

Sparsh Marwah

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

Work Experience

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

- Developed classification and regression models, including **XGBoost**, **Linear Regression**, and **Random Forest**, using **Python** to predict Customer Lifetime Value (CLTV) and churn, to improve customer retention, for a top U.S. retail client
- Engineered and deployed Databricks-based **MLOps** pipelines for automated model retraining, monitoring, and version control using **MLflow** and **CI/CD workflows**, reducing model deployment time by 40% and enhancing performance tracking by 30%
- Applied data preparation using **PySpark** on an e-commerce dataset of 10M+ customers, followed by **feature engineering**, **EDA** and **model development** for predictive analysis, achieving a 91.2% accuracy in customer behavior forecasting
- Executed **A/B and multivariate testing** for consumer behavior analysis across three product categories (Alcohol, Food & Beverages, and Cigarettes), leveraging Python and **SQL** to analyze customer segmentation, increasing conversion rates by 15%
- Designed interactive **Tableau** dashboards to visualize financial and customer insights, enabling data-driven pricing strategies that boosted conversion rates by 25% and improved revenue generation
- Collaborated with engineering teams to optimize model deployment for scalability and efficiency in real-time applications, while working with product teams to align model outputs with business needs and enhance user experience

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

- Collected and analyzed energy inventory data by developing complex **SQL** queries, optimizing stock level assessments and identifying sales trends, leading to a 20% reduction in stockouts and a 15% improvement in inventory turnover
- Drafted data integration workflows documentation to decide the entire lifecycle of the project, ensuring seamless dataflow
- Conducted data quality audits and troubleshooting, ensuring accuracy, integrity, and consistency across datasets, reducing data errors by 30% and enhancing decision-making efficiency by 25%

Technical Skills

Programming & Frameworks: Python (Scikit-learn, TensorFlow, PyTorch, LLaMA.cpp), SQL (PostgreSQL, MySQL), NoSQL

AI/ML & Modeling: Generative AI, Large Language Models, Time Series Forecasting, PCA, Random Forest, XGBoost, MLE

Cloud-Native & MLOps: Kubernetes, OpenShift, Kubeflow, Docker, MLflow, CI/CD, GitHub Actions, Databricks, Airflow

DevOps & Cloud: REST APIs, Cloud Functions, Git, AWS(S3, EC2, Lambda), Google Cloud Platform, YAML

Visualization & Storytelling: Tableau, Power BI, Microsoft Excel (VBA, Macros), NLG for Automated Reporting

Certifications: Python ([Programming](#), [Data Structures](#)), [Data Science & AI](#), [Intro to Cloud Data Analytics](#), [ETL in Python and SQL](#)

Academic Projects

AI-Powered Assistant for Data Analysis & Visualization Feb 2025 – March 2025

- Designed and implemented an AI-driven system that utilizes LLM APIs (**OpenAI**, **Gemini**) to automate data analysis, uncover insights, and generate dynamic visualizations
- Developed a multi-agent framework to process large datasets, perform **exploratory data analysis** (EDA), and generate interactive dashboards for trend identification
- Automated data storytelling and reporting using **natural language generation** (NLG), enabling seamless interpretation of complex data for decision-making

Air Quality Prediction ([View Project](#)) Sep 2024 – Dec 2024

- Designed and deployed a cloud-native MLOps pipeline on **GCP** for real-time time-series data ingestion and preprocessing through **ETL processes**, ensuring efficient data handling and scalability
- Optimized model deployment by integrating **GitHub Actions** & MLflow for automated retraining, reducing deployment time by 50% and maintaining consistent accuracy
- Implemented **REST API** integration to provide real-time air quality predictions, enhancing end-user accessibility and usability
- Automated model monitoring using **MLflow** and **Cloud Functions** for drift detection, ensuring proactive retraining and sustained model performance

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 **Tableau** dashboard with dynamic visuals, slicers, and filters, enabling users to explore and analyze movie data, providing stakeholders with personalized insights and actionable results

Education

Northeastern University, Boston, MA Sep 2023 - 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, Applied Gen-AI

Relevant Experience: Teaching Assistant for **Computational and Visualization for Analytics**

SRM Institute of Science and Technology, Chennai, India Jul 2017 - 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](#))