BINFENG XU

+1 (336) 744-6809 || bx2010@nyu.edu || in

EDUCATION

New York University Sep 2020 – May 2022
- Master of Science in Data Science;
- Relevant Coursework: Deep Learning (*Yann Lecun*), Natural Language Processing & Understanding.

Wake Forest University Sep 2016 – May 2020
- Bachelor of Science in Computer Science (Honors) and Mathematical Statistics (Honors);
- Relevant Coursework: Data Structure & Algorithms, Machine Learning, Computer Vision, Parallel Programming, Numerical Computation, Mathematical Statistics, Probability Theory, and Optimization.

HONORS & CREDENTIALS

Kaggle: Competitions Master (*Global Rank: #805 of 166,227*) May 2018 – Jun 2021
- Santander Customer Transaction Prediction (Banking, Classification): Rank #24 of 8,802 (Gold);
- Santa 2020 Contest (Competitive Reinforcement Learning): Rank #17 of 792 (Silver);
- BirdCLEF 2021 Birdcall Identification (Signal Processing, Time-Series): Rank #15 of 816 (Silver);
- Predicting Molecular Properties (Scientific, Time-Series): Rank #182 of 2,749 (Bronze).

ACM ICPC: Regional 4th Place in North Carolina Mar 2019 – Mar 2019

Udacity Nanodegree in Deep Learning Sep 2018 – Oct 2018

PROFESSIONAL EXPERIENCE

Data Scientist Intern @ eBay May 2021 – Aug 2021
- Built and deployed a classification model to detect fraudulent sellers who mislead buyers into cancelling orders to avoid stock-out penalty, based on M2M chat messages and other item/order/user features;
- Queried, analyzed, and pre-processed data from database; designed modeling pipeline and targets for labeling;
- Trained and fine-tuned a two-level classification model consisted of 1) a RoBERTa sentiment classification with M2M messages and 2) a Gradient Boosting Decision Trees model for predicting fraud activities with item/order/user level features; Balanced Accuracy improved from 78.4% to 92.7%;
- Created API and documentation for model deployment, including database querying, Docker image, training/retraining, inference, auto-hyperparameter tuning, performance comparison, and pipeline testing; tested the model on eBay's AI cluster and worked with web service to design output usage at the production level.

Student Researcher @ New York University Nov 2020 – May 2021
- Project title: *Prediction and Policy-Learning under Uncertainty* Advisor: *Yann LeCun*
  - Modified and trained a policy model of autonomous vehicles while storing agent states at each timestamp;
  - Programmed environmental cars in dense traffic to update their trajectories based on agent states, in order to avoid crushing into the agent; average success rate of agents increased by 2.3% in our primary traffic simulator.

Teaching Assistant @ New York University Sep 2020 – Dec 2020
- Gave lab lectures and led practice sessions for a class of 26 (*Fundamental Data Science*) for Fall 2020.

Student Researcher @ Wake Forest University Sep 2018 – Dec 2019
- Project title: *Neural Activity Tensor Decomposition and Data Mining* Advisor: *Grey Ballard*
  - Developed a Tucker Decomposition-based algorithm to reduce the time complexity of decomposing neural Functional Magnetic Resonance Imaging (f-MRI) data by 60+ times and memory usage by 12,000+ times;
  - Built classification and clustering models to compare model capacity at different f-MRI granularity levels.
- Project title: *Object Recognition in Peru Forest* Advisor: *Paul Pauca*
  - Trained image classification models on mining sites with our original drone-taken images from Peru;
  - Programmed drones to detect and trace local mining activities with YOLOv3 and A* search algorithm.

Data Scientist @ iesLab (Shanghai) Mar 2018 – Aug 2018
- Collected data and built time-series models to predict real-time electricity prices and local industrial load usage;
- Optimized custom loss to tune Recurrent Neural Network (RNN) and Gradient Boosting Decision Trees; stacked multiple model outputs and achieved a theoretical saving of industries' annual power cost by 18.3%;
- Created API in Java to embed models and the power-purchasing recommendation algorithm into cloud service.

Data Scientist @ Honeywell (Beijing) Apr 2017 – Jun 2017
- Produced proof of concept to test feasibility of redistributing customer data to the Cloudera Hadoop ecosystem.

TECHNICAL SKILLS

Areas of Strength: Machine Learning, Data Analytics, NLP, Computer Vision, and Reinforcement Learning.
Programming Skills: Python3*, Pytorch*, C++, SQL, PySpark, C#, JAVA, Matlab, R, and Linux Programming. [1]

[1] Last update of document: June 2022