REAL-TIME STREAMING FEATURE GENERATION

RESPONSIBLE RESEARCH AND INNOVATION OF RECOMMENDATION SYSTEMS

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Param Damle

December 4, 2023

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Rider Foley, Department of Engineering and Society

Rosanne Vrugtman, Computer Science

Introduction

Behind each action we take is a network of hidden algorithms designed to influence our desires. Companies have increasingly earned greater profits by predicting user behavior with models trained on large quantities of data, a process ingrained into the technology that we interact with every day. The scope of these systems is immense—anything from the rental you book on Airbnb to what you watch on Netflix to everyday Google searches is factored into constructing a profile of you across various databases (Jeckmans et al., 2013). As platforms such as these grow more ubiquitous in countries like the United States, user data monetization is dominating other revenue streams within the cloud services space (Dessemond, 2020). The collection, retention, and distribution of this data is just the tip of the iceberg in terms of how it can be used for profit, but there's deeper concern in the less regulated domain of using this data for recommending products, people, and programs of action. For example, targeted recommendations generated Google $147 billion in revenue in 2020 and account for more than 35% of Amazon's sales, but data regulations involving informed consent rarely extend to the systems that use the data. When outcry over recommendation privacy first started in the late 2010s and Apple introduced an opt-out feature for having one's data factored into personalization, 96% chose to withdraw their personal information (Küçükgül et al., 2022).

I witnessed this in a previous summer internship, where I built a data streaming pipeline that channeled user data from a banking company's website to a machine learning model that would predict the best products to market to the user, all in real-time. The data collected ranged from the obvious (such as buttons the user clicked) to the obscure and somewhat deceptive, such as user demographic data and their interactions from *other* websites obtained via cookies and linked to their unified profile. Not only is it unclear to most users that data is being collected in

this manner when they visit an inconspicuous website, it is less clear that the information is being leveraged to instantaneously recommend new products to them. As engineers like me design the intrusiveness of these systems, they become increasingly aware of their responsibility to implement safeguards against malicious data handling and behavior influence into the systems. To ensure this role is being fulfilled, I seek to examine how the proliferation of recommendation systems has affected advocacy for and protection of user data privacy. Although the technical specifications of these systems require ever-increasing sources of data to boost performance, I expect the consideration of social implications to have increased attention towards ethical concerns and implementations of privacy-conscious mechanisms.

Societal Role of Recommendation Systems

The invention of social media in the 1990s produced a chronological "timeline" biased heavily in favor of spam, requiring developers to implement their own guardrails for recommending content (Meserole, 2022). As these hard-coded rules didn't scale, developers incorporated machine learning and behavioral modeling. Although this evolution seems straightforward, the implications on the communities involved are not. Understanding why these systems are as pervasive and intrusive as they are requires analysis of all actors, from those designing the system (engineers) and prescribing its greater goal (executives) to those affected by it (users) and those with a capacity to change it (lobbyists and legislators).

Framed using Hughes' lens (1987), these recommendation engines are technological systems that underwent a series of innovations to produce more accurate predictions at the cost of gathering a greater breadth and depth of personal information. We are now at the stage where various entities (in the e-commerce and social media spaces, primarily) are competing to give you the best recommendations by leveraging even more of your data. This competition

incentivizes each company to transfer (copy and adapt) the most effective methods from its rivals, such as how both Facebook and Twitter invested in replicating TikTok's successful strategy of "inventorying" data from users across their platform to influence the timeline of any one user (Meserole, 2022). Given the size of the user bases, these applications and their underlying models are embedded in the way people shop for goods, consume content, and connect with each other. Companies like Google, Amazon, and Facebook are among the largest drivers to the US economy, and other firms riding their algorithmic coattails hire a significant portion of the software labor force. Under Hughes's framework, these recommendation algorithms have acquired momentum—their ubiquity and perceived necessity allows them to grow without legitimate challenge. It's difficult to argue against data farms if every major store and media company relies on them for traction, and it'll get increasingly difficult as data collectors consolidate this power and make using their data addictive (DeLeon, 2019). This is compounded by the influence platforms like Facebook have on people who don't use their service but live in a region where Facebook is popular (Kotliar, 2021).

Although this classifies as an indirect impact on the user, advocates have found greater success in targeting direct impacts such as a breach of privacy (breadth, depth, or lifetime) and unethical recommendations (revealing connections, funneling users into extremist echo chambers, and so on). The extensive privacy concerns have led to an increase in digital rights advocacy over recent years, as well as technical changes to the systems, such as randomization, data anonymization, and end-to-end encryption when transmitting personal information. The problem of echo chambers has real world consequences such as the spread of COVID misinformation, far-right extremism, and perhaps an insurrection against one's government (Meserole, 2022). Hughes would deem the governments affected by and responsible for these

issues reverse salients, or entities that respond to new developments far slower than they arise. Since legislation lags behind a technological landscape accelerated by the iterability of software, advocates often expect the companies instituting these systems to proactively address their ethical consequences. Despite this expectation, there's incentives for companies to pursue projects that collect data more aggressively and, potentially, nefariously.

Technical Incentives in Recommendation Platforms

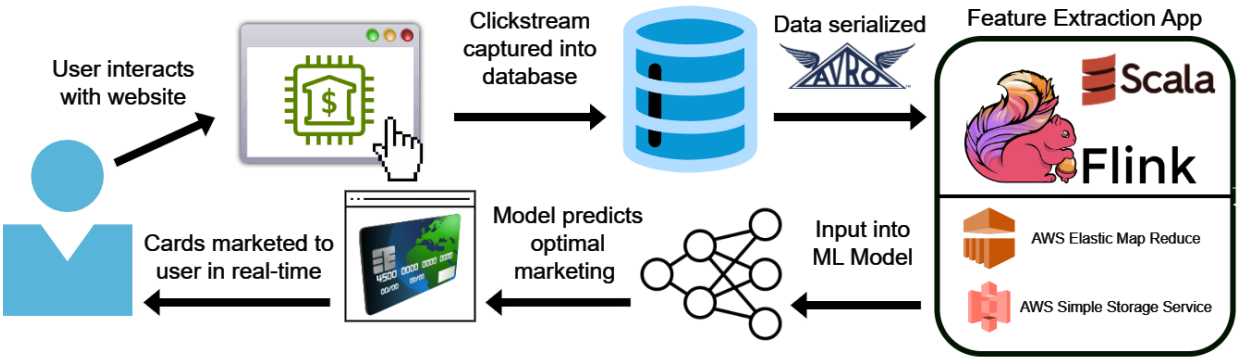
The largest predictor of machine learning performance is not the model's hyperparameters, but the quality of the training set and the features derived from it (Wang & Shah, 2021). This has created a "data imperative" for maximizing information collection, whereby algorithms can craft the most specific profile for each user (Seaver, 2021). Specificity provides more personalized recommendations, leading to a uniquely memorable customer experience, which in turn produces more loyal sources of revenue for a company (Vas, 2021).

To generate this revenue, the bank I interned at captured user data from its website such as the clickstream (cursor movement), cookies capturing user preferences across all websites, and information about the user's device and network. After market hours every night, a model transformed this data into features to input into a model, which predicted the optimal credit cards to advertise to the user the next time they visited the website. This presented an issue of staleness, measured by how much time passes between the collection of data and the input of relevant features into the algorithm (Talati et al., 2023). My solution was streaming feature generation, which Iguazio regards as the contemporary "Holy Grail" of data science because of its ability to adapt to changes in user behavior in real-time (2023). With this setup, the backend could calculate the best credit card offers to market to a user as they browse the website and display the optimal ads relating to these offers before the user even leaves the website.

To achieve this, I deployed an instance of Apache Flink, a stream-processing framework, using Scala, a language that allowed chaining transformations so that raw clickstream data was iteratively refined into the relevant feature set for my model. These transformations varied from stateless changes, like converting a timestamp from 12-hour to 24-hour format, to stateful calculations, like averaging the number of pages visited by a given IP address in the last 10-minute window. The logic of this feature extraction app alone was insufficient to meet the requirements of the full system, as careful construction of the data pipeline was necessary to ensure the turnaround time of the entire process was under a second.

Figure 1

Real-time Streaming Feature Generation Pipeline



The data pipeline I built first connected to an endpoint that served live clickstream data in Avro format, which was optimized for compressed data encoding, allowing transmission of the "raw" data to be as efficient as possible. The Flink app ran on an AWS Elastic Map-Reduce cluster, which I chose because it allocates work across multiple instances to parallelize stream processing and is supported by Simple Storage Services with the fastest disk access and modification times. Lastly, the extracted features were connected to the preexisting prediction model with DynamoDB as a temporary storage. After my app was integrated with all these

components, the overall latency from data ingestion to prediction computation was about 100 milliseconds, representing a major improvement from the 1-day latency of the batch system.

The project's success was evaluated directly on technical metrics like latency and throughput and indirectly using changes in sales, a standard measure for the business value of a recommender system (Jannach & Zanker, 2022). No attention was given to the social and personal impacts of the system I was tasked with designing, even though I saw issues in how easily I could access people's information in real time when working on my project. As these systems expand, they introduce vulnerabilities posed by the centralization of data and ingestion from external sources like third-party cookies and Internet of Things devices (Wang, 2021; Himeur et al., 2022). When engineers are incentivized to favor prediction quality over privacy preservation as I was, rapid innovation leads to escalation in ethically ambiguous business practices.

Analyzing Attention to Privacy Concerns

As government regulations lag behind the aforementioned innovation-violation duality, engineers bear increasing responsibility to institute guardrails against unethical system design (Zhang et al., 2014). So, is the technical community behind these systems adequately responding to concerns posed by new advances in the field? To assess this, I will thematically analyze the corpus of proceedings from The ACM Conference Series on Recommender Systems (RecSys) between 2007-2023 for mentions of data privacy and user protection. I'll employ Stavrakis's extraction method (2023) to pull raw text from each submitted paper and workshop PDF. I will survey a variety of embeddings to convert this textual dataset into a numerical dataset with which papers can be compared with each other for semantic analysis, which at its simplest is the Euclidean distance between vectors (Pawar & Mago, 2018). I plan to pursue both supervised

approaches, like correlating numerical similarity scores to conference year, and unsupervised approaches, like clustering the embedded vectors and analyzing similarities within clusters. After normalizing by the number of submissions to each conference year, I expect to see an increase in references to privacy topics within the average paper. If this hypothesis is accepted, RecSys can serve as a model technical field where attention to social issues has, at minimum, kept pace with increased attention and innovation. This would mean that coverage within academic circles is not a reverse salient with respect to the growing technological momentum of recommendation systems. If my claim is rejected, however, my research will highlight a need for the RecSys community to invest more resources into addressing the societal consequences of such systems.

Conclusion

As recommendation systems play an increasing role in online interaction, commerce, and democracy, they pose concerns such as user privacy, declining public trust, echo chamber incubation, and algorithmic discrimination (Jannach & Zanker, 2022). After my streaming feature extraction pipeline was labeled a success due to its achievement of business objectives, I'm seeking to illuminate whether the field is achieving ethical objectives. This research will serve as a litmus test to how well the industry is acknowledging threats to user privacy. Are societal considerations being covered more in technical literature, or do engineers continue to build more powerful systems without sufficient guardrails? The UN describes global development fueled by a "cloud economy", where public welfare is tied to the systems engineers deploy on the internet (UNCTAD, 2013). Ensuring recommendation systems remain conscious of the full range of social implications will provide an ethically sustainable path to global development.

References

DeLeon, H. (2019, April 24). The Ethical and Privacy Issues of Recommendation Engines on Media Platforms. *Towards Data Science*. https://towardsdatascience.com/the-ethical-and-privacy-issues-of-recommendation-engines-on-media-platforms-9bea7bcb0abc

Dessemond, E. G. (2020). Restoring Competition in "Winner-Took-All" Digital Platform Markets. *United Nations Conference on Trade and Development*, *40*. https://doi.org/10.18356/5ee7948e-en

Himeur, Y., Sohail, S. S., Bensaali, F., Amira, A., & Alazab, M. (2022). Latest trends of security and privacy in recommender systems: A comprehensive review and future perspectives. *Computers & Security*, *118*, 102746. https://doi.org/10.1016/j.cose.2022.102746

Hughes, T. (1987). The Evolution of Large Technological Systems. In W. Bijker, T. Hughes, & T. Pinch (Eds.), *The Social Construction of Technological Systems* (pp. 45–76). MIT Press.

Iguazio. (2023, February 7). How to Build Real-Time Feature Engineering with a Feature Store. *AI Infrastructure Alliance*. https://ai-infrastructure.org/how-to-build-real-time-feature-engineering-with-a-feature-store/

Jannach, D., & Zanker, M. (2022). Value and Impact of Recommender Systems. In F. Ricci, L. Rokach, B. Shapira, & P. Kantor (Eds.), *Recommender Systems Handbook 3rd ed.* (pp. 519–546). Springer.

Jeckmans, A. J. P., Beye, M., Erkin, Z., Hartel, P., Lagendijk, R. L., & Tang, Q. (2013). Privacy in Recommender Systems. In N. Ramzan, R. van Zwol, J.-S. Lee, K. Clüver, & X.-S. Hua (Eds.), *Social Media Retrieval* (pp. 263–281). Springer. https://doi.org/10.1007/978-1-4471-4555-4_12

Kotliar, D. M. (2021). Who Gets to Choose? On the Socio-algorithmic Construction of Choice. *Science, Technology, & Human Values*, *46*(2), 346–375. https://doi.org/10.1177/0162243920925147

Küçükgül, C., Özer, Ö., & Wang, S. (2022). Recommender Systems under Privacy Protection. *SSRN*. https://doi.org/10.2139/ssrn.4138757

Meserole, C. (2022, September 21). How do recommender systems work on digital platforms? *Brookings*. https://www.brookings.edu/articles/how-do-recommender-systems-work-on-digital-platforms-social-media-recommendation-algorithms/

Pawar, A., & Mago, V. (2018). *Calculating the similarity between words and sentences using a lexical database and corpus statistics* (arXiv:1802.05667). arXiv. https://doi.org/10.48550/arXiv.1802.05667

RecSys – ACM Recommender Systems. (n.d.). The ACM Conference Series on Recommender Systems. Retrieved October 25, 2023, from https://recsys.acm.org/

Seaver, N. (2021). Seeing like an infrastructure: Avidity and difference in algorithmic recommendation. *Cultural Studies*, *35*(4–5), 771–791. https://doi.org/10.1080/09502386.2021.1895248

Stavrakis, G. (2023, September 21). Extracting text from PDF files with Python: A comprehensive guide. *Towards Data Science*. https://towardsdatascience.com/extracting-text-from-pdf-files-with-python-a-comprehensive-guide-9fc4003d517

Talati, A., Parkhe, M., & Lukiyanov, M. (2023, February 16). Best Practices for Realtime Feature Computation on Databricks. *Databricks*. https://www.databricks.com/blog/2023/02/16/best-practices-realtime-feature-computation-databricks.html

United Nations Conference on Trade and Development. (2013). The cloud economy ecosystem. In *Information Economy Report* (pp. 1–14). United Nations. https://doi.org/10.18356/63479764-en

Vas, G. (2021, September 21). How Recommendation Systems Comply with Privacy Regulations. *Gravity Research & Development*. https://www.yusp.com/blog-posts/recommendation-systems-comply-with-privacy-regulations/

Wang, A., & Shah, K. (2021, March 4). Building Riviera: A Declarative Real-Time Feature Engineering Framework. *DoorDash Engineering*. https://doordash.engineering/2021/03/04/building-a-declarative-real-time-feature-engineering-framework/

Wang, Y. (2021, December 15). How We Can Get Better Recommendations Without Giving Up Our Privacy. *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2021/12/15/how-we-can-get-better-recommendations-without-giving-up-our-privacy/?sh=317760315a0a

Zhang, B., Wang, N., & Jin, H. (2014). *Privacy Concerns in Online Recommender Systems: Influences of Control and User Data Input*. 159–173. https://www.usenix.org/conference/soups2014/proceedings/presentation/zhang