Shuyi (Ryan) CHEN
(412) 909 7683 | shuyic@alumni.cmu.edu

EDUCATION

Carnegie Mellon University May 2023
Master of Information Systems Management Pittsburgh, PA
- Courses: Statistical ML (PhD); Introduction to ML (PhD); Introduction to Deep Learning; Causal Inference; Advanced NLP; Distributed System

The Chinese University of Hong Kong, Shenzhen May 2021
Bachelor of Business Admin. in Global Supply Chain Management (Minor in Statistics), First Class Honors Shenzhen, China
- Courses: Fundamentals of Machine Learning, Techniques for Data Mining, Data Structure (Java), Statistical Inference, Time Series and Forecasting (R), Probability and Stochastic Process, Quantitative Analysis in Public Policy

University of California, Berkeley (Summer Visiting Program) June – August 2018
- GPA: 4.0/4.0 | Relevant courses: Politics and Social Change; Gender and Women's Studies

RESEARCH EXPERIENCE

Uncertainty-Aware Robust Learning on Noisy Graphs Jan. – June 2023
- Developed a novel uncertainty-aware graph learning framework that addresses challenges caused by noisy measurements in real-world graphs; Implemented the graph learning framework using PyTorch.
- Conducted experiments and evaluated the proposed framework, which demonstrates its superior predictive performance compared to state-of-the-art baselines (including GCN and GAT) in diverse noisy settings.
- Collaborated on the writing of the paper, especially for the Experiment and related work section.

API-Assisted Code Generation for Question Answering on Arbitrarily Structured Tables Feb. – June 2022
- Proposed and implemented a small number of API functions to incorporate external knowledge and address complex relational questions in table question answering (QA) systems. This approach resulted in substantial improvements over previous state-of-the-art systems specifically designed for different table QA datasets.
- Leveraged few-shot prompting techniques to develop code generation models that translated natural language questions into executable Python programs on the multi-index data frames, imprpove Table QA capabilities for LLMs.

Learning Generalized Audio Representation Through Batch Embedding Covariance Regularization Feb. – May 2022
- Discovered and tested the positive connection between the representation layer's projection separability and its downstream general-purpose performance by replicating the SOTA PaSST model and three other classic audio representation models (OpenL3, Hubert, CREPE) on NeurIPS HEAR2021 dataset.
- Based on the findings, designed a regularization term that encourages the generalization capability of embedding representation.
- Implemented the proposed regularization with PaSST model and pretrained on FSD50K dataset. Experiments show it improves the performance in all four downstream tasks from three different domains by a margin from 2% to 33% given the same pre-training.

WORK EXPERIENCE

ByteDance April – July 2022
Data Science Intern Beijing, China
- Developed a Python toolbox for Multiple Touchpoint Attribution models, containing 4 data-driven models of Shapley Value, Markov, Additive Hazard Survival Model, and Bagged Logistic Regression, as well as 5 rule-based attribution models.
- Based on the toolbox's evaluation report of robustness and accuracy, proposed an attribution-based data-driven ad feature selection scheme and positive sample weighting method to mitigate the delayed feedback issue in ad recommendation. Preliminary results show it uplifts ADVV by 2% and CVR by 0.5% in affected ad scenarios with 95% confidence (equivalent daily gain of $5K).
- Employed statistical methods of Synthetic Control Method for causal impact evaluation with time series data, PSI for audience similarity analysis, and exponential model for decaying impact modeling. Reports impacted 5 advertisers' advertising plans totaling $1M budget by the sales team I supported.

Kuaishou Technology July – Sept 2020
Data Scientist Intern, Department of Data Science Beijing, China
- Estimated the net increase in total views depending on the timing of similar-content recommendations using conditional probability. Therefore, optimized the timing of the algorithm which resulted in a 3% increase in users' daily watch time.
- Implemented Rubin and Imbens' optimal propensity score trimming technique and various matching algorithms (including XGBoost, Random Forest, GBM, and Light GBM) to reduce the Propensity Score Matching estimator's variability and mitigated pre-treatment heterogeneity. The comparison showed it boosted estimation by 15%, reduced the estimator's variance by 40%, and improved the soundness of pre-treatment parallel assumption, effectively reducing the necessary sample size needed to half.
- Encapsulated optimal trimming into the company's Python causal inference package that benefited a DS team of 60 employers.