Pranab Islam

My Website ◦ GitHub ◦ LinkedIn
pfi203@nyu.edu
M: (917) 574-0680

Education

Sep 2020 – Jan 2023 **New York University** *New York, NY*
Master of Science in Data Science | GPA: 3.88

Skills

Languages:	Python ◦ SQL ◦ JavaScript
ML Modeling:	PyTorch ◦ Hugging Face ◦ scikit-learn ◦ Pandas ◦ Dask ◦ DataRobot
Cloud / ML Ops:	Airflow ◦ Prefect ◦ PySpark ◦ AWS (SageMaker, Lambda) ◦ GCP ◦ Docker
Version Control / Other:	Git ◦ Linux / Bash ◦ Looker ◦ React ◦ Next.js ◦ Hadoop ◦ LangChain

Professional Experience

Mar 2023 – Present **Bardess | Data Scientist** *New York, NY*
Data Science Team

- Created end-to-end multivariate time series training / inference pipeline (~10,000 lines of code) to predict financial line items using DataRobot's AutoML tool
  - ~20% improvement from financial planning organization's line item estimates
  - Wrote a rich set of software features to run ML feature selection, experimentation, and deconstruct the target time series into 81 segments to model separately and aggregate afterward

Aug 2023 – Nov 2023 **Patronus AI | Machine Learning Engineer (Contractor)** *New York, NY*
Co-led the creation of a real-world financial question answering (QA) dataset. Paper here

- Led the labeling efforts as Patronus's finance expert by managing a team of 20 annotators and offering strict guidelines for question and answer construction, eventually creating 10,231 QA pairs
- Single-handedly created a programmatic labeling pipeline that generated 78% of the dataset (the other 22% was completely human-made from annotation guidelines).
- Conducted benchmarking, evaluations, ablations, and data analysis which involved running various state of the art models (GPT-4 Turbo, Llama2, Claude2) and retrieval configurations

Sep 2022 – Dec 2022 **Cash App | Machine Learning Modeler Intern** *New York, NY*
Search & Discovery Machine Learning Engineering Team

- Developed query intent model using XGBoost to classify user search queries
  - Created daily training / feature engineering pipeline from logging data using SQL and Airflow
  - Deployed model to run batch inference daily via Prefect. 95% ROC-AUC achieved with a precision-recall AUC of ~50% on 1.5 months of post-training data
  - Constructed and back-tested two low-latency model approximators (using embeddings from matrix factorization) to deploy for real-time customer search

Jul 2019 – Jul 2020 **Mizuho Securities | Investment Banking Analyst** *New York, NY*
Financial Sponsors Group

Machine Learning Publications and Research

Sep 2022 – Present **MarkupMnA – MultiModal Legal Contract Document Segmentation**
Constructed an open-source multimodal (HTML + natural language) document segmentation dataset and ran benchmarking with (MarkupLM, XDoc & RoBERTa). Model weights, dataset, and results here

- Co-first author for paper submitted to various legal and NLP conferences

Feb 2022 – May 2022 **Analyzing Bagging Methods for Large Language Models**
Research project analyzing whether bagged ensembles of large language models could outperform single language model baselines, holding model parameter count constant. Detail and results here

- Developed an automated pipeline that fine-tuned large language models, created various bagged ensembles of them, and evaluated ensemble performance using SuperGLUE benchmark

Projects

Bayesian Multivariate Time Series Forecasting ◦ Classifying and Clustering Metropolitan Areas
Music Recommender System ◦ Capitalizing on Mispriced Odds in NBA Betting