Reuben Chatterjee

LinkedIn Github San Diego, CA reuben.a.chatterjee@gmail.com +1 (858) 220-2422

EDUCATION

University of California San Diego — San Diego, USA
Master of Science in Data Science (GPA: 3.82/4.00) — Sep. 2023 – Present
Courses: Statistical Models, Deep Learning, Scalable Data Systems, Fraud Analytics, Causal Inference

University of Mumbai — Mumbai, India
Bachelor of Engineering in Computer Engineering (GPA: 3.93/4.00) — Aug. 2019 – May 2023
Courses: Database Management & Warehousing, Machine Learning, NLP, Big Data Analytics

EXPERIENCE

Data Science Research Assistant — Jan. 2025 – Present
Gender Based study on COGS108, UC San Diego — San Diego, CA

- Executed advanced statistical analysis on 500+ student survey responses using R, applying ANOVA and Wilcoxon Rank Sum tests to identify significant gender-based skill disparities with p-value < 0.01
- Developed end-to-end data pipeline for demographic analysis using tidyverse (dplyr, tidyr) to process 4,000+ survey responses, reducing processing time by 40% while maintaining experimental integrity
- Designed PostgreSQL database schema for longitudinal survey data, enabling complex time-series analysis across multiple student cohorts, and used ggplot to translate findings into actionable insights

Lead Teaching Assistant - Data Science in Practice — Oct. 2024 – Present
Cognitive Science Department, UC San Diego — San Diego, CA

- Led team of 16 TAs for 840-student course; implemented structured grading workflow using Jupyter and Nbgrader resulting in 30% faster assignment turnaround and 95% student satisfaction
- Automated grading pipeline by developing multiple scripts using Python (Pandas, NumPy) and Bash , reducing runtime by 20% and standardizing evaluation across 4,200+ assignments per quarter

Data Scientist Intern — Jul. 2024 – Sep. 2024
Datamatics Global Services — San Diego, CA

- Built ETL pipelines using Python (PySpark) and SQL to process 50GB+ data for ML model training on the RAKEZ - UAE government SEZ project, maintaining MySQL databases with optimized query writing
- Developed ML models (XGBoost, LightGBM) achieving 92% accuracy in demand forecasting, utilizing feature engineering techniques including encoding and feature scaling

PROJECTS

Fraud Detection for Credit Card Transactions | *Python, SciKit-learn, XGBoost* — Apr. 2024 – Jun. 2024

- Built fraud detection model using RandomForest and XGBoost with SMOTE for imbalanced data handling, achieving 92% accuracy while reducing false positives by 10%
- Processed 97,852 transactions, engineering 3,200 features through time-based aggregations and statistical transformations, improving model precision by 15%

Climate Change Analysis | *Python, Time-Series Forecasting, Data Visualization* — Sep. 2023 – Dec 2023

- Developed a predictive model using NorESM2, processing 250 years of climate data across 150 countries, implementing time-series forecasting techniques with ARIMA and Prophet models, while achieving 90% accuracy
- Engineered a sophisticated data preprocessing pipeline that interpolated complex CO2 emissions datasets, transforming raw monthly data into cumulative carbon emissions metrics with 100% data completeness

TECHNICAL SKILLS

Programming Languages: Python, SQL (Advanced), R (tidyverse, ggplot2), PL/SQL, C++, Bash
Data Science & ML: TensorFlow, PyTorch, scikit-learn, Supervised/Unsupervised learning, Clustering, Decision Trees, Neural Networks, Feature Engineering, XGBoost, AWS, Azure, A/B Testing, SMOTE, Cross-validation
Databases & Big Data: MySQL, PostgreSQL, Spark, Databricks, AWS, Hadoop, Data Warehousing
Data Visualization: Tableau (Advanced), PowerBI, Matplotlib, Seaborn, Interactive Dashboards, Data Storytelling