

PROFESSIONAL SUMMARY

Data Scientist with 2+ years of experience applying Python, SQL, machine learning, and statistical methods to large-scale customer, behavioral, and longitudinal data. Experienced in predictive modeling, causal inference, experimentation, and building scalable analytics workflows to identify key drivers and measure business impact. Adept at translating complex analytical findings into actionable insights for stakeholders.

EXPERIENCE

Research Assistant | Causal Inference & Econometric Analysis

May 2025 - Aug 2025

Duke University

Durham, NC

- Evaluated 7-wave longitudinal survey data (2004–2011) on Indonesian tsunami survivors using Stata, quantifying the long-term effects of disaster exposure on marriage decisions and family formation outcomes.
- Constructed a publication-ready panel dataset by integrating 7 survey waves and 20+ derived demographic, marital-transition, and education variables to support longitudinal econometric analysis.
- Assessed 48+ econometric specifications using linear probability models with interaction effects, identifying heterogeneous impacts across 3 key demographic dimensions (gender, age, and education) while automating regression reporting workflows.

Data Analyst | Sports Subscription Analytics & Customer Insights

Aug 2023 - Jun 2024

DAZN

Hyderabad, India

- Built scalable Python + Snowflake ETL pipelines integrating 8+ data sources, enabling near real-time analytics across 30+ key performance indicators (KPIs): engagement, churn, CLTV, ARPU and reducing pipeline runtime by 30%.
- Engineered a 30-day churn prediction model across 5+ markets, using behavioral, payment, and customer-support features across 5M+ subscribers, identifying high-risk customer segments and enabled targeted retention campaigns that improved retention by 5%.
- Optimized 15+ production SQL workflows through query restructuring, clustering, and partition-pruning strategies, reducing average runtime by 20% and accelerating downstream analytics and experimentation.
- Developed a segmented Difference-in-Differences (DiD) framework to estimate the impact of subscription price changes across 6 pricing tiers, quantifying heterogeneous churn and retention elasticities for new and renewing users to guide pricing strategy.
- Automated cohort and longitudinal behavioural analyses through Looker dashboards tracking 25+ retention and engagement metrics, enabling continuous monitoring of subscriber lifecycle and market-level trends.
- Processed and structured 10K+ customer complaint records into issue categories and behavioral signals, identifying recurring product and payment friction points to inform prioritization of customer-experience improvements.

Research Assistant | Applied Machine Learning Research

Oct 2022 - Jun 2024

Indian Institute of Science Education and Research, Bhopal

Bhopal, India

- Formulated multivariate LSTM forecasting models in TensorFlow using 920 location-year observations across 6 locations and 44 years (1972–2016), incorporating climate and soil NPK treatment variables to predict agricultural yields.
- Benchmarked stacked, bidirectional, and attention-based LSTM architectures, improving predictive performance by 32% (R-squared) while enhancing model robustness under varying environmental conditions.
- Devised a data randomization framework to mitigate distribution shift and improve model generalization, reducing prediction variance by 5% across repeated train-test evaluations.

Associate Business Analyst | Customer Analytics & Retention Strategy

Feb 2023 - Aug 2023

Lognormal Analytics

Hyderabad, India

- Delivered customer analytics for 4+ enterprise sports-betting operators, applying cohort and segmentation analysis on bettor-level data to identify high-LTV bettors, churn-risk users, and promotion-sensitive segments for CRM and retention strategy.
- Estimated the incremental impact of promotional bonuses on player retention and engagement using Propensity Score Matching (PSM), constructing comparable treatment-control cohorts from historical behavioral features.
- Co-developed an in-house KPI-driven analytics product (Falcon Dive), providing day-level monitoring and predictive signals across 20+ performance metrics, improving reporting efficiency and decision turnaround by ~30% for internal teams and sportsbook clients.

SELECTED PROJECTS ([Link](#))

Hallucination-Resistant RAG QA Assistant for Policy & Economics Papers

- Architected a hallucination-resistant RAG system, implementing a hybrid retrieval pipeline (dense BGE embeddings + BM25 + Reciprocal Rank Fusion + cross-encoder reranking) over a Qdrant vector store, served via a FastAPI backend with PDF upload endpoints and a Streamlit chat UI.
- Fine-tuned DeBERTa-v3-base for claim verification using QLoRA on 185K FEVER claim-evidence pairs, integrated as a post-generation NLI verification layer that classifies cited claims as SUPPORTED/REFUTED/NEI and flags hallucinations before surfacing answers to users.

Multilingual Clinical Trial Summarizer

- Engineered an NLP pipeline to generate plain-language clinical trial summaries from ClinicalTrials.gov, combining text simplification with neural machine translation for multilingual accessibility.
- Automated preprocessing and evaluation workflows incorporating sentence segmentation, medical terminology normalization, ROUGE, and BERTScore metrics, validating summary quality across 100 clinical trials and 10 languages.

Interpretable Uplift Modelling for Digital Advertising Optimization (Criteo dataset)

- Implemented an interpretable uplift modeling framework on 13.9M ad exposure records using Uplift Random Forest (causalml), improving AUUC by 29% vs. baseline propensity targeting and identifying high-value customer cohorts via a rule-based surrogate model for incremental conversion optimization.
- Established a rule-based targeting policy from uplift scores, validated against a randomized holdout, reducing audience size by 80% while retaining ~97% of incremental conversions - quantified with bootstrapped confidence intervals.

Causal Impact of Buy Now, Pay Later on Household Financial Outcomes

- Measured causal effects of BNPL usage on household financial outcomes such as debt burden, late payments, and liquidity (2022 Survey of Consumer Finances).
- Leveraged 3 causal inference estimators (AIPW, IPW, and Causal Forests) to detect heterogeneous treatment effects, identifying the strongest adverse impacts among consumers under age 35.

EDUCATION

Master of Science in Economics and Computer Science

Duke University, Durham, NC.

Teaching Assistant for Causal Inference, ML for Finance, Programming for Data Analytics, Market Research.

- Relevant Courses:** Data Science, Theory & Algorithmic ML, NLP, Causal Inference, Econometrics, Algorithmic Trading, Asset Pricing and Risk Management, Quant Marketing, Probability and Statistics, Database Management Systems, Computer Vision.

SKILLS

- Programming:** Python, SQL, R, Stata | **Data Engineering:** Snowflake, ETL Pipelines, Data Modelling
- Machine Learning & AI:** Scikit-learn, PyTorch, TensorFlow, Transformers, NLP, RAG, LLM Fine-Tuning (LoRA/QLoRA)
- Statistics & Analytics:** Regression, A/B Testing, Uplift Modelling, Segmentation, Causal Inference, Econometrics
- Data Visualization:** Tableau, Looker, MS Excel, Streamlit