Reuben Chatterjee

San Diego, CA | reuben.a.chatterjee@gmail.com | Linkedin - Reuben Chatterjee | github.com/ReubenChatterjee

Professional Summary

Data Scientist with a Master's in Data Science from UC San Diego, specializing in building scalable ML and analytics pipelines that drive measurable business outcomes. Expertise in data warehousing (Snowflake, Fivetran), supervised and unsupervised learning, customer segmentation modeling, churn prediction, and fraud detection. Skilled in feature engineering, statistical analysis, and time-series forecasting using Python, R, SQL, and cloud platforms (AWS, GCP). Proven track record of delivering $2M+ business impact and 60% cost reductions in both academia and industry settings

Experience

Alcamo Marketing — *Jul 2025 - Oct 2025*

Data Scientist

- **Reduced operational costs by 60% ($20K annually)** by migrating data infrastructure from Adverity to Snowflake, building Snowflake tables and views with 100+ field mappings and transformations for 4 clients.
- Enabled **real-time reporting with 99% uptime** for 4 clients by **engineering 15 automated ETL pipelines** (Fivetran to Snowflake to Looker Studio) processing 50GB+ daily data from Google Ads, LinkedIn Ads, Meta, Salesforce, and HubSpot
- Engineered **ML-based customer segmentation** with clustering algorithms **achieving 90% accuracy scores**, analyzing **behavioral patterns across 100K+ customer records** to optimize engagement strategies and prioritize high-value opportunities

Ellis Lab, UC San Diego — *Sep 2024 - Jun 2025*

Graduate Research Assistant - Data Science

- Developed **automated data cleaning and validation pipeline** in R using Regex and NLP to standardize 5,000+ free-text responses, **reducing manual processing time by 90%**
- **Created longitudinal panel dataset tracking 1,000+ students** over time, and conducted demographic analysis across gender, major, and experience and visualized equity gaps in course outcomes using ggplot2 to inform course structure redesigns
- Applied **ANOVA, Tukey HSD** tests and correlation analysis to examine gender based differences male vs female dominant groups across team contributions, leadership roles, and project outcomes

Datamatics Global Services — *Jun 2024 - Sep 2024*

Data Scientist Intern

- Built **Python and SQL based ETL pipelines** to process 50GB+ of economic zone data for RAKEZ, enabling demand forecasting of industrial land lease activity
- Trained a Random Forest model to optimize zone planning decisions, achieving 92% accuracy and validating results via A/B testing

Halicioglu Data Science Institute — *Dec 2023 - Jun 2024*

Data Analyst

- **Increased online engagement by 30%** by building interactive Tableau dashboards to visualize social media metrics
- **Analyzed platform-specific metrics** across Twitter, Instagram, and Facebook, generating recommendations that improved user retention and conversion rates by 15%

Projects

Afterpay Customer Retention Prediction Model

Python, XGBoost, LightGBM, Scikit-learn, Pandas, SHAP, Snowflake, SQL

- Built **gradient boosting churn prediction model** achieving **89% accuracy and 0.92 ROC-AUC** by engineering **50+ behavioral features** including RFM metrics, payment patterns, and temporal trends from 350K+ transaction records.
- Identified **$2.1M in at-risk revenue** by **scoring 12,500 high-risk customers** through 4-tier risk segmentation system, enabling targeted retention campaigns that **improved customer lifetime value by 18%**.

Credit Card Anomaly Detection using Gradient Boosting

Python, SciKit-learn, Hugging Face, CNN, XGBoost, LightGBM, Random Forest, Seaborn

- 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), achieving 92% accuracy and 0.59 OOT AUC
- Reduced false positives by 10% via threshold tuning and SMOTE, contributing to $2M+ projected annual savings

Education

University of California, San Diego

Master of Science in Data Science (CGPA:3.82) — *Sep 2023 - Jun 2025*

Courses: Statistical Models, Scalable Data Systems, Fraud & Pricing Analytics, Deep Learning

University of Mumbai

Bachelor of Engineering in Computer Engineering (CGPA:3.93) — *Aug 2019 - May 2023*

Courses: Data Structures, Big Data Analytics, DBMS, AI & Machine Learning, NLP

Technical Skills

Programming: SQL, Python (Pandas, NumPy, Scikit-learn, PySpark), R, C++, JavaScript

Data Science & Machine Learning: Data Cleaning, EDA, Classification, Regression, Clustering, Feature Engineering, Statistical Modeling, Model tuning & deployment, ETL pipelines, A/B Testing, Customer Segmentation

Big Data & Cloud Platforms: Snowflake, Databricks, AWS, Azure, Hadoop, dbt, Ray, Docker, Kubernetes, Fivetran

Visualization & Communication: Tableau, Looker Studio, Power BI, Excel, ggplot2, Shiny, D3.js, Seaborn