Dr. Jingtian 'Josh' Wang

Data Scientist

✉ 13jtjoshua@gmail.com 🌐 joshjingtianwang.github.io/ ☎ 9174366823
in joshjingtianwang/ JoshJingtianWang

EMPLOYMENT

Roblox, *Data Science Intern*, San Mateo, CA May 2023 - Aug. 2023

- Slashed analysis time by 99% by developing an automatic **root cause analysis** tool, reducing time from days to minutes.
- Enabled 3 teams to adopt the tool after presenting its effectiveness to over 60 colleagues across 4 sessions.
- Boosted data integrity and cut error resolution time by 90% with a rapid data quality check tool.
- Built the first dashboard that tracks the usage of Roblox avatar animations, facilitating the shift towards data-informed decision making for the avatar movement team.

IFF, *Data Science Intern*, Hazlet, NJ Jan. 2023 - May 2023

- Boosted matching accuracy by 2% in the **entity resolution** pipeline through integrating **active learning** techniques, leading to reduced data labeling costs.
- Enhanced model evaluation precision and statistical power by engineering a **ML model comparison** library with robust statistical methods.
- Presented model optimization strategies to the digital operations team on 3 distinct occasions.

University of California, Irvine, *Graduate Student Researcher*, Irvine, CA Aug. 2018 - Aug. 2022

- Identified potential gene therapy targets from a detailed analysis of 80GB patient alternative splicing data, driving a data-centric approach to cancer treatment.
- Elevated the efficiency of a high-throughput RNA-seq data pipeline, improving QC, alignment, and differential expression analysis capabilities.
- Publications: https://scholar.google.com/citations?hl=en&user=TrF6tCkAAAAJ

PROJECTS

Kaggle Challenge (1st place): Classification of Tweets from Northern Europe (NLP)

- Achieving an 85% accuracy on the leaderboard by fine-tuning several multilingual **Transformer** models (BERT/RoBERTa) to classify tweets into political orientations. Achieved first place in the Kaggle Challenge.

Kaggle Challenge (1st place): Campaign Contributions and Elections in the US (Regression)

- Engineered 140 unique features from network datasets. Utilized **Pycaret** for model selection, while employing **Bayesian optimization** for hyperparameter tuning. Achieved first place in the Kaggle Challenge.

Classifying Tweets about Natural Disasters (NLP)

- Built an **LSTM** model to predict the authenticity of tweets about natural disasters.
- Conducted in-depth textual analyses encompassing topic modeling, sentiment extraction, and visualization via word clouds.

Citi Bike Station Inventory Forecasting (Time Series, Databricks)

- Executed ETL and EDA using **PySpark** on **Databricks** for data preparation and optimized a Prophet **Time Series** model via grid search and cross-validation, tracking and deploying models with **MLflow**.
- Created a pipeline adaptable to streaming data, providing visual prediction capabilities.

EDUCATION

University of Rochester Aug. 2022 - Dec. 2023

M.S. Data Science

Courses: Time Series, Deep Learning, Computational Intro to Stats, Database Systems, Statistical Machine Learning, Data Science at Scale

University of California, Irvine Aug. 2017 - Aug. 2022

Ph.D. Molecular Biology and Biochemistry 2022

Courses: Foundamentals of Genomics, Intro to Bioinformatics, Regulation of Gene Expression

Metis Data Science Bootcamp Jan. 2022 - May 2022

Completed an immersive 5-month data science bootcamp with a strong emphasis on project-oriented skill-building in problem-solving, data wrangling, statistical modeling, machine learning, and communication of deliverables.

SKILLS

Bash, Databricks, Deep Learning, Machine Learning, NLP, Numpy, Pandas, PySpark, Python, PyTorch, SQL, Tableau, Tensorflow, Time Series