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 — *Sep 2020 - Jan 2022*

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

Internship Experience

Data Mining Intern — June 2020 – Aug. 2020
Saint Gobain — *Shanghai, China*

- Extract data through data mining and clawing methods from test reports in Python.
- Develop function to integrate newly collected data with history data and store in Oracle automatically.
- Visualize test progress through Tableau.
- Analyze manufacturing data using XGBoost method, with F1 score 0.85.

Data Analyst Intern — Dec. 2019 – May 2020
Lufax — *Shanghai, China*

- Analyze data using retention analysis model and funnel analysis with MySQL and Tableau.
- Tune parameters and develop abnormal detecting model based on time series data.
- Visualize the abnormal change and standardize the output report.
- Extract data from database using MySQL.

Competitions and Related Personal Projects

Test Answer Prediction(Kaggle top 18%) | *Python* — Dec. 2020 – Jan. 2021

- 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 0.781.

Jane Street Market Prediction(Kaggle Bronze Medal) | *Python* — Jan. 2021 – Feb. 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.

Recommendation System for Speed Dating | *Python* — Nov. 2020 – Dec. 2020

- https://github.com/XiaoyuLiu198/Speed-Dating
- Recommend potential participants that match certain conditions and share similar interest or background.
- Use target encoding to encode the categorical features.
- Impute the missing value using MICE and Decision Tree according to the relationship between features.
- Tune parameters using grid search method.
- Cluster users using KNN model according to their interest and background.

Analysis of Distribution of Charging Piles(MCM Second Award) | *Python, R* — Jan. 2018 – May 2018

- Scrape traffic data and map data using API.
- Build regression model to predict the total number of charging piles.
- Solve the maximum coverage problem using genetic algorithm.
- Use Q-type clustering method based on level of development of the country, density of popularity, and other indexes.

Technical Skills

Languages: Python, SQL, Java
Software and System: R, SAS, Tableau, Linux, Hadoop
Libraries: matplotlib, ggplot, sklearn, tensorflow, pytorch, keras, dplyr, tidyverse, pandas, numpy