

Qifan Zhang

Yale University, New Haven, CT, USA, 06511
qifan.zhang@yale.edu — +1 (xxx) xxx-xxxx — <https://zh-qifan.github.io>

RESEARCH INTERESTS

Machine Learning Theory and Applications

EDUCATION

Yale University, New Haven, CT, USA Aug. 2023 — Dec. 2024
Master of Science in Statistics and Data Science Courses with Honors Grade: 5/5

Nanjing University, Nanjing, China Sept. 2019 — June 2023
Bachelor of Science in Statistics Cumulative GPA (WES): 3.98/4.00
Honors: National Scholarship, Cheung Kong Graduate School of Business Scholarship, Suzhou Scholarship
Thesis Title: Cost-sensitive Set-valued Classifiers with Expected Size Control

National University of Singapore, Singapore Aug. 2021 — Dec. 2021
Student Exchange Program Cumulative GPA: 5.00/5.00

ACADEMIC EXPERIENCE

School of Medicine, Yale University New Haven, CT, USA
Data Analyst, Deep Learning for Antimicrobial Peptides (AMPs) Generation Oct. 2023 — Present

- Implemented a generation pipeline for AMPs by training a neural network to fit minimum inhibitory concentration (MIC) on peptide sequences and optimizing it with Simulated Annealing method; The neural network includes a ProBert (a pre-trained model) and a CNN model, achieving 92% precision on average on 8 datasets on average.

Lee Kong Chian School of Business, Singapore Management University Singapore (Remote)
Research Assistant, Predict House Prices with Machine Learning and Automated Valuation Model Nov. 2021 — June 2023

- Designed a searching algorithm for similar house transactions in Singapore based on comparison of geolocation and purchasing time to generate about 30 main features describing purchasers' pricing reference behaviors; Engineered over 60 supplemental features on multiple data sources such as public facility location, list price, buyer and seller information.
- Trained and optimized 15 XGBoost models on different data samples and feature combinations; The best ensemble model achieves less than 2% MAPE in predicting house prices in Singapore from Jan. to Feb. 2023, outperforming other models in the literature.

PUBLICATIONS

Academic Paper

In-Progress

- Minimax Set-valued Classifiers with Expected Size Control with Professor Lihong Wang

Learning Notes

- Notes in Statistics and Machine Learning

INDUSTRIAL EXPERIENCES

Lingjun Investment Shanghai, China
Quantitative Researcher Intern, High Frequency Trading Team Feb. 2023 — May 2023

- Created 53 features (calculated per 3s) by analyzing time-series and cross-sectional patterns of limit order book and tick data of A-share stocks and filtered features using SHAP value; Implemented selected features by Python and C++ to improve the correlation score of high-frequency ensemble deep learning models by 6%.
- Developed and deployed a selection strategy on ensemble models using ANOVA method with 24 indicators, such as liquidity and momentum, increasing the models' performance by 2% on average.
- Optimized the run time of fill-and-kill (FAK) order identification from 1 millisecond to 10 microsecond by designing a comparison method based on order sequence number to replace the hashing method.

Huawei Shenzhen, China
Product Data Engineer Summer Intern, Consumer Business Group June 2022 — Aug. 2022

- Built 5 Hive tables to model over 100 users' behavior-related features such as application usage and ad click history by SQL, utilized by 2 teams to improve the accuracy of predictive models by 15%, resulting in 8% improvement on ad click rate.
- Optimized the sampling methods for 5 users' interest predictive models by combining hard negative sampling and popularity-based negative sampling, improving the relevant ad click rate by 16% on average.
- Built LightGBM models (precision and recall achieving 90% on average) with over 100 features on around 1 million user samples to predict users' interest in financial products such as credit cards, improving the ad click rate about 5 times in A/B testing.

PROJECTS

Explore Stock Graphs for Stock Price Prediction [Github Repo]

New Haven, CT, USA

Course Project for CPSC 583 Deep Learning on Graph-Structured Data

Sept. 2023 — Dec. 2023

- Designed and implemented a LSTM + GAT framework with the basis learning method to predict 5-day-ahead stock prices in NYSE and NASDAQ from 2016 to 2017. The stock graphs includes a supplier-customer graph (constructed by Wikipedia data), a sector-industry graph and a comovement graph (constructed using historical return correlation). The proposed model can reduce the MSE loss of LSTM by over 50%.

SELECTED COURSES

Master's Courses

- Statistical Inference
- Statistical Decision Theory (Expected)
- Advanced Optimization (Expected)
- High Dimensional Statistics (Expected)
- Intermediate Machine Learning
- Deep Learning on Graph-Structured Data
- Data Science Software System

Bachelor's Courses

- Advanced Mathematical Statistics
- Advanced Probability
- Measure and Integration
- Functional Analysis
- Stochastic Processes
- Partial Differential Equations
- Time Series Analysis

ENGLISH & GRE TESTS

TOEFL: 108

Listening: 29 — Reading: 28
 Speaking: 23 — Writing: 28
 Test date: Sept. 2022

GRE General Test: 326

Quant: 170 — Verbal: 156
 Analytical writing: 4.0
 Test date: May 2022

SKILLS & INTERESTS

- **Programming:** Python, C++, R, \LaTeX , Git, Bash, SQL
- **Statistical and Data Science Packages (Python & R):** pandas, numpy, scikit-learn, scipy, PyTorch, multiprocessing, cython, SHAP, matplotlib, seaborn, shiny
- **Interests:** Soccer (Team captain and champion of Nanjing University Soccer Premier League in 2021), Swimming