Students Performance Prediction by Mining Behavioural Pattern

Huanliang Sun*, Shi He and Olaoluwa Esho

School of Information and Control Engineering, Shenyang Jianzhu University, Liaoning, China

*Corresponding author e-mail: sunhl@sjzu.edu.cn

Abstract. Analysing and monitoring students’ progress and performance is an active research area in educational data mining. Some research work uses direct construct relating to academic achievements such as GPA, SAT in predicting performance while others make use of psychometric construct in measuring skill and knowledge, ability and educational achievement from data gotten from questionnaires. In this paper, we propose the use of psychometric construct as a method of extracting non-cognitive features from campus check-ins data in predicting academic performance of college students. A P-FRAME framework was developed and non-cognitive attributes extracted through psychological theories were used as the predicting variables. The extracted variable was applied to various penalized regression algorithm and the results were compared. The result of our experiments showed a high correlation between the actual score and the predicted score which is within 2:0 of reported score.

Keywords: Academic Performance; Psychology; Behavioural Pattern

1. Introduction
In the educational system, the implementation of smart ID cards has been very useful in an administrative task, staff, student and visitor’s management, time/attendance tracking, access control and so on. The mundane tasks of registration, book purchases, meal programs that require an army of administrators to manage piles of paperwork are made simple through different applications and databases incorporated with the smart cards and sensor network. Many higher institutions have adopted the use of smart ID card both with students and staffs because of its cost-effectiveness and its simplicity.

It is important to analyze these data to gain insights that can be useful to the school management. The ability to uncover hidden patterns in large databases will enable higher institution develop models that can be used to predict different desirable outcomes. By acting on these predictive models educational institutions can effectively address issues ranging from performance, retention, transfer, marketing, and intervention. Many factors influence the academic performance of college student. It is a general believe that cognitive attributes are most important in determining academic performance but with on-going research, it has been discovered that non-cognitive attributes may play an important role in reversing or limiting delays in cognitive development and academic achievement and they may complement direct efforts to improve academic learning [1].
The information from the students' smart card reflects various activities and patterns undertaken each day that includes visit to different locations they visit and at Time when swiping the ID card to make purchase, to gain access into a building or check-ins for attendance. Using the location clusters and the time information, it is possible to analyze behavioral pattern and activities of an individual to determine persistence, motivation and behavioral pattern. For example, we can determine the interest or enjoyment of a task or goal. Engaging in reading activity when school is not in section or engaging in a task with a reward attached to it. Cash reward for performing well in a sport competition or marks awarded to a top sport performer which can be classified as a form of motivation.

This paper aims to investigate the underlying patterns within the campus big data and using psychology theories to extract variables which can be used to predict academic performance. The P-FRAME model was developed in predicting academic performance based on the extracted attributes that includes Motivation, behavioral pattern and attitude. A penalized regression algorithm (lasso/ridge, forward stepwise and regression algorithm) was used to test the data set to discover important features and propose a model to be used in predicting the academic performance of college students.

2. Related Works
Early stages of educational research, the transition from high school to college or universities has received much attention[1,2]. There are three important indicators when it comes to predicting future success during this transition period: high school GPA, intelligence quotient (IQ) and self-efficacy. In [3], there was analysis to measure the correlations between the high school GPA and undergraduate grades in the range of 0.26 to 0.53. Also [2] emphasize that high school grades are the best known predictors of student’s readiness for undergraduate studies, regardless of the quality and type of high school attended, while a standardized admission test provides useful supplementary information.

The indicative values of several types of explanatory variables have also been assessed during the transition from undergraduate to graduate studies in predicting academic success. The Indicators of undergraduate achievements that predicts future graduate level performance has been shown in many studies [4]. Applying a factor analysis on school grades, [5] find two components that determine achievement: ability and adaptation to the school system. That second dimension is also identified as student non-cognitive behavior [6], academic ethic [7], or common grade dimension ([8]. It is related to non-cognitive constructs such as motivation, effort, self-efficacy, perseverance, and locus of control. Nevertheless, disagreement has also arisen about the quantification of those constructs [9,10]. Here, we refer to that second dimension as adaptation to the academic culture.

3. Model Framework
This section gives a detailed description of the model developed. The P-FRAME (fig.1) was developed and psychological factors/theories were used to extract attributes for predicting academic performance. The first level of the model is the Data input and this is where part of the preprocessing takes place. At this level, the first stage of preprocessing takes place on the raw data and other additional data that will be useful for the model. The raw data, meta-data or the group data are at the base of the model. At this stage, the cleaning of data, removing incomplete data set, and making sure the overall data is uniform and conforming to the models requirement.

3.1. Meta Data
Meta data contains all the domain information about the data which includes the size, time and date creation. Here the granularity is very important, which helps get in-depth and more structured information.

3.2. Activity and Group Data
Activity data and group data are other data that contains information about extracurricular activities, Dormitory information, Domain information etc. This is the most important stage of this model. This is the level at which reference is made to the psychological factors and theories, which will help us in
extracting the attributes needed for the prediction model. Here, the psychology factors can be predefined or existing theories can be incorporated with the data set with the aim to help select attributes.

The clear understanding of these theories is very important because it gives various dimensions in extracting and attributes labeling. Firstly, these p-factors are used for data filtering before attribute selection. Secondly, they can also be used to discover patterns and also used to give labels to variable with the purpose of better understanding of the features. The last level of the P-FRAME model is where the appropriate algorithm is applied to get a predictive model. After features engineering from the two previous levels, here an appropriate penalized regression algorithm is applied which is because of the nature of the target variable? Section III-B gives the explanations of the algorithms that will be applied to these variables extracted from the previous level. The forward pairwise, ridge, lasso and LARS algorithm uses the features extracted from the level 3 of the P-FRAME model to train the predictive models. There will be more iterative flow between the prediction model and model testing which is dependent on the threshold error required to obtain a good predictive model.

4. Experiment
The data consist of the check-in data of students from the Shenyang Jianzhu university over a period of 4 years, and a total of 8 semesters. There are 4 classes used for this experiment, each class have at least 30 students. The daily usage of student’s id cards is continuously logged into the database when a student activates the card either by buying, checking-in, accessing a building, subscribe for internet, visit the medical center etc. Majority of the data clusters accounted for are at the eating locations across the campus because students tend to activate their ID card more than once while buying food at the cafeteria.

The features extracted from the data-set 36 different features. The experiment was performed on the attributes before normalization and after normalization. From the result obtained using Forward step-wise algorithm, the prediction model developed was able to predict the error in the range of ±0.9 RMS while using 12 attributes. The iterative loop was 666 times. In table 1 we see the best attributes that gives the minimum error. For example if we the top 5 attributes, the error will be ±1.37, if we are to predict a score of 50, it will fall with the range (48.63 to 51.37). Table 2 shows the result of the actual score and the predicted scores with a variance of 0.92.
Table 1. Best attributes of forward step-wise

| Number | Best Attributes | Error | Coefficients |
|--------|----------------|-------|--------------|
| 1      | Em             | 3.183 | 1.2          |
| 2      | Cafe           | 2.259 | 0.004        |
| 3      | Library        | 1.852 | 0.22         |
| 4      | Building       | 1.449 | -0.0019      |
| 5      | Breakfast      | 1.371 | 0.0017       |
| 6      | Early          | 1.310 | -0.20        |
| 7      | Lunch          | 1.339 | 0.023        |
| 8      | Eating         | 1.292 | -0.11        |
| 9      | Bath           | 1.187 | 0.350        |
| 10     | Dinner         | 0.996 | -0.182       |

Table 2. Actual score VS predicted score

| Number | Actual Score | Predicted Score |
|--------|--------------|-----------------|
| 1      | 76.38        | 77.92           |
| 2      | 79.44        | 79.63           |
| 3      | 74.76        | 73.65           |
| 4      | 80.57        | 80.81           |
| 5      | 74.31        | 75.09           |
| 6      | 76.98        | 76.06           |
| 7      | 74.04        | 74.18           |
| 8      | 85.79        | 85.72           |
| 9      | 73.29        | 74.73           |
| 10     | 78.98        | 80.72           |

Table 3. Experiment summary

| Number | Algorithm | Loop Number | Best Model | MSE   | RMS   | Penalty |
|--------|-----------|-------------|------------|-------|-------|---------|
| 1      | F-step wise | 660         | 78         | 0.81  | 0.90  | -       |
| 2      | Lasso      | 10 folds    | -          | 0.54  | 0.76  | 0.085   |
| 3      | LAR        | 1000        | 995        | 0.66  | 0.81  | -       |
| 4      | Ridge      | 182         | 80         | 6.01  | 2.46  | 0.001   |

Various penalized regression algorithms (forward step-wise, LARS, LASSO, Ridged) were used for experimental analysis to determine the best attributes. Form the table 3, Lasso and LARS algorithm gave a better result while ridge regression algorithm performed poorly on the test data. The important variables agree with the ground truth from [69] that indicates breakfast have positive influence on academic performance. Also, the extrinsic motivation influences future outcomes than intrinsic motivation. The Mean of the errors on the test data was used as the measuring metrics and lowest error for the best algorithm (LASSO) is 0.54. The result from this experiment proves that there are non-cognitive factors that contribute to the academic performance of college students.

5. Conclusion
This paper investigated the underlying patterns within the student’s check-in data set. These patterns were extracted by the help of psychological theories and P-FRAME model was developed which helps to integrate the check-in data into suitable algorithms for analysis. Various penalized regression algorithms (forward step-wise, LARS, LASSO, Ridged) were used for experimental analysis to determine the best attributes. The comparison results of the experiments show Lasso and LARS algorithm gave a better result while ridge regression algorithm performed poorly on the test data.
References
[1] Walling D R. College Knowledge: What It Really Takes for Students to Succeed and What We Can Do to Get Them Ready (review)[J]. Review of Higher Education, 2006, 29(4):541-542.
[2] Atkinson R. C., Geiser S. Reflections on a century of college admissions tests. Educational Researcher 2009,38(9):665-676.
[3] Trapmann S , Hell B , Weigand S. The validity of school-grades for academic achievement: a meta-analysis [J]. Ztschrift für Pdagogische Psychologie, 2007, 21:11-27.
[4] Koys, Daniel. GMAT versus Alternatives: Predictive Validity Evidence from Central Europe and the Middle East[J]. Journal of Education for Business, 2009, 85(3):180-185.
[5] Cecilia, Thorsen. Dimensions of Norm-Referenced Compulsory School Grades and their Relative Importance for the Prediction of Upper Secondary School Grades[J]. Scandinavian Journal of Educational Research, 2014.
[6] Bowersa A J. What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school[J]. Educational Research and Evaluation, 2011, 17(3):141-159.
[7] Durand R A. The Academic Ethic and College Grades: Does Hard Work Help Students to "Make the Grade"?[J]. Sociology of Education, 2000, 73(1):19-38.
[8] Lekholm A K, Cliffordson C. Discrepancies between school grades and test scores at individual and school level: Effects of gender and family background[J]. Educational Research & Evaluation, 2008, 14(2):181-199.
[9] Juergen K, Schweiger A H, Michael B, et al. To replicate, or not to replicate - that is the question: how to tackle nonlinear responses in ecological experiments[J]. Ecology Letters, 2018, 21.
[10] Hans, Bonesrønning, Leiv. How Much is Students' College Performance Affected by Quantity of Study?[J]. International Review of Economics Education, 2012.