CHENYUE JIN

Arlington, VA | 202-725-4544 | cj576@georgetown.edu | GitHub | LinkedIn

EDUCATION

Georgetown University — M.S., Data Science and Analytics, GPA:4.0/4.0 Aug 2021 - May 2023

- Courses: Database and SQL, Machine Learning, Big Data and Cloud Computing, Data Visualization, Stat Computing, Natural Language Processing, Time Series

Huaqiao University, China — B.S., Statistics (Financial Statistics Concentrate), GPA: 3.95/4.00 Sept 2017 - Jul 2021

WORK EXPERIENCE

Global A.I. *New York, NY*

Data Scientist Intern — *Python, MySQL* Jul 2022 - Sept 2022

- Analyze the impact of MSCI US Index stocks on the Sustainable Development Goals with 2 GB historical data for research purposes.
- Processed big sparse datasets and handled missing values. Performed predictive analysis on stock price through machine learning and a deep learning model (LSTM), resulting in an accuracy improvement of 7.8%.
- Established data pipeline between raw data sources and database through continuous data scraping, structure transformation, cleaning, and formatting in Python, improving efficiency by 20%.
- Utilized MySQL queries, including aggregations, window functions, and OLAP, to provide precise small datasets for other departments' needs.
- Generated reports with data visualization focused on Information Technology Sector to distill highly technical information into actionable information to help with decision-making.

Phalanx Analysis Group *San Francisco, CA*

Data Analyst Intern — *R, MySQL, Tableau* May 2022 - Jun 2022

- Collected and pulled the industry time series data to build charts for the top 5 international smartphone companies' market development and competition using Tableau.
- Implemented ARIMA, GARCH, and VAR models to forecast future market share and global shipments for customer investment reference, achieving models' MSE values below 10.
- Performed data cleaning, mathematical modeling, digital storytelling, and database management with multiple programming languages such as R and SQL to meet client business objectives.

Ipsos Market Consulting Co. Ltd. *Beijing, China*

Research Analyst Intern — *MySQL, MS Excel, Tableau* Jul 2020 - Sept 2020

- Provided technical consulting to companies ranging from college research teams to large corporations on various topics such as urban studies, business studies, and new retailing.
- Implemented data aggregation and filtering using Vlookup with Excel. Proficient in creating user-friendly pivot tables and Tableau dashboards.
- Assisted in creating and presenting informational reports for management based on SQL databases.

PROJECT EXPERIENCE (Other projects link)

Bank Customer Churn Prediction Analysis — *Python, R* Nov 2021 - Dec 2021

- Designed a logistic regression classification model, Random Forest, and KNN that can predict customer churn intention based on labeled data with the average accuracy of 83.76%.
- Used k-fold cross-validation technique to find the best models, implemented Lasso and Ridge regression to find optimal hyper-parameters, and overcome over-fitting via regularization with optimal parameters.
- Generate feature importance to find the main elements that influenced the outcomes. The top 5 features by Random Forest Model are age, estimated salary, credit score, balance, and number of products.
- Evaluated model performance for Logistic Regression and Random Forest are 0.772 and 0.845.

San Francisco Crime Analysis — *Spark SQL, Databricks* May 2022 - Jul 2022

- Performed spatial and time series analysis on SFPD's dataset of 15 years of reported events.
- Built a data processing pipeline based on Spark RDD, Dataframe, and Spark SQL for OLAP of big data.
- Trained an ARIMA model to predict the number of burglary events per month.

Natural Language Processing and Topic Modeling on User Review Dataset — *Python* Jan 2022-Mar 2022

- Preprocessed unlabeled textual documents of user reviews for a watch on Amazon by tokenizing, stemming, stop-word removing, and extracted features using TF-IDF approach.
- Constructed unsupervised learning models of K-Means Clustering and LDA.
- Identified latent topics and keywords of each document for clustering and calculated document similarity.

COMPUTATIONAL SKILLS/OTHERS

Programming Languages: Python (3 years), R (4 years), MySQL (3 year), SPSS, Spark, Hadoop, AWS (s3, EC2)

Languages: Mandarin (native), English (fluent)