Boya Zeng

Chicago, IL 60611●bzeng7@uchicago.edu ● (608) 622-3620 ● LinkedIn ● Personal Website

EDUCATION

University of Chicago Expected Dec 2024
Master of Science in Applied Data Science | GPA: 4.0/4.0

University of Wisconsin-Madison May 2022
B.S. in Computer Science & B.S. in Data Science | GPA: 3.59/4.0

TECHNICAL SKILLS

- **Programming**: **Python** (Pandas, NumPy, Scikit-Learn, Pytorch, TensorFlow), **Java, JavaScript**, **SQL,** HTML, CSS, React, R, C/C++
- **Data Tools**: **BI Tool(Tableau, Power BI**, **Looker), Cloud databases, AWS Cloud** Spark**.** Azure, Git, VBA, Computer Vision
- **Data Analytics:** Statistical analysis, Excel (Pivot Tables, advanced filtering, charting), A/B Testing, Data Visualization, Database

PROFESSIONAL EXPERIENCES

Capstone Data Scientist Chicago, IL
University of Chicago & Royal Cyber Inc. Mar 2024 - Present

- Engineered a personalized recommendation engine leveraging **Large Language Models (OpenAI GPT and LLaMA)** within the LangChain framework, significantly enhancing product suggestion relevancy and user satisfaction in e-commerce platforms.
- Developed an advanced **chatbot** using **Retrieval-Augmented Generation (RAG)** technology, facilitating dynamic information retrieval and improving user interactions by integrating generative AI models.
- Implemented the Pretrain, Personalized Prompt & Predict Paradigm (P5) in conjunction with a **Vector Database** to devise a customer journey framework, enabling seamless navigation through complex E-commerce interactions and increasing customer engagement.

Business Analyst Intern Shanghai, CN
The Trade Desk (Business Intelligence Team) Sept 2022 - Aug 2023

- **Led a team** of 3 to design and implement a web-based automation project using **Python**(**Flask, Streamlit**), **JavaScript**, **HTML**, and **CSS**. Managed all aspects through **GitLab**, including streamlined file uploads, integration with multi-response checks, and enhanced data handling. The initaitve helped **150+** internal stakeholders to save **+90%** of manual precessing time.
- Deployed an **E2E data pipeline** and developed and maintained full-cycle analytics dashboards through **Tableau** to uncover business opportunities, which guide decision-making process, securing an **$88 million+** deal with Samsung in the US.
- Built **ETL** process infrastructure using **SQL in Vertica (relational database)** and analyzed data from **Google Analytics 360** followed by requests from **JIRA** to improve workflow efficiency. Created visualizations such as waterfall charts, KPI cards, and drill-down reports using **Power BI**, presented to senior leaders and external partners.
- Created a custom **ad-hoc reporting** platform using **Python and Streamlit**, allowing users to generate tailored performance reports for internal stakeholders. Presented analysis and recommendations to senior management.

Data Scientist Intern Shanghai, CN
Tencent (Youtu Lab) Jun 2021 – Sep 2021

- Implemented advanced NLP techniques to analyze 200K unstrunctured text data, performing topic detection with **LDA** and **BERTopic**, **sentiment analysis** with a custom logistic model, and entity identification using Big Data Cloud Computing, yielding detailed insights into industry-wide AI sentiments and trends.
- Developed and optimized an automated image classification system using **TensorFlow and Keras**, processing over 50,000 labeled images with data augmentation and advanced CNN architectures like ResNet and VGG, achieving 92% accuracy.

Data Analyst Intern Hangzhou, CN
Alibaba Group (Freshippo Grocery) *Aug 2020 - Oct 2020*

- Built a user **churn prediction** model using **Random Forest** and **XGBoost** through Python's Scikit-Learn, achieving **+95%** accuracy. Further validated its effectiveness through **A/B testing**, resulting in a **5PP** improvement in user **retention** during customer reengagement campaigns with promo codes.
- Utilized **K-means** clustering and **Apriori** algorithms to analyze **over 3 million** records of customer transaction data, demographic data, and product SKUs. Deployed a customer classification system based on shopping patterns (segmented customer into 5 tiers) and provided product recommendations tailored to individual preferences via a searching dashbaord.
- Collaborated with **cross-functional** and **marketing** teams to launch a series of **targeted** campaigns through **A/B testing**, leveraging previously defined **power user** classification. This initiative led to **10%** increase in repeated purchases (**stat sig**).

PROJECT EXPERIENCES

Time Series Demand Foresting for Divvy Bike Usage in Chicago Apr 2024 – May 2024

- Analyzed four years of data (2019-2023) to develop and implement predictive models **(SARIMA, Prophet, Orbit, LSTM)** using Python, enhancing bike availability and reducing operational costs.
- Model Optimization: Identified LSTM as the optimal model for its accuracy in capturing complex demand patterns, significantly improving forecast reliability.

- Operational Improvements: Enabled precise bike stock management and improved service delivery during peak demand periods, boosting customer satisfaction.
- Technical Expertise in Machine Learning: Utilized Python, Bayesian Optimization, and various machine learning algorithms to refine forecasting accuracy, contributing to data-driven decisions that bolstered business operations.

Computer Vision Project: Automated Image Classification System Mar 2024 - Apr 2024

- Utilized a dataset of over 50,000 labeled images, applying data augmentation techniques to increase the robustness and generalizability of the model. Preprocessed images by resizing, normalizing, and applying transformations to improve model performance.
- Implemented and fine-tuned **convolutional neural networks (CNNs)** using **TensorFlow** and **Keras.** Experimented with various architectures, including ResNet, Inception, and VGG, to identify the optimal model.
- Trained the model on a high-performance computing cluster, employing early stopping and learning rate scheduling to prevent overfitting. Achieved a classification accuracy of 92% on the test set.

Credit Card Fraud Detection System using Deep Learning Apr 2024 - May 2024

- Developed a deep learning model to identify and classify fraudulent transactions using a dataset with over 30k rows, characterized by extreme imbalance (0.17% fraudulent transactions).
- Implemented both a **Multilayer Neural Network (NN)** and a **Convolutional Neural Network (CNN)** leveraging techniques such as Batch Normalization, Dropout, L2 Regularization, Early Stopping, Bayesian Optimization, and Random Search.
- Conducted extensive data preprocessing including random under-sampling to balance the dataset, feature engineering through PCA transformation, and normalization features using StandardScaler.
- Achieved significant improvements in financial security and customer trust by enhancing the detection of fraudulent transactions, using advanced layers such as Conv1D, Flatten, and MaxPooling1D, reaching **97.50% accuracy and 100% precision** without overfitting.

AI Trend Analysis on Industries with Natural Language Processing Feb 2024 - Mar 2024

- Conducted a comprehensive analysis on the impact of AI across various industries, identifying both positive and negative effects and sentiments towards AI.
- Analyzed approximately 200K news articles using **advanced NLP techniques**, including article clean-up and filtering with regular expressions and lemmatization.
- Performed topic detection with **Latent Dirichlet Allocation (LDA) and BERTopic**, and **sentiment analysis** using a Customized Logistic Model with Yelp review/News data.
- Utilized Google Vertex AI for **entity identification (NER),** providing a detailed overview of AI trends and sentiments.

Amazon Review Big Data Analysis Nov 2023 - Dec 2023

- Utilized Hive for initial data ingestion and transformation for over 200,000 Amazon review, uploaded to **Google Cloud Storage (GCS)** and extracted insight using **Google Cloud's BigQuery**, enabling seamless access to large datasets and facilitating quick data retrieval and visualization.
- Utilized Spark NLP's pretrained model for sentiment analysis and Apache Spark's ALS algorithm for developing a personalized product recommendation system based on user preferences and behavior.
- Applied K-Means clustering to segment Amazon customers, enhancing marketing strategies by targeting distinct customer groups based on their purchasing behaviors and preferences.

Chicago Crimes Big Data Exploration and Analysis using PySpark Nov 2023 - Dec 2023

- Utilized Hive for initial data ingestion and transformation from a 5 GB dataset uploaded to Google Cloud Storage (GCS) and extracted insight using **Google Cloud's BigQuery**, enabling seamless access to large datasets and facilitating quick data retrieval and visualization.
- Implemented PySpark API for advanced data exploration, transforming and partitioning the dataset for optimized analysis.
- Developed Data Visualizations: horizontal bar charts and heatmaps in Python to visualize community crime counts, crime type distributions and monthly time series and year-over-year comparisons for top crime types.

APAC Region Campaigns KPI Real-Time Interactive Dashboard Jan 2023 - Mar 2023

- Engineered an automation data pipeline integrating 10M+ rows from AWS S3 and and **third-party APIs** using AWS Redshift.
- Scheduled regular data updates for **AWS Redshift** and the **relational database Vertica** to maintain data freshness. These updates were seamlessly connected to Tableau, ensuring the dashboards reflected the most current data.
- Designed and implemented a performance tracking system with **interactive dashboards using Tableau**, providing insights into key KPIs and metrics, which improved decision-making efficiency by 30%.

Hotel Cancellation Rate Prediction Nov 2022 - Feb 2023

A end to end analysis to prodict hotel cancellation rate through Python

- Explored and pre-processed 1 million+ records and 100+ features, conducted feature engineering through PCA.
- Performed exploratory analysis and visualized behavioral difference between cancelled and active customers.
- Implemented **Random Forest**, **Boosting**, and **SVM** models and fine-tuned model parameters using random search and grid search. Optimized the best-performing model through parameter tuning and achieved a final model with 95% accuracy.