Pranab Islam

M: (917) 574-0680
pfi203@nyu.edu
github.com/pranabislam
linkedin.com/in/islampranab

Education

Sep 2020 – Dec 2022 **New York University** *New York, NY*
Master of Science in Data Science | GPA: 3.85
Coursework: Deep Learning, Natural Language Understanding, Machine Learning, Big Data, Probability & Statistics, Optimization, Linear Algebra, A/B Testing

Sep 2015 – Jun 2019 **The University of Chicago** *Chicago, IL*
Bachelor of Arts in Economics | GPA: 3.81

Professional Experience

Jun 2022 – Present **Roku** *San Jose, CA*
Data Science Intern, Core Analytics Team
- Constructed a last touch attribution model which allowed the team to understand how various web traffic channels relate to customer purchasing behavior
  - Executed project from end-to-end (data gathering, cleaning, model-building, visualizations / reporting, productionizing via Airflow, and future model improvement prototype creation)
- Supplied analysis and documentation to engineers to remedy bugs found in data logging process

Jun 2021 – Nov 2021 **Algorand** *Boston, MA*
Data Science Intern, Product Team
- Created the first weekly clustering system that grouped 10,000+ decentralized apps on Algorand's blockchain using set similarity on application metadata. This allowed Algorand to use common cluster characteristics to collaborate more systematically with app developers
- Designed the first iteration of event tracking metrics with seven separate event types for Algorand's developer documentation website. These metrics enable better data-driven decision making
- Led weekly product analytics meetings by creating presentations of various data visualizations and in-depth analyses of the Algorand blockchain network and user activity for 15 separate weeks

Feb 2021 – May 2021 **Global Association of Risk Professionals** *Jersey City, NJ*
Data Analyst Intern
- Developed a pipeline in Python to parse large volumes of non-standardized financial pdf files, extract relevant tabular data, and output the results in a shared repository for the data visualizations team

Technical Projects

Feb 2022 – May 2022 **Analyzing Bagging Methods for Large Language Models**
Natural language research project analyzing whether various bagged ensembles of large language models could outperform single language model baselines, holding model parameter count constant; project detail and results: https://arxiv.org/abs/2207.09099
- Developed an automated model fine-tuning, bagging, and evaluation pipeline that created various bagged ensembles of DeBERTa, GPT2, RoBERTa and compared performance across three of the eight natural language understanding tasks in the SuperGLUE benchmark
- Determined that (i) bagged ensembles generally do not perform better than single language models when holding model size constant, (ii) lightly pruning models in ensembles could be beneficial for certain use cases, and (iii) prediction variance generally decreases when bagging

Apr 2021 – May 2021 **Music Recommender System**
Implemented a recommender system with the Million Song Dataset using collaborative filtering via alternating least squares in PySpark; achieved a mean average precision of 30x the popularity baseline model (0.08 vs 0.0026); project detail and results: https://tinyurl.com/musicRecSys

Nov 2020 – Dec 2020 **NBA Underdog Matchup and Betting Analysis**
Developed tree-based classifiers and conducted the entire extract, transform, load process in order to analyze how to create profitable betting positions focusing on significant NBA underdog matchups; achieved an investment return of ~9%; project detail and results: https://tinyurl.com/nbaUnderdogs

Skills

Python (NumPy, Pandas, scikit-learn, PySpark, PyTorch, Transformers), Airflow, SQL, MapReduce, Hadoop, Dask, Tableau, Google Cloud Platform, Google Analytics, Linux, Bash, Git