JONATHAN ESPERANZA

jonathanesperanza67@gmail.com | linkedin/jonathanesperanza | github/jon-esperanza

EXPERIENCE

Credit Karma Charlotte, NC

Software Engineer III, Recommendations Platform *September 2024 - Present*

- Spearheaded the development of an AI-powered observability agent to debug complex recommender system request logs. The agent uses RAG architecture on BigQuery log data, reducing manual debugging time for on-call engineers by an estimated 70%. This project won our company-wide hackathon and was adopted as an official internal tool.
- Led the re-architecture of the ML training data pipeline, migrating from JSON to Parquet format. This initiative reduced annual storage costs by 25% ($850k/year) and accelerated feature retrieval for model training by ~4x, enabling faster experimentation and more frequent model refreshes.
- Delivered a recommendation blending system that accepts multiple candidate lists and applies a co-ranking function. This unified the user's journey, contributing to a holistic app optimization strategy.
- Designed and built a self-serve method for real-time model formula experimentation, empowering data scientists to independently define, deploy, and A/B test objective functions for deep learning rankers. This reduced the iteration cycle for new ranking strategies from weeks to hours.
- Migrated the real-time model serving architecture to NVIDIA Triton, reducing p99 inference latency by 60ms. This unlocked the ability to productionize diverse models which were previously infeasible, like PyTorch and ONNX.

Software Engineer II, Offline Recommendations Platform *February 2023 - September 2024*

- Owned the ML pipelines driving 1.5B personalized notifications per month, attributing to 24% of sitewide traffic.
- Deployed several ML and data features through cross-functional efforts with marketing, product, and data science teams resulting in notable lifts to engagement and revenue.
- Achieved a remarkable 70% cost reduction on Cloud Dataflow, saving $1.8 million annually, by optimizing performance for large, inefficient data joins in our batch inference ML pipelines.
- Designed, built, and delivered a pre-production environment, significantly reducing outages associated with major feature releases by 95% and unlocking performance benchmarks and tuning.

Software Engineer I, Recommendations: Prediction Services *August 2022 - February 2023*

- Contributed to scaling, maintaining, and operating critical model serving and feature store services to serve recommendations to 150M+ members in real time.
- Played a key role in migrating the Tensorflow model scoring service from TF 1.x to TF 2.x, ensuring a smooth transition and improved latency/performance.

EDUCATION

University of North Carolina in Charlotte Charlotte, NC

B.S Computer Science, Software Engineering *August 2022*

SKILLS

Programming Languages	Java, Scala, Python, SQL
Cloud	Google Cloud Platform (Cloud Storage, BigQuery, BigTable, Cloud Dataflow, Cloud Functions, Pub/Sub); Experience with AWS and Microsoft Azure
Technologies and Frameworks	Apache Beam, Tensorflow, Apache Spark, Kubernetes, Docker, Terraform, Airflow
Development Tools	Git, CircleCI, GitLab, GitHub