Yin Zhang

(509) 339-3262
yin.zhang@wsu.edu

Data Scientist / Economist / PhD Student

GitHub: YinZhang0810
LinkedIn: yinzhang0810

Education

Ph.D. Economics, *Washington State University* May 2023[Expected]
M.S. Statistics, *Washington State University* May 2023[Expected]
M.S. Mathematics, *Washington State University* May 2020
M.A. Financial Economics, *University of San Francisco* May 2017
B.A. Finance (2nd Major: Management Operation), *Washington State University* May 2014

Expertise

Time-Series Forecast; Anomaly Detection; Natural Language Processing;
Data Mining; Data Visualization; Predictive Modeling; Causal Inference

Skills

Tools and Languages	Python, R, SQL, SAS, Stata, LaTeX, MarkDown
Packages	Pandas, NumPy, NLTK, scikit-learn, PyOD, StatsModels, Matplotlib, Git
Communication	English(Fluent), Chinese (Native)

Technical Experience

Incident Topic Similarity Matching and Its Application to Accounting Information Retrieval [Python] **February-May 2022**
Microsoft *Remote*

- Performed dictionary build, frequency count, and supervised classifier with incident data from Microsoft Enterprise Data Lake House.
- Applied NLTK POS tagger. Evaluated the improvement of classification accuracy by removing irrelevant words.
- Adopted Sentence-Bert for sentence classification. Adopted Facebook fast text for multi-intent classification.

Anomaly Detection and Future Usage Forecasting on Storage [Python] **February-May 2022**
Microsoft *Remote*

- Performed Exploratory Data Analysis (EDA) for daily time series data obtained from Microsoft Enterprise Data Lake House. Visualization were plotted using matplotlib and plotly.
- Detected anomalies using PyOD API (Median Absolute Deviation, K Nearest Neighbors, Local Outlier Factor, Isolation Forest)
- Per-processed data by replaced the anomalies and missing values.
- Predicted future storage capacity, read/write operation size using (1) ARIMA with per-processed data (2) Prophet with change points detected. Selected the best combination of a grid of parameters via Time-Series split cross-validation.

NLP Sentiment Analysis on Prediction of Financial Asset Return [Python] **Spring 2022**
Dissertation Chapter *Washington State University*

- Scraped large sets of text data from Twitter (10 million+ observations) to investigate the relationship between online investor sentiment and cryptocurrency returns.
- Applied tokenization and normalization by Natural Language Processing (NLP) technique to clean text data.
- Classified sentiment by both unsupervised-learning and supervised-learning approaches, including logistic regression, SVM, and own-developed lexicon-based tool. Assigned polarity scores of sentiment for each Tweet.
- Adopted VAR and RNN methods to forecast multivariate Time-Series data. The significant predictive power of Twitter sentiment on cryptocurrency returns was obtained.

Pricing European Options Under Heston Model [R] **Spring 2020**
Project of M.S. Mathematics *Washington State University*

- Estimated parameters by Maximum Likelihood, calculated option price, checked validation and discussed properties of Heston model.
- Increased the forecasting accuracy by 13%, compared with classical Black-Scholes option pricing model.

Project of PNW Container Optimization Model **Fall 2019**
Research Assistant *Washington State University*

- Collected, cleaned, prepped data using Stata, SQL, and Excel to help Professor build container rental and transportation optimization model.
- Contributed towards the development of optimization model and conducted econometric analysis on operational costs reduction problem.

Other Course Projects

> Credit card fraud detection using supervised classifiers > Forecasting exchange rate by ARIMA model