SIYING (ELLIE) YANG

Cambridge, MA

617-821-9334 | ellie.yang02siying@gmail.com| https://ellie-yang-siying.github.io/

EDUCATION

MASSACHUSETTS INSTITUTE OF TECHNOLOGY Cambridge, MA

Candidate for Master of Business Analytics (Data Science), Operations Research Center, August 2025 2024 - Present

- ***Coursework:*** Machine Learning, Optimization, Natural Language Processing, Computer Vision, Time Series, Statistical Analysis, Experimentation (A/B testing), Causal Inference, Generative AI (LLMs, RAG), Reinforcement Learning
- ***ML Project:*** Developed a multi-modal stacking model to enhance demand prediction in online marketplaces; integrated enriched feature sets by extracting text embeddings using TF-IDF, SpaCy, FastText and image embeddings using ResNet50

TSINGHUA UNIVERSITY Beijing, China

Double Major: Bachelor of Science in Mathematics and Physics, Bachelor of Engineering in Energy 2020 - 2024

- Academic: 3.86 / 4.00 GPA, Academic Excellence Scholarship (top 5%)
- Leadership: President of Student Union's Learning Department; organized 20 events for 1,300 students

TECHNICAL SKILLS

- Programming & Tools: Python, R, SQL, Julia, C++, Git, MATLAB, Tableau, Power BI, ThinkCell, Azure, Databricks, AWS
- Libraries & Frameworks: NumPy, pandas, Matplotlib, scikit-learn, PyTorch, TensorFlow, Keras, Seaborn, Statsmodel

EXPERIENCE

THE CAMPBELL'S COMPANY / MIT Generative AI Lab Cambridge, MA

Data Scientist (Python, SQL) Feb 2025 - Present

- Build a multi-agent Retrieval-Augmented Generation (RAG) system with LLaMA on Databricks; integrate 18 datasets and implement retrieval logic, and tool-calling orchestration to answer real-time marketing questions through a chatbot interface
- Perform statistical analysis to examine marketing spend over time and across campaigns, build regression models to evaluate relationship between spend and ROI; integrate a diagnostic agent to interpret current campaign performance
- Develop a LightGBM forecasting pipeline to simulate expected returns under different budget allocation scenarios; integrate forecasting outputs with a recommendation agent for future budget reallocation strategy suggestions

MASSMUTUAL / MIT Analytics Lab Cambridge, MA

Data Scientist (Python) Fall 2024

- Optimized 6-step website journeys by fine-tuning a RoBERTa-based transformer to extract conversion-aware journey sequence embeddings and conducted counterfactual simulations to curate high-submission journey flows, contributing to a 1.5x increase in form submissions and a projected $7M ROI lift
- Visualized latent structures in high-dimensional customer journey embeddings using t-SNE, enabling interpretation and identification of journey paths that drive higher lead-form submissions to connect with financial advisors
- Segmented users into three journey-based clusters using K-Means and interpreted behavioral patterns with Optimal Classification Tree and LIME; identified a high lead-form submission cluster that served as the focus for journey redesign
- Cleaned 2M+ records of customer's browsing data from MassMutual's website to prepare for user journey modeling; removed invalid sequences, identified outliers, resolved mislabeled records, standardized steps using fuzzy matching (difflib)

POIZON *(Leading Chinese e-commerce platform specializing in shoes)* Shanghai, China

Data Scientist Intern (Python, SQL, Tableau) Summer 2023

- Conducted A/B testing on page redesign strategies, performed hypothesis testing to confirm statistical significance, resulting in a 15% increase in CTR, 0.5% lift in per-user value, and a 13-point Net Promoter Score improvement
- Applied a two-model uplift modeling framework to estimate heterogeneous treatment effects of the page redesign; identified a high-response user segment with a predicted 33% uplift in click-through rate, informing customer targeting strategies
- Identified friction points and content-level barriers by detecting abnormal CTR drops across customers' in-app page-to-page browsing funnel stages through statistical variance analysis; proposed navigation, content, and layout redesign strategies
- Performed ETL and queried 100M+ in-app user session logs using Spark SQL; aggregated clickstream data into funnel-stage metrics such as page-to-page click-through rates (CTR) and browsing durations across user journey

OLIVER WYMAN Remote

Data Scientist, Part-Time Assistant (Python) Spring 2023

- Engineered 125 features via transformations, statistical summaries, and sliding-window; performed feature selection using correlation, regularization and permutation methods
- Built time-series models with LSTM and XGBoost to forecast 10+ interest rate volatilities over 3-, 5-, and 10-year horizons; tuned hyperparameters with Optuna
- Classified predicted volatilities into no-risk, low-risk, and high-risk levels based on historical baseline and standard deviations, achieving a low classification error of 4%
- Improved model interpretability by analyzing feature importance and identifying key feature changes in historical data that triggered 'high-risk' predictions; presented findings to a leading bank client to support their risk measurement initiatives