Kartikeya Agarwal — ML Engineer 2

Bengaluru, India 560103
+91-9718502020 • kartikeya72001@gmail.com • kartikeya72001

Summary

Machine Learning Engineer with 3 years of experience building production ML systems at Navi Technologies. I've worked across the full ML lifecycle - from fine-tuning LLMs for SMS entity recognition to developing credit risk models that improved approval rates by 150bps. I enjoy optimizing systems and processes, having reduced model training costs by 31% and saved over $250,000 monthly through AWS resource optimization. Currently focused on MLOps and GenAI applications in financial services.

Experience

Navi Technologies **Bangalore**
Machine Learning Engineer *July 2023 – Current*

- **GenAi**
  - **Fine-tuning LLaMA and Mistral for SMS Entity Recognition**
    I proposed automating the process of SMS tagging with the help of parameter-efficient fine-tuning of Mistral 7B and LLaMA3 8B for SMS entity recognition. I overcame resource constraints with the help of Quantized Low-Rank Adaptation. I then trained BiLSTM-CRF using BIO Tagging to overcome the inherent problem of latency and costs with LLMs. This helped in reducing the dependency on manual tagging and reduced turnaround time by 36 hours.
  - **LLM-Based Underwriting user journey**
    Developed an end-to-end guided re-verification platform for rejected loan applicants, combining an in-house ITR & GST parser (IGNIS) with a multi-agent LLM underwriting engine to transform "dead-end" rejections into empathetic, step-by-step conversational journeys. Integrated a multi-lingual chatbot for personalized document guidance and orchestrated a dynamic 3–5 model ensemble, generating 150+ features and achieving a +36 bps approval rate improvement, INR 180 crores in yearly disbursal uplift, and INR 10–15 lakhs in monthly cost savings.
- **Credit-Underwriting**
  - **Risk and Income Model Development**
    I improved upon the previously deployed PrismV6 model by engineering new temporal features, replacing the simple XGBoost retraining pipeline with multi-modal pipeline. This resulted in a direct increase in the approval rate by 150bps. I also built a new income model using XGBoost and LightGBM, improving acc_within_10percent by 6 points.
  - **Model Retrain Optimization**
    I identified various bottlenecks in the current retraining pipeline and improved the system efficiency by optimizing various feature selection methods, data fetch pipelines, and hyperparameter tuning. These changes led to a 46% reduction in time and a 34% cost reduction in daily model retraining.
- **MLOps**
  - **Model Updation Service Development**
    I designed and developed a comprehensive MLOps platform for automated model deployment and management. Built a FastAPI-based service with PostgreSQL backend, MLflow integration for model versioning and artifact management, and automated pipeline switching between qa and prod environments. Implemented various endpoints, centralized logging with runtime configuration, and integrated external services (JIRA, Slack) for workflow notifications. The system completely removed manual deployment effort and improved model updation reliability.
  - **SPADE Real Time**
    Spade Real Time is a family of 3 services, architected to serve high-volume data at extremely low latency in production. The service can serve sms, app, device and location features at sub-20ms p99 latency at over 10,000rps. The service architecture leverages Reactive Kotlin, Postgres Database, EFS and S3 as data-sink and Kafka for data streaming. I also identified and removed various chokepoints in SMS Parser Jar and implemented Trie for efficient Template searching within the SPADE Jar to cut runtimes by upto 70%.

- **Automated Feature Store Development**
  I designed and built an entirely automated feature store complete with metrics and alerts while ensuring 99.99% feature consistency across prod and dev environments. The latest feature store is more than 50% faster and 2x more cost efficient than the legacy system. This was made possible by rearchitecting the pipeline and feature tables consumed by the entire Data Science, Risk, Analytics etc teams.
- **Process Improvements and Cost Optimization in Databricks and AWS**
  I designed and built an access control and cost attribution framework over Databricks. It was accompanied with an Databricks resource monitor to optimise compute usage. Using a regression framework on the compute usage pattern, I was able to recommend resource optimisations to minimize wasted compute and time. This resulted in over $250,000 in yearly savings while providing more optimal hardware usage.
- **Feature Selection Optimisation**
  I optimized key feature-selection jobs (RFE/IV/Boruta/Mutual Correlation), by using ray and tensors. This helped reduce the model training costs and time by over 80

Projects

Auto Code Sequence Generator *Oct 2022 – Apr 2023*

- I developed an Attention model to generate production-ready code from natural language prompts and wireframe diagrams. I created a custom token vector for wire-frame to code model to support multiple front-end frameworks. The model predicted generalised tokens which can be translated into different frameworks like React or Angular.

Visual Aid for Blind *Mar 2022 – Apr 2022*

- We built a small device, using an ESP32 module and arduino-nano. We setup a server which received raw image signals from the camera over wifi and used YOLOv3 and Transformer-based model for real-time field descriptions. We trained our model over Microsoft's COCO dataset.

AI Image Caption Bot *Mar 2021*

- Used LSTM and CNNs for making an Image Captioning Model with over 1.5 million parameters using the Flickr30k dataset.

Education

Netaji Subhas University of Technology **New Delhi, India**

Bachelor of Science in Computer Science *May 2023*

- CGPA: 8.72 (Graduated with Distinction).
- Recipient of the CVSPK Scholarship (2020): 100% Tuition Fee Waiver.

Research

Early Detection of Covid-19 using Machine Learning *May 2021 – July 2021*

- We utilized the ResNet50 model to extract features and distinguish COVID-19 from normal lung X-rays and pneumonia, achieving 99% accuracy. Our findings were published in the Indian Journal of Computer Science.

Classification of Skin Cancer Images using Convolutional Neural Networks *Apr 2021 – May 2021*

- We created a CNN based model to detect and classify skin lesions into Benign and Malignant. We were able to achieve an accuracy of over 86%. We used XceptionNet for image segmentation and feature extraction and created a small Dense Network for actual classification. (https://arxiv.org/abs/2202.00678)