Datafest 2024

Daniel Illera, George Daher, Isabella Wang, Sreekar Challa

March 24, 2024

Intro

One of the goals of CourseKata is to “improve important learning outcomes for all students.” Through our analysis of the dataset we sought to evaluate the success of this goal, and determine whether there is a way to predict student academic success based on the data. If CourseKata as a platform would be able to project academic outcomes for students, then they could potentially offer extra resources to students predicted to struggle in order to ensure that all students are able to master the subject being taught.

Finding a Performance Metric

The first step of our analysis was to identify a suitable metric for academic performance. Due to the nature of the data, common academic performance metrics such as grades, GPA or standardized test scores were not available. Within the data, we found two candidates for potential performance metrics: end of chapter (EOC) assessment results and Pulse Check survey answers. Although engagement is important to understand how students learn, we did not consider engagement as a suitable performance metric because lack of engagement could equally mean quick mastery of the material or a lack of care to learn, and more context would be needed. Specifically, we would want to know the importance of the CourseKata books in the way subjects were being taught. As a result, we reserved analysis of Engagement for prediction of performance, rather than using it as a performance metric. Instead we focused on comparing two main performance metrics: EOC assessment results and Pulse Check self-survey answers.

End of Chapter Assessments

The first candidate performance metric we examined was EOC assessment results, due to their assessment structure and performance evaluation per chapter. However, we determined that EOC results were flawed as a performance metric. First, we found that because EOC assessments allow an unlimited number of attempts, the proportion correct could not be considered a true performance metric without some dependency structure including number of attempts. When we used a linear mixed model to account for this dependency, we found that EOC assessment proportion correct and number of attempts were very weakly related. Additionally, using another linear mixed model, we found a negative relationship between EOC assessment percentage correct and chapter number, suggesting that EOC results may be more useful as a measure of difficulty than success. This was supported by the fact that the mean number of attempts per chapter was fairly consistent regardless of chapter number. These findings indicate that the goal of students taking EOC assessments was not necessarily to maximize EOC proportion correct, because despite constant attempts from chapter to chapter, EOC proportion correct decreased. In a better performance metric we would expect to see some demonstrated interest in improving performance. Given the formative nature of these assessments and our analysis of the results, we concluded that EOC results were better interpreted as a metric of difficulty, and we investigated other performance metrics.

Pulse Check Self-Survey

Another potential performance metric we examined was the Pulse Check self-survey responses for each student–chapter. While we were initially skeptical of using a performance metric that was self-reported, multiple published studies support the validity of self-prediction as a predictor of academic success. Several studies have found that students' self-prediction of grades is significantly related to their semester-wide GPA, which is a generally accepted performance metric for academic success. Additionally, one study found that in the context of predicting college students' academic success, self-prediction can be a better predictor of academic achievement than a standardized test or high school GPA. This is an especially important finding because predicting academic success based on available data is generally difficult, and there is a large body of research that seeks to answer this question. Being able to isolate self-prediction as a usable and sound performance metric was therefore a key aspect of our analysis.

Within the Pulse Check self-survey responses at the end of each chapter, students answered four different questions on a scale from 1–6. Of these, Expectancy and Cost both relate to students' self-predictions about their academic performance because they survey self-confidence in the materials and whether they were able to "put time to do well" on the material, respectively. Meanwhile, Intrinsic and Expected Value are still important to consider for understanding dependency structures in the data, but are not themselves measures of performance. We found that the wording of the survey question for Cost was ambiguous. Students rated their agreement with the statement "I was unable to put in the time needed to do well in the previous chapter." Agreeing with this statement could mean that they did not have time, were not able to put enough time into understanding difficult material, or simply did not do well on the chapter. Additionally, due to the survey structure we hypothesized and confirmed that there was substantial multicollinearity between all four "Pulse metrics." As a result, and due to the ambiguity in Cost, we chose to focus solely on Expectancy as a performance metric. Expectancy measures students' self-confidence in their understanding of the material and is itself a form of self-prediction of academic performance. Combining this with existing research on the validity of self-prediction, we chose Expectancy as our main performance metric and focused our analysis on trends, missingness, imputation, and modeling for Expectancy.

Performance Analysis Methods

The most important feature present in the data is its clustering structure, which makes both visualization and modeling more challenging. Since we have limited information about the institutions that were sampled, we assumed that the institutions are fixed, meaning our inference applies only to the 11 institutions in the dataset. Within each institution, classes were sampled, and within each class we observe a cluster of students, so students are nested within classes. The data are collected from the same student at multiple time points, indicating that the student should be treated as a random effect. Students within the same class are likely to be similar in terms of their understanding of the material, major, and other demographics. Therefore, properly accounting for the nesting structure is important in order to draw valid inferences from the data.

With the goal of modeling how well a student thinks they understand the material, we used a Cumulative Logit Mixed Model (CLMM) to account for the multilevel structure in the data. We used the number of minutes students engaged with a chapter and how useful the student found the material

as predictors for the students' self-reported understanding of the material (Expectancy). The model also contains dummy variables to indicate institution, in order to account for variability between institutions, and dummy variables for the book and chapter with which the student engaged.

The Pulse Check response variables contained moderate amounts of missingness. Initially, we hypothesized that the missingness might be Missing Completely At Random (MCAR), meaning that missing values were due to random chance. However, closer inspection revealed that the amount of missingness varied by class, institution, and book, suggesting that the MCAR assumption was not appropriate. To investigate the impact of missingness, we fit our model twice: first using only complete cases, and then using data sets in which missing values were imputed.

Missingness in Understanding Self-Reported Metric

Book	Proportion Missing by Book
College / Advanced Statistics and Data Science (ABCD)	0.5157466
College / Statistics and Data Science (ABC)	0.1522832
High School / Advanced Statistics and Data Science I (ABC)	0.1803279

For imputation we decided to use Multiple Imputation to ensure that our estimates of the standard errors were accurate and accounted for the variability introduced by the imputation method. We used the MICE package in R to impute five complete data sets using Predictive Mean Matching (PMM). This algorithm was chosen because it can impute multinomial data (the self-reported metrics) while preserving the multilevel structure in the data. According to Rubin's rules, our estimate of each slope coefficient is the average of that coefficient across the five complete data sets, and the total variance is a linear combination of the within-imputation variance (the variance of the estimator within each complete data set) and the between-imputation variance (the variance of the five point estimates).

Because the dataset is large—containing over 1,000 students, each with at least 12 observations—we bootstrapped our standard errors using a sample of 100 students over 1,000 bootstrap iterations. At each iteration, we sampled students in a way that preserved the observed proportions of institutions and classes so that each bootstrap sample would resemble the original dataset structure. For example, about 55% of the data come from Institution 2, and approximately one-third of those observations belong to a particular class in that institution, so around 17 students were sampled from that class in each bootstrap replicate. For computational reasons, each of the five imputed data sets was used for 200 bootstrap iterations, for a total of 1,000 iterations. We do not believe that reducing the number of iterations per imputed data set meaningfully affected our conclusions, as this number of iterations should be sufficient to obtain stable estimates.

Results

We found that none of the variables we considered were statistically significant predictors of Expectancy. Even utility value—which we hypothesized would be associated with whether or not the student understood the material better—did not emerge as a significant predictor. The results did not change dramatically between the complete case analysis and the imputed analysis. The main exception was the coefficient for the amount of minutes a student engaged with the chapter: under multiple

imputation, this coefficient moved closer to zero and its confidence interval became considerably tighter.

Imputation Analysis Results

	grandMeans	lowerBound	upperBound
1	0.1280812	-4.679854	4.936016
2	17.1058071	-17.395779	51.607393
3	21.0414215	-15.137164	57.220007
4	23.6266350	-13.528120	60.781390
5	27.3665184	-14.930039	69.663076
6	28.8865353	-23.175564	80.948635

Limitations

There are some time dependencies within the data that, due to the nature of this competition, we did not have time to deal with appropriately. The initial solution we had was to include a random effect for the day relative to the beginning of class, however some institutions functioned on a quarterly system, others included summer sessions as well as the spring semester, and some students were not in the course for its entire duration. Another large issue that we failed to handle was censorship. The students that did not make it to the end of the course should be considered censored data points because we did not get to see how they performed in later chapters of the book.

References

Gadzella, B. M., Cochran, S. W., Parham, L., & Fournert, G. P. (1976). Accuracy and Differences among Students in Their Predictions of Semester Achievement. The Journal of Educational Research, 70(2), 75·81. http://www.jstor.org/stable/27536965

Keefer, K. E. (1969). Self-Prediction of Academic Achievement by College Students. The Journal of Educational Research, 63(2), 53·56. http://www.jstor.org/stable/27535912

Vink, G., Lazendic, G., & van Buuren, S. (n.d.). Partitioned predictive mean matching as a multilevel imputation technique. Stefvanbuuren.Name. Retrieved March 24, 2024, from https://stefvanbuuren.name/publications/2015%20Partitioned%20PMM%20-%20PTAM.pdf