Yiwen Li

Boston, MA | +1 857-270-8656 | yiwenli@fas.harvard.edu | linkedin.com/in/yiwen-li-8671b7209

EDUCATION

Harvard University Sept 2024 - Jun 2026

Master of Science in Data Science

- **Relevant Courses:** Machine Learning (ML), Data Science, Natural Language Processing (NLP), Multilevel Modeling

University of Toronto Sept 2019 - Jun 2023

Bachelor of Science in Statistics and Economics (double majors) | ***GPA: 3.99/4.00***

- **Relevant Courses:** Advanced Data Analysis, Python Programming, Method for Multivariate Data, Time Series Analysis
- **Awards**: Science & Mathematics Scholarship (Top 1%), Reuben Wells Leonard Scholarship (Top 1%), C. L. Burton Open Scholarship, U of T Special Admission Scholarships, Dean's List Scholar (3 years)

Programming & Software: R (4 years), Python (2 years), SQL (2 years), Microsoft Office Suite (Excel, PowerPoint, Word)

INTERNSHIPS

Jd.Com, Inc Aug 2022 - Dec 2022

Data Analyst *Beijing, China*

- Implemented automatic workflows in **Power Query** to standardize large-scale data and generate weekly reports within 5 seconds, empowering faster data-driven marketing decisions through 20+ KPIs like conversion rate
- Optimized processing efficiency for 100K+ eCommerce data rows using advanced Excel functions (**VLOOPUP**, **Pivot Tables)** and Python **Pandas**, enabling accurate data extraction and reducing human error rates by 30%
- Created 10+ **Tableau** validation dashboards to ensure consistency between **SQL** queries and reporting logic
- Increased business sales by 25% through optimizing logistics resource allocation with an in-depth quantitative analysis of regional gross profit and customer engagement, supporting cross-functional teams in establishing a new warehouse

UnionPay May 2022 - Jul 2022

Data Analyst *Shenzhen, China*

- Designed 20+ **ETL** (Extract, Transform, Load) data pipelines in **SQL** to process 600K transactional data from financial institutions, increasing reporting efficiency by 40% and saving the team 8 hours per week
- Assessed annual company performance in **R** by benchmarking 200+ business indicators, with Exploratory Data Analysis (**EDA**) to examine distributions and the Kruskal-Wallis test to identify statistically significant indicators
- Built a risk prediction model (**Logistic Regression**) to detect suspected cash-out activities with a 90% accuracy, informing the executive board of high-risk merchants and recommended solutions through clear **presentation**

RESEARCH EXPERIENCES

Surrogate Assisted Positive Unlabeled Learning on EHR data May 2023 – May 2024

Research Assistant | Supervisor: Prof. Jessica Gronsbell *Toronto, Canada*

- Developed a **semi-supervised ML** algorithm for phenotype prediction with an **AUC** score exceeding 93 that outperformed all 7 baseline models, offering a substantial and accurate solution to reduce manual chart-review efforts for data labeling
- Achieved robust feature selection with adaptive **LASSO** and automated hyperparameter tuning with **R** (*glmpath*)
- Conducted **NLP** analysis on real doctor notes to extract disease-indicative terms using the Unified Medical Language System, improving the model's predictive accuracy by 10%
- Demonstrated model effectiveness in handling **high-dimensionality** through extensive model robustness testing on 42K Electronic Health Records (EHR) from the MIMIC database with a generation of 1100 covariates
- Presented research poster to 300+ professionals and published the model as an **R package** (*SAPUL*), contributing to the open-source development and allowing the research community to reproduce statistical findings

Causal-Debias for Demographic Bias Mitigation in NLP Models Sept 2024 – Present

Capstone Project | Supervisor: Prof. Jacob Andreas *Boston, MA*

- Investigated demographic bias in language models by applying a novel Causal-Debias approach to reduce biases related to sexual orientation, religion, and disability during model fine-tuning to improve model fairness in NLP tasks
- Developed a perturbed dataset with neutral or counterfactual terms over 180,000 sentences from WikiText and Stanford Sentiment Treebank (SST) to train a **BERT-based model** focused on minimizing reliance on bias-related features
- Evaluated model performance on **SST-2 sentiment classification** task, achieving an accuracy of 93% while significantly reducing bias as measured by SEAT and CrowS-Pairs benchmarks
- Conducted comprehensive benchmarking against existing debiasing methods, achieving a 27% improvement in SEAT scores over baseline, with consistent results in CrowS-Pairs indicating enhanced model stability and reduced bias

Analysis of FDA MAUDE System for AI/ML Device Safety Reporting Apr 2023 – Present

Research Assistant | Supervisor: Prof. Boris Babic *Toronto, Canada*

- Conducted a comprehensive analysis of the FDA's MAUDE database, investigating adverse event reports for 823 **AI/ML medical devices** approved from 2010 to 2023, evaluating 54 features per device
- Identified critical gaps in the MDR reporting system by analyzing patterns of missing data, instances of event misclassification, and the lack of dedicated tracking mechanisms for AI/ML-specific issues
- Co-authored a manuscript proposing improvements for postmarket surveillance and reporting practices, targeted for submission to **Nature Biomedical Engineering**