2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) | 978-1-6654-1460-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICT4SD50815.2021.9396868

A Comprehensive Study of Machine Learning algorithms for Predicting Car Purchase Based on Customers Demands

Anamika Das Mou[1], Protap Kumar Saha[2] Sumiya Akter Nisher[3] and Anirban Saha[4]
Department of Computer Science and Engineering, East West University
Dhaka, Bangladesh[123]
Department of Mechanical and Materials Engineering, Florida International University
Florida, USA[4]
[1]anamikamou9@gmail.com,[2]protap.k.saha@gmail.com, [3]nisher.sumiyaakter@gmail.com, [4]asaha017@fiu.edu

***Abstract*—The automobile industry is one of the prominent industries for the national economy. Day by day car is getting popular for the private transport system. The customer needs review when he wants to buy the right vehicle, especially the car. Because it is a very costly vehicle. There are many conditions and factors matter before buying a new car like spare parts, cylinder volume, headlight and especially price. So deciding everything, it is important for the customer to make the right choice of purchase which can satisfy all the criteria. Our goal is to help the customer to make the right decision whether he will buy a car or not. Therefore we wanted to build a technique for decision making in-car buy system. That's why we propose some well-known algorithms to get better accuracy for a car purchase in our paper. We applied those algorithms in our dataset which contains 50 data. Among them, Support Vector Machine(SVM) gives the best result with 86.7% accuracy of prediction. In this paper, we have also revealed the comparative results using different algorithms precision, recall and F1 score for all data samples.**

***Index Terms*—Supervised Machine Learning, Naive Bayes, Random Forest tree, Support Vector Machine, KNN, Accuracy, Cosine Similarity**

I. Introduction

In this smart era of technology, people like to make those ideas and decisions which are not only for their current benefits as well as price but also for their future advantage. For example, if a person desires to choose a job or if he is planning for where he will stay or planning for a fine vacation, all are important in his life decisions. Because these are needed to consider about the utility that will increase in the future. People always like to make those types of decisions which maximizes the utility [1]. Since the utility is linked with the financial system in our day to day life.

At present, the automobile industry is one of the most important businesses in the world. Though Bangladesh is a small country in South Asia, the demand for automobiles is increasing with each passing day. People are using private vehicles to move one location to another location. The four-wheeled private car is good and flexible among those. A good portion of the economic development of a country depends on the transportation system. Because when the transport system is efficient, it affords economic as well as a social opportunity which gives a positive effect to markets [2]. So earlier than buying a new car, customers prefer to be assured of the money they spend to be worthy. Because to buy a new car, it is the matter of a good amount of money in the perspective of the Bangladeshi economy. Thats why it is important to get the information about cars which are good or bad based on customers experiences, who bought those before. The lifecycle of a modern car depends on so many different parts.

In this paper, we want to predict the probability of buying a car based on price, spare part, customer review, cylinder volume, resale price. Predicting the likelihood or probability of purchase for vehicles is a superb and much-needed problem [3]. We applied four popular algorithms Naive Bayes, SVM, Random Forest Tree and K-nearest neighbour to do the comparison that, which algorithm gives better accuracy for predicting purpose.

The rest of the paper is assembled as follows. Section II of the paper analysis related works. Section III describes the particulars of the proposed method for finding better accuracy. Section IV evaluates the experiment and displays the experiment results phase-wise. Section V winds up the paper in a nutshell and highlights some future work that can be done.

II. Related Work

Some people preferred good parts, some are high or low price with all of their needed features, some are only weak for famous brands of the car only. To select the perfect car is still a difficult task though some parameters like color, comfort, seating capacity, etc are known [2]. Thats why we tried to compare some algorithm for predicting car buying purpose that which one gives better accuracy.

An implementation of Nave Bayes Classification method is proposed by Fitrina et. al [3]. Naive Bayes is known as a simple probabilistic classifier. They applied this method for predicting purchase. They used a dataset on 20 car purchase

978-1-6654-1460-9/21/$31.00 ©2021 IEEE

data and got 75% accuracy. Srivasta et. al [4] applied the powerful learning method, Support Vector Machinen(SVM) on different types of data like Diabetes Data, Heart Data, Satellite Data and Shuttle Data. Those datasets have multi classes. They are also proven the analysis of the comparative consequences the use of divers kernel functions on their paper.

A comparative analysis of machine learning algorithm was proposed by Ragupathy et. al [5]. In their paper, they tried to identify and classify sentiment, conveyed in main text. They have collected their data from social media like Twitter, comments, blog posts, news, status updates etc. They also applied Naive Bayes, Decision Tree, K-Nearest Neighbour and Support Vector Machine classifiers for their comparison purpose. Their goal was to find out the most efficient classification technique and SVM came out with 72.7% which was the best accuracy. Another prediction system using supervised machine learning technique was proposed by Noor et. al [6]. They used multiple linear regression method and predicted vehicle price. They got 98% accuracy on their system. Pal et. al [7] proposed a methodology for predicting used cars costs. In their paper, they used Random Forest classifier to predict the costs of used cars. To train the data, they created a Random Forest with 500 Decision trees. Finally, they got 95.82% as training accuracy and 83.63% as testing accuracy. Pudaruth et. al [8] proposed another methodology for predicting used cars prices. In that paper, he applied multiple linear regression analysis, k-nearest neighbours, Naive Bayes and Decision Tree which were used to make the predictions. Osisanwo F.Y. et. al [9] proposed Supervised machine learning technique. They compared seven different Supervised learning algorithms and described those. They also found out the most effective classification algorithm established on dataset.

A different work on car purchase was proposed by R.Busse et. al [10]. In their paper, they prioritized the psychological effect of weather. They applied projection bias and salience as two major psychological mechanism.

A new defect classification technique was proposed by veni et. al [11] to predict the class label of ”severity” tuple. Those data tuples were described by various attribute like Phase attribute, Defect, Phase Defected, Impact and Weight. They applied Naive Bayes classifier for prediction purpose. Jayakameswaraiah et. al [2] developed a data mining system to analyze cars. They proposed TkNN clustering algorithm to predict the right car. They also shown the comparison of KNN and their proposed novel TkNN clustering. Another car price prediction technique was proposed by Gegic et. al [12] where they used three machine learning techniques. They got 92.38% accuracy on combination of all ML methods.

Another medical work was proposed by Jabbar et. al [13] to predict heart disease in diagnosis system. They used K-nearest neighbour(KNN) algorithm to predict it. The algorithm performs tremendously with 100% accuracy. Peerun et. al [14] presented a technique to predict rice of second-hand cars. In their paper, they used Artificial Neural Networks. They applied it on dataset of 200 records cars and compared different kinds of machine learning algorithm. Yuan et. al [15] offered a study on prediction. He tried to predict the car sales based on some web search records.

In spite of these well-known works, there also exist some more challenging works. As a result, we focus on the comparison of four types of well-known machine learning algorithms and try to find out which algorithm gives the best accuracy for our dataset.

III. Methodology

The aim of this research is to analyze the accuracy of different predictive algorithms that can predict the probability of purchasing a car. Figure 1. depicts the workflow of the proposed methodology.

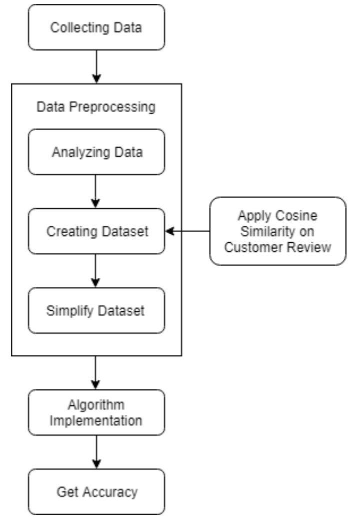

Fig. 1: Workflow of Whole Process

A. Simplifying Dataset:

In this phase, we again do a update of our dataset. We assume some numeric value for our data. We assume Expensive as 3, Affordable as 2 and Normal as 1 numeric value for our Price attribute. Same as Low(1), Medium(2), High(3) for Spare Part and Cylinder Volume, again same as Expensive(3), Affordable(2), Normal(1) as Resale Price attribute and Yes(1),No(0) for Buy attribute.

TABLE I: A Portion of Simplified Dataset

Price	Spare Part	Cylinder Volume	Resale Price	Car’s Review	Buy
3	1	2	3	1	1
2	2	2	2	1	1
1	3	1	2	0	0
3	1	2	3	1	1
2	2	3	1	0	0

We also apply Cosine Similarity for mining the text as customer review in our Cars Review attribute. Applying Cosine Similarity, we found 1 for positive review and 0 for negative review. For finding cosine similarity, we have a dataset which contains customer review, according to the review their is an output which was positive or negative. Then according to the dataset, we measure our collected customer's review positive or negative. Cosine similarity measure two sentence according to the function,

$$\text{Cosine Similarity} = \frac{A.B}{|A|.|B|}$$

The Cosine Similarity measure the value between 0 to 1. After measuring two sentences Cosine Similarity, if the value is greater than or equal to 0.5, then the review is positive. If the value is smaller than 0.5, then the review is negative. We have assumed the threshold value, for better performance of the similarity measurement.

B. algorithm Implementation:

To predict something, first of all, we have to learn our machine. Those machines can learn with the proper algorithm. There are three types of machine learning algorithms. They are supervised learning, Unsupervised learning, Semi-supervised learning. Among those, we choose supervised learning algorithms. Those are Nave Bayes, Support Vector Machine (SVM), K-nearest neighbor algorithm(KNN) and Random Forest tree.

1) Naive Bayes: Naive Bayes is known as an arithmetical classification method based on the Bayes Theorem for classification problems. It is a simple learning algorithm which is not only known for its easiness but also its effectiveness. It is regarded as nave because of its assumption shortens calculation[1].The overall formula can be written as, .

$$P(c/f) = \frac{P(c) * P(f/c)}{P(f)} \qquad (1)$$

here, c = class, f = features

- P(c/f) : Posterior Probability
- P(c) : Class Prior Probability
- P(f/c) : Likelihood
- P(f) : Predictor Prior Probability

2) Support Vector Machine(SVM): The SVM classifier is known as a stirring algorithm and its concepts are comparatively humble. It is also known as a discriminative classifier. It bids very high accuracy associated with other well-known classifiers[2].

There are many applications of SVM such as handwriting recognition, classification of emails in a mail account, human or other animal face detection, etc[3].

3) k-nearest neighbor algorithm(KNN): The KNN algorithm deals with similar things which are near or close to each other. Thats mean, it assumes to compute all those similar things which occur in close nearness area[4].K-NN is also known as a lazy learner. Because it only memorizes the training dataset easily, doesnt want to learn a discriminative function system from the training data[5].

$$\text{y} = \frac{1}{x}\sum_{1}^{i} y_i \qquad (2)$$

where y_i is known as the ith case of the examples of every sample and y is the prediction (outcome) of the query point.

4) Random Forest tree: Its name implies what is Random Forest is. A large range of individual decision trees that work as a group, makes the Random Forest. Each particular tree of a Random Forest dribbles out a class prediction in the dataset. Most of the votes owner class will be our systems prediction. So as the number of tree in the forest will be, it will give the high accuracy as well. Random Forest allows a large amount of number which is weak or weakly-correlated classifiers to build a strong enough classifier[6].

Our Random Forest model system deals with ID3 algorithm to train and uses the Gini index to measure it. Gini index is used for calculating the uprightness of split criteria. The Gini impurity measure can be written as,

$$\text{Gini(Xn)} = \sum kpnk\text{(1-pnk)} \qquad (3)$$

where pnk is the fraction of times. k is an element of class occurs in a split. Xn is an element of set X.

5) Get Accuracy: In this phase, we apply above mentioned four algorithms for our dataset. We select one algorithm as our desired algorithm which provides best result for the dataset.

IV. Experiments and Results

A. Implementation

For implementing car prediction algorithm, we have used anaconda[7], which is an environment of python 3.7 with a lot of packages of machine learning. Our processor is Intel core i3 with clock rate 2.4GHz, RAM 4GB. We used Windows 10(64 bit)as operating system.

B. Evaluation Dataset

We have collected our data from different shops in Bangladesh and also from social media. After successfully creating our dataset, we evaluate algorithms by splitting dataset. We use 70% of data as training and 30% of data as testing. A simple statistics of dataset given in Table **??**,

TABLE II: Simple Statistics of Dataset

Attributes	**Number of Count**
Data Collected	50
Training Data	35
Testing Data	15

C. Evaluation Measurement

To evaluate the results, we have used precision-recall, execution time, accuracy measurement of the algorithms.

[1] https://www.datacamp.com/community/tutorials/naive-bayes

[2] https://towardsdatascience.com

[3] https://www.datacamp.com/community/tutorials/svm

[4] https://towardsdatascience.com/machine-learning-basics

[5] http://www.statsoft.com/Textbook/k-Nearest-Neighbors

[6] https://towardsdatascience.com/understanding-random-forest

[7] https://www.anaconda.com/

1) Precision, recall and f1 score: Precision is a ratio of accurately predicted positive observations. Recall is a ratio of accurately predicted positive observations to all observations in actual class-yes. F1 score is weighted average of the precision and recall. This scores evaluate how a model has performed. Precision, recall and f1 score of several algorithms are given in Table iv,

TABLE III: Precision, Recall and F1 Score of Mentioned algorithms

algortihm	Precision	Recall	F1 Score
Random Forest	0.60	0.60	0.60
KNN	0.75	0.73	0.73
Naive Bayes	0.39	0.43	0.41
SVM	0.89	0.87	0.86

2) Accuracy of several algorithms: Accuracy is a measurement of how a model predicts correctly to the total number of input samples. In our proposed method, all algorithms have split the dataset into 70% training and 30% testing. A simple equation of accuracy calculation is given below,

$$Accuracy = \frac{No.of correct classification}{No.of total input}$$

Several algorihms accuracy comparison are given in figure 2,

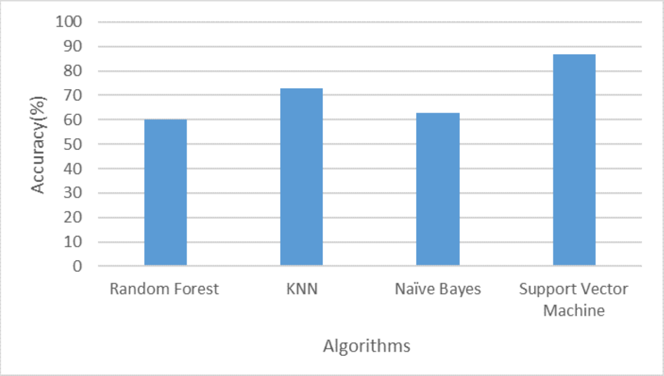

Fig. 2: Accuracy of Several algorithms

From figure-2, we can see Support Vector Machine gives highest accuracy(87.6%) than Random Forest, K-nearest Neighbour(KNN) and Naive Bayes. That means Support Vector Machine can classify approximately 44 car purchase data of 50 dataset.

3) Comparison with other Methods: There are many study about car price prediction from many years. Several study use several machine learning techniques. A methodology for predicting purchase used Naive Bayes algorithm and get 75% accuracy [3]. Another work for predicting used cars prices used Random Forest and get test accuracy 83.63% [7]
On the other hand, our proposed method used Cosine Distance for review analysis and after that using Support Vector Machine got 86.7% accuracy. There are some obstacles determined in different classifiers like Naive Bayes classifier cannot cope with more amounts of data with ease. A simple comparison shown in figure 3,

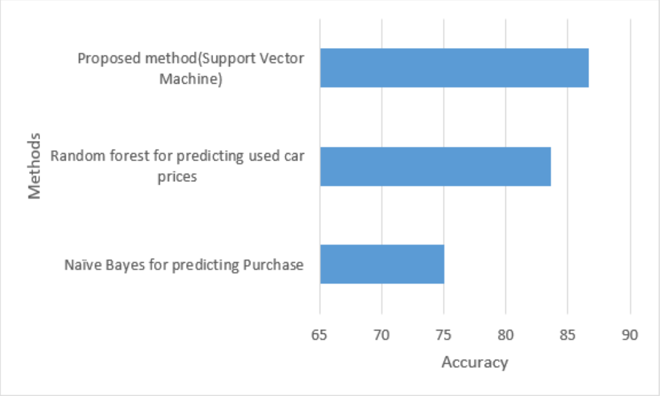

Fig. 3: Comparison With Other Methods

V. Conclusion

In our study, we mainly focused on Customer's review and we have used Cosine Similarity to analyze customer's review. Then we applied several algorithms on our dataset. We have compared those algorithms according to their accuracy. Support Vector Machine has given the highest accuracy among all the algorithms.

VI. Limitations And Future Work

We have used 5 features for predicting the final output. in future we will collect more features for prediction and increase our dataset. We will use a more efficient technique to get better accuracy. We will intend to apply some more advanced machine learning techniques like Fuzzy logic, Decision Tree, Artificial Neural Network, Ordinary Least Squares Regression (OLSR), Fuzzy logic etc as our future work. This work can be enlarged by choosing more features for classification also.

References

[1] J. C. Pope and J. Silva-Risso, "The psychological effect of weather on car purchases* meghan r. busse devin g. pope," *The Quarterly Journal of Economics*, vol. 1, no. 44, p. 44, 2014.

[2] M. Jayakameswaraiah and S. Ramakrishna, "Development of data mining system to analyze cars using tknn clustering algorithm," *International Journal of Advanced Research in Computer Engineering Technology*, vol. 3, no. 7, 2014.

[3] F. Harahap, A. Y. N. Harahap, E. Ekadiansyah, R. N. Sari, R. Adawiyah, and C. B. Harahap, "Implementation of naïve bayes classification method for predicting purchase," in *2018 6th International Conference on Cyber and IT Service Management (CITSM)*. IEEE, 2018, pp. 1–5.

[4] K. S. Durgesh and B. Lekha, "Data classification using support vector machine," *Journal of theoretical and applied information technology*, vol. 12, no. 1, pp. 1–7, 2010.

[5] R. Ragupathy and L. Phaneendra Maguluri, "Comparative analysis of machine learning algorithms on social media test," *International Journal of Engineering and Technology(UAE)*, vol. 7, pp. 284–290, 03 2018.

[6] K. Noor and S. Jan, "Vehicle price prediction system using machine learning techniques," *International Journal of Computer Applications*, vol. 167, no. 9, pp. 27–31, 2017.

[7] N. Pal, P. Arora, P. Kohli, D. Sundararaman, and S. S. Palakurthy, "How much is my car worth? a methodology for predicting used cars prices using random forest," in *Future of Information and Communication Conference*. Springer, 2018, pp. 413–422.

[8] S. Pudaruth, "Predicting the price of used cars using machine learning techniques," *Int. J. Inf. Comput. Technol*, vol. 4, no. 7, pp. 753–764, 2014.

[9] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi, “Supervised machine learning algorithms: classification and comparison,” *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, 2017.

[10] M. R. Busse, D. G. Pope, J. C. Pope, and J. Silva-Risso, “The psychological effect of weather on car purchases,” *The Quarterly Journal of Economics*, vol. 130, no. 1, pp. 371–414, 2015.

[11] S. Veni and A. Srinivasan, “Defect classification using naïve bayes classification,” *Interbational Journal of Applied Engineering Research*, vol. 12, no. 22, pp. 12 693–12 700, 2017.

[12] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, “Car price prediction using machine learning techniques,” 2019.

[13] M. Jabbar, “Prediction of heart disease using k-nearest neighbor and particle swarm optimization,” *Biomed. Res*, vol. 28, no. 9, pp. 4154–4158, 2017.

[14] M. C. Sorkun, “Secondhand car price estimation using artificial neural network.”

[15] Q. Yuan, Y. Liu, G. Peng, and B. Lv, “A prediction study on the car sales based on web search data,” in *The International Conference on E-Business and E-Government (Index by EI)*, 2011, p. 5.