

Shrestha Bhattacharya

San Jose, CA | iamsb98@gmail.com | +1 (408) 464-3379 | linkedin.com/in/sbhattacharya98 | shrestha57.github.io

SUMMARY

Data engineer with 3+ years of experience building, optimizing, and supporting SQL data pipelines and validating data quality for brokerage and insurance claims data at Cognizant. Pursuing an M.S. in Analytics at Georgia Tech, specializing in Computational Data Analytics (GPA 3.83/4.00). Builds independent projects in causal inference and experimentation — marketplace switchback testing, variance reduction and sequential testing, and advertising incrementality and uplift modeling — with experience tracing data issues to their source and communicating findings to technical and business partners.

TECHNICAL SKILLS

- **Data Engineering:** Data Pipelines, ETL, Data Modeling (Slowly Changing Dimensions Type 1 & 2), Data Validation, SQL Optimization, Informatica PowerCenter, AQT, Autosys
- **Languages & Libraries:** Python (pandas, NumPy, scikit-learn, Matplotlib, NLTK), SQL (PostgreSQL)
- **Statistics & Experimentation:** Experimental Design, A/B Testing, Sequential Testing, Variance Reduction (CUPED), Causal Inference, Uplift Modeling, Hypothesis Testing, Regression Modeling, Forecasting, Predictive Modeling
- **AI Tools:** Claude for coding, debugging, and data exploration

DATA SCIENCE PROJECTS

Testing a Ride-Hail Promotion When Riders Compete for Drivers: Switchback Experiments on 19.7M NYC Trips

- Diagnosed a critical flaw in marketplace A/B testing: simulating a ride-hail market on 19.7M real NYC trips (validated against Erlang's exact loss formula) proved standard user-level tests overstate a promotion's true causal impact by 3.3x (+15.0% reported vs. +4.6% actual), because riders compete for the same drivers.
- Designed a switchback (time-window) experimental framework — window length, carryover washout, and demand-adjusted regression — that cut measurement error 11x and held false-positive rates at the 5% target (vs. 15% for standard methods), delivering a production-ready fix for marketplace experimentation bias.

A/B Tests With 32% Fewer Users: Variance Reduction & Safe Early Stopping

- Reduced the sample size required for A/B tests by up to 22% by engineering and validating variance-reduction techniques (CUPED, CUPAC) on a real 200K-customer experiment, directly cutting experimentation cost and time-to-decision.
- Eliminated false positives from early stopping ("peeking"), where naive monitoring reached a 38% error rate, by engineering statistically valid sequential-testing boundaries — cutting the required sample by 32% while raising detection power from 79.5% to 86.2%.

Advertising Incrementality & Uplift Modeling

- Corrected a 60% overstatement in ad-effectiveness measurement on a 14M-user ad campaign by applying causal inference (instrumental variables) to isolate the true incremental lift (+3.2pp), enabling more accurate ad-spend decisions.
- Built machine learning uplift models that increased targeting efficiency 4x over random allocation, capturing 82% of incremental conversions by targeting only the top 20% of users (95% CI 78.5% to 84.8%).

Experiment Doctor: Automated A/B Test Diagnostics | Independent Open-Source Project

- Built an open-source Python diagnostic tool that automatically audits A/B test results for five common failure modes — sample ratio mismatch, peeking-driven false positives, marketplace interference, Simpson's-paradox mix-shift, and missed CUPED variance-reduction opportunities — before a team ships a decision.
- Packaged the statistical safeguards from my switchback and variance-reduction projects into a reusable Streamlit app with documented methodology, so other teams can run the same checks on their own experiments.

Variance-Reduced Streaming PCA with Parameter-Free Momentum | Independent Research, submitted to NeurIPS 2026

- Solved a key limitation in streaming PCA algorithms by deriving a closed-form, parameter-free tuning method, eliminating manual hyperparameter search for real-time, high-dimensional data streams.
- Reduced subspace estimation error by 22-27% over the standard algorithm (Oja's), validated on 40+ years of real S&P 500 financial data and synthetic non-stationary data streams.

YouTube Trending Content Analytics | Georgia Tech Course Project

- Identified the key drivers of video performance by building predictive models on 56,640 YouTube videos, improving prediction accuracy 19% over baseline (RMSE 1.41 vs. 1.75).
- Quantified the limits of virality prediction through residual analysis, showing where post-publication engagement data improves forecasts (RMSE 1.57 to 1.47) and where it does not.

PROFESSIONAL EXPERIENCE

Cognizant Technology Solutions, India | Associate | July 2021 - November 2024

- Reduced data pipeline runtime by 87.5% by filtering records earlier and removing unnecessary joins from SQL transformations.
- Built and supported SQL pipelines for brokerage and insurance claims data, using SCD Types 1 and 2 to maintain current and historical records for downstream reporting.
- Improved data reliability by tracing recurring production failures to client source-file errors and redesigning a problematic load around a weekly refreshed intermediate table; partnered with technical and client teams to translate business requirements into production transformations, supporting testing, deployment, and post-release validation under time-sensitive requirements.

EDUCATION

Georgia Institute of Technology | M.S. in Analytics, Computational Data Analytics

August 2025 - August 2027 (Expected) | GPA: 3.83/4.00

University of Engineering & Management, Kolkata, India

B.Tech in Computer Science & Engineering | August 2017 - May 2021 | CGPA: 8.58/10.00