Shashank Srinivasan

+1 404 519 4936 | contact@knhash.in | knhash.in | github.com/knhash | linkedin.com/in/knhash

Seattle, WA | Authorized to work in the US (STEM OPT, 36 months); no immediate sponsorship required

Senior Machine Learning Engineer with 6 years building real-time recommendation and ranking systems. Served personalized content to **100M DAU** at **sub-50ms p99**, lifting **conversion +12%** and engagement +5%. MS in CS (HPC specialization), Georgia Tech, May 2026.

Experience

Senior Machine Learning Engineer, Walmart | *Dec 2024 – Jul 2025 (concurrent with MS)*

- Cut per-experiment compute cost by **40%** on floor-price and adtech recommendation ML pipelines by profiling PySpark bottlenecks, rewriting shuffle-heavy transformations, and right-sizing batch processing.
- Built a **Text2SQL agentic system** with plain-English querying, visualizations, and active chat, giving non-technical teams self-serve analytics over business and ML pipeline data.
- Improved production ML reliability and release velocity by building an **observability framework** and standardizing model build, test, and deployment with GitHub Actions.

Machine Learning Engineer III, Glance (InMobi) | *Jul 2022 – Dec 2024*

- **Founding engineer** on the rebuilt Glance feed personalization platform serving **100M DAU**; owned cold-start and sparse-user personalization end to end.
- Designed the sparse-user modeling stack: multinomial Thompson-sampling bandits, **Elo-style exploration** adapted from game matchmaking, and **lookalike models** (SVD/K-Means, then node2vec on app-ownership signals) that beat production baselines on time spent and reward; **+18.38%** time spent for cold users.
- Delivered content at **sub-50ms p99** via low-latency serving and streaming pipelines (prediction services, Vertex AI feature store, Kafka, ELK logging, Airflow retraining); **+12% conversion** and **+5% engagement**.
- Built and optimized a rewards pipeline over **billions of clicks**, cutting runtime from **5 hours to 20 minutes**.
- Standardized deployment, on-call, and automated canary across the team; mentored engineers and led an org-wide paper-reading session.

Data Scientist II, Glance (InMobi) | *May 2021 – Jun 2022*

- Built Glance's **first live-commerce recommendation system** (Roposo) on graph embeddings at sub-50ms, and the feed team's first Python prediction service (sub-100ms, in two weeks): **+6.54%** lift over control and the best-performing model in production.
- Shipped real-time, NLP-based **quiz generation over live video** (GCP), demoed at the company AI Demo Day; integrated OEM and OTT content with external partners (Docubay, EpicON, Zee5).

Founding Data Scientist, Bright Money | *Jun 2019 – Apr 2021*

- Founded the data science function and built **Debt Manager**, Bright's flagship AI product that analyzes a user's finances and autonomously schedules debt payments; on Airflow and Django, it served **20K+ users/day** at **$100K+ daily transaction volume** and drove close to **100% of revenue** by 2020.
- Shipped multiple 0-to-1 ML products: bill identification, affordability scoring, overdraft prediction, and liability inference for risk and retention workflows.
- Built time-series forecasting for income, expense, and account balance (Prophet, SARIMA) and segmentation models powering activation, retention, and monetization.

Education

Master of Science, Computer Science, Georgia Institute of Technology | *Aug 2024 – May 2026*

- Specialization: High Performance Computing.
- Research: Analog compute-in-memory error mitigation with Prof. Hyesoon Kim
- Coursework: GPU Hardware & Software, High-Performance Parallel Computing, HPC Architecture, Scientific ML, Deep Learning, Designing for Curiosity.

Bachelor of Engineering, Computer Science, Sir M. Visvesvaraya Institute of Technology | *May 2018*

- First Class with Distinction.
- Thesis: Image Regeneration with Generative Models; used CapsNet as a GAN discriminator for semantic inpainting on MNIST and face images.

Selected Systems & Research

Embedded Drift Sentinel (under review at ASPLOS 2027): Improved day-360 ImageNet **top-1 accuracy by +14 pp** on average over GDC across ResNet-18/34/50 and GoogLeNet on a 360-day PCM drift horizon (ResNet-50: 68% vs. 53% GDC, vs. 17% uncorrected). Designed a companion-crossbar architecture for online drift correction in analog compute-in-memory accelerators; eliminates main-array probe downtime at 0.3–1.7% cell overhead.

Physics-Informed RL for Plasma Control: Trained PPO policies in TORAX (auto-differentiable 1D core transport simulator) on a 50-parameter partially-observed state for ITER tokamak fusion control; achieved **1.7 ms average inference** (3.5 ms max), well under the 50 ms real-time threshold.

Meibomian Gland Segmentation: Published CNN-based meibography segmentation work in Biomedical Signal Processing and Control; developed custom augmentation and gland-health metrics validated against clinical expert annotations across tabletop and prototype handheld imagers.

Automated Debt Management Patents: R&D contributor; US20220261886A1, US20230075411A1.

Honors & Teaching

ACM ICPC 2015: Qualified to regional level; honorable mention.

Section Leader, Code in Place, Stanford University: Taught Python to 30+ students per cohort in 2021, 2023, and 2024.

Technical Skills

Languages: Python, C/C++, Go, SQL, PySpark, Shell

Artificial Intelligence & Machine Learning: Recommendation Systems, Candidate Generation, Retrieval & Ranking, Multi-Armed Bandits, Collaborative Filtering, Graph ML (node2vec, embeddings), Deep Learning, NLP, LLMs & Agentic Workflows (LangChain), Reinforcement Learning, Time-Series Forecasting (Prophet, SARIMA), A/B Testing; PyTorch, scikit-learn, XGBoost/LightGBM

MLOps & Infrastructure: Docker, Kubernetes, Airflow, Argo, Kafka, Spark, Vertex AI Feature Store, Neptune, GitHub Actions (CI/CD), ELK, Prometheus, Grafana; CUDA, MPI, OpenMP, GPU Profiling (Nsight Systems, nvprof), Distributed Systems, Scientific ML (PINNs)

Cloud & Data: GCP, AWS, Azure, PostgreSQL, Cosmos DB, Firebase, Tableau, Metabase, Django, Linux, Git