Curiginal

Document Information

Analyzed document	loksundar.doc (D97301999)
Submitted	3/5/2021 10:52:00 AM
Submitted by	NAGARAJ P
Submitter email	nagaraj.p@klu.ac.in
Similarity	0%
Analysis address	klulibrary.kalu@analysis.urkund.com

Sources included in the report

Entire Document

TABLE OF CONTENTS

ABSTRACT

In any School, if a significant number of students drop from school with a short notice period, it may lead to a reduction in overall throughput which in turn will certainly have an impact on the school . School Management need to spend additional efforts in terms of time and cost to fill up the vacant position without any substantial loss to the School . To avoid these situations, we can use machine learning techniques to predict students who are planning to leave the school with the help of some related data. One more way is to identify the features which inspire students to leave their school. Refining such features in the school also will result in reducing the student dropout rate of the school . In this paper, we attempted to predict school dropout analysis using the classification algorithms, Decision tree, Random Forest, Neural Networks and eXtreme Gradient boosting . We also have applied regularization for every algorithm to find the precise parameters to predict the student dropout rate considering the School data set from Survey which consists of 15 features .

Keywords: Dropout rate, Classification algorithms, Decision tree, Random Forest, MultiLayer Neural Networks, eXtreme Gradient boosting, Regularization parameters, School dataset from survey.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1. DT Decision Tree 2. RF Random Forest 3. XGB Xtrenme Gradient Boost 4. NN Neural Networks 5. PCA Principle Component Analysis

CHAPTER 1

INTRODUCTION

Student dropout will affect school and students in many ways. If any student leaves the school, more focus is towards knowledge transfer than any other productive work. Finding a suitable school with all kind of comfort may not be easy . Students may take time to adjust with new school environment and friends.

To avoid such scenarios, the school management tries different ways to know which student is not content in the current academics. Then, the teachers motivates such students to avoid the movement. Accurate prediction of student dropout rate and taking preventive measures play a valuable role in maintaining the proper strength of the school . So, an automated approach using machine learning algorithms to predict student dropout rate has become the essential need for every school.

CHAPTER 2

LITERATURE SURVEY

Literatre Table 2.1 :

S.No Paper Author Journal Description 1 Bayesian Optimization with Unknown Search Space Huong Ha, Santu Rana, Sunil Gupta, Thanh Nguyen, Hung Tran-The, Svetha Venkatesh 33rd Conference on Neural Information Processing Systems (NeurIPS 2019) In this paper, they Implemented the new way of optimizing the features space which is unknown to the

best fit data can do. To address this problem, they propose a systematic volume expansion strategy for Bayesian optimization. They devise a strategy to guarantee that in iterative expansions of the search space, our method can find a point whose function value within Epsilon of the objective function maximum. Without the need to specify any parameters, our algorithm automatically triggers a minimal expansion required iteratively.

Literature Table 2.2 :

S.No Paper Author Journal Description 21 Learning Feature Engineering for Classification Nargesian, Horst Samulowitz, Udayan Khurana Elias B.Khalil, Deepak Turaga The Twenty-Sixth International Joint Conference on Artificial Intelligence has published its proceedings (IJCAI-17)

In this paper, they discussed the automated feature engineering techniques. Existing approaches to automating this process rely on either transformed feature space exploration through evaluation-guided search or explicit expansion of datasets with all transformed features followed by feature selection. Such approaches incur high computational costs in runtime and/or memory. For this, they presented a novel technique, called Learning Feature Engineering (LFE).

CHAPTER 3

SURVEY QUESTIONS AND NEED ANALYSIS REPORT

3.1 SURVEY QUESTIONS:

Place: Municipal Commissioner, 63-Sakkarai kulam street, Srivilliputtur. Pincode- 626 125. Phone: Office: 260257, Personal: 261222 STD Code: 04563 E-Mail: commr.srivilliputhur@tn.gov.in

.

Interacting with staff:

• How many students are there in the school ?

There are 200 students in the high school.

• How many students are dropped out in the last two years?

10 students were dropped out in last two years.

• Did any of the dropped students mention the reason for dropout?

No , None of the students were mentioned reason for dropout.

• How many language teachers are there?

3 teachers are there to teach English.

• How many are Science teachers?

5 teachers are there to teach Science.

• Is Internet is available in the school?

No, Internet was not available.

3.2 NEED ANALYSIS:

• Our analysis will useful to school management to prevent dropout of students.

• By using Machine Learning algorithms we found the main reasons for dropout of students.

• Based on that ,the school management can overcome the reasons for dropout of students.

• Our analysis will help to the school management to overcome the loss for dropout of students.

• It will help to the management to maintain the proper strength of students in the school.

CHAPTER 4

OBJECTIVES

• The main objective of this project is to find the school dropout Students in advance and Reasons for Dropping the school using machine learning algorithms.

CHAPTER 5

METHODOLOGY

5.1 Major Design Constraints:

The data set that we consider in this paper is School Dataset which we collected and created from the survey .The flow of the methodology is as follows.

Initial Model :

Fig 5.1.1 : First Model

Final Model :

Fig 5.1.2 : Improvised Final Model

Data-pre-processing:

The School data set from Survey contains 15 features in which 14 are independent features and the remaining one is a dependent feature which is our output. Figure.1 lists all the features of the data set along with the possible values each feature can take and then its data type. The features are almost self-explanatory from their names itself.

The features in this data set are of different data types like categorical, binary and numerical. This required us to convert every feature into numerical form ,in addition to pre-processing the data. The feature engineering can give a better accuracy score as it removes the unwanted features. To start with, we first converted the categorical data to binary numeric data using an encoding script which uses a dictionary concept.

5.2 Algorithms Used :

Decision Tree:

Decision trees are useful when data possess nonlinear patterns present in it. It is represented as a flow chart in which internal nodes represent the test on a particular-feature and the branches represent output of the test and the child node represents the class in which the input / item is classified. We consider Entropy as the splitting constraint. The decision tree which we get for the employee attrition prediction is shown in Fig 3.

Figure 5.2.1 : Decision Tree

Random Forest : The working principle is given in Figure.4 for random forest algorithm It constructs a number of decision trees using CART procedure and the output will be either mean or mode of all trees present in a random forest. This helps the algorithm to avoid overfitting. The diagram representation of the Random forest algorithm is shown below.

Figure 5.2.2 : Illustration of Random Forest Woking Principle

Neural Networks : Neural networks are mostly used for image classifications and computer vision to learn patterns on the unstructured data. As we know that Neural Networks works best for unstructured data, it is capable of taking out difficult patterns from data. Since our data is biased, we selected neural networks to find how best this works to our problem.

The input for each individual neuron can be given by z=wTx+b where WT is weight vector and b is bias. The output of the neuron can be given by the function y = g(z) the function g is called sigmoid function . The loss function can be given by

eXtreme Gradient Boost (XGB) : XGB algorithm is an advanced version of Random forest. In random forest, each sub tree is independent but in this model the trees are interlinked and error in the current tree is adjusted through constructing consecutive trees by reducing errors. Every tree has some weight in the final prediction of the model according to its error rate .

The main formulae's we use in the XGB is :

similarity score = ΣΘ / (ë +λ)

λ is used for regularization which varies between 0-1

Θ = error residual= difference between the predicted and expected values

è = error samples

5.3 Experiment:

Regularization :

For each model, we apply regularisation by changing the parameters and drawing graphs with respect to the accuracy and parameters. After analysing it, we can select the best parameters and include these model parameters for final comparison.

The data set used in this paper contains 15 features in which 11 features are useful for the prediction of dropout rate not including school id, total students, total tiolets, establishment year. The output variable is a binary variable having value of Yes / No indicating the prediction of student movement. For our experiment, we used 70% data from the data set for training and 30% for testing. The graphs are as follows:

Decision Tree : Random Forest:

Figure 5.3.1: Regularized Parameters Decission Tree vs Random Forest

From Figure 5.3.1, we could finalize regularized parameters for D-tree algorithm as 5 nodes and for Random forest regularized value for number of estimators is around 3.

XG Boost :

Figure 5.3.2 : XGBoost Regularized Parameters

From Figure 5.3.2,we take regularized parameters as Depth = 5 and Estimators =5 for XGBoost algorithm.

From the above graphs we got the regularized parameters for all the algorithms. Now we will use these parameters for the final comparison of all the algorithms.

CHAPTER 6

RESULTS AND DISCUSSION

6.1: Simulation Results:

The multiple regression results are :

Figure 6.1: Analysis of Multiple Regression algorithm on Data.

After wards we will build the stack model and fits the all four models into ensemble learning model. Then we will test the all models and we will take mode of the four models as the output. The accuracy is as follows

After these we will get the important features which are more involving in the student dropout and save them in the excel sheet .The features are as follows

Excel Sheet:

6.2 Implementation :

The results which we get when we use harmonic mean of classes are as follows. From Fig. 17, we can understood that XGBoost performs best compared to the remaining algorithms with an accuracy of 97% and F1 score of 0.72. Since its Recall score is greater than the Precision score, it says that false negative is greater than false positives. So Type 2 errors are very low. One more observation here is that the

precision, recall and f1-score values are little low because we are not giving importance to only classify into one's (attrition positive) but also for zero's (attrition negative).

Figure 6.2 : Heatmap of the Results of all Models used.

CHAPTER 7

CONCLUSION AND FUTURE WORK

In this paper, we implemented certain classification algorithms on the dataset taken from Survey having 15 features for the prediction of School Dropout students. We are able to obtain 88% accuracy when we use eXtreme Gradient Boosting algorithm. When we observe our model, it is giving best and balancing precision and recall which tells that our model considered both type 1 error and type 2 error reductions.

In future we will try to build an deep learning model with Convolutional Neural network with more data. The dataset we used in this paper consists of around 1500 rows which are less. If we get more data from any research organization or from any competition then our model is regularized very well in the upcoming model. As we know the logic of machine learning is more data, more precision in prediction. Afterward, we can implement deep learning models when we are having more features which are complex but gives more precision in prediction algorithm.

CHAPTER 8

REFERENCES

• Algorithm is Learnt from Udemy - Frank Kane :- Founder, Sundog Education.

• Ensemble Learning to Improve Machine Learning Results , Stats and Bots Available at: https://blog.statsbot.co/ensemble-learning-d1dcd548e936

• Towards Data science Multivariate regression models : https://towardsdatascience.com/data-science-simplified-part-5-multivariateregression-models-7684b0489015

From the figure we observe that the last three features are having negligible weights that is ~0. which tells that these features are collinear features .they may cause overfitting of the model so we will remove these three features.

Curiginal

Hit and source - focused comparison, Side by Side

Submitted text	As student entered the text in the submitted document.
Matching text	As the text appears in the source.