Hum Engagement Time Machine

- Griffin McCauley (kzj5qw)
- Theo Thormann (nbx5kp)
- Eric Tria (emt4wf)
- Jake Weinberg (jaw7cd)

Capstone Work II (DS 6013)

Dr. Judy Fox

25 February 2023

Progress Report 1

Introduction

Since October, our team has been working closely with our sponsor, Hum, to develop a model which will aid academic publishers identify users that are likely to maintain high levels of engagement with their platform. As a stretch goal, our team is also trying to identify which users may be high-quality candidates for peer reviewers in the future. Hum is a customer data platform (CDP). This means that it collects first-party data from clients' online interfaces and then uses this information to help the clients glean valuable insights into how users are engaging with their virtual content. This insight provides marketing teams with actionable information on how to better serve their users. The data being collected comes in the form of "events". An event might be a "pageview," "post-read-(start/mid/end)," "cite," or "pdf-click." These events also contain other salient features such as what time they were performed, an ID of the visitor who performed them, and what content the action was performed on. Taken together, this data offers a stream of activity which has occurred on the publisher's platform, and, if tailored correctly, should be able to form the input to a powerful, predictive, deep learning model.

Related Work

Although there are numerous platforms that evaluate their users' interaction patterns to enhance retention and promote higher levels of loyalty, many of these techniques are proprietary, and the data they have access to is formatted differently than ours. Due to these limitations, we can loosely learn from previously implemented methodologies, but most of our work has been research-driven and novel in nature. Given that most of our data is both categorical and composed of sequential events, our initial instinct was to begin exploring various recurrent neural network (RNN) architectures, such as the gated recurrent unit (GRU) and long short-term memory (LSTM). We hoped to be able to provide a sequence of

events and their corresponding timestamps as inputs and to receive a set of meaningful predicted events or a sequence classification as the output by using one of the previously mentioned model types.

Initial Methodology

Based on our understanding of the data at the time, we proposed the following data analysis and modeling framework at the start of this semester: since every user has their own sequence of events and associated timestamps, we wanted to break these long sequences into shorter subsequences delimited by "idle" periods, which we empirically determined should be approximately three days in duration. This would make it so that we could then label subsequences that were followed by another subsequence as resulting in re-engagement, while subsequences which were followed by a gap of more than two weeks and no other subsequence as resulting in dropout or churn. Thus, by training a sequence classification LSTM on these subsequences, the model would be able to learn the patterns in user behavior which would likely result in re-engagement versus dropout or churn, and, by applying this model to a users' most recent activity cycle, we would be able to predict whether they were likely to re-engage on their own or whether the publisher should try to intervene in some way if they hoped to retain that user. We proceeded to successfully process the data in this format and were able to train a baseline model on just the time gap subsequences by the end of January. However, we discovered that the data was not rich or expressive enough to allow our model to generate reliable predictions with meaningful accuracy. It was able to produce an accuracy of roughly 67% on the testing set, but this was only about 10% higher than the score which could have been achieved with a naive, majority classifier.

A Process of Iteration

More concerning than the sub-optimal accuracy, however, was our discovery that, due to the format of the online platform that generated our data and the way in which events are registered to Hum's database, every time a user visited an article, the events of "pageview" and "post-read-(start/mid/end)" were all triggered at once, essentially rendering the highly granular timestamps and sequential patterns in the data meaningless. This realization made us take a step back and re-envision the scope and methodology of our project, since the properties of the data we relied heavily on were now jeopardized. After taking the time to consider other options, we came up with the following approach: instead of conducting our analysis using the sequence of events a user has performed, we would derive a set of attributes for each user based on their events sequence.

More explicitly, by consolidating and linking the germane information from the Content, Event, and Profile tables in our database, we would be able to collect not only the events and their timestamps but also robust contextual information. This included data on the content that the event pertained to (i.e.

using identifying information such as the URL), what type of content it was (e.g. whether it was an abstract, an article, a figure, etc.), where the content was reached from (e.g. Google, PubMed, RUPress, etc.), and how popular the content was based on its volume of engagement across all users. Through feature engineering, we would be able to determine the number of unique articles a user has looked at, the percentage of content that was reached through Google as opposed to a more scholarly source, the percentage of content which was actually an article, the number of engagement cycles a user has had, the time it took them to complete a certain number of events, and the average popularity score of the content they engage with.

These six new features could now serve as inputs to what will be a multilayer perceptron (MLP) model designed to answer the question, "If a user has performed 40 events, will they perform another 40?" This is the question we have chosen to focus on since, in our minds, the notion of being a "good", engaged user mandates that a user continues to perform a decent volume of activities for an extended period of time. The other criteria we considered using to classify someone as maintaining high engagement was how long they continued to perform actions after they reached 40 events and whether they voluntarily supplied additional data to the publisher such as their email address or other identifying information. We also decided to apply a filter on our dataset to include only users who have performed 40 or more actions on the platform to focus on the most engaged, and thus important, subsegment of profiles. Since our data has only been captured since March of 2022, we decided to avoid relying too heavily on time dependent criteria since not enough time has passed to observe long-term trends. At this point, we have received confirmation that this new model seems feasible and would provide value to our client. Our sponsor has expressed excitement over the fact that an MLP framework will potentially be more intuitive and easier to modify and extend to other use-cases; this additional flexibility will be highly beneficial going forward since this model will hopefully serve as a framework upon which future models can be constructed.

Overcoming Challenges & Client Satisfaction

Although there has been reconceptualization during the course of this semester, this is all a part of the iterative process of data science. Since our sponsor is a small, start-up enterprise, we are subject to constantly shifting business priorities and data collection methods, so it is not unexpected for there to be a need for dynamism and adaptability during this practicum. While it would have been nice for our initial model to have performed better, our newly determined framework should adhere to the needs of our sponsor and their clients more closely and offer a better solution to the problem of modeling user engagement. Since we have weekly meetings with our Hum liaison, Dr. Will Fortin, we have been able to maintain a high level of transparency and an open and constant dialogue throughout this experience. Dr.

Fortin has indicated that he has been satisfied with the progress we have made on the project and is excited to see what we are able to produce. This week we confirmed with him the ambitions and scope of this project. He clearly laid out that as long as the data pipeline from their database into our model is sound and that the MLP produces reasonable results, this will be a valuable contribution to their code base and that it will be able to act as the foundation for other projects he is hoping to expand to in the near future.

Next Steps & Project Timeline

As ideal as it would be for our progress to have been linear, this is not the nature of data science. Between identifying the flaws in the data collection methods used on the customer platform and grappling with the sparsity of some fields in the database, this semester has posed a number of challenges we have had to overcome. By remaining nimble, open-minded, and tenacious, we have managed to deftly maneuver the ever shifting landscape of our project and come up with a proposed path forward that should satisfy the needs of our sponsor and client, the needs of the class and teaching staff, and the needs of our own education and academic experience. By the end of this month, we hope to have the final data organization process and preliminary MLP model build complete. We look forward to March where we can fine-tune the model parameters and modify our model to better suit the needs of Hum's clients. This should afford us a decent period of time in April to ensure that all of the necessary documentation is in place for both our final reports and presentations and also to curate the Github and all files in a manner that makes them highly understandable, usable, and reproducible for others who need to access our work in the future. Hum's virtuous goal has always been to increase the interconnectivity and accessibility of human knowledge, and, hopefully, through the course of this project, we will succeed in forwarding this goal and expanding the frontier of academic publishers' user engagement.