

Xiaoyu Liu

608-320-6596 | xliu969@wisc.edu | <https://github.com/XiaoyuLiu198>

EDUCATION

University of Wisconsin Madison

Master of Science in Data Science (GPA: 3.91/4.00)

Madison, WI

Sept. 2020 - May 2022

Hunan University

Bachelor of Science in Statistics

Changsha, China

Sept. 2016 - Jun. 2020

INTERNSHIP EXPERIENCE

Data Developer Intern

Windriver

July 2021 –

California, US

- Develop and deploy data preprocess functions and data migration functions on AWS EC2.
- Build ELT pipelines for Snowflake database. Retrieve XML format data with API calls, parse to structured data and send to S3 bucket. Load and transform in Snowflake with SQL.
- Develop data validation functions for BI team.

Data Mining Intern

Saint Gobain

Jun. 2020 – July 2020

Shanghai, China

- Construct ETL process. Extract manufacturing data through mining and clawing methods from reports in Python. Complement data wrangling and transformation.
- Create and manage data model to store collected data.
- Build pipeline to integrate newly collected data with historical data and store in Oracle automatically.

Data Product Management Intern

Lufax

Dec. 2019 – May 2020

Shanghai, China

- Develop demo function of abnormal detection on time series data to automatically produce detection report.
- Design retention analysis function and funnel analysis function.
- Participate in designing website for data warehouse managing.

COMPETITIONS AND RELATED PERSONAL PROJECTS

Streaming Data Analysis | *Spark+Kafka+Airflow+AWS*

July 2021 – Aug. 2021

- <https://github.com/XiaoyuLiu198/Streaming-Data-Analysis>
- Set up Kafka topic and feed raw twitter JSON data into Kafka cluster from Twitter API.
- Build ETL process. Preprocess data read from Kafka using Spark SQL functions and window operation. Store data with AWS S3.
- Build machine learning pipeline with user defined function and LDA in pySpark.
- Develop dashboard showing analysis result with Python dash.
- Deploy ETL tasks and machine learning pipeline with Airflow.

Stock Data Pipeline Construction on AWS | *AWS+Java*

Aug. 2021 –

- Retrieve stock data with REST api. Put records to AWS kinesis.
- Parse stock data and analyze with AWS Lambda.
- Store processed structured data in AWS S3 bucket. Visualize with Amazon Athena and Quicksight

Test Answer Prediction(Kaggle top 18%) | *Python*

Dec. 2020 – Jan. 2021

- <https://www.kaggle.com/xiaoyuli123123/lightgbm-sakt>
- Preprocess the historical performance of over 100 million students in Riid lab and the metadata about the questions and the lectures. Create features on user-level and content-level.
- Transform and group tags using truncated SVD. Predict the probability of answering correctly using LightGBM.
- Predict the accuracy of answer in Self-Attentive-Knowledge-Tracing model.
- Embed the prediction using bagging method. Reach accuracy of 0.785.

Jane Street Market Prediction(Kaggle Silver Medal) | *Python*

Jan. 2021 – Feb. 2021

- Analyze real stock market data larger than 5GB with billions samples.
- Carry out exploratory analysis and preprocess numerical values with feature scaling.
- Build Autoencoder and feed encoded data to Multilayer Perceptron. Tune hyperparameters of max-depth and learning-rate in XGBoost and train data with cross validation to avoid overfitting.
- Embed the prediction from XGBoost and MLP. Reach accuracy of 0.762.

TECHNICAL SKILLS

Languages: Python, SQL, Java, Scala, R

Platforms: Spark, AWS, Kafka, Tableau

Libraries: matplotlib, ggplot, sklearn, tensorflow, pytorch, keras, dplyr, tidyverse, pandas, numpy