THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

A New Interface For Novice Researchers Employing Various Algorithms: Machine Learning Insights

Jeba Nega Cheltha
School of Computer Science and Engineering,
LPU, Punjab, India
Jeba.25111@lpu.co.in

Manish Choudhary
School of Computer Science and Engineering,
LPU, Punjab, India
cmanish365@gmail.com

Pukhraj Itara
School of Computer Science and Engineering,
LPU, Punjab, India
pukhraj2911@gmail.com

Sahil Sharma

School of Computer Science and Engineering,
LPU, Punjab, India
Sahil.24886@lpu.co.in

ABSTRACT

Machine learning is an application of artificial intelligence (AI) that provides systems, the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. This paper focus on machine learning, which helps the new learner to use the interface without writing thecode. The utilizer with just a few clicks can go back and forth, select some other algorithm, check the precision, make a note of it and reiterate the process until he/she gets the best algorithm of all. The conception of going back and forth might sound vexing, but the time taken to check for all the algorithms is less than 30 seconds. The main motive of this paper as mentioned earlier is to help user to analyze some of the data, and give the ability to make prognostication without indicting any kind of code. The projected work is solely dependent on machine learning. This component was a simultaneous process with the back-end. In this section, we take the dataset from the user, along with the technique and corresponding harvest of algorithm by the utilizer and perform training. The results (precision) of the trained model is

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

exhibited with some consequential visualization plots which give utilizer an intuitive understanding of the data

Keywords: Machine Learning, Decision Tree, Logistic Regression, Random Forest, KNN

I. INTRODUCTION

Machine Learning Insights is a Web application that is based on the machine learning paradigm which intends to solve the problem for novice programmers and machine learning practitioners and to all those who want to start their journey in machine learning. Hence it is common in machine learning to try multiple models and find one that works best for a particular problem but this task can be cumbersome as well as time taking, moreover to find an algorithm that works best for a given dataset and then to further fine-tune it can be computationally expensive as well. The rigid solution is always to run multiple algorithms for a given dataset and to choose the best among them, and to help our audience with that, we have built this tool which will help them get the intended output without writing a piece of code and is faster than what it normally takes while training a model .The anticipated exertion eases the progression of choosing an effective algorithm and further-more gives an idea to the audience about the model concert which familiarizes the consumer with the dataset.

II. LITERATURE SURVEY

Rob Law in 1998concernabout neural networks towards occupancy debt for the quarters of Hong Kong inns and stumble on that neural network outperforms raw extrapolation edition and furthermoregreater to various regression [1]

Hua et al. in 2006portray supportvector machines loom to calculateoccasion of non zero stipulate or load occasionstipulate of auxiliarycomponent which use in petrochemical venture in china for catalogadministration[2] [3]. Wang in 2007 portray the use oflearning method through geneticalgorithm (GA)-SVR with real value GA[4],[11].

III. PROPOSED WORK

Following algorithms are used in this paper.

a. Logistic Regression:

Logistic Regression is a learning algorithm used for classification tasks that predicts the probability of a dependent categorical variable based on the values of other independent variables. Logistic regression makes use of MLE(maximum likelihood estimation) to get the

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

version coefficients that relate the impartial variables (predictors) to the established variables [12]. Logistic Regression is shown in the following Fig 1. The algorithm uses the Logit function which is defined as follows:

Logit = Log(p/(1-p)) = log (P (event happening) / P (event not happening)) = log(odds).

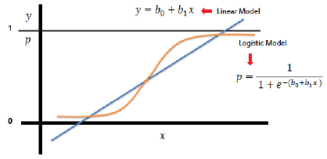


Fig 1: Logistic Regression

b.**Decision Trees:**

Decision Trees may be used to build each decision in addition to regression models with the help of tree structures.The algorithm chunks the data into smaller subsets as the depth of the tree increases until the structure reduces to the decision nodes and leaf nodes. Decision trees can be integrated to be used with both numeric as well as categorical records. The root node is the high-quality predictor that slits the facts infurther two sets and the same happens recursively with every decision node. The leaf node represents the decision [6].

c. **Support Vector Machines:**

SVM's or support vector machines are a discriminating classifier to split dataset. This algorithm includes planar transformation with the aid of including an additional axis .The algorithm allows tackling outliers and overlapping factors with the help of a consequence parameter referred to as regularization parameter, tuning the optimization parameters enables us to get a robust version[11]

d.**K-Nearest Neighbours:**

KNN or K nearest neighbor is a supervised learning algorithm and for this reason we must recognize the target variable for a given data record as shown in Fig 1K- stands for the number used to identify similar neighbors for the new data point. The choice of K highly determines the model and the performance of the algorithm. To find the optimal number of clusters, we generally use the elbow method/metric.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

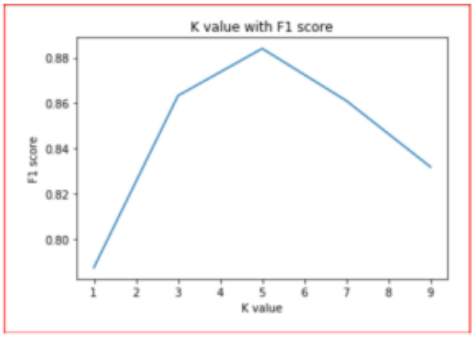


Fig 2 K-Nearest Neighbours

e.**Naive Bayes**

Naive Bayes is a probabilistic machine learning algorithm that allows us to do classification tasks. The gist of the Algorithm is based on the Bayes Theorem.

f.**Adaboost:**

AdaBoost is a Meta machine learning algorithm that incorporates more than one machine learning algorithm to improve the performance of the model. Adaboost is a super example of an ensemble model that makes use of what we call as susceptible freshmen.It outputs a strong gaining knowledge of set of rules that has higher overall performance that the weakclassifier, which is also referred to as robust boosting classifier[7].

g.**Perceptron:**

A perceptron is an algorithm for supervised learning which gives neurons the functionality. It is a neural network unit that does necessary calculations for making a prediction which may be able to help with business [8] solutions.

h. **Random Forest**

Random forest is an ensemble learning framework that incorporates multiple decision trees to uplift the overall performance of the algorithm. It can be used for both regression analysis as well as classification tasks. Random Forest is popular among the machine learning community because of its performance as without any hyper parameter tuning; the performance is exceptionally well as compared to other ensemble methods. Random Forest creates more randomness in the learning process, which makes the model more generalized and has more accurate predictive power.

IV. IMPLEMENTATION & RESULT

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

1. On clicking the button 'technique' drop down, the user can select between 'Classification' and 'Regression' which in turn will bring out another drop down:

a. If the user selects 'Classification', the other drop down will consist of 8 classification algorithms to choose from.

b. If the user selects 'Regression', the other drop down will consist of 4 regression algorithms to choose from.

2. The user will have to copy and then paste the path of the dataset in the Path field.

3. On clicking 'submit' button, the division below (column name) will be filled with the columns that are consisted in the dataset. The user must know the target column beforehand.

4. The user must pass the target column as a string value in the 'Enter the target field'.

5. On clicking the 'Next' button, the user will be taken to new page that will be loaded on the previous page itself. This page will consist of 4 different interactive visualizations along with training and testing accuracies.

6. The accuracies that are calculated will go through a generalized machine learning exploratory data analysis process.

7. First null values are checked. If there is any null values, the empty field will be interpolated by mean values.

8. After solving null value problem, we do standardization of the whole column using mean centering and variance scaling.

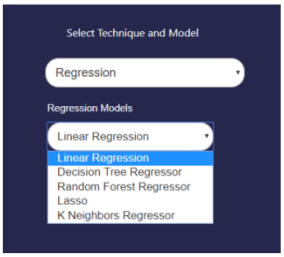
9. After that we split the whole data using 70-30 rule in which we use 70 percent of whole data for training and the remaining 30 percent for validation or test our model performance in unseen data.

10. After splitting we pass the algorithm according to the user's choice and the model is trained. Finally, the results are displayed in the following Figures.

Fig 3: Home Page

Fig 4: Path

ISSN: 0971-1260
Vol-22-Issue-3-July-2019



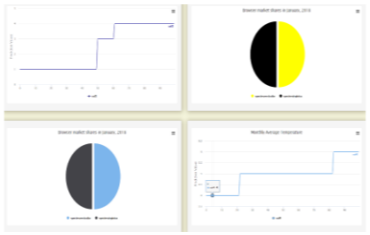


Fig 5: Algorithm Selection Fig 6: Interactive plots Fig 7: Accuracy report

V.CONCLUSION

This proposed work helps the users to use this interface without writing any code. According to our dataset accuracy of Random Forest is accurate. It will take time to write code for each and every algorithms but this proposed work helps the users to use this interface without writing code. The conception of going back and forth might sound vexing, but the time taken to check for all the algorithms is less than 30 seconds. The main motive of this paper as mentioned earlier is to help user to analyze some of the data, and give the ability to make prognostication without indicting any kind of code. The projected work is solely dependent on machine learning.

REFERENCES

[1]. Zhongsheng Hua , Bin Zhang, "A hybrid support vector machines and logistic regression approach for forecasting intermittent demand of spare parts" ,Applied Mathematics and Computation 181, pp 1035–1048, 2006.

[2]. Real Carbonneau, Kevin Laframboise, Rustam Vahidov, "Application of machine learning techniques for supply chain demand forecasting " , European Journal of Operational Research 184, pp 1140 1154, 2008.

[3]. Kuan-Yu Chen, Cheng-Hua Wang, "Support vector regression with genetic algorithms in forecasting tourism demand" , Tourism Management 28, pp 215–226, 2007.

[4]. Wei-Chiang Hong, Yucheng Dong, Li-Yueh Chen, Shih-Yung Wei, " SVR with hybrid chaotic genetic algorithms for tourism demand forecasting", Applied Soft Computing 11, pp 1881–1890, 2011.

[5]. M. H. Fazel Zarandi, Esmaeil Hadavandi,B. Turksen,"A Hybrid Fuzzy Intelligent Agent-Based System for Stock Price Prediction", International Journal Of Intelligent Systems, Vol. 01, pp 1–23, 2012.

[6]. https://api.highcharts.com/highcharts/

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

[7].https://flask-cors.readthedocs.io/en/latest/

[8]. https://docs.mongodb.com/

[9]. https://cloud.google.com/automl/

[10].https://www.tutorialspoint.com/mongodb/index.htm

[11]. https://towardsdatascience.com/introduction-to-pytorch-biggraph-with-examples-b50ddad922b8

[12].https://noflojs.org/documentation/graphs/