Reuben Chatterjee

reuben.a.chatterjee@gmail.com | Linkedin - Reuben Chatterjee | github.com/ReubenChatterjee | Portfolio

Education

- **University of California, San Diego**
  Master of Science in Data Science *Sep 2023 - Jun 2025*
  Courses: Statistical Models, Scalable Data Systems, Machine Learning, Deep Learning
- **St. Francis Institute of Technology**
  Bachelor of Engineering in Computer Engineering *Aug 2019 - May 2023*
  Courses: Data Structures, Big Data Analytics, DBMS, AI & ML, NLP

Technical Skills

- **Programming Languages:** Python, R, SQL, C++ (Object-Oriented Design), Javascript
- **Tools and Frameworks:** MySQL, PySpark, Tableau, Git, Docker, AWS, Bash, REST API, CI/CD
- **Machine Learning:** Pytorch, Tensorflow, Keras, MLFlow, Scikit-Learn, XGBoost, Scipy, Hugging Face, NLTK, TF-IDF

Professional Experience

- **Ellis Lab, UC San Diego** *Jan 2025 - Present*
  Graduate Research Assistant - Data Science
  - Developing a **data-cleaning and preprocessing pipeline** for 1,200+ peer evaluation records; standardizing over 5,000 free-text responses to enable downstream analysis of team dynamics and collaboration equity.
  - **Conducting demographic analysis on 1000+ students** enrolled in COGS108 by aggregating and visualizing pre-course survey data across gender, major, and prior experience, uncovering equity gaps and informing course improvements for future cohorts.
  - Designing a **classification pipeline** in R to analyze 1,800+ open-ended student responses and identify contribution types across 80+ teams; leveraging **regex** to reveal behavioral patterns correlated with team composition and gender diversity.
- **Cognitive Science Department, UC San Diego** *Sep 2024 - Present*
  Lead Graduate Teaching Assistant - COGS108
  - Leading a team of 16 TAs for a 800+ student course for the 3rd consecutive quarter, holding 2 discussion sections of 60+ students each and dedicated office hours for students.
  - Scripting **automated scalable grading pipelines** for grading, feedback generation and release, and grade posting in Python using NBGrader with Canvas API, eliminating need for manual grading and **increasing the grading process speed by 85%**.
  - **Mentored 20+ project groups** each quarter for the comprehensive Data Science Project, providing technical guidance from data collection through model deployment with **95% projects receiving distinction**.
- **Datamatics Global Services** *Jun 2024 - Sep 2024*
  Data Scientist Intern
  - Architected and implemented **ETL pipelines using Python and SQL** to process data on the 'RAKEZ - UAE government SEZ' project, enabling seamless integration of multiple data sources for ML model training.
  - Developed production-ready machine learning models **achieving 92% accuracy** in demand forecasting, implementing **A/B testing frameworks** to validate model performance improvements.
- **Halicioglu Data Science Institute** *Sep 2023 - Jun 2024*
  Data Analyst (Part time)
  - Built **interactive Tableau dashboards** to visualize social media metrics, leading to a **30% increase in online engagement** by optimizing content strategy and posting schedules.
  - **Analyzed platform-specific metrics** across Twitter, Instagram, and Facebook, generating recommendations that improved user retention and conversion rates by 15%.

Projects

- **Credit Card Fraud Detection using Gradient Boosting**
  Python, SciKit-learn, Hugging Face, CNN, XGBoost, LightGBM, Random Forest
  - Engineered 3,200+ behavioral features from 97,852 credit card transactions using domain-specific encodings and behavioral signals.
  - Tuned LightGBM via multi-model comparison (RF, XGBoost, CatBoost); achieved 92% accuracy and 0.59 OOT AUC.
  - Reduced false positives by 10% via threshold tuning and SMOTE, contributing to $2M+ projected annual savings.
- **Document Summarization with LSI & BERTSUM**
  Python, SVD, TF-IDF, NLTK
  - Built a scalable pipeline to preprocess 9,000+ CNN news articles; applied TF-IDF, tokenization, and vectorization.
  - Compared SVD-based LSI and transformer-based BERTSUM models, improving ROUGE-L F1 score from 0.24 to 0.42.
  - Optimized document-term matrix generation for efficient downstream summarization.
- **Student Accommodation App**
  Flutter, Node.js, MySQL, Python, K-Means
  - Built a cross-platform app to match roommates based on personality and lifestyle compatibility using the OCEAN model.
  - Applied K-Means clustering on survey data to generate behavioral clusters; integrated results via a RESTful backend.
  - Achieved 85% match satisfaction in user trials; aligned compatibility through ML-driven profiling and real-time matching.