

Yu-An(Aaron) Chen

412-214-2260 • yuanche2@andrew.cmu.edu • [GitHub](#) • [LinkedIn](#)

Education & Skills

Carnegie Mellon University, GPA: 3.81/4.00

9/2020 - 12/2023

B.S. in Statistics and Machine Learning • Minor in Computer Science

Pittsburgh, PA

Programming Languages: Python, R, Apache Spark, SQL

Tools/Frameworks: dbt, Airflow, Databricks, Redshift, AWS, sklearn, Hightouch, Stitch, PowerBI

Skills: Data engineering, statistical analysis, feature engineering, time-series analysis, data visualization

Certifications: Financial Engineering, Risk Management, Finance/Financial Markets, AWS Technical Essentials

Work Experience

Analytics Engineer | Insurify

2/2024 - Present

Insurance Comparison Startup (600M Valuation)

Cambridge, MA

- Scaling 5+ departments' initiatives via engineering novel data transformations/pipelines, creating on-demand reporting, and building syncs between external platforms and internal data warehouse
- Directing MLOps in real-time ad bidding (\$9MM+ monthly spend) and customer LTV estimation via enhancing feature engineering and post-processing methods, scaling traffic volume by 65% in 5 months
- Managing 750+ dbt models and 50+ Airflow DAGs by conducting quality assurance, developing cross-referential tests, and implementing refactors to improve quality and robustness of existing tech stack

Teaching Assistant(s) | Carnegie Mellon University

8/2023 - 12/2023

Machine Learning Dept., Dept. of Statistics & Data Science

Pittsburgh, PA

- Facilitated [Machine Learning with Large Datasets](#) (graduate/PhD-level, 80+ students) by developing assignments and exams, grading student work, hosting office hours, and directing recitations
 - Key topics: Spark, distributed ML, GPUs, hashing, deep learning optimization, dimensionality reduction
- Facilitated [Modern Regression](#) (senior/graduate-level, 230+ students) by grading student work, hosting office hours, providing feedback, and proctoring exams
 - Key topics: regression theory, model inference & diagnostics, feature & model selection, regularization

Data Science Intern | Federated Hermes

5/2023 - 8/2023

Investment Manager (750B+ AUM)

Pittsburgh, PA

- Engineered 120+ features from sales and CRM data (19+ million rows, ~10.5 GB) with aggregation and NLP methods to determine optimal client contact strategies to increase mutual fund inflows
- Built and automated data pipelines with PySpark on Databricks to streamline ML development, established benchmark model performance of 93.7% accuracy predicting purchase activity & suggesting next steps
- Segmented client base based on purchase history with regression and determined ideal combination of interactions across client types for maximizing conversion rates
- Presented key findings to management, upgrading sales' targeted outreach & client retention capabilities

Machine Learning Intern | Behavior

5/2022 - 8/2022

Addiction Recovery via AI

Pittsburgh, PA

- Built frameworks for evaluating craving-predicting ML models through F-score and K-fold cross-validation, facilitating context-specific model evaluation
- Parallelized hyperparameter tuning process of CNN and XGBoost models by incorporating multiprocessing, decreased evaluation runtime by up to 800%
- Employed and tested multiple feature engineering techniques (Gaussian Smoothing, Autoencoding, etc.) to enhance model performance, increased cross-validation accuracy by 3.5%

Research Assistant | LearnLab

1/2021 - 5/2023

CMU Research Lab (School of Computer Science)

Pittsburgh, PA

- Developed data pipelines & reporting involving NLP, mixed-effects modeling, and hypothesis testing to reveal associations between course material presentation methods and student performance
- Led the data analysis team of Podsie, an educational start-up, and analyzed the efficacy of its products under low-sample size constraints and presented results to board of directors
- Presented findings to research leads and grant reviewers, leading to renewed lab funding in both review cycles

Personal Projects

[Optimal ETF Allocation Solver \(R\)](#) • [Neural Wavelets for Time Series Compression \(Python\)](#) •

[Distributed ML with Million Song Dataset \(PySpark\)](#) • [Government Bias in Affordable Housing \(R\)](#)