Surya Dutta

708-712-9625 · suryabrata.dutta@gmail.com · linkedin.com/in/suryadutta · github.com/suryadutta

Education

University of California, Berkeley — Berkeley, CA
Master of Data Science (MIDS) | GPA 3.88/4.00 — *Jan. 2019 – Apr 2021 exp*

Yale University — New Haven, CT
Bachelor of Science, Physics (Intensive) | GPA 3.74/4.00 — *Aug. 2014 – May 2018*

Experience

McMaster-Carr Supply Company — Elmhurst, IL

Lead Data Engineer, Measuring the Digital Strategy — *Feb 2020 – Present*

- Productionized ETL pipelines and Neo4j graph databases to capture customer interaction data
- Architected and implemented design to capture exposure metrics and tie back to specific internal publishing data
- Present regularly to company leadership and stakeholders using data narratives and visualizations
- Developed suite of post-processing jobs to extract standardized OKRs from diverse customer search patterns. Built a REST API to dynamically retrieve these metrics, filtering by keywords, datetime range, and metadata
- Led department-wide training sessions and office hours on Neo4j and graph databases
- Organize and facilitate team conversations on culture and inclusive language

Technical Training Co-Instructor — *May 2020 – Aug 2020*

- Created and implemented 3 week full-stack software development curriculum for fully-remote training program
- Instructed cohort of 8 newly-hired software developers with a wide range of previous technical experience
- Managed new developers to complete a production-ready capstone project utilizing their newly-acquired skills

Software Engineer, Performance — *Aug 2018 – Feb 2020*

- Helped reduce the median response time of McMaster.com to 100 milliseconds
- Implemented a variety of predictive algorithms to pre-fetch resources before users completed their next action
- Orchestrated ongoing A/B tests to experimentally verify success of performance optimizations made by the team

Projects

Cognition Tracker | *Tensorflow, Python, Docker, React Native* | Website — March 2020

- Developed audio models for detecting cognitive state and progression of dementia through spontaneous speech
- Productionized end-to-end architecture serving multiple models and ensembling to give instantaneous score
- Ensemble model components include LSTM layers, BERT embeddings, and manual feature engineering

Mitigating Gender and Racial Bias in BERT | *Tensorflow, Python* | Github Repository — Dec 2019

- Surveyed current state of bias measurements and mitigation for word embeddings, discussed how they are ineffective with contextualized word embeddings like ELMo and BERT
- Defined a new metric measuring bias of contextualized embedding models in NER tasks, the ACEAT
- Showed how an adversarial debiasing model can reduce gender and racial bias simultaneously in BERT embeddings

Predicting Airline Delay with Weather Data | *Spark, Databricks, Python* — Apr 2019

- Utilized BTS dataset of 31.75 million domestic US flights over 3 years, combined with an NOAA weather dataset of 631 million records, to predict the probability of a substantial (15+ minute) flight delay 2 hours beforehand
- Conducted distributed computing processing and ML tasks in Spark on the Databricks platform
- Feature engineering included optimized windowing joins and distributed PageRank of airline travel

Technical Skills

Languages and Databases: Python, SQL, Neo4j (Cypher), MongoDB, C#/.NET, JavaScript
Frameworks and Tools: Tensorflow, Spark, Docker, React / React Native, Git
Libraries: NumPy, pandas, scikit-learn, spaCy, Dask, matplotlib, seaborn