Colorado State University

Grade Point Average Prediction Based on Student's Lifestyle

Jim Xu, Minzhuo Jin, Ziran Min

Department: Computer Science

Course：Introduction to Distributed Systems

Instructor: Shrideep Pallickara

Date: 22 April 2017

Introduction

College life is colorful and meaningful for every single student. During this stage, students learn knowledge to achieve their academic success as a main goal. At the same time, they engage in various kinds of extracurricular activities to find their interests, get in touch with society and make friends to form their own personality, and finally they develop different lifestyles during their college life. Students want to seek a balance between extracurricular activities and academic success. In the paper “Relationship between Undergraduate Student Activity and Academic Performance” released by Purdue University, AmyL. argues that “Participation in student organizations can lead to the development of social and leadership skills, higher retention rates, heightened self-confidence, improved satisfaction with college, the ability to see course curriculum as more relevant, and further success after college.” (AmyL, 3) Though attending activities to a certain extent is help for good academic performance, how to manage the schedule is also meaningful for university students. Many of the students could not figure out what kind of lifestyle they live will have a potential influence on their academic performance. There are many factors affecting how a student performs academically. For example, it is commonly believed that a student who doesn’t go partying a lot and sleep well will probably do better in study. However, what roles these factors play in student’s academic performance are hard to understand thoroughly. Our problem is derived from those issues and tries to figure out how those potential factors contribute to individual’s Grade-Point Average (GPA). One common limitation on survey-based studies is that the students may not be able to give precise information about their all day activities. Fortunately, smartphones nowadays are equipped with various sensors, which can be used to monitor their owner’s activities on a regular basis. Valuable information can be extracted from those sensor data. In our research, we use distributed system and machine learning algorithms to process a large amount of sensor data as well as survey results and other data and estimate how a student will perform based on his or her behavior, emotion state, etc. In our project, we are processing a large amount of data presented above to construct an attribute list for each student in the sample. The main dataset we are using is “StudentLife Dataset” (*"StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones."*). This dataset includes behavior-related data from students at Dartmouth. The types of data include:

1. Smartphone sensor data: these are collected from smartphone sensors: accelerometers, light sensors, microphone, etc.
2. GPS location logs: these enable us to find out both students’ indoor and outdoor activity and other information such as when and where they eat.
3. Other smartphone data: these include mainly app usage, phone charge log, screen lock and unlock log, etc.
4. Survey data: these can be used to evaluate students’ mood, stress, personality, etc.
5. Academic performance: the main indicator for academic performance we will be using in this research is the term GPA. The dataset also provides other indicators such as cumulative GPA.
6. Academic activity data: these include activities on piazza, class information, deadlines, etc.

This paper makes the following contributions. First, we use decision tree regression algorithm to train a model that estimates a student’s academic performance based on his or her behavior. Second, we are able to identify some key factors in affecting academic performance.

Further, we can also provide suggestions based on our algorithms for students orientating a reasonable lifestyle to achieve their academic goal. The resulting model can be extremely useful in helping college students to find a balance between study and other activities.

Problem characterization

The dataset we are using in this analytical task, “StudentLife Dataset” ("StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones."), is large in size (about 6.2 GB). Data are stored in various forms and schemas. The first challenge of our project is to dealing with different file forms in the given dataset. For example, some files are consisted of information for all participated students in this program while some files only contain a single student’s information. What’s more, some files are in csv format, others are in json and SQLite format. It is challenging to extract useful information from those data files. We need to compute statistics that can reflect students’ behavior in many aspects by processing this huge amount of data. Those statistics are important because they are essential to produce a high-performance final model. The second challenging task is to merge different behaviors information of a single student and construct an attribute vector (lifestyle) for each student for our machine learning model. Considering students’ behavior information are discontinuous, distinctive and arbitrary datasets which are scattered in different files, how we approach this task may have a huge impact on the overall performance of this system. We must be careful to use Spark’s API properly to avoid any unnecessary big performance issue. These first two tasks are also important in that how they are implemented will make a difference in the convenience in tweaking the model later. Through this process, we are using Spark’s built-in Dataset and DataFrame data structure (“Spark SQL, DataFrames and Datasets Guide.”). These are the facilities Spark provides to deal with structured data files, database tables or existing RDDs. They are distributed and lazy-evaluated data structures which come with the handy feature of RDDs. Plus, they support SQL queries and basic visualizations which can be helpful in gaining some insights in data. However, these structures are very versatile, which introduces challenges in manipulating these data structures efficiently and gracefully.

Dominant approaches to the problem

One successful approach to analysis the dataset is proposed by Dartmouth Studentlife study group, which is the provider of our Dataset. They successfully built a model based on linear regression with lasso regularization to accurately predict cumulative GPA, which results in a deviation within ±0.179 of the reported grades. The main advantage using linear regression algorithm is that the results are highly built on the changes of each components during the time series, which means they put more emphasis on the trend of data changes instead of the range of data. That makes the predicted GPA more accurate and reasonable than using decision tree. Another advantage of the approach is that it is more sensible to the data changes across time period than other approaches. According to the algorithm they provided in their paper (refer to paper), changes between two successive data field will be amplified by the resulting number and then have an large scope of accumulated effect on the predicted GPA. However, in our implementation, we choose to put more emphasis on the value of each given component instead of the changes of value. As a result, it may shrink the effect of changes in data especially in significant time period, but it could also exaggerate the effect of no significant data. Besides, their model can easily detect which component plays a more contributing role to the expecting result. Based on their two advantages discussed above, when there is a change for a given component and the resulting GPA deviates the normal value, we can easily infer that specific component has a larger influence to the outcomes and vise versa. During the machine learning operation, those coefficients will be recalculating until it approaches the reasonable and realistic value. However, one advantage that decision tree possesses over the linear regression with lasso regularization is the first one has the ability to take all contributing factors into account even if they are discontinuous, deviated data, which could not be easily used in pure linear regression algorithm (refer to paper).

Methodology

As an overview, there are 4 main steps to achieve our goal. Firstly, we extract metrics such as average, standard deviation, and variance from relevant sensor data that reflect different aspects of students' lifestyle. Secondly, we create a feature vector to store the above attributions for each sample student. In addition, the feature vectors also include students' response to EMA surveys and activity summary on Piazza. These feature vectors together with the label column (GPA) forms the table that serve as the training and testing data for next steps. Thirdly, we use the Random Forest regression algorithm to process most feature vectors to get a resulting model. Lastly, we use the result model to process the remain feature vectors. Then we evaluate the resulting model with metrics such as root mean squared error.

The sensor data we make use of in the analysis are inferred from raw data captured by microphones, accelerometer and light sensor on students' smartphone.

- Physical Activity
    - Information in these data files are students' continuous activity inference (explained below) with corresponding timestamps.

  - Activity inference are integer values from 0 to 3 that are assigned with different meanings. 0 is for “stationary”, 1 is for “walking”, 2 is for “running”, and 3 is “unknown”. As we can see, higher number excluding 3 indicates more active state.
  - We first filter out the entry with unknown states. The remaining data are used to extract each student’s average activity inference, its standard deviation and variance. Higher average means that the student is more active.

- Audio
  - Data in this category are similar to physical activity data explained above. Audio inference indicates how noisy the environment the phone is in. With the unknown inference number filtered out, higher number indicates noisier environments.
  - Similarly, we extract the average noise level, its standard deviation and variance for each student. Higher average means that the student spend more time in a noisy environment.
- Dark
  - Data in this category have entries consisting the start and end time for each time interval when the phone is in a dark environment for more than 1 hour.
  - We transform the data such that each line represents the time duration for each recorded dark period. Then, we extract the average of time duration for each student. Higher average means that the student spends longer time in dark environments.

- Conversation
  - Data in this category are similar to dark data explained above except now it tells the start and end time for each recorded conversation.
  - We extract the average conversation duration for each student. Higher average means that the student spends more time for each conversation.

We also analyzed system logs about phone charging time and phone lock time logged by students' smartphone.

· Phone Charge

o These data files include the start and end time of each charging cycle that is more than 1 hour.

o We extract the average phone charging duration for each student and its standard deviation.

· Phone Lock

o These data files include the start and end time of each period when the phone is locked for more than 1 hour.

o We extract the average phone lock duration for each student and its standard deviation.

The EMA survey responses reflect students' impression about their exercise, sleep, stress, study space and class quality periodically.

· Class

o Students' answer to the questions:

§ How you enjoyed the class?

§ How many hours did you spent on coursework outside class since the last class?

§ Do you have a due today?

· Stress

o Stress level

In Moneta research, he insists that positive affect was associated with higher grades and and GPAs, while negative affect during the second half of the semester was associated with lower grades and GPAs. In our research, it also provides the result about the relation between stress and GPA.

· Sleep

In Kelly's research, "long sleepers reported significantly higher GPA" (Kelly 85). In the response, we can extract the following information

o Number of hours slept last night

o Quality of the sleep last night

These data together with the logged phone charging and dark duration can reflect the quality of sleep

· Exercise

In the article "Vigorous Exercise Linked with Better Grades", Tara argues that "College students who want to boost their grades can start by boosting their level of exercise." (Tara, 1) A number of people agree with this point, however, whether this point is convincing, we are able to use our research result to explain this.

- Exercise duration
- Exercise intensity

We extract the average value and standard deviation of these responses for each student.

Finally, we join all the extract metrics with the activity summary for each student on Piazza, which includes (1) number of days the student logged in the class page, (2) number of posts the student has viewed, (3) number of posts, responses, edits, etc., (4) number of questions the student has asked, (5) number of notes the student has posted, (5) number of questions the student has answered.

All the metrics we extract are put into the features vector. The data table used for machine learning has 3 main attributions, Uid (student identifier), GPA (Grade-Point Average) and features (calculated based on the metrics). The content of table in detail is as follows.

Uid	GPA	feature
u02	4.0	10340.1089743589...
u04	3.5	14496.5518672199...
u08	3.333333333	8540.22463768116...
u10	3.777777778	9943.01661129568...
u14	3.888888889	11295.1382488479...
u16	4.0	10656.0371900826...

Previous research presented a prediction model get from linear regression fitting algorithm. We realize that some features of dataset may not have a great impact on students' academic performance, which leads us to consider fitting a model with decision tree algorithm, where less relevant features are excluded in the predicting process. The reason we switched to random forest algorithm is that we are having over fitting issue. Random forest algorithm uses multiple decision trees at training time. In addition, it produces a model where several different decision trees are used and their predictions are averaged to compute a final prediction. Random forest algorithm alleviates decision trees' tendency of over fitting to their training set.

Experimental Benchmarks

To evaluate our model, we take different evaluation aspects into account. First of all, we use the mean absolute error (MAE) to measure the accuracy of outcome prediction. Mean absolute error (MAE) is a popular measurement of accuracy which represents an average of the absolute differences between observed value and predicted value. The mean absolute error is an average of the absolute errors e_{i}|=|y_{i} - x_{i}|, where y_{i} is the prediction and x_{i} is the true value. A smaller MAE indicates that the predictions are closer to the real observed values. In our prediction model, we used an array of observation attributes to predict the GPA and compared it with students' real outcome GPA. The MAE of our prediction is 0.287, which means the average difference between our predicted GPA and the eventual GPA of a student is ±0.287. And the MAE fluctuates in ±0.15 range because the algorithm randomly choosing learning subsection and testing subsection, where a evenly distribution of learning subsection plays an important role on how the prediction model behaviors. Second, we also find that our model has a good performance considering the algorithm throughput. It takes 54.843 seconds to extract useful data attributes from a total of 2.57G files. The throughput of our extraction algorithm is 47.99 MB/s. Then, the extracted dataset was randomly splitter into 2 parts, one for training and one for evaluating. We used the typical split ratio 0.7:0.3 on the random split. The results showed that 1.933 seconds were used on training a prediction model based on random forest algorithm using training dataset, and 1,224 seconds on evaluating the predicted model using testing dataset. However, there are some benchmarks we did not achieved during our method. First, we did not calculate the coefficient of correlation (r value) for each attributes in the feature, that means we can not have a statistics idea of how and what depth an attribute contributes to the eventual prediction of our result. If the r value is a high positive number, that means the specific attribute in feature array have a larger portion of influence. Another thing is that we data we used is not objective. For example, in the EMA dataset, all the data are collected by Q&A, that makes the data less objective and reliable. However, some of those data played an important role in training our prediction model. Given that, we should use scientific monitoring and detecting methods to obtain those data, which is an potential improvement of our methodology will be discussed further in the next part. Overall, our method has an expected performance using the dataset, and the predicted results are in acceptable deviation range.

Insights Gleaned

During the project, several things that we did not thought of before we started became a crucial factor that influence our prediction accuracy. First thing is that our dataset had really bad performance under decision tree algorithm because the number of students' sample in our dataset is small. However, the large data attributes relating to each student makes our model much more complex to analysis and thus causes over fitting problem. Compared to decision tree, forests of trees splitting with oblique hyper planes. If randomly restricted to be sensitive to only selected feature dimensions, can gain accuracy as they grow without suffering from overtraining[Tin]. The random forest algorithm has a better performance than decision tree when dealing with small data samples while with large data attributes. By taking the feature of random forest, we can correct the decision trees' habit of over fitting to the training set and provides us with a more accurate GPA prediction. Second, we came to know that different methods for dealing with different types of data have significant impact on the error rate of our result. For example, in the field of physical activity, the prediction error rate turns out to be the lowest when taking the standard deviation and average value of activity durations into account. In the contrast, the sum of durations in the field of sociability results in smaller prediction error rate. What's more, it seems there is no correlation between class attendance and academic performance. That means we need to find a better reasonable and describable way to deal with the data in order to achieve the more accurate prediction results. It is especially the case considering we using random forest algorithm, since we need to manually feed in a feature array describing students' lifestyle. And the feature array must be well designed using different mathematical algorithm to describe the characteristics. Lastly, when analyzing the outcomes of our predictions, we used a view table built on DataFrame to provide a intuitive graphical interface. The view table supports more operations than data frame, such as database operations, graphic presentation distribution analysis. Instead of viewing results as pure number sets, using a view table helps us gain a more depth understanding of how our algorithm performs in the given dataset.

How the problem space will look like in the future?

Knowing factors which have effects on GPA is always meaningful to university students, not only those who pursue higher GPA, but also those who suffer from poor one. In order to improve GPA, students can adjust their lifestyle based on the relative factors. For instance, if students know that more workout time may help to GPA, they will do more sports in the future. In addition, except students, a number of researchers and scientists may also benefit from our research result. The result is able to give them more inspiration. For example, sociologists may want to figure out the relation between higher GPA and communication time. Though our project is able to predict a decent result and the value of RMSE is small, there still have room for improvement. The improving methods are referring to the size of dataset, a more appropriate algorithm, advanced detection technique and other influence factors. To begin with, though our prediction value is close to the real value, our dataset is to small, since not every one want to share their lifestyle information. If possible, we can find more volunteers who are from different country, region and geographic condition. In this case， the research result may be more universal. In addition, this is also able to help to minimize the RMSE. Furthermore, improving the algorithm we applied is also significant. We can achieve this by emphasizing on the scope of data during the whole time period as well as the trend of data changes across time. Given that, the GPA predictions are able to take the student's benchmark and his changing trends into account at the same time. That will make our prediction more accurate and reasonable compared to the current implementations. Lastly, by taking advantage of the technology development, smartphone application can be helpful to take in new data records from students. Some data which could not be subjective expressed can be possibly recorded by machine sensors. At the same time, our prediction model could take in more up-to-date data records based on the real time transportation and that enriches our sample size for machine learning and make our prediction more accurate. Then the machine learning modification can keep updating into a more suitable one for students based on the current world. Also, we can take feedbacks from students to further perfect our decision tree model as well as avoid some unexpected situations. Third, more influential factors should be taken into account. Since student academic performance is a very complex model, even student’s lifestyle is one of those many influential factors. We plan to connect a student with other potential stakeholders like teachers, families and friends by sharing data. By introducing new interventions into students' academic performance, we could probably figure out a new complex way to build the relationship model for the GPA prediction and find out more interesting relationship among them.

Conclusions

A conclusion is not a summary. You must make a set of assertions about your work.

Through this analysis task, we realize that despite the fact that some factors are considered more important on students' academic performance, all the features we collect play a role in the final random forest model. In the forest, we can see various features vector components are used to determine the student's GPA. That means multiple factors should be considered to predict a student's academic performance. If we are able to access more students' data, we can make a more accurate model that predicts student's academic performance. Such data can be collected from students' smartphones and logs of online course assistance software such as Canvas and Piazza. Furthermore, data collected from e-learning websites can be immediately used for training the model. Those websites are a good source of data because they are objective and rich in detail. From those websites, we can know about the information about the course arrangement. These includes how often the assignments are published, what the due dates are, whether it is a big class or small class. Plus, students' most recent grade can be accessed as well as the points for each assignment and quiz. We can also investigate how active students are communicating with peers and instructor on those online platforms. All these data can be analyzed to form a features vector for regression algorithms. After several iterations, the model can be used to predict the students' GPA based on his behavior and performance. Systems can be set up to alert the students whose lifestyle may have a potential negative effect on academic performance. Students can be informed whether they are having serious sleep deviation that may cause negative effects on academic performance, or whether they are lacking physical activities, partying too much, etc.

Educators can also benefit from this system because it can be used to reveal the relations between those factors with academic performance. Machine learning algorithms can be used to identify a student's stress level, how engaged or interested the student is in certain courses. All the information can be used by instructors and administration to find out the potential problems in the course arrangement. For example, they can find out when most students are stressed out too much and when challenging works can be required to make the course more interesting.

Our project's methodology can be used to build better learning management systems where machine learning algorithms are used on distributed dataset to perform analysis tasks that evaluate registered students' lifestyle and give feedbacks.

Bibliography

Hawkins, Amy L., "Relationship between Undergraduate Student Activity and Academic Performance" (2010). *College of Technology Directed Projects*. Paper 13. Web. 23 Apr. 2017. <http://docs.lib.purdue.edu/techdirproj/13>

Wang, Rui, Fanglin Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, and Tia Zhou. "StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones." *In Proceedings of the ACM Conference on Ubiquitous Computing* (2014) Web. 23 Apr. 2017.

"Spark SQL, DataFrames **and Datasets Guide."** ***Spark SQL and DataFrames - Spark 2.1.0 Documentation*****, spark.apache.org**/docs/latest/sql-programming-guide.html. Accessed 23 Apr. 2017.

Wang, Rui, Gabriella Harari,†Peilin Hao, Xia Zhou, and Andrew T. Campbell . " SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students." Web. 23 Apr. 2017. < http://studentlife.cs.dartmouth.edu/smartgpa.pdf>

Rogaten, J., Moneta, G., and Spada, M. Academic performance as a function of approaches to studying and affect in studying. Journal of Happiness Studies 14, 6 (2013), 1751–1763.

Kelly, William E., Kathryn E. Kelly, and Robert C. Clanton. "The relationship between sleep length and grade-point average among college students." *College Student Journal* 35.1 (2001): 84-86. Web. 23 Apr. 2017. < http://media.biobiochile.cl/wp-content/uploads/2015/03/215-4929907074434892796-The_Relationship_Between_Sleep_Lenght_and_Grade_Point_Average.pdf>

Parker-Pope, Tara. "Vigorous Exercise Linked With Better Grades." The New York Times. The New York Times, 03 June 2010. Web. 28 Apr. 2017. <https://well.blogs.nytimes.com/2010/06/03/vigorous-exercise-linked-with-better-grades/comment-page-4/?_r=0 >

Ho, Tin Kam (1995). *Random Decision Forests* (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.