Feb, Pg 14.

Hastie,T., Tibshirani,R. and Friedman,J. (2003) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer, 1st edn, 3rd print. Springer-Verlag, New York.

Huang, W., Zhang, Q., Xu, W., Fu, H., Wang, M. and Liang, X., (2015). A Novel Trigger Model for Sales Prediction with Data Mining Techniques. *Data Science Journal*, 14, p.15. DOI: http://doi.org/10.5334/dsj-2015-015

I. Markou, F. Rodrigues and F. C. Pereira, "Real-Time Taxi Demand Prediction using data from the web," (*2018), 21st International Conference on Intelligent Transportation Systems (ITSC)*, Maui, HI, 2018, pp. 1664-1671, doi: 10.1109/ITSC.2018.8569015.

Institute of Information Systems, University of Hamburg(2013), Resale Price Prediction in the Used Car Market, Available at: https://www.semanticscholar.org/paper/Resale-Price-Prediction-in-the-Used-Car-MarketVo%C3%9F/edd2191dae1b2f10052c51fd61a908e0b91fbb74

Knofczynski **&** Mundfrom(2008), 'Sample Sizes When Using Multiple Linear Regression for Prediction'Sage Journals, Volume 68 Number 3,Pg:431-442

Lacetera.et.al.(2012). 'Heuristic Thinking and Limited Attention in the Car Market'. *American Economic Review*, 102 (5): 2206-36.

Limsombunchai, V. (2010). 'House Price Prediction : Hedonic Price Model vs . Artificial Neural Network'. *American Journal of Applied Sciences* 1 (3): 193-201, 2004 ISSN 1546-9239

Ljung, L(2002).' Prediction error estimation methods'. *Circuits Systems and Signal Process* 21, Pg**:**11–21

Molinaro.et.al(2005), 'Prediction error estimation: a comparison of resampling methods', *Bioinformatics*, Volume 21, Issue 15, , Pages 3301–3307

Nairn, A., Berthon, P., & Money, A. (2003). THROUGH THE PARADIGM FUNNEL: A CONCEPTUAL TOOL FOR LITERATURE ANALYSIS. International Journal of Market Research, 49(2), 257–274.

Özer Çelik, Osmanoğlu(2019) ,Prediction of the Prices of Second-Hand Cars, European

Pal, N.et.al. (2017). How much is my car worth? A methodology for predicting used cars prices using Random Forest. *ArXiv, abs/1711.06970*.

Provost,Foster and Fawcett,Tom(2013) 'DATA SCIENCE AND ITS RELATIONSHIP TO BIG DATA AND DATA-DRIVEN DECISION MAKING', DOI: 10.1089/big.2013.1508

Puteri ,Safitri(2020), Analysis of linear regression on used car sales in Indonesia, Journal of Physics: Conference Series ,1469,doi : 10.1088/1742-6596/1469/1/012143

ResearchAndMarkets (2019) ,*UAE Used Car and Auto Classified Market 2018-2022*. Newyork: Cision PR Newswire , Available at: https://www.prnewswire.com/news-releases/uae-used-car-and-auto-classified-market-2018-2022-300840808.html(Accessed: 12.07.2020).

Rodriguez J.D.(2010),' Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation',*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 569-575, doi: 10.1109/TPAMI.2009.187.

Roeder, Theresa M. K. et al. (2016). 'REAL-TIME PREDICTIVE AND PRESCRIPTIVE ANALYTICS WITH REAL-TIME DATA AND SIMULATION'.' *Winter Simulation Conference',Availableat:*https://pdfs.semanticscholar.org/af34/ef860bf06cfe2f1338281717b6a05f1561cc.pdf?_ga=2.153409659.217006139.1595444084-528633918.1595444084

Rouse,Margaret(2015),' ensemble modelling',Tech Target,02 January. Available at: https://searchbusinessanalytics.techtarget.com/definition/Ensemble-modeling (Accessed on 13 July 2020)

S. Gongqi, W. Yansong and Z. Qiang, (2011)'New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit', *Third International Conference on Measuring Technology and Mechatronics Automation*, Shangshai, 2011, pp. 682-685, doi: 10.1109/ICMTMA.2011.455.

Sameerchand Pudaruth(2014), Predicting the Price of Used Cars using Machine Learning Techniques , International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753-764

Stanford University |(2012), Don't Get Kicked - Machine Learning Predictions for Car Buying,Available at: http://cs229.stanford.edu/proj2012/HoRomanoWuKickedCarPrediction.pdf

Stanford University |(2019), Predicting Used Car Prices,Available at: http://cs229.stanford.edu/proj2019aut/data/assignment_308875_raw/26463410.pdf

T. Petereke et.al(2013), "Performance evaluation of Random Forest regression model in tracking Parkinson's disease progress," *13th International Conference on Hybrid Intelligent Systems* ,Gammarth, 2013, pp. 83-87, doi: 10.1109/HIS.2013.6920459.

techopedia(2020)*Real-TimePredictiveAnalytics*. Available at: https://www.techopedia.com/definition/31743/real-time-predictiveanalytics#:~:text=Real%2Dtime%20predictive%20analytics%20is%20based%20on%20a%20predictive%20model,of%20large%20amounts%20of%20data.&text=This%20model%20is%20then%20made,data%20is%20fed%20to%20it..

Wayne W. Eckerson(2007) *PREDICTIVE ANALYTICS.* United States: TDWI

X. Lu and X. Geng, (2011), "Car Sales Volume Prediction Based on Particle Swarm Optimization Algorithm and Support Vector Regression," *2011 Fourth International Conference on Intelligent Computation Technology and Automation*, Shenzhen, Guangdong, pp. 71-74, doi: 10.1109/ICICTA.2011.25

Yiu,Tony(2019),' Understanding Random Forest How the Algorithm Works ,towards data science ,12 June. Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (Accessed on 13 July 2020)

Zhang,Zixuan(2019),'Boosting Algorithms Explained Theory, Implementation, and Visualization',towards data science ,26 June. Available at:

https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30 (Accessed on 13 July 2020)

8.APPENDICES

1.FINAL PRODUCT

```
#* @filter cors
cors <- function(res) {
  res$setHeader("Access-Control-Allow-Origin", "*")
  plumber::forward()
}
#* Return the price estimate
#* @param req - request object
#* @post /estimate
function(req) {
  input=(req$body)
  brandname=toupper(input$brand)
  modelname=toupper(str_replace_all(input$model," ","+"))
  baseurl="https://dubai.dubizzle.com/motors/used-cars/?page=%d"
  map_df(1:15,function(i){
    url1=sprintf(baseurl,i)
    url=paste(url1,"&price__gte=&price__lte=&year__gte=&year__lte=&kilometers__gte=&kilomet

    webpages=read_html(url)
    print(webpages)
    year=html_text(html_nodes(webpages,'.features:nth-child(1) li:nth-child(1) strong'))

    mileage=html_text(html_nodes(webpages,'.features:nth-child(1) li+ li strong'))

    print(mileage)
    mileage=gsub(",","",mileage)
    mileage=gsub("\n","",mileage)
    mileage=gsub(" ","",mileage)
    mileage=gsub("\n","",mileage)
    print("mileage is")
    print(mileage)

    #  .normal-field:nth-child(4) .normal-field-strong
    colour=html_text(html_nodes(webpages,'.features+ .features li+ li strong'))

    colour=gsub("\n","",colour)
    colour=gsub("  ","",colour)
    colour=gsub("\n","",colour)

    price=html_text(html_nodes(webpages,'.selling-price__amount'))
    price=gsub("\n","",price)
    price=gsub("  ","",price)
    price=gsub("\n","",price)
    price=gsub("AED","",price)

    #added below
    price=gsub(",","",price)
    price=gsub(" ","",price)

    print(price)
    print(colour)


    trim=html_text(html_nodes(webpages,'.features+ .features li:nth-child(1) strong'))
    trim=gsub("\n","",trim)
    trim=gsub("  ","",trim)
    trim=gsub("\n","",trim)
    trim=gsub(" doors","",trim)
    trim=gsub(" door","",trim)
    trim=gsub("\\+","",trim)
```

```
    print(trim)
    data.frame(year,mileage,trim,colour,price)
     # Sys.sleep(30)
  })->traindata
  write.csv(traindata,"D:/MSc/Study/@5Project/@6Data/car.csv")
  }
  print("ScrapedData")
  print(traindata)

  traindata$colour=as.factor(traindata$colour)
  traindata$year=as.factor(traindata$year)
  traindata$mileage=as.numeric(traindata$mileage)
  traindata$year=as.factor(traindata$year)
  traindata$trim=as.factor(traindata$trim)
  traindata$price=as.numeric(traindata$price)
  print("The Specification of Car for which Price is to be Predicted")
  test=data.frame(year=input$year,mileage= input$mileage,trim=input$door,colour=input$colc
  test$mileage=gsub(",","",test$mileage)
  test$mileage=gsub("\n","",test$mileage)
  test$mileage=gsub(" ","",test$mileage)
  test$mileage=gsub("\n","",test$mileage)
  print(test)

  gradientNormal=gbm((price) ~mileage+trim+year+colour, data = traindata, distribution = "
  summary(gradientNormal)
  probgradientNormal=round(predict(gradientNormal,test))
  print("Predicted Value of Car is ")
  print(probgradientNormal)
  a=list(success= unbox(TRUE),value= unbox(probgradientNormal))

}
```

2.SCRAPING CODE

```
baseurl="https://dubai.dubizzle.com/motors/used-cars/?page=%d"
map_df(1:10,function(i){
url1=sprintf(baseurl,i)
url=paste(url1,"&price__gte=&price__lte=&year__gte=&year__lte=&kilometers__gte=&kilometers_
webpage=read_html(url)
webpage=read_html(url)
 #Extracting the urls from landing page
   lr <- webpage%>% # feed `main.page` to the next step
   html_nodes(".results-listing-title a") %>% # get the CSS nodes
   html_attr("href") # extract the URLs
data.frame(lr)
})->dfs

#Scraping from inner page using the urls in the lr
for (i in 1:nrow(dfs))
{
webpages=read_html(dfs$lr[i])
         print(dfs$lr[i])
         year=html_text(html_nodes(webpages,'.cursor-pointer+ .border .ellipsis'))
         mileage=html_text(html_nodes(webpages,'.border:nth-child(4) .ellipsis'))
         colour=html_text(html_nodes(webpages,'.is-flex+ .plain_details_section .normal-fie

         colour=gsub("\n","",colour)
         colour=gsub("  ","",colour)
         colour=gsub("\n","",colour)

         price=html_text(html_nodes(webpages,'#price'))
         prices=gsub("\n","",price)
         prices=gsub("  ","",price)
         prices=gsub("\n","",prices)
         prices=gsub("AED ","",prices)
         print(colour)

         warranty=html_text(html_nodes(webpages,'.border:nth-child(5) .ellipsis'))
         specs=html_text(html_nodes(webpages,'.border:nth-child(6) .ellipsis'))
         transmission=html_text(html_nodes(webpages,'.border:nth-child(7) .ellipsis'))
         body=html_text(html_nodes(webpages,'.border:nth-child(8) .ellipsis'))
         bodycond=html_text(html_nodes(webpages,'.is-flex+ .plain_details_section .normal-f
         bodycond=gsub("\n","",bodycond)
         bodycond=gsub("  ","",bodycond)
         bodycond=gsub("\n","",bodycond)

         mechcond=html_text(html_nodes(webpages,'.is-flex+ .plain_details_section .normal-f
         mechcond=gsub("\n","",mechcond)
         mechcond=gsub("  ","",mechcond)
         mechcond=gsub("\n","",mechcond)

         fuel=html_text(html_nodes(webpages,'.is-flex+ .plain_details_section .normal-field
         fuel=gsub("\n","",fuel)
         fuel=gsub("  ","",fuel)
         fuel=gsub("\n","",fuel)

         trim=html_text(html_nodes(webpages,'.details_show_more .normal-field:nth-child(1)
         trim=gsub("\n","",trim)
         trim=gsub("  ","",trim)
         trim=gsub("\n","",trim)
         trim=gsub(" doors","",trim)
         trim=gsub(" door","",trim)
         trim=gsub("\\+","",trim)
```

```
        print(trim)

        cylinder=html_text(html_nodes(webpages,'.details_show_more .normal-field:nth-child
        cylinder=gsub("\n","",cylinder)
        cylinder=gsub("  ","",cylinder)
        cylinder=gsub("\n","",cylinder)

        seller_type=html_text(html_nodes(webpages,'.details_show_more .normal-field:nth-ch
        seller_type=gsub("\n","",seller_type)
        seller_type=gsub("  ","",seller_type)
        seller_type=gsub("\n","",seller_type)
       print(seller_type)

    if(isempty(year)||isempty(seller_type)||isempty(cylinder)||isempty(trim)||isempty(fuel)
       next

#storing the scraped values in the data frame
  dfs$year[i]=year
  dfs$mileage[i]=mileage
  dfs$colour[i]=colour
  dfs$price[i]=prices
  dfs$warranty[i]=warranty
  dfs$specs[i]=specs
  dfs$transmission[i]=transmission
  dfs$bodycond[i]=bodycond
  dfs$body[i]=body
  dfs$mechcond[i]=mechcond
  dfs$fuel[i]=fuel
  dfs$trim[i]=trim
  dfs$cylinder[i]=cylinder
  dfs$seller_type[i]=seller_type
}
```

In [22]:

```
#Setting Directory and path
setwd("D:/MSc/Study/@5Project/@6Data")
.libPaths("D:/MSc/Study/@5Project/@5library")
getwd()
```

'D:/MSc/Study/@5Project/@6Data'

In [ ]:

```
#Libraries Used
library("rvest")
library("purrr")
library("jmotif")
library("caret")
library("Metrics")
library("stringr")
library("randomForest")
library("gbm")
library("ggplot2")
```

3.MODELING

```
carDf=read.csv("D:/MSc/Study/@5Project/@6Data/dubizzledataKD.csv")
```

In [189]:

```
#dimension of carDf
dim(carDf)
```

243 18

In [ ]:

```
head(carDf)
```

In [ ]:

```
complete.cases(carDf)
```

In [197]:

```
#Percentile 95%
carDf<- subset(carDf,carDf$prices<quantile(carDf$prices,seq(0.95,1))+50000)
carDf<- subset(carDf,carDf$year>2000)
```

In [195]:

```
#Omit Na values
carDf=na.omit(carDf)
```

In [196]:

```
dim(carDf)
```

243 18

```
#changing type of model
carDf$year=as.numeric(carDf$year)
carDf$colour=as.factor(carDf$colour)

carDf$warranty=as.factor(carDf$warranty)
carDf$prices=as.numeric(carDf$prices)

carDf$specs=as.factor(carDf$specs)
carDf$mileages=as.numeric(carDf$mileages)

carDf$transmission=as.factor(carDf$transmission)

carDf$bodycond=as.factor(carDf$bodycond)

carDf$body=as.factor(carDf$body)

carDf$mechcond=as.factor(carDf$mechcond)

carDf$fuel=as.factor(carDf$fuel)

carDf$trim=as.numeric(carDf$trim)
carDf$cylinder=as.numeric(carDf$cylinder)

carDf$seller_type=as.factor(carDf$seller_type)
```

In [15]:

```
typeof(carDf)
```

'list'

In [199]:

```
# Test Train Normal Split
traindatas=carDf[1:225,]
tests=carDf[226:nrow(carDf),]
```

3.1 Linear Regression

Linear Regression Normal Split

```
#The linear model training
linearNormal=lm(prices~mileages+year+colour+warranty+bodycond+body+mechcond+trim,data = tra
```

In [ ]:

```
summary(linearNormal)
```

```
#Code for Printing the accuracy,RMSE,MAPE,MAE
actuals_predss <- data.frame(cbind(actuals=traindatas$prices, predicteds=probLinearNormal))

actuals_preds <- data.frame(cbind(actuals=tests$prices, predicteds=probLinearNormal))
print("actual")
actuals_preds

#Train accuracy
min_max_accuracy <- mean(apply(actuals_predss, 1, min) / apply(actuals_predss, 1, max))
print("Min_Max Train Accuracy")
min_max_accuracy

#Test accuracy
min_max_accuracy <- mean(apply(actuals_preds, 1, min) / apply(actuals_preds, 1, max))
print("Min_Max Test Accuracy")
min_max_accuracy

mape <- mean(abs((actuals_preds$predicteds - actuals_preds$actuals))/actuals_preds$actuals)
print("MAPE")
mape

Avg_error <- mae(tests$prices,probLinearNormal)
print("MAE")
Avg_error

R_error <- rmse(tests$prices,probLinearNormal)
print("RMSE")
R_error
```

Linear Regression Normal Split -Log10(prices)

```
linearNormal=lm(priceslog~mileages+year+colour+warranty+bodycond+body+mechcond+trim,data =
```

In [ ]:

```
summary(linearNormal)
```

In [ ]:

```
probLinearNormal=predict(linearNormal,tests)
probLinearNormal
```

Random Split with Normal (price) value

```
#Splitting data Randomly
trains = sample(1:nrow(carDf),floor(nrow(carDf)*0.80))
tests = -trains

trainRandom=carDf[trains,]
dim(trainRandom)

testRandom=carDf[tests,]
dim((testRandom))

trainRandom$colour=as.factor(trainRandom$colour)
trainRandom$trim=as.factor(trainRandom$trim)
trainRandom$year=as.factor(trainRandom$year)
trainRandom$mileage=as.numeric(trainRandom$mileage)
testRandom$colour=as.factor(testRandom$colour)
testRandom$trim=as.factor(testRandom$trim)
testRandom$year=as.factor(testRandom$year)
testRandom$mileage=as.numeric(testRandom$mileage)
```

192 19

48 19

In [84]:

```
linearNormal=lm(prices~mileages+year+colour+warranty+trim,data = trainRandom)
```

In [ ]:

```
summary(linearNormal)
```

In [ ]:

```
probLinearNormal=predict(linearNormal,testRandom)
probLinearNormal
```

```
linearNormal=lm(priceslog~mileages+year+colour+warranty+trim,data = trainRandom)
```

In [ ]:

```
summary(linearNormal)
```

In [ ]:

```
probLinearNormal=predict(linearNormal,testRandom)
probLinearNormal
```

Cross Validation

```
folds = createFolds(carDf$prices, k = 10)
cv = lapply(folds, function(x) {
  training_fold = carDf[-x, ]
  #print(training_fold)
  test = carDf[x, ]
  classifier0 = lm(formula = prices ~ mileages+warranty+trim+cylinder,
                   data = training_fold
                  )
  y_pred0 = predict(classifier0, newdata = test[-3])
  print(y_pred0)

actuals_preds <- data.frame(cbind(actuals=test$prices, predicteds=y_pred0))
print("actual")
actuals_preds

min_max_accuracy <- mean(apply(actuals_preds, 1, min) / apply(actuals_preds, 1, max))
print("min_max")
min_max_accuracy

mape <- mean(abs((actuals_preds$predicteds - actuals_preds$actuals))/actuals_preds$actuals)


Avg_error <- (mae(test$prices,y_pred0))
print(Avg_error)
print("RMSE")

R_error <- (rmse(test$prices,y_pred0))
print(R_error)
})

print("Min_Max test Accuracy")
accuracy = mean(as.numeric(min_max_accuracy))
m= mean(as.numeric(mape))
print(accuracy)
print("mape")
print(m)
print("MAE")
print(Avg_error)
print("RMSE")
print(R_error)
```

3.2 Gradient Boost Model

Normal Spilt

```
#Gradient model fitting
gradientNormal=gbm(prices ~mileages+year+colour+warranty+bodycond+body+mechcond+trim, data
```

In [ ]:

```
summary(gradientNormal)
```

In [ ]:

```
probgradientNormal=predict(gradientNormal,tests)
probgradientNormal
```

Normal Split-log10(Price)

In [125]:

```
gradientNormal=gbm(priceslog ~mileages+year+colour+warranty+bodycond+body+mechcond+trim, da
```

In [ ]:

```
probgradientNormal=predict(gradientNormal,tests)
probgradientNormal
```

Random Split

In [136]:

```
#Normal Value of
gradientNormal=gbm(prices~mileages+year+colour+warranty+bodycond+body+mechcond+trim, data =
```

Random Split

In [136]:

```
#Normal Value of
gradientNormal=gbm(prices~mileages+year+colour+warranty+bodycond+body+mechcond+trim, data =
```

In [ ]:

```
summary(gradientNormal)
```

In [ ]:

```
probgradientNormal=predict(gradientNormal,testRandom)
probgradientNormal
```

Random split with Log10(Price)

```
gradientNormal=gbm(priceslog~mileages+year+colour+warranty+bodycond+body+mechcond+trim, dat
```

3.3 Random Forest

Normal Split

In [157]:

```
#Normal Split used
randomNormal <- randomForest(prices ~mileages+colour+warranty+specs+transmission+bodycond+b
)
```

In [ ]:

```
summary(randomNormal)
```

In [ ]:

```
probrandomNormal=predict(randomNormal,tests)
probrandomNormal
```

Normal Split with log10(Price)

In [168]:

```
#Normal Split log10(Price)
randomNormal <- randomForest(priceslog ~mileages+colour+warranty+specs+transmission+bodycon
)
```

In [ ]:

```
summary(randomNormal)
```

```
probrandomNormal=predict(randomNormal,tests)
probrandomNormal
```

Random Split

```
randomForestRandom<- randomForest(priceslog ~mileages+colour+warranty+specs+transmission+bc
)
```

In [ ]:

```
summary(randomNormal)
```

In [ ]:

```
probrandomNormal=predict(randomNormal,tests)
probrandomNormal
```

4 HYPERPARAMETER TUNING

Normal Value of Price

In [ ]:

```
set.seed(71)
rf <-randomForest(prices~mileages+colour+specs+warranty+body+bodycond+fuel+mechcond+trim+cy
print(rf)
#Evaluate variable importance
importance(rf)
varImpPlot(rf)
```

In [ ]:

```
# Random Search
control <- trainControl(method="repeatedcv", number=10, repeats=3, search="random")
#set.seed(seed)
mtry <- sqrt(ncol(traindata))
rf_random <- train(price~mileage+colour+specs+warranty+body+bodycond+fuel+mechcond+trim+cyl
print(rf_random)
plot(rf_random)
```

Log10(Price)

```
set.seed(71)
rf <-randomForest(priceslog~mileages+colour+specs+warranty+body+bodycond+fuel+mechcond+trim
print(rf)
#Evaluate variable importance
importance(rf)
varImpPlot(rf)
```

In [ ]:

```
# Random Search
control <- trainControl(method="repeatedcv", number=10, repeats=3, search="random")
#set.seed(seed)
mtry <- sqrt(ncol(traindata))
rf_random <- train(price~mileage+colour+specs+warranty+body+bodycond+fuel+mechcond+trim+cyl
print(rf_random)
plot(rf_random)
```

4.2 Gradient Boosting

In [ ]:

```
fitControl <- trainControl(method = "cv",
                           number = 3,
                           repeats = 5,
                           classProbs = TRUE)

set.seed(825)
rda_fit <- train(price~mileage+colour+specs+warranty+body+bodycond+fuel+mechcond+trim+cylin
                 method = "gbm",
                 tuneLength = 50,
                 trControl = fitControl)
rda_fit
fitControl
```

5. FINAL MODEL

In [ ]:

```
gradientNormal=gbm(priceslog ~mileages+year+colour+warranty+bodycond+body+mechcond+trim, da
                   distribution = "gaussian",n.trees = 200, interaction.depth = 29, shrinka
```

In [ ]:

```
summary(gradientNormal)
```

In [ ]:

```
probgradientNormal=predict(gradientNormal,testRandom)
probgradientNormal
```