Sarath Lella

☎+1 5714718234 / ✉ sarathlella.work@gmail.com / in sarath-lella / sarathlella / 🌐 sarathlella.portfolio

PROFESSIONAL SUMMARY

Results-driven AI/ML Engineer with 6+ years of experience in software development, data science, and machine learning. Adept at leveraging advanced ML and Generative AI techniques to deliver impactful business solutions, particularly in financial and telecommunications domains. Built a strong foundation in .NET and transitioned into AI and data science through hands-on experience with real-world healthcare and financial datasets. Demonstrated success at leading institutions like BNY Mellon and Charter Communications.

- 6+ years of professional experience developing scalable **Python** applications, leveraging key ML libraries such as **NumPy**, **Pandas**, **Scikit-learn**, **TensorFlow**, and **Keras** for end-to-end model development and deployment.
- Experienced in **prompt engineering** techniques (zero-shot, few-shot, instruction-tuned prompts) and evaluation strategies (BLEU, ROUGE, perplexity) to guide **LLM** behavior for production-grade use cases.
- Fine-tuned large language models (**LLMs**) using **PEFT/LoRA** techniques on domain-specific datasets such as **MedQuAD** and **PubMed** to enable context-aware QA and summarization.
- Architected and implemented **Generative AI** solutions to automate financial report summarization and client advisory assistance
- Designed and developed robust, scalable, and secure enterprise applications using **.NET Core / .NET 6+**, **C#**, and **ASP.NET MVC / Web API**, focusing on modular architecture, performance optimization, and maintainability across large codebases.
- Proficient in building ML models using **TensorFlow**, **PyTorch**, and scikit-learn with deployment in Dockerized environments - Specialized in **deployment, inference, tuning, and performance measurement** of ML models across production environments.
- Strong hands-on experience with **MLOps** tools like **MLflow**, **SageMaker**, and **DVC** for model lifecycle management
- Familiar with **Snowflake** data pipelines, including **Streams**, **Tasks**, **UDFs**, and the Cortex platform for deploying ML models at scale.
- Developed internal tools for ML explainability and model diagnostics using **SHAP**, **LIME**, and **ELI5**
- Built scalable and production-ready APIs using **FastAPI**, **Flask**, and **REST** principles to serve models in real time
- Solid foundation in cloud platforms (AWS, Azure) with experience deploying ML solutions using **SageMaker** and **Azure DevOps**
- Created insightful dashboards using **Power BI** and **Seaborn** to communicate model findings to stakeholders and leadership teams
- Built autonomous Agentic AI workflows using **LangChain** agents and tool integrations for multi-step reasoning, dynamic API calls, and decision-making automation.
- Strong statistical foundation with hands-on experience applying hypothesis testing, time series analysis, and probabilistic modeling across regulated financial and healthcare datasets.
- Designed and deployed **cloud-native ML applications** using AWS services like **SageMaker**, **Bedrock**, **S3**, **Lambda**, **AWS Batch**, **Step Functions**, **SNS**, **SQS**, **EventBridge**, and **CloudFormation** templates (CFT).
- Built and maintained infrastructure as code (IaC) using **AWS CloudFormation**, **Terraform**, and **OpenTofu** for scalable cloud resource management.
- Automated build, test, and deployment processes using **Jenkins** for continuous integration and delivery (CI/CD).
- Hands-on experience using **Git** for source control, branching strategies, and collaborative development in Agile teams.
- Expertise in infrastructure as code (IaC) services, including AWS CloudFormation and tools like Terraform or OpenTofu, for managing and provisioning cloud resources
- Built, tested, and deployed containerized applications using **Docker**, optimizing for highly scalable distributed systems based on open-source technologies.

TECHNICAL SKILLS

Category	**Skills**
Programming Languages	Python, C#, JavaScript, SQL
ML/DL Frameworks	TensorFlow, PyTorch, Keras, Scikit-learn
Generative AI	LLMs, Transformers, Diffusion Models, GPT-4, RAG Systems
Cloud & DevOps	AWS (SageMaker, EC2, S3, RDS), Azure DevOps, GitHub Actions, Docker
Tools & Libraries	Pandas, NumPy, OpenCV, HuggingFace, MLflow, Airflow, Visual Studio
Databases	PostgreSQL, SQL Server, MongoDB, Oracle RDS
Visualization	Power BI, Matplotlib, Seaborn
Web & API	Flask, FastAPI, RESTful APIs
Version Control	Git, GitHub, GitLab

PROFESSIONAL EXPERIENCE

BNY Mellon, New York NY
Sr. AI/ML Engineer
May 2023 – Present

- Designed and implemented a **Generative AI**-based intelligent document summarization system tailored for regulatory reports and investment memos by leveraging large language models (**LLMs**) and retrieval-augmented generation (**RAG**) pipelines, significantly reducing manual compliance review time and enabling financial analysts to gain faster insights, while adhering to the strict auditability, interpretability, and security requirements expected in the financial sector.
- Engineered multi-agent systems using **LangChain** to automate legal document processing, integrating **LLMs** with external tools for **RAG**-based search, summarization, and report drafting.
- Applied **prompt engineering** strategies for dynamic user intent recognition, retrieval reranking, and answer formatting in chatbot and search applications.
- Developed explainable AI models (XAI) using **SHAP** and **LIME** to meet regulatory compliance requirements for transparency in financial decision-making.
- Led development of **ML pipelines** for fraud detection and risk scoring, significantly improving model accuracy through ensemble learning
- Integrated **NLP** models and transformer-based architectures to classify, summarize, and extract key information from internal documents
- Conducted LLM fine-tuning and evaluation for GPT-based models to meet task-specific performance benchmarks, using datasets like MedQuAD and evaluating with BLEU and ROUGE.
- Implemented model versioning and rollback strategies to ensure production model stability and compliance with audit and regulatory guidelines (e.g., SOX, Basel, GDPR).
- Implemented an **LLM**-powered chatbot with RAG, enhancing internal response efficiency by over 60%
- Developed scalable and automated ML workflows using **Airflow**, **MLflow**, and **Docker**, reducing model retraining time by 40%
- Applied feature selection techniques and advanced hyperparameter tuning to improve prediction models across asset classes
- Built predictive models to optimize **loan default risk**, **policy lapse prediction**, **loss ratio modeling**, **mortality and morbidity risk**, and **customer segmentation**.

- Applied statistical analysis and ML models to optimize business operations and detect anomalies, fraud, and customer churn patterns
- Collaborated with cross-functional teams to align AI systems with business **KPIs**, compliance, and audit-readiness requirements
- Developed asynchronous inference APIs using FastAPI with Uvicorn and Gunicorn to support low-latency, high-availability deployments in financial environments.
- Monitored model drift and deployed continuous evaluation pipelines using **MLflow** and **SageMaker**
- Deployed secure and optimized REST APIs for ML microservices with **FastAPI** and OAuth2 authentication
- Experienced in scaling ML pipelines and real-time inference systems to handle large volumes of data and predictions.
- Developed and deployed ML services in **AWS**, leveraging **S3** for data storage, **SageMaker** for model training/inference, **Step Functions** for orchestration, **Lambda** for event-driven processes, and **Athena/Redshift** for big data analytics.
- Built and exposed **RESTful APIs** and **microservices** integrated with **Azure App Services**, **AWS API Gateway**, and **OAuth2-based authentication** (OpenID Connect), supporting millions of client requests with high availability and low latency.

Charter Communications, Stamford, CT
Data Analyst/System Developer
Aug 2022 – May 2023

- Developed and deployed machine learning pipelines that processed multi-terabyte-scale customer network usage logs in near real-time, applying time-series forecasting and anomaly detection models to proactively identify service degradation and preempt outages, resulting in a 25% improvement in network reliability metrics and a major reduction in customer churn, critical to maintaining service-level agreements in a high-availability telecom environment.
- Designed predictive models to forecast network load, detect anomalies, and optimize service uptime using time-series data
- Developed end-to-end data science solutions including data wrangling, model development, and deployment via **Azure ML**
- Built customer segmentation and churn prediction models using **XGBoost** and **Random Forests** integrated with Power BI for visualization
- Leveraged **SQL** and **Python**-based pipelines for data ingestion and preprocessing of streaming customer usage data
- Enabled A/B testing and uplift modeling to guide product development and network investment strategies
- Used **.NET** Core and **Python** to build microservices integrating ML outputs with front-end dashboards and internal tools
- Applied advanced techniques like **ensemble learning (Random Forests, XGBoost)** and **deep learning (DNNs, RNNs, CNNs for time series forecasting)** for improving model accuracy and robustness.
- Designed asynchronous Python APIs to serve time-series anomaly detection models, supporting concurrent requests with optimized inference latency.
- Created robust ETL workflows in Azure Data Factory and automated model refresh jobs.
- Contributed to internal research on generative modeling techniques for synthetic log generation and rare event simulation, driving innovation in predictive maintenance.
- Conducted feature importance analysis and implemented explainable AI techniques to promote stakeholder trust
- Delivered presentations to leadership explaining ML model outcomes, insights, and actionables with data storytelling
- Engineered real-time fraud detection systems using **Kafka Streams**, **AWS Kinesis**, and **serverless Lambda functions** to monitor transactions for anomalies.

Hansa Solutions, Hyderabad, India
.NET Developer & Data Analyst
Dec 2018 – Aug 2021

- Led the design of an **ML-powered** decision support system for a Insurance services client that integrated logs, diagnostic codes, and historical sales performance to generate dynamic pathway suggestions and cross-sell opportunities using classification algorithms, Bayesian optimization, and clustering techniques—bridging the gap between Insurance intelligence and business strategy in a data-scarce environment with heavy compliance constraints.
- Transitioned from traditional .**NET** development into AI/ML by building predictive and classification models on client sales and health data
- Designed ML systems adhering to financial sector regulations, ensuring **data privacy**, **model explainability**, **audit trails**, and **risk modeling guidelines** (e.g., OCC guidelines, IFRS 17, Solvency II).
- Containerized ML models and APIs using **Docker** and deployed them to **AWS ECS/Fargate**, **Azure Kubernetes Service (AKS)**, and **Google Kubernetes Engine (GKE)**.
- Led statistical modeling efforts for pricing strategy optimization, utilizing **A/B testing**, uplift modeling, and regression-based elasticity estimation in compliance-heavy environments.
- Deployed **ETL** pipelines that supported training of low-resource classification models fine-tuned on customer intent datasets using **HuggingFace** transformers.
- Developed **.NET MVC** web apps that incorporated embedded **ML** inference engines for real-time recommendations
- Designed anomaly detection systems to monitor transaction spikes and flag unusual business activity
- Used **Python** for building **ETL** scripts and pandas-based analytics on structured and semi-structured data
- Collaborated with product owners, actuaries, underwriters, risk officers, and compliance teams in Agile SCRUM environments to iteratively deliver impactful ML solutions in production.
- Created dashboards in **Power BI** to support operational and strategic decisions informed by **ML model** results
- Applied machine learning to improve client sales forecast accuracy using regression and time-series models
- Collaborated with data scientists to deploy recommendation engines and customer insights modules
- Integrated rule-based and **ML-driven** decision engines into existing enterprise reporting pipelines
- Conducted multivariate statistical analyses to support product development and pricing strategies
- Improved report delivery systems using **SSRS** and integrated visual **ML model** outputs
- Led model retraining and validation workflows for evolving customer behavior models

EDUCATION

Master of Science in Data Science
University of Missouri-Kansas City, 2022

Bachelor of Technology in Computer Science
GMR Institute of Technology, Rajam IN, 2019

PROJECTS & OPEN-SOURCE

- **Heart Disease Risk Predictor** – Deployed SHAP-explainable model with 81% ROC AUC
- **Early Diabetes Detection** – Wearable-style feature dataset with >90% accuracy
- **Breast Cancer Classifier** – Patch-level IDC detection using ensemble of CNNs (EfficientNet, ResNet, DenseNet) with test-time augmentation, explainability overlays (GradCAM), and >94% accuracy.
- **Medical Chatbot using LLMs –** Fine-tuned transformer models on MedQuAD and PubMedQA, applied prompt engineering with LangChain, and deployed with RAG for multi-turn QA over custom knowledge base.
- **Clinical NER** – Fine-tuned BioBERT for identifying medical entities in discharge summaries, integrating output into downstream pipelines for EHR processing.

References Available Upon Request