Yiming (Tim) Zhang

yimingzh@umich.edu | (213)-343-8125 | Ann Arbor, MI

EDUCATION

University of Michigan - Ann Arbor Ann Arbor, MI
M.S. in Quantitative Finance and Risk Management (GPA: 3.8/4.0) Aug. 2024 – May. 2026 (Expected)
- Courses: Real Analysis, Stochastic Analysis, Stats for Fin Data, Math Algo Trading, Financial Mathematics I/II, Data Sci in Python

University of Edinburgh Edinburgh, UK
M.S. in Statistics with Data Science (GPA: 3.44/4.0) Sep. 2021 – Nov. 2022
- Courses: Machine Learning/ Pattern Recognition, Multivariate/ Bayesian Data Analysis, Optimization Method in Fin, Time Series

University of Nottingham Ningbo, China
B.S. in Mathematics and Applied Mathematics (GPA: 3.69/4.0, First Class w. Honors, Dean's Scholarship) Sep. 2017 – Jun. 2021
- Courses: ODE, PDE, Math Analysis, Topology and Metric Spaces, Mathematical Finance, Optimization, Statistics, Probability

RESEARCH INTEREST

- Stochastic Differential Equation, mean-field games, queueing theory, control- and game-theory

RESEARCH EXPERIENCES

LLM – Detect AI Generated Text
University of Michigan Oct. 2023 – Jan. 2024
- Built leakage-safe pipeline with prompt-grouped K-fold, train-only normalization, and adversarial validation to control drift; engineered snapshot datasets for code-competition constraints. Balanced log loss reduced by 12–18% versus a strong TF-IDF + logistic regression baseline, with stable out-of-time curves.
- Engineered phrase-centric and lexical features inspired by public winning approaches: curated phrase dictionaries and collocations, TF-IDF over char and word n-grams, token/word/char length stats, capitalization and punctuation ratios; trained linear SVM and L2-regularized logistic regression with isotonic/Platt calibration, and stacked/blended the models. AUC improved by 3–5 points, calibration error decreased, top-decile precision improved under fixed recall.
- Productionized inference for speed and reproducibility: scikit-learn pipelines with frozen vectorizers, versioned preprocessing artifacts (joblib), deterministic seeding, lightweight rank-averaging ensembling; CPU-only scoring suitable for batch/API deployment. End-to-end inference latency under 50 ms per essay on CPU, reproducible scores across notebook re-runs, and consistent LB/CV alignment under grouped folds.

The Impact of DNA Methylation Sample on Forecasting Human Ages
University of Edinburgh (Thesis, Advisor: Dr. Bruce J. Worton) May. 2022 – Jul. 2022
- Formulated age prediction from high-dimensional DNA methylation profiles (on the order of 10^2 subjects and 10^5 CpG sites) as a penalised regression problem, using elastic net to handle strong feature correlations and enforce sparsity.
- Split the data into training and held-out testing cohorts; on the training set, implemented a leakage-free leave-one-out cross-validation pipeline (normalisation, feature selection, model fitting) for hyperparameter tuning.
- Achieved competitive out-of-sample performance on the test set, with mean absolute error on the order of 3–4 years, root mean squared error of similar magnitude, and coefficient of determination R^2 typically in the range 0.80–0.85.
- Compared alternative feature-selection strategies to identify a compact subset of age-informative CpG sites, reducing the feature space by over an order of magnitude while keeping test-set MAE degradation below 0.2 years.
- Fitted linear, quadratic, and cubic basis expansions of age within the elastic-net framework; analysed residuals, calibration plots, and influential outliers to characterise bias–variance trade-offs and potential biological age acceleration effects.

Potential Solution of an Aspect of the Derrida–Retaux Model
University of Nottingham (Advisor: Dr. Vladimir Toussaint) Sep. 2020 – Oct. 2020
- Reviewed spin-glass, Ising and hierarchical depinning models and the max-type recursive Derrida–Retaux system, reformulating the conjecture in a probabilistic framework via the recursion, the free energy F_∞ and its critical manifold.
- Implemented the Derrida–Retaux recursion and associated moment generating functions in MATLAB, running large-scale Monte Carlo experiments to estimate the free energy numerically and to visualise the Berezinskii–Kosterlitz–Thouless-type depinning transition under different initial laws.
- Empirically investigated the sustainability probability $P(X_n > 0)$ and the critical decay $P(X_n > 0) \approx n^{-2}$, as well as conditional distributions $P(X_n = k \mid X_n > 0)$, and compared simulation output with recent rigorous results and conjectured scaling limits.
- Exploited the hierarchical tree / open-subtree representation of the model to design resampling strategies on open paths, linking the recursion to branching processes, recursive distributional equations and ideas that naturally appear in stochastic analysis and control of random media.
- Summarised the literature on Derrida–Retaux-type models in a short internal note, highlighting universality classes (finite-moment vs heavy-tailed regimes), possible continuous-time limits and how similar techniques could extend to controlled stochastic systems and SDE-based models.

PROFESSIONAL EXPERIENCES

Qinyuan Private Equity Fund ($3B AUM) Shanghai, China
Data Scientist Intern (Python/Pandas/PyTorch/CNN/MLP) Jun. 2025 – Aug. 2025
- Built a multi-factor alpha framework (20+ OHLCV signals: momentum, turnover & volume families, K-Bar, MACD, market cap, etc.) large-scale feature engineering pipeline processing 15M+ data; implemented panel operations with leakage-safe transforms, train-only standardization, winsorization, and date-stamped snapshots for full reproducibility.
- Engineered K-fold training/validation harness with permutation-/SHAP-based feature importance to guide architecture and regularization choices; the resulting structure (tuned lookback, weight decay, early stopping) improved log-loss by 18%, raised AUC by 4–6 pts, enhanced calibration, translating into higher top-decile signal lift and smoother PnL under transaction-cost controls.
- Developed CNN-MLP hybrid architecture for time-series prediction using PyTorch; implemented custom loss functions, AdamW optimization with gradient clipping, and early stopping; achieved 48% MSE reduction compared to baseline models (XGBoost, LSTM) through systematic hyperparameter tuning.

China Merchants Securities (National top securities) Beijing, China
Applied Scientist Intern (Python/Pandas/Pytorch/LSTM/Autoencoder) Jun. 2024 – Aug. 2024

- Built a leakage-safe multi-factor asset-pricing pipeline: factor curation (value, quality, momentum, size, volatility), winsorization and z-scoring, industry and market-cap neutralization, and time-ordered rolling splits with walk-forward evaluation; ex-post results: cross-entropy/log-loss reduced by 15–20% versus baseline, AUC increased by 3–5 points, calibration error decreased, and ranking quality (NDCG@10) improved, delivering higher top-decile signal lift under turnover and slippage constraints.
- Designed an Autoencoder + LSTM to compress cross-sectional factors and capture temporal dynamics, guided by permutation and SHAP feature importance to prune noisy inputs; results: feature dimensionality reduced by about 70% while preserving about 95% information (variance or mutual information), validation AUC improved by 4–6 points with better stability across folds; deployed batched CPU serving (ONNX/FastAPI) with median inference latency under 100 ms.
- Ran rolling out-of-sample evaluation with full diagnostics (prediction versus actual, residuals, drift, capacity and turnover, drawdowns) and ablation-driven model selection (latent dimension, lookback, weight decay, early stopping); outcomes: MAE reduced by 10%, R^2 improved by 3%, directional accuracy at 53–55%, lower maximum drawdown, and smoother PnL under transaction-cost controls; shipped reproducible reports and stability plots.

Bohai Securities (Government-backed comprehensive securities) Beijing, China
Machine Learning Engineer Intern (Python/Pandas/PyTorch/Transformers/LIT) Jun. 2023 – Nov. 2023
- Built high-throughput news NLP pipeline for 100k+ daily Wind articles with share filtering, headline–body merging, spaCy/NLTK preprocessing, and snapshot "time-travel" datasets to eliminate training–serving skew; added adversarial validation to detect distribution shift and late-arrival handling to prevent leakage. Results: cleaner training data, lower drift alerts, and log-loss reduced by about 8–12% versus naive chronological splits.
- Fine-tuned a BERT sentiment model using Hugging Face Transformers/Datasets with class-imbalance handling (weighted loss), early stopping, and threshold calibration; audited model behavior with Google LIT for salience/attention checks. Results: out-of-sample accuracy 0.9826, AUC 0.9746, F1 improved by about 3–4 points. Deployed a REST API (FastAPI + ONNX Runtime) with autoscaling; sustained 1000 QPS throughput with median latency under 100 ms.
- Ran architecture and loss-function ablations: GRU and CNN baselines, SCINet-style temporal blocks for long-context aggregation, and a ticker-aware dual-encoder to enrich entity embeddings; compared cross-entropy, focal loss, and label smoothing with temperature scaling for better calibration. Results: best variant improved AUC by about 2–4 points, PR-AUC by about 3–5 points, expected calibration error reduced, and a distilled small model cut memory by about 30% with comparable accuracy.

Volunteer

Laligurash Bright Future School Dolakha, Nepal
Physics Teacher Feb. 2018
- Designed and organized four modules in physics, including lecture, experiment, and examination.
- Constructed a scoring system which reflects students' performance in classes; held revision seminar for students.
- Held future-oriented lectures including the basic concepts for career path, the importance of education, etc.

Skills

- Programming: Python(6 years), SQL(6 years), R(8 years), Matlab(8 years), C++(2 years), C(11 years), C#, Java, LaTeX.
- Machine Learning: Supervised Learning (Linear/ Logistic Regression, SVN, Tree-Based Models, Gradient Boosting), Unsupervised Learning (K-Means, PCA, DBSCAN), Deep Learning (CNN, RNN, LSTM, Transformer), NLP (BERT, TF-IDF, NLTK, GPT)
- Cloud & Tools: AWS (Redshift, RDS, SQS, IAM, KMS), Google Cloud (BigQuery), Docker, Spark, Kafka, Airflow, Hadoop, PyTorch, TensorFlow, Keras, scikit-learn, Hugging Face Transformers, ONNX, MLflow, TensorBoard, NumPy, Pandas, SciPy, OpenCV, Git, Jenkins, GitLab CI, VS Code
- Databases: MySQL, MongoDB, Oracle Database, SQLite, Amazon Redshift, Snowflake, Redis
- Certifications: CFA Level II Candidates, FRM Level II Candidates