Jiahui Wu

Phone number: (+1) 202-716-3466 | Email: jw1987@georgetown.edu | Arlington, Virginia, United State

Education

Georgetown University 08/2021 till now

Master of Science, Data Science and Analytic

The Chinese University of Hong Kong, Shenzhen 09/2017 - 05/2021

Bachelor of Science, Mathematics and Applied Mathematics, Minor in Statistic

Core courses

Machine learning, Database systems and SQL, Big data and cloud computing, Regression analysis, Mathematical statistics, Categorical data analysis, Stochastic process, Statistics experiment design, Optimization, Probability theory, Data structure.

Skills

Python (proficient): Data Analysis(pandas, sikit-learn, numpy); Data Visualization(plotly, matplotlib, seaborn geopandas, folium); Web Crawler(urllib, bs4); Database Management(pymysql); multiprocessing, pyspark; Data Structure.

R (proficient): Statistical Analysis & Simulation; Machine Learning; Data Visualization(ggplot2), rmarkdown.

Mysql (proficient): Relational Model; Data Querying(window function, indexing)

AWS: Cloud9, Sagemaker, EC2, S3

Familiar with: **Hadoop, Git, Linux, Tableau , HTML, CSS, Java script, Matlab, Java, Stata**.

Statistics: Regression Analysis; Hypothesis Test; Categorical Data Analysis; Financial statistical Modeling; Monte Carlo simulation.

Mathematics: Optimization; Numerical Analysis; Probability Theory

Language: English (proficient), Chinese (native)

Project

Service demand analysis and prediction of for-hire vehicle in New York (Python, Mysql) 09/2021-11/2021

- Data gathering of raw service record and weather data through API, followed by Mysql database management and data cleaning
- Data mining with ARM Network, decision tree based methods (Random Forest, Xgboost).
  - Average service demand prediction corresponding to location, time point and day of week within a month.
  - Specific service demand prediction corresponding to location, time point with hourly weather data taken into account.
  - Interactive network showing relationship between time slots and locations that will generated large service demand.
- Geographic and interactive visualization of predictions and true values.
- Capture over 85% variance of service demand with respect to time and locations, provided efficient traffic planing.

Movie recommendation system (Python, Mysql, AWS) 01/2022-04/2022

- Offline recommendation for old user based on their previous rating records: dimension deduction with SVD and ALS algorithms, calculated and stored cosine similarity matrix of movies. Predicted score for unrated movie for each user. Selected the top 20 movies with highest predicted rating.
- Online recommendation for old user with their newly rated movie taken into consideration: refresh predicted result with old similarity matrix of movie and new rating records.
- Cold-start of new users: initialize with top 50 highest average rated movies.
- User interface design (TkinterGUI)
- 70% recommended movies intuitively meet with users' preference based on historical rating record in several tests.

Statistical analysis of drug consumption risk based on personality (R) 04/2020-05/2020

- Pairwise independence test (Likelihood ratio) on drug and personality & Conditional dependence test (Cochran-Mantel-Haenszel) between drug and personalities. For each individually dependent personality, find its effect modifier(other personalities).
- Logistic regression of drug consumption on personality.

Internship

Data analytics intern, Tian An Financial Holding Limited Company, Shenzhen 07/2020 - 08/2020

- Conduct experiment on R to verify financial investment strategies with thousands of financial statement of Chinese companies, including experimental CAPM test, calculation of value factor, exploring announcement effects on market reaction of Chinese companies.
- Reinforce data processing skill with financial data and coding skill of R.