Yifei Chen

443-418-3210 | ychen780@jh.edu | Baltimore, MD | LinkedIn

CAREER ASPIRATION

2+ years data professional with a strong foundation in mathematics and statistics, experienced in leveraging Python and machine learning techniques to extract actionable insights from complex datasets. Proven results improvements in model accuracy and insight deliveries to cross-function teams in both tech and non-technical settings

EDUCATION

Johns Hopkins University, Whiting School of Engineering — Baltimore, MD
Master of Science in Engineering in **Applied Mathematics and Statistic** — Sep. 2025-Dec. 2026
Concentration on Machine Learning and AI model optimization
Core Courses: Introduction to Data Science, Mathematics of Data Science, Introduction to Convexity

China Agricultural University — Beijing, China
Bachelor of Science in **Mathematics and Applied Mathematics |** 3.47/4.00 — July 2025

TECHNICAL SKILLSET

Programming Languages & Tools: Python (Pandas, NumPy, Matplotlib, Scikit-learn), MATLAB, C++, Latex, R, Excel (VBA), Wind, Stata, Looker Studio, SQL (BigQuery, AWS)
Analytics Techniques: Classification & Clustering Modeling (K-Means, Decision Trees, Random. Forest), Linear/Ridge Regression, PCA & Factor Analysis, Data Visualization, A/B Testing, LLM

PROFESSIONAL EXPERIENCE

Financial Data Analyst — July 2023-Aug. 2023
Lakefront Asset Management CO., LTD (Annual managed assets exceed 2 billion USD) — Beijing, China

- ✧ **Reporting & Analysis:** Built 3 financial statements, and delivered reporting (ROA, ROE, P/E) across 31 industries to support senior management in strategic decision-making through data-rigorous insights
- ✧ **Portfolio Modeling & Optimization:** Developed a Python-based quantitative model based on Markowitz Portfolio Theory, and applied to data from 10,000+ funds, which identified optimal portfolios and increased the Sharpe ratio by 20% compared to the market benchmark, enhancing investment decision-making
- ✧ **Data-Driven Cluster Analysis:** Utilized unsupervised machine learning models (Elbow Method/ Hierarchical/K-Means Clustering) to develop industry tailored metrics, consolidating 31 raw industries into 7 distinct verticals to reduce noises and pinpoint key insights to inform macro-economic analysis

DATA SCIENCE PROJECT

Data Scientist/Project Manager | Authored and Published in China Swine Industry — Nov. 2022-Dec. 2024
Model Innovation*: Two-Stage Prediction Method for Pig Growth Traits Based on Group Feature Selection*

- ✧ **Objective:** Addressed the limitations in traditional genomic selection to model SNP data, spearheading a novel two-stage method called genetic data group effect to improve pig breeding value accuracy
- ✧ **Data Processing:** Built a Python pipeline to process and integrate high-dimensional genomic datasets (PIC, S21, S22), covering 9,400+ real-world pig samples and 120,000+ SNP markers, to predict 5+ complex traits with varying heritability
- ✧ **Modeling Building & Analysis:** Developed a two-stage ML pipeline combining K-means clustering of SNP-grouped genetic data with Kernel Ridge Regression, reducing computational complexity by 60% while simultaneously improving prediction accuracy
- ✧ **Insights & Economy Impact:** Benchmarked the new model against 8 methods (GBLUP, SVR, LASSO, etc.), achieving a 4% accuracy gain over industry-standard GBLUP and identifying 5+ key traits to improve pig breeding success rates, with potential to enhance efficiency and productivity in the Chinese swine industry, which accounts for approximately 5% of China's GDP and over half of global pork production

2nd Author | Published in EAI WiSATS 2024 Proceedings — May 2024-Aug. 2024
Algorithm Optimization: *Semi-Supervised Multi-View Dimensionality Reduction*

- ✧ **Algorithm Development**: Designed a novel multi-view semi-supervised projection (MSSP) algorithm by introducing a new regularization term based on a similarity-correlation metric to effectively integrate incomplete, high-dimensional datasets
- ✧ **Model Optimization & Evaluation**: Evaluation & Impact: Tested on YALE, ORL, COIL-20 datasets using Python, achieving 12.4% higher classification accuracy over 7 benchmarks and demonstrating robust feature extraction for multi-view, partially labeled data
- ✧ *EAI WiSATS: EAI International Conference on Wireless Innovations and Advances in Satellite Communications*