Pavan Pratapagiri | Data Scientist

Texas | +1 (469) 892- 8858 | saipavanp2504@gmail.com| LinkedIn

SUMMARY

Data Scientist with 5+ years of experience building and operationalizing machine learning solutions for fraud prevention, risk assessment, and financial forecasting across insurance and capital markets. Specialized in Agentic AI frameworks, Retrieval-Augmented Generation (RAG), and deep learning using TensorFlow and PyTorch. Delivered measurable business impact, including 25% uplift in predictive accuracy and 40% reduction in fraud losses through cloud-based ML architectures on AWS and GCP (BigQuery, GCS). Strong expertise in distributed data processing (Spark, Hadoop), real-time decision systems, explainable AI (SHAP), and full ML lifecycle governance.

TECHNICAL SKILLS

Programming & Distributed Data Processing: Python (Pandas, NumPy, SciPy), Advanced SQL, PySpark, Apache Spark, Hadoop (HDFS, MapReduce), ETL Development, Data Modeling

Machine Learning & Statistical Modeling: Scikit-learn, XGBoost, TensorFlow, PyTorch, ARIMA, LSTM, Time Series Forecasting, Anomaly Detection, Risk Modeling, Fraud Detection, Feature Engineering, Hyperparameter Tuning, Cross-Validation, RMSE, MAE, Precision-Recall, ROC-AUC

Generative AI & Agentic Systems: Agentic AI Architectures, Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Transformer Models, Prompt Engineering, Embeddings, Vector Retrieval, Sentiment Analysis

Cloud & Big Data Platforms: AWS (SageMaker, S3, EC2), Google Cloud Platform (GCP), BigQuery, Google Cloud Storage (GCS), Distributed Data Processing, Scalable Model Deployment

MLOps & Model Governance: MLflow, Model Deployment, Model Monitoring, Drift Detection, SHAP (Explainable AI), CI/CD Pipelines, Docker, Kubernetes, REST APIs

Streaming & Real-Time Systems: Kafka, Event-Driven Architectures, Real-Time Inference Pipelines

EXPERIENCE

Assurant, Texas, USA **Jul 2025 – Present**
Data Scientist

- Directed end-to-end development of XGBoost and TensorFlow models for vehicle claim risk scoring and fraud analytics, achieving 25% accuracy gains and decreasing false positives by 30% based on ROC-AUC and RMSE evaluation.
- Built event-driven risk assessment pipelines using PySpark, Kafka, and AWS SageMaker, shortening claim adjudication cycles by 35% across multi-state portfolios.
- Developed distributed ETL architectures with Spark and Hadoop to handle high-volume mobility and claims datasets spanning 50+ insurance programs.
- Advanced premium pricing methodologies through ARIMA and LSTM-based forecasting techniques, strengthening long-term revenue planning and actuarial insights.
- Implemented transformer-driven NLP and RAG frameworks to extract contextual intelligence from unstructured policy documents and customer communications.
- Integrated SHAP-based interpretability mechanisms and compliance controls to support audit readiness and transparent decision-making.
- Deployed containerized inference services with Docker and REST APIs while instituting automated drift detection and performance tracking workflows.

Morgan Stanley, Texas, USA **Jun 2024 – Jun 2025**
Data Scientist

- Constructed predictive frameworks for market risk evaluation and trading anomaly identification using XGBoost and TensorFlow, delivering 20% forecasting improvement and lowering unauthorized trading exposure by 40%.
- Designed streaming anomaly detection systems with PySpark and Kafka to supervise high-frequency transaction patterns and strengthen regulatory safeguards.
- Performed comparative modeling across ARIMA, LSTM, and gradient boosting techniques to anticipate forex and commodity price fluctuations for hedging optimization.
- Engineered high-throughput data processing workflows utilizing Spark and advanced SQL to accelerate quantitative research pipelines.
- Applied transformer-based sentiment modeling on earnings releases and financial news to refine short-term trading signals by 15%.
- Partnered with quantitative analysts and trading desks to integrate ML-driven insights into proprietary execution platforms, boosting operational throughput by 35%.

University of Missouri, Kansas City, USA **May 2023 – May 2024**
IS Technical Student Assistant

- Created and validated machine learning prototypes that elevated research model effectiveness by 30% across academic datasets.

- Produced SQL- and PostgreSQL-driven dashboards delivering analytical visibility to 20+ faculty and administrative stakeholders.
- Streamlined preprocessing pipelines with Spark and Pandas, decreasing preparation time for downstream modeling by 40%.
- Implemented forecasting and anomaly detection experiments within AWS SageMaker for research data validation.
- Developed Flask-based APIs enabling near real-time visualization and analytical reporting.

Hexaware Technologies, India **Aug 2020 – Dec 2022**
Machine Learning Engineer

- Built distributed ML workflows processing structured and unstructured healthcare data, increasing throughput by 40%.
- Developed predictive healthcare models using XGBoost and PyTorch, lowering diagnostic error rates by 25% and reaching 92% precision.
- Orchestrated feature engineering and SQL-based integration across heterogeneous clinical systems.
- Delivered secure inference services via Dockerized Flask applications deployed on AWS SageMaker, ensuring HIPAA-compliant data handling.
- Enhanced generalization performance through systematic validation strategies and hyperparameter optimization, cutting iteration cycles by 20%.

EDUCATION

Master of Science in Computer Science **May 2024**
University of Missouri, MO

Bachelor of Science in Computer Science **May 2021**
Koneru Lakshmaiah University, India

PROJECTS

Topic Discovery & Trend Analysis on New York Times Articles (NLP, LDA)
Technologies: Python, NLTK, SpaCy, Gensim, WordCloud, pyLDAvis

- Processed and normalized 50K+ news articles using advanced NLP preprocessing (lemmatization, stop-word filtering, token normalization) to improve topic coherence and semantic consistency.
- Implemented Latent Dirichlet Allocation (LDA) to identify 10 dominant editorial themes and evaluated model quality using topic coherence scores.
- Delivered interactive visualizations via pyLDAvis to enable exploratory trend analysis and interpretable topic insights.

Real-Time Crowd Density Estimation via Semantic Segmentation
Technologies: TensorFlow, YOLOv5 / Faster R-CNN, Flask, OpenCV

- Designed and deployed a dual-stage deep learning pipeline combining object detection and semantic segmentation for accurate crowd density estimation in high-traffic environments.
- Evaluated model performance using MAE, MSE, and F1-score, achieving high precision in dense scene detection scenarios.
- Operationalized inference through a Flask-based REST API to enable real-time video stream processing and scalable deployment.

CERTIFICATION

Microsoft Certified: Azure AI Engineer Associate
Stanford (Coursera): Machine Learning