Students' Academic Performance Analysis

Anonymous

ABSTRACT

This project aims at exploring and analyzing a dataset called âĂIJStudentsâĂŹ Academic Performance DatasetâĂİ which is an educational dataset collected from learning management system (LMS) called âĂIJKalboard 360âĂİ. This dataset has 480 instances and 16 attributes which contains both categorical and numerical attributes. The target variable is the âĂIJClassâĂİ which shows how the students are classified into three numerical intervals based on their total grade. The programming language used is Python. Packages used are pandas, scikit-learn and seaborn. Data mining tool used is Rapid Miner.

1. MOTIVATION

As it is know to all, grade is the technical way to evaluate how well a student study as a result. However, is there a way of predicting students' academic performance based on how they perform before the test? I think so, and this project is how I try to achieve this goal.

The motivation for this project was two fold. The first was the opportunity to deal with a dataset filled with the combination of categorical and numerical attributes by utilizing the power of data mining methods. The second motivation was to build a predictive model that could provide probabilities for students to obtain the grade they deserve.

2. METHODOLOGY

This project can be decomposed into several parts:

1. Cleaning and Preparation
2. Exploration
3. Verification
4. Prediction
5. Improvement
6. Comparison and Conclusion

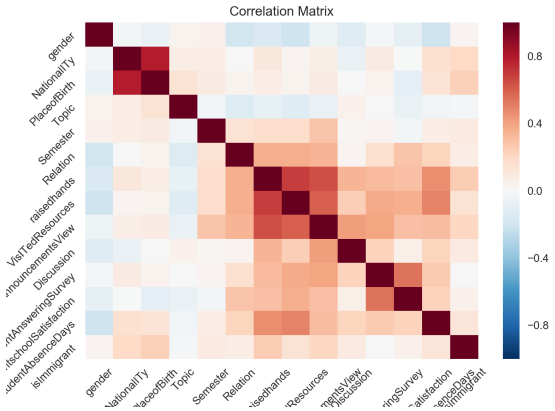
2.1 Cleaning and Preparation

First of all, I use pandas to read in the csv file and obtain the dataframe. Secondly, replace "KW" with "KuwaIT" since another attribute use this name to present the same thing. The third, I use LabelEncoder() function to convert categorical attributes into numerical by assigning all of them in an integer interval from 0 to (category number - 1). The last but the most important is to drop all the ID attributes such as "StageID", "GradeID" and "SectionID" because these are totally meaningless to this data mining project.

2.2 Exploration

Since I have no idea which ones among the fifteen attributes are the most important for us to predict studentsâĂŹ academic performance, data exploration is vital to this project. Many plots have been done to help me have a better understanding of the dataset, but I will not show all of them in the report. The correlation matrix which is shown as follows helps me to see which attributes are more relevant to each other. As we can tell, "raisedhands", "visitedresources" and "announcementsview" are the three most relevant attributes. Therefore, it is highly likely that these three attributes would have more importance in the classification and prediction task.

Figure 1: Correlation Matrix

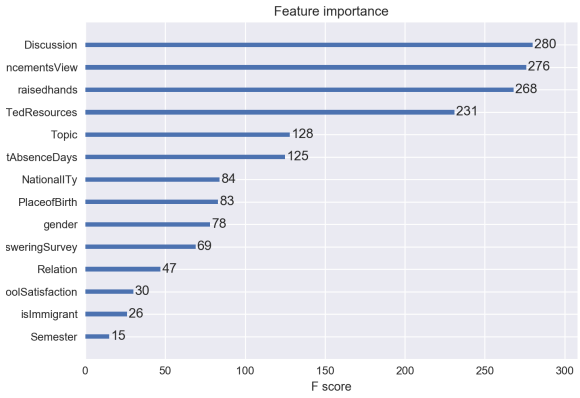


2.3 Verification

In order to verify our guess in previous step, we plot the

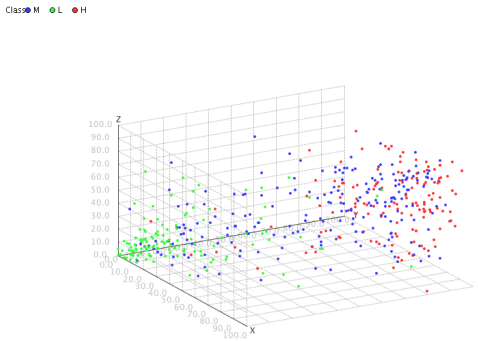
feature importance by using XGBoost package. Apparently the three attributes mentioned above are the top three attributes according to feature importance. Apart from this, "discussion" also seems highly relevant since it scores the 4th highest.

Figure 2: Feature Importance



Even if this shows how important they are to achieve our goal, I still need more persuasive and intuitive evidence to prove that it is true. Therefore, I plot the three most important features using Rapid Miner.

Figure 3: 3D Plot



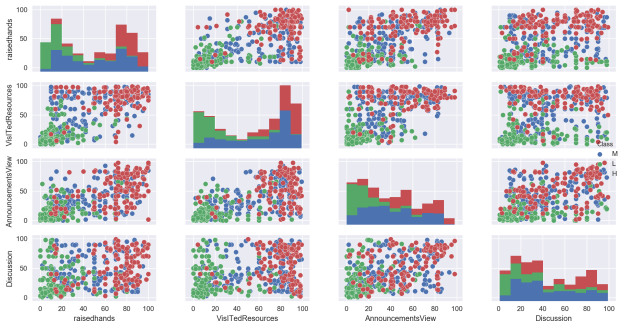
As we can tell, the data points are linear separable and they do have a positive correlation with grade.

Considering the importance of "discussion", I also create a pair grid to show the distance relationship with each pair of attributes. It provides me a better understanding of relationships between each two of them.

Above all, since it is very clear to me that these four features will be vital to our prediction task, I decided to use these four features only instead of all the attributes to do the job.

However, I made the wrong decision. After several times of tests, it proves that prediction accuracy is higher when I use all the attributes instead of the top four important ones. The final results shown in this project uses all the available features.

Figure 4: Pair Grid



2.4 Prediction

Five fold cross validation is used to test the data. I decide to try several different methods to train and test our model including two boosting algorithms: AdaBoost (Adaptive Boosting) and XGBoost (Scalable Tree Boosting). The rest of models are logistic regression, decision tree and Random forest. I also use try-them-all method to try different combination of parameters to see which one of them score the highest on boosting algorithms. Moreover, for all of these random methods, I use 42 as the random state or seed number, since it is the answer to the ultimate question.

Each of the testing results will be shown as a table with its precision, recall, f1-score and support. "H","L","M" represents "High", "Medium" and "Low" in student's academic performance. Average/mean score is also shown in convenience of reading.

- High: 90-100
- Medium: 70-89
- Low: 0-69

Table 1: Logistic Regression

	precision	recall	f1-score	support
H	0.56	0.64	0.60	22
L	0.81	1.00	0.90	26
M	0.79	0.65	0.71	48
AVG	0.75	0.74	0.74	96

Table 2: Decision Tree

	precision	recall	f1-score	support
H	0.67	0.64	0.65	22
L	0.76	0.85	0.80	26
M	0.74	0.71	0.72	48
AVG	0.73	0.73	0.73	96

Table 3: Random Forest

	precision	recall	f1-score	support
H	0.77	0.77	0.77	22
L	0.89	0.92	0.91	26
M	0.85	0.83	0.84	48
AVG	0.84	0.84	0.84	96

Table 4: AdaBoost

	precision	recall	f1-score	support
H	0.69	0.91	0.78	22
L	0.86	0.92	0.91	26
M	0.86	0.75	0.82	48
AVG	0.83	0.83	0.83	96

Table 5: XGBoost

	precision	recall	f1-score	support
H	0.83	0.73	0.80	22
L	0.84	0.92	0.89	26
M	0.82	0.88	0.86	48
AVG	0.83	0.85	0.85	96

2.5 Improvement

The tables shows us that random forest gives us the best accuracy of around 84%, while XGBoost scores only a little bit less. Since XGBoost is a boosting method, I think maybe I can tweak a little bit to make it achieve better performance than random forest does. What I did was change the values of depth, learning rate and estimators within certain range. The following table shows a part of testing results, and it turns out among all the parameters I tried, 85.4167% accuracy is the top score XGBoost can give us.

Table 6: XGBoost Improvements

accuracy	learning_rate	max_depth	n_estimators
0.854167	0.1	3	100
0.843750	0.1	4	50
0.843750	0.1	4	150
0.843750	0.1	4	200
0.833333	0.5	4	50
0.833333	0.1	4	100
0.822917	0.1	3	150
0.812500	1.0	4	100

2.6 Comparison and Conclusion

After improvement of XGBoost, it is found to be slightly better than random forest does. Although random forest is much simpler than XGBoost, it surprisingly did a good job on prediction which is very impressive, and also it proves not every task necessarily need a complex model to fit. To conclude, XGBoost is chosen to be the best model in this project and we successfully achieve 85% prediction accuracy of students' academic performance.

Table 7: Model Comparison

Model	Accuracy Score
Logistic Regression	0.739583
Decision Tree	0.729167
Random Forest	0.843750
XGBoost	0.833333 ->0.854167
AdaBoost	0.833333

3. LESSONS LEARNED

This project is a very good way to apply what I have learned this semester into practice. It is a challenging task for me especially with a lot of categorical data. I tried different ways to achieve the goal but only one of them worked out well. As I mentioned above, I totally focused on choosing the best features to do the work but I ignored the fact that all the attributes after data cleaning are positively correlated with the target variable, which means even the least important ones are contributing to predicting the results. Moreover, I get my own hands on the project which makes me much more familiar with all kinds of data mining techniques and tools, even the ones I have never used before.

4. FUTURE WORK

In this paper I have detailed how to use the right model to predict students' academic performance. It might be useful for teaching in the future because I do not think final grade should be the only metric that shows how well this student studies. With this method, teachers can predict each student's grade based on their daily performance as a reference so that each student could get grade they deserve.

Future work can include improvement of AdaBoost, and also the study of how immigrants does in comparison with locals. In this project, I have created an attribute called "isImmigrant" and tried to find something interesting about this. However, because the dataset is not big enough, I am not able to find a pattern out of it. I hope this could be done when the dataset grows larger in the future.

5. REFERENCES

1. Amrieh, E. A., Hamtini, T., Aljarah, I. (2016). Mining Educational Data to Predict StudentâĂŹs academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 119-136.

2. Amrieh, E. A., Hamtini, T., Aljarah, I. (2015, November). Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.