Haoyue (Diane) Yang

116 Oak Ave, Ithaca, NY 14850 | (424) 489-1312 | hy569@cornell.edu

EDUCATION

Cornell University 08/2020 – 12/2021
Master of Professional Studies, Applied Statistics

University of California, Los Angeles 09/2015 – 06/2018
Bachelor of Science, Mathematics/Economics with Specialization in Computing
Bachelor of Science, Statistics
Vice President of Chinese Literature and Art Creation Association (CLACA), UCLA
Dean's Honors List

University of California, Irvine (Transferred out) 06/2014 – 06/2015
Dean's Honors List

PhD-LEVEL COURSEWORK

Operations Research
  Bayesian Statistics and Data Analysis (A+), Applied Stochastic Processes(A-), Statistical Principles(A),
  Mathematical Programming (Audit), Online Decision Making and Market Design (A)

Statistical Science
  Computationally intensive statistical methods (A), Theory of Linear Models (A+)

PUBLICATIONS IN PREPARATION

Maximizing Portfolio Predictability with Machine Learning
  Michael Pinelis, Haoyue Yang, and David Ruppert (2022), in preparation.

Control Burn: Feature Selection by Sparse Forests
  Brian Liu, Miaolan Xie, Haoyue Yang, and Madeleine Udell (2022), in preparation.
  Target: Journal of Statistical Software, January 2022

Data-Efficient Random Forest with Feature Space Sampling
  Haoyue Yang, Yucheng Lu, Chengrun Yang and Madeleine Udell (2022), in preparation.
  Target: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), February 2022.

RESEARCH EXPERIENCE

Research Assistant 07/2021 – Present
Cornell University Ithaca, NY
Advisor: Dr. Madeleine Udell, School of Operations Research and Information Engineering, Cornell University

- Present statistical inference and explainable machine learning topics in group meeting.
- Contribute to project *ControlBurn: Feature Selection by Sparse Forests*: evaluate sampling methods at bagging stage, propose metrics for interaction effects, and research on statistical inference of current methods.
- Lead project *Data-Efficient Random Forest with Feature Space Sampling*; techniques including simple random sampling with/without replacement, Quasi Monte Carlo sampling, bisection sorting, and copula transformation.

Research Assistant 05/2021 – Present
Cornell University Ithaca, NY
Advisor: Dr. David Ruppert, School of Operations Research and Information Engineering, Cornell University

- Conduct research collaboratively on Maximizing Portfolio Predictability with Machine Learning.
- Produce the entire model pipeline including: data collection and preprocessing, feature engineering and analysis, hyperparameter tuning, model construction and interpretation.

Research Assistant   03/2021 – 08/2021
Cornell University   Ithaca, NY
Advisor: Dr. Jura Liaukonyte, SC Johnson College of Business, Cornell University

- Coded Stata, R and Python programs to process and analyze data; reported the work progress to Prof. Jura Liaukonyte and provided improvements and suggestions for current research work.
- Part of the research work contributed to the paper "How Viewer Tuning, Presence and Attention Respond to Ad Content and Predict Brand Search Lift", M. McGranaghan, J. Liaukonyte, K.C. Wilbur (2021), forthcoming in *Marketing Science.*

On Demand Stock Return Forecasting Model Optimization (Python)   01/2020 – 05/2021
Gravity Investment   Remote
Advisor: Dr. Thomas DiCiccio, Department of Statistics, Cornell University

- Retrieved financial data via API from Tiingo.com; conducted data preprocessing to convert raw datasets into quarterly rolling-window data.
- Designed feature engineering program to add informative financial characteristics; performed feature selection to identify statistically significant features.
- Developed a two-step rolling-window model to predict alpha and beta; constructed the most profitable profolios based on the ranking of assets by predicted excess return.
- Commented on the performance of the two-step model and proposed future improvements.
- The project resulted in a 32-page research report.

Dynamic Hierarchical Models (R)   03/2021 – 05/2021
Advisor: Dr. David Ruppert, School of Operations Research and Information Engineering, Cornell University

- Conducted an independent research project for the PhD course ORIE 6780, Bayesian Statistics and Data Analysis.
- Illustrated the basic concepts of dynamic hierarchical models using both simple examples and rigorous definition; commented on an extension of dynamic hierarchical model, Conditionally Gaussian Dynamic Model.
- Modified the dynamic linear model package "dlm" in R and implemented a dynamic hierarchical model on real data, PM2.5 in Five Major Cities in China.
- Evaluated the model performance and discussed future improvements.
- The project resulted in a 21-page research paper.

Measure Theory and the Foundation of Probability Theory   09/2017 – 12/2017
Advisor: Dr. Andrew Manion, Department of Mathematics, UCLA

- Conducted an 11-week independent research of honors contract on measure theory and probability theory.
- Revisited probability theory from the real analysis perspective and explored the parallel relationship between measure theory and probability theory.
- Wrote a 10-page research paper that summarizes the understanding of basic concepts of measure theory and how some of the concepts contribute to the foundation of probability theory.

WORK EXPERIENCE

Mile Express Inc.   Remote
Data Analyst   04/2020 – 02/2021
- Performed quantitative data analysis by implementing statistical methods and visualizing analysis results.
- Conducted international industry research in the emerging markets of Fintech, Biotech, last mile delivery and healthcare, etc.; presented business insights to clients based on research results.
- Provided research in global VC firms, incubators, and business accelerators to boost startup growth and proceed to match funding.

Just The Berries PD Corporation   Los Angeles, CA
Executive Assistant to Chairman/CEO   11/2018 – 02/2020
- Pioneered large-volume industry research work of launching and marketing the new products.
- Led a team of four departmental managers to orchestrate broad market investigation of health supplements in Japan, China, and the USA.

- Compiled formula, sales, cost, and price data to identify product specialties and assess pricing strategy; presented 7 industry and market research reports to facilitate go-to-market decisions.
- Performed complex data analysis work by applying statistical methods on a weekly basis.
- Queried raw data from websites and database; evaluated customer profiles and performance of E-blast advertisements, promotional campaigns, and sales activities.
- Translated the analysis into actionable insights and made recommendations to CEO/Chairman.
- Acted as the primary contact for internal department managers and external constituencies on all matters pertaining to CEO/Chairman.
- Oversaw and reported daily activities of sales, marketing, and operations departments in a fast-paced environment.
- Presented business insights to CEO/Chairman and managers based on the overall knowledge of the company.

Sky Vision Insurance   San Marino, CA
Financial Service Representative   05/2018 – 11/2018
- Conducted financial market research and compared return of products from different insurance companies.
- Forecasted the risk and potential profit of products and provided the consulting reports.
- Helped clients manage their wealth distribution, retirement plan and other personal accounts; promoted suitable insurance and financial products based on customers' needs and requirements.

TEACHING EXPERIENCE

UniStudy USA   Remote
Mathematics and Statistics Tutor   03/2020 - Present
- Teach private mathematical, statistical and econometrics lessons to groups of 1-3 undergraduate students.
- Help review course materials, solve problems in students' homework, and prepare for midterms and finals.

Teaching Assistant: Special Topic in Planning (CRP 3850)   02/2021 – 05/2021
Cornell University   Ithaca, NY
- Examined and graded students' homework in applying machine learning in city and regional planning.

Math Tutor, ECC Academy, Los Angeles, CA   01/2016 – 09/2016
Math Tutor, Save Our Youth, Costa Mesa, CA   09/2014 – 06/2015
- Volunteered to teach 2.5-hour Math lessons every week to middle and high school students in a non-profit organization, whose mission is to provide academic and social support for low-income youth.
- Made study plans for students and helped with their homework; answered questions related to AP Calculus, algebra, pre-calculus, etc.

SELECTED COURSE PROJECTS AND COMPETITIONS

Methodology of Modifying Random Forest for Imbalanced Classification (R)   03/2021 – 05/2021
- Conducted the collaborative research project under supervision of Dr. Sumanta Basu, for the PhD course STSCI 6520, Computationally Intensive Statistical Methods.
- Explored model accuracy problems lying in imbalanced classification scenarios; researched and commented on prevailing methodologies, including oversampling, undersampling, and data augmentation, etc.
- Introduced improvements for current methods, such as Weighted Random Forest and Balanced Random Forest.
- Examined and evaluated the above methods using 5 real-world datasets; developed metrics to compare and comment on these methods; discussed pros and cons, and suitable scenarios for different methods.

Beijing PM2.5 Time Series Analysis (R)   04/2020 – 05/2021
- Built and compared three time series models on Beijing PM2.5 dataset: local level model, local linear model and smooth trend model.

US Election Prediction (Python)   12/2020
- Programmed to create features from three datasets: 2012 and 2016 election results and geographical neighbors.
- Built and compared logistic regression, SVM and neural network models; selected the best model, multi-layer Perceptron classifier, which increased the prediction accuracy from 68% to 82%.

Statistical Consulting: Liver Transplant Study (R)   09/2017 – 12/2017
- Explored liver transplant data by plotting boxplots, distributions, and correlations between variables using ggplot package in R.
- Applied backward selection to choose related variables; selected best logistic regression model with the smallest AIC to predict patients' survival time and the utilization rate of livers.

Los Angeles Fire Department Reaction Time Prediction (R)   05/2017 - 06/2017
- Predicted Los Angeles Fire Department reaction time by experimenting data with xgboost, random forest and linear models; obtained best model by using gradient boosting algorithm.
- Top 20 as an outstanding team in the Kaggle competition of 500 students with a 25% more accurate model than average; provided suggestions based on prediction results.

Statistical Consulting: Dog Adoption Project (R)   03/2017 – 06/2017
- Compared a logistic model with a Bayes' classifier to determine the key factors to the outcomes of sheltered dogs.
- Interpreted the models and visualized the analysis results using ggplot package in R.
- Co-authored a 16-page final research paper.

Financial Markets and Institutions: Chinese Economy: Today and the Future   10/2016
- Collaborated with five classmates to conduct insightful research in Chinese economy with three major focuses on financial bubble, housing market, and depreciation of the Chinese Yuan.
- By discussing macroeconomics topics, presented prediction on the future Chinese economy and proposed an investment portfolio consisting of Ferex, Reits, stocks, commodities, and bonds.

American Statistical Association DataFest 2017: Expedia Hotel Booking Project (R, Tableau)   05/2017
- Analyzed millions of users' hotel booking records from Expedia and learned customers' preferences of booking in summer vocations by applying machine learning models.
- Presented and visualized our analysis results using Tableau and ggplot package in R.

UNIVERSITY AND COMMUNITY SERVICE

Caregiver for Patient with Huntington's Disease   11/2018 – 06/2019
- Performed day care work for patient with Huntington's Disease, such as cooking, serving meals, accompanying with walk and exercise, etc.
- Quelling and comforting patient's emotional shifts resulted from mental problems, such as rage, depression and suicidal tendencies.

Vice President, Chinese Literature and Art Creation Association, UCLA   09/2016 – 06/2017
- Led the club to promote Chinese music, arts and literature creations to UCLA community; took charge of weekly activities for the club, including holding meetings and seminars, organizing club activities, etc.
- Managed financial activities of the club, including allocating activity expenses, applying for club funding, etc.
- Led a team of five members to work on social media marketing activities; managed WeChat and Facebook accounts by publishing weekly articles and artwork from the club.

Member, Chinese Students and Scholars Association, UCLA   09/2015 – 06/2017
- Attended weekly meetings and promoted academic exchange between PhD students, faculty members and scholars from other universities.
- Managed WeChat account by summarizing weekly seminars in math and engineering departments; helped hold academic seminars and presentations sponsored by the club.

Member, Anteater for Autism, Irvine, CA   02/2015 – 06/2015
- Helped hold social parties for adults with Autism and their caregivers.
- Accompanied adults suffering from Autism, played games with them, and brought care and love to their lives.

Technical Skills

- **Programming skills**: C++, Java, Python, R, Stata, MATLAB, JavaScript, HTML, PHP, CSS.
- **Database Management Skills**: SQL, Hadoop, Pig, Spark, etc.

Hobbies

- Electronic Piano
  Level 9 (The highest level of amateur electronic piano players in China)

- Chinese Calligraphy
  First Prize, National Teenager Calligraphy Contest, Chinese Ministry of Education, China

- Writing
  First Prize, 14th Chinese Newspaper National Writing Competition, China Association of Newspapers