Kiran Nagabhushan Kumar

krankumar@gmail.com|Phone: 361-522-2789| kirankumar.ai
linkedin.com/in/kirankumarn|github.com/knkumar

Summary

Enterprise AI/ML leader with 10+ years of experience designing and deploying production-scale machine learning, LLM, and decision-support systems across enterprise platforms. Proven track record translating AI capabilities into measurable business outcomes through architecture design, prototyping, experimentation, and cross-functional stakeholder leadership.

EXPERIENCE

ML Engineer Tech Lead, Chewy **Sep 2024 – Present**

- Shaped enterprise AI strategy for marketing use cases by architecting an LLM/agent prototype focused on orchestration, memory, and contextual retrieval, helping define the organization's AI roadmap.
- Designed and executed robust experimental measurement strategies (A/B tests, geo/holdout designs) and communicated limitations/confounders to senior stakeholders to drive data-informed decisions.
- Led marketing measurement and advertising analytics initiatives supporting $500M+ annual media spend; built causal MMM + incrementality tooling to estimate channel lift and marginal ROI for budget allocation and pacing decisions.
- Developed production guardrails (data quality checks, drift/backtest monitoring, observability dashboards) that improved reliability and reduced model R&D iteration time by 60% across the team.
- Led and mentored 5+ engineers/scientists; set analytical standards (experiment reviews, metric definitions) and accelerated delivery across multiple cross-functional workstreams

ML Engineer Tech Lead, EquipmentShare **May 2023 - July 2024**

- Architected LLM-powered search and recommendation engine using RAG, projected to increase net income by 15% ($2M+ annually)
- Built production LLM gateway on Kubernetes with Flask/Docker, handling 10K+ daily requests with 99.9% uptime
- Designed ROI analysis framework using vector databases and AWS infrastructure, validating $500K+ product investment
- Improved asset uptime data quality by 25% through operational state modeling, reducing maintenance costs
- Established on-call processes with PagerDuty/Datadog, reducing incident response time by 75%

Research Scientist, Convoy **December 2021 - March 2023**

- Spearheaded the development of a cost forecasting platform using Tab Transformer Foundation models, adapter modules, fine-tuning, and transfer learning, improving operational efficiency by 2x.
- Prototyped a human-in-the-loop explainability architecture incorporating SHAP values, permutation importance, and ensembling to create a feedback loop, leading to a 3% increase in revenue.
- Led the development and implementation of feature flags for A/B test experiments, collaborating with the front-end team to productionize these features, achieving a 2% increase in revenue.
- Authored a whitepaper proposing an innovative two-stage auction model for a clean-miles-upgrades program.

Senior ML Engineer, Metromile **August 2021 - December 2021**

- Directed stakeholder discussions to define requirements for AI-driven renewal underwriting, integrating factors such as endorsements, prior driving history, credit, and demographics to achieve an **5%** reduction in loss ratio.
- Authored and presented detailed proposals, strategic roadmaps, and implementation plans for deploying advanced underwriting models and led the development of a feature store.

Data Scientist, Viaduct **February 2021 - August 2021**

- Refined and translated requirements into precise modeling and technical statements across multiple product lines, reducing turnaround time by **2x**. Collaborated closely with Product Managers to ensure alignment and efficiency.
- Enhanced vehicle failure prediction recall by **12%** by generating 20 context features through extensive experimentation with telematics variables.
- Designed and architected a robust process for analyzing feature importance and trends across multiple model deployments, which refined feature inclusion criteria and improved model performance by **2%**.

- Optimized performance of spark job with pandas UDF for feature generation, resulting in **30%** faster experimentation.

Data Scientist, SAP **August 2019 - January 2021**

- Aligned customer objectives across operations, finance, and strategy, delivering impactful presentations to stakeholders that enhanced revenue by 6 million through the implementation of Industry 4.0 solutions.
- Collaborated with a team of UX and front-end engineers to integrate NLP analysis into an order status chatbot, enhancing its ability to recognize Named Entity elements for improved conversational intelligence in procurement, resulting in an annual ACV of 0.2 million.
- Led innovation assessment projects to align with organizational objectives and current capabilities, producing comprehensive presentation materials and detailed reports for executive management.
- Drove an organizational explainability initiative by developing a novel method for feature importance with concentration bounds, achieving a 6x gain in efficiency and facilitating easier adoption and market readiness.

Doctoral Researcher, Indiana University **August 2013 - August 2019**

- Conducted in-depth research on visual attention through experimental and modeling approaches resulting in the acquisition of research grants, the publication of two peer-reviewed articles, and six conference presentations.
- Invented an experimental paradigm that enabled the collection of **4000** data points within **20** minutes of testing per subject, significantly enhancing distributional fits by a factor of 4x, thereby advancing methodologies for upcoming clinical trials.
- Supervised and mentored **two** undergraduate research assistants in hypothesis development, experimental design, and data analysis, culminating in an **award-winning poster presentation**.
- Coordinated and moderated Cognitive science seminars for **12** months, fostering active participation and facilitating scholarly discussions.

Software Engineer, Altran Praxis **February 2010 - July 2010**
Software Engineer, Accenture **June 2007 - February 2010**

SKILLS

Languages: Python, pytorch, pyspark, pandas, scikit-learn, numpyro, SQL
Tools: git, Jupyter, AWS, Spark, Airflow, MLFlow, SAP Data Intelligence
Machine learning toolkit: LLM Fine tuning, adapter networks, TabTransformer, Hidden Markov Models, Time series signal processing, Optimization, Statistical testing, Bayesian methods, Statistical models

EDUCATION

Indiana University, Bloomington, **PhD** in ***Cognitive Science and Psychology*** **June 2020**
Indiana University, Bloomington, **M.S.** in ***Computer Science*** **May 2012**
M.S. Ramaiah Institute of Technology, Bangalore, India, **B.Eng** in ***Computer Science*** **June 2007**

PUBLICATIONS (https://scholar.google.com/citations?user=F6jxKfoAAAAJ&hl=en)

- Kumar, K.N., Harding, S.M., & Shiffrin R.M. (2018). Inferring attention through cursor trajectories. Proceedings of the 39th Annual Conference of the Cognitive Science Society.
- Kumar, K. N., Chandramouli, S. H., & Shiffrin, R. M. (2015). Salience, perceptual dimensions, and the diversion of attention. *The American journal of psychology, 128(2), 253-265*

Presentations

- Kumar, K.N., Harding, S.M., & Shiffrin R.M. (2018) Continuous dynamics of cursor trajectory. 2018 Midwest Cognitive Science Conference.
- Kumar, K. N., Chandramouli, S. H., & Shiffrin, R. M. (2013) Popout Attention with two foils: linear dependence and dimensional interaction. 34th Annual Conference of the Cognitive Science Society
- Kumar, K. N., Chandramouli, S. H., & Shiffrin, R. M. (2013) Popout Attention with two foils: linear dependence and dimensional interaction. Context and Episodic Memory Symposium (CEMS)