Final Project Report:
Engagement Time Machine

The Capstone Experience (DS 6011)

Griffin McCauley, Theo Thormann, Eric Tria, Jake Weinberg

I. Background

Since early October, our capstone group has been working in collaboration with our sponsor, Hum, to gain deeper insights into the academic publishing industry and to begin developing models that will enhance the experiences of Hum's clientele. As a customer data platform (CDP), Hum helps organizations and, in particular, academic publishers better understand their audiences and derive valuable insights about their users through the collection of first-party data. By tracking user engagement across a company's relevant platforms and digital interfaces, Hum enables marketing teams to determine patterns in user behavior that drive high value actions and which result in different levels of product loyalty. Through gathering information about individual users' activity such as page views, clicks, reads, citations, likes, and subscribes, Hum hopes to create personalized user profiles that allow for more targeted interventions that will generate greater engagement and a more efficient distribution of knowledge. As part of this process, Hum has also developed a proprietary text analysis model called CueBERT which derives salient topics from articles and assigns tags to them based on their content and subject matter. Thus, by tagging articles as being related to specific fields of research and tracking how users interact with different materials, Hum produces unique and specialized features for each profile which highlight what people's areas of interest are and to what degree they tend to be driven towards papers of that nature. This affords publishers the ability to leverage their corpus of content more effectively and to cater to the needs of their users with greater precision. In this way, Hum provides organizations with the necessary tools to increase user engagement, strengthen customer loyalty, drive up high value actions, and quicken the spread of information.

While one of the primary goals is to uncover meaningful relationships within patterns of user behavior that offer companies the ability to grow their audience and better retain existing users, one of the other major aspirations of Hum is to reduce friction within the academic publishing process and to increase the speed with which new knowledge is created and proliferated throughout the world. Academic publishing is an essential part of many institutions since grants and funding for research generate large revenue streams in the form of overhead costs that go directly to the university, but there are notable inefficiencies in the research and subsequent publication process. Hum hopes to reduce these inefficiencies. The company is beginning to consider how the data it collects to form these personalized user profiles can be harnessed to identify qualified peer reviewers and to recommend these individuals to the editors of major academic journal publications. Currently, the steps required to go from research to published article can be arduous. One of the tightest bottlenecks in this series of events is finding appropriate peer reviewers for new articles. Oftentimes, it can be challenging to locate peer reviewers who both are qualified to assess a paper about a niche subject in a highly technical field and have the bandwidth to take on the additional time commitment of carefully and thoroughly reading and reviewing another article on top of their other obligations. By having access to potential reviewers' engagement histories along with their specific fields of research

and interest, Hum hopes to design and develop a model which will tease out individuals who are likely to be highly qualified reviewers and to accept the offer to take on such a role. In order to achieve this goal, additional external data from clients will need to be procured, and there are currently discussions between the Hum team and several of their academic publishing clients to address this. Hopefully, these deliberations will result in Hum gaining access to the information necessary to turn this idea into a reality.

Through its CDP, Hum aspires to "turn a cacophony of internet noise into one cohesive Hum," and, by achieving the goals of increasing user engagement and streamlining the peer reviewing process, Hum will forward the overarching mission of "making human knowledge more interconnected and accessible."

II. Communications

Since the project began this semester, we have had weekly meetings with our sponsors from Hum. We have been in communication with our main sponsor, Will Fortin, along with Niall Little and Dylan DiGioia. The initial kick off meeting was productive. It gave our group a detailed explanation of what the project's scope and requirements are, which will be discussed in detail in the next section. Our Hum sponsors also gave us an introduction to Hum as a company. It was insightful since our group, for the most part, was not familiar with the academic publishing industry, especially with how data science can be applied. Our Hum sponsors also discussed the work done by the previous capstone group and how our group will branch off of their work. At the end of the meeting, our sponsors also provided source material that our group could read in order to prepare for the following meetings and the project in general.

The course of our weekly meetings with our sponsors has generally followed a regular pattern. It starts with our sponsors explaining technical details on the assigned reading material, which leads to a conceptual group discussion. During this conversation, our group also presents the work we have done in the previous week. After this collaborative briefing, we conclude the meeting by brainstorming next steps and determining what we plan to accomplish prior to the following session. This set up has been very effective so far since the weekly timeframe has proven to be optimal for completing the assigned deliverables in an efficient and productive manner. As we continue this cycle in the spring semester, we expect our group to remain on track for accomplishing the project goals.

In addition to our weekly meetings, we also have a Slack channel with our sponsors. This is helpful since our group can ask questions throughout the week whenever we get stuck. It also allows our group to be comfortable reaching out to and maintaining regular contact with our sponsors. Our main sponsor, Will, has also corresponded with each group member individually to know more about which specific parts of the project most interest them. This allows us to more deeply evaluate and explore our personal data science passions through this project.

Overall, the meetings and communication channels with our sponsors have been fruitful. Our group has learned a lot about the academic publishing industry and how CDPs work with them. We have also gotten an introduction to how the machine learning models we learned in class get applied to professional projects. This includes the entire process of planning out the models, including how to deploy them on various cloud services. Looking forward, since we will be meeting with our academic

advisor, Dr. Judy Fox, more frequently, our group is expecting our brainstorming sessions with our sponsors to be even more effective and productive.

III. Project Scope and Requirements

Our project seeks to use first party data collected by our client to create two models: one that seeks to maximize user retention and another that will suggest reviewers for scientific articles. Our data is hosted on the data warehouse Snowflake. There are about 200 events to search through, but we will only be using about 10 of these events in our two models. A couple of examples of these events are pageview, when a user views a page, and cite, when a user cites a paper. We will have to figure out how to interpret these events and use other tables related to users' profiles to see how events factor into our two models. While the events and user tables will be the focus of our project, Hum's schema is much larger.

In order to accomplish our goals, we are required to meet each week with our sponsor, Hum, and do whatever readings and video viewings are assigned to us. We first had to understand Snowflake which we did by completing Snowflake tutorials and video watching. We then had to get Snowflake integrated to Snowpark which will allow us to use Python to wrangle and model our data. We were also required to complete readings on the publishing industry to get a deep understanding of our project topic and to conduct research about possible data models which could be used to create the deliverables for our project.

IV. General Project Plan

We have already completed several steps in our project plan. Our goal for the fall was to have our kickoff meeting, gain subject matter expertise, and perform exploratory data analysis. Through our readings we have gained a deeper understanding of the complex field of academic literature publishing. We were also able to complete exploratory data analysis by using Snowflake and Snowpark to wrangle and visualize our data.

Our goal over winter is to discuss what types of models we think would best predict user retention and determine future peer reviewers. We are starting to investigate different model types and believe a type of neural net would work well to predict which users will disengage from the publishers' content. We believe that a regression or clustering model might work well to assess which users could be potential reviewers.

During the spring, we plan to refine and test our models as well as have them ready for deployment by our sponsor.. Although no model is going to run perfectly the first time, since we have our data from our capstone sponsor and a plan for our project, we will have plenty of time to create and refine our model. Once our model can make predictions well, it should be ready for deployment. At that point, our mission will have been accomplished and our sponsor will take over to handle the actual implementation on live user data.

V. Defining Success

When it comes to defining success for our project, we have it broken down into two categories: the more functional, client-side component and the more experiential, student-side

component. On the client side, the goals are clearly established; we hope to design and develop two adaptable but undeployed models which will address the areas of increasing user engagement and retention and identifying and recommending qualified peer reviewers. The first of these models will attempt to determine which users are at high risk of dropping out or disengaging from the publisher's content through analysis of their online event history that is captured by Hum's CDP. Additionally, it will seek to offer actionable insights to the organization's marketing team on how best to try and provide new value to said user and hopefully cause them to continue interacting with the publisher's platforms and digital media. At this point, we have begun conceptualization of how this model might be implemented, and, based on the form of the data we have access to for each user being primarily categorical and time-serialized in nature, we believe a recurrent neural network (RNN) model will be a good starting point at which to begin our model construction endeavors. So far, we have initiated our research into how these RNN frameworks are built using libraries such as TensorFlow and Keras in Python, and we hope to produce some preliminary code in the coming weeks in order to put us in the best position for success come spring. If done correctly, we are confident that this RNN model will be able to predict the future behaviors and actions of users based on their history, thus enabling us to recognize the signs of disengagement or unsubscription, and, hopefully, by building on top of this prediction model foundation, we will be able to categorize the symptoms related to different types of content apathy and recommend specific intervention approaches to the marketing teams that are tailored to the user's unique style of reduced interaction.

The second model, which strives to evaluate the qualifications of potential peer reviewers and to recommend the top, most relevant candidates to editors at academic journal publications, is a bit less well defined at this point due to the fact that the data currently at our disposal does not contain all of the salient information required to judge whether a specific individual has experience with peer reviewing and adequate expertise in a given field. The team at Hum believes this data does exist, however, and is in the process of attempting to acquire access to it, so, despite the potential hindrance of not having it available at present, we are optimistic that the information will be usable in the not so distant future and have decided to begin using clustering and regression techniques to get a sense of how we may structure a model once all of the pertinent data is consolidated in one location.

By creating these two models, we will hopefully be able to contribute meaningfully to Hum's overarching mission to "make human knowledge more interconnected and accessible."

On the student side, our main goals are to apply concepts from our coursework to a real-world problem domain, to use teamwork skills to design and manage a project from start to finish, and to learn about a unique industry. Since the start of the summer, we have been gaining extremely valuable and robust skillsets, and, while we have had numerous opportunities to implement our newly acquired knowledge through small projects and on toy datasets, this is our first chance to gain experience working with organic and authentic industry data, and we are all excited to see how well the theory translates into practice. We also come from diverse and unique backgrounds and experiences, and the capstone project provides us with a tremendous medium through which we can combine these varied perspectives and bodies of knowledge in novel ways, and, through the use of open dialogue, good communication practices, and strong teamwork and collaboration, we intend to drive this project to success. Finally, this project involves a topic domain in which none of us have direct, previous experience, so, through our engagement with this project and the members of the Hum team, we hope to gain a richer understanding of the academic publishing industry and to better

appreciate all that goes into generating transformative knowledge which is carried forward into future generations. If we are able to achieve the sets of goals we have laid out on both the client and the student sides, we will gain a wealth of hands-on data science experience while simultaneously advancing the aspirations and mission of an academically inspired and morally grounded start-up company.

VI. Data and Methods

As discussed in the previous sections, our group will make use of customer data platform (CDP) events such as page views, citations, and scrolling behavior. The data we will be using for our project is from Rockefeller University Press. Since these events are at a very detailed level, our group will have access to millions of rows of data. It is part of our task to make sense of which specific events will be useful in building the two models: user retention and reviewer recommendation. In addition to event data, our group will also make use of user profile information such as when users started, how long they have been active, and others. This is especially interesting for the reviewer recommendation model since there are different criteria for being asked to review a paper.

Since the dataset we will be using is massive, the data is being stored in Snowflake, a cloud-based data warehouse. The Hum team set up this warehouse, and they provided us access to it so that we can do quick queries on the data using the intuitive Snowflake console. (Given that the Hum team provided this access, our group did not have to make a budget request for this purpose. Snowflake pricing works on a usage basis where one only gets billed based on how much computing resources are used. This is still dependent on which account tier Hum is part of, as can be seen on Snowflake's pricing information.) In addition to accessing the data on the Snowflake console, our group can also connect to the warehouse using the Snowpark API and our Python scripts. This makes it convenient for us to easily analyze the data. Our sponsors have also discussed using Amazon SageMaker, a cloud-based machine learning platform, for this project, to which they will be providing us access as well. We have not used this platform in the fall, but we are looking to explore it in the spring as we prepare our models for testing and deployment. (SageMaker cost is also usage-based as seen on the pricing calculator, but it again depends on Hum's account tier.)

As for the data science methods, our group is currently exploring various machine learning models that may be useful for our goals. The focus now is on learning about recurrent neural networks and how they will perform for our first model: user retention. Since the reviewer recommendation model is more complicated, our group is currently focusing on starting with performing regression and clustering analysis to see how we can proceed with that model. We are looking forward to building two effective models in the spring with the guidance of our sponsors and academic advisor.

VII. Risks and Mitigation Plans

Potential pitfalls in our project could emerge from three general areas – the client data, our model building, and the technical resources we will be using. While we are set up well with a communicative client that has functional experience, we want to ensure that we have mitigation strategies in place for any disruptions that may arise out of these categories.

Hum is providing us with our data after it has already been acquired and processed. It is an advantage for our group that the data cleaning is done for us, but this does prevent us from having much influence on its wrangling. The data that we will receive for our RNN, as of right now, will only have between five and ten features. To build a successful model that can predict high-value events that keep subscribers engaged, we will need to hope that this data set will have enough information for us to extract helpful insights.

A possibly greater data risk surrounds the reviewer model. The only data that we have at present is about user events, which is only a sliver of the information that we would need to build an effective regression model. Data describing who has recently published and reviewed papers would be essential to have, as well as user profile data on candidate reviewers. The best we could do for now is to look for clusters that could suggest likely and qualified reviewers based on users' behavior. A clustering analysis such as this would help us to craft the next steps of our project, but it would not be a marketable asset to clients. To mitigate these potential issues, we will leverage the open communication channels we have with Hum. If we explain the data concerns that we have, Hum will do everything they can to provide us with what we need. At the very least, we will be clear on our expectations for the final products that we can deliver given the data that is available.

Additionally, a potential area for slippage is the model building process. Most of our group does not have any experience with building an RNN, and the current state of the reviewer model means that we are lacking a clear path forward. Both models will have a bit of a laborious beginning that we will need to clear. If we are not cognizant of our timelines, we could be in danger of our delivery date slipping. Because of the hard stop in the spring, it is essential that we avoid that. Even though to date our client has given us week-to-week guidance on what needs to be done, we should not be relying on them to manage our project timelines. We need to be proactive about achieving success. To ensure that we do this, we have used our ramp up period to read as much as we can on RNN techniques and to thoroughly explore the data we have. We will continue to build on this over the winter to achieve a high level of conceptual understanding that will allow us to expedite the model building process as much as possible once we enter the heart of the project.

Lastly, we will need to be prepared for any technical issues that will inevitably emerge throughout the project. Our clients have been very helpful in providing us with instruction on Snowflake and Snowpark, as well as any other tools that we are likely to need. We have already experienced and overcome a few bumps in the road, but we have yet to reach any true heavy lifting. Part of our mitigation tactics were to tackle bugs early, so, while there will be new technologies that emerge throughout the project which might cause issues, we hope that many of the problems that arise from new applications will be benefited by what we have set up already. It is still possible that we will lose time to technical complications, however, so we will account for this in our project by building a buffer around our deadlines. We will force ourselves to stay ahead of our work and to resolve complications as quickly as possible. When obstacles do present themselves, we will be sure to leverage each other and our client to get back on track as quickly as possible.

Hum has provided us with a unique opportunity to learn about a new industry and build complex, interesting models that will directly contribute to their business. We are excited to dive into the analytical process this winter. Our team is looking forward to the value that we can deliver to our client and to the learning opportunities that will provide a successful capstone experience to our group.