www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

# Research on Machine Learning Algorithms for Demand-Based Automobile Purchase Predictions

[1] S. Akhila, [2] Ch. Sowmya,

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

## Abstract

One of the most visible sectors of the American economy is the automotive sector, which produces automobiles. The use of cars for private transportation is on the rise. Before purchasing any vehicle, but notably a car, the buyer should do his research. Reason being, it's a very expensive car. Parts availability, engine displacement, lighting, and most importantly, budget are just a few of the many considerations that should go into the purchase of a new vehicle. The onus is on the buyer to make an informed decision that satisfies all requirements before making any purchases. Our objective is to assist the client in making a well-informed choice about the purchase of a vehicle. Thus, we aimed to develop a method for the in-car purchase system's decision-making process. For this reason, our research suggests a few popular algorithms that might improve the accuracy of automobile buying. Our dataset has 50 data points, and we used those techniques on them. Support Vector Machine (SVM) outperforms the others with an 86.7% prediction accuracy rate. Also included in this study are the findings of comparing all data samples using various algorithms in terms of accuracy, recall, and F1 score.

## Index Terms

Machine Learning using Supervision, Naive Bayes, Support Vector Machine, Cosine Similarity, KNN, and Random Forest tree

## I. INTRODUCTION

People nowadays want to think and act in ways that benefit them in the long run, rather than only in the short term. This is especially true in this technologically advanced age. Examples of significant life choices include picking a career path, deciding where to live, and organizing a vacation. Reason being, they are essential for thinking about future utility increases. The option that maximizes utility is the one that people choose to make in every situation [1]. Due to the fact that the usefulness is associated with the financial system in our day-to-day lives. There are few industries as consequential as the automotive sector right now. Despite Bangladesh's tiny size, the demand for vehicles is growing daily across South Asia. People are getting from one place to another using their own cars. Among them, the four-wheeled private vehicle is both reliable and versatile. Much of a nation's economic growth is dependent on transportation network. Because markets benefit from an efficient transportation infrastructure because it creates social and economic opportunities [2]. Therefore, consumers want guarantees that their money will be well-spent when they purchase a new automobile. Considering the state of the Bangladeshi economy, purchasing a new vehicle is an expensive affair. That is why it is essential to gather information on the quality of automobiles based on the experiences of previous buyers. A contemporary car's lifespan is dependent on a myriad of factors. We aim to forecast the likelihood of purchasing a vehicle in this study by analyzing factors such as price, spare part availability, customer reviews, cylinder volume, and resale price. An excellent and much needed issue is the prediction of the possibility or probability of purchase for automobiles [3]. To determine which of four well-known algorithms—Naive Bayes, Support Vector Machines, Random Forest Tree, and K-Nearest Neighbor—produces the most accurate predictions, we put them through their paces. Here is the remaining portion of the document. The paper's second section examines relevant literature. Section III delves into the specifics of the suggested approach to achieving improved accuracy. Part IV provides an analysis of the experiment and a sequential presentation of the findings. In Section V, we

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

summarize the material and point out several areas where further research may be beneficial.

## II. RELATED WORK

others individuals want high-quality components, others like cheap parts with all the characteristics they require, and some only like parts for well-known automobile manufacturers. Although some criteria like as color, comfort, seating capacity, etc. are recognized, choosing the ideal automobile remains a challenging endeavor [2]. That is why we set out to determine which algorithm provides the most accurate predictions for the purpose of purchasing an automobile. Fitrina et al. [3] suggests using the Naive Bayes Classification technique. One example of a probabilistic classifier is Naive Bayes. They used this strategy to forecast sales. Twenty vehicle purchase datasets were used. found 75% accuracy in the data. Data from several sources, including satellites, hearts, diabetes, and shuttles, were subjected to the powerful learning approach Support Vector Machinen (SVM) by Srivasta et al. [4]. There are several classifications in those datasets. Additionally, they have shown in their work that they analyzed the comparative impacts of using various kernel functions. According to Ragupathy et al. [5], it might be beneficial to compare different machine learning algorithms. They attempted to categorize the feelings expressed in the major text of their article. Twitter, blog comments, news articles, status updates, and other social media have all contributed to their data collection. When comparing, they also used decision trees, support vector machines, naive bayes, and K-nearest neighbours. By comparing several classification techniques, they were able to determine that SVM had the highest accuracy at 72.7%. Using a supervised machine learning approach, Noor et al. [6] presented an additional prediction method. They forecasted the price of the automobile using multiple linear regression. Their approach had a 98% success rate. A system for forecasting the costs of secondhand automobiles was presented by Pal et. al. [7]. Predicting the prices of secondhand vehicles was done using a Random Forest classifier in their article. They built a Random Forest using 500 Decision trees to train the data. At the end of the day, their training accuracy was 95.82% and their testing accuracy was 83.63%. One alternative approach to used-car price prediction was put forward by Pudaruth et al. [8]. He employed k-nearest neighbours, Naive Bayes, Decision Tree, and multiple linear regression analysis to generate predictions in that article. The supervised machine learning method was suggested by Osisanwo F.Y. et al. [9]. They detailed and contrasted seven distinct supervised learning methods. Additionally, they discovered the best classification system that has been tested on the dataset. R. Busse et al. [10] offered an alternative study on the topic of automobile acquisition. The importance of weather's psychological impact was highlighted in their research. They used two important psychological mechanisms, projection bias and salience. Venenet al. [11] presented a novel method for defect classification that uses class label prediction for the "severity" tuple. Some of the attributes used to characterize these data tuples were Phase, Defect, Impact, and Weight. In order to make predictions, they used a Naive Bayes classifier. In order to study automobiles, Jayakameswaraiah et al. [2] created a data mining system. In order to forecast the correct vehicle, they suggested the TkNN clustering technique. Also shown was a comparison between KNN and the new TkNN clustering algorithm they had developed. Gegic et al. [12] presented an additional method for predicting vehicle prices using three machine learning algorithms. The combined accuracy of all ML algorithms was 92.38%. To further improve the diagnostic system's ability to foretell cardiac illness, Jabbar et al. [13] suggested more medical work. To make the prediction, they used the Knearest Neighbor (KNN) method. With perfect precision, the algorithm works wonders. In order to forecast the value of used vehicles, Peerun et al. [14] devised an approach. Artificial Neural Networks were used in their publication. They used it to evaluate several machine learning algorithms on a dataset of 200 recorded vehicles. With regard to forecasting, Yuan et al. [15] provided research. Based on certain search information from the web, he attempted to forecast the sales of cars. Despite these famous works, there are additional works that are more difficult. Consequently, we want to determine which of four popular machine learning algorithms provides the highest level of accuracy for our dataset by comparing their performance.

## III. METHODOLOGY

Examining the efficacy of various prediction algorithms that may foretell the likelihood of buying a vehicle is the primary goal of this study. The suggested methodology's process is shown in Figure 1.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

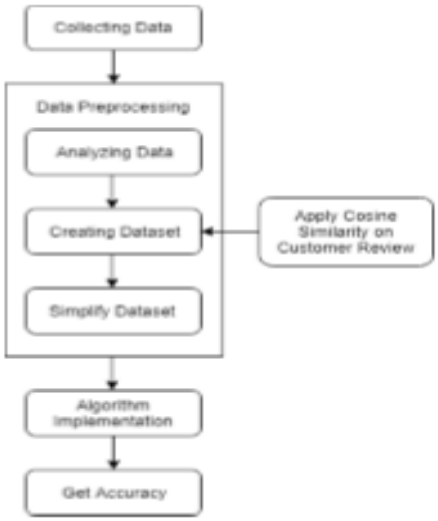


Fig. 1: Workflow of Whole Process

Step one, simplifying the dataset, involves updating the dataset once again. We take it as read that our data has a numerical value. The values we're working with for the Price attribute are Expensive(3), Affordable(2), and Normal (1). Resale Price: Expensive(3), Affordable(2), Normal(1), and Buy: Yes(1), No(0), are identical to the Low(1), Medium(2), and High(3) values for Spare Part and Cylinder Volume, respectively.

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 |

Additionally, we use Cosine Similarity to extract customer reviews from the language in our Cars Review feature. We discovered 1 favorable review and 0 negative review using Cosine Similarity. In order to determine the cosine similarity, we have a dataset that includes customer reviews. Each review indicates whether the output was favorable or bad. Then, we determine whether the client reviews we gathered were good or negative based on the data. Two sentences are compared using the cosine similarity function,

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

The range of values for the Cosine Similarity metric is from zero to one. The review is considered favorable if the value of the two-sentence Cosine Similarity is equal to or higher than 0.5. The review is considered unfavorable if the value is less than 0.5. For a more accurate similarity assessment, we have presumptively used the threshold value. The first step in making a prediction is to train your computer, which brings us to point B. method implementation. The right algorithm can teach such machines new things. Three distinct categories of algorithms are used in machine learning. There are three types of learning: supervised, unsupervised, and semi-supervised. We choose supervised learning techniques from that group. These include Random Forest, K-nearest neighbor algorithm (KNN), Naive Bayes, and Support Vector Machine (SVM). 1) Bayes Naive: One mathematical approach to classification issues is Naive Bayes, which is based on the Bayes Theorem. Not only is it well-known for being easy to use, but it is also a straightforward learning algorithm. The belief that it reduces calculating time is what gives it its "naval" reputation. One way to express the whole formula is as,

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

here, c = class, f = features

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

called a stirring algorithm and using quite simple ideas. A discriminative classifier is another name for it. When compared to other popular classifiers, it offers very high accuracy. Among SVM's many uses are the following: handwriting recognition, email account classification, facial identification (of humans or other animals), etc3. Thirdly, the K-Nearest Neighbor (KNN) program: Things that are close together or comparable are what the KNN algorithm focuses on. It takes it as a given that it will calculate all the occurrences of comparable objects in the immediate vicinity. 4. The term "lazy learner" describes K-NN as well. This is due to the fact that it learns the training dataset quickly and refuses to use it to construct a discriminative function system.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

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

that is, yi is the ith instance of each sample, and y is the predicted result for the query point. fourthly, a random forest tree You can tell what Random Forest is by looking at its name. The Random Forest is a collection of several separate decision trees that cooperate to get a final result. A Random Forest's individual trees each drizzle forth a dataset class prediction. Our algorithm predicts that the majority of votes will come from the owning class. Therefore, it will provide very accurate results regardless of the quantity of trees in the forest. A large number of weak or weakly-correlated classifiers may be used to generate a strong enough classifier using Random Forest. 6. Using the Gini index as a metric, our Random Forest model system trains using the ID3 method. The uprightness of split criteria may be determined using the Gini index. Here is another way to express the Gini impurity measure:

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

where pnk represents the frequency. k is a class element that happens in a split. Each Xn belongs to the set X. In the fifth and final step, "Get Accuracy," we apply the four algorithms described earlier to our dataset. Once we find an algorithm that works well with the dataset, we choose it as our preferred algorithm.

# IV. EXPERIMENTS AND RESULTS

Section A: Execution Anaconda7, a Python 3.7 environment with several machine learning libraries, was utilized to develop the automobile prediction algorithm. Our CPU is a 2.4 GHz Intel Core i3, and we have 4 GB of RAM. Window 10 (64 bit) was the OS that we used. Section B. Assessment Files We gathered our information from several Bangladeshi stores as well as social media platforms. We test algorithms by dividing the dataset once we've successfully created it. Our data is divided as follows: 70% for training and 30% for testing. Table?? provides some basic statistics on the dataset,

TABLE II: Simple Statistics of Dataset

| Attributes | Number of Count |
|---|---|
| Data Collected | 50 |
| Training Data | 35 |
| Testing Data | 15 |

Assessment C. Measuring We have measured the algorithms' accuracy, execution time, and precision-recall to assess their performance.

F1 score, recall, and precision: The ratio of correctly predicted positive observations is called precision. For a class-yes prediction, recall is the ratio of all positive observations to the number of successfully predicted ones. A weighted average of recall and accuracy is the F1 score. These ratings measure the effectiveness of a model. Table iv lists the algorithms' f1 scores, recalls, and precisions.

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) Algorithm accuracy: Model accuracy is defined as the proportion of input samples that the model properly predicts. All of the algorithms in our suggested system divide the dataset into a training set of 70% and a testing set of 30%. The following is a basic equation for calculating accuracy:

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

Several algorihms accuracy comparison are given in figure 2,

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

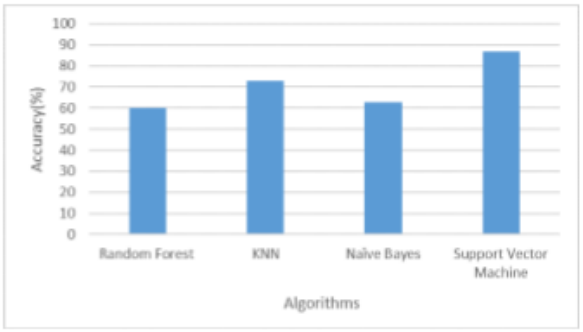


Fig. 2: Accuracy of Several algorithms

Support Vector Machine outperforms Random Forest, K-nearest Neighbor (KNN), and Naive Bayes in terms of accuracy (87.6%), as shown in figure 2. This indicates that out of 50 datasets, Support Vector Machine is able to categorize 44 instances of vehicle purchase data. Step Three: Examining Alternative Approaches: Predicting the price of an automobile has been the subject of several studies throughout the years. Multiple studies use various machine learning approaches. A 75% success rate was achieved by a purchase prediction system that made use of the Naive Bayes algorithm [3]. Another study that used Random Forest to forecast the cost of secondhand automobiles had a test accuracy of 83.63%. [7] A review study employing Cosine Distance yielded an accuracy of 86.7% when combined with Support Vector Machine, according to our suggested strategy. various classifiers have various problems; for example, a Naive Bayes classifier will struggle to handle larger datasets. Figure 3 displays a straightforward comparison.

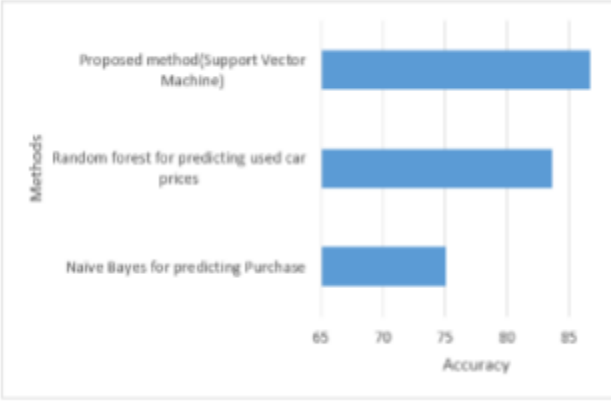


Fig. 3: Comparison With Other Methods

# V. CONCLUSION

Customer reviews were the primary focus of our research, and we analyzed them using Cosine Similarity. After then, our dataset was subjected to a number of algorithms. We have evaluated those algorithms based on how accurate they are. The most accurate algorithm out of all of them is Support Vector Machine.

# VI. LIMITATIONS AND FUTURE WORK

The ultimate output was predicted using five characteristics. We want to expand our dataset and gather more characteristics for prediction in the near future. If we want better results, we need to employ a more efficient method. In our next projects, we want to use more sophisticated machine learning methods such as Fuzzy logic, Decision Tree, Artificial Neural Network, Ordinary Least Squares Regression (OLSR), and more. Adding additional features for classification would also make this effort bigger.

# 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 na¨ıve 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 psychologicaleffectofweatheroncarpurchases,"TheQuarterlyJournal of Economics, vol. 130, no. 1, pp. 371–414, 2015.

[11] S. Veni and A. Srinivasan, "Defect classification using na¨ıve bayes classification," Interbational Journal of Applied Engineering Research, vol. 12, no. 22, pp. 12693–12700, 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.