Xiaoyu Liu

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

EDUCATION

University of Wisconsin-Madison — **Madison, WI, US**
Master of Science in Statistics (Data science concentration), GPA: 3.81/4.00 — **Sept 2020 - Jan 2022**
Hunan University — **Changsha, China**
Bachelor of Science in Statistics — **Sept 2016 - Jun 2020**

INTERNSHIP EXPERIENCE

Data Mining Intern — **Shanghai, China**
Saint Gobain — **Jun 2020 - Aug 2020**

- Extract manufacturing data through mining and clawing methods from test reports in Python. Complement data wrangling and transform to structured data.
- Construct ETL process.
- Develop pipeline to integrate newly collected data with history data and store in Oracle automatically.

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

- Develop function of abnormal detection based on time series data to find out the reason of abnormal change.
- Implement retention analysis function and funnel analysis function with Python.
- Develop function of extracting data from database using SQL.
- Participate in designing website for data warehouse managing.

RESEARCH PROJECT

Streaming Data Analysis Dashboard Development | Spark+Kafka+Airflow+AWS — **Mar 2021 –**

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

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

- https://www.kaggle.com/xiaoyuliu123123/lightgbm-sakt
- Preprocess the historic 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 accuracy of answer in Self-Attentive-Knowledge-Tracing model and LightGBM.
- Embed the prediction using bagging method. Reach accuracy of 0.785.

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

- https://www.kaggle.com/xiaoyuliu123123/xgboost-mlp-for-beginners
- Analyze real stock market data larger than 5GB with billions samples. Create features on user-level and content-level.
- Carry out exploratory analysis and preprocess numerical values with feature scaling.
- Tune hyperparameters of max-depth and learning-rate in XGBoost and train data with cross validation to avoid overfitting.
- Build Autoencoder and Multilayer Perceptron.Embed the prediction from XGBoost and MLP. Reach accuracy of 0.702.

Natural Language Processing with Twitter Data | Python+R — **Nov. 2020 – Dec. 2020**

- https://github.com/XiaoyuLiu198/NLP
- Apply tokenization and deleted stopwords to data set include 50000 twitter comments..
- Build MLP model with word2vec embedding layer, dense layer with activation function and dropout layer to classify the comment into positive or negative group.
- Develop analysis dashboard with Rshiny, including visualization with bar plot and topic analysis with LDA.
- Evaluate the result using classification metrics. Accuracy of classification is 0.802

SKILLS & INTERESTS

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