

Xiaoyu Liu

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

EDUCATION

University of Wisconsin Madison

Master of Science in Data Science

Madison, WI

Sep 2020 - Jan 2022

Hunan University

Bachelor of Science in Statistics

Changsha, China

Sep 2016 - Jun 2020

INTERNSHIP EXPERIENCE

Data Mining Intern

Saint Gobain

June 2020 – Aug. 2020

Shanghai, China

- Construct ETL process.
- Extract data through data mining and clawing methods from test reports in Python.
- Develop pipeline to integrate newly collected data with history data and store in Oracle automatically.
- Visualize test progress through Tableau.
- Analyze manufacturing data using Random Forest method, with F1 score 0.81.

Data Product Intern

Lufax

Dec. 2019 – May 2020

Shanghai, China

- Develop demo function of abnormal detecting model based on time series data.
- Visualize the abnormal change and standardize the output report in demo function.
- Turn retention analysis model and funnel analysis into software pattern.
- Develop function of extract data from database using MySQL.
- Design the data warehouse managing process.

COMPETITIONS AND RELATED PERSONAL PROJECTS

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

March. 2021 –

- Set up Kafka topic and feed raw twitter data into Kafka cluster.
- Preprocess data from Kafka using Spark SQLtext. Store data with AWS S3.
- Apply sentiment analysis and topic analysis to streaming data using user defined function and LDA in Spark.
- Deploy analysis tasks with Airflow.
- Developing dashboard showing EDA of hashtags with Python dash.

Gene Network APP Development | *R*

Feb. 2021 –

- Develop interactive analysis platform for genetic usage.
- Visualize gene network with igraph and visNet.
- Analyze network data with centrality measures and gene ontology enrichment analysis.

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

Dec. 2020 – Jan. 2021

- <https://www.kaggle.com/xiaoyuliu123123/lightgbm-sakt>
- 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 SAKT model, which is a deep learning model specified in learning trace.
- Combine the prediction using bagging method. Reached accuracy of 0.785.

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

Jan. 2021 –

- <https://www.kaggle.com/xiaoyuliu123123/xgboost-mlp-for-beginners>
- Exploratory analysis and pre-process with feature scaling.
- Tune hyper parameters in XGBoost and train data with split sets to avoid overfitting.
- Build Autoencoder and Multilayer Perceptron.
- Combine the prediction from XGBoost and MLP.

TECHNICAL SKILLS

Languages: Python, SQL, Scala, Java

Software and System: R, SAS, Tableau, Linux, Spark

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