Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/27/20230119

Predicting consumer acceptance of automobiles based on deep learning and traditional machine learning algorithms

Linrang Yang

Science Facility, University of Alberta, 116 St & 85 Ave, Edmonton, AB, Canada, T6G 2R3

linrang@ualberta.ca

Abstract. Researchers have made significant progress in machine learning in recent years. Machine learning can learn and predict large and complex data sets. Researchers have divided machine learning algorithms into two categories: deep learning and traditional machine learning. Every problem can be predicted in both ways. This paper uses the "Car Data" dataset to investigate deep learning and traditional machine learning. In order to find a machine learning algorithm that is more conducive to analyzing and predicting consumers' acceptance of different cars, this paper mainly explores the differences in the prediction accuracy of the three methods of Neural Networks, Random Forest and Support Vector Machine (SVM). We construct 3-hidden layers neural networks and 4-hidden layers neural networks. After testing, it is known that the result predicted by Random Forest is the worst. The prediction accuracy of 3-hidden layers Neural Networks is similar to that by SVM. When we added an extra layer of hidden layers on the basis of 3-hidden layers, the prediction accuracy was higher than that of SVM. Adding a hidden layer can improve the prediction accuracy, and both SVM and Neural Network can be used to analyze Car Data. But not all methods have similar predictive accuracy.

Keywords: Deep Learning, Machine Learning, Neural Network, Random Forest, SVM.

1. Introduction

Cars have changed people's lifestyles and become the primary ways of travel. As the auto industry matures, automakers offer consumers different kinds of cars for different purposes. Automakers need to know what kind of cars consumers are willing to buy so that they can increase the number of corresponding cars produced in a planned way to maximize profits. A reliable machine learning model can help manufacturers quickly and accurately analyze and predict consumers' acceptability for different types of vehicles. However, the maintenance of the traditional machine learning model is very complicated and requires professionals to carefully adjust the parameters, so that the traditional machine learning model is difficult to modify and update according to the real situation [1]. The algorithms of machine learning are very diverse, but the core idea is the same. The algorithm will improve its accuracy in the task according to the experience of processing the task. Classification tasks are the most common in traditional machine learning [2]. The categorization task is to predict which category the new data belongs to based on the new input data. Both neural network and traditional machine learning belong to machine learning, but the most important difference is that traditional machine learning requires a large amount of data to become reliable, while neural network requires enough neurons and a large amount of data to become reliable [3]. Random forest is another traditional machine learning algorithm commonly used.

In order to further study traditional machine learning and deep learning, this paper will use neural networks, SVM and random forest, respectively, to predict people's acceptability of different cars based on car data sets. Finally, finding the most suitable prediction method and the corresponding prediction accuracy can help automobile manufacturers better understand the specific needs of consumers. And promoting the increase of the supply and demand equilibrium in the automobile market.

2. literature review

Many scholars have done a lot of research on random forests. Xie et al. evaluated the renewable use of water resources by a random forest algorithm [4]. Joe et al. constructed the random forest model to predict the possibility of snowstorms, which can help people avoid the possibility of snowstorms in winter to a great extent [5]. Eslami R et al. predicted the possibility of forest fire in an area through random forest, and people can investigate the forest earlier and find the factors that may cause fire [6]. Berk and Hyatt used random forest to predict the sentencing frequency of prisoners and the types of criminal laws. This method can reduce the negative impact of non-objective factors on the sentencing of prisoners as much as possible [7]. Liu et al. used neural networks and Support Vector Machine (SVM) to predict the consumables of different buildings respectively, which fills the vacancy of machine learning in building consumables [8]. Machine learning has a huge impact in all areas of research, although there is a gap in the research on the acceptability of different cars.

3. Data set and theoretical framework

3.1. Data set

The Car Dataset contains 1728 sets of data, which have six input attributes: buying, maintenance, doors, persons, lug_boot, safety, and one output attribute: car acceptability. Buying corresponds to four levels: vhigh (very high), high, med and low respectively represent the prices consumers need to buy a car. So Maint corresponds to four levels, which are vhigh, high, med and low, representing the price consumers need to maintain their cars. The Doors correspond to four levels (2, 3, 4, and 5more), which represents the number of doors. Persons correspond to three levels, 2, 4 and more, representing the maximum number of people the car can carry. Lug_Boot corresponds to three levels: small, med and big, representing car boot sizes. Safety corresponds to three levels, which are low, medium, and high , representing the safety of the car. Car Acceptability corresponds to the four unacceptable levels of consumers toward automobiles: unacc (unacceptable), acc (acceptable), good, and vgood(very good).

3.2. Neural Network

Neural networks are also known as deep learning. The structure of the neural network is similar to the structure of neural connections in the biological brain. The top layer is the input layer, through which new data enters the neural network, and the last layer is the output layer, which outputs results based on the new data. The hidden layer is between the input layer and the output layer. The input layer has the weight to connect the hidden layer, and the hidden layer has the weight to connect the output layer. After calculating each hidden layer, the corrected linear unit (Relu) is used to return to the output layer finally. The weight of connections can be adjusted through continuous learning of neural network, which makes the prediction of the neural network more and more accurate. The data set is divided into a training set and a test set [9]. Neural networks can learn the relationship between nonlinear input and output.

3.3. Support Vector Machine (SVM)

SVM is a common classification algorithm in traditional machine learning. SVM algorithm is to find the plane that can correctly separate data sets [2]. Since classification is not only divided into two categories but also into multiple categories, consider using planes to classify data sets in multidimensional space. The advantage of SVM is that it is not easy to overfit [2]. Overfitting means that data can be well classified on the plane of the training data set, and classification errors will be generated once using data outside the training data set.

3.4. Random Forest

Random forest is another common classification algorithm. It is mainly composed of several decision trees; each decision tree is independent and unrelated. The nodes of the decision tree are randomly selected K attributes from the attributes of the input data as nodes for splitting until they can no longer be split. A decision tree generates its own training data set by randomly drawing replacements from the training data set. A new input data is classified by all decision trees at the same time, and each decision tree has a classification result for the new input data. Among all classification predictions, the classification result that appears most frequently is the classification prediction finally obtained by random forest. The advantage of a random forest is that it is not easy to over-fit [10].

4. Method

4.1. Preprocessing Data

Because there are numeric variables and category variables in the Car Data dataset, all of them should be unified into numerical variables. In this paper, 3 represents vhigh, 2 represents high, 1 represents medium, low represents 0, 5more represents 1, more represents 3, small represents 0, big represents 3, vgood represents 3, good represents 2, acc represents 1, and unacc represents 0. Figure 1 is the original data. Figure 2 shows the pre-processed data.

Table 1. Original data set [11]

vhigh	vhigh	2	2	small	low	unacc
vhigh	vhigh	2	2	small	med	unacc
vhigh	vhigh	2	2	small	high	unacc
vhigh	vhigh	2	2	med	low	unacc
vhigh	vhigh	2	2	med	med	unacc
vhigh	vhigh	2	2	med	high	unacc
vhigh	vhigh	2	2	big	low	unacc
vhigh	vhigh	2	2	big	med	unacc
vhigh	vhigh	2	2	big	high	unacc
vhigh	vhigh	2	4	small	low	unacc
vhigh	vhigh	2	4	small	med	unacc
vhigh	vhigh	2	4	small	high	unacc
vhigh	vhigh	2	4	med	low	unacc
vhigh	vhigh	2	4	med	med	unacc
vhigh	vhigh	2	4	med	high	unacc
vhigh	vhigh	2	4	big	low	unacc
vhigh	vhigh	2	4	big	med	unacc
vhigh	vhigh	2	4	big	high	unacc
vhigh	vhigh	2	more	small	low	unacc
vhigh	vhigh	2	more	small	med	unacc
vhigh	vhigh	2	more	small	high	unacc
vhigh	vhigh	2	more	med	low	unacc
vhigh	vhigh	2	more	med	med	unacc
vhigh	vhigh	2	more	med	high	unacc

Table 2. Pre-processed data set

3	3	'2'	'2'	0	0	0
3	3	'2'	'2'	0	1	0
3	3	'2'	'2'	0	2	0
3	3	'2'	'2'	1	0	0
3	3	'2'	'2'	1	1	0
3	3	'2'	'2'	1	2	0
3	3	'2'	'2'	3	0	0
3	3	'2'	'2'	3	1	0
3	3	'2'	'2'	3	2	0
3	3	'2'	'4'	0	0	0
3	3	'2'	'4'	0	1	0
3	3	'2'	'4'	0	2	0
3	3	'2'	'4'	1	0	0
3	3	'2'	'4'	1	1	0
3	3	'2'	'4'	1	2	0
3	3	'2'	'4'	3	0	0
3	3	'2'	'4'	3	1	0
3	3	'2'	'4'	3	2	0
3	3	'2'	3	0	0	0
3	3	'2'	3	0	1	0
3	3	'2'	3	0	2	0
3	3	'2'	3	1	0	0
3	3	'2'	3	1	1	0
3	3	'2'	3	1	2	0

4.2. Construct Neural Network

The input layer has six inputs corresponding to six attributes. There are three hidden layers. The first hidden layer expands the 6 input attributes to 128 neutrons. The second hidden layer expands 128 neutrons into 256 neutrons hidden layer shrinks 256 neutrons into 64 neutrons. The output layer is linear, which will come out of 4 numerical outputs corresponding to 4 acceptability. Finally, using Softmax can normalize the output of the neural network to the (0, 1) interval. Figure 1 shows the neural network.

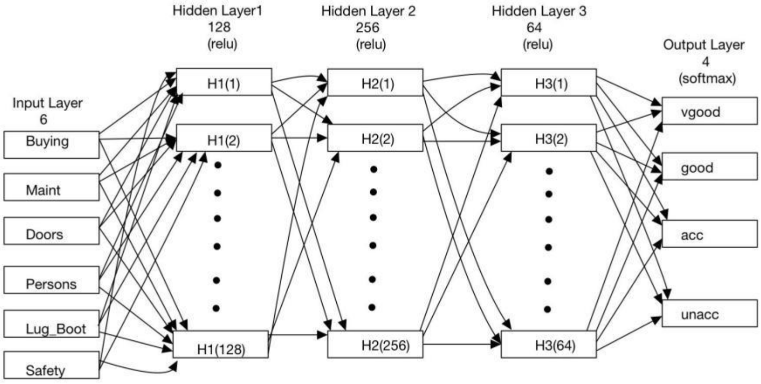

Figure 1. Neural Network.

This paper used cross-entropy as a loss function; the optimizer is Adam. In this paper, 70% of the Car Data is randomly used as the training set and the remaining 30% is used as the test set. This paper trains for 250 epochs. Figure 2 shows the change in the loss function.

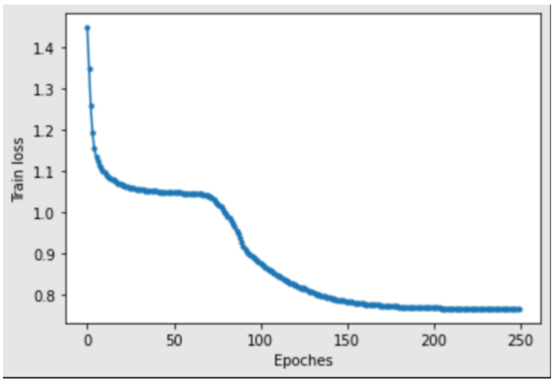

Figure 2. The loss function of three hidden layers neural network.

In the first 50 epochs of training, the loss curve drops quickly, indicating that the neural network is still learning. From 50 epochs to 80 epochs, the loss curve is a straight line, indicating that the loss function has reached the local minimum. Then the loss function jumps out of the local minimum and continues to decline, indicating that the loss function has found another smaller local minimum . In this paper, based on 3 hidden layers, 1 hidden layer is added as the second neural network. The fourth hidden layer shrinks 256 neutrons into 128 neutrons . The loss curve of the 4 hidden layers neural network is similar to the loss curve of the 3 hidden layer neural network (Figure 3).

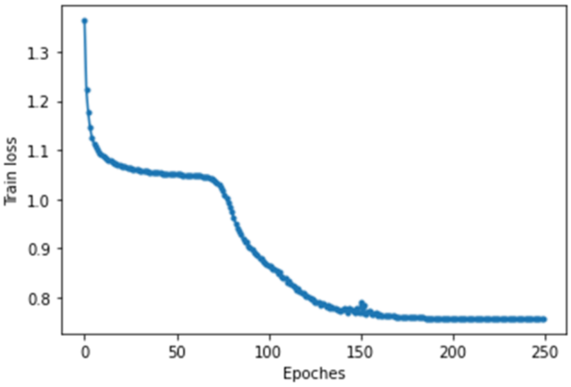

Figure 3. The loss function of four hidden layers neural network.

5. Results and discussion

This paper predicts people's acceptance of different cars by using 3 hidden layer neural networks, 4 hidden layer neural networks, SVM, and Random Forest. The prediction accuracy rates of the four methods are 0.9287, 0.9364, 0.9210 and 0.6821 respectively. According to the confusion matrix, the prediction accuracy under different methods can be displayed more intuitively. The lighter the color in the confusion matrix, the more frequently the data appears. On the diagonal are the number of pairs predicted by the network. Figure 4 shows the confusion matrix of a neural network with 3 hidden layers. According to the six attributes, the network predicts that 350 results are 0, 96 results are 1, 21 results are 21, and 14 results are 3. Figure 5 shows the confusion matrix of the 4 hidden layers neural network. A neural network with 4 hidden layers is more accurate at predicting an output of 1 than a neural network with 3 hidden layers. Figure 6 shows the confusion matrix of SVM. The prediction results with the SVM method are similar to the results of the 3-hidden layers neural network but not as effective as the 4-hidden layers neural network. Figure 7 shows the confusion matrix of the random forest. The prediction accuracy of random forest is the lowest, at only 0.6821, indicating that random forest is not very effective for this task. The 4 hidden layer neural network is more predictive than the 3 hidden layer neural network. SVM has almost the same predictive ability as 3-hidden layer neural network but is weaker than 4-hidden layer neural network.

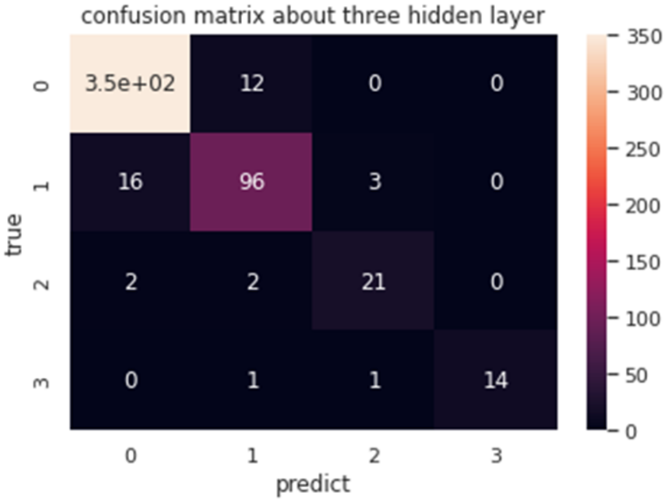

Figure 4. Three hidden layers confusion matrix.

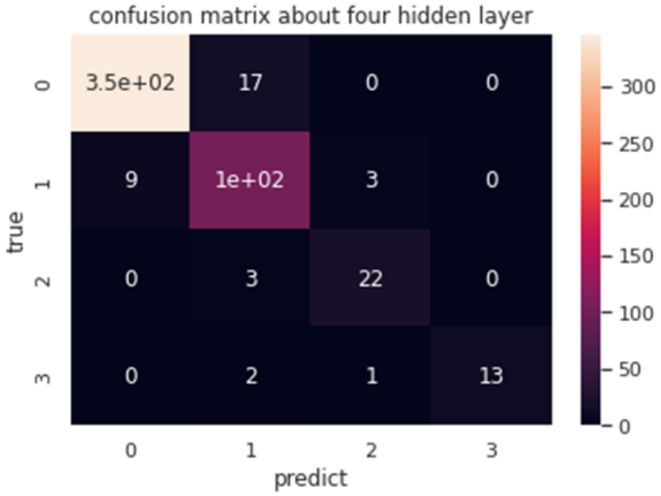

Figure 5. Four hidden layers confusion matrix.

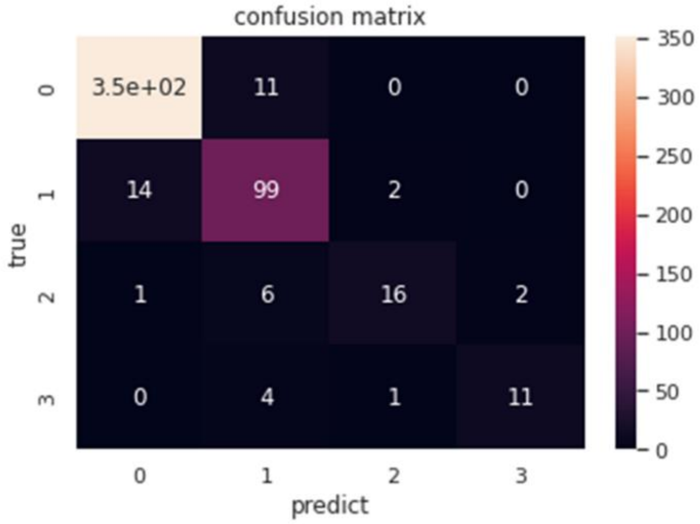

Figure 6. SVM confusion matrix.

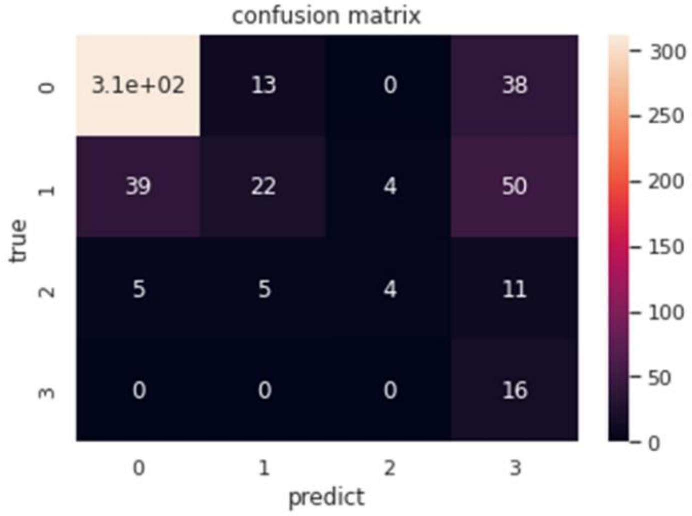

Figure 7. Random forest confusion matrix.

6. Conclusion

This paper uses neural network, SVM and random forest to analyze the Car Data dataset. The prediction accuracy rate of the 3-hidden layers neural network is 0.9287. The prediction accuracy rate of the 4-hidden layer neural network is 0.9364. The prediction accuracy of SVM is 0.9210. The prediction accuracy of random forest is 0.6821. The accuracy of the three-layer neural network is almost the same as that of SVM. The accuracy of 4-hidden layers neural network is better than SVM's. According to the test set accuracy, the effect of applying Random Forest is not ideal. This paper needs to be improved by using the traditional machine learning method, which is not comprehensive enough, and the Naive Bayesian algorithm or K-Nearest Neighbor (KNN) should be used to find their prediction accuracy and find the most suitable prediction method. Another study should be done on the 5-layer neural network to see if the prediction accuracy continues to increase.

References

[1] John Healy, Leland McInnes, Colin Weir. Bridging the Cyber-Analysis Gap: The Democratization of Data Science, The Cyber Defense Review, 2017, Vol. 2, No. 1:109-118

[2] Benjamin C. Jantzen. Discovery without a miracle, Synthese, October 2016, Vol. 193, No. 10: 3209-3238

[3] Mike Ananny.Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness, Science, Technology, & Human Values, January 2016, Vol. 41, No. 1: 93-117

[4] Chengyu Xie, Lei Chao, Dongping Shi, Zhou Ni. Evaluation of Sustainable Use of Water Resources Based on Random Forest, Journal of Coastal Research, WINTER 2020: 134-136

[5] Joe P, Melo S, Burrows W, Casati B, Crawford R, Deghan A, Gascon G, Mariani Z, Milbrandt J, Strawbridge K. Supersite at Iqaluit, Bulletin of the American Meteorological Society , April 2020, Vol. 101: 305-312

[6] R Eslami, M Azarnoush, A Kialashki and F Kazemzadeh. GIS-BASED FOREST FIRE SUSCEPTIBILITY ASSESSMENT BY RANDOM FOREST, ARTIFICIAL NEURAL NETWORK AND LOGISTIC REGRESSION METHODS, Journal of Tropical Forest Science, April 2021, Vol. 33, No. 2: 173-184.

[7] Richard Berk, Jordan Hyatt. Machine Learning Forecasts of Risk to Inform Sentencing Decisions, Federal Sentencing Reporter, Vol. 27, No. 4: 222-228.

[8] Zhijian Liu, Di Wu, Yuanwei Liu, Zhonghe Han, Liyong Lun, Jun Gao, Guangya, Jin, Guoqing Cao. Accuracy analyses and model comparison of machine learning adopted in building energy consumption prediction, Energy Exploration & Exploitation, July 2019, Vol. 37, No. 4: 1426-1451.

[9] Nico Orlandi. Predictive perceptual systems, Synthese, Vol. 195, No. 6: 2367-2386

[10] Daniel R. Gambill, Wade A. Wall, Andrew J. Fulton, Heidi R. Howard. Predicting USCS soil classification from soil property variables using Random Forest, Journal of Terramechanics 65 (2016) : 85–92.

[11] Data Source, Car Evaluation Data Set, Retrieved from: https://archive.ics.uci.edu/ml/datasets/Car+Evaluation