Fei Wu

fw2411@columbia.edu | (347) 472-1782 | New York, NY | linkedin

EDUCATION

Columbia University New York, NY
Master of Science, Major in Applied Analytics *[Expected]Sep 2023- Dec 2024*

University of Toronto Toronto, Canada
Bachelor of Science, Major in Statistical Science, Minor in Computer Science and Math *Sept 2018- June 2023*
Relevant Courses: Machine Learning, Deep Learning and Neural Networks(PyTorch), Software Design (java), System Programming(C), Data Structures, Algorithm Design, Multivariate Data Analysis, Database(PostgreSQL), Time Series Analysis(R), Cloud Computing.

TECHNICAL SKILLS

*Programming Languag*es: Python(Pandas, Matplotlib, scikit-learn, NumPy, PyTorch), R, SQL, C, Java, HTML5/CSS, JavaScript, Node.js.
Technology & Frameworks: Tableau, Power BI, Spark, Tensorflow, Git, AWS, dbt, Airflow, A/B Testing.

PROFESSIONAL EXPERIENCE

New York Life New York, NY
Data Science Intern *June 2024– Aug 2024*

- Refreshed candidate assessment model with 150k new data points, conducted thorough **validation** of data pipeline, analyzed model performance utilizing term effects and the Gini metric, ensuring model robustness in candidate evaluation.
- Introduced a new variable into the **generalized additive model** through extensive feature exploration. This initiative led to a significant enhancement in model accuracy, resulting in a 0.02 increase in the **Gini** coefficient.
- Extended the current GAM to include 11 new features engineered from 80k background resume data using **text mining** in R, enhancing the team's understanding of the dataset, improving the model's ability to predict candidate success based on text columns.
- Supported work into research and implementation of analytic methods and transitioned data ingestion pipeline from on-prem to cloud.

Huazhong Blockchain Technology Center Wuhan, China
Data Science Intern *Feb 2023– June 2023*

- **Automatically scraped** user accounts by creating a Python program triggered by daily scheduled cron job and input 2020 electric meters to **MySQL** database; guaranteed robust data integrity, coupled with scalability for accommodating growing data needs.
- Conducted **feature engineering** by integrating temperature and holiday indicator as independent variables for electricity usage forecasting using **SARIMAX time series** model; optimized through **grid search** hyperparameter tuning, achieving an average MAPE of 3.5%; constructed interval for **anomaly detection**, reducing energy fraud to safeguard revenue while enhancing customer service.
- **Streamlined** visualization in **Tableau** interactive **dashboard** to monitor a live summary of input data and receive alerts for detected anomaly usage, generating real-time insights that empower in-time decision making and operational efficiency.

Chinese Rowing Association Beijing, China
Data Science Intern *Jun 2020–Aug 2020*

- Built a MySQL **relational database** encompassing 8 entities and 49 attributes, conducting validation checks during the migration of 2000+ Excel entries; documented **ETL** process to ensure transparency and facilitate efficient data management.
- Developed a ML solution for classifying injury risk levels by performing **EDA** on 830 observations and building a **random forest** model on 12 transformed features, which assisted the design of tailored training plans and reduced monthly injury cases by 15%.
- Formulated web reports to the executive team by applying **HTML5, CSS** and employing **Node.js** to establish database connections to extract and display up-to-date statistics, enabling user centric interface, adaptability across devices and long-term usability.

PROJECT EXPERIENCE

Educational Insights Dashboard *Jan 2024-Mar 2024*
An end-to-end student question-solving dashboard and lecture diagnosis system from the EdNet-KT1 datasets.

- Engineered an ETL pipeline using **PostgreSQL**, integrated dbt to refine data transformations and implement comprehensive testing.
- Optimized efficiency of complex queries with **Spark** to extract key insights from 131 million quiz answering interactions and extracted insights of lecture qualities, enhancing the analysis and reporting capabilities.
- Delivered 2 **Flask**-based dashboards for visualizing user quiz performances and lecture quality assessment, employing HTML and CSS for intuitive and responsive user interface design, facilitating **real-time** data interaction and insight dissemination.

Cloud Based Cyrillic Recognition API *Sep 2023-Nov 2024*
Deployed a hybrid deep learning model architecture on AWS to enhance the accuracy and efficiency of Cyrillic handwriting recognition.

- Leveraged AWS S3 for robust data management, fine-tuned and created a model endpoint using customized model registry functions, successfully deploying a combined **ResNet** and **Transformer** model to enable real-time inference capabilities in **SageMaker**.
- Engineered and deployed a **RESTful API** using AWS API Gateway and **Lambda**, ensuring robust data handling and seamless model invocation to support scalable real-time data operations and improved user interactions.