

SATHVIK RAJU

sathuraju@gmail.com | +1 628 688 7828 | San Francisco, CA | linkedin.com/in/sathvikraju | github.com/SATHVIKRAJU | sathvikraju.com

PROFILE

Staff-level Data Scientist with 8+ years building deep learning and optimization systems for resource-constrained production environments. Partnered directly with infrastructure engineering to turn operational signals into measurable outcomes including \$15M+ in annual savings. Drove executive-level narratives.

PROFESSIONAL EXPERIENCE

Senior Data Scientist Microsoft, San Francisco, USA

(Sept 2022 to Present)

Windows 365 Frontline: Intelligent Pre-start (ML)

- Designed, built and shipped the Intelligent Pre-start system end-to-end as the sole data scientist. Owned modeling, pipelines, evaluation, reporting and maintenance. Used a regressor combined with Gaussian KDE estimation to capture multimodal user activity and drove the system to GA in April 2026.
- Impact: roughly a 90-second reduction in connection time and about 70% of Frontline users experiencing faster logins. Delivered within a COGS budget under 16%.

Windows 365 Hibernation: Ensemble Modeling & Production

- Developed hybrid deep-learning transformer and ensemble models that improved device coverage and cost efficiency. HR models now drive about 45% of monthly COGS savings and \$15M in annual savings. Designed rollout strategies grounded in user-level performance analysis to avoid degrading end-user experience.
- Owned the business metrics, evaluation scripts and hibernation-hour tradeoff analysis used to validate models and gate production rollout. Optimized data preprocessing to cut backfill time and speed up iteration.
- Pushed models into production with a focus on scalability and provided on-call pipeline support. Established ML Production Rigor standards covering validation gates, data quality checks, monitoring and incident handling.

Copilot Scenarios: Applied LLMs

- Co-architected a Retrieval-Augmented Generation (RAG) pipeline using vector embeddings and semantic search over a vector database. Built an early evaluation framework for it, powering a custom support-case triage tool for a Windows virtualization product.
- Helped design the end-to-end workflow and build a data query engine accessible through Microsoft Teams, powered by GPT-3.5.
- Fine-tuned and prompt-engineered an early version of GPT-3.5 to automate incident-ticket summarization and topic labeling. Built a thorough validation framework that scored outputs against human-labeled ground truth to catch drift before rollout.

WCX ML Platform (MLOps) & Cross-Team Collaboration

- Built the end-to-end machine learning model pipeline spanning data generation, training, validation and inference on the team's unified MLOps platform, contributed to platform capabilities for experimentation and deployment reliability, and led cross-team workflow discussions across DS, engineering and product.

Senior Data Scientist OnPoint Solutions, San Francisco, USA

(Aug 2020 to Aug 2022)

- Served as senior technical lead on Cortex, the company's core ML platform, owning full-stack workflows from extraction through statistical inference, modeling and deployment. Partnered with product, design and data engineering to shape the roadmap, win enterprise buy-in for ML proofs of concept, and mentored data science interns as the team scaled.

Abnormal Behavior Detection & Recommendation Diagnostics Engine

- Led the end-to-end causal-inference and anomaly-detection framework across 10+ heater types, from ingestion through hypothesis testing to deployment. Used Gaussian Mixture Models, autoencoders and copula-based outlier detection to flag high-impact anomaly points.
- Delivered the company's highest-revenue ML product. Each detected anomaly cut operating cost by \$50K and reduced maintenance turnaround by 6 months.

Vegetation Failure Forecasting & Inference

- Built probabilistic forecasting models using custom loss functions and Probabilistic Neural Networks to predict equipment failure across multiple electrical plants, averting over \$10M in operating costs and wildfire risk.
- Engineered data pipelines on AWS SageMaker and EC2 unifying weather, imagery and tabular data, with SQL-based monitoring for data quality and model performance.

Data Scientist OnPoint Solutions, New York City, USA

(Nov 2017 to Aug 2020)

Machine Learning Platform (Cortex)

- Built core ML algorithms from scratch including KNN, Random Forests, Neural Networks, Bayesian Networks and LSTM, and implemented Deep Auto-Encoding GMMs for unsupervised anomaly detection. Productionized the approach into Cortex.
- Architected automated ETL and monitoring pipelines for deployed models. Defined the customer and product metrics through A/B testing and statistical modeling used to scope subsequent ML projects.

Failure-Type Detection & Smart Combustion Control

- Combined 1D-CNN waveform and XGBoost models with DBSCAN clustering and autoencoder feature engineering to improve multi-class failure detection, and researched unsupervised failure-prediction models using custom-loss LSTM networks for early event detection.
- Led the data science solution for real-time NOx forecasting and control on Solex using SVR, XGBoost, Markov Decision Processes and Model Predictive Control, deployed at edge scale via PLCs and saving about \$1M in year-one costs.

EDUCATION

M.S., Electrical & Computer Engineering University of Illinois at Chicago

(2015 to 2017)

B.E., Electronics & Communication Engineering Visvesvaraya Technological University

(2011 to 2015)