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, Financial Management for Engineers

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: Python (Pandas, Scikit-learn, Matplotlib, Seaborn, Plotly), SQL (Hive, Redshift, PostgreSQL), MySQL
Big Data & Tools: HDFS, PySpark, Databricks, GitHub, Docker, Airflow
Machine Learning: Neural Networks, Gradient Boosted Machines, Supervised/Unsupervised Modeling Techniques
Visualization Tools: Tableau, Power BI
Cloud Platforms: Google Cloud Platform, AWS (S3, EC2, Lambda)
Other Skills: Statistical Analysis, A/B Testing, Data Storytelling, KPI Measurement
Certifications: Python (Programming, Data Structures), Data Science & AI, Intro to Cloud Data Analytics, ETL in Python and SQL

Work Experience

Teaching Assistant, Northeastern University **Sep 2024 – Dec 2024**

- Instructed students in **Python**, database management and data analysis, offering tailored guidance towards data visualization
- Directed labs and workshops on **Tableau**/storytelling with data; resolved problems for students to deliver 20+ projects

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

- Gathered information about their 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 decide the 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 (View Project) **Sep 2024 – Dec 2024**

- Developed **machine learning** models to predict PM2.5 and PM10 levels using OpenAQ API
- Applied advanced **data preprocessing, feature engineering,** & **model selection** techniques to create a reliable prediction system
- Designed and implemented a comprehensive **MLOps** pipeline using **Airflow**, automating data ingestion, model retraining, and deployment of new data seamlessly through **Google Cloud Platform**
- 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

IMDb Movie Data Analysis & Visualization (View Project**)** **Jan 2024 – May 2024**

- Executed comprehensive **data preprocessing, exploratory data analysis (EDA)**, & time series analysis on IMDb movie data, followed by predictive modeling using ML algorithms **Linear Regression and XGBoost**, achieving an accuracy of 91.2%
- Developed an interactive **PowerBI** dashboard with dynamic visuals, slicers, and filters, enabling users to explore and analyze movie data, providing stakeholders with personalized insights and actionable results

Cricket Auction Player Performance Tracking System (View Project**)** **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