Xinyu Chen

+86 18611777299 | cc1120173514@163.com | Hefei, China | Homepage

EDUCATION

University of Science and Technology of China (USTC) — Hefei, China
M.S. in Statistics, School of Management — *Sept. 2021 – Nov. 2024*
National University THE Ranking: #6

Beijing Institute of Technology (BIT) — Beijing, China
B.S. in Statistics, School of Mathematics and Statistics — *Sept. 2017 – Jun. 2021*
National University THE Ranking: #14

AWARDS & HONORS

Outstanding MSc Thesis Award, USTC — *2024*
1st-Tier Postgraduate Scholarship, USTC — *2021, 2023*
Distinguished Graduate Award, BIT (TOP 10%) — *2021*
China National Encouragement Scholarship (TOP 5%) — *2018*
Outstanding Undergraduate Scholarship, BIT (TOP 15%) — *2017 – 2021*

PUBLICATION & PRESENTATION

Chen X., Yu D., & Zhang X. (2024), Optimal Weighted Random Forests, *Journal of Machine Learning Research (JMLR),* 25(320), 1-81.

Chen X. (2023), Statistical Prediction and Machine Learning, 20-minute presentation, First International Conference on Machine Learning and Statistics.

INTERNSHIP EXPERIENCE

Anhui Province Key Laboratory of Contemporary Logistics and Supply China — Hefei, China
Student Research Assistant — *Mar. 2023 – Jun. 2023*

- Worked on predicting one-day package delivery based on customer and goods data from JD.com.
- Processed big data from multi-source, merging, cleaning, and feature engineering with Numpy and Pandas.
- Implemented asymmetric binary classifiers (RF, XGBoost, and ANN models) in ensemble frameworks, achieving an F1 score of 0.78, surpassing JD.com's conventional forecasting strategy with an F1 score of 0.72.

Shanghai Gene Asset Management Co., Ltd. — Shanghai, China
Quantitative Stock Analysis Intern — *Jul. 2022 – Oct. 2022*

- Cleaned financial data from HeidiSQL and Wind databases.
- Engaged in mining alpha factors for stock selection using Python by capturing patterns in financial market.
- Participated in collaborative meetings, reviewed financial research reports, and contributed to brainstorming sessions to enhance alpha factor generation strategies.

RESEARCH PROJECTS

Large Language Models (LLMs) Combined with Statistical Methods — Xi'an, China
Research Project – Numerical Experiments in Progress — *Aug. 2024 – Present*

- Proposed an ensemble framework for LLMs that distills knowledge into conventional models, reducing computational costs while preserving diverse and powerful inference capabilities.
- Proposed a framework for fine-tuning LLMs in multivariate time series forecasting, incorporating copula functions and graph knowledge to represent multivariate structures, addressing the gap in univariate-focused LLM research.

Transfer Learning for Random Forests — Hefei, China
Research Project – Manuscript in Progress — *Feb. 2024 – Present*

- Proposed two novel transfer learning methods for random forests, one transferring "variable importance in branches" and the other transferring predictions with cross-validated weights, using AUC as the objective.

- Conducted numerical experiments on real-world datasets, demonstrating that the first method excels with domain shift, while the second is more effective without domain shift. Both methods outperform random forests trained solely on the target domain.

Optimal Weighted Random Forests Hefei, China
Research Project – Published *Feb. 2022 – Apr. 2024*

- Proposed a novel weighting algorithm that integrates regression trees within random forests, utilizing weights derived from a Mallows-type loss function.
- Designed an iterative algorithm to accelerate convex optimization by reducing its computational order.
- Established the asymptotic optimality of the proposed algorithms, demonstrating that their forecasting performance approaches that of the theoretically optimal but infeasible weighted random forests under certain conditions.
- Performed comprehensive numerical studies on real-world and semi-synthetic datasets using R, consistently outperforming equal-weight forests and other weighted random forest methods.
- Offered practical guidance on the real scenarios where the proposed algorithms are most effective.

Weighted Random Forests for Classification Hefei, China
Research Project – Manuscript Completed *Sept. 2021 – Dec. 2021*

- Developed weighting algorithms for random forests, integrating classification trees with objective functions based on metrics like weighted AUC, KL distance, and classification accuracy to balance performance and complexity.
- Conducted numerical experiments on real-world and synthetic datasets using R, achieving improvements of 1% to 4% in AUC, accuracy, recall, precision, and F1 score.

Bayesian Model Averaging for Linear Regression Models Hefei, China
Course Project *Oct. 2021*

- Compared Bayesian model averaging methods (Occam's Window and Markov Chain Monte Carlo) with frequentist model averaging and traditional model selection techniques.
- Conducted numerical experiments on the UScrime dataset using R, evaluating the performance of these methods in analyzing the effects of punishment regimes on crime rates, and produced a comprehensive report.

Weighted Random Forests for Regression Beijing, China
Bachelor Thesis *Oct. 2020 – May 2021*

- Reviewed and compared various weighting algorithms for regression trees from the literature.
- Implemented these algorithms in Python and evaluated their performance against conventional methods through Monte Carlo simulations. Results showed superior forecasting accuracy, especially in cases of homoskedasticity and smaller sample sizes.

Multivariate Linear Analysis of Residents' Consumption Levels Beijing, China
Course Project *May 2020*

- Analyzed the impact of various factors on residents' consumption expenditure across 31 regions, identifying 9 key influencing factors for modeling based on consumption economic theory.
- Developed a multiple stepwise regression model using SPSS, conducting diagnostic tests such as outlier detection, heteroscedasticity, autocorrelation, and multicollinearity to enhance model accuracy.
- Provided government recommendations to boost consumption levels during COVID-19, based on the model's findings.

City Selection for Graduates Using Multivariate Statistical Methods with SAS Beijing, China
Course Project *Nov. 2019*

- Conducted a multivariate analysis to guide living city selection for graduates based on housing prices, applying principal component and factor analysis to identify key indicators: economic development and housing supply.
- Utilized cluster analysis to classify 31 provincial capital cities into 3 tiers by housing prices of China, and performed discriminant analysis to categorize 5 additional metros.
- Developed regression models using stepwise and principal component regression to predict future housing prices.
- Created a structured decision-making framework for practical application.

SKILLS & INTERESTS

Programming: R, Python, MATLAB, SAS, SPSS
Languages: Mandarin (Native); English (Fluent: IELTS 6.5); Korean (Basic)
Research Interests: Multi-source Learning, Model Selection and Averaging, Machine Learning, Large Language Model
Other Interests: Chinese Kungfu, Hiking, Cycling, Painting, Volunteering at Animal Shelters.