The basis of this challenge was to identify and address ways to improve the student experience of learning statistics and data science with CourseKata. Students often struggle with inefficiency and persistence in an online learning environment, which can be detrimental to learning outcomes. Our analysis focused on finding a solution to these challenges through the use of data provided by CourseKata.

The six data files provided described students' interactions with online interactive textbooks in statistics and data science courses. This study explores the factors that most significantly contribute to the proportion of accuracy regarding end-of-chapter (EOC) questions. In addition, we chose to investigate how different measurements of course engagement contributed to the understanding of chapter content. Various predictive models were built to predict student learning outcomes in a chapter, and model performance was compared using the root mean square error (RMSE).

We started by looking at data from checkpoints_eoc_sample.csv, and we pulled variables from media_views_sample.csv, page_views_sample.csv, and checkpoints_pulse_sample.csv in order to identify any potentially significant relationships. The data were merged using student identification and chapter number as common factors. Missing values were omitted, and some variables were modified to facilitate their analysis. After choosing which specific variables to investigate, we continued our analysis with the comprehensive datasets rather than the samples. Our preprocessing resulted in a primary dataset of 3,915 observations described by 6 predictor variables derived from checkpoints_eoc.csv and page_views.csv.

Preliminary analysis indicated positive correlations between EOC and time engaged, number of page reviews, and page resets. Data visualization displayed a noticeable trend in later chapters exhibiting lower EOC scores and lower engagement. Some variables, such as access_count, had potential significance, but usability was restricted by limited availability (e.g. data for visual media access was only included in chapters 4 and 5). Descriptive statistics revealed the significance at the 0.5 level in every predictor (chapter_number, n_attempt, t_chap_engaged, n_review, n_tried_again_clicks). This result indicated that active course engagement in online learning is important and strongly recommended.

Lastly, fitting a linear regression model to predict EOC revealed heavily skewed residuals. In order to satisfy the model assumptions, a log transformation was used on the response variable to fix the skewness in the model. We then fitted multiple predictive models (lasso, ridge, and random forest) with less restrictive assumptions. The predictive model comparison analysis revealed superior performance in regularized regressions (lasso and ridge regression). We observed that both the lasso and ridge regression models exhibited the lowest RMSE values. These models can be employed to predict EOC scores throughout a student’s interactions with course material, based on measures of engagement. We believe that our model would be helpful in creating a more personalized learning experience, where students are not only encouraged to review their weak points multiple times but also increase their engagement in additional activities and example problems.

In conclusion, course engagement is clearly tied to learning outcome, and this is evidenced by the significant relationships between our predictor variables and EOC response. Our specific suggestions for CourseKata include synchronously predicting EOC scores as students progress through course material and providing additional reinforcement in response to those predictions. We would like to see CourseKata further prioritize and facilitate student engagement with their online learning materials. As supported by our analysis, this would benefit student learning outcomes in statistics and data science.