Shafayet Khan Shafee

sshafee@isrt.ac.bd kshafayet.netlify.app shafayetShafee shafayetshafee Google Scholar
+880-1791-051104 Dhaka, Bangladesh

Data Scientist specializing in experimentation, causal inference, and machine learning for data-driven decision-making, with experience building analytics systems and ML solutions in fintech. Experienced in owning the full analytical lifecycle, from framing ambiguous business problems to delivering measurable outcomes through experimentation, causal analysis, Bayesian methods, and predictive modeling, including shipping ML and data pipelines to production. Interested in applying rigorous statistical thinking to large-scale, product-led environments, particularly at the intersection of causal inference and Bayesian methods.

Professional Experience

Data Scientist Dhaka, Bangladesh
Pathao Pay Jan 2025 – Present

- Led the design and execution of 5 end-to-end randomized experiments across fintech product and marketing initiatives, spanning both user and merchant populations. Translated experiment findings into product and growth decisions by identifying actionable interventions, detecting ineffective interventions early, and preventing unnecessary promotional spend.
- Evaluated the causal impact of a fintech product rollout on merchant business outcomes in a non-randomized setting using quasi-experimental methods, including propensity score matching, inverse probability weighting, and Difference-in-Differences. Generated causal evidence to support rollout decisions when randomized experimentation was not feasible.
- Improved controlled rollout planning by applying Bayesian experimental design to determine sample sizes using probabilistic criteria, making tradeoffs between statistical confidence, operational cost, and experiment efficiency explicit and communicable to stakeholders.
- Evaluated product adoption across small and unequal-sized cohorts using hierarchical Bayesian modeling, enabling more reliable estimates under limited data by quantifying uncertainty and supporting risk-aware rollout decisions.
- Architected and maintained a centralized analytics transformation layer using dbt and BigQuery on GCP, converting fragmented raw data sources into structured analytical models. Built reusable production datasets that simplified downstream analysis, reduced typical query scans from GB-scale to MB-scale, lowered BigQuery costs, and standardized metrics across Growth, Product, Operations, and Finance analytics teams.

Data Scientist Dhaka, Bangladesh
Pathao Limited Jul 2023 – Dec 2024

- Designed and owned 2 end-to-end randomized experiments to optimize BNPL repayment engagement, testing notification strategies across thousands of users. Demonstrated the impact of increased repayment communications in reducing post-due unpaid user rate, contributing to a 4% reduction after broader rollout without negatively impacting monthly active users.
- Developed a causal ML-based personalization framework using Double Machine Learning and causal forests to estimate heterogeneous treatment effects across user subgroups. Enabled targeted promotion strategies by identifying high-response segments, contributing to a 35% reduction in monthly promotional spend while maintaining comparable user activity and business performance.
- Co-developed a repayment collection prioritization framework using Random Survival Forests to estimate time-to-repayment and stratify overdue BNPL users by repayment likelihood. Optimized limited CX call capacity by prioritizing users with higher likelihood of repayment following outreach while reducing unnecessary collection efforts.
- Co-developed a credit risk scoring framework for BNPL applicants using XGBoost to estimate probability of default. Enabled risk-based user segmentation and dynamic credit limit assignment, allowing the business to expand BNPL access while managing credit risk exposure.
- Built automated reporting workflows using dbt, BigQuery, Looker Studio, and Google Apps Script, delivering executive dashboards used by Fintech and Finance leadership to monitor key business metrics. Reduced manual reporting overhead and improved access to timely business insights.

Technical Skills

- **Programming Languages:** Python, R, SQL.
- **Machine Learning:** Ensemble Methods, Survival Modeling, Conformal Prediction, Model Calibration, Causal ML (DML, Causal Forest, BCF).
- **Statistics:** Hypothesis Testing, Bayesian Modeling, Causal Inference (IPW, Matching, DiD), A/B Testing, Time Series Analysis.
- **Data & ML Engineering:** dbt, Kedro, MLflow, Bash Scripting.
- **Tools & Platforms:** BigQuery, GCP, Docker, Git, CI/CD (GitHub Actions, GitLab CI).
- **Analytics & Visualization:** Looker Studio, MixPanel, Shiny.

Education

M.Sc. in Applied Statistics Dhaka, Bangladesh
GPA 3.97 / 4.00 2022–2023
ISRT, University of Dhaka

B.Sc. in Applied Statistics Dhaka, Bangladesh
CGPA 3.96 / 4.00 2018–2022
ISRT, University of Dhaka

Publications

[1] **S. K. Shafee** and M. S. Rahman. *On the estimation of the median odds ratio for measuring contextual effects in multilevel binary data from complex survey designs*. [Under review, 2026]. arXiv: 2606.15145 [stat.ME].

[2] **S. K. Shafee** et al. *G-computation for causal effect estimation from observational hierarchical data with unmeasured cluster context*. [Under review, 2026]. arXiv: 2606.14131 [stat.ME].

[3] **S. K. Shafee** et al. "Investigating the causal effect of maternal continuum of care on child's minimum acceptable diet: A multilevel approach using partially pooled propensity score weighting". In: *PLOS ONE* 20.11 (2025), pp. 1–13. DOI: 10.1371/journal.pone.0335972.

Open Source

- skmiscpy — Python package for causal inference diagnostics, including Love plots, mirror histograms, and SMD computation for IPW workflows.
- randomizer — Dockerized Shiny app for experimental unit randomization with SRM testing and covariate balance diagnostics.
- kedrogen — CLI tool for scaffolding reproducible data science projects from cookiecutter templates.
- python-uv-gcloud — Minimal Docker base image with Python, uv, and Google Cloud SDK for CI/CD pipelines and cloud workflows.
- MOR — R package implementing post-estimation functions to compute the Median Odds Ratio and confidence intervals from fitted multilevel binary logistic models.

Awards & Achievements

- **Dean's Award**, *Faculty of Science, University of Dhaka* **2025**
- **Conference Award for Scientists**, *45th Annual Conference of the ISCB, Thessaloniki, Greece* **2024**
- **Datathon Finalist**, Top 10 of 300 teams; built a purchase-behavior prediction model for mobile data packages, *Robi Axiata Limited, Dhaka, Bangladesh.* **2024**
- **National Science and Technology (NST) Fellowship**, M.Sc. thesis research in multilevel modeling, *Ministry of Science and Technology (MoST), Government of Bangladesh* **2023**