Engagement Time Machine

McCauley, G., Thormann, T., Tria, E. , Weinberg, J.
University of Virginia, School of Data Science, Charlottesville, VA

***Abstract*—In this paper, we will discuss the processes of using machine learning to engineer a novel set of user-level features and construct a model to accurately recognize high-quality, valuable users early on in their life cycles. Emerging technologies can be applied to the academic publishing space in many exciting ways: it can transform the selection process for peer review, use targeted pokes to keep subscribers engaged, and tailor advertising to be highly relevant for each user. Our project builds the framework for all of these applications. We use a K-means clustering model to determine what constitutes a highly-engaged user and a multilayer perceptron (MLP) model to predict whether new users who have performed a threshold number of events are likely to remain highly engaged. Using first-party data collected from the customer data platform (CDP) Hum, we developed a model that can accurately classify the online readers of an academic publisher as being high- or low-quality based on their early-stage engagement profiles. Using Hum's relational database containing over 100 features, we engineered four new variables to serve as the basis of our analysis that illuminate differences between high-value and low-value user behavior. Through a combination of K-means clustering for determining training labels and an MLP for predicting which of our client's users belong to each cluster, we were able to identify what characteristics are indicative of high- versus low-quality engagement. Based on our derived features and labels determined by K-means clustering, our MLP model is able to predict whether a user is high- or low-quality with 95% accuracy. These engineered features and this model framework will now be able to serve as foundational components in the burgeoning field of digital academic publisher engagement.**

I. Introduction

The academic publishing industry could benefit from the use of machine learning in recognizing high-quality users early on in their life cycles to determine which readers to target with tailored interactions. Although there are numerous platforms that evaluate their users' interaction patterns to classify high-quality users, 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.

Our K-means clustering and multilayer perceptron (MLP)-prediction structure will assist academic publishers both to determine high-quality users and to recruit new reviewers, benefiting the industry as a whole. Working with data provided to us by Hum, a first party customer data platform (CDP), our team has developed a model which will aid academic publishers in identifying users that are likely to maintain high levels of engagement with their platform. A first party CDP collects first-party data from clients' online interfaces and then uses this information to help their 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)," "citation," 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 summary of activity which has occurred on the publisher's platform, and, when tailored correctly, can form the input to a powerful, predictive, deep learning model.

Establishing what makes a "good" user is the most subjective part of our project as well as most the novel. To our knowledge, there is no universally agreed upon metric for determining user quality in the academic publishing industry. Through in-depth discussions with our sponsor, extensive exploration of the features available in the database, and several tests with different feature combinations, we derived four features that, together, are highly indicative of high-value user behavior.

II. Related Work

Many industries such as telecommunications[2], banking[3], and e-commerce[4] are implementing machine learning to predict customer churn and increase user retention. Publishers could use the emergence of new and more accessible technologies to aid marketing teams which seek to limit subscriber churn. Similar to the marketing space[5], academic publishers could use information about subscribers to "poke" them with desirable content and keep them engaged and subscribed to their platform.

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 architectures, such as the gated recurrent unit and long short-term memory (LSTM). One similar method [1] uses a recursive LSTM to learn embedding of users to predict their interactions. 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.

As our group transitioned into using clustering methods to segment the users, we decided to use K-means clustering as it is a very common method for this use-case. Pradana et al. presents a way of using K-means clustering to segment mall customers in order to aid in marketing strategy[6]. Kansal et al. also discuss how K-means clustering can be used to segment users into groups based on good or bad customer behavior[7]. These methods showed to be effective in our project.

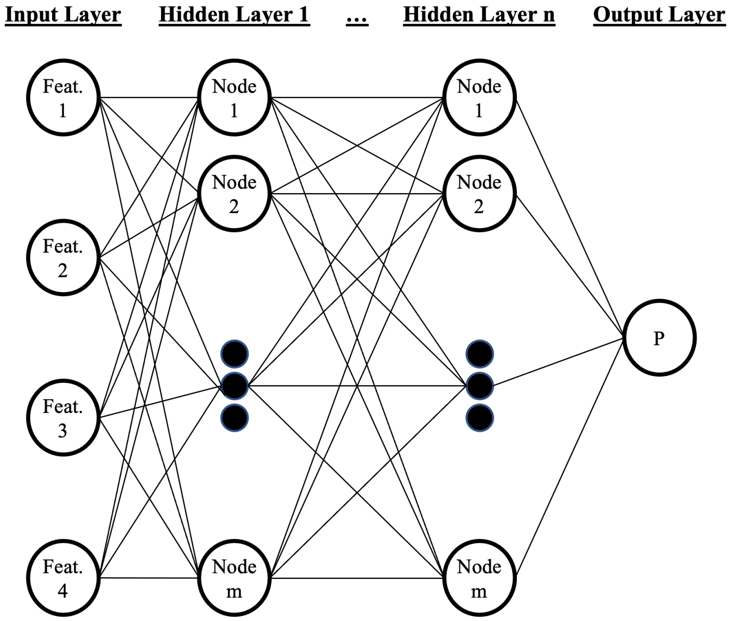
III. Data Description

Our model is built off of first-party data in a database maintained by our sponsor, Hum, that is sourced from a mid-size educational publisher with four journals. The data tracked user activity on the platforms for this publisher over approximately one year. While we received access to a very large database, we narrowed the salient tables down to three: Event, describing details on the activity that a user performed on the publisher's platform; Content, providing information on the category and quality of the content a reader engaged with, along with the type of event a user performed with the content; and, Profile, identifying characteristics of each specific user of the platform. Full schemas of these tables, along with descriptions of each variable, can be found in Fig. 8 & Fig. 9 & Fig. 10 (Appendix).

The earliest data tracked for this publisher was during the end of March 2022. For the year 2022, there were around 1.8 million users interacting with the platform and roughly 10.8 million total user events. For data until March 2023, roughly one year of data, the totals go up to 2.2 million users and 13.4 million user events. This shows the total possible size of data available to us for this project. As part of model building, filters will be applied to this total dataset in order to obtain a good training set for our model.

IV. Methodology

To accomplish our goal of predicting user-level future engagement, we first had to figure out what constitutes the behavior of a high-quality user. By consolidating and linking the germane information from the Content, Event, and Profile tables in our database, we were 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.), and where the content was reached from (e.g. Google, direct link, same journal navigation, external journal link, etc.). Through feature engineering, we were able to derive four primary features which we believed would be highly valuable metrics with regards to online publication engagement; these four features were the number of articles read per event, the percentage of articles reached through Google (as opposed to a more scholarly source), the percentage of content the user read that is an article, and the number of events a user performed per day on the platform. Although there have not been previously documented accounts of the usage of these specific measurements, through discussions with our sponsor, we were able to conclude that high- versus low-quality users should display starkly different distributions across this feature space. For instance, high-quality users would interact with fewer articles per event and have a lower percentage of their content being articles since these behaviors would suggest more thorough involvement with supplemental material such as abstracts and figures. These high-quality users would also be less likely to come from Google if they are more invested in seeking content from scholarly sources, and they would perform more events per day since this is clearly a sign of more active and regular engagement. Therefore, these four new features were chosen to comprise the input layer of our MLP. Fig. 1 describes the model structure in detail.



Feature 1: Number of unique articles read by the user per day
Feature 2: Percentage of articles reached through Google
Feature 3: Percentage of content read by the user that was an article
Feature 4: Number of events per day performed by a user

P: Probability that a user belongs to Cluster A (high-engagement user)

Fig. 1: Visualization of MLP model

To achieve these high- versus low-quality behavioral groupings, we performed K-means clustering on features using users' full event sequences. This would ensure that a user's entire engagement history and their holistic behavior is taken into consideration when forming the clusters. (Note that we also decided to apply a filter on our dataset to include only users who have performed at most 100 actions on the platform in order to remove outliers from the analysis and to avoid skewing the data user for clustering in a significant way. 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.)

After running K-means clustering to partition our users into two groups and deliberating with our sponsor, we all agreed that the notion of being a high-quality and engaged user was being represented in the characteristics of the resulting behavioral clusters. This conclusion was primarily reached because the expected and hypothesized distributional patterns were clearly present in the algorithmically determined groups. We then constructed an MLP model designed to accomplish the goal of predicting whether users who have already performed at least 16 events belong to certain behavioral clusters that are indicative of their user quality. Since our labels were determined by a spatial, centroid-based clustering method, training our model will simply cause it to learn the hyperplane

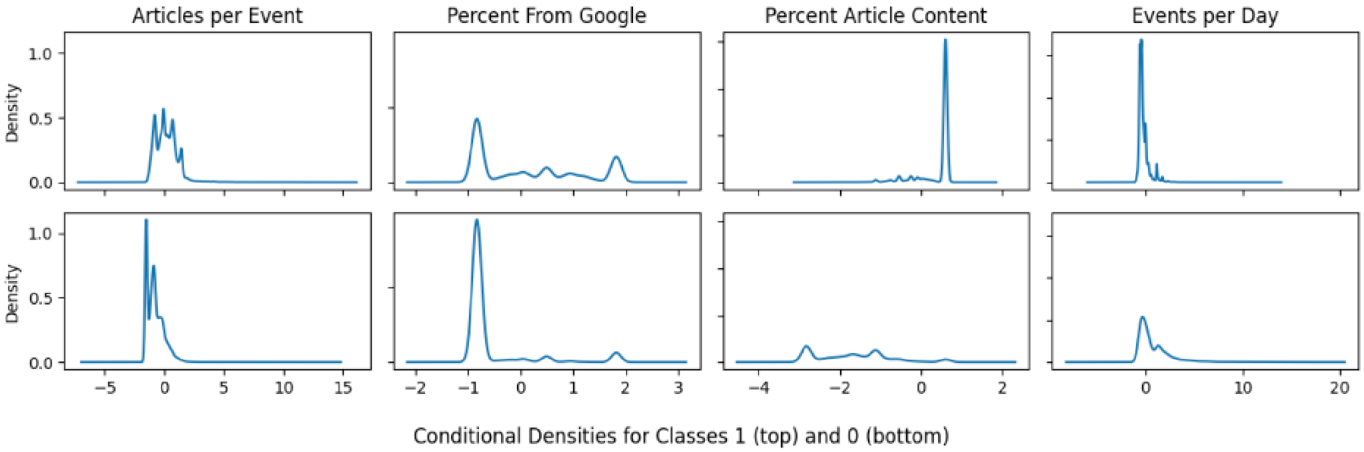


Fig. 2: Class conditional densities for four features used in K-means clustering model applied to all user events

decision boundary produced by the K-means clustering and will essentially be telling us whether a user's initial behavior on the platform is representative of how they will engage over their entire lifecycle. Although other statistical models such as linear discriminant analysis, support vector machines, random forests, and others may also be effective at learning the linear decision surface, our sponsor has expressed excitement over the fact that an MLP framework will potentially be more generalizable and easier to extend to other use-cases. This additional flexibility will be highly beneficial going forward since our sponsor plans to use this model to serve as a framework upon which future models can be constructed for their other clients. For these reasons, an MLP framework was established as the optimal model for the purposes of this project.

From the K-means clustering and MLP model, we were able to generalize behaviors of high-engagement users. As predicted, these users have a low number of articles read per event, which indicates them diving deeper into the article they are reading, and a lower percentage of articles in the content they consume, meaning they interact with other supplemental content such as abstracts and figures. They also have a low percentage of articles reached through Google, which means they are reaching the journals through more scholarly sites instead of casual searches. And, lastly, these users have a high number of events performed per day, which clearly demonstrates a greater level of activity and engagement.

V. Results

A. Clustering

Using K-means clustering with K set to 2 on the dataset composed of features derived from users' entire engagement cycles, we attempted to assess whether there was a spatial division within our data which might represent the chasm between high- versus low-quality users. To evaluate the behavioral characteristics of the two generated clusters, we plotted and examined the class conditional densities for each of the four features. These distributions revealed a contrast in the engagement patterns between the users assigned to each cluster and supported the notion that each cluster assignment could be used as a proxy for being a high- versus low-quality user (Fig. 2). Our clustering determined that high-quality users would be those who interact with fewer articles per event, have a lower proportion of references from Google, a lower proportion of content that are articles, and more events per day; these would be the users belonging to Cluster 0. This cluster consists of 19.4% of users in our dataset and is consistent with heuristic-based assumptions regarding the composition of online users. From these observations, we concluded that the clusters determined via K-means were sound approximations for the ground truth user quality labels.

In addition to the class conditional distributions, we also visualized the results of our clustering via principal component analysis (PCA). By applying PCA to our dataset, we were able to then plot the lower dimensional projection of our data along its highest variance axes and color the points by their assigned labels to visually gauge the spatial distinction between the clusters (Fig. 3).

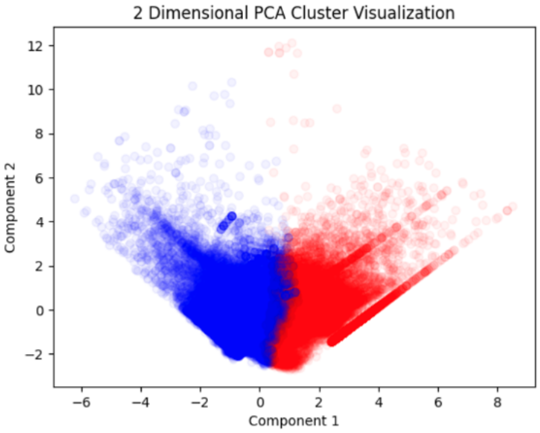


Fig. 3: PCA visualization between K-means clusters

While a region of overlap was present towards the center of the projected data cloud, there was a well defined cusp delimiting the boundary between the two groups. Since this

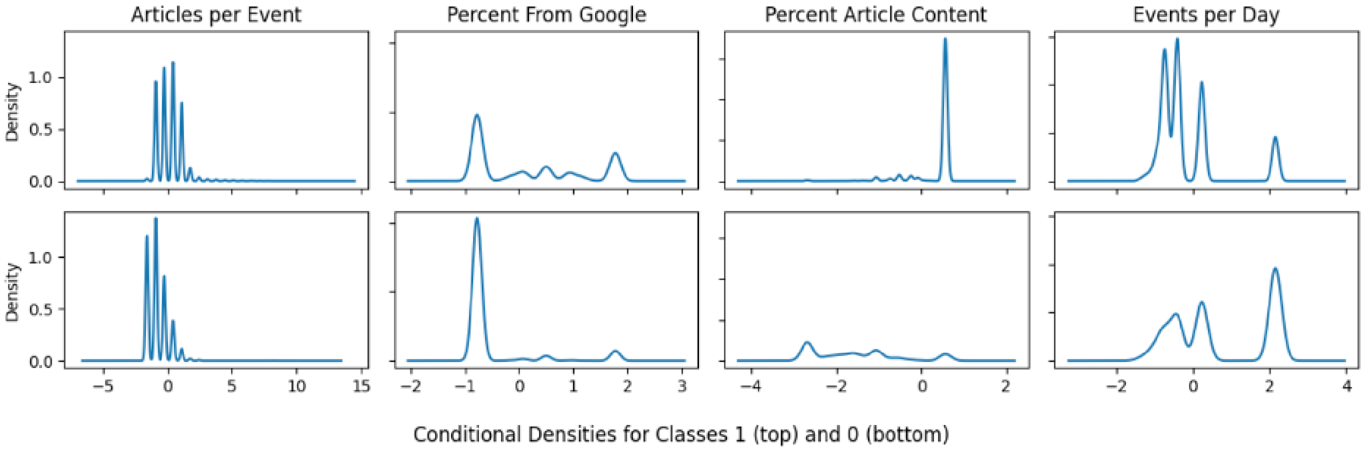


Fig. 4: Class conditional densities for four features determined by K-means clustering model when applied to users first 16 events

geometric feature is created through the realignment of the axes induced by PCA, it is not feasible to directly interpret the meaning of this cusp, but it does visually capture the criteria upon which these clusters seem to be formed. In an attempt to investigate whether there was any deeper meaning that could be extracted from the projected components, we looked at the eigenvectors and their corresponding eigenvalues (table I).

	Articles per Event	Percent from Google	Percent Article Content	Events Per Day
1	-0.510444	-0.429043	-0.611513	0.425935
2	0.629167	-0.615004	0.205002	0.42882
3	-0.017198	0.587668	0.155031	0.793924
4	0.585918	0.303872	-0.748325	-0.066110

TABLE I: Eigenvectors from PCA Analysis

These seemed to express that the number of events per day was negatively correlated and inversely related to the other three features. This further verified what we had previously observed in the class conditional feature distributions. We also found that the first three eigenvectors captured the vast majority of the variance explained through this decomposition (Table II & Table III).

Component 1	Component 2	Component 3	Component 4
1.618	0.942	0.842	0.599

TABLE II: Eigenvalues from PCA Analysis

Component 1	Component 2	Component 3	Component 4
0.404	0.235	0.210	0.150

TABLE III: Proportion of Explained Variance from PCA Analysis

B. Classification

The inputs to our MLP were the same features as the inputs to our clustering model; however, the MLP features included data from only the first 16 events of a user, while the clustering model derived training labels using the full event activity for each user. The MLP model architecture consisted of a four-dimensional input layer, two ten-dimensional hidden layers, and a one-dimensional output layer for binary classification (Fig. 1). Although this neural network is not as deep nor complex as many other neural networks, it was sufficient to use in our analysis because the original decision boundary learned through K-means clustering was linear in nature. Even though the new boundary is based on the features derived from just the first 16 events of each user, we were able to confirm that the general behavioral patterns should be similar by again reviewing the class conditional distributions for these features (Fig. 4). The training quickly converged to a stable level that produced comparable accuracies of roughly 94% for both the training and validation sets (Fig. 5), so the last facet of the analysis was to analyze the model's predictions.

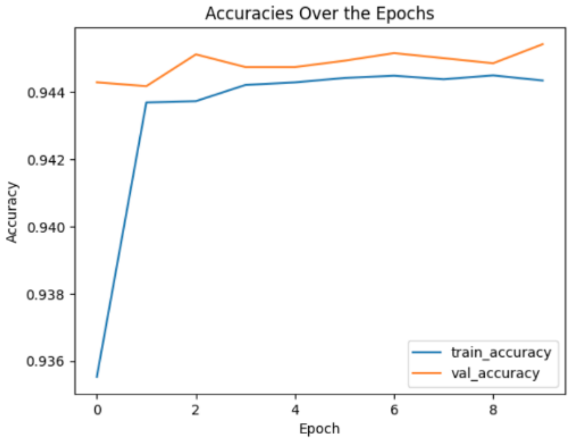


Fig. 5: Plot of training and test accuracy for MLP model

The predictions generated by our model for the test set produced a concave ROC with an AUC of 0.9623 (Fig. 6). This score is indicative of an accurate model that is able to simultaneously achieve a high true positive rate and low false positive rate. While different criteria can be used to determine the optimal classification threshold for assigning our outputted prediction scores, one standard approach is to try and jointly maximize true positive and true negative rates such that they intersect. This approach yielded an optimal threshold of 0.8838 for our data, and this resulted in the confusion matrix seen in Fig. 7. The associated true positive and true negative rates were 91.0% and 90.3%, respectively. All of these performance metrics demonstrate the ability of our model to correctly identify which users will have high-quality engagement behaviors over their entire lifecycle based on only their first 16 events.

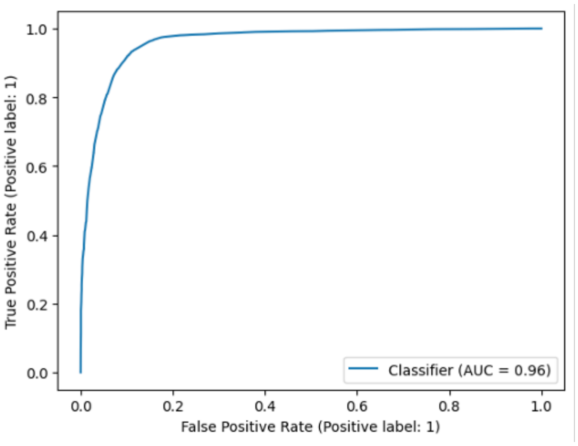


Fig. 6: ROC curve with an AUC of 0.96 for MLP model

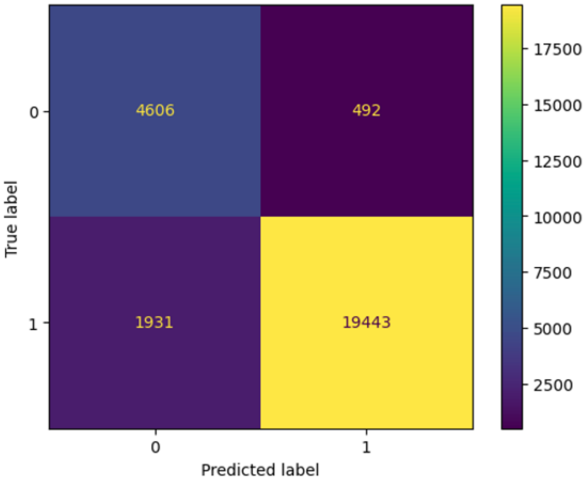


Fig. 7: Confusion matrix for MLP model results

VI. Discussion

Throughout the course of this project, we focused on two central tasks: deriving a set of novel features which could serve as valuable metrics for evaluating and predicting the quality of users' online engagement with academic publishers and designing and constructing a simple model framework which would validate the predictive power of our newly engineered features. By methodically considering each attribute available in Hum's extensive database, we were able to arrive at a final collection of four calculated features: the number of articles read per event, the percentage of content reached through Google, the percentage of content that was an article, and the number of events performed per day on the platform. Although our sponsor's knowledge of the academic publishing industry suggested that these novel features should be indicative of the quality of a user's engagement behaviors, it was not until performing K-means clustering and empirically confirming our hypotheses that we could feel confident in the discriminative power of these features.

The most innovative aspect of our research into subscriber behavioral analysis was the establishment of the four engineered and derived features upon which the rest of the analysis and modeling was constructed. Without the development of a set of novel features which could differentiate between high-versus low-quality users, the rest of the results would not have been possible. Additionally, these features can be taken at face value to build a profile for each user that summarizes their engagement quality, which will bolster the understanding that academic publishers have of their client base. The features are intuitive, explainable, and easily measurable, meaning that they can become key performance indicators in the publishing industry going forward.

While our prediction process using an MLP model may be simplistic, it aligned with Hum's goals of creating a skeleton model that can be easily altered to accommodate the needs of their other academic publishing clients. Not only does our project establish a novel criteria for what features determine an engaged vs. a disengaged user, it also creates a comprehensive pipeline to run predictions on new users as they join and subscribe to academic publishers. We were able to host this pipeline entirely on the cloud. This begins with querying the data from Hum's database hosted on Snowflake, a cloud-based data warehouse. The feature engineering is already executed in the SQL query to save on time since the full query only takes about less than a minute to execute. This query is run on Amazon Web Services (AWS) using a SageMaker notebook and the data is stored in a designated S3 bucket. The next step of the pipeline is loading the data file from S3 and running the K-means clustering and MLP models in order to generate the final output. We have also kept an extensive archive on GitHub along with a comprehensive README to make interpreting and altering our model as simple as possible for our client.

VII. Conclusion and Future Work

After extensive discussions with Hum, we have both determined our project to be a resounding success. We produced a solution that provided key behavioral insights on the users of one of our sponsor's clients, and we provided our sponsor with a skeleton model that can be easily augmented to fit the data of each of each publisher. According to Hum, the industry is in the midst of a changing attitude toward machine learning-powered solutions, so our analysis is both highly valuable and timely.

For future work, our model can be reframed to improve the process of peer reviewer recruitment, which is a crucial aspect of academic publishing. Due to limitations with the data we had accessible, we could not implement a peer review component of our model, but this is highly feasible going forward. Users who qualify as highly-engaged and valuable represent high-quality targets for peer review. When eventually coupled with additional user profile information, our model should springboard technology that can essentially rank candidates for peer review of a paper on a given topic. Our sponsor is keen on the importance of machine learning in reviewer recruitment[8], and they are excited for this development. The refinement and implementation of these techniques into the publishing space only stands to benefit the overall peer review process. We believe that our model provided Hum a good baseline through which they can alter the code we provided them to create a robust reviewer recruitment model.

All code for this project can be found on our GitHub at the following address: https://github.com/Data-ScienceHub/ETM

Acknowledgment

We would like to acknowledge the contributions of the Hum staff, specifically Dr. Will Fortin, Niall Little, and Dylan DiGioia, to this project. We would also like to thank our capstone advisor, Dr. Judy Fox, for her assistance with this project.

References

[1] Yin C., Wang S., Miao H. 2020. "Recursive LSTM with Shift Embedding for Online User-Item Interaction Prediction," 2020 IEEE 13th International Conference on Cloud Computing (CLOUD), Beijing, China, 2020, pp. 10-12, doi: 10.1109/CLOUD49709.2020.00010. Retrieved April 1, 2023 fromhttps://ieeexplore.ieee.org/document/9284319.

[2] P. Tang, "Telecom Customer Churn Prediction Model Combining K-means and XGBoost Algorithm," 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 2020, pp. 1128-1131, doi: 10.1109/ICMCCE51767.2020.00248. Retrieved April 8, 2023 fromhttps://ieeexplore.ieee.org/document/9421755.

[3] M. Alizadeh, D.S. Zadeh, B. Moshiri, and A. Montazeri, "Development of a Customer Churn Model for Banking Industry Based on Hard and Soft Data Fusion," IEEE Access, vol. 11, pp. 29759-29768, 2023. Retrieved April 9, 2023 from https://ieeexplore.ieee.org/document/10070749.

[4] X. Xiahou and Y. Harada, "B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM," Journal of Theoretical & Applied Electronic Commerce Research, vol. 17, no. 2, pp. 458-475, Jun. 2022. Retrieved April 8, 2023 from https://www.mdpi.com/0718-1876/17/2/24.

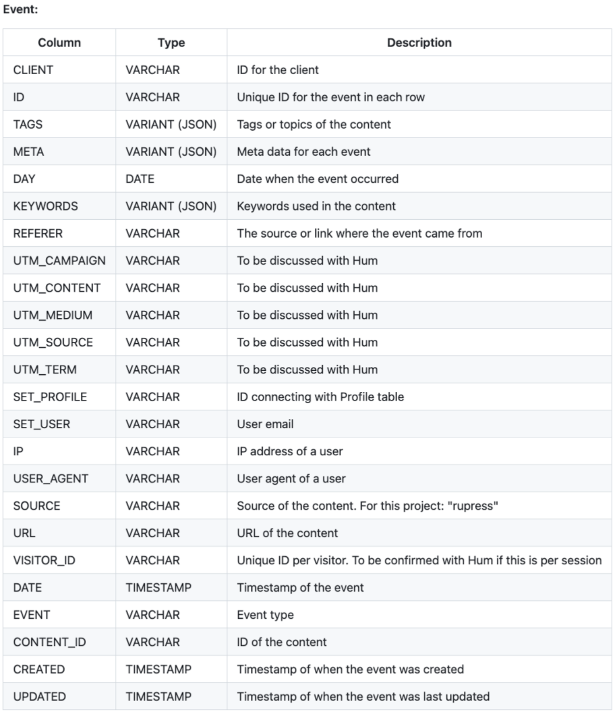
[5] V. Duarte, S. Zuniga-Jara and S. Contreras, "Machine Learning and Marketing: A Systematic Literature Review," in IEEE Access, vol. 10, pp. 93273-93288, 2022, doi: 10.1109/ACCESS.2022.3202896. Retrieved April 8, 2023 from https://ieeexplore.ieee.org/document/9869838.

[6] M. Pradana and H. Ha, "Maximizing Strategy Improvement in Mall Customer Segmentation using k-means Clustering," Journal of Applied Data Sciences, vol. 2, no. 1, pp. 19-25, 2021, doi: 10.47738/jads.v2i1.18. Retrieved April 1, 2023 from http://bright-journal.org/Journal/index.php/JADS/article/view/18.

[7] T. Kansal, S. Bahuguna, and T. Choudhury, "Customer Segmentation using k-means Clustering," in 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 135-139, doi: 10.1109/CTEMS.2018.8769171. Retrieved April 1, 2023 from https://ieeexplore.ieee.org/abstract/document/8769171.

[8] Kerzendorf, W.E., Patat, F., Bordelon, D. et al. Distributed peer review enhanced with natural language processing and machine learning. Nat Astron 4, 711–717 (2020). https://doi.org/10.1038/s41550-020-1038-y Retrieved April 1, 2023 from https://www.nature.com/articles/s41550-020-1038-y.

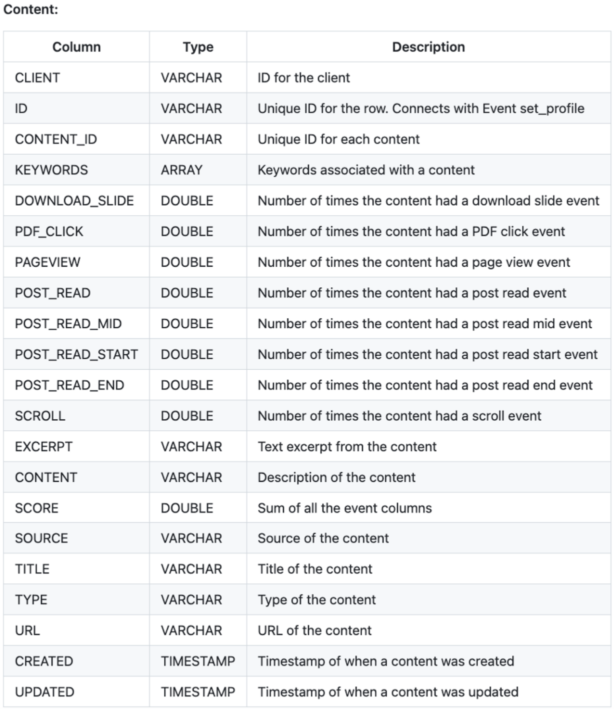
Appendix



Event:

Column	Type	Description
CLIENT	VARCHAR	ID for the client
ID	VARCHAR	Unique ID for the event in each row
TAGS	VARIANT (JSON)	Tags or topics of the content
META	VARIANT (JSON)	Meta data for each event
DAY	DATE	Date when the event occurred
KEYWORDS	VARIANT (JSON)	Keywords used in the content
REFERER	VARCHAR	The source or link where the event came from
UTM_CAMPAIGN	VARCHAR	To be discussed with Hum
UTM_CONTENT	VARCHAR	To be discussed with Hum
UTM_MEDIUM	VARCHAR	To be discussed with Hum
UTM_SOURCE	VARCHAR	To be discussed with Hum
UTM_TERM	VARCHAR	To be discussed with Hum
SET_PROFILE	VARCHAR	ID connecting with Profile table
SET_USER	VARCHAR	User email
IP	VARCHAR	IP address of a user
USER_AGENT	VARCHAR	User agent of a user
SOURCE	VARCHAR	Source of the content. For this project: "rupress"
URL	VARCHAR	URL of the content
VISITOR_ID	VARCHAR	Unique ID per visitor. To be confirmed with Hum if this is per session
DATE	TIMESTAMP	Timestamp of the event
EVENT	VARCHAR	Event type
CONTENT_ID	VARCHAR	ID of the content
CREATED	TIMESTAMP	Timestamp of when the event was created
UPDATED	TIMESTAMP	Timestamp of when the event was last updated

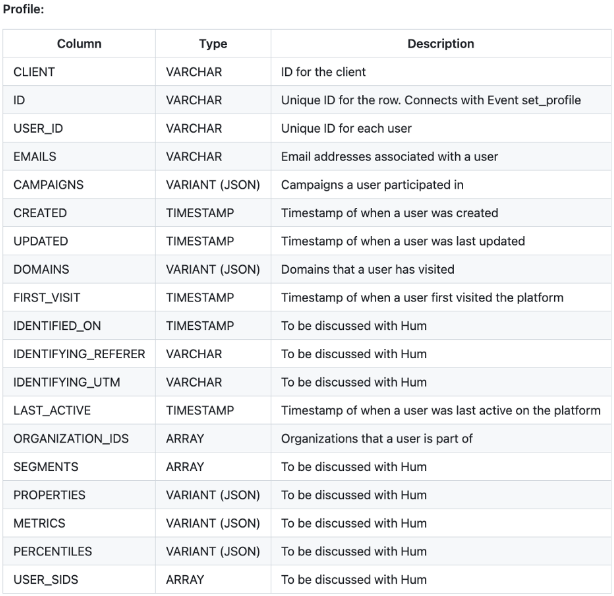
Fig. 8: Event table from schema



Content:

Column	Type	Description
CLIENT	VARCHAR	ID for the client
ID	VARCHAR	Unique ID for the row. Connects with Event set_profile
CONTENT_ID	VARCHAR	Unique ID for each content
KEYWORDS	ARRAY	Keywords associated with a content
DOWNLOAD_SLIDE	DOUBLE	Number of times the content had a download slide event
PDF_CLICK	DOUBLE	Number of times the content had a PDF click event
PAGEVIEW	DOUBLE	Number of times the content had a page view event
POST_READ	DOUBLE	Number of times the content had a post read event
POST_READ_MID	DOUBLE	Number of times the content had a post read mid event
POST_READ_START	DOUBLE	Number of times the content had a post read start event
POST_READ_END	DOUBLE	Number of times the content had a post read end event
SCROLL	DOUBLE	Number of times the content had a scroll event
EXCERPT	VARCHAR	Text excerpt from the content
CONTENT	VARCHAR	Description of the content
SCORE	DOUBLE	Sum of all the event columns
SOURCE	VARCHAR	Source of the content
TITLE	VARCHAR	Title of the content
TYPE	VARCHAR	Type of the content
URL	VARCHAR	URL of the content
CREATED	TIMESTAMP	Timestamp of when a content was created
UPDATED	TIMESTAMP	Timestamp of when a content was updated

Fig. 9: Content table from schema



Profile:

Column	Type	Description
CLIENT	VARCHAR	ID for the client
ID	VARCHAR	Unique ID for the row. Connects with Event set_profile
USER_ID	VARCHAR	Unique ID for each user
EMAILS	VARCHAR	Email addresses associated with a user
CAMPAIGNS	VARIANT (JSON)	Campaigns a user participated in
CREATED	TIMESTAMP	Timestamp of when a user was created
UPDATED	TIMESTAMP	Timestamp of when a user was last updated
DOMAINS	VARIANT (JSON)	Domains that a user has visited
FIRST_VISIT	TIMESTAMP	Timestamp of when a user first visited the platform
IDENTIFIED_ON	TIMESTAMP	To be discussed with Hum
IDENTIFYING_REFERER	VARCHAR	To be discussed with Hum
IDENTIFYING_UTM	VARCHAR	To be discussed with Hum
LAST_ACTIVE	TIMESTAMP	Timestamp of when a user was last active on the platform
ORGANIZATION_IDS	ARRAY	Organizations that a user is part of
SEGMENTS	ARRAY	To be discussed with Hum
PROPERTIES	VARIANT (JSON)	To be discussed with Hum
METRICS	VARIANT (JSON)	To be discussed with Hum
PERCENTILES	VARIANT (JSON)	To be discussed with Hum
USER_SIDS	ARRAY	To be discussed with Hum

Fig. 10: Profile table from schema