PAVAN YELLATHAKOTA

pavan.yellathakota.ds@gmail.com | **+1 (929) 278-4589** | **Seattle, WA** | Linkedin : yellatp | github.com/yellatp

PROFESSIONAL SUMMARY

Data Analyst turned **ML Engineer** with a Master's in Data Science and 3+ years of experience building end-to-end data pipelines and ML models. Currently a **Founding Engineer** architecting retrieval systems and LLM-driven backends, specializing in transitioning R&D research into production-grade infrastructure using **FastAPI, pgvector, and Vertex AI.**

EDUCATION

M.S., Applied Data Science | **Clarkson University**, Potsdam, NY, *USA* | *Aug 2023 – May 2025*

B.Tech., Computer Science | **Yogi Vemana University**, Proddatur, India | *Aug 2016 – Dec 2020*

TECHNICAL SKILLS

Programming: **Python** (Pandas, NumPy, Scikit-learn, XGBoost, **PyTorch**, **CausalML**, Scipy), **SQL**, R.
Gen AI & Vector DB: Gemini (**Vertex AI**), DeepSeek, LangChain, **MCP**, **pgvector**, ANN Search.
ML & Data Science: A/B Testing, **Time-Series** Forecasting, Statistical Modeling, Statsmodels.
Data Engineering: **PySpark**, AWS Glue, Airflow, Schema Normalization, **ETL/ELT**.
Backend & MLOps: **FastAPI**, **Docker**, DigitalOcean VPS, **SageMaker**, MLflow, AWS Lambda.
Cloud Analytics & Visualization: **AWS** (S3, Redshift, Athena), Tableau, Looker Studio, Matplotlib.

PROFESSIONAL EXPERIENCE

Founding ML Engineer | **Alphonso AI (Early-Stage)**
backed by Shipley Center for Innovation, Clarkson University | Remote, USA *Jul 2025 – Present*

- **Designed a 0→1 Backend Ecosystem** using **FastAPI and PostgreSQL**, deploying a scalable microservices architecture on **DigitalOcean VPS** to optimize infrastructure costs while bridging Java-based core services with Python ML workloads.
- **Engineered a Multi-Model "Text-to-Query" (TTQ) Engine** leveraging **Gemini (Vertex AI) and DeepSeek APIs** to enable dynamic, prompt-driven semantic search across high-dimensional talent data.
- **Developed a Domain-Aware Recommender System** using **Vectorized Embeddings**; optimized ranking logic to prioritize industry-specific sector expertise over generic roles, improving candidate-to-company fit.
- **Optimized Search Performance** by deploying a multi-stage retrieval pipeline: utilizing **pgvector for Approximate Nearest Neighbor (ANN)** search and **CUDA-accelerated Cross-Encoders** for high-precision re-ranking

Graduate Quantitative Researcher | **SMIF, Clarkson University** | Potsdam, NY *Sep 2024 – Apr 2025*

- Automated SEC EDGAR data extraction and built a **BERT**-based **NLP sentiment pipeline** for Reddit and earnings call transcripts, reducing manual **data processing time** by ~**70**%.
- Developed Monte Carlo simulations to stress-test portfolio allocations; the team's fund delivered a **51% return** on a **$650K** portfolio, **outperforming the S&P 500 benchmark** during the period.

Data Science Consultant | **HAVK Mladost (Elite Athletics Club)** | Remote *Oct 2023 – May 2025*

- Architected a **centralized data lake** on **AWS S3**, migrating legacy athletic performance records to a cloud-queryable environment, reducing average data retrieval latency by **30**%.
- Built **PySpark** ETL jobs on **AWS Glue** to process athletic performance event data; applied **partition pruning** to reduce query costs and improve processing speed.
- Applied **uplift modeling** to segment the fan base and identify high-propensity audience segments for merchandise marketing campaigns.

Business Data Analyst | **eAppSys Limited** | India *Jul 2022 – Dec 2022*

- Developed Prophet and **SARIMAX** demand-forecast models for 1,500+ SKUs incorporating exogenous variables (promotions, seasonality), improving **forecast accuracy** (**MAPE**) by **15**%.
- Automated procurement **KPI** reporting via **Oracle Analytic**s, saving 12 manual hours per week and **streamlining stakeholder decision-making**.

Data Analyst | **Kantar GDC India** | India *Sep 2021 – May 2022*

- Built Python/**PySpark** pipelines to integrate 10M+ survey records from 30+ data sources, reducing processing latency by 30% for global research projects.
- Developed **regression models** and **statistical significance-testing frameworks** (pandas, statsmodels) to validate data representativeness across FMCG and Telecom markets.
- Engineered **Power BI** dashboards and **semantic analysis** tools to convert unstructured survey data into actionable **go-to-market insights for stakeholders**.