Ruize Xuan

📞 +1 (858) 333-0081   ✉ rayxuan2000@gmail.com   in LinkedIn   ⓖ Github

Objective

A graduate student who aims to find **data analyst internship**, with **more than one year experience** in handling data sets using **machine learning**, **big data framework** and **data visualization tools** to make data-driven decisions. Solid programming skills in **Python** and **SQL**.

Education

University of California, San Diego   *CA, United States*   **Sept 2023 – May 2025 (Expected)**
Master's major: Electrical and Computer Engineering   *GPA: 3.6/4*
Core coursework: Programming for Data Analysis, Probability and Statistics for Data Science, Scientific Data Analysis and Statistical Learning, Introduction to Visual Learning, Web Mining and Recommender Systems

Southeast University   *Jiangsu, China*   **Aug 2019 – Jun 2023**
Bachelor of Engineering, Information Engineering (with honored degree, Presidential Fellowship)   *GPA: 3.75/4*

Project Experience

Sales Data Analysis and Visualization for Business Intelligence | *File*
- Executed SQL queries on a comprehensive PostgreSQL database with **12** tables and **88** columns, utilizing the pgAdmin tool interface to conduct in-depth analysis of company sales, supply chain, shipping, and customer orders.
- Generated an ERD to visualize database structure; designed **25+** OLAP queries to extract valuable business insights.
- Developed a Tableau dashboard to present sales KPIs, providing an interactive and clear overview for stakeholders to drive strategic initiatives and optimize business operations.

Taxi Rides time series analysis and Statistical Modeling | *File*
- Developed a Python workflow for collecting, cleaning, and analyzing **10GB+** of cab trip records from; Created Tableau dashboards to visualize the data, supporting high-level decision-making in strategic route planning.
- Conducted exploratory data analysis; performed ADF stability testing and differential smoothing to ensure data integrity and reliability for business insights.
- Generated autocorrelation and partial autocorrelation plots; fine-tuned and compared multiple time series models; achieved a **12% reduction** in AIC, enhancing the model's predictive accuracy for demand forecasting.
- Proposed and validated a hypothesis testing model for cab passenger flow using external data and Poisson regression; established a statistically significant correlation with commercial activities.

Customer Reviews Text Mining and Topic Modeling | *File*
- Developed clustering and LDA models on an **8MB+** E-commerce customer reviews dataset using Python; built a topic-based keyword indexing table to uncover customer shopping patterns and provide insights for enhancing retention.
- Preprocessed text data by sentiment-based segmentation, tokenization, and stemming; vectorized review texts using TF-IDF for enhanced sentiment analysis.
- Developed semantic prediction models using Naive Bayes and SVM, achieving an AUC score of **over 0.92**, supporting targeted marketing campaigns and customer satisfaction initiatives.

Geo-Visualization and Modeling in Healthy Home Prediction | *File*
- Implemented geo-related coding and map visualization using Python on a **4MB+** public health dataset to derive data-driven insights into areas with heightened health risks, supporting strategic health interventions and policy-making.
- Conducted data cleaning on GeoDataFrame and developed custom functions to integrate spatial information effectively.
- Mapped city infrastructure on the plot and generated a comprehensive heatmap depicting hazard air pollution levels.
- Integrated cross-validated ensemble learning models with various feature importance metrics to identify indicators of air pollution, achieving an optimal **1.38** RMSE value, contributing to better community health management.

E-commerce Fraud Detection and Risk Analysis | *File*
- Engaged in the creation of a hierarchical alert system tailored for a **17MB+** E-commerce enterprise's transaction database; contributed to the formulation of strategies aimed at mitigating fraudulent activities and reducing financial risk.
- Implemented IP address analysis, resulting in a **22% reduction** in lookup time and enhancing operational efficiency.
- Conducted data preprocessing and feature engineering to extract critical insights from raw data.
- Played a pivotal role in machine learning modeling; employed various resampling techniques; achieved a remarkable **75% improvement** in recall score, significantly boosting the accuracy of safeguarding revenue.

Technical Skills

Programming Languages and visualization:
- Python (*sklearn, pandas, numpy*), SQL , MATLAB, Tableau

Machine Learning and Statistics Analysis:
- Classical & Penalized Regression Methods, Decision Tree, Clustering, KNN, PCA; A/B Testing, Hypothesis Testing