Jingtian 'Josh' Wang

Data Scientist

13jtjoshua@gmail.com joshjingtianwang.github.io/
in joshjingtianwang/ JoshJingtianWang

EMPLOYMENT

Roblox, *Data Science Intern* **May 2023 - Aug. 2023**

International Flavors & Fragrances, Inc., *Digital Operations Intern* **Jan. 2023 - May 2023**

- Leveraged active learning in the entity resolution pipeline's matching step, boosting accuracy by 2% over passive learning, reducing data labeling costs.
- Developed a Python package that provides a statistically robust and streamlined ML model comparison process and offers estimations of minimum test data size and statistical power of model comparisons.
- Delivered presentations to the entire digital operations team on three separate occasions, showcasing key results and progress.

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

- Extrapolated new cancer gene therapy targets from over 80GB of cancer patient alternative splicing data
- Developed an alternative splicing-based cancer gene therapy
- Optimized a pipeline for RNA-seq data QC, alignment, and differential expression analysis for running on the high-performance computing cluster
- Publications: https://scholar.google.com/citations?hl=en&user=TrF6tCkAAAAJ

PROJECTS

Kaggle Challenge: Election Contributor Success Rate Prediction (Regression)

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

Citi Bike Station Inventory Forecasting (Time Series)

- Executed ETL and EDA using PySpark on Databricks for data preparation.
- Developed 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.

Stroke Onset Prediction (Classification)

- Developed an SVM classification model to probabilistically predict stroke onset, utilizing patient physical attributes and lifestyle data.
- Addressed class imbalance with Imblearn, enhancing the Brier Skill Score tenfold through GridSearchCV optimization.

EDUCATION

University of Rochester **Aug. 2022 - Dec. 2023**

M.S. Data Science

Courses: Data Mining, 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

DATA CLEANING	NLTK, Numpy, Pandas, PySpark, SpaCy
MACHINE LEARNING	Hyperopt, MLflow, PyTorch, Scikit-learn
DATA VISUALIZATION	Matplotlib, Seaborn, Tableau
OTHER	Bash, Cloud Computing, Databricks, Excel, Python, R, SQL, Streamlit