International Journal of Innovative Research in Engineering & Management (IJIREM)
ISSN: 2350-0557, Volume-9, Issue-2, April 2022
https://doi.org/10.55524/ijirem.2022.9.2.97
Article ID IJIRP-2300, Pages 617-621
www.ijirem.org

An Approach of Machine Learning to Get the Popularity Vaticinator of Vehicle

P.V.Subba Reddy[1], Dr. A. Seshagiri Rao[2], P.Ramalingamma[3], and S. Giribabu[4]

[1,3]Assistant Professor, Department of Information Technology, PACE Institute of Technology and Sciences, Ongole, Andhra Pradesh, India
[2] Professor, Department of Information Technology, PACE Institute of Technology and Sciences, Ongole, Andhra Pradesh, India
[4]Assistant Professor, Department of Computer Science & Engineering, PACE Institute of Technology and Sciences, Ongole, Andhra Pradesh, India

Correspondence should be addressed to P.V.Subba Reddy; ithod@pace.ac.in

ABSTRACT- Today could be a world of technology with a predicted way forward for a machine reacting and thinking same as human. during this method of rising computing, Machine Learning, information Engineering, Deep Learning plays an important role. during this paper, drawback the matter} is known as regression or classification downside and here we've resolved a true world problem of recognition vaticinator of a auto company persecution machine learning approaches.

KEYWORDS- Machine Learning, Regression, Classification, Supervised Machine Learning, Logistic Regression, KNN, Random Forest.

I. INTRODUCTION

In the era that we have a tendency to sleep in, technology encompasses a huge impact on our lives. computing [6], data engineering, Machine learning, Deep learning [4][5], tongue processing[7][8] area unit rising technologies that plays a vital role within the leading comes of today's world. computing is a vicinity or branch that aims or emphasizes on making machine that works showing intelligence and their reactions is comparable to it of human. In computing, Machine learning is a necessary and core half providing the flexibility of learning and rising by itself. the main focus of this method is on creation of programs which may decide the information and learn from it by itself. Earlier, statistician and developers worked along for predicting success, failure, future etc. of any product. This method crystal rectifier to delay of the merchandise development and launch. Maintenance of such product within the ever-changing technology and knowledge is additionally one amongst the main challenges. Machine learning created this method easier and quicker. There area unit varied Machine learning pseudo-codes loosely classified into four paradigms: Supervised learning [7] [9] [10]: This learning rule provides a perform therefore on create vaticinators for output values, wherever method starts from analysis of a far-famed coaching dataset. This rule will be applied to the past learned knowledge to new knowledge persecution labels therefore on predict future events. Unsupervised learning: This rule is employed on coaching dataset and informs that is neither classified nor labeled . It conjointly studies to infer a perform from a system to explain a hidden structure from unlabeled knowledge. cluster is Associate in Nursing approach of unattended learning. Semi supervised learning [6] [11]: It takes the characteristics of each unattended learning and supervised learning. These pseudo-codes use quantity|bit|touch of labeled knowledge and enormous amount of unlabeled knowledge. Reinforcement [12]: during this rule, interaction is formed to setting by actions and discovering errors. It permits machines and computer code agents in deciding ideal behavior in a very specific context such performance may well be maximized.

Regression and Classification issues area unit sorts of issues in supervised learning. In classification, conclusion is drawn persecution values that area unit obtained by observation. A distinct output variable say y is approximated by this downside employing a mapping perform say f on input variables say x. The output of category clarification is usually distinct however it may be continuous for each class label within the kind of likelihood. A regression downside has output variable as a true or continuous price. a nonstop output variable say y is approximated by this downside employing a mapping perform say f on input variables say x. The output of regression is usually continuous however it may be distinct for Associate in Nursing category label within the kind of an number with several output variables is mentioned variable regression problem.

In this paper we'll be specializing in a retardant picked from hacker rank wherever a corporation is making an attempt to launch a brand new automobile changed on the idea of the popular options of their existing vehicles. the recognition are foretold persecution machine learning approach. It will be classified as regression downside particularly a variable regression downside and also the downside will be classified beneath supervised learning. so varied supervised learning pseudo-codes are used for this vaticinator.

II. RELATED WORKS

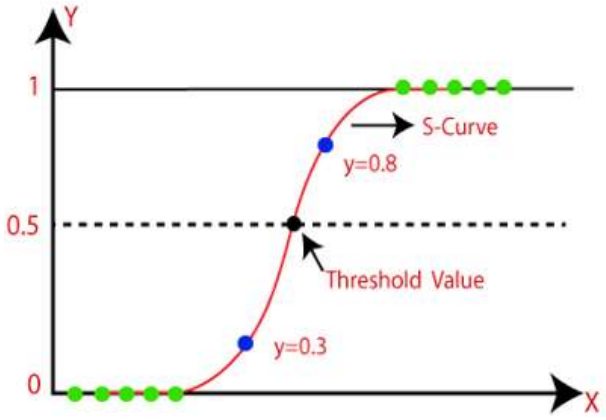
In paper "Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks[1]", author has reviewed some classification pseudo-codes such as random forest, gradient boosted trees, artificial neural network and logistic regression to predict 463 stocks of the S&P 500. In order to study the predictability of these stocks, author has performed multiples of experiments with these classification pseudo-codes. The obtained result of predicting future prices from the past available data was not up to the mark as the expected result, The author wanted to obtain. However, they successfully showed the vast growth in predictability of European and Asian indexes closed a little while back.

2. Arithmetic mean is calculated for r-

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

3. Return as output value for t

A. Logistic Regression [14]



In paper "Performance analysis of prophetical models for missing information imputation in weather data[2]", author has urged a brand new approach to manage the missing information in weather information by playacting numerous tests with NCDC dataset to assess the vaticinator error of 5 methods: statistical regression, SVM, random forest, KNN Implementation and kernel ridge. so as to handle the missing values of dataset they performed 2 actions: one.removing the complete row that contains missing worth and a couple of. Impute the missing information. They performed each the ways to handle the missing information and compared the ascertained result. In paper "Amazon EC2 cash price Vaticinator mistreatment Regression Random Forests [3]", author has planned Regression Random Forests (RRFs) model to forecast the Amazon EC2 cash price one week ahead and one month ahead. This vaticinator model would facilitate in designing once to amass the spot instance, the model conjointly predicts the execution price and it conjointly suggests the user once to bid so as to reduce the execution price.

III. PSEUDO-RELATED IMAGE

B. KNN (K-Nearest Neighbor) [13] KNN

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

here p denotes probability if characteristic of interest is present.

The logit transformation is defined as logged odds.

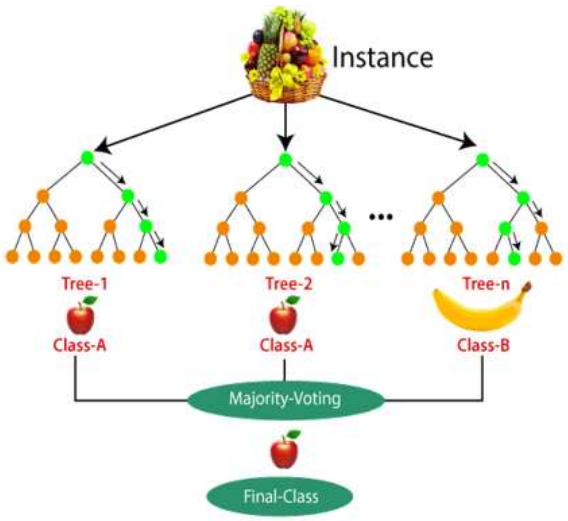
$$\text{Osdds} = \frac{p}{(1-p)} = \frac{\textit{Probability of presence of characteristic}}{\textit{Probability of absence of characteristic}} \qquad (3)$$

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

In logistic regression, estimation is made by choosing parameters that maximizes likelihood of observing the sample values.

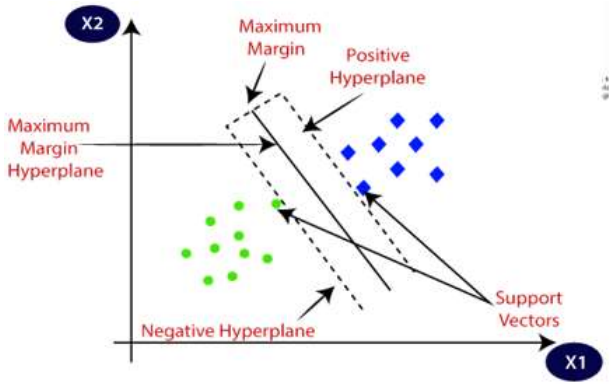
C. Random Forest [15] [16]

Random forest is a type of a supervised classification algorithm for creating a forest and making it random by some way.



Final vaticinator is considered to be highest voted predicted target.

D. Support Vector Machine [17] [18]



IV. EXPERIMENT DETAILS AND NUMERICAL SIMULATIONS

There are two data sets available in a .csv file which is comma separated file with useful information

- Train.csv This is a file that is used as training dataset whose each row provides information on each vehicle. With values such as buying_price, maintenance_cost, number_of_doors, number_of_seats etc. Some of the attributes are explained as follows-
- buying_price The buying_price attribute is used to describe the buying price of the vehicles. It ranges from [1…4] where 1 represents the lowest price and 4 is representing highest price.
- maintenance_cost The maintenance_cost attribute is used to describe the maintenance cost of the vehicles. It ranges from [1…4] where 1 represents the lowest maintenance cost and 4 is representing highest maintenance cost.
- Number_of_doors The number_of_doors attribute is used to describe the number of doors in the vehicle, and the values ranges from [2...5], where each value of number_of_doors represents the number of doors in the vehicle.
- Number_of_seats The number_of_seats attribute is used to describe the number of seats in the vehicle, and the values are [2, 4, 5], where each value of represents the number of seats in the vehicle.
- Luggage_boot_size The luggage_boot_size attribute is used to denote the luggage boot size , and its values ranges from [1..3]. Value 1 smallest and 3 is largest luggage boot size.
- Safety_rating The safety_rating attribute is used to describe the safety rating of vehicles. Its value ranges from [1...3] where 1 represents low safety and 3 is high safety.
- popularity: The popularity attribute is used to describe the popularity of the vehicles. Its values ranges from [1…4] where 1 represents the unacceptable vehicle, 2 represents an acceptable vehicle, 3 represents a good vehicle, and 4 represents the best vehicle. We have performed the experiment in python programming language. We have used pandas, numpy, matplotlib, seaborn, sklearn python libraries for solving the problem. The snippet of training data is shown in Table 1.The schema of training data is shown in Table 2.Brief description of training data is shown in Table 3.

Table 1: 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

Table 2: Training Data Schema

```
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
```

Table 3: Training Data Description

```
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	0.654766
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

Training data visualization Figure 1 represents bar chart of parameter popularity where x axis represents popularity on the scale of 1 to 4 and y represents total count of vehicles belonging to a particular scaling parameter. Figure 2 represents hexplot of parameter popularity where x axis is representing safety_rating on the scale of 1 to 3 and y axis is representing popularity on the scale of 1 to 4. Figure 3 represents stacked plot of parameter popularity where x axis is representing buying_price, maintainence_cost on the scale of 1 to 4 and y axis is representing safety_rating, popularity on the scale of 0 to 3.5.

- Test.csv

It is the test dataset of vehicles along with above attributes excluding popularity. The goal is to predict the popularity of vehicles of test dataset based on their remaining attributes.

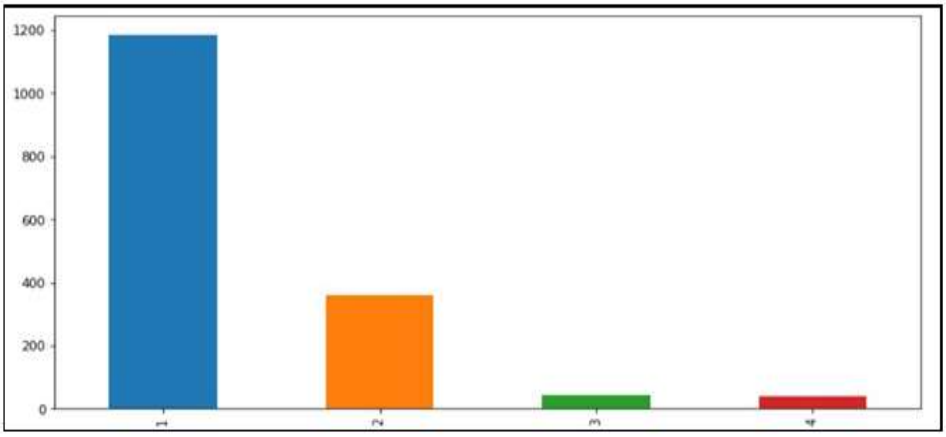


Figure 1: Bar chart representation of popularity parameter

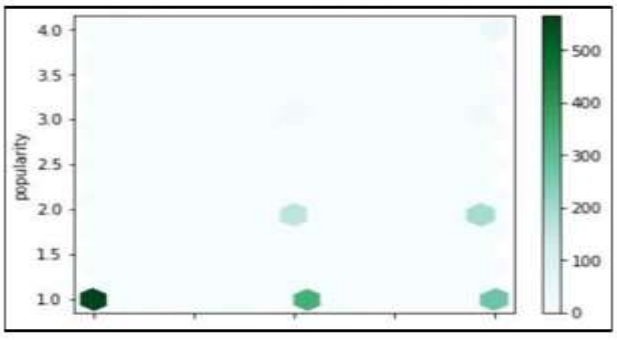


Figure 2: Hexplot of popularity

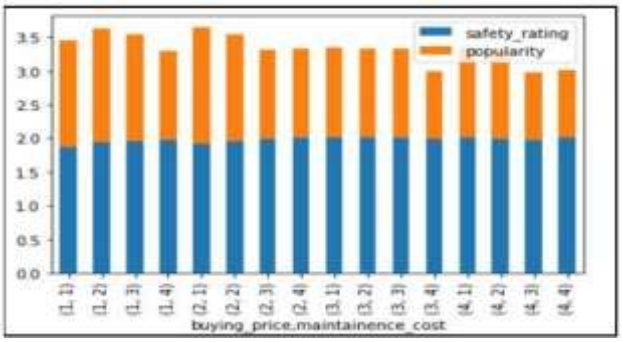


Figure 3: Stacked Plot of parameter popularity

V. RESULT AND DISCUSSION

After executing the Machine Learning Algorithm the next step is to find out the effectiveness of model based on various performance metrics. Different performance

metrics are used for different Machine Learning Pseudo-codes. For example: For classification we use different performance metrics [19] such as Accuracy, Cross Validation, Precision, Recall, and f1 Score. If the machine learning algorithm is used for vaticinator (for example: stock price vaticinator, housing vaticinator and like in our case vehicle popularity vaticinator) we use Root Mean Square Error (RMSE) [20], Mean Square Error (MSE) [20]. Because of absence of output data, we are unable to measure the performance of the Machine Learning Pseudo-codes we applied in this problem. However, we have stored the predicted output values in .csv file we received after performing the pseudo-codes we implemented in this paper. We have calculated the accuracy of the machine learning models we implemented which is shown in table 1.

VI. CONCLUSION

Machine Learning is a fast growing approach to solve real world knots. This paper focused on some of the supervised learning pseudo-codes such as Logistic Regression, KNN, SVM and Random Forest for vaticinator popularity on a scaling measure of [1…4] for a vehicle company. From table 1 it is clear that SVM is giving us the best result. Thus for future work, our focus would be on modifying SVM model used and will try to make the vaticinator more accurate. Also implementing the problem using deep learning deep learning and neural network pseudo-codes will be our focus, as they provide more generalization of knots.

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 Vaticinator 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." (2007): 3-24.

[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 pseudo-codes for text-documents classification." Journal of advances in information technology, 1(1), pp.4-20.