

Shreyas Mysore Sudesh

AI/ML Engineer

+1 (518) 558 3895 | shreyas.s@cvtechmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

SUMMARY

- AI/ML Engineer with 4+ years of experience across financial services, machine learning, predictive analytics, and production AI solutions, translating complex business requirements into scalable, data-driven applications.
- Hands-on expertise in Generative AI, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), agentic AI, LangChain, LangGraph, LlamaIndex, prompt engineering, and tool/function calling.
- Experienced in developing and evaluating machine learning models using Python, Scikit-learn, XGBoost, Pandas, NumPy, and PyTorch, including classification, regression, forecasting, feature engineering, and model validation.
- Skilled in building scalable data pipelines and distributed processing workflows using Apache Spark, PySpark, Databricks, Delta Lake, and Apache Kafka to support enterprise analytics and ML workloads.
- Proficient in deploying and operationalizing AI/ML applications using AWS, Amazon Bedrock, SageMaker, Docker, Kubernetes, MLflow, and CI/CD, with a focus on model serving, monitoring, scalability, and production reliability.
- Strong background in NLP, computer vision, and data visualization, with experience in Transformers, BERT, Hugging Face, spaCy, OpenCV, YOLO, Power BI, and Tableau.

SKILLS

Programming & Scripting: Python, SQL, Bash, Pandas, NumPy, SciPy

Generative AI & LLM Engineering: Generative AI, LLMs, RAG, Agentic AI, LLM Agents, LangChain, LangGraph, LlamaIndex, Prompt Engineering, Tool/Function Calling, OpenAI APIs, Embeddings, Vector Databases, Semantic Search, LLM Evaluation, MCP

Machine Learning: Supervised & Unsupervised Learning, Feature Engineering, Model Evaluation, Classification, Regression, Clustering, Time-Series Forecasting, Scikit-learn, XGBoost, LightGBM

Deep Learning & NLP: PyTorch, TensorFlow, Keras, Transformers, Hugging Face, BERT, GPT Models, NLP, NER, Text Classification, Sentence Transformers, spaCy

Computer Vision: OpenCV, YOLO, Image Classification, Object Detection, OCR, Image Segmentation, Transfer Learning

Data Engineering & Distributed Computing: Apache Spark, PySpark, Apache Kafka, Databricks, ETL/ELT, Data Modeling, Delta Lake, Parquet, Data Lakes, Data Warehousing

Databases & Storage: PostgreSQL, MySQL, MongoDB, Redis, Snowflake

MLOps & Model Deployment: MLflow, Kubeflow, Docker, Kubernetes, CI/CD, GitHub Actions, Jenkins, Airflow, Model Registry, Model Monitoring, Feature Stores

Cloud & AI Platforms: AWS, SageMaker, Bedrock, Lambda, ECR, EC2, S3, Azure Machine Learning, Azure AI, Google Cloud, Vertex AI

Model Serving & Optimization: ONNX, ONNX Runtime, Triton Inference Server, vLLM, Model Serving, Inference Optimization, Model Quantization, A/B Testing, Hyperparameter Tuning, Optuna

Monitoring & Responsible AI: AI Observability, Model Performance Monitoring, Data/Model Drift, Model Explainability, Bias Detection, Responsible AI, Logging

Visualization & Engineering: Power BI, Tableau, Plotly, Matplotlib, Git, GitHub, GitLab, Agile, Scrum, SDLC, Code Reviews, Unit Testing, Integration Testing

EXPERIENCE

Morgan Stanley

Jan 2026 - Present

AI/ML Engineer

New York

- Developed an enterprise RAG platform for financial research and document intelligence using Python, LangChain, LangGraph, OpenAI APIs, embeddings, and vector search, reducing analyst information retrieval time by 45%.
- Engineered agentic AI workflows for financial document analysis and research automation using LLM tool calling and MCP, reducing manual investigation effort by 35%.
- Implemented LLM evaluation pipelines, semantic retrieval validation, and AI observability to improve response consistency and answer accuracy by 28% across prioritized use cases.
- Deployed scalable AI services using Docker, Kubernetes, AWS Bedrock, Lambda, Amazon ECR, and CI/CD pipelines, reducing deployment cycles from days to hours.
- Collaborated with product, data, and business teams to translate financial research and operational requirements into production-ready RAG and agentic AI applications.
- Established reusable engineering standards for prompt engineering, model evaluation, monitoring, and responsible AI, reducing duplicated development effort by 25% across AI initiatives.

HCLTech

Mar 2023 - Aug 2024

AI/ML Engineer

India

- Developed machine learning solutions for financial data analysis, transaction classification, and business forecasting using Python, Scikit-learn, XGBoost, and feature engineering, improving model performance by 22%.
- Built scalable ETL/ELT pipelines using PySpark, Apache Spark, Databricks, and Delta Lake to process financial and customer datasets, reducing data preparation time by 50%.
- Implemented classification, regression, and clustering models to identify customer behavior patterns, support financial insights, and improve decision-support accuracy by 20%.

- Optimized distributed data processing and Spark transformations for large datasets, reducing pipeline execution time by 38% and accelerating downstream analytics and model refreshes.
- Developed reusable ML pipelines for feature engineering, model training, validation, and experiment tracking, reducing development effort by 30% across analytics projects.
- Built predictive analytics and time-series forecasting solutions using Python, Pandas, NumPy, and XGBoost to support business planning and reduce forecast error by 18%.

Zensar Technologies

Apr 2021 - Feb 2023

Data Scientist

India

- Developed predictive modeling solutions using Python, Pandas, NumPy, Scikit-learn, and XGBoost to support business forecasting and classification, improving prediction performance by 18%.
- Applied supervised and unsupervised learning techniques, including classification, regression, and clustering, to identify customer segments, behavioral trends, and operational patterns.
- Built NLP pipelines for text classification, entity extraction, and document processing using BERT, Hugging Face Transformers, spaCy, and Sentence Transformers, reducing manual review effort by 40%.
- Developed computer vision solutions using OpenCV, YOLO, and OCR for image classification and automated document or image analysis, reducing manual inspection time by 30%.
- Created interactive dashboards and analytical visualizations using Power BI, Tableau, Matplotlib, and Plotly to communicate model results and business trends, reducing reporting turnaround by 25%.
- Implemented model validation and performance monitoring workflows to compare algorithms, improve prediction reliability, and reduce production model rework by 20%.

PROJECTS

Enterprise Financial Research Assistant | Generative AI & RAG

- Built a financial document intelligence platform using Python, LangChain, LangGraph, OpenAI APIs, embeddings, and vector databases to retrieve relevant information from financial documents and generate context-grounded responses.
- Implemented agentic workflows, MCP-based tool integration, retrieval evaluation, and LLM observability to support multi-step financial research and improve response reliability.

Financial Data Intelligence & Predictive Modeling | Machine Learning

- Developed predictive models for transaction classification, customer analytics, and financial forecasting using Python, Scikit-learn, XGBoost, and feature engineering to support data-driven business decisions.
- Engineered scalable data pipelines with PySpark, Databricks, and Delta Lake, integrating data transformation, model validation, and experiment tracking into reusable ML workflows.

Intelligent Document Processing & Text Analytics | NLP and Computer Vision

- Developed NLP pipelines for document classification and entity extraction using BERT, Hugging Face Transformers, and spaCy to automate the analysis of unstructured business documents.
- Integrated OCR and computer vision techniques using OpenCV and YOLO with Python-based analytics workflows to streamline document and image processing and improve information extraction.

EDUCATION

Master of Science in Business Analytics

University at Albany

May 2026

SUNY, USA