Sparsh Marwah (Open to relocate)

marwah.sp@northeastern.edu | +1 (857) 225-9142 | LinkedIn | GitHub | Portfolio

Summary

Data Science Analyst with 3+ years of experience in Python, SQL, machine learning, and data visualization. Proven ability to clean, model, and interpret large datasets, build deployable ML models, and generate actionable insights. Passionate about leveraging statistical learning and real-time data pipelines to solve real-world problems

Work Experience

Graduate Teaching Assistant, Northeastern University, Boston, MA **Sep 2024 – May 2025**

- Guided 40+ students in advanced data mining and model-building processes using Python and SQL on large transactional datasets
- Developed comprehensive data frameworks and predictive modeling exercises focused on extracting actionable business insights
- Created performance testing methodologies for data-driven solutions and established metrics for model evaluation and optimization

Data Science Analyst, Tredence Analytics Solutions Pvt. Ltd., Bengaluru, India **Jun 2021 - Jul 2023**

- Developed **predictive models (Random Forest, XGBoost)** on 10M+ transactional and customer records to identify fraud and optimize retention, improving retention rates by 18%
- Built scalable **ETL data ingestion** and transformation processes using PySpark and AWS Glue, ensuring data accuracy and consistency across marketing and financial datasets
- Applied **SHAP** explainability and feature engineering to enhance model interpretability and align predictive signals with credit behavior, supporting business strategy
- Designed and evaluated **A/B tests** and statistical analyses to inform marketing and product decisions, improving model personalization precision by 15%
- Developed interactive **Tableau dashboards** to communicate fraud detection KPIs, spending trends, and retention metrics to clients and senior stakeholders
- Partnered with engineering and product teams to productionize predictive models on **AWS** and monitor performance, driving continuous model improvements

Data Science Intern, SJVN Ltd., Shimla, India **Jun 2019 - Aug 2019**

- Developed **SQL** scripts to optimize inventory data processes, reducing stockouts by 20% and improving demand forecasting accuracy by 25% for supply chain operations
- Applied **regression techniques** to forecast inventory demand, enabling procurement and supporting cross-functional teams

Education

Northeastern University, Boston, MA **Sep 2023 - May 2025**
Master of Science in Data Analytics Engineering, GPA: 3.8/4.0
Coursework: Data Management in Analytics, Operations Research, Machine Learning Operations, Applied Gen-AI

SRM Institute of Science and Technology, Chennai, India **Jul 2017 - May 2021**
Bachelor of Technology in Computer Science Engineering
Publication: AI Music Generator. (2022). Journal of Pharmaceutical Negative Results, 67-71 (Research paper)

Technical Skills

Programming Languages/Tools: Python, R, SQL, Git
Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, SHAP
ML/Stats: Regression, Classification, Tree-based Models, Cross-validation, Model Evaluation
Data Engineering & Visualization: PySpark, AWS (S3, Glue, Redshift), APIs, Tableau, Power BI, Streamlit
Deployment & DevOps: Docker, Jenkins (familiar), GitHub Actions, MLflow

Projects

Medical Non-Adherence Prediction System **Apr 2025 – May 2025**

- Built a predictive model on UCI diabetes data to classify insulin non-adherence with 78% precision
- Deployed scoring pipeline on **AWS SageMaker** and Lambda for real-time risk notifications and automated alerts
- Engineered end-to-end cloud pipeline for continuous scoring and model monitoring on **AWS EventBridge**

AI-Powered Resume & Job Matching Assistant – "404: Job Not Found" (View Project) **Feb 2025 – May 2025**

- Developed a **RAG** architecture with **LangChain** and **MongoDB** to recommend resumes with 90%+ accuracy
- Integrated **ChromaDB** vector search to quickly retrieve candidate profiles and match them with job postings

Equipment Price Prediction using Tree-Based Model **Sep 2024 – Dec 2024**

- Utilized **XGBoost and Random Forest** to predict heavy machinery prices from 50K+ listings, optimizing for RMSE and interpretability.
- Engineered features like usage hours, depreciation rate, and seasonal factors to improve model accuracy.
- Built an interactive **Streamlit dashboard** to simulate pricing, supporting better customer and dealer decisions.

Churn Prediction (View Project) **Apr 2024 – May 2024**

- Developed **XGBoost and Random Forest models** to predict churn on telecom datasets, achieving 92% accuracy and 0.91 AUC
- Preprocessed 5,000+ records by handling missing values, encoding categorical features, and engineering medically relevant predictors from 19 clinical attributes to enhance model interpretability