Kaibo Zhang

+1 (438)8650693 | kb.zh@outlook.com | 1188 Saint Antoine O H3C1B4 | LinkedIn | Personal Website | GitHub

EDUCATION

MCGILL UNIVERSITY Montreal, Canada

*B.Com. in **Business Analytics,** Desautels Faculty of Management* Aug. 2022 – May 2025 (Expected)

- *Minor in **Statistics,** Department of Mathematics and Statistics* GPA: **3.96/4**
- Coursework: ETL Workflow, Data Mining, Optimization, Data Structure, Machine Learning, Statistical Modelling, Simulation, Time Series Forecasting, Linear Algebra, Multivariate Calculus, Probability, etc.
- Scholarships and Awards: Dean's Honour List – McGill (top 10%), Accenture Prize in Management ($1,590)

PROFESSIONAL EXPERIENCE

Research Assistant at Professor Kyunghee Lee's Lab Montreal, Canada

ML/DL Model for Patent and Tech Company in the U.S. Jan. 2024 – Now

- Constructed a Regression model based on patent data and the corresponding vector space from the extracted coefficients to realize mapping firms' relative spatial relationships for comparison analysis.
- Collected patents and tech companies' data to create Entity Embedding and built a vector database and 2 ETL pipelines for data preparation and model building.
- Coded pre-processing Python scripts for data cleaning and joining individual patent dataframes, aggregating company names as unique IDs and storing patent-holding information in dictionaries.
- Collaborated with Professor Lee to research and design novel validation methods, boosting the efficiency and effectiveness of the resulting vector space.

Machine Learning and AI Experimental Lab Intern Shenzhen, China

Bosera Asset Management Co. Limited May. 2024 – Aug. 2024

- Designed and implemented relational database and APIs with SQLAlchemy and FastAPI to facilitate efficient LLM data retrieval, achieving 85% critical information retrieval accuracy and outperforming sample responses.
- Researched and enhanced Retrieval Augmented Generation performance with Landmark Embedding and tree-organized Natural Language Retrieval workflow, improving semantic and answering accuracy by 45%.
- Integrated K-Means algorithms into Multi-Agent workflow using Python to extract and group performance metrics from research reports to compare companies' performance in different time horizons with LLM.
- Collaborated with 4 senior Data and AI Engineers in drafting AI and Agent Engineering Cookbook and leading 12 peers for task arrangements and progress communications.

Data Analyst Intern Remote

VariFlight Technology Ltd Sept. 2023 – Dec. 2023

- Utilized Mixed-Integer Optimization and sensitivity analysis on resource constraints to discover revenue management insights for airlines and airports, projecting a 2-4% expected increase in simulated profit.
- Designed 3 MySQL relational databases and built 4 ETL pipelines with Python to load API data into databases for model training and deployment, extracting, managing, and cleaning data for weekly report publications.
- Implemented interactive features for 2 existing visualization dashboards with Power BI and Python, simplifying graphs and enhancing readability.

Data Analyst Intern at Online Ride-hailing Area (Urban Central Platform) Tianjin, China

Beijing DiDi Infinity Technology and Development Co. May. 2023 – Aug. 2023

- Built data automatic processing models that execute SQL queries to update datasets from source data at preset times, resulting in a 30% increase in efficiency.
- Cleaned 5 datasets with SQL and built data visualization and Regression models with Excel for online ride-hailing operational needs, unifying data formats and helping to put 13% of stale data into use.
- Led and delegated tasks for colleagues to call online ride-hailing to collect first-hand data manually and used hypothesis testing for measuring price-related market standings with competing brands.

PROJECTS & COURSEWORK

Bridging Emotions and LLMs: From Random Forest to BERT in Emotion Classification GitHub

Sentiment Classification | RF | Naïve Bayes | LLM Fine-Tuning

- Fine-tuned BERT and GBT2 models with Transformers for 28 class emotion classifications on GoEmotions, achieving a test accuracy of 63.03% and 59.59%, respectively.
- Conducted in-depth attention mechanism analysis of BERT, visualizing token-level importance and identifying model limitations in handling conflicting sentiments within texts.

- Trained a Random Forest model as the baseline and implemented Multinomial Naïve Bayes from scratch with Python, achieving a test accuracy of 44.49% with optimized hyperparameters on the validation set.

Exploring Neural Network Architecture Choices on OrganAMNIST Images GitHub

Image Classification | MLP | CNN | Pre-Trained Model Tuning

- Fine-tuned ResNet101 and ViT models with PyTorch to compare convergence and accuracy performance with the self-designed CNN model.
- Implemented and designed a CNN with two convolutional layers and one dense layer with PyTorch, achieving a test accuracy of 89%.
- Implemented MLP with Python from scratch and explored the impact of architectural decisions on neural networks, achieving 75.75% accuracy despite the algorithm's disadvantages in image classification.

Traffic Prediction and Optimization of Bike-sharing System in New York City GitHub

Time Series | Bayes Rules | Network Analysis | K-Means | Regression | Optimization

- Developed a Regression model to analyze bike-sharing demand dynamics, building on prior HKUST research.
- Utilized the Adjacency Matrix to capture relationships between individual stations and applied a K-Means model to group stations into 3 clusters to enhance result generalizability.
- Conducted Time Series forecasting using Conditional Probabilities to predict people's bike travel habits.
- Implemented a Linear Programming model to optimize resource allocation for the bike-sharing system in NYC.

LinkedIn Database Replication Design GitHub

Relational Database Design | DDL | DML | MySQL

- Conducted a comprehensive analysis and replicated databases and 85% of its original functionalities.
- Developed conceptual, relational, and physical database models using a top-down approach with MySQL.
- Crafted 20 advanced SQL queries to demonstrate the database's capability to support complex user operations.

Social Speculation for Harnessing Reddit to Forecast Bitcoin Fluctuations GitHub

ETL | API | NLP | MLP

- Built two ETL workflows with Python using Alpha Vantage and Reddit APIs to gather and store relevant data.
- Transformed textual comment data into sentiment and subjectivity scores with TextBlob for deeper analysis.
- Trained an MLP model using Pytorch to predict Bitcoin close price, improving baseline MSE by 22.62%.

Regression Analysis on the Factors Affecting the Monthly Spending of Desautels Students GitHub

Data Collection | Exploratory Analysis | Multiple Regression

- Conducted a statistical analysis to identify factors influencing students' monthly spending.
- Collected more than 300 records of first-hand data using a self-administered questionnaire on Qualtrics.
- Performed data preprocessing, exploratory analysis, hypothesis testing, and multiple regression with Excel and concluded that dine-out, groceries, and subscriptions were the significant factors influencing monthly spending.

LEADERSHIP & EXTRACURRICULAR

Teaching Assistant Montreal, Canada

MGCR 271 Business Statistics &. MGSC 404 Founds of Decision Analytics Aug. 2023 – Now

- Prepared sample final exam solution explanations, helping students apply different probability distribution models under the right context.
- Held tutorials, and designed and graded assignments and exams, innovating an intuitive way to explain the Z-score notion by making analogies with the unit transformation between meters and centimeters.
- Streamlined communication between students and the professor with an online forum, efficiently managed inquiries, and resolved issues, saving supervisors significant time and effort in administrative tasks.

VP of Finance Montreal, Canada

McGill Data Network May. 2023 – Now

- Spearheaded the development and management of the club's annual budget, ensuring financial sustainability and responsible allocation of resources for cost-saving and expenditure optimization.
- Maintained accurate financial records and facilitated data-driven decision-making for projects and activities.
- Collaborated with McGill professors and professionals from BNP and EY to hold panel discussions and other interactive workshops on analytics techniques to connect the student body with the analytics industry.

SKILLS&ACTIVITIES

- Technical Skills: MySQL, Neo4j, R, Python (Pandas, Sktlearn, Requests, Matplotlib, etc.), Minitab, Git, Slurm, JavaScript, Java, HTML, CSS, Canva, Figma, Microsoft Office Suites
- Language Skills: English and Mandarin
- Interests: Photography, Guitar, Alto-saxophone, and Piano