Car Sales Prediction Using Machine Learning Algorithmns

Madhuvanthi.K, Nallakaruppan.M.K, Senthilkumar N C, Siva Rama Krishnan S

Abstract: *Sales prediction is the current numero trend in which all the business companies thrive and it also aids the organization or concern in determining the future goals for it and its plan and procedure to achieve it. The data about car sales are derived from various sources .sales of cars does not contain any independent variable since various factors such as horse power; model, width, fuel type, height, price, city-mileage, highway-mileage and manufacturer are the various features that influence the sales. In car sales prediction we first implement the methodology of analytic hierarchy process in order to get varied idea about how well the various criteria's in our dataset works and after this we apply the machine learning algorithms such as Linear regression, Random tree to get the best clusters and we process them in to random forest to get best accurate feature out of it. which is ultimately followed by* ***Technique for Order of preference by similarity to ideal solution (TOPSIS) an tool which helps the researcher to arrive at a verdict when he/she faces the one or more pattern selection problem and the*** *final resultant derived from all these methods gives the fittest feature which influences the customer in purchasing the car which indirectly gives the company or the research market a result in predicting the future sales for cars. Hence this paper not only provides its users with some stats, it also serves an guiding guardian by providing accurate results for purchasing a car.*

Keywords: *Sales prediction, horse power, model, engine power, clusters, Accurate.*

I. INTRODUCTION

Every human being has got his/her business in this world .so the term business comes with certain tagged words such as sales, profit and loss. A concerns success is determined by its sales and the performance of its product in the industry is also identified by its sales. Hence in order to improve the standard of the firm /concern the strategies, the techniques the methodologies are built .in this process of development forecasting of certain key terms such as profit, loss, return of interest may pave the way for the successful venture of the concern yet when we look deep in to the details all these key terms are conveniently related to the term called sales .sales prediction is one of the master trades of business which may open the gateways for obtaining knowledge about the existing market trends and the ways to conquer the market .planning is the first step in every activity we perform and hence knowing what lies ahead in terms of sales hugely aids the concern/organization in this planning process . sales prediction can be a massive support in order to perform tracking,cashflow and purchasing .sales forecasting provides the business minds an idea of about how much to buy and how much not to buy ,what are the risk can be taken in terms of revenue ,how to plan budget ,market trends, introduction of new products according to the organizations capability and ability ,what changes may happen if the plan fails are the areas where it helps to develop our ideas ,let us shift our focus from the generic business term to its detailed chunk of streams such as education,medical,automobile and lots of other streams. Once great interest in recent days lies in automobile industry. The first boon to the automobile industry came only after the industrialization renaissance period. Now the automobile industry creates, packs, runs, moves the world on its wheels and also allows other fields to flourish parallel along with them .yet it also has its own defaults since it is not cost friendly ,any single default can affect the whole system and it cause a huge failure in the market for its brand ,there is a huge amount of revenue being spend and huge amount of chunks being produced and exploited on daily basis hence sales forecasting in this industry can lead to the organized pattern of selling,buying,producing goods and even the taxes to be implemented also comes in to role .forecasting these sales in automobile industries can be performed with various and variety of technologies and one among them is the machine learning technique. That may help in classifying the prediction of an automobile for a say lets us take it as car the yearly sales of it if know before then it will provide the manufacturer the huge boost in designing it, getting spare parts, getting key parts and reducing the waste products and tracking its revenue model its generation and various other activity. The classifiers used such as logistic regression, decision tree and random forest provides us with accurate prediction results

II. LITERATURE SURVEY:

Machine learning models and bankruptcy prediction is a paper work which talks about the improvement that takes place in academics industry with the aid of machine learning algorithms in predicting bankruptcy .The data is derived from integrated resource of Salomon center database which contains the details about the North American firms from the period between 1985 to 2013 . This paper implements the usage of algorithms such as bagging, boosting, random forest and support vector machine for

Revised Manuscript Received on March 10, 2019.

Madhuvanthi.K, M.Tech (software engineering), Vellore institute of Technology University, Vellore, India.

Nallakaruppan.M.K, SITE, Vellore institute of Technology University, Vellore, India.

Senthilkumar N C, SITE, Vellore Institute of Technology University, Vellore, India.

Siva Rama Krishnan S, SITE, Vellore Institute of Technology University, Vellore, India

predicting bankruptcy even before the event occurs and a greater span of comparative study takes place with the performance of these results with the results of logistic regression and neural networks [9]. Original Altman's Z-score variables are used as predictive variables with addition of extra variables such as the operating margin,sales,growth measures related to assets, change in return-on-equity, change in price-to-book, and number of employees based on carton and Hofer(2006). And a comparison is made between the models and these variables, the machine learning techniques and the algorithm with most accuracy is determined. Handling class imbalance in customer churn prediction by j.Burez and D. van den poel suggests the customer the various ways to handle class imbalance in churn prediction. AUC and lift are the evaluation metrics with which the sampling methods are interrogated .the modeling techniques such as weighted random forest, gradient boosting are compared with other techniques. The better evaluation metrics and the best modelling techniques are found out with the help of each techniques accuracy and from past studies [6]. Calling communities analysis and identification using machine learning techniques is the work that determines the worth of a particular customer with respect to his/her general pattern trait of the community that he/she hails from. The customer calling impressions can be told beforehand by making use of a classifier model and cluster analysis for detail selection. The attributes such as accuracy and computational performance are taken in to consideration for comparison of various machine learning techniques [1]. Customer churn prediction using improved balanced random forests is the paper that explains the real time working model that had been used in china. Improved balanced random forest is the hybrid version of balanced random forest and weighted random forest, two interval variables had been introduced such as e and f where e is the middle point and f is the length of interval .Random distribution of these classes are maintained with the help of these variables .Hence it produces more accurate results than its other counterparts [2]. A sampling based sentiment mining approach for e-commerce applications paper puts the limelight on how the customers are being influenced by the Online reviews which is a part of marketing strategy of the e-commerce platforms. Hence this issue is attempted with the help of mining techniques .The two sampling methods are also used for classification of imbalanced data. A modified support vector machine based ensemble algorithm is the methodology used by the researches to identify the performance prediction [10]. On the differential benchmarking of promotional efficiency with machine learning modelling (II): Practical applications presents two different databases of different categories such as non-seasonal and heavy seasonal and models are analyzed here .The detailed performance of four famous machine learning techniques that has huge complexity is been dissected in this work. Certain features of various machine learning algorithms do not perform well because of these databases. In order to gain more accurate dissection results and feature extraction there is a need to implement certain correct procedures that may influence the specificity of the behavior of certain categories and product ranges [11]. Linguistic features for review helpfulness prediction by Srikumar Krishnamurthy analyses what makes an Online review with the help of a predictive model .This model follows the methodology of extracting linguistic category features such as adjective feature, state verb feature and action verb features it also takes in to account the readability related features for prediction. Hence the hybrid set of features that are obtained after the analysis on two real-life review datasets gives the researches the best accuracy rate of all time [8]. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data paper puts the focus on how this work can contribute to the real estate industry that may cause an adverse effect on the US housing market.an eight step methodologies are used for the dataset of 5359 townhouses in Fairfax County, Virginia. The dataset has been segregated in to training, validation and testing set then training parameters such as C4.5, RIPPER, Naïve Bayesian, and Adaboost are set and the model is trained and evaluated using training set and validation set respectively and this process is iterated until it gains an optimal error in training, validation and testing.Finaly these results are compared to gain the optimal accuracy results [7]. Explaining machine learning models in sales prediction is a generic manuscript that discusses about the recent trends of predictive models, real time scenarios in order to gain a deep insight about buyers and seller's interaction and the forecasting of sales [5]. Early churn prediction with personalized targeting in mobile social games is a manuscript that explains Customer churn .churn is defined by the act of a customer leaving a product for good. This churn are reduced to a greater extent by following the procedure of mapping the feature with the interest of the customer and pushing the notifications in order to drag back the customer in to the game .this manuscript implements the methodologies such as logistic regression for the simple object linear model ,decision trees for extracting redundancy from features random forest to be used in various situations .Naive Bayes for generating the models and gradient boosting for its popularity [4]. Distributed customer behavior prediction using multiplex data: A collaborative MK-SVM approach is a paperwork that explains the challenges that are encountered by the traditional method in predicting customer behavior. It contains a huge datasets which are categorized as static, symbolic sequential, textual data and time series in its database collaborative multiple kernel support vector machine (C-MK-SVM) is the new technique that is used for distributed customer behavior prediction with the aid of multiplex data. Various sub models in this technique is used for global optimization. The results obtained through Computation tell the researches that it is best suited for customer behavior prediction performance and for its maximum computational speed [3].

Customer churn prediction using improved balanced random forest is also one of the works on churn prediction.it undergoes the disadvantage of imbalanced in the data distribution. This improved random balanced forest also uses some other sampling techniques with it. IBRF's features are misclassified minority class with higher penalties are iteratively learned by altering the class distributions. It works with the real time data such as bank customer database. When compared with other methodologies such as artificial neural networks, decision trees, and class-weighted core support vector machines (CWC-SVM) IBRF has more accurate prediction features [2].The process /model of the system is kicked off to action by collecting the data for car sales prediction from renowned data repositories and available databases around the world and then these data are preprocessed, then all these data sets are selected for training and the best dataset are classified and these classifiers are further trained using various methodology and the results are predicted and its performance are evaluated and finally the results are being displayed.

III. IMPLEMENTATION

AHP-The Analytic hierarchy process is a Boon to the data industry since it helps people in taking the complex decisions .any decision depends on at what point of time the decision is taken and how well it helps in growth of the product. Hence this AHP helps in taking not just one equitable decision but it stocks up the customer with the most desirable decision solution that can help them in achieving their objective to their problem. It can be implemented in various sectors where we can expect a huge amount of data .For example in automobile industry people can have a problem related to purchasing an automobile in this case AHP comes in to action where out of various types of automobiles it gives the customer a detailed report of desirable solutions. When we look deep down in to AHP process certain steps needs to be followed. We take the Automobile dataset and the attributes of the dataset are given a grading based on the weights of those attributes and in the next process these weighted attributes are been

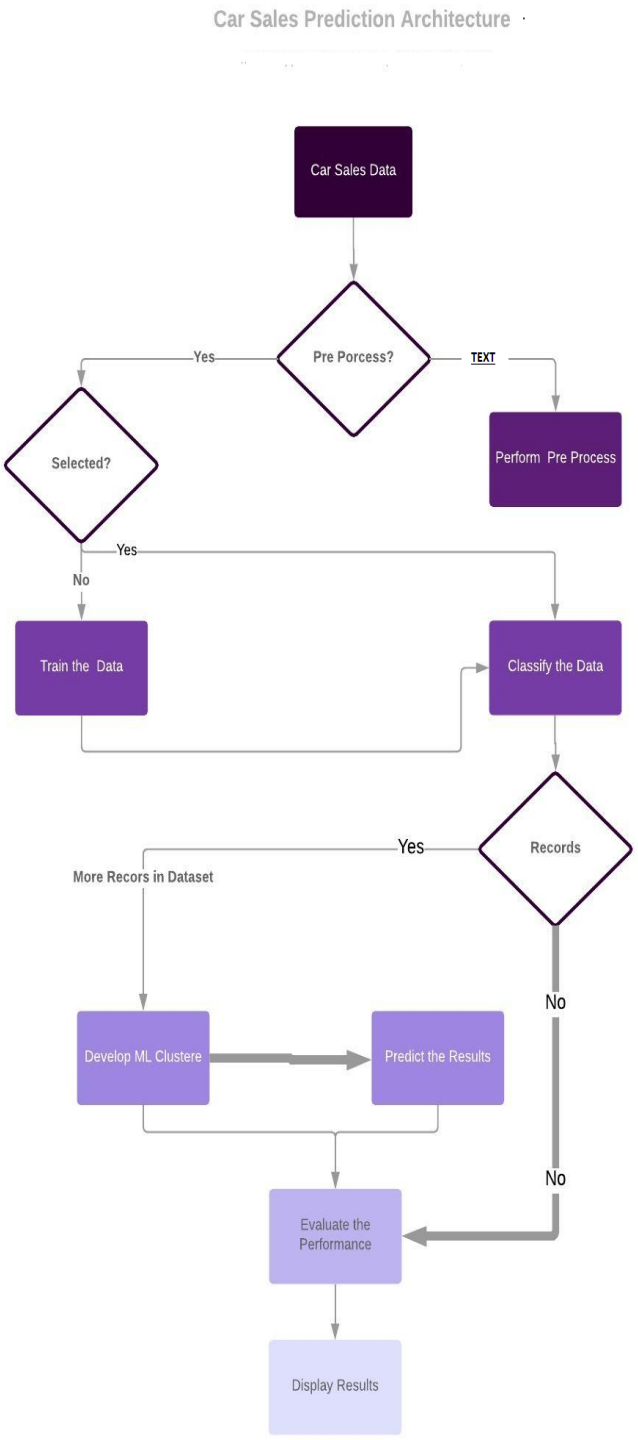

Figure1. Car Sales Architecture for Decision Making Process

assessed based on its values .The points are pitted against each other in pairs and it is checked whether they can achieve its objective .The number of pairs to be drawn against each other are first entered and then these pairs are pitted against AHP priorities and each attributes are given parallel values in the integer range of one to nine where one represents significant, three represents moderate significance ,five represents strong significance, seven represents very strong significance ,nine represents .in our automobile dataset eight attributes such as city-mileage, highway-mileage, model name,height,width,price,fueltype,horsepower in which price serves with the value of high significance and the difference between each of these eight values are pitted against each other to gain the consistency rate .if that value is less than ten percent it is said to be of with good consistency. And the following results are achieved:

Table 1.Summary of Analytic Hierarchy Process

Category		Priority	Rank	weights
1	Mileage City	22.00%	2	0.220058
2	Mileage Highway	15.90%	3	0.15939
3	Model name	2.40%	8	0.024315
4	height	3.40%	7	0.033876
5	width	5.00%	6	0.049616
6	price	33.30%	1	0.333246
7	Fuel Type	10.70%	4	0.107498
8	Horse Power	7.20%	5	0.072001

After the obtaining the suggestion for the most likely decision like price should have the highest priority the machine learning algorithms are applied to this dataset to obtain a better predictive mode.

A.Linear Regression:

The interaction that happens between any two fixed components which are also coined as variables and the method used to define or calculate the weightage of their bond is called regression analysis Their main focus is to make the novice understand the process that happens when there is a drastic change and how these changes happen due to the modification of predictors and the adverse effect it has on criterion variable .at the preliminary level the data generation serves as a key in order to understand this model. The always three vital parts needed are the criterion variable, the predictors and the disguised parameters. We also categorize them in to two simple and multiple (more than one variable).The tools or the metrics to measure the linear variable varies within these two types an unstoppable scale and categorical one .to put in precise terms correlation is what the linear regression is made of whereas the difference lies in their ability to distinct things. The straight line formula:

G(u)=t(0)+t(1)f(u)+d(u).u=1,…..,n.

Hence forms this simple linear regression is estimated as:

G^(u)=t^(0)+t^(1)f(u).

And the remaining or the left over part gives the ability of a variance which is also known as mean square error (MSE)

DDF=summation of j (u) ^2 where u =1 to n. S^2=DDF/n-2.

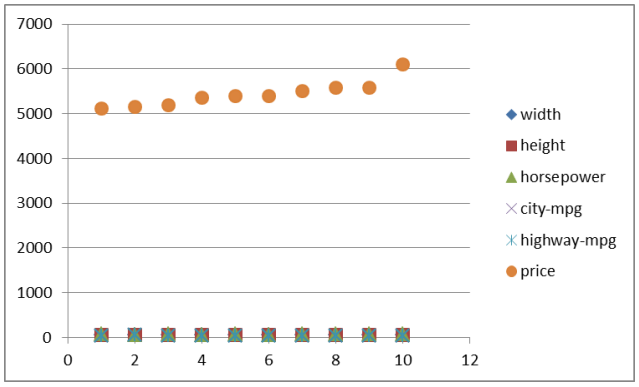

Figure 2. Linear Regression of Price with other attributes

Table 2.Linear Regression for Price vs Rank

Correspondence Coefficient	1
Average absolute error	0
Root Average squared error	0
Comparative absolute error	0%
Root comparative Squared error	0%
Total number of instances	10

A figure depicting the linear regression graph on the dataset.

Table 3 Linear Regression for Price vs Highway

Correspondence Coefficient	1
Average absolute error	0
Root Average squared error	0
Comparative absolute error	0%
Root comparative Squared error	0%
Total number of instances	10

Table 4. Linear Regression for Price vs City

Correspondence Coefficient	1
Average absolute error	0
Root Average squared error	0
Comparative absolute error	0%
Root comparative Squared error	0%
Total number of instances	10

Table 5: Linear Regression for Price vs Horse Power

Correspondence Coefficient	1
Average absolute error	0
Root Average squared error	0
Comparative absolute error	0%
Root comparative Squared error	0%
Total number of instances	10

Table 6: Linear Regression for Price vs height

Correspondence Coefficient	1
Average absolute error	0
Root Average squared error	0
Comparative absolute error	0%
Root comparative Squared error	0%
Total number of instances	10

Table 7: Linear Regression for Price vs width

Correspondence Coefficient	1
Average absolute error	0
Root Average squared error	0
Comparative absolute error	0%
Root comparative Squared error	0%
Total number of instances	10

B.Random tree:

A random tree is a tree/arborescence which is formed by a random process and can deal with both classification and regression problems. The classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of votes. For regression, the classifier response is the average of the responses over all the trees in the forest. All the trees are trained with the same parameters but on training sets. In random trees, there is no need for an accuracy estimation procedure, such as cross – validation or, a separate test set to get an estimate of the training error.

Highway Mileage:

```
width < 61.85 : 53 (1/0)
width >= 61.85
|  city-mpg < 33
|  |  city-mpg < 30.5 : 31 (1/0)
|  |  city-mpg >= 30.5
|  |  |  make = subaru : 36 (1/0)
|  |  |  make = chevrolet : 0 (0/0)
|  |  |  make = mazda : 38 (1/0)
|  |  |  make = toyota : 0 (0/0)
|  |  |  make = mitsubishi : 0 (0/0)
|  |  |  make = honda : 0 (0/0)
|  |  |  make = nissan : 37 (1/0)
|  |  |  make = dodge : 0 (0/0)
|  |  |  make = plymouth : 0 (0/0)
|  city-mpg >= 33
|  |  city-mpg < 36 : 39 (1/0)
|  |  city-mpg >= 36
|  |  |  city-mpg < 37.5 : 41 (3/0)
|  |  |  city-mpg >= 37.5 : 42 (1/0)
Size of the tree : 20
```

Summary:

Table 8: Highway Mileage Summary

Correspondence Coefficient	1
Average absolute error	0
Root Average squared error	0
Comparative absolute error	0%
Root comparative Squared error	0%
Total number of instances	10

City Mileage:

```
width < 61.85 : 47 (1/0)
width >= 61.85
|  highway-mpg < 38.5
|  |  highway-mpg < 33.5 : 30 (1/0)
|  |  highway-mpg >= 33.5 : 31 (3/0)
|  highway-mpg >= 38.5
|  |  make = subaru : 0 (0/0)
|  |  make = chevrolet : 0 (0/0)
|  |  make = mazda : 0 (0/0)
|  |  make = toyota : 35 (1/0)
|  |  make = mitsubishi : 37 (1/0)
|  |  make = honda : 38 (1/0)
|  |  make = nissan : 0 (0/0)
|  |  make = dodge : 37 (1/0)
|  |  make = plymouth : 37 (1/0)

Size of the tree: 16
```

Summary:

Table 9 :Summary of City Mileage

Correspondence Coefficient	1
Average absolute error	0
Root Average squared error	0
Comparative absolute error	0%
Root comparative Squared error	0%
Total number of instances	10

Price:

width < 63.5
| width < 61.85 : 5151 (1/0)
| width >= 61.85 : 5118 (1/0)
width >= 63.5
| city-mpg < 30.5 : 5195 (1/0)
| city-mpg >= 30.5
| | make = subaru : 0 (0/0)
| | make = chevrolet : 0 (0/0)
| | make = mazda : 6095 (1/0)
| | make = toyota : 5348 (1/0)
| | make = mitsubishi : 5389 (1/0)
| | make = honda : 5399 (1/0)
| | make = nissan : 5499 (1/0)
| | make = dodge : 5572 (1/0)
| | make = plymouth : 5572 (1/0)
Size of the tree : 16

Table 10. Summary of Price

Correspondence Coefficient	1
Average absolute error	0
Root Average squared error	0
Comparative absolute error	0%
Root comparative Squared error	0%
Total number of instances	10

Height:

width < 64.3
| highway-mpg < 40
| | make = subaru : 53.7 (1/0)
| | make = chevrolet : 0 (0/0)
| | make = mazda : 54.1 (2/0)
| | make = toyota : 54.5 (1/0)
| | make = mitsubishi : 0 (0/0)
| | make = honda : 0 (0/0)
| | make = nissan : 54.5 (1/0)
| | make = dodge : 0 (0/0)
| | make = plymouth : 0 (0/0)
| highway-mpg >= 40
| | make = subaru : 0 (0/0)
| | make = chevrolet : 53.2 (1/0)
| | make = mazda : 0 (0/0)
| | make = toyota : 0 (0/0)
| | make = mitsubishi : 0 (0/0)
| | make = honda : 52.6 (1/0)
| | make = nissan : 0 (0/0)
| | make = dodge : 50.8 (1/0)
| | make = plymouth : 50.8 (1/0)
width >= 64.3 : 50.8 (1/0)
Size of the tree : 23
Summary:

Table 11; Summary of Height

Correspondence Coefficient	1
Average absolute error	0
Root Average squared error	0
Comparative absolute error	0%
Root comparative Squared error	0%
Total number of instances	10

Horse power:

width < 61.85 : 48 (1/0)
width >= 61.85
| highway-mpg < 41.5
| | make = subaru : 69 (1/0)
| | make = chevrolet : 0 (0/0)
| | make = mazda : 68 (2/0)
| | make = toyota : 62 (1/0)
| | make = mitsubishi : 68 (1/0)
| | make = honda : 0 (0/0)
| | make = nissan : 69 (1/0)
| | make = dodge : 68 (1/0)
| | make = plymouth : 68 (1/0)
| highway-mpg >= 41.5 : 60 (1/0)
Size of the tree : 14
Summary:

Table 12. Summary of Width

Correspondence Coefficient	1
Average absolute error	0
Root Average squared error	0
Comparative absolute error	0%
Root comparative Squared error	0%
Total number of instances	10

Width:

horsepower < 54 : 60.3 (1/0)
horsepower >= 54
| horsepower < 68.5

Table 13. Summary of Make

Correctly Classified Instances	10	100%
Incorrectly Classified Instances	0	0
Kappa statistic		1
Average absolute error		0
Root Average squared error		0
Comparative absolute error		0%
Root comparative Squared error		0%
Total Number of Instances		10

| | make = subaru : 0 (0/0)
| | make = chevrolet : 0 (0/0)
| | make = mazda : 64.2 (2/0)
| | make = toyota : 63.6 (1/0)
| | make = mitsubishi : 64.4 (1/0)
| | make = honda : 64 (1/0)
| | make = nissan : 0 (0/0)
| | make = dodge : 63.8 (1/0)
| | make = plymouth : 63.8 (1/0)
| horsepower >= 68.5
| | make = subaru : 63.4 (1/0)
| | make = chevrolet : 0 (0/0)
| | make = mazda : 0 (0/0)
| | make = toyota : 0 (0/0)
| | make = mitsubishi : 0 (0/0)
| | make = honda : 0 (0/0)
| | make = nissan : 63.8 (1/0)
| | make = dodge : 0 (0/0)
| | make = plymouth : 0 (0/0)

Size of the tree: 23

Summary:

Table 14. Summary of City vs Width

Correspondence Coefficient	1
Average absolute error	0
Root Average squared error	0
Comparative absolute error	0%
Root comparative Squared error	0%
Total number of instances	10

Make:

city < 0.07
| width < 0.02
| | height < 0.01 : subaru (1/0)
| | height >= 0.01 : toyota (1/0)
| width >= 0.02
| | width < 0.02 : nissan (1/0)
| | width >= 0.02 : mazda (2/0)
city >= 0.07
| height < 0.01
| | width < 0.02
| | | horse < 0.02 : dodge (1/0)
| | | horse >= 0.02 : plymouth (1/0)
| | width >= 0.02 : mitsubishi (1/0)
| height >= 0.
| | width < 0.02 : chevrolet (1/0)
| | width >= 0.02 : honda (1/0)

Size of the tree : 17

Summary:

Table 15: Summary of make detailed accuracy class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Subaru
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Chevrolet
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	mazda
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Toyota
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Mitsubishi
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	honda
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	nissan
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	dodge
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Plymouth
Weighted Avg.	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	

Confusion Matrix

```
a b c d e f g h i   <-- classified as
 1 0 0 0 0 0 0 0 0 | a = subaru
 0 1 0 0 0 0 0 0 0 | b = chevrolet
 0 0 2 0 0 0 0 0 0 | c = mazda
 0 0 0 1 0 0 0 0 0 | d = toyota
 0 0 0 0 1 0 0 0 0 | e = mitsubishi
 0 0 0 0 0 1 0 0 0 | f = honda
 0 0 0 0 0 0 1 0 0 | g = nissan
 0 0 0 0 0 0 0 1 0 | h = dodge
 0 0 0 0 0 0 0 0 1 | i = plymouth
```

C. Random Forest*:*

This algorithm can be used with ease since it is a cluster of decision trees put together to form a forest so it can be put in to purpose in both classification and regression problems which occur in the field of machine learning. It searches in which part of an attribute set contains the best feature so that it can randomly split them in to a node since there are varsity of attributes they will load us with better performances .and the difference here is only randomly selected attribute set is taken in to consideration for spiting and can be achieved by benchmarking a threshold value for every particular attribute. This methodology also prunes out the most inappropriate tress in the forest to minimize its impurity. During prediction one can easily identify the Significant importance of each feature and a vast difference is also being detected between decision tree and random forest .in order to gain an absolute magnificent performance one needs to tune its hyper parameter hence in this methodology they are max_features, prediction rate, random state and the final parameter for cross validating the random forest. They may be so easy to train but while executing they react in a very slow manne The results are as follows

Table 16: Summary of Price in Random Forest**:**

Correspondence Coefficient	0.9699
Average absolute error	0.0019
Root Average squared error	0.0026
Comparative absolute error	49.9066%

Root comparative Squared error	50.0255%
Total number of instances	10

Table 17: Summary of Highway Mileage in Random Forest

Correspondence Coefficient	0.9742
Average absolute error	0.002
Root Average squared error	0.0032
Comparative absolute error	42.6656%
Root comparative Squared error	47.6472%
Total number of instances	10

Table 18: Summary of city Mileage in Random Forest :

Correspondence Coefficient	0.982
Average absolute error	0.0024
Root Average squared error	0.0039
Comparative absolute error	32.1675%
Root comparative Squared error	40.9679%
Total number of instances	10

Table 19: Summary of Horsepower in Random Forest

Correspondence Coefficient	0.9934
Average absolute error	0.0007
Root Average squared error	0.0011
Comparative absolute error	39.4152%
Root comparative Squared error	48.493%
Total number of instances	10

Table 20: Summary of Height in Random Forest

Correspondence Coefficient	0.9629
Average absolute error	0.0022
Root Average squared error	0.0022
Comparative absolute error	32.1675%
Root comparative Squared error	40.9679%
Total number of instances	10

Table 21: Summary of Width in Random Forest

Correspondence Coefficient	0.9654
Average absolute error	0.0001
Root Average squared error	0.0002
Comparative absolute error	32.1675%
Root comparative Squared error	40.9679%
Total number of instances	10

Table 22: Summary of Make in Random Forest

Correctly Classified Instances	10	100 %
Incorrectly Classified Instances	0	0
Kappa statistic	1	
Average absolute error	0.072	
Root Average squared error	0.1349	
Comparative absolute error	36.6429%	
Root comparative squared error	43.1042%	
Total number of instances	10	

Table 23: summary of detailed accuracy by class of make

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Subaru
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Chevrolet
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	mazda
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Toyota
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Mitsubishi
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	honda
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	nissan
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	dodge
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Plymouth
Weighted Avg.	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	

Confusion Matrix

```
a b c d e f g h i   <-- classified as
 1 0 0 0 0 0 0 0 0 | a = subaru
 0 1 0 0 0 0 0 0 0 | b = chevrolet
 0 0 2 0 0 0 0 0 0 | c = mazda
 0 0 0 1 0 0 0 0 0 | d = toyota
 0 0 0 0 1 0 0 0 0 | e = mitsubishi
 0 0 0 0 0 1 0 0 0 | f = honda
 0 0 0 0 0 0 1 0 0 | g = nissan
 0 0 0 0 0 0 0 1 0 | h = dodge
 0 0 0 0 0 0 0 0 1 | i = plymouth
```

D.TECHNIQUE FOR ORDER OF PREFERENCE BY SIMILARITY TO IDEAL SOLUTION (TOPSIS):

In car sales prediction to give the customer a knowledge about which car can be rated as the best we have certain points of reference where we rate them according to the ranks and present them before the customer that is fuzzy logic. In order to gain the ranks of the car we attribute those points of references with the weighted values. We had already given the numeric preference to the values in AHP method itself and we also follow this here. Mainly we take six attributes such as width, height, horsepower, city, highway and price. We perform vector normalization using formulas such as E(kl)=E(kl)/Square root of E^2(kl) where summation is from l=1 to n. The square root part of this formula is calculated by adding up the square values of a particular column and taking the square root of final results .now the normalized value of each cell is being gained by dividing the performance values with square rooted values and those performance values are replaced by these normalized values hence we get a normalized matrix

Table 24: Normalized Matrix from Topsis method

Make	fuel-type	Width	Height	horsepower	city	Highway	price
Subaru	Gas	0.315432	0.320823	0.335133	0.274347	0.282782	0.297483
Chevrolet	Gas	0.300009	0.317836	0.233136	0.415945	0.416317	0.299401
Mazda	Gas	0.319413	0.323213	0.330276	0.265497	0.243506	0.301959
Toyota	Gas	0.316427	0.325603	0.301134	0.309747	0.306347	0.310852
Mitsubishi	Gas	0.320408	0.303497	0.330276	0.327446	0.322057	0.313235
Toyota	Gas	0.316427	0.325603	0.301134	0.309747	0.306347	0.310852
Honda	Gas	0.318417	0.314251	0.29142	0.336296	0.329912	0.313816
Nissan	Gas	0.317422	0.325603	0.335133	0.274347	0.290637	0.319629
Dodge	Gas	0.317422	0.303497	0.330276	0.327446	0.322057	0.323872
Plymouth	Gas	0.317422	0.303497	0.330276	0.327446	0.322057	0.323872
Mazda	Gas	0.319413	0.323213	0.330276	0.274347	0.298492	0.354271

Then we multiply the weights obtained through the AHP method with each column of the cell and we get the weighted normalized matrix as the table below:

Table 25: Weighted Normalized Matrix from Topsis method

Make	fuel-type	width	height	Horse	city	highway	price
Subaru	Gas	0.01565	0.010868	0.02413	0.060372	0.045073	0.099135
Chevrolet	Gas	0.014885	0.010767	0.016786	0.091532	0.066357	0.099774
Mazda	Gas	0.015848	0.010949	0.02378	0.058425	0.038812	0.100627
Toyota	Gas	0.0157	0.01103	0.021682	0.068162	0.048829	0.10359
mitsubishi	Gas	0.015897	0.010281	0.02378	0.072057	0.051333	0.104384
Honda	Gas	0.015799	0.010646	0.020983	0.074005	0.052585	0.104578
Nissan	Gas	0.015749	0.01103	0.02413	0.060372	0.046325	0.106515
Dodge	Gas	0.015749	0.010281	0.02378	0.072057	0.051333	0.107929
plymouth	Gas	0.015749	0.010281	0.02378	0.072057	0.051333	0.107929
Mazda	Gas	0.015848	0.010949	0.02378	0.060372	0.047577	0.11806

In next step we calculate the ideal best and the ideal worst values which are o(kl)+, o(kl)-. In case of price the smallest value is the ideal best and the highest value is the ideal worst since it is the non-beneficial point of reference, whereas in other columns the maximum is treated as ideal best and minimum is treated as ideal worst. The Euclidian distance is calculated using o(kl)+=((f(kl)-f(k)^2)0.5 ,differences between each normalized performance value and the ideal best and a total summation is taken then that value is squared ,this process is done for each row .The same procedure is followed for even ideal worst and o(kl)- is obtained . The table below depicts the best and worst ideal results:

Table 27: summary of final priories of make:

make	U(k)	rank
subaru	0.361374	7
chevrolet	0.862721	1
mazda	0.303969	8
toyota	0.411442	6
mitsubishi	0.488928	3
honda	0.520594	2
nissan	0.294555	9
dodge	0.45981	5
plymouth	0.45981	4
mazda	0.218122	10

Table 26: summary of ideal best and ideal worst tables:

O(kl)+	O(kl)-
0.037737	0.021354
0.007446	0.046791
0.043095	0.01882
0.029652	0.020729
0.025164	0.024074
0.023164	0.025154
0.037772	0.015771
0.026135	0.022246
0.026135	0.022246
0.041011	0.011441

Now the performance score is calculated by U(k)=o(kl)-/o(kl)+ +o(kl)- and then they are ranked according to those scores. The final results are given in the table below

IV. RESULTANALYSIS

From these studies we infer that price is the attribute that makes an impact on predicting the car sales values, we also rank Chevrolet brand or make of the car as the leading sales in the car industry with the rank of number one which is followed by Honda after wards. This Rank is being proposed after we perform the "TECHNIQUE FOR ORDER OF PREFERENCE BY SIMILARITY TO IDEAL SOLUTION ". Hence with the help of Analytic hierarchy process, several machine learning algorithms such as linear regression, random forest we have inferred results that are so specific and accurate. The following picture depicts the percentages of the price attribute of each make

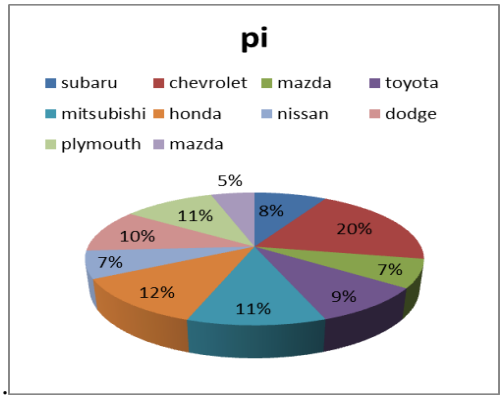

Figure 3.Final results of make after Topsis method

REFERENCES

1. Keivan Kianmehr a, Reda Alhajj a,b,*2008 , Calling communities analysis and identification using machine learning techniques, journal homepage: www.elsevier.com/locate/eswa.
2. Yaya Xie a, Xiu Li a,*, E.W.T. Ngai b, Weiyun Ying c,2008, Customer churn prediction using improved balanced random forests, journal homepage: www.elsevier.com/locate/eswa.
3. Zhen-Yu Chen , Zhi-Ping Fan,2012, Distributed customer behavior prediction using multiplex data: A collaborativeMK-SVM approach, journal homepage: www.elsevier.com/ locate/knosys.
4. Miloˇs Miloˇsevi´ca,b, Nenad ˇZivi´ca,d, Igor Andjelkovi´ca,c,2017, Early churn prediction with personalized targeting in mobile social games. PII: S0957-4174(17)30304-4 DOI: 10.1016/j.eswa.2017.04.056,Reference: ESWA 11289
5. To appear in: Expert Systems With Applications
6. Marko Bohaneca,b,_, Mirjana Kljaji´c Borˇstnarb, Marko Robnik-ˇSikonjac,2016, Explaining machine learning models in sales predictionsPII:S0957-4174(16)30632-7,DOI: 10.1016/j.eswa.2016.11.010,Reference: ESWA 10981
7. To appear in: Expert Systems with Applications
8. J. Burez, D. Van den Poel *,2008, Handling class imbalance in customer churn prediction, journal homepage: www.elsevier.com/locate/eswa.
9. Byeonghwa Park a, Jae Kwon Bae b,,2014, Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data, journal homepage: www.elsevier.com/locate/eswa
10. Srikumar Krishnamoorthy,2015, Linguistic features for review helpfulness prediction, journal homepage: www.elsevier.com/locate/eswa.
11. Flavio Barboza a , ∗, Herbert Kimura b , Edward Altman c,2017,Machine learning models and bankruptcy prediction,http://dx.doi.org/10.1016/j.eswa.2017.04.
12. G Vinodhini ∗, RM Chandrasekaran ,2016, A sampling sentiment mining approach for e-commerce applications, journal homepage: www.elsevier.com/locate/infoproman.
13. Cristina Soguero-Ruiz a,ñ, Francisco-Javier Gimeno-Blanes b, Inmaculada Mora-Jiménez a, María Pilar Martínez-Ruiz c, José-Luis Rojo-Álvarez a,2012, On the differential benchmarking of promotional efficiency with machinelearning modelling (II): Practical applications, journal homepage: www.elsevier.com/locate/eswa.