Improvising multinomial Classification Accuracy of the model using Feature Selection and Class Imbalance algorithms

[1]Sujith Jayaprakash, [2]Jaiganesh V.
[1] Research Scholar, Dr. N.G.P Arts and Science College, [2] Professor, Department of PG & Research, Faculty of Computer Science, Dr. N.G.P Arts and Science College.

Abstract— Higher education institutions across the globe are tirelessly making efforts to improvise the student performance and curb down the attrition rates. Academic institutions hold a massive amount of information like students personal details, classroom activities, academic achievements, etc., Thus obtaining insights from this data is highly imperative to take crucial decisions for education institutions. Embracing technology is the only solution to address these challenges and this in turn attributed to the usage of Education Data Mining. In this research work, we made a conscious effort to predict the performance of students using various attributes like socioeconomic, academic and demographic details. To overcome the class imbalance problem, we applied the Synthetic Minority Over-sampling Technique (SMOTE) technique. We conducted Pearson's Chi-Squared Test to identify important attributes and Ensemble methods were used to build the prediction model. The results showed that Random Forest performed better than the Bagging and Boosting models.

Index Terms— ***SMOTE, Chi-Square Test, Random Forest, Bootstrap Aggregation, AdaBoost***

1. Introduction

Technology is in its inclining phase in every area of business and this growth is highly attributed to the application of artificial intelligence and machine learning algorithms. Globally, companies around us are adopting Machine learning algorithms to take prudent measures and achieve greater profit margins. Machine learning algorithms not only aids organizations in taking pragmatic decisions but also helps them to reduce the time involved in decision making process. Hence, be it a media, trading, technology or education industry we have seen an in-depth involvement of machine learning in every phase of business. Though we are yet to sense the fully involved ML applications in Education industry, it's obvious that the institutions are moving towards it and soon we can see the involvement of these ML applications in academic decisions. Increasing amount of data in the education industry urges the need for the involvement of technology to extract meaningful data from the loads of information and aid the institutions in taking timely decisions. One of the emerging disciplines in Data mining is Education Data mining and this is due to the availability of data in education sector which also attracts quite a number of researchers.

EDM aims at devising and using algorithms to improve educational results and explain educational strategies for further decision making [1]. Many of the research works carried out in EDM addresses the following problems in today's education sector,

a) Improving Student's satisfaction level
b) Predicting the academic progression of students and provide timely feedback for improvisation
c) Identifying Internal and External factors influencing the performance of a student
d) Controlling the students attrition rate
e) Analyzing the lecturer feedback and involvement of students in lecture halls to increase the academic quality delivery.

Apparently, in recent years there is plethora of research works carried out to address the aforementioned problems and many have come out with an optimal solutions. Tech industries like IBM and Oracle are tirelessly working on developing applications that can help the educational institutions to address these challenges. In this paper, we tried to address the major challenges faced by every higher education institution i.e., student prediction and identifying the influential factors that directly or indirectly affect students performance [2]. Addressing these core issues will help institutions to improve their academic delivery and curb the student attrition rate to a greater extent. To accomplish this task, we collected student's admission details, socio economic details and also demographic details through surveys and from student's admission record.

2. Related Research Works

Hussain et al., [3] attempted to predict student's performance based on socioeconomic, demographic and academic data. Data Collected from multiple sources are preprocessed and influential attributes are identified using feature Selection methods. Internal assessment attribute in the continuous evaluation process makes the highest impact in the final semester results of the students. Correlation-based Attribute evaluation is used in identifying the influential parameters. Selected classification algorithms such as J48, PART, BayesNet and Random are applied to the processed data. Based on the experimental results it has been observed that Random Forest classification algorithm outperformed the rest and shown a high accuracy rate. Lastly, Apriori algorithm was applied to the dataset to identify the most important relationships. Luhaybi et al., [4] attempted to evaluate the student's performance and identify the key parameter which influences the predictive model using the classification algorithm. Attributes used in this research are classified into three categories: admission information, module-related data, and final year grades. J48, Decision tree, and the Naïve Bayes classifications are used to train the model to predict a low, medium and high risk of failures of students. It has been observed that the accuracy rate of the Naïve Bayes algorithm is higher compared to the other two classification algorithms. The study also revealed that the entry qualification of a student is one of the influential parameters which determine the failure rate. Polyzou, A. and Karypis, G. [5] adopted a two-fold approach to identify the student's at risk and the factors influencing their academic performance. Big data approach was handled in this research work to extract the features from historical grading data. Binary classification is used in the predictive model, Decision Tree and Linear support vector machine are used as base classifiers, and Random Forest and Gradient Boosting are used as ensemble classifiers. Area under the receiver operating characteristic (ROC) curve and F1 score proves that Gradient Boosting and Random Forest Classifiers are the best performing methods. Mueen et al., [6] analyzed the academic performance of students using their academic record and forum participation statistics. Classification algorithms (Decision tree (C4.5), Multilayer Perception, and Naïve Bayes) were used to build the training and model and to predict the outcome. Naïve Bayes classifier produced the highest accuracy in comparison with the other two classification algorithms. Data collected from an e-learning platform is used for this research work. Objective of this research is to assist the teachers to early detect students who are poor performing and provide special attention so as to avoid failures in future or student attrition. Guyon, I. and Elisseeff, A. [7] discussed the importance of variable and feature selection. In their preliminary study it has been observed that variable selection improves the prediction performance of the predictors, provide faster and cost effective predictors, and also provides better understanding of the underlying process that generate the data. Their research also states that wrapper or embedded methods are highly advantageous comparing to simpler variable ranking methods. Karthikeyan, K. and Kavipriya, P. [7] proposed a Student Performance Prediction System (SPSS) based on enhanced feature selection and ensemble classification algorithms using student academic data. In this research work swam optimization technique is used to enhance the feature extraction and the classification accuracy is improvised using cluster-based enhanced support vector machine (CESVM). By applying the SVM and CESVM algorithms in the training model, it has been observed that proposed CESVM algorithm outperforms the conventional classifiers. Villagra-Arnedo et al., [8] made an attempt to analyze the academic performance of students using the behavioral and learning data. A learning platform has been built to record the students learning behavior and also the effective usage of the system. The predictive model is built based on three possible outcomes: high, medium and low class performance. Support vector machine is used to build the training mode and it has been observed that by supplementing behavioral data with learning data provides better and more stable predictions about the student performance.

3. Pre-processing Techniques

In this research work, we have used a supervised learning mechanism in machine learning. The outcome or the target variable will be predicted using several independent variables. A training model will be created using classification algorithms and will be tested. The algorithm which provides the high accuracy will be considered to implement the recommendation model. In our earlier research works, we have used Naïve Bayes, Support Vector machine and ensemble algorithms to predict the performance and found that ensemble outdid the rest of the algorithms. However to improvise the accuracy level, we intend to use a trifold approach in this research work by:

1. Applying SMOTE algorithm to overcome the class imbalance problem. We found that the output class is imbalanced due to the high number of Second class division compared to the first class and the third class division. To address this problem, we have used Synthetic Minority Over-sampling Technique.
2. After preprocessing, feature selection algorithm is used to identify the feature that best contributes to the accuracy of the model.
3. Finally, Classification algorithms are applied and predicted the student's performance in Semester 3. We used different ensemble models to attain the high level accuracy.

a. SMOTE Algorithm

Class imbalance problem occurs when the classes are not represented equally. In this research we used a multi class classification problem. Student's final GPA is classified as First class, Second Class and Third Class. We identified that majority of the student falls under Second class and Third Class and whereas very few fall under the rest. Hence, we faced a class imbalance problem and to overcome this we tried using SMOTE algorithm which increased the number of instances of minority class. This technique is used to increase the accuracy level of the model. SMOTE is a hybrid approach which combines oversampling and under sampling approach. Unlike oversampling, SMOTE creates multiple instances of the minority class to balance the dataset.

b. Feature Selection:

Model performance is highly influenced by the relevance of features used in the dataset. Irrelevant or partially relevant features may adversely affect the performance of the model and hence it is imperative to identify the relevant feature to build a successful model. In this research work, we have used several independent features that may affect the performance of a student. However, it is important to identify the right features that can contribute to the performance of the model. Using feature selection method we have tried to select those features that contribute most to the prediction variable. Feature selection aids in removing redundant data and reduces over fitting. Also, it enhances the training time and accuracy of the model.

In our research work, we used Pearson's chi-squared statistical hypothesis which is generally applied to the groups of categorical features to evaluate the likelihood of correlation or association between them using their frequency distribution.

Chi-Square statistic is calculated as follows:

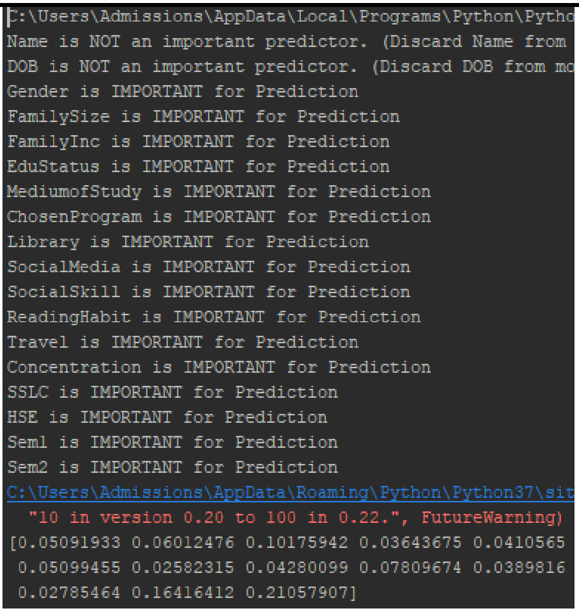
$$\chi = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$Expected = \frac{RowTotal \times ColumnTotal}{OverallTotal}$$

$$DF = (Rows - 1) \times (Columns - 1)$$

Fig. (1) Chi-Square algorithm.

Using Chi-Square statistic, we identified the features that best contributes to the model. Fig (2) shows both the important and non-important features for prediction with its ranking,



```
C:\Users\Admissions\AppData\Local\Programs\Python\Pytho
Name is NOT an important predictor. (Discard Name from
DOB is NOT an important predictor. (Discard DOB from mo
Gender is IMPORTANT for Prediction
FamilySize is IMPORTANT for Prediction
FamilyInc is IMPORTANT for Prediction
EduStatus is IMPORTANT for Prediction
MediumofStudy is IMPORTANT for Prediction
ChosenProgram is IMPORTANT for Prediction
Library is IMPORTANT for Prediction
SocialMedia is IMPORTANT for Prediction
SocialSkill is IMPORTANT for Prediction
ReadingHabit is IMPORTANT for Prediction
Travel is IMPORTANT for Prediction
Concentration is IMPORTANT for Prediction
SSLC is IMPORTANT for Prediction
HSE is IMPORTANT for Prediction
Sem1 is IMPORTANT for Prediction
Sem2 is IMPORTANT for Prediction
C:\Users\Admissions\AppData\Roaming\Python\Python37\sit
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)
[0.05091933 0.06012476 0.10175942 0.03643675 0.0410565
 0.05099455 0.02582315 0.04280099 0.07809674 0.0389816
 0.02785464 0.16416412 0.21057907]
```

Fig. (2) – Output of Chi-Square test.

c. Building a classification model

In this research work, we built a classification model to predict the performance of students based on their historic data and other relevant variables. Multiple classification algorithms are used to ensure that high level of accuracy is attained.

4. Experiment Results

Ensemble methods are used to combine several machine learning models to improvise the prediction accuracy. These methods are used to identify "weak learners" and combine them to create a "strong learner". An ensemble classifier is a supervised learning algorithm which can train and combine several models to predict the output. Different types of ensemble classification methods are,

- Bagging
- Boosting
- Randomization
- Bucket of models
- Bayes optimal classifier
- Stacking

In this research work, we used three ensemble models: Bagging, Boosting and Randomization to measure the prediction accuracy.

a) Bagging algorithms are also known as Bootstrap algorithms which essentially build multiple models from the subsamples of the given training set.

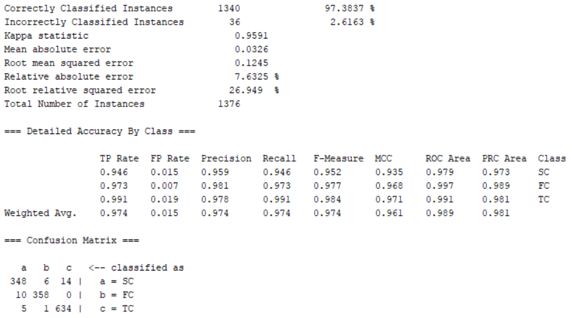
b) Boosting algorithms are also used to improvise the weak learners to strong learners. In boosting algorithm the weak learners are trained sequentially to tune the performance of the model in which each model will correct the predecessor’s value.

c) Randomization is an ensemble model which uses Random Forest which is also an extension of Bagging for Decission trees. It uses greedy algorithm to find the best split point at the each step in tree building process.

Ensemble learning models helps to reduce the variance and bias. A models performance can be measured based on its prediction accuracy and how well it generalizes a test dataset. If the training dataset used to build the model is relatively small or having more variations then it leads to error. So, error is described as the difference between predicted value and actual value. Hence, the main objective is to reduce the error and build a model which can provide highest accuracy rate. Error rates in machine learning are classified as reducible and irreducible errors. Reducible errors are further classified into Bias and Variance. Bias is an error which occurs due to the erroneous assumptions and it is mostly systematically discriminatory. Variance is a difference in the performance of a model in trained and non-trained datasets.

Bagging Algorithm:

In this work, we used WEKA tool to apply the Bagging algorithm on the training dataset. It is also called as Bootstrap aggregation which is a statistical technique. Training data is subdivided into multiple random samples and different learning models are applied to find the predictions. Final result of each models are averaged to make an accurate prediction.



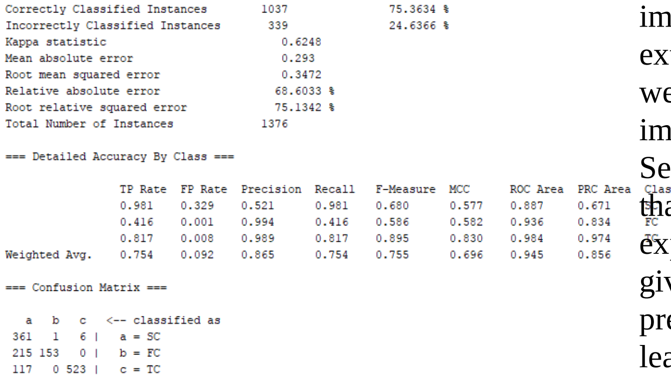
```
Correctly Classified Instances        1340               97.3837 %
Incorrectly Classified Instances        36                2.6163 %
Kappa statistic                          0.9591
Mean absolute error                      0.0326
Root mean squared error                  0.1245
Relative absolute error                  7.6325 %
Root relative squared error             26.949  %
Total Number of Instances             1376

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.946    0.015    0.959      0.946    0.952      0.935    0.979     0.973     SC
                 0.973    0.007    0.981      0.973    0.977      0.968    0.997     0.989     FC
                 0.991    0.019    0.978      0.991    0.984      0.971    0.991     0.981     TC
Weighted Avg.    0.974    0.015    0.974      0.974    0.974      0.961    0.989     0.981

=== Confusion Matrix ===

   a   b   c   <-- classified as
 348   6  14 |   a = SC
  10 358   0 |   b = FC
   5   1 634 |   c = TC
```

Fig. (3) – Output of Bootstrap Aggregation

We used 10-fold cross validation approach in evaluating the dataset. Using bagging algorithm the model achieved 97% accuracy in which Kappa statistic performance was 95%. We also observed that the Cohen's Kappa coefficient (k) is measured to .95 which shows a "near perfection" agreement. Kappa coefficient is a metric that compares the observed and expected accuracy of the model.

Boosting Algorithm:

Adaboost or Adaptive boosting is a machine learning algorithm which is used to improve the performance of a model. It is called adaptive boost because the subsequent weak learners are improvised based on the performance of its predecessors.



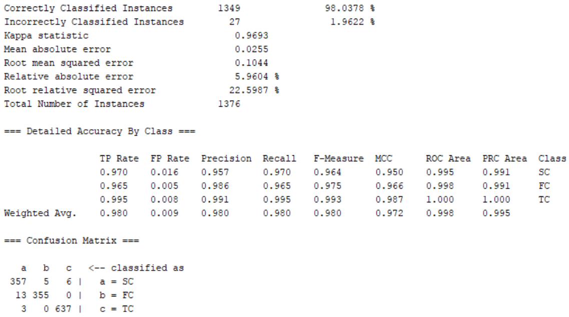
```
Correctly Classified Instances        1037               75.3634 %
Incorrectly Classified Instances       339               24.6366 %
Kappa statistic                          0.6248
Mean absolute error                      0.293
Root mean squared error                  0.3472
Relative absolute error                 68.6033 %
Root relative squared error             75.1342 %
Total Number of Instances             1376

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.981    0.329    0.521      0.981    0.680      0.577    0.887     0.671     SC
                 0.416    0.001    0.994      0.416    0.586      0.582    0.936     0.834     FC
                 0.817    0.008    0.989      0.817    0.895      0.830    0.984     0.974     TC
Weighted Avg.    0.754    0.092    0.865      0.754    0.755      0.696    0.945     0.856

=== Confusion Matrix ===

   a   b   c   <-- classified as
 361   1   6 |   a = SC
 215 153   0 |   b = FC
 117   0 523 |   c = TC
```

Fig. (4) – Output of Adaboost Algorithm

Using this algorithm, our model achieved 75% accuracy and which shows that this algorithm does not fit for this model. Also, the kappa statistic metric is .62 which shows a "substantial agreement".

Random Forest:

Random Forest is an extension over bagging. In this work, we identified that Random Forest has outperformed the Bagging and Boosting algorithm as we achieved an accuracy rate of 98%. Also, Kappa statistic merit shows that the model has achieved "near perfection" classification. Using the greedy algorithm, Random Forest algorithm not only selects the random subset of data but also selects random features to grow the trees. Also, algorithm maintains the accuracy level for missing data which make the model robust.



```
Correctly Classified Instances        1349               98.0378 %
Incorrectly Classified Instances        27                1.9622 %
Kappa statistic                          0.9693
Mean absolute error                      0.0255
Root mean squared error                  0.1044
Relative absolute error                  5.9604 %
Root relative squared error             22.5987 %
Total Number of Instances             1376

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.970    0.016    0.957      0.970    0.964      0.950    0.995     0.991     SC
                 0.965    0.005    0.986      0.965    0.975      0.966    0.998     0.991     FC
                 0.995    0.008    0.991      0.995    0.993      0.987    1.000     1.000     TC
Weighted Avg.    0.980    0.009    0.980      0.980    0.980      0.972    0.998     0.995

=== Confusion Matrix ===

   a   b   c   <-- classified as
 357   5   6 |   a = SC
  13 355   0 |   b = FC
   3   0 637 |   c = TC
```

Fig. (5) – Output of Random Forest algorithm

5. Conclusion:

In our earlier research works, we used Naïve Bayes, SVM and Decision Tree algorithms and created a single model classification. Naïve Bayes outperformed the rest of the algorithms. However, the bias and variance were higher due to the less number of instances used. Also, class imbalance problem affected the output to an extent. Hence to overcome all these problems, we used SMOTE Technique to overcome class imbalance problem. Also, we used Feature Selection technique to identify the best features that contributes to the prediction. Finally, the experiments conducted on the selected feature gives highest accuracy comparing to our previous works. Moreover, use of ensemble learning models like Random Forest and Bootstrap aggregation has reduced the error rate to a greater extent. Model developed using Random Forest is robust and gives a high accuracy rate in prediction. We proposed this model to be implemented in educational institutions as recommender system which can help institutions to identify the features that affect the performance of students and reduce the student attrition rate. Institutions across globe suffer to a great extent due to the student attrition and produce quality graduates. In future research work we will use cluster algorithms to group the features that makes great impact on students performance.

6. References:

1. Silva, Carla & Fonseca, Jose. (2017). Educational Data Mining: A Literature Review. 10.1007/978-3-319-46568-5_9.
2. Devasia, Tismy, et al. “Prediction of Students Performance Using Educational Data Mining.” 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), 2016
3. Hussain, S., Dahan, N. A., Ba-Alwi, F. M., & Ribata, N. (2018). Educational Data Mining and Analysis of Students’ Academic Performance Using WEKA. Indonesian Journal of Electrical Engineering and Computer Science, 9(2), 447. doi:10.11591/ijeecs.v9.i2.pp447-459
4. Luhaybi, M. A., Tucker, A., & Yousefi, L. (2018). The Prediction of Student Failure Using Classification Methods : A Case study. Computer Science & Information Technology. doi:10.5121/csit.2018.80506
5. Polyzou, A. and Karypis, G. (2018). Feature extraction for classifying students based on their academic performance. Eleventh International Conference on Educational Data Mining.
6. Mueen, A., Zafar, B. and Manzoor, U. (2016). Modeling and Predicting Students’ Academic Performance Using Data Mining Techniques. International Journal of Modern Education and Computer Science, 8(11), pp.36-42.
7. Guyon, I. and Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3(1), pp.1157-1182.
8. Karthikeyan, K. and Kavipriya, P. (2017). On Improving Student Performance Prediction in Education Systems using Enhanced Data Mining Techniques. International Journal of Advanced Research in Computer Science and Software Engineering, 7(5), pp.935-941.
9. Villagrá-Arnedo, C., Gallego-Duraìn, F., Compañ-Rosique, P., Llorens-Largo, F. and Molina-Carmona, R. (2016). Predicting academic performance from behavioural and learning data. International Journal of Design & Nature and Ecodynamics, 11(3), pp.239-249.