

AMEER SOHAIL SHAIK

+1 (813)-817-1935 skameersohail148@gmail.com LinkedIn GitHub Portfolio

Education

University of Maryland – College Park

Master of Science in Data Science

Aug 2025 – Present

College Park, MD

Vellore Institute of Technology, AP

Bachelor of Technology in Computer Science and Engineering

2020 – 2024

AP, India

Experience

Machine Learning Engineer

August 2024 – August 2025

Procedit – (Remote)

Barcelona, Spain

- Identified that **lagging sector-ETF returns outperformed 8 other candidate features** (rolling volatility, RSI, MACD) in predicting next-week price direction across 200+ tickers; communicated feature importance rankings to the trading signals team, directly informing their shift from an ARIMA baseline to a supervised learning approach.
- Designed an **offline backtesting evaluation framework** comparing 10 model architectures on directional accuracy, Sharpe ratio, and max drawdown, surfaced that gradient-boosted models matched deep learning precision (65%) at 12x lower inference cost, preventing over-engineering of a production LSTM pipeline.
- Cut data preparation workflows by **92%** (60+ min to 3 min per sync) by building an automated ETL pipeline integrating PostgreSQL with Google Sheets API; wrote **50+ SQL queries** (joins, window functions, CTEs) to reconcile daily OHLCV data across 4 source tables, freeing 8+ analyst-hours per week.

Data Science Intern

January 2024 – May 2024

SBP Consulting Pvt Ltd

Hyderabad, India

- Analyzed 12 client accounts' SAP S/4HANA records and identified **recency-weighted purchase frequency and seasonal spend ratios** as top revenue predictors, engineered these into an XGBoost model achieving 87% accuracy (MAPE: 13%), with results directly shaping quarterly C-suite planning recommendations.
- Designed **5 interactive Power BI dashboards** mapping sales trends, churn risk segments, and 7+ KPIs to executive questions; replaced a 2-day manual Excel reporting cycle with 4-hour automated refresh, enabling stakeholder self-service and reducing ad-hoc data requests by ~60%.
- Reduced data preparation time by **45%** by building Python scripts (pandas, NumPy) that consolidated 4 SAP modules into analysis-ready formats via standardized cleaning logic (null imputation, duplicate detection, schema validation), eliminating recurring quality errors across 3 downstream pipelines.

Projects

StreamTest: A/B Testing & Causal Inference on Streaming Data | Python, SciPy, Statsmodels

- Framed the product question — “Does genre diversification improve user satisfaction?” — and operationalized it on 162K MovieLens users (25M ratings), running a **6-test hypothesis suite** (Welch's t, Mann-Whitney, chi-square, ANOVA, bootstrap CIs); designed sample-size simulations showing 8,200 users/group needed at 80% power on heavy-tailed data.
- Estimated the **causal effect of genre diversity on ratings (+0.12 stars, 95% CI: [0.06, 0.18])** via propensity score matching controlling for 3 confounders; validated with Bayesian A/B testing (98.3% P(B>A)), translating findings into a clear product recommendation.

SparkFlow: Scalable ETL & Feature Store Pipeline | PySpark, SQL, Apache Airflow

- Processed **2M+ clickstream events and 500K+ transactions** through a 4-stage PySpark pipeline in under 3 minutes; wrote sessionization via SQL window functions, built 35 aggregated features, and orchestrated data quality checks (null rates, freshness, drift detection) via an Airflow DAG with SLA alerting.

Dynamic SQL Assistant | Python, LangChain, Groq API, SQLite

- Built a **Text-to-SQL engine** enabling natural language database queries with 90%+ semantic accuracy across 50+ tested patterns; architected prompt orchestration with dynamic schema injection and SQL validation to guarantee syntactic correctness.

Technical Skills

Languages & Databases: Python, SQL (window functions, CTEs, correlated subqueries), Java, PostgreSQL, MySQL, SQLite, MongoDB

Data Science & Analytics: Pandas, NumPy, Scikit-learn, XGBoost, SciPy, Statsmodels, A/B Testing & Experiment Design, Causal Inference (PSM, DiD), Feature Engineering, EDA

Deep Learning & GenAI: TensorFlow, PyTorch, LangChain, Llama, RAG, SHAP

Visualization & BI: Power BI, Tableau, Matplotlib, Seaborn, Plotly

Tools & Cloud: Flask, GraphQL, Git, Docker, Apache Airflow, PySpark, Streamlit, AWS (S3, Redshift, QuickSight, Lambda)