Zhehan Shi

zs1113@nyu.edu | (917) 703-6627 | linkedin.com/in/zh-shi | Personal Projects cyberzzhhss.github.io/projects
54 Swedesford Rd, Malvern, PA, 19355

EDUCATION

New York University | *New York, NY* *Sep. 2021-May 2023*
M.S. in Data Science | **GPA:** 3.75/4.00
Relevant Courses: Algorithmic Trading, Deep Learning, Big Data, Natural Language Processing, Computer Vision, Causal Inference.

New York University | *New York, NY* *Sep. 2017-May 2021*
B.A. in Computer Science and **B.A**. in Mathematics | **GPA:** 3.59/4.00 | Minor in Business Studies
Relevant Courses: Machine Learning, Object-Oriented Programming, Parallel Computing, Partial Differential Equations

TECHNICAL SKILLS & CERTIFICATIONS

ML & Statistic Analysis Skills: Deep Learning Models, Regression Models, Decision Tree Models, Clustering Models, Time Series Models, Cross Validation, Bayesian Statistics, A/B Testing, Hypothesis Testing, Data Visualization, Exploratory Data Analysis
Programming Languages: Python (PyTorch, TensorFlow, Scikit-learn, Matplotlib), SQL, NoSQL, R, JavaScript, Hive, Spark, Hadoop
Platform & Tools: Jupyter, Git, MySQL, R Studio, Power BI, Tableau, Linux, AWS, Docker, LaTeX, Microsoft Office

RESEARCH EXPERIENCE

Data Scientist, The Vanguard Group, Malvern, PA [*CatBoost* | *Gradient Boosting*] *Oct. 2023-Current*
- Built a high-performance loan-level prepayment model for 30-year fixed-rate mortgages using CatBoost with an 8-GPU setup, processing over 150+ GB of data from 2014 to 2023 and outperforming the YieldBook model.
- Optimized data handling with Polars for Amazon S3 streaming and implemented memory-saving techniques, reducing data footprint by over 40%.

Researcher, NYU Courant Institute of Mathematical Sciences, New York, NY [*TensorFlow* | *Deep Learning*] *Feb. 2022-May 2022*
- Developed a deep learning framework in TensorFlow for pricing financial instruments, basket options of 10 stocks, by solving high-dimensional stochastic differential equations.
- Incorporated transactional costs into neural network model trained on 20,000 mini-batches with a 512 batch size to achieve state-of-the-art model, overcoming curse of dimensionality.

ACADEMIC PROJECTS

Modeling Covariance Matrix Estimators Performance via Markowitz Portfolio, New York University [*Python*] ***GitHub***
- Compared 4 covariance matrix estimators, including Exponential Weighting Covariance Estimator.
- Processed a subset of 30GB+ millisecond-level high-frequency trades data to construct custom-built 5-min sliding window covariance matrix estimators.
- Achieved insightful conclusions on Optimal Shrinkage Estimators for daily and high-frequency situation.

Handwritten Digits Recognition Web Application, Individual Project [*PyTorch* | *Streamlit*] ***Demo***
- Constructed a deep learning model using Convolutional Neural Network (CNN), trained on 60,000 28x28 grayscale images of single digit handwriting dataset.
- Achieved 99.17% validation accuracy on handwritten digit recognition.
- Deployed on public cloud platform Streamlit for interactive user digit sketching and recognition.

Modeling the Price Impact of Large-Scale Trades Impact Model, New York University [*Regression* | *Python*] ***GitHub***
- Processed over 100GB+ of 3-month, millisecond-level, high-frequency NYSE trades and quotes tick data from over 1000 tickers.
- Built an Almgren-Chriss market impact model to understand the execution of large-scale trades on the NYSE.
- Obtained useful parameters using parametrized non-linear regression.

Question Answering for Reading Comprehension via NLP, Individual Project [*Transformer*| *BERT* | *PyTorch*] ***Demo***
- Conducted data pre-processing pipeline such as that included tokenization of questions and context, handling long contexts using stride, and mapping correct answer positions into tokenized sequences.
- Finetuned a pre-trained Transformer model, BERT, for a question-answering task on SQuAD (Stanford Question Answering Dataset) dataset, consisting of over 107,000 question-answer pairs.
- Evaluated BERT on the question-answering task from a given context with F1 score of 88.65% and an exact match score of 81.04%.
- Published the model on Hugging Face platform for interactive access.

Personalized Movie Recommendation System, New York University [*Apache Spark*] ***GitHub***
- Developed a personalized movie recommendation system using Model-Based Collaborative Filtering.
- Utilized Latent Factor Model integrated with Alternating Least Square (ALS) Matrix Factorization.
- Engineered the parallelized Spark Machine Learning model on a dataset of 25 million movie ratings from MovieLens and attained Root Mean Square Error (RMSE) of 0.819 during grid search.

Object Detection and Classification in Everyday Images, New York University [*PyTorch*] ***GitHub***
- Constructed an object detection model using RetinaNet for classifying items within everyday images.
- Leveraged VICReg for pretraining the model on 512,000 unlabeled images, and performed finetuning on 30,000 labeled images.
- Achieved a mean Average Precision (mAP) score of 0.154 on the test dataset.