Engagement Time Machine

User Engagement Analysis for Academic Publishers

Griffin McCauley, Theo Thormann, Eric Tria, and Jake Weinberg

University of Virginia | School of Data Science

Abstract

In this project, we will discuss the processes of using machine learning to classify and predict user engagement for an academic publisher. Our project consists of a K-means clustering model to determine what constitutes an 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 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. Our MLP model is able to predict whether a user is high- or low-quality with 95% accuracy.

Background

Our sponsor, Hum, operates in the academic publishing industry and utilizes its proprietary CDP to collect first-party data across clients' online content. This industry is just now experiencing the big data revolution, and greater understanding of user engagement patterns has massive business implications. By leveraging contemporary data science techniques, the hope is to enhance and optimize the currently inefficient peer reviewer selection process.

Goal: Engineer a novel set of user-level features and construct a model to accurately recognize high-quality, valuable users early on in their lifecycles

Data & Pipeline

- Our model is built off first-party data in a database maintained by our sponsor, Hum, that is sourced from a mid-size educational publisher with four journals.
- Using 3 main tables: **EVENTS**, **PROFILES**, **CONTENT**, we engineered 4 new features:

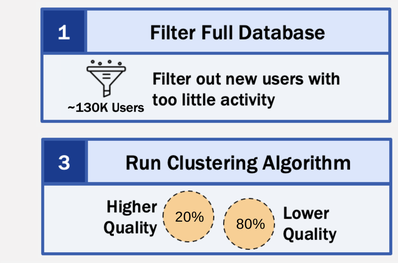
Feature	Description
Articles per Event	Ratio of distinct articles for each user action
% of Articles from Google	Articles read that were found through Google search
% of Content that are Articles	Consumed content that is are articles
Event Density	The number of events per day

- Data is accessed from a cloud data warehouse in **Snowflake**. The feature engineering and storage process is deployed on Amazon Web Services (AWS) **SageMaker** and **S3**.

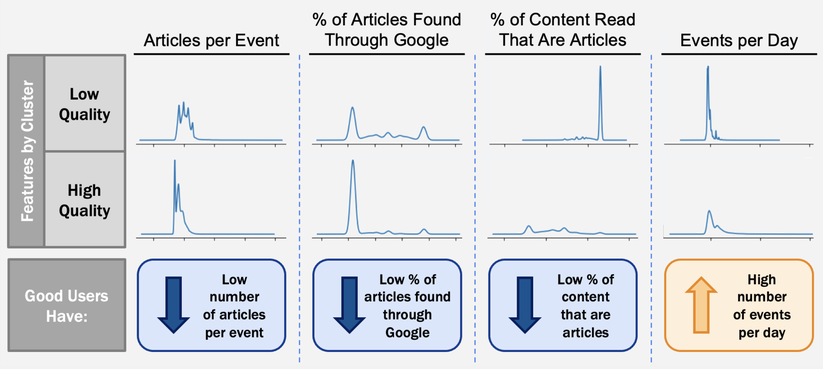
Snowflake | SageMaker | S3

Cluster Analysis

- No educational publishing industry standard for defining or quantifying a high-quality user
- Needed to define criteria to use as labels for training our classification model that distinguishes between "good" and "bad" users
- Used K-means clustering to split users into high- and low-quality groups (2 total clusters)
- Filtered total database to only include users reaching activity threshold of 16 events
- Clustered on 4 derived features: Articles per Event, % of Articles from Google, % of Content that are Articles, and Event Density
- Successfully derived distinct profiles by group; our features represent possible industry KPIs

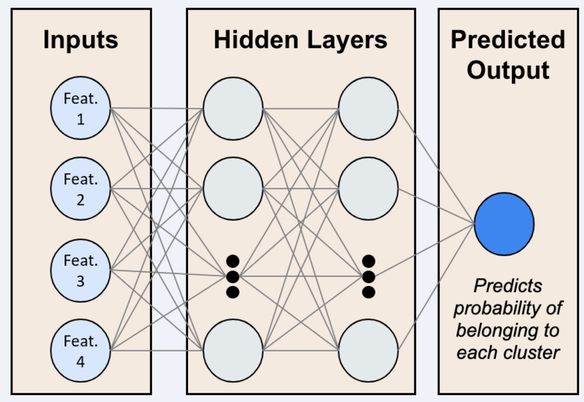


Note: for ease of interpretability and commercial viability, we created only two clusters rather than deriving the optimal number



Deep Learning Implementation

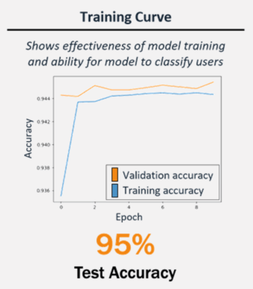
- Needed to design a robust model framework to perform cluster classification
- Used deep learning techniques to implement a multilayer perceptron (MLP) model
- The model architecture can be seen on the right and can be summarized as follows:
  - 4 input features
  - 2 10-dimensional hidden layers
  - 1 output node
- The input features are the same engineered ones previously discussed except now derived only from users' first 16 events

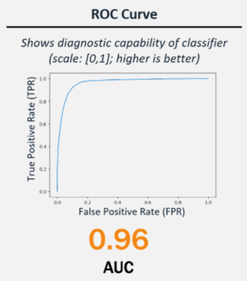


MLP Model Architecture for Binary Cluster Classification

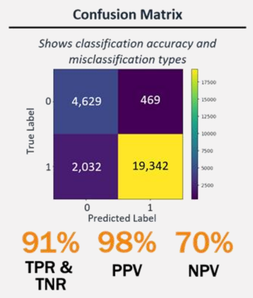
Results

- Training produced swift parameter convergence
- MLP model demonstrated strong ability to distinguish classes
- Needed to assess thresholds other than 0.5





- The ROC curve shows our model's impressive capacity to produce predictions which simultaneously yield low false positive rates and high true positive rates
- Threshold can be tuned to optimize tradeoff between false positives and negatives
- When jointly minimizing FPR and FNR, we get a threshold of 0.877
- This produces the adjacent confusion matrix



Summary/Future Work

- **Found that user lifetime behavior can be predicted very early on**
- Constructed a robust model framework that can be easily extended to other academic publishers
- Classified user engagement with high accuracy based on novel features
- **Next: Identify potential peer reviewers based solely on reading behaviors**
- Next: Tailor recommended content and ads based on user activity
- Next: Incorporate information for other clients and more granular user data

References

[1]Pradana, M., Ha, H. 2021. "Maximizing Strategy Improvement in Mall Customer Segmentation using K-means Clustering." Journal of Applied Data Sciences, 2(1), 19-25. doi:https://doi.org/10.47738/jads.v2i1.18. Retrieved April 1, 2023 from http://bright-journal.org/Journal/index.php/JADS/article/view/18

[2]Kansal, T., Bahuguna, S., Choudhury, T. 2018. "Customer Segmentation using K-means Clustering." 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

Acknowledgements

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.

University of Virginia | Data Science