Sparsh Marwah

Boston, MA 02130 | marwah.sp@northeastern.edu | +1 (857) 225-9142 | LinkedIn | GitHub | Portfolio

Education

Northeastern University, Boston, MA **May 2025**
Master of Science in Data Analytics Engineering, GPA: 3.75/4.0
Relevant Coursework: Data Management in Analytics, Data Mining in Engineering, Machine Learning Operations

SRM Institute of Science and Technology, Chennai, India **May 2021**
Bachelor of Technology in Computer Science Engineering
Relevant Coursework: Data Structures, Data Science and Big Data Analysis, Object Oriented Analysis and Design
Publication: AI Music Generator (Research paper)

Technical Skills

Programming & Databases: Advanced SQL (PostgreSQL, Redshift, Hive), Python, R, Java, NoSQL (MongoDB)
Statistical Analysis & Modeling: Hypothesis Testing, Predictive Analytics, A/B Testing, Statistical Measurement
Data Analytics & Visualization tools: Tableau, Power BI, Excel, storytelling through data, KPI optimization
Big Data & Tools: Hadoop, PySpark, Databricks, Alteryx, MLflow, Docker
Cloud & Infrastructure: AWS (S3, Glue, Timestream, SageMaker, Bedrock, Athena, lambda), Google Cloud Platform
Libraries & Frameworks: Pandas, NumPy, Scikit-learn, XGBoost, Matplotlib, Seaborn
Certifications: Python (Programming, Data Structures), Data Science & AI, Intro to Cloud Data Analytics, ETL in Python and SQL

Work Experience

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

- Developed and maintained large-scale e-commerce data pipelines for a top U.S. retail client using **Python** to enable robust data integration in alignment with scalable data science solutions
- Conducted **A/B testing** with cross-functional teams to analyze product adoption trends and user retention, showcasing insights through **Tableau** dashboards that informed growth strategies and increased customer retention by 20%
- Executed comprehensive **data preprocessing**, **exploratory data analysis (EDA)**, **feature engineering, and feature selection**, followed by predictive modeling using **ML algorithms** such as **Linear Regression & XGBoost**, achieving an accuracy of 91.2%
- Visualized feature contributions to model predictions using **SHAP** feature importance graphs, providing insights into the features that most impacted the model's predictions and enhancing interpretability, which increased overall accuracy by 10%
- Fine-tuned models using **k-fold cross-validation** to ensure robustness and optimal performance and reduce over-fitting

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

- Designed and implemented a cloud-native **MLOps** pipeline on Google Cloud for real-time air quality data ingestion and preprocessing, ensuring high data consistency and reliability.
- Automated model retraining workflows using CI/CD pipelines integrated with **GitHub** Actions and **MLflow**, enabling continuous accuracy monitoring and timely updates for predictive models.
- Developed **REST API** integrations for end-user accessibility and real-time updates, enhancing usability and ensuring seamless delivery of air quality insights to stakeholders.

Project Experience

Air Quality Prediction **Sep 2024 – Dec 2024**

- Developed a cloud-native MLOps pipeline using **Google cloud platform** for real-time data ingestion and preprocessing
- Automated model retraining with **CI/CD pipeline integration** using **GitHub actions** & **MLflow** and data pipeline using **Airflow**, ensuring consistent accuracy across deployments and consistent flow of preprocessed data
- Integrated the pipeline with **REST APIs** for end-user access and real-time updates, enhancing product usability
- Leveraged MLflow for model tracking, drift detection, and version control on GitHub, automating drift detection to flag accuracy deviations and enable timely retraining, maintaining performance standards across deployments using **Cloud functions**

Liver Cirrhosis Survival Prediction (Link) **Apr 2024 – Jun 2024**

- Performed **exploratory data analysis (EDA)** to identify trends and patterns, guiding feature engineering and model development
- Predicted liver cirrhosis survival using **Random Forest**, achieving 81.9% accuracy by creating features like Symptom & Risk Score

Face Mask Detection (Link**)** **Feb 2024 – Apr 2024**

- Developed a face mask detection system using a **CNN-based classifier** and **SSD** for real-time face detection, ensuring high-speed, accurate performance
- Integrated the solution into real-time surveillance, employing semantic segmentation for pixel-level precision and optimizing for real-time video processing, enhancing its use in public safety and health monitoring

Cricket Auction Player Performance Tracking System (Link**)** **Sep 2023 – Dec 2023**

- Collected, cleaned, and optimized player stats, linking performance data with team outcomes for key relationship analysis
- Developed SQL queries and used **Python** with **Matplotlib** and **Seaborn** for visualizations for performance insights like batting average, top 10 batsmen, top 10 bowlers & bowling average. Utilized **NoSQL** database for unstructured data in the dataset