Prahlad Siwakoti

Data Scientist/Engineer

Harrisburg, PA

✉ prasiwakoti@gmail.com ☎ (225) 210-0199 🌐 Portfolio in LinkedIn

Summary

I am a passionate Data Scientist/Engineer with a background in physics and mathematics. I have built statistical and machine learning models to derive insights from complex datasets using Python, R, SQL, and other open-source tools. I have leveraged modern data engineering tools to automate and scale data pipeline tasks and deployed scalable solutions with FastAPI. I am looking forward to integrate my educational experience, and programming skills to solve complex problems.

Skills

Languages: Python, R, SQL, SAS
Visualization: Power BI, ggplot, Matplotlib, R-Shiny, Streamlit, Seaborn
Machine Learning: Scikit-learn, Tensorflow, Pytorch, Statsmodels, spaCy, NLTK
Database: Database Design, PostgreSQL, MongoDB, DuckDB, Neo4j
Engineering: Docker, SFTP, Data Ingestion, Migration, SQL Scripting, dbt, Airflow, FastAPI, Cloud Functions (Azure, GCP), Cloud Storage (Azure Blob, MINIO), Snowflake, Databricks, PySpark
Cloud Platforms: AWS
Additional: Debugging, Time-Series Analysis, Deep Learning, Version Control (Git/GitHub), CI/CD (GitHub Actions)

Education

- **Masters in Data Science**, University of Texas at Austin 01/2024–05/2025
  Courses: Machine Learning, Probability and Statistical Inference, Data Visualization, Algorithms, Advanced Predictive Models, Deep Learning, NLP, Reinforced Learning, Data Science for Health.
- **PhD in Physics**, Louisiana State University, Baton Rouge, LA 08/2015–11/2021
  Dissertation: Effects of Structure, Crystallographic Orientation, and Dimensionality on Emergent Properties of Transition Metal Oxide Thin Films

Data Experience

- **Data Engineer Assistant Instructor** (Part Time), Nashville Software School 10/2025–Present

– Assist in teaching data engineering concepts to students
– Provide guidance on projects involving modern data stack tools
– Support curriculum development and delivery
– Help students troubleshoot technical issues

- **Data Engineer Apprenticeship**, Nashville Software School 05/2025–09/2025

– Automated extraction, transformation, and loading (ETL) of structured and semi-structured data from REST APIs and local storage sources
– Designed and implemented data pipelines using Airflow, DBT, and Snowflake, following the medallion architecture for scalable and reliable data processing
– Explored streaming and event-driven architectures using pub/sub, Kafka, and webhooks to handle real-time data ingestion and processing
– Developed a modern data architecture using Dremio, DBT, iceberg, and Nessie to optimize data storage and retrieval
– Containerized data applications using devcontainers and Docker for consistent deployment across environments

- Implemented data quality checks and monitoring to ensure data integrity and reliability throughout the pipeline using github actions
- Built a FastAPI-based frontend with integrated Swagger documentation for interactive API exploration and testing

- **Data Scientist Apprenticeship**, Nashville Software School 09/2023–07/2024

- Wrangled data and performed exploratory data analysis using Python's pandas library and R's tidyverse packages
- Created data visualizations using matplotlib, seaborn, and ggplot2 to understand complex datasets and problems
- Built and evaluated statistical and machine learning models using scikit-learn and statsmodels
- Developed and evaluated machine learning models for classification and clustering tasks, interpreting confusion matrices, ROC curves, and precision-recall metrics
- Applied natural language processing using nltk and spaCy to enhance text analysis capabilities
- Performed network analysis on graph data using Neo4j to identify key relationships and patterns
- Built and deployed interactive data visualizations using R Shiny
- Managed source code version control with Git/GitHub, ensuring code integrity and facilitating team collaboration

Professional Experience

- **Researcher**, University of Tennessee at Knoxville 11/2021–12/2023

- Developed and maintained a data analysis pipeline for large-scale synchrotron data using Python and R
- Wrote Python scripts to simulate observed data and perform statistical analysis
- Collaborated with researchers from various disciplines to analyze, interpret data, and deduce conclusions
- Provided mentorship and training to graduate students in research, instrumentation, and troubleshooting

- **Research Assistant**, Louisiana State University, Baton Rouge 01/2018–11/2021

- Explored non-trivial physics of transition metal oxide perovskite thin films with respect to symmetry and growth orientation; studied various two-dimensional defects

Selected Projects

- **DE Capstone: International Trade Insights for United States** (GitHub)
  Developed a data engineering solution to analyze and visualize US international trade data, integrating multiple APIs (U.S. Census Bureau, WITS) and automating ingestion, transformation, and analytics workflows. Built a dashboard for stakeholders and a RESTful API for flexible data access. Utilized containerized architecture (MinIO, Dremio, dbt, Prefect) for scalable, reproducible deployment.
  Skills: Data Ingestion (APIs), ETL, Prefect Orchestration, MinIO, Iceberg Tables, dbt, Dremio, FastAPI, Streamlit, Data Visualization, Containerization (devcontainer, Docker)

- **Bank Classifier MLflow** (GitHub)
  Demonstrated a complete machine learning workflow for bank marketing classification using FastAPI, MLflow, PyTorch, and Docker. Implemented data ingestion and preprocessing, accelerated feature engineering with DuckLake, and trained a neural network classifier with hyperparameter optimization. Tracked experiments and model artifacts using MLflow, and served predictions via a REST API supporting both raw and processed features. Containerized the workflow for reproducibility and easy deployment.
  Skills: ML Pipeline (PyTorch, Optuna), FastAPI, MLflow, DuckLake, Docker, Feature Engineering, REST API, Experiment Tracking

- **Marvan Research Data Pipeline** (GitHub)
  Built an automated data pipeline for the Marvan research project using Airflow, dbt, and Snowflake to streamline data ingestion, transformation, and analysis. Developed a FastAPI data access layer with Swagger documentation for easy querying and interaction with research datasets. Implemented medallion architecture for scalable and maintainable data workflows.

Skills: Data Pipeline Automation (Airflow, dbt), Snowflake, Medallion Architecture, FastAPI (Swagger Docs)

- **Air Quality Prediction** (GitHub)
  Constructed a predictive model for air-quality monitoring from data obtained from inexpensive PurpleAir sensors and meteorological sources. Utilized tree-based spatio-temporal models and neural networks to predict air quality.
  Skills: Time-Series Analysis, Spatial Regression, Kriging Interpolation, Machine Learning, Deep Learning (PyTorch), Data Visualization
- **Other Projects:** Portfolio

Certifications

- **Data Engineering Certificate**, Nashville Software School 09/2025
- **Data Science Certificate**, Nashville Software School 07/2024
- **Databricks Fundamentals**, Databricks 08/2025
- **Data Lakehouse Basics**, Dremio 08/2025

Peer Reviewed Publications

Link to Google Scholar