Xiaoyu Liu

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

Education

University of Wisconsin Madison — Madison, WI
Master of Science in Data Science (GPA: 3.91/4.00) — *Sept. 2020 - Dec. 2021*

Hunan University — Changsha, China
Bachelor of Science in Statistics — *Sept. 2016 - Jun. 2020*

Internship Experience

Data Developer Intern — July 2021 –
Windriver — *California, US*
- Develop and deploy data preprocess functions and data migration functions on AWS EC2, improving the efficiency of BI team by 5 hours per person per month.
- Build automatic 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.

Data Mining Intern — Jun. 2020 – July 2020
Saint Gobain — *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 — Dec. 2019 – May 2020
Lufax — *Shanghai, China*
- Develop demo function of anomalies analysis on time series data to automatically produce detection report.Increase the efficiency of producing anomalies report by 3 hours per person per day.
- Design retention analysis function and funnel analysis function.
- Participate in designing website for data warehouse managing.

Competitions and Related Personal Projects

Streaming Data Analysis with Big Data Engine | *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 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 –
- Obtain stock data with REST api. Put records to AWS kinesis.
- Parse stock data and transform with AWS Lambda.Store processed structured data in AWS S3 bucket.
- Retrieve data with Amazon Athena and visualize with Quicksight

Test Answer Prediction(Kaggle top 18%) | *Python* — Dec. 2020 – Jan. 2021
- Preprocess the historical performance of over 100 million students in Riiid 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
- https://www.kaggle.com/xiaoyuliu123123/xgboost-mlp-for-beginners
- 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