Juye Xiao

jx2479@columbia.edu | 13436826896

EDUCATION

COLUMBIA UNIVERSITY **New York, NY, USA**

Master of Science in Applied Analysis Sep.2021–Feb.2023

Relevant Coursework: Managing Data, Strategy and Analytic, Applied Analysis Framework and Methods, Storytelling with Data, Applied analysis in organizational context, Cloud Computing, Analytics and leading change, Solving Real World Problems with Analytics (Capstone sponsored by S&P Global)

UNIVERSITY OF MICHIGAN **Ann Arbor, MI, USA**

Bachelor of Science in Computer Science; Bachelor of Science in Data Science Aug.2018–Dec.2020

Relevant Coursework: Programming and Data Structure, Data Structure and Algorithm, Introduction of Computer Organization, Foundations of Computer Science, Mobile App Development for Entrepreneurs, Database Management Systems, Web Systems, Introduction to Machine Learning, Human-Centered Software Design and Development, Computer Vision, Data mining, Applied Regression Analysis

University Honors ***May.2019***

PENNSYLVINNA STATE UNIVERSITY **University Park, PA, USA**

Bachelor of Science in Mathematics; Bachelor of Science in Computer Science Aug.2017–May.2018

Relevant Coursework: Linear Algebra, Discrete Math, Calculus, Differential Equation, Introduction of Digital System and Design, Introduction of Python, Introduction of Java, Elementary Statistics

The President's Freshman Award ***April.2018***

Dean's List ***FA 2017 &SP 2018***

EXPERIENCES

Data Analyst Specialist, C.O.D.E-Automotive, Camarillo, California, Sep.2023

- Managed inventory records and shipping processes, analyzing product sales data to provide procurement recommendations and inventory management.
- Evaluated advertising performance on eBay, Walmart, and Amazon, adjusting placement strategies to improve ROI..
- Optimized advertising budget allocation to drive sales growth while ensuring competitive product pricing in the market.

Database Management and Data Entry Specialist, China United Transport, Walnut, California, Mar.2023

- Participated in the establishment of a comprehensive database for historical freight data, significantly enhancing the company's data retrieval capabilities and analytical insights.
- Responsible for the daily maintenance and update of the database, ensuring the accurate entry of new shipping itineraries to maintain the database's integrity and relevance.

Front-End Internship, DiDi, Beijing, China, June.2019 – August.2019

- Contributed to the development of a website performance monitoring platform using Lighthouse and Puppeteer, leveraging technologies such as React and Node.js.
- Utilized data collection and analysis techniques to extract valuable insights from performance data, aiding in the optimization of web applications.
- Involved in DiDi's internationalization process, gaining exposure to global web deployment practices and cross-cultural collaboration.

Data Analysis Intern, Gcores, Beijing, China, May.2018 – July.2018

- Involved in a database optimization project by identifying and reorganizing tags, enhancing data accessibility and query efficiency.
- Applied data extraction and cleansing techniques to preprocess player reviews collected from prominent game websites using web crawlers.
- Collaborated closely with cross-functional team members to extract actionable insights from data and leverage them in content creation and article composition for the website. (HTML, Sass)

PROJECTS

Climate Change Monitoring and Prediction using Satellite Data

- Executed a project sponsored by **S&P Global** to bridge the gap between climate change and economic data.
- Gathered **MODIS** Land Surface Temperature and **NEX-GDDP** near-surface temperature data from **NASA**.
- Worked with **netCDF** and **HDF** file formats, extracted and cleaned to build data visualization, including **Heat Map**, to elucidate temperature patterns.
- Conducted comprehensive **time-series analysis** and **decomposition** to reveal systematic variations between MODIS and NEX models.
- Developed, and trained an **ARIMA model** to forecast the land surface temperature based on the MODIS data.

Banking Customer Churn Prediction and Analysis

- Built a machine learning model to predict customer churn, achieving 86.35% accuracy and a high AUC score using **Random Forest**.
- Preprocessed data using **One-Hot Encoding** and **standardization**, and identified key factors like age, account activity, and geographic location.
- Evaluated models (**Random Forest, Logistic Regression, KNN**) through **cross-validation** and optimized with **GridSearchCV**.
- Enabled the bank to identify high-risk customers, implement targeted retention strategies, and improve customer retention rates.

LendingClub Loan Default Predictions

- Developed a machine learning model to predict loan defaults on LendingClub's platform, improving risk management.
- Used **Logistic Regression**, **KNN**, **Random Forest**, and **SVM**, with **GridSearchCV** for hyperparameter tuning.
- Identified key predictors of loan default such as income, interest rate, and credit utilization.
- Achieved 83.52% accuracy with **Logistic Regression**, enabling more accurate loan risk assessment and better investment decisions.

Amazon User Review Analysis and Topic Modeling

- Conducted Amazon user review analysis using Python, leveraging **K-Means clustering** and **Latent Dirichlet Allocation (LDA)** to uncover hidden semantic structures and customer insights.
- Preprocessed review data through **tokenization**, **stopword removal**, and **stemming**; extracted **TF-IDF features** to identify the top keywords representing customer sentiment in clusters and topics.
- Identified 5 key user interest themes, including product aesthetics, comfort, quality, price-value, and gifting potential, enabling actionable marketing and product development strategies.
- Delivered insights to optimize pricing strategies, design enhancements, and targeted advertising, improving customer satisfaction.

San Francisco Crime Analysis

- Utilized **Spark SQL** to identify the most dangerous districts in San Francisco (Southern, Mission, and Northern) and analyze crime distribution by category and time, highlighting peak hours and dominant crime types.
- Conducted **time-series decomposition**, revealing higher crime rates in January and October each year, along with a declining trend observed after 2018.
- Built and applied **grid search** to determine the optimal parameters for the **SARIMA model**, successfully forecasting San Francisco's monthly average crime rates for the next 24 months to aid resource planning.
- Evaluated crime resolution rates, focusing on low-resolution cases like theft and burglary, and proposed targeted improvement measures.
- Recommended strategies such as optimizing patrol schedules, enhancing evidence collection methods, and promoting community safety programs to reduce crime rates and improve resolution rates.

Professional Photo Editing App, PicassoXS

- Developed and managed a professional photo editing application that simulated industrial development of a data-driven project.
- Conducted comprehensive customer research, utilizing the **Value Proposition Canvas**, to assess and address customer needs and preferences.
- Gathered, cleaned, and analyzed research data, creating an **affinity map** and **user personas** for precise targeting.
- Estimated market size and performed a thorough **competitor analysis**, guiding product strategy and positioning.
- Successfully built a **Minimum Viable Product (MVP)** and conducted **unit tests** to validate the value proposition and user satisfaction.
- Designed an intuitive Front-end User Interface using **Figma** and implemented it using **Swift**, ensuring a seamless user experience.
- Established robust communication between the front-end and back-end systems for efficient data handling.

ADDITIONAL SKILLS

Technical Skills: Python (Sklearn, Pandas, NumPy), Java, C++, Swift, Unix, AWS, Pytorch, Apache Spark, SQL, MongoDB, React, HTML, CSS, Node.JS, R, Figma, Tableau, Metabase

Machine learning: Classical and Penalized Regression methods (Lasso and Ridge), Decision Tree, Clustering, KNN, K-means Clustering, Principal Component Analysis, Convolutional Neural Network, Recurrent Neural Network

Statistics Analysis: Hypothesis Testing, Text mining, Time-series analysis