# 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
**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**
- Constructed scalable data pipelines using **Python** and **Databricks** for a major U.S. retail client, integrating data for robust analytics over e-commerce data
- Automated workflows by developing machine learning models, including Linear Regression and XGBoost, achieving 91.2% accuracy
- Conducted A/B testing and presented results using Tableau dashboards, improving customer retention by 20%
- Fine-tuned models using k-fold cross-validation and employed SHAP for feature importance analysis to enhance interpretability
- Implemented end-to-end solutions integrating ML algorithms to automate decision-making and business processes

**Data Integration Intern**, SJVN Ltd., Shimla, India **Jun 2019 - Aug 2019**
- Gathered information about different energy forms, analyzed their powerhouse tools inventory data by developing **SQL** queries to understand the stock levels and sales trends
- Developed data integration workflows documentation to determine entire lifecycle of the project, ensuring seamless dataflow
- Performed data quality audits and troubleshooting to ensure data accuracy, integrity, consistency, contributing to improved decision making and operational efficiency

## Project Experience

**Air Quality Prediction** (Link) **Sep 2024 – Dec 2024**
- Developed a cloud-native MLOps pipeline using **AWS** 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 AWS **lambda**

**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