Collaborators: Peter Kinder and Michael Ghattas
Domain expert name: Kai Larsen
Domain expert affiliation: University of Colorado at Boulder Professor
Domain expert contact info: kai.larsen@colorado.edu
Date of initial meeting: 04/20/23
Date(s) of additional meetings: N/A
Date of report: 05/05/2023

Q1: Qualitative
Q1 is the project's initial Qualitative component, which sets the foundation for the Quantitative component of the project (Q2) and the implementation of the solution (Q3). Specifically, there are seven aspects of Q1 relevant for every project.

A: What is the domain problem? Be sure to write down their overall research, business, or policy goals and their specific scientific questions.

The domain of the problem relates to an instructional software called Simplyx that is used by students to prepare for a certification in another software called Alteryx. The goal is to develop a predictive model for student performance on the certification based on Simplyx log data.

B: Why is this problem important or interesting? This should be answered individually by each collaborator on the project. Why is your domain expert's research, business, or policy question interesting to you? If it's not interesting, make up a plausible reason.

Peter: This problem is interesting because I took the class before something like this existed and created it with Kai to help other students learn data analytics. I have also been a TA during multiple semesters and it is gratifying to see others learn this content and enjoy it as much as I do. So, anything that can help students come to learn and love data analytics is interesting and important to me.

Michael: The project presents an opportunity to explore the development of a Machine Learning (ML) model aimed at identifying students' projected performance to ensure they stay on a good academic track. Additionally, such a model could have the potential to be amended and scaled into a product that can benefit academic institutions and their students.

C: How will the eventual solution be used? How they will use the answers to their research questions (i.e., what is their intended outcome of the research) and how will this help achieve the overall goal of the project?

The solution will be used to further support student learning by allowing the course instructor to quickly assess which students are predicted to perform less than desirable on the certification. The course instructor can use this information to reach out to students and offer support in a variety of ways. Currently, there are some tools that a

professor can use for this, but a single, actionable prediction would be much more effective.

Additionally, it might be set up that the prediction for individual students is available to them through the Simplyx website with the hopes that seeing how they are predicted to do could reinforce good habits or motivate additional effort.

D: What potential data could solve the domain problem? What data, if it were available and accessible, would help answer the underlying research questions or guide the business or policy decisions? This is an important hypothetical exercise.

Some data that could potentially help with the predictive model would be information related to the students history at the university. For example, a student's major, GPA, and past courses could be helpful. Additionally, data related to a students familiarity with computers could be helpful, i.e. if they have any experience with coding and such.

E: The actual data (only if data have already been collected)

E1: What data have been collected?

Log data for each interaction a student has with Simplyx. This data includes a unique id for each student, problem number, belt number, whether the schema and the problem was correct, and a timestamp.

E2: Why were the data collected originally? (For what purpose?)

The data were collected with the hopes of being able to improve the product and provide feedback to instructors. Additionally, the idea of gamification of competition to encourage students is employed using the data. Lastly, the data may be used in the future to identify whether students are submitting their own work.

E3 and E4: When and where were the data collected?

The data were collected over three semesters from Spring of 2022 through Spring of 2023.

E5: Who or what collected the data?

The owners of Simplyx.

E6: How were the data collected? With what instrumentation/methods?

The data were collected automatically through user interaction with Simplyx.

F. What may be the qualitative relationships between variables, for those observed and unobserved?

- A possible relationship between specific questions and better performance
- A possible relationship time-related variables and better performance
- A possible relationship between number of attempts and better performance
- Differences in performance amongst different tiers (Belts)

G: Which types of statistical analyses or techniques would be most useful to the domain expert? Which would not be useful?

We plan to explore a variety of predictive models for this task. To list what we are currently thinking about: Tree based methods, BART, lasso, ridge regression, PLS, and PCR.

Q2: Quantitative
Summarize the statistical collaborators' quantitative contribution or advice. Did the domain expert understand the statistics? This can be whatever (if any) quantitative contribution or advice you provided during the initial meeting or in a subsequent follow-up email. If there has not been any Q2 advice so far, indicate your thoughts of potential directions for Q2.

See Q1 G.

Q3: Qualitative
Did the contribution, advice, or solution answer the researchers' questions? Will it help the domain expert achieve his or her overall research goal(s)? Are there any practical constraints limiting the effectiveness of the proposed Q2 statistical solution? What is the answer to the research question(s)? Note: It is uncommon for an initial meeting and follow-up activities to result in Q3 conclusions or recommendations. If these have already occurred, please detail them here. If they have not occurred, just state that Q3 has not yet occurred.

Yes! The DE was happy with the results and indicated that the model would be deployed for the fall 2023 semester. We also highlighted how more data would be valid for future consideration of future applications.

Shared Understanding Statement:
After completing the previous report sections, you must send the report to the domain expert(s) for any edits or additions and obtain their concurrence on the report. After your team has reached an agreed upon version with the domain expert, you can proceed to submission.

This has been reviewed by the Domain Expert (Yes/**No**)
The Domain Expert made edits or additions (Yes/**No**)
The Domain Expert agrees that shared understanding has been created (**Yes**/No)

Meeting Notes:
Append your initial (or follow-on) meeting notes here or provide a link.

LISA Meeting Notes

Attending: Kai Larsen, Michael Ghattas, Peter Kinder
Date: 04/020/23
Location: Zoom (https://cuboulder.zoom.us/my/kai.larsen)
Video Recording: No (Tech. Issues)
Time: 2:30PM - 2:45PM

Wants/Goals for the Meeting:

1. Review types of questions in the survey. Do the types of questions used elicit useful responses?
    - Gage usefulness of predictive model for instructors and students
2. Discuss best methods for analyzing the results of each type of question (multiple choice, rank order, etc.).
    - Discuss possible features' engineering to be used in predictive modeling

Items Discussed/Decisions Made:

1. Feature engineering is needed to develop components to measure metrics to be used in the analysis which can be used to derive useful insight
2. Possible predictive model to identify how students behavior is related to performance
3. Average time on question / context of question / score on specific questions
4. Share MoM with DE for alignment
5. Make suggestions on possible models/approach that can be used and why
6. Very high passing rate. Not sure if there is a cheating element
7. Predictive model os to be the main deliverable
8. Measure cheating factor
9. Measure completeness of tasks
10. Individual problem completion relationship with certification performance

To do:

1. Data cleaning and scrubbing for consistency and correctness
2. Identify the needed features and metrics
3. Explore possible ML/statistical methods to use for predictive model
4. Features' engineering and build model framework
5. Develop and test predictive model on training set
6. Test predictive model on test set to ensure lowest possible MSE
7. Investigate results and draft analysis report
8. Provide further recommendations and possible limitations to completed analysis
9. Finalize presentation on project

Timeline:

1. *Complete to-do items from one through seven by Friday, Apr 28th*
2. *Complete to-do items from eight through nine by Tuesday, May 2nd*

Next Meeting: TBD.