Yujia (Sian) Jin

Tel: (202) 6296597 | yj234@georgetown.edu | Linkedin | Github | Portfolio | Washington, DC

EDUCATION

Georgetown University | M.S in Data Science and Analytics | GPA: 3.9/4.0 DC, USA | Aug 2021 – May 2023

Relevant Coursework**:** Big Data and Cloud, Machine Learning, Data Structures & Algorithms, Data Mining

Queen's University Belfast | M.S in Pharmacy | GPA: 3.55/4.0 Belfast, UK | Sept 2017 – July 2021

SKILLS

Programming: SQL, Python (Numpy, Pandas, Scikitlearn), SAS, R, HTML, JavaScript, Linux, CSS,

Models: Regression, Classification, Clustering, PCA, RNN, CNN, Natural Language Processing, Time Series

Tools: Jira, Hadoop, Spark, PyTorch, TensorFlow, Azure DataBricks, Data Flow, AWS, Git, MS Access, Excel

Visualization: Tableau, leaflet, d3.js, bokeh, Altair, PowerBI

WORK EXPERIENCES

Data Scientist | *Pear Tree Consulting, LLC* PA, US | July 2023 – Present

- Developed dynamic stock price information dashboards using Tableau and MS Excel; generated and delivered detailed reports to clients, summarizing analytical findings, and facilitating data-driven discussions and strategies
- Ensured data quality and integrity through rigorous data cleaning and preparation tasks in Python and R
- Incorporated advanced visualizations of stock market and generative AI to provide clear narratives for visual
- Applied time series and neural network models to forecast the performance of 100+ stocks and Bitcoin daily; engineered 10+ new data features to amplifying predictive power, increasing prediction accuracy by 10%

Data Scientist Intern | *Civilience* DE, US | May 2022 – Aug 2022

- Conducted web scraping to download 10000+ records of online disease text info and applied topic modeling and sentiment analysis on TF-IDF and Word2Vec output to identify trends and focus areas in infectious diseases
- Designed and implemented machine learning ETL data pipeline using AWS (EC2, lambda, EventBridge, S3) that processed billions of records per day, reducing manual effort by 80%, and ensuring timely data processing
- Developed a recommendation system using cosine similarity and collaborative filtering based on user info and inputs, fetched via REST API, to automatically assign mentors to users, resulting in a 5% increase in subscriptions
- Communicated the findings from customer satisfaction analysis to the product development teams, fostering cross-functional collaboration and ensuring that data-driven insights were incorporated into product improvements

Research Analyst | *Queen's University Belfast* Belfast, UK | Sept 2020 – Jun 2021

- Designed and executed SQL queries and MS Excel to construct databases containing over 50 records of proteins
- Utilized principle component analysis to reduce data dimension; visualized the dimension-reduced results through interactive plots, facilitating intuitive interpretation and enabling valuable discoveries for subsequent research
- Conducted hypothesis testing on protein data, and applied K-means clustering to categorize protein into 5 types using silhouette score for determining K, which achieved an impressive 75% accuracy rate in protein classification

Project Data Analyst | *Worley Parsons* Shanghai, China | June 2019 – Sept 2019

- Conducted data governance practices, ensuring proper documentation of data sources, and analysis methods
- Utilized the relational database to store and access the data needed for analysis; designed and implemented complex SQL queries and Snowflake views to join and aggregate data from multiple sources
- Leveraged Power BI and Tableau dashboard to track project data, enabling data-driven decision-making for resource allocation and project performance, leading to improved project outcomes and operational efficiency

PROFESSIONAL PROJECTS

Big Data: Examining Political Engagement and Sentiment (git) DC, US | Sept 2022 – Dec 2022

- Conducted exploratory analysis on subreddits data to understand textual patterns of different political Reddit users
- Implemented data cleaning and text processing pipelines on Azure DataBricks, via PySpark to process 8TB data, including tokenization, feature extraction using CountVectorizer and IDF, and label encoding
- Constructed machine learning pipelines with decision tree, random forest with hyperparameters tuning, and Gradient Boosting to predict political label of the subreddits, achieving a 30% improvement in accuracy

Data Visualization: Volcano Analysis & Visualizations (git) DC, US | Jan 2022 – May 2022

- Built websites using HTML/Javascript to present a comprehensive data science story on volcanoes
- Conducted exploratory data analysis and created interactive visualization using Python (matplotlib, plotly), R and JavaScript to uncover insights in geospatial and quantitative data