What can we learn from college students' network transactions? Constructing useful features for student success prediction.

Ian Pytlarz
Purdue University
West Lafayette, IN
ipytlarz@purdue.edu

Shi Pu*
Purdue University
West Lafayette, IN
spu@purdue.edu

Monal Patel
Purdue University
West Lafayette, IN
patel@purdue.edu

Rajini Prabhu
Purdue University
West Lafayette, IN
rajini@purdue.edu

ABSTRACT

Identifying at-risk students at an early stage is a challenging task for colleges and universities. In this paper, we use students' on-campus network traffic volume to construct several useful features in predicting their first semester GPA. In particular, we build proxies for their attendance, class engagement, and out-of-class study hours based on their network traffic volume. We then test how much these network-based features can increase the performance of a model with only conventional features (e.g., demographics, high school GPA, standardized test scores, etc.). We labeled students as "above median" and "below median" students based on their first term GPA. Several machine learning models were then applied, ranging from logistic regression, SVM, and random forests, to AdaBoost. The result shows that the model with network-based features consistently outperforms the ones without, in terms of accuracy, f1 score, and AUC. Given that network activity data is readily available data in most colleges and universities, this study provides practical insights on how to build more powerful models to predict student success.

Keywords

Student success prediction, Engagement, Attendance, Study time, Network activity.

1. INTRODUCTION

Students' academic performance is of interest for important practical reasons. To start with, one's college GPA is related to an individual's labor market performance [9, 16] and future educational pursuits [3]. More importantly, studies have shown that academic performance, especially in the early stage, is a strong predictor of students' retention [1, 5, 11]. Therefore, it could be used to identify at-risk students.

Unfortunately, predicting students' early academic performance is a challenge, essentially because it is difficult to obtain informative data. In this study, we propose to use students' on-campus network traffic volume to infer their location and behaviors. Through the inferred location and behavior, we construct several features that have been shown to be related to students' academic success, namely, attendance, in-class engagement, and out-class study effort. We then demonstrate that including these features into predicting models will improve the model performance in all conventional performance metrics.

Specifically, our research questions are:

1. How accurate is the location inferred from students' network traffic?
2. How much gain could we obtain by incorporating students' network inferred behavior in predicting their academic success?

2. RELATED STUDIES

Empirical studies have accumulated considerable evidence on the effect of attendance, engagement, and study time on a student's academic performance. The most rigorous literature comes from the Economics discipline, where experimental or quasi-experimental designs were applied. To name a few, in a randomized experiment, Chen and Lin [6] found that attendance increases students' final exam course grade by 9.4% – 18%. In another field experiment, Marburger [14] showed that mandatory attendance policy improves exam performance through reducing absenteeism. Using an instrumental variable approach, Stinebrickner and Stinebrickner [18] showed that college students' study time has a positive impact on their first year grade. In another study, Andrietti and Velasco [2] used first difference to remove time-invariant confounding variables, such as ability, in the estimating of effects of study time. They also found that study time had a large impact on students' final grades in two econometrics courses. Credé, Roch, and Kieszczynka [7] conducted a recent meta-analysis on the effect of attendance on grades. They found that attendance has strong relationships with both course grades and GPA.

In correlational studies, the well-cited work by Kuh, Cruce, Shoup, Kinzie, and Gonyea [13] showed that the time spent studying per week and the engagement in educational purposeful activities like asking questions in class are positively correlated to a student's first-year GPA. In a recent literature review, Trowler [19] concluded that studies in engagement in general found it to be positively correlated to student learning.

Though significantly correlated with academic performance, a student's behavioral data is difficult to obtain. Recent effort usually relies on measuring individuals' interaction with the learning management system as a proxy for their study effort [4, 8, 12, 15, 17]. Such practice has value, especially for the courses that are pre-

* Shi Pu is the corresponding author.

dominantly online. However, when the interested population are students taking courses on a traditional campus and learning management systems are mainly used as a mean to disseminate lecture notes and collect homework, interaction with the learning management system is unlikely to be an informative proxy for study effort.

To our best knowledge, this paper is the first study to use network traffic to build meaningful features in predicting students' academic success. A few previous works have demonstrated the possibility of inferring students' attendance through smartphone GPS and WiFi connections [10, 20, 21]. In general, they use smartphones to track individuals' location and check if students appear to be in class when they should be. These studies shed light on an innovative approach to collect real-time students' attendance data. However, all of them involve installing a third-party software, which provides an extra roadblock for scaling. As we will demonstrate later, students' attendance can also be inferred from their on-campus network traffic. This approach utilizes the existing network data, thus is arguably more scalable.

3. DATA AND METHOD

The study utilizes the data collected for an advanced learning analytics endeavor at Purdue University, namely Academic Forecast[1]. The project built cutting-edge machine learning models for students' course performance and accumulative GPA. Academic Forecast intends to identify student behaviors that are positively correlated with their academic performance and to encourage students to increase such beneficial behaviors. Though utilizing a part of the data from Academic Forecast, the models we experiment in this study are not directly related to the ones implemented for Academic Forecast.

The study utilize students' individual-level administrative and network traffic data from Purdue University[2]. The sample included all first-time, full-time freshmen that entered the university in fall 2017, with 7555 students in total. The response variable of interest is a student's fall semester GPA. The response is coded as 1 if a student's GPA is larger than the median, 0 otherwise. Notice that the choice of median ensures that the label is balanced. The network traffic volume provides two pieces of important information about students: 1) a student's approximate location (the campus building name) when s/he is connected to the network, and 2) a student's network traffic volume during a time period.

The first research question concerns how accurate the network inferred location is. To validate the inferred location, we need some form of ground truth. Fortunately, as many first-year students live on campus in Purdue, we can safely assume that most students should be in their residential buildings during early morning hours. Thus, we can compare the network-inferred location with students' on-file residential buildings[3].

The second research question concerns the contribution of network-inferred behavior data to prediction models. The follow paragraphs will briefly cover the construction of the network-inferred behaviors.

A student i is considered attending a registered course j's session k if the student appears to be in the building where the session k is held during the class time. Then, the average attendance rate for student i in the first semester is inferred by averaging student i's attendance across sessions and across all courses:

$$attend_{ijk} = \begin{cases} 1, & if\ bld_{it} = bld_{jk}, t \in [jk\ start, jk\ end] \\ 0, & o.w \end{cases}$$

$$attend_i = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{m_j}\sum_{k=1}^{m_j} attend_{ijk}\right)$$

Note that n is the total number of courses a student i has in the first term. m_j stands for the total number of sessions for course j. bld_{it} indicates a student i's campus building id at time t , and bld_{jk} indicates the campus building id for course j at session k. Essentially, $bld_{it} = bld_{jk}$ if and only if a student i shows up in the class building during the scheduled class time.

A student's out-class study time is approximated by the total time spent in buildings that are predominantly used for learning purposes (indicated by bld_{study} in the formula), for example, libraries and active learning centers. Out-class time is obtained by excluding the time when a registered course is taking place. Formally:

$$study_i = \sum t \times \left(bld_{it} == bld_{study}\right)$$

Where t does not belong to any scheduled class time for student i. Note that $bld_{it} == bld_{study}$ if and only if a student i is in a "study related" building at none-class time t.

In-class engagement for a student i in course j session k is inferred by the network traffic volume a student has during that class session. The average in-class engagement during the first semester is again averaged across sessions and across all courses. Noting that the higher the traffic volume, the more likely that a student is dis-engaged[4] in the class:

$$eng_i = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{m_j}\sum_{k=1}^{m_j} eng_{ijk}\right)$$

The network-inferred behaviors, along with a set of pre-college variables, were then fed to several common machine learning algorithms to predict if a student is going to score higher than the median. The common pre-college variables include high school GPA, high school quality, standardized test scores, gender, residency, race, etc. A 20-fold cross-validation is applied to

[1] Website: https://www.academicforecast.org

[2] The scope and procedure of this study strictly follow a proved IRB. All of the analysis of the data occurs within the existing Purdue data security infrastructure and guidelines controlling data utilized for campus daily operations. Data can only be accessed via a machine controlled by Purdue data security protocols.

[3] Students' locations after the late night and before early morning were never used in any of our predictive models, due to potential privacy concerns. However, for the purpose of validating the merit of network inferred index, we checked *at the aggregate level* if students' night locations agree with their residential buildings on the book. We did not further investigate which students' inferred location and theoretical location did not match.

[4] This is not necessarily true for classes that entail the use of internet.

estimate the model performance on unseen data. All models used pre-defined hyper-parameters to avoid being over-optimistic on performance estimation.

4. RESULT

To validate the accuracy of our network-inferred location, we choose an early Tuesday morning in September 2017 that is neither a public holiday nor a university holiday. Recall that students should be in their dormitory rooms at this time, thus their on-file residential building could be used as a ground truth to validate our network inferred location.

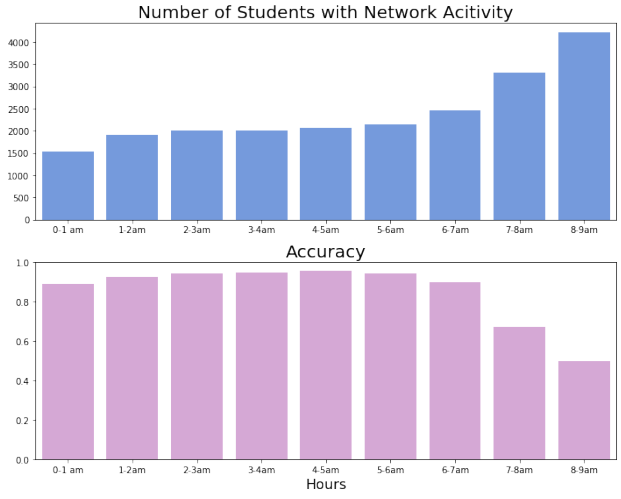

Figure 1. Validate network inferred location

Figure 1 demonstrate the result of this validation. As shown in the upper part of the graph, only a portion of all students living on campus have network activity before 6am, and the number increases rapidly after 7am. The lower part of the graph demonstrates the accuracy of the network inferred location. The accuracy is defined as the percentage of students whose on-file residential building agrees with the network inferred location. As we expected, the accuracy is high during the early morning, ranging from 89.20% - 95.70% between 0 to 6 am. The accuracy dropped rapidly after 7am; this plunge is likely due to the fact that students start to leave their residential buildings, thus it can no longer serve as a ground truth.

Students' on-file residential location never fully agrees with the network-inferred location. This does not necessary suggest that there is some noise in the inferred location, as we cannot be fully sure that all students are in their residential buildings at any given time. However, this result indicates that network-inferred location should be a good proxy for a student's real location. Thus, it can provide useful information on students' behaviors.

In Table1, we compare the classification accuracy between models with network features and the ones without. A 20-fold cross-validation is applied to estimate the model performance on unseen data. As the label is balanced (exactly 50% of students score higher than the median), accuracy serves a good performance metric. We experiment on several common algorithms to check if the performance gap is model dependent. All models used pre-defined hyper-parameters.

As shown in Table 1, models with network-inferred behaviors consistently outperform the models without network-inferred behaviors. The difference in accuracy ranges from 0.016 to 0.021. The right-most column records the t-statistics [5] for improved accuracy. The improvement is statistically significant at 0.05 level with one-side t-test for logistic model, random forest model, and AdaBoost model. The improvement on SVM model is only significant at 0.1 level. After including Bonferroni correction, only the improvement on AdaBoost remains statistically significant.

Table 1: Accuracy comparison : with/out network behaviors (t-test with Bonferroni correction, α = 0.05)

Classifier	No Network Behaviors	Network Behaviors	Diff	t-stat
Logistic	0.669 (0.02)	0.686 (0.03)	0.017	2.01*
SVM	0.667 (0.03)	0.683 (0.03)	0.016	1.67
Random Forest	0.678 (0.03)	0.696 (0.03)	0.018	1.96*
AdaBoost	0.676 (0.03)	0.696 (0.03)	0.021	2.39*

Note: standard errors in parentheses

Table 2: Model performance comparison: with/out network behaviors

Classifier	Network	F1	Precision	Recall	AUC
Logistic	No	0.68	0.654	0.711	0.732
	Yes	**0.696**	**0.671**	**0.724**	**0.751**
SVM	No	0.672	0.661	0.687	0.726
	Yes	**0.683**	**0.678**	**0.691**	**0.747**
Random Forest	No	0.679	0.671	0.688	0.735
	Yes	**0.687**	**0.702**	**0.674**	**0.761**
AdaBoost	No	0.67	0.675	0.668	0.734
	Yes	**0.690**	**0.698**	**0.682**	**0.757**

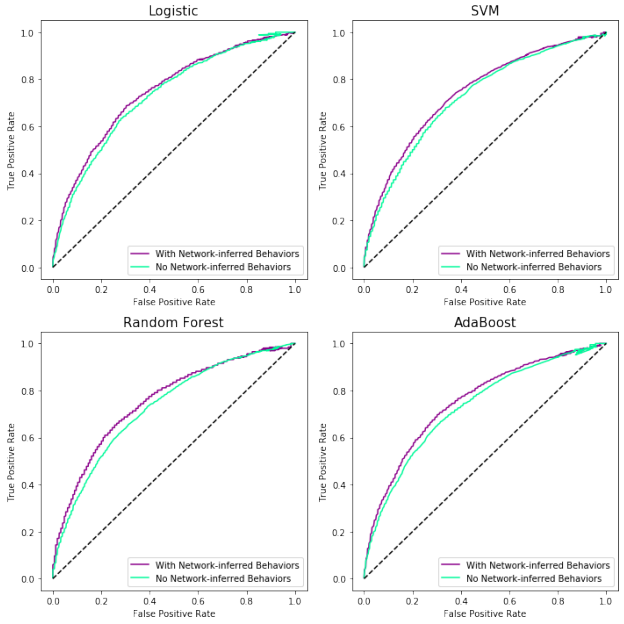

Figure 2. Mean ROC curves for different models, with v.s. with-out network-inferred behaviors

[5] Paired sample t-tests are used here to compare the difference between models with network features and the ones without.

Table 2 and Figure 2 report further performance comparison. Models with network-inferred behaviors again consistently outperform the models without such information on F1 score, precision, recall, and AUC. The differences are small but consistent in the ROC curves.

At last, Table 3 demonstrates the top 10 important feature importance in Random Forest and AdaBoost. Network-inferred behaviors are always among the top important features. In AdaBoost, student engagement is the single most important feature, followed by students' high school GPA, high school quality[6], study time, and attendance. The Random Forest relies disproportionally on students' high school GPA. Students' attendance, study time, and in-class engagement are more informative for the model than the rest predictors. Note that network-inferred behaviors are always more important than standardized test scores[7] in the two models.

Table 3: Feature importance[8]

Random Forest		AdaBoost	
Feature Name	Importance	Feature Name	Importance
High school GPA	0.385	**Engagement**	0.224
Attendance	0.157	High school GPA	0.145
Engagement	0.107	School quality	0.143
Study time	0.106	**Study time**	0.143
Std test score	0.091	**Attendance**	0.141
Zip code income	0.067	Zip code income	0.120
School quality	0.061	Std test score	0.059
Female	0.009	International	0.009
International	0.009	Asian	0.006
Asian	0.005	Hispanic	0.004

5. DISCUSSION

This study proposed a novel way to utilize on- campus network traffic data to improve student success prediction models. In particular, we have demonstrated that network-inferred location is a good proxy for students' actual location. Experimenting on a randomly chosen early morning, we found that 89.20% - 95.70% of students' network-inferred location matches their on-file residential location. In addition, we demonstrate that including the network traffic data improves the model performance in conventional performance metrics. Interestingly, the improvement is consistent across different models, ranging from the basic logistic regression models to more complicated ensemble classifiers.

The network-inferred behaviors are rooted in existing literature on student success. Namely, attendance, engagement, and study time have been found to be related to a student's GPA in various researches. Therefore, we believe the result should not be a peculiarity in Purdue's data but can be generalized to other colleges and universities.

In addition to generalization, our approach has two important practical advantages. First, models based on network-inferred behavior provide actionable suggestions for student advisors. To elaborate, the pre-college predictors can only tell advisors whether a student is well-prepared for college. Other analytical models usually only inform the advisor how well a student is doing in each class. Neither type of predictor could provide suggestions on *why* a student is having trouble. The network-inferred behaviors, on the contrary, could possibly pinpoint the student's action that directly leads to their poor performance, e.g., poor attendance. Second, network-inferred behaviors are based on existing network data in each university, thus the scaling cost is arguably low.

The study, nevertheless, has several important limitations. First, the chosen response is median GPA instead of more meaningful classifications, e.g., retention and academic probation. Therefore, it is unclear if the network features are still informative for detecting at-risk students. Second, the improvement in accuracy is limited. Future study should seek to uncover deeper pattern from the location data to improve model performance.

6. REFERENCES

[1] Allen, J. et al. 2008. Third-year college retention and transfer: Effects of academic performance, motivation, and social connectedness. *Research in Higher Education*. 49, 7 (2008), 647–664. DOI:https://doi.org/10.1007/s11162-008-9098-3.

[2] Andrietti, V. and Velasco, C. 2015. Lecture Attendance, Study Time, and Academic Performance: A Panel Data Study. Journal of Economic Education. 46, 3 (2015), 239–259. DOI:https://doi.org/10.1080/00220485.2015.1040182.

[3] Astin, A.W. 1993. What matters in college? Liberal Education.

[4] Brinton, C.G. and Chiang, M. 2015. MOOC performance prediction via clickstream data and social learning networks. *2015 IEEE Conference on Computer Communications (INFOCOM)* (2015), 2299–2307.

[5] Cabrera, A.F. et al. 1993. College Persistence: Structural Equations Modeling Test of an Integrated Model of Student Retention. *The Journal of Higher Education*. 64, 2 (1993), 123. DOI:https://doi.org/10.2307/2960026.

[6] Chen, J. and Lin, T.F. 2008. Class attendance and exam performance: A randomized experiment. *Journal of Economic Education*. 39, 3 (2008), 213–227. DOI:https://doi.org/10.3200/JECE.39.3.213-227.

[7] Crede, M. et al. 2010. Class Attendance in College: A Meta-Analytic Review of the Relationship of Class Attendance With Grades and Student Characteristics. *Review of Educational Research*. (2010). DOI:https://doi.org/10.3102/0034654310362998.

[8] Jiang, S. et al. 2014. Predicting MOOC Performance with Week 1 Behavior. *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*. (2014), 273–275.

[9] Jones, E.B. and Jackson, J.D. 1990. College Grades and Labor Market Rewards. *Journal of Human Resources*. 25, 2 (1990), 253–266. DOI:https://doi.org/10.2307/145756.

[10] Kassarnig, V. et al. 2017. Class attendance, peer similarity,

[6] Measured by the average Purdue GPA for students come from that high school.

[7] Constructed based on SAT and ACT scores.

[8] A feature's importance in Random Forest is the average decrease in impurity by that feature across all trees, the higher the better. AdaBoost's feature importance depends on the base learner, which are decision trees in this study. Therefore, its feature importance is again calculated by averaging the decrease in impurity by each feature.

and academic performance in a large field study. *PLoS ONE*. (2017). DOI:https://doi.org/10.1371/journal.pone.0187078.

[11] Kern, C. et al. 1998. Correlates of College Retention and GPA- Learning and Study Strategies, Testwiseness, Attitudes and ACT. *Journal of College Counseling*.

[12] Kloft, M. et al. 2014. Predicting MOOC Dropout over Weeks Using Machine Learning Methods. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (2014), 60–65.

[13] Kuh, G.D. et al. 2008. Unmasking the Effects of Student Engagement on First-Year College Grades and Persistence. *The Journal of Higher Education*. 79, 5 (2008), 540–563. DOI:https://doi.org/10.1353/jhe.0.0019.

[14] Marburger, D.R. 2006. Does Mandatory Attendance Improve Student Performance? *The Journal of Economic Education*. 37, 2 (2006), 148–155. DOI:https://doi.org/10.3200/JECE.37.2.148-155.

[15] Phan, T. et al. 2016. Computers & Education Students ' patterns of engagement and course performance in a Massive Open Online Course. *Computers & Education*. 95, (2016), 36–44. DOI:https://doi.org/10.1016/j.compedu.2015.11.015.

[17] Sinha, T. et al. 2014. Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions. (2014).

[18] Stinebrickner, R. and Stinebrickner, T.R. 2008. THE CAUSAL EFFECT OF STUDYING ON ACADEMIC PERFORMANCE. The BE Journal of Economic Analysis & Policy. 8, 1 (2008), 14. DOI:https://doi.org/10.1017/CBO9781107415324.004.

[19] Trowler, V. 2010. Student engagement literature review. *Higher Education*. (2010), 1–15.

[20] Wang, R. et al. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones. DOI:https://doi.org/10.1145/2632048.2632054.

[21] Zhou, M. et al. EDUM: Classroom Education Measurements via Large-scale WiFi Networks. DOI:https://doi.org/10.1145/2971648.2971657.