Zehui (Bella) Gu

guzehuibella@gmail.com • (858) 888-2178 • www.linkedin.com/in/zehuigu • New York, NY / Cambridge, MA

EDUCATION

New York University — New York, NY
Master of Science in Data Science, GPA 3.92 — 09/2022 - 05/2024
Coursework: Time Series Analysis, Machine Learning, Big Data, Natural Language Processing, Computer Vision

Harvard University — Cambridge, MA
Bachelor's Degree in Mathematics — 09/2017 - 05/2021
Coursework: Data Science & Machine Learning, Computer Science, Optimization, Probability & Statistical Modeling

SKILLS

Programming: Python, SQL, R, MATLAB
Framework & tools: Git, Linux, Pandas, Scikit-learn, PyTorch, HuggingFace, Hadoop, Spark, AWS, Tableau, Power BI
Machine Learning: Regressions, Decision Tree, Random Forest, unsupervised learning (k-means, PCA), Deep Learning

PROFESSIONAL EXPERIENCE

Delt4 — Cambridge, MA
Data Scientist Intern — 06/2023 - 08/2023
Topic Modeling and Integration with Large Language Model (LLM)

- Conducted deep textual analysis on biochemical papers, creating knowledge clusters using **clustering** algorithms like **BERTopic, t-SNE+HDBSCAN, Top2Vec**; Extracted keywords to label clusters and visualized insights through **word clouds**
- Integrated custom clustering algorithms with **PaperQA** tool to enable targeted Q&A for biologists, boosting research efficiency and achieving a **7x reduction** in cost
- Developed a multi-source web scraper using **Selenium** and **Requests** to augment paper collection methods, further streamlining the process through API integrations

Clinical Trial Outcome Prediction

- Engineered features on unstructured data, generating embeddings for molecule structures, diseases and trial protocols
- Developed and optimized machine learning models (**XGBoost, CatBoost, LightGBM**) with **grid search** to predict clinical trial outcomes across Phase 1/2/3 trials, achieving an overall **84% accuracy**

TikTok — Shanghai, China
Data Analyst — 10/2021 - 06/2022

- Performed Exploratory Data Analysis on transaction data using **SQL** and **Python pandas**, and applied RFM metrics to engineer features for Customer Lifetime Value **(CLV)** modeling via BG-NBD and Gamma-Gamma methods
- Utilized CLV predictions to translate data into **customer segmentation** and targeted retention strategies
- Leveraged segmentation analysis for client cohort identification, conducted web scraping to compile a list of potential clients, resulting in a **20% increase in regional sales** and improved market penetration
- Developed real-time, interactive **dashboards** to visualize core business **KPIs**, providing actionable insights to executives and business teams, using Python, SQL, Aeolus (TikTok's in-house **BI** platform)

Alibaba Group — Beijing, China
Data Analyst Intern, Digital Media & Entertainment Department — 07/2019 - 08/2019

- Analyzed video streaming data using SQL and Python, and built ML models to predict customer churn (**logistic regression, random forest**), with an 81% accuracy score
- Collected data via Python web scraping with **Beautiful Soup** for designing department recruitment plans

SELECTED PROJECTS

Automated Assessment of Epilepsy Using Patient Language Data — 10/2023 - 01/2024

- Automated the creation of a validated fact dictionary using **TF-IDF** and **KeyBERT** from transcribed speech. **Fine-tuned** BERT models to predict epilepsy diagnosis, reaching 83% accuracy
- Implemented Auto Speech Recognition (**ASR**) with **WhisperX** to expedite the research, reducing the error rate by 5-fold

Walmart Sales Forecasting — 10/2022 – 12/2022

- Utilized **ARIMA, Prophet**, and **Gaussian Process** time series models to forecast Walmart's weekly sales and fine-tuned hyperparameters via cross validation and grid search
- Achieved a Mean Absolute Percentage Error (MAPE) of less than 3% for the best model, indicating high accuracy

Improved Text Summarization on Financial Text Data (NLP) — 03/2023 - 05/2023

- Implemented **BART** and **T5** models for initial **abstractive summarization** on financial earnings call transcripts
- Integrated FinBERT embeddings with TextTiling techniques to add **segmentation** tasks, outperforming the state-of-the-art model with a **1.73% increase in ROUGE** scores

Spotify Music Recommender — 10/2019 - 12/2019

- Designed a collaborative-filtered **ALS** matrix factorization model in **PySpark**, processing billions of listening records to learn latent factors for personalized song recommendations, with an **8x improvement** on precision@k to the popularity baseline