Wenjing (Jenny) Shan

wshan2@andrew.cmu.edu | LinkedIn | 412-613-4908 | Pittsburgh, PA 15213

EDUCATION

Carnegie Mellon University Pittsburgh, PA

- *B.S. in* ***Statistics and Machine Learning*** Expected May. 2023

Minor in ***Business Analytics and Optimization / Computer Science***

- GPA: 3.91/4.00, College Dean's List 2019 - 2021 with High Honors
- Relevant Coursework: Machine Learning, Statistical Computing, Functional Programming, Statistical Graphics and Visualization, Business Analytics, Algorithms and Advanced Data Structures, Data Management, Deep Learning

WORK EXPERIENCE

Analytics For Everyone LLC. (a4e.tech) Pittsburgh, PA

Data Analytics Intern Sept. 2021 – Nov. 2021

- Designed asynchronous data loading process for podcast analytics that aims to simplify the access, searching, and manipulation of podcasts, including data preprocessing, feature engineering, and text analytics via **Python**
- Created **HBase** Database Docker Image with **NoSQL** tables for podcast file analysis storage, built inverted-index **PySpark** function and unit tests, and populated the database tables with the keywords

Stem-Away Inc. Remote

Data Analyst Intern Jun. 2021 - Aug. 2021

- Extracted, transformed, and loaded 1GB of user posts data from Discourse Hub Community via **Python** to **SQL** database, and calculated the similarity between posts by vectorizing words and computing the similarity matrix
- Built logistic regression and machine learning models (classification trees, random forest, natural language processing) to categorize user posts based on related factors in **Python (Pandas, NumPy)**
- Improved model processing efficiency by 20% by feature extraction and increased model accuracy by 20% via machine learning model selection and parameter tuning
- Built a web application to recommend forum posts by **Flask** and maintained data processing in **Amazon AWS**

Ernst & Young Shanghai, China

Business Consulting Intern Dec. 2020 - Feb. 2021

- Collaborated with Project Managers to facilitate GlaxoSmithKline to maintain effective internal Enterprise Resource Planning (ERP) in order to improve the efficiency and reduce potential risks
- Conducted data cleaning and develop descriptive statistics in Excel on 100k rows of data to perform **Exploratory Data Analysis** to show the system risks and related factors
- Built 5+ visualization charts in **Excel** and **Tableau** to present analysis result, and automated the assessment of key risks and potential impact on the client's business
- Presented key findings and propose strategic solutions to 10+ project stakeholders

PROJECTS

Snowflake – Data Warehouse Architecture Nov. 2021

- Built a data pipeline using **Dbt** on subscribed data sets to calculate profit and loss for financial reporting and insight
- Conducted basic testing, deployment, and materialization to run reliable and high-performance transformation

Database Design for Nile.com Oct. 2021

- Developed a **SQL** database system via Valentina Studio that supports the capture, storage, and management of all data required for online order, fulfillment center process scheduling, package delivery, and customer support
- Loaded sample data from fulfillment centers, checked for redundancies and anomalies to ensure schema correctly captures the specific business entities and events, and shared the database on **AWS** Aurora database server

Classification of Point-Like X-ray Sources in the ROSAT 2RXS Survey Sept. 2020 - Dec. 2020

- Processed data of astronomical objects with 4,000+ observations and 26 features on **R**, conducted data cleaning, exploratory analysis, and principal component analysis (PCA)
- Built multiple models (Lasso/ridge regression, decision tree, KNN, and generalized additive model) to identify the most efficient classifier and key factors, improved classification accuracy by 5%

SKILLS

Programming Languages: SQL, Python (scikit-learn, pandas, numPy, PyTorch), R (dplyr, ggplot2, biplot), C, SML
Tools and Frameworks: Tableau, NoSQL, Agile, Git, Docker, Spark, MS Office (Excel Pivot, PowerPoint), A/B testing
Languages: Proficiency in Mandarin and English, Spanish