ROSY XU

201-978-7301 | rosyxu@nyu.edu | LinkedIn | Personal Website

SUMMARY

Data Scientist with a strong foundation in Machine Learning, Statistical Analysis, Predictive Modeling, and Data Visualization. Proficient in Python, SQL, and R to analyze complex datasets and uncover actionable insights. Skilled in using machine learning and statistical tests for data-driven decision-making.

Professional Experience: **New York Life Insurance, GardenStar Group, China Construction Bank, Clear Creek Capital**

EDUCATION

New York University **Aug 2023 - May 2025**
M.S. Data Science, GPA: 3.95 New York, NY
Relevant Coursework: Machine Learning, Visualization for Machine Learning, Natural Language Processing, Computational Cognitive Modeling, Capstone, Big Data, Programming for Data Science, and Practical Training for Data Science.

University of California - San Diego **Sep 2019 - June 2023**
B.S. Probability and Statistics, Minor in Data Science and Economics, GPA: 3.98 (**top 2%**) San Diego, CA

TECHNICAL SKILLS

Programming: Python (Pandas, Scikit-learn, PyTorch, Numpy, matplotlib), SQL, R, Excel VBA, Java, HTML.
Data Tools: Tableau, Advanced Excel, Databricks, Pyspark, PowerBI, Git, LaTeX, SAS, AWS, Snowflakes.
Data Science Methods: Machine Learning, Deep Learning, NLP, Statistical Analysis, Database Management, Data Visualization.

WORK EXPERIENCE

GardenStar Group **Sep 2024 - Dec 2024**
Data Scientist Intern Jersey City, NJ

- Analyzed emerging trends in the hospitality and real estate markets, and identified investment opportunities for luxury resort.
- Developed and automated **ETL pipelines** for over 400 housing records in Georgia and North Carolina via **Python Selenium**.
- Conducted rental market analysis and used **Lasso Regression** to predict rental prices and provide interpretable insights.
- Designed and implemented a scalable data pipeline using **AWS D3, Glue and Athena** to transform raw data into high-quality, analytics-ready datasets for market trends analysis.

New York Life Insurance **Jun 2024 - Aug 2024**
Data Scientist Intern Tampa, FL

- Collaborated with marketing team to deliver actionable business insights for direct mail campaigns. Built a robust marketing targeting model for direct mail using **Logistic Regression**. Final model enhanced the campaign effectiveness by **50%**.
- Queried and analyzed 100,000+ data using **Toad SQL**, ensured data quality and relevance for **300+ variables** through rigorous screening, exploratory data analysis, and data engineer.
- Employed a comprehensive **feature selection** method, incorporating both statistical, machine learning (**random forest**) and visualization techniques, to downsize model variables to 10+ for the final model.

China Construction Bank **Jul 2023 - Aug 2023**
Data Analyst Intern Suzhou, China

- 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 **Jan 2022 - Jun 2023**
Instructional Assistance San Diego, CA

- 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 **Jun 2022 - Sep 2022**
Data Analyst Intern Los Angeles, CA

- Applied robust intermarket analysis on macroeconomic cycles, refined profitability models and investment strategies.
- Collected 50,000+ data points on crude oil, gold, and treasury bonds using **Python BeautifulSoup**.
- Updated datasets with **SQL queries** for market sentiment reports, optimized operations by window function to lower runtime.

SELECTED PROJECTS

Mitigating Overconfidence in LLMs through Knowledge Transfer (Python PyTorch, OpenAI)

- 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 and ECE for 15 datasets.
- Ensured accuracy of verbalized confidence levels and answers by data cleaning using Python **Numpy** and **OpenAI API**.

Movie Rating Analysis (Python, Statistical Tests)

- Utilized **Mann-Whitney U tests** and found significant rating differences 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%**.