SAI PRATHAP REDDY CHELURI

Seattle, WA | +1 (857)-214-9758 | saiprathap.1061@gmail.com | linkedin.com/in/sai-prathap-reddy | sai-prathap-reddy-cheluri.github.io

SUMMARY

Software Development Engineer with 7+ years building large-scale distributed systems and data pipelines. MS in AI/ML with hands-on experience applying machine learning to production systems. Transitioning engineering expertise in scalable data systems into Data Science and ML Engineering roles.

TECHNICAL SKILLS

- **Programming & Data:** Python, SQL, Java, Pandas, NumPy
- **Machine Learning & AI:** PyTorch, scikit-learn, XGBoost, LLMs (RAG, CLIP, Prompt Engineering, API Integration)
- **Cloud & Infra:** AWS (S3, Lambda, EC2, SQS, SNS, DynamoDB, CDK, IAM, CloudFormation), Git, CI/CD
- **Databases:** PostgreSQL, DynamoDB, Redis
- **Data Science & Analytics:** Supervised & Unsupervised Learning, Feature Engineering, Hyperparameter Tuning, Model Evaluation, A/B Testing, Statistical Significance, Data Visualization (Matplotlib, Seaborn, Tableau)
- **Prototyping & Tools:** Streamlit, Gradio, Jupyter

PROJECTS

Multimodal Movie Recommender | GitHub
- Built a movie recommender using dense embeddings + BM25 + FAISS reranking, improved NDCG@10 by ~12% over baseline.
- Implemented personalization and language-aware ranking in a Streamlit demo, supported by a reproducible CSV, Parquet data pipeline with CI validation.

Home Credit Default Risk | GitHub
- Built an end-to-end ML pipeline on Home Credit dataset (~7M rows) with data cleaning, feature engineering, and model training.
- Engineered 200+ features and tuned XGBoost model (ROC-AUC ≈ 0.78), outperforming logistic regression baseline (~0.72) and highlighting key predictors such as external scores and credit history.

AI Recipe Generator | GitHub
- Developed an LLM-powered recipe assistant (OpenAI + Gemini) with Streamlit UI, generating recipes from user-provided ingredients and preferences.
- Applied prompt engineering with multi-model fallback and caching to achieve 90%+ recipe coverage on test prompts and ~40% latency reduction, with export options (PDF/CSV) for usability.

WORK HISTORY

Software Development Engineer | **AWS (Amazon Web Services)** – Seattle, WA 07/2022 - Current
- Engineered tabular data extraction and indexing for Amazon QBusiness, enabling structured querying; direct S3 retrieval reduced query latency by 25% and expanded data coverage.
- Implemented a multi-criteria ranking algorithm in Kendra, improving top-10 ranking accuracy by ~12% and reducing manual filtering for enterprise users.
- Engineered file-upload support to QBusiness chat, improving context ingestion and boosting answer accuracy by ~15%.
- Increased ingestion reliability by 30% in QBusiness by adaptive throttling and race-condition handling, reducing tickets by 20%.

Application Development Analyst | **Accenture Solutions Pvt Ltd** – Bengaluru, India 03/2017 - 12/2020
- Engineered the upgrade of OpenText systems from v10 to Content Server v16.4 and Archive Server v16.2, integrating them with SAP XECM to enable a petabyte-scale repository for 10,000+ users.
- Engineered infrastructure optimizations for OpenText Content Server, resulting in a 30% increase in application performance and enabling the system to process over 2 million documents daily through OTIC, DMS, and Document Pipeline services.

EDUCATION

Master of Professional Studies: Artificial Intelligence & Machine Learning 06/2022
Northeastern University - Boston, MA | GPA 3.95

Bachelor of Technology: Computer Science 07/2016
VTU - Bengaluru, India.