ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

Predicting Which Cars Will Be Popular: A Machine Learning Approach

[1] V. Vasudha, [2] T. Bharghavi

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

Abstract

It is widely anticipated that in the near future, machines will be able to mimic human reactions and thought processes. Deep Learning is very crucial to the emergence of AI, ML, and Knowledge Engineering. This article presents a solution to a real-world issue of predicting a vehicle company's popularity using machine learning methodologies. The topic is characterized as a regression or classification challenge. categorized or labelled. Additionally, it investigates methods for deducing a system's function that may characterize an underlying structure from unlabeled data. Clustering is a method of learning without supervision.

Keywords—Machine Learning, Regression, Classification, Supervised Machine Learning, Logistic Regression, KNN, Random Forest.

INTRODUCTION

Technology has a significant influence on our daily life in this day and age. Emerging technologies such as artificial intelligence [6], knowledge engineering, machine learning, deep learning [4][5], and natural language processing [7][8] are key to today's most prominent initiatives. The goal of artificial intelligence research is to design computers with cognitive abilities and emotional regulation capabilities comparable to those of humans. When it comes to AI, machine learning is fundamental since it gives AI the capacity to learn and improve itself. Building algorithms that can autonomously choose data and learn from it is the main goal of this approach. When it came to forecasting a product's future success or failure, statisticians and engineers used to collaborate. Because of this procedure, the development and release of the product were postponed. One of the biggest obstacles is keeping up with the product as data and technology evolve. This was facilitated and accelerated by machine learning. Many different types of machine learning algorithms fall into one of four main categories: One kind of learning algorithm is supervised learning, which begins with analysis of a given training dataset and then gives a function to predict future output values [7, 9, 10]. In order to foretell what's to come, this algorithm may be trained on fresh data with labels and applied to previously learnt data. Using a training dataset, unsupervised learning algorithms reveal which ones aren't • Semi-supervised learning combines elements of both unsupervised and supervised learning [6] [11]. Most of the data used by these algorithms is unlabeled, while a tiny portion is labelled. The Reinforcement Algorithm [12] interacts with the environment via actions and error discovery. Machines and software agents may use it to learn how to behave optimally in a given situation, which in turn boosts their performance. Some examples of supervised learning tasks are classification and regression. Classification relies on values derived from observation to form conclusions. In this task, we use a mapping function f on input variables x to approximate a discrete output variable y. Classification typically produces discrete output, but it has the potential to provide continuous probability for each class label. A real or continuous value is used as the output variable in a regression issue. This issue involves approximating a continuous output variable y from a set of input variables x using a mapping function f. In most cases, regression produces continuous output; however, with integer-based class labels, discrete output is also possible. A multivariate regression issue is one that involves many output variables. This article will center on an issue selected from hackerrank in which a vehicle manufacturer is attempting to introduce a new model that incorporates some popular elements from their current lineup. Using a machine learning method, we can forecast how popular it will be. This issue falls within the purview of supervised learning and may be characterized as a regression problem, more specifically a multivariate regression problem. Because of this, a number of supervised learning techniques will be used to make this forecast.

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

RELATED WORKS

To forecast 463 S&P 500 stocks, the author of the article "Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks[1]" used several classification techniques, including logistic regression, gradient boosted trees, artificial neural networks, and random forests. The author has run a battery of tests using these classification algorithms to investigate the stocks' predictability. Unfortunately, the author's desired outcome was not achieved while attempting to forecast future prices using the existing prior data. But they were effective in demonstrating the recent dramatic increase in the predictability of Asian and European indices. The author proposes a novel method for handling missing weather data in the article "Performance evaluation of predictive models for missing data imputation in weather data[2]" by comparing five different approaches—linear regression, support vector machine (SVM), random forest, KNN implementation, and kernel ridge—using the NCDC dataset. The two steps they used to deal with the dataset's missing values were (1) deleting the whole row that had the missing value and (2) impute the missing data. To deal with the missing data, they ran both procedures and compared the results. To predict the Amazon EC2 Spot Price one week and one month in advance, the author of the article "Amazon EC2 Spot Price Prediction using Regression Random Forests [3]" suggests a model based on Regression Random Forests (RRFs). In addition to predicting the execution cost and suggesting when the user should bid to reduce execution cost, this prediction model would be useful for planning when to buy the spot instance.

ALGORITHMS

K-Nearest Neighbor (KNN) [13]: KNN has a second strength; it skips the training phase altogether, or at least has a short and painless one. What this implies is that KNN is not good at generalizing from training data, and that testing is where you should be directing your attention. This is why KNN has a reputation for being a leggy algorithm. Working with KNN—Presuming The data set is an NXP dimensional matrix. P stands for situations, which may be defined as. At least one characteristic (N) is present in every scenario (={,……., }). In each case, the output values are represented by a vector O, where o={ ….. }. What to do: 1. Via a loop that repeats X times, the output values for query scenario q of the X closest neighbors are saved in vector r = { …… }. a. The iteration in the domain {1… P} is continuing, symbolized by I in the following scenarios derived from the dataset. b. If not, then set Next up are t and f. c) Repeat as necessary until the dataset is exhausted. d. Keepingt in vector c and f in vector r.

$$\bar{r} = \frac{1}{X}\sum_{i=1}^{X} r_i \qquad (1)$$

Arithmetic mean is calculated for r-

When it comes to making predictions, logistic regression is a good choice [14]. Logistic regression is used when the dependent variable is a binary one. This approach is useful for describing data and explaining relationships between independent and dependent variables. This statistical approach uses dependent variables that are binary, with values as low as 1 as high as 1. The goal of this approach is to find the best-fitting model by explaining the relationship between independent variables and binary attributes of interest. In order to forecast a logistic transformation, logistic regression generates confidences of a formula:

$$\text{logit(p)} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k \qquad (2)$$

If the feature of interest is present, the probability is denoted by p. The logged odds define the logit transformation.

$$\text{Odds} = \frac{p}{(1-p)} = \frac{\textit{Probability of presence of characteristic}}{\textit{Probability of absence of characteristic}} \qquad (3)$$

$$\text{And} \qquad \text{logit (p)} = \ln\left(\frac{p}{1-p}\right) \qquad (4)$$

Estimation in logistic regression is achieved by selecting parameters that optimize the chance of witnessing the sample values. Block C. Random Forest: The goal of random forests, a subset of supervised classification algorithms, is to generate a forest with a certain degree of randomness. The findings are more accurate when there are more trees. You may use random forest to conduct regressions as well as classifications. Random forest classifiers are able to model for categorical data and can manage missing inputs. There are two steps to how Random Forest works: 1. Making a Random Forest is the first step. a. From the set of m characteristics, choose K at random. b. Using the optimal split point, determine which of the 'K' characteristics belongs to node 'd'. c.

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

The node is divided into two younger nodes. d. Continue steps a through c until you achieve a total of 1 node. e. To make n trees, just repeat steps a through d n times. So, a forest is established. 2. The next step is to use a random forest classifier to produce predictions: a. The testing characteristics and rules of each randomly formed decision tree are used to forecast and record an outcome. b. For every target that is predicted, votes are computed.

The final forecast is the one with the most votes. D. A V-Machine for Support [17] Among the supervised machine learning algorithms used for classification and regression issues is Support Vector Machine, or SVM for short [18]. With successful training data classification and improved generalizability to new data, support vector machines (SVMs) aim to maximize the margin of the training data by locating the ideal separation hyper plane. Section IV: Information on the Experiment and Its Numerical Models An information-rich.csv file contains two data sets: 1. Train.csv — This file contains information on each automobile and is used as a training dataset. Using variables like purchase price, operating cost, number of doors, number of seats, and more. Here are a few of the characteristics: a. buying_price: The purchasing price of the vehicles is described by the buying_price property. The range is [1.....4], with 1 being the lowest price and 4 being the highest. b. maintenance_cost: You may specify the automobiles' maintenance cost using the maintenance_cost attribute. The maintenance cost goes from 1 (the lowest) to 4 (the highest), with 1 being the lowest and 4 the highest. c. number of doors: This property describes the number of doors in the automobile. It takes values between 2 and 5, with each value representing the number of doors in the car. d. number_of_seats: This characteristic shows how many seats are in the automobile; it may take on the values [2, 4, 5] to indicate the number of seats. e. baggage boot size: Set to a value between 1 and 3, this characteristic indicates the size of the luggage boot. Value 1 represents the lowest boot size and 3 represents the greatest. f. Safety rating: What constitutes a car's safety rating is described by the safety rating feature. It may take on values between 1 and 3, with 3 indicating very high levels of safety. g. popularity: An automobile's level of popularity may be expressed using the popularity characteristic. On a scale from 1 (the worst automobile ever) to 4 (the finest car ever), its values vary from 1 (the worst) to 4 (the best). The experiment was carried out using the Python programming language. To resolve the issue, we used the following Python libraries: pandas, numpy, matplotlib, seaborn, and sklearn. Figure 1 displays the sample of training data. The training data format is shown in figure 2. Figure 3 provides a concise overview of the training data.

	buying_price	maintainence_cost	number_of_doors	number_of_seats	luggage_boot_size	safety_rating	popularity
0	3	2	4	2	2	2	1
1	3	2	2	5	2	1	1
2	1	4	2	5	1	3	1
3	4	4	2	2	1	2	1
4	3	3	3	4	3	3	2

Fig 1: Training Data

```
In [5]: train.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1628 entries, 0 to 1627
Data columns (total 7 columns):
buying_price          1628 non-null int64
maintainence_cost     1628 non-null int64
number_of_doors       1628 non-null int64
number_of_seats       1628 non-null int64
luggage_boot_size     1628 non-null int64
safety_rating         1628 non-null int64
popularity            1628 non-null int64
dtypes: int64(7)
memory usage: 89.1 KB
```

Fig 2: Training Data Schema

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

```
In [6]: train.describe()
Out[6]:
```

	buying_price	maintainence_cost	number_of_doors	number_of_seats	luggage_boot_size	safety_rating	popularity
count	1628.000000	1628.000000	1628.000000	1628.000000	1628.000000	1628.000000	1628.000000
mean	2.532555	2.528256	3.493857	3.633292	1.987101	1.977887	1.348280
std	1.109626	1.116920	1.120557	1.257815	0.816520	0.819704	0.654786
min	1.000000	1.000000	2.000000	2.000000	1.000000	1.000000	1.000000
25%	2.000000	2.000000	2.000000	2.000000	1.000000	1.000000	1.000000
50%	3.000000	3.000000	3.000000	4.000000	2.000000	2.000000	1.000000
75%	4.000000	4.000000	4.250000	5.000000	3.000000	3.000000	2.000000
max	4.000000	4.000000	5.000000	5.000000	3.000000	3.000000	4.000000

Fig 3: Training Data Description Training data visualization :

Figure 4 is a bar chart showing the popularity of various parameters. The x-axis shows the popularity of the parameter on a scale from 1 to 4, and the y-axis shows the total number of automobiles that correspond to that parameter. Figure 5 illustrates a histogram of parameter popularity, with safety rating (x-axis) and popularity (y-axis) scaled from 1 to 4, respectively. Figure 6 shows a stacked plot of parameter popularity, with purchase price and maintenance cost on the x-axis and safety rating and popularity on the y-axis, scaled from 0 to 3.5, respectively. 2. CSV file -: Automobiles with the aforementioned features (but not popularity) make up the test dataset. Using the remaining variables, we want to forecast which vehicles in the test dataset will be the most popular.

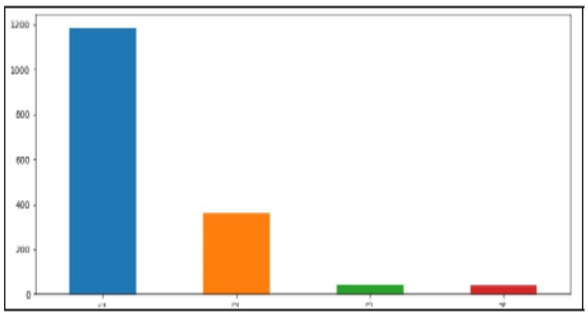

Fig 4: Bar Chart representation of Popularity parameter

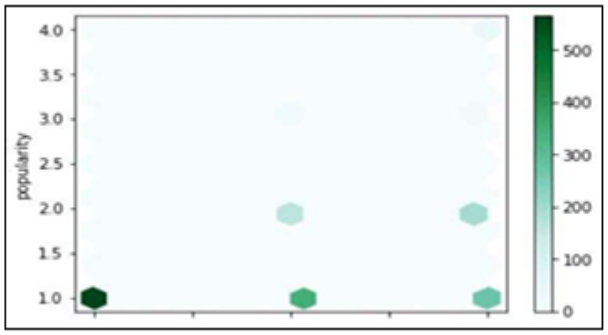

Fig 5: Hexplot of popularity

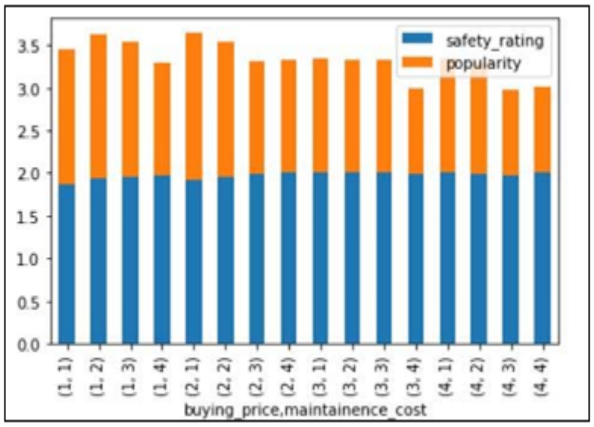

Fig 6: Stacked Plot of parameter popularity

RESULT AND DISCUSSION

Finding out the model's efficacy using different performance indicators is the next step after running the Machine Learning Algorithm. Every Machine Learning Algorithm has its own unique set of performance measures. To illustrate: We use a variety of performance measures, including Accuracy, Precision, Recall, Cross Validation, and f1 Score, for classification [19]. We employ Root Mean Square Error (RMSE) [20] and Mean Square Error (MSE) [20] when we apply a machine learning algorithm to make a forecast, such as when we sell a vehicle or try to estimate the value of a stock. The performance of the Machine Learning Algorithms used to this topic cannot be measured due to the lack of output data. Nevertheless, we have documented the anticipated results in a.csv file that was generated after the execution of the methods described in this article. Table 1 shows the results of our accuracy calculations for the machine learning models we used.

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

TABLE I. TRAING TESTING ACCURACY OF MODELS

Model	Training Accuracy	Test Accuracy
KNN	0.97	0.94
Logistic Regression	0.83	0.99
Random Forest	**0.86**	**0.98**
SVM	0.97	0.99

CONCLUSION AND FUTURE WORK

Machine learning is an emerging method for addressing practical issues. The supervised learning algorithms discussed in this article, including Logistic Regression, KNN, SVM, and Random Forest, were used to forecast how popular a vehicle company's predictions would be on a scaling scale of [1...4]. It is evident from table 1 that SVM is providing us with the best outcome. In light of this, we want to improve the accuracy of the predictions in our future work by tweaking the SVM model now in use. Because they allow for a greater generalization of difficulties, deep learning and neural network methods will also be our emphasis while solving the challenge.

REFERENCES

[1]. Jiao, Yang, and Jérémie Jakubowicz. "Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks." Big Data (Big Data), 2017 IEEE International Conference on. IEEE, 2017.

[2]. Gad, Ibrahim, and B. R. Manjunatha. "Performance evaluation of predictive models for missing data imputation in weather data." Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on. IEEE, 2017.

[3]. Khandelwal, Veena, Anand Chaturvedi, and Chandra Prakash Gupta. "Amazon EC2 Spot Price Prediction using Regression Random Forests." IEEE Transactions on Cloud Computing, 2017.

[4]. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436.

[5]. Le, Quoc V., Jiquan Ngiam, Adam Coates, Abhik Lahiri, Bobby Prochnow, and Andrew Y. Ng. "On optimization methods for deep learning." In Proceedings of the 28th International Conference on International Conference on Machine Learning, pp. 265-272. Omnipress, 2011.

[6]. Zhu, Xiaojin. "Semi-supervised learning literature survey." (2005).

[7]. Olsson, Fredrik. "A literature survey of active machine learning in the context of natural language processing." (2009).

[8]. Cambria, Erik, and White B. "Jumping NLP curves: A review of natural language processing research." IEEE Computational intelligence magazine 9.2 (2014): 48-57.

[9]. Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160 (2007): 3-24.

[10]. Khan, A., Baharudin, B., Lee, L.H. and Khan, K., 2010. "A review of machine learning algorithms for text-documents classification." Journal of advances in information technology, 1(1), pp.4-20.

[11]. Jiang J. "A literature survey on domain adaptation of statistical classifiers." URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey. 2008 Mar 6;3.

[12]. Kaelbling, L.P., Littman, M.L. and Moore, A.W., 1996. "Reinforcement learning: A survey." Journal of artificial intelligence research, 4, pp.237-285

[13]. Ban, Tao, Ruibin Zhang, Shaoning Pang, Abdolhossein Sarrafzadeh, and Daisuke Inoue. "Referential knn regression for financial time series forecasting." In International Conference on Neural Information Processing, pp. 601-608. Springer, Berlin, Heidelberg, 2013.

[14]. Dutta, A., Bandopadhyay, G. and Sengupta, S., 2015. "Prediction of stock performance in indian stock market using logistic regression." International Journal of Business and Information, 7(1).

[15]. Liaw, A. and Wiener, M. "Classification and regression by randomForest." R news (2002), 2(3), pp.18-22.

[16]. Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. and Feuston, B.P. "Random forest: a classification and regression tool for compound classification and QSAR modeling." Journal of chemical information and computer sciences (2003), 43(6), pp.1947-1958.

[17]. Smola, A.J. and Schölkopf, B. "A tutorial on support vector regression." Statistics and computing (2004), 14(3), pp.199-222.

[18]. Gunn, S.R. "Support vector machines for classification and regression." ISIS technical report (1998), 14(1), pp.5-16.

[19]. Williams, N., Zander, S. and Armitage, G. "A preliminary performance comparison of five machine learning algorithms for practical IP traffic

flow classification." ACM SIGCOMM Computer Communication Review (2006), 36(5), pp.5-16.

[20]. Willmott, C.J. and Matsuura, K. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance." Climate research (2005), 30(1), pp.79 82.