Rahul Mehta

Senior Machine Learning Engineer · Ranking & LLM Systems

rahul.mehta@example.com | (650) 555-0185 | Palo Alto, CA | linkedin.com/in/rahulmehta-example | github.com/rahulmehta-example

SUMMARY

Machine learning engineer with 8 years building ranking, recommendation, and LLM systems in production. Led a ranking model that raised marketplace conversion from 3.4% to 4.1% ($22M a year) and cut LLM serving cost by 55% at the same latency.

EXPERIENCE

Larkspur Marketplace Jun 2022 - Present

Senior Machine Learning Engineer *Palo Alto, CA*

- Led the move to a two-tower retrieval and transformer ranking model, raising search conversion from 3.4% to 4.1% (about $22M a year)
- Owned LLM serving for listing generation on vLLM, cutting GPU cost per 1k requests by 55% at the same p95 latency
- Designed an offline-online evaluation framework that made 80% of offline wins hold up in A/B tests, up from 45%
- Mentored 5 engineers and set the team's model review process across 3 ML squads

Lumen Analytics Jul 2019 - May 2022

Machine Learning Engineer *Seattle, WA*

- Built a churn prediction model in XGBoost scoring 4M accounts a week, helping retention cut churn from 9% to 6.5%
- Created a feature store on Feast and Spark that cut new-model feature work from 3 weeks to 4 days
- Shipped model monitoring for drift and data quality across 14 production models

Orchard Health Jun 2017 - Jun 2019

Data Scientist *Seattle, WA*

- Built a no-show prediction model that let clinics overbook carefully, raising filled appointments from 86% to 93%

EDUCATION

Carnegie Mellon University Aug 2015 - May 2017

M.S. Machine Learning *Pittsburgh, PA*

University of California, San Diego Sep 2011 - Jun 2015

B.S. Computer Science *San Diego, CA*

SKILLS

ML & Modeling: PyTorch, XGBoost, Transformers, Two-tower retrieval, LLM fine-tuning (LoRA), A/B testing

Platform: Python, Spark, Ray, vLLM, Feast, Kubernetes, MLflow, GCP (Vertex AI)