Xingyin Xu

201-978-7301| xingyinxu@gmail.com | linkedin.com/in/xingyin-xu | yinyin0916.github.io/

EDUCATION

New York University — **New York, NY**
Master of Science in Data Science — ***Aug. 2023-May. 2025 (Expected)***
GPA: 3.95/4.0
Relevant Courses: Machine Learning, Visualization for Machine Learning, Natural Language Processing, Computational Cognitive Modeling, Capstone, Big Data, Programming for Data Science, Practical Training for Data Science.

University of California, San Diego — **San Diego, CA**
Bachelor of Science in Probability and Statistics, Minor in Data Science and Economics — ***Sept. 2019-Jun. 2023***
GPA: 3.98/4.0 (**top 2%**)
Relevant Courses: Data Analysis, Data Structures, Data Management, Machine Learning, Recommender Systems, Mathematical Statistics, Stochastic Processes, Mathematics of Finance, Econometrics, Linear Algebra.

WORK EXPERIENCE

GardenStar Group — **Jersey City, NJ**
Data Scientist Intern — ***Sep. 2024-present***
- Scraped and analyzed 400+ real estate data in Hampton, GA and Wendell, NC using **Python Selenium**.
- Conducted a comprehensive rental market analysis and predicted rental prices using **Lasso Regression**.
- Collaborated with other business partners to analyze emerging hospitality market and investment opportunities for a luxury resort project in GA.

New York Life — **Tampa, FL**
Data Scientist Intern — ***Jun. 2024-Aug. 2024***
- Queried and analyzed 100000+ data using **Toad SQL**, ensured data quality and relevance for **300+ variables** through rigorous screening and exploratory data analysis.
- Employed three **feature selection** methods, incorporating both statistical and machine learning techniques, to evaluate the essential features for the final model.
- Built a robust marketing targeting model for direct mail using **Logistic Regression**, **Random Forest**, and **XGBoost** iteratively in **R**, final model enhanced the campaign effectiveness by **50%**.
- Collaborated closely with marketing team to derive actionable **business insights** from direct mail campaigns.

China Construction Bank — **Suzhou, China**
Data Analyst Intern — ***Jul. 2023-Aug. 2023***
- Developed and implemented a **VBA-based automated system** for daily deposit reports, enhancing the detection of anomalies and inconsistencies. Reduced manual workload by approximately **50%**.
- Created and deployed an interactive data dashboard using visualization techniques (line and pie graphs, slider controls) in **Tableau**. Improved customer relationship management, leading to **40%** increase in operational efficiency.

Halıcıoğlu Data Science Institute – UC San Diego — **San Diego, CA**
Instructional Assistance for Data Structures and Algorithms — ***Jan. 2022-Jun. 2023***
- Collaborated with faculty to craft course materials, ensured clarity and comprehensibility in assignments and exams.
- Led office hours each week and explain complex concepts into understandable segments in **Python** and **Java**.

Clear Creek Capital — **Los Angeles, CA**
Data Analyst Intern — ***Jun. 2022-Sept. 2022***
- Scraped and collected **50,000+ data points** on crude oil, gold, and treasury bonds using **Python BeautifulSoup**.
- Updated existing datasets with **SQL queries** for market sentiment reports, optimizing join operations and utilizing window functions to lower runtimes and costs.
- Applied robust intermarket analysis on macroeconomic cycles, refined profitability models and investment strategies.

PROJECTS

Mitigating Overconfidence in LLMs — ***Oct. 2023-Dec. 2023***
- Employed **Vicuna** and **prompt engineering** to generate verbalized confidence levels and answers of multiple choice and sentiment analysis datasets with **PyTorch**.
- Knowledge transferred from GPT-4 to Vicuna by finetuning, improved LLM accuracy, ratio of overconfidence bias, and ECE for **15 datasets**.
- Ensured accuracy of verbalized confidence levels and answers by data cleaning using **Python** Numpy and **OpenAi**.

Movie Ratings Analysis — ***Oct. 2023-Dec. 2023***
- Utilized **Mann-Whitney U tests** and found **significant** rating difference on engaged and not engaged audience.
- Built **Linear Regression Model** to predict movie ratings and utilized **grid search** to build **Lasso regularization**.
- Examined quality consistency of franchise movies using **Kruskal-Wallis H-test**, and utilized consistent franchise movies ratings to build regression model. Improved R square by more than **150%**.

SKILLS

Programming Languages: SQL, Python, R, Excel, VBA, Java, HTML.
Data Processing & Modeling: Pandas, Numpy, Pytorch, Pyspark, Scikit-learn, Seaborn, MongoDB, NLTK, matplotlib.