Xinyi Yang

christinaxy131@gmail.com | +1 (551) 338-1648 | https://www.linkedin.com/in/xinyi-yang-b831a928a/| https://christinaxy31.github.io/

Jersey City, NJ, 07302

EDUCATION

New York University | *New York, the United States* *Sep. 2023-May. 2025*

M.S. in Data Science | **GPA:** 3.89/4.0

Coursework: Probability and Statistics for Data Science, Machine Learning, Big Data, Deep Learning, Probabilistic Time Series Analysis, Natural Language Understanding

Beijing University of Technology | *Beijing, China* *Sep. 2019-Jul. 2023*

B.S. in Computer Science and Technology | **GPA:** 3.79/4.0

Coursework: Object-oriented Programming, Network Programming, Operations Research, Advanced Mathematics, Linear Algebra, Data Structures and Algorithm, Database System, Software Engineering, Distributed Systems

Publication: Yang, X., *Prediction of Credit Risk Based on Logistic Regression and Random Forest Algorithm*. In Proceedings of the 2021 International Conference on Computer Engineering and Information Processing. Paper ID: CEIP-521375.

TECHNICAL SKILLS

ML & Statistical Analysis Skills: Deep Learning Models (Transformer, CNNs, RNNs), Machine Learning Models (Regression Models, Tree Models, Clustering Models), Time Series Models, Bayesian Statistics, A/B Testing, Hypothesis Testing, Data Visualization

Programming Languages: Python (Scikit-learn, PyTorch, Matplotlib, Seaborn, Pandas, Numpy), SQL, Java, C++, C, R, Scala, Spark, MATLAB, LINGO, SPSS, Stata, CSS, HTML

Platform & Tools: Git, MySQL, Tableau, Power BI, AWS

PROFESSIONAL EXPERIENCE

Data Analyst Intern, JD.com, Inc., Beijing, China *Aug. 2022-Nov. 2022*

- Preprocessed user behavior data and applied **Ensemble Learning** to optimize a weighted fusion model with the existing **XGBoost** models to reduce operating cost and use the inventory in hand, achieving 5% improvement in forecast accuracy.
- Conducted the Exploratory Data Analysis (**EDA**) and trained Machine Learning models to analyze customer profile and boost coupon redemption rates, resulting in a formulated precise coupon delivery strategy to improve ROI by 20%.
- Designed and implemented **A/B Testing** on multiple controlled variable samples using Python to assess the significant impact of the promotional content of the landing page on customer behavior, which led to a 12% higher conversion rate.

Data Analyst Intern, CAS Institute of Geographic Sciences and Natural Resources Research, Beijing, China *Jul. 2022-Aug. 2022*

- Utilized **Python** Scrapy and Beautiful Soup to crawl 50,000+ comments on Ctrip and Mafengwo travel websites.
- Applied **Spark SQL** for data preprocessing and **Pyspark MLlib** to train **Logistic Regression**, **SVM**, **Decision Tree**, **GBDT** and **Random Forest** models. Established an automated sentiment analysis workflow in **AWS SageMaker** for real-time predictions.
- Leveraged **AWS** S3, Athena, and QuickSight to create personalized SQL queries and develop an interactive dashboard with WordClouds, HeatMaps and Tree Charts, and other data visualizations.
- Effectively communicated statistical insights to non-technical teams using **Gephi** for network analysis and topic modeling, as well as **Tableau** for data exploration and reporting, resulting in an 11% enhancement in tourist satisfaction.

Data Analyst Intern, Tencent Technology Co., Ltd, Beijing, China *Jun. 2020-Jul. 2020*

- Performed **Customer Segmentation** to improve marketing strategies through the use of a weighted **RFM** model and **K-means** clustering with **Python**, and refined operations based on behavior preferences, increasing ARPU by 17% monthly.
- Preprocessed 2 million user behavior data, built **AAARR** model and Funnel Charts for **Hypothesis Testing** using **SQL** to analyze user behavior trend based on different time periods and behavior routes, improving the conversion rate by 25%.

RESEARCH EXPERIENCE

Research Assistant, Beijing University of Technology, Beijing, China *Dec. 2022-May 2023*

- Utilized multimodal data from physiology, vehicles, and the environment to design a model based on **Transformer** and **Dynamic Graph Convolution** that extracted global and local features. Achieved real-time detection of driver emotions by a **Hybrid Attention** mechanism with dynamic weight allocation to fuse multimodal features.
- The light-weighted model was achieved through Multi-scale Depth Separable Convolution, which could reduce inference time by 15%. The Transformer-based model improved the generalization performance of the model and enabled cross-individual detection of driver emotions with 89% accuracy.

Research Assistant, CAS Research Center On Fictitious Economy & Data Science, Beijing, China *Aug. 2021-Sep. 2021*

- Applied **SQL** and **Python** to automate data cleaning and preprocessing on an unbalanced loan records dataset, made a Exploratory Data Analysis (**EDA**), and conducted **Feature Engineering** to transform features to improve forecasts granularity.
- Constructed a Probability of Default (**PD**) Model using **Logistic Regression**, **Decision Tree**, and **Random Forest** models, solved the problem of imbalanced data classification by using penalized learning algorithms.
- Selected features to guarantee computing efficiency, and performed Hyper-parameter Tuning to find the best threshold and improve the AUC value of the **PD model** to 0.86, with a 10% improvement from the credit risk baseline model.

COMPETITIONS

Team Leader, (Kaggle) Feedback Prize-English Language Learning (ELL) — ranked 98/2654 (Silver Medal) *Sep. 2022-Nov. 2022*

- Developed a Multi-dimensional Score Model in **Pytorch** for 8th-12th grade ELL essays, used **Semi-supervised Learning** to train the Multi-label Regression Model, and adopted **Ensemble Learning** with **SVR** to have higher predictive accuracy.
- Performed **Transfer Learning** with **DeBERTa**, **RoBERTa** and **ELECTRA** to train a small dataset, and extracted Pre-trained Contextualized Embedding to fine-tune parameters for downstream tasks. Implemented Average Pooling and used Layer-wise Learning Rate Decay to ensure the efficiency of gradient descent methods, obtaining a final MCRMSE score of 0.436108.