Jinyi Xu

jix223@lehigh.edu • (610)-570-5154 • Bethlehem, PA 18015 • https://www.linkedin.com/in/jinyi-xu-b55647232/

EDUCATION

Lehigh University ***Bethlehem, PA***

Bachelor of Science in Business and Economics & Bachelor of Science in Mathematics ***Expected 2024.05***

GPA: 3.7/4.0, Major in Finance & Mathematics

Core Course: *Linear Models in Statistics; Business Analytics I & II; Statistical Computing & Application; Principles of Analysis I; Intro to Theory of Probability; Time Series & Forecasting*

PROFESSIONAL EXPERIENCE

THE BIG DATA RESEARCH CENTER OF TSINGHUA UNIVERSITY ***Beijing, China***

Research Assistant ***2023.05 – Present***

- Researched on the architecture and application effectiveness of medical databases through literature review
- Conducted a comprehensive comparative analysis involving PostgreSQL database, cluster, and Z-Order indexing, utilizing query time and data retrieval volume as primary metrics; executed multi-condition MySQL queries with identical data dimensions but varying quantities, and compared query times across different methods
- Concluded that Z-Order indexing is the most efficient one, particularly in scenarios characterized by substantial data volumes

GAOTU TECHEDU INC. (NYSE: GOTU) ***Beijing, China***

Data Analyst Intern ***2023.05 – 2023.07***

- Collected 200,000+ pieces of educational data (e.g., average score and ranking in different regions, minimum and maximum score for entry into different majors and colleges) about Gaokao (Chinese college entrance exam) using web scrapping in Python
- Conducted data cleaning and preprocessing tasks using Excel, including removing duplicates, handling missing values, standardizing data formats and merging tables for further analysis
- Performed statistical analysis to identify trends and patterns in educational data; developed insightful dashboards to track and monitor Gaokao performance metrics (e.g. regional admission rate & acceptation rate); provided recommendations on making school list and preference during application period for education agents, students and parents to help improve application results

KPMG CHINA ***Beijing, China***

Product Analyst Intern ***2022.06 – 2022.10***

- Built insightful Excel dashboards to generate daily and monthly reports encompassing customer complaints, technical support, activation rate, and product ratings for Mercedes-Me products; presented findings and recommendations to product directors to enhance customer experience and product performance
- Assisted in organizing and delivering training sessions to over 700 Mercedes-Benz dealers, driving product sales and usage
- Collaborated with a team of 20+ customer service representatives to ensure timely response to customer compensation

SELECTED PROJECT

Time series forecasting of Favorita store sales performance (R) ***2023.09 – Present***

- Extracted and processed 90,000+ store records spanning from 2013 to 2017, conducting data cleaning (e.g., log transformation) in preparation for subsequent machine learning & time series modeling
- Utilized Python to generate time series plots, unveiling the inherent seasonality of store sales and identifying performance outliers or anomalies that raised business concerns
- Built an ARIMA time series model for store sales prediction, achieving a Root Mean Squared Error (RMSE) of 0.2
- Undertook a comprehensive literature review to identify optimal use cases for various time series models, consulted AI search engines, and compared the methodologies with project's initial analysis

Predicting the cost of claim for heart disease (SAS: Regression, Model Selection) ***2022.03 – 2022.05***

- Collected 800+ pieces of claim data and performed data cleaning, including removing duplicates, checking multicollinearity and variable transformation, to prepare further regression analysis in SAS
- Built multilinear regression models in SAS to explore the relationship between heart disease and influencing factors such as age, gender, intervention and emergency room visits
- Utilized automatic backward and forward selection to select the proper features for the model, tuning the model accuracy from 0.55 R^2 to 0.62 R^2

Effects of environmental factors on the species richness patterns of plants with different growth forms in North China (R: Visualization, Regression) ***2019.09 –2019.10***

- Conducted fieldwork to collect data about 60 plant species and relevant factors (e.g., annual average temperature, altitude, rainfall precipitation), and employed R for data cleaning and preprocessing, ensuring data quality and integrity for subsequent analysis
- Used multilinear regression to investigate the relationships between climate factors and diversity, created visualized trends and correlations using scatter plots and bar charts in ggplot2, and fine-tuned the regression model to achieve 0.76 adjusted R^2
- Concluded and substantiated positive correlations between species diversity and specific climate factors, and published the research findings in the Journal of Central China Normal University

SKILL & INTEREST

Technical Tools: Microsoft Office, SAS**,** Java, Python, NumPy, Pandas, R, MySQL

Language: Mandarin (Native); English (Fluent)