Miao **Fu**

Data Scientist · Machine Learning · Big Data Engineer

No.78, Section 3, Furong Middle Road, Changsha, Hunan Provience, Rep. of China

(+86) 185-6949-9324 | miaofu1997@gmail.com | miao-fu.com

"Torture the data, and it will confess to anything."

Education

Hunan University of Technology and Business (HUTB) — *Hunan, China*

B.S. in Information and Computing Science, GPA 3.43/4 (Top 7%) — *Sep 2015 - June 2019*

- **Mathematics**: Mathematical Analysis, Differential Equations, Operations Research, Analytic Geometry
- **Computer Science**: Data Structure, Database Systems, Operating System Engineering, Introduction To Algorithms, Machine Learning

Brown University — *Rhode Island, USA*

Summer Session — *June 2017 - Aug 2017*

- **Mathematics**: Linear Algebra, Probability and Statistical Inference

Skills

Programming	Python, Go, Scala, SQL, Bash, JavaScript, Ruby, LaTeX
Frameworks	Hadoop, Spark, Flink, TensorFlow, PyTorch, Vue, Jekyll
Tools and Platforms	Git, Hive, Clickhouse, MySQL, Kafka, HBase, MATLAB, ChatGPT
Languages	Chinese (Native), English (Proficient, capable of reading and writing academic papers), Spanish (Beginner)

Experience

Samoyed Cloud Technology Group Holdings Limited — *ShenZhen, China*

Intelligent Advertising Decision System — *Mar. 2021 - Aug. 2023*

- Designed and developed a performance calculation process for different MOB (Month on Book) scenarios, aggregating profit and loss data by month across 11 advertising channel types. Successfully processed data for over 3,000 channels monthly, significantly improving data processing efficiency and accuracy.
- Trained linear regression (LR) with growth rate scaling models to accurately predict the cost metrics' trends, assessing the profit and loss performance of each advertising channel for future periods. The prediction accuracy exceeded 93%, effectively supporting the optimization of future advertising decisions.
- Filed 2 patents related to the computational process, enhancing the company's competitive edge in technological innovation.

Device Fingerprinting

- Processed device fingerprint data streams from Kafka using Spark for distributed batch prediction tasks. Trained a triplet neural network to generate high-dimensional feature vector fingerprints, which were then matched in the Cubetree vector database.
- After deployment, the system handled 10-20 millions of data entries daily, achieving a matching response time within 50 milliseconds, with an accuracy rate of 96.8% in single scenarios and 71.24% in cross-scenarios, saved the company 400,000 RMB per year by replacing the previously purchased device fingerprinting system.
- Utilized FlinkSQL to import millions of new fingerprint records daily into Clickhouse, establishing and maintaining materialized views to achieve sub-second query response and improve data storage efficiency.
- Developed monitoring systems for key performance metrics of the device fingerprint system, including Accuracy, Precision, Recall, and F1 score.

User Segmentation and Analysis for the Shengbei App

- **User Persona and Conversion Rate Analysis**: Analyzed the conversion rate performance of borrowers in terms of login, application completion, credit approval, and transactions, based on data from the Shengbei App between 2019 and April 2023. Calculated the conversion ratio for over 10 segmented user groups. These analyses identified higher-quality borrower groups, leading to adjustments in marketing and lending strategies that successfully improved customer quality and reduced customer acquisition costs.
- **The "Gold Mining" Plan**: Analyzed the last login times of over 3 million inactive users and segmented them based on login time and attributes such as application status, credit rejection, gender, and occupation. By targeting these segmented user groups through marketing campaigns, we increased the activation rate of existing users by approximately 13%, successfully achieving re-engagement goals.
- **User Retention and LTV Analysis**: Extracted login, transaction, and revenue data for various user groups that failed to complete transactions (incomplete applications, approved credit but no transactions, and rejected) from the past 30 days. Calculated 30-day retention rates and average revenue per user, and accurately assessed user lifetime value (LTV), providing data support for optimizing user management strategies.

Intermediate Revenue Analysis for Shengbei App

- **Membership Card Daily Report**: Developed an automated daily report tracking key metrics for membership cards, including click-through rates, activation revenue, payment success rates, and refund rates. Integrated results into the Guanyuan BI system, significantly improving operational efficiency.
- **Subprime Lending Referral Project**: Developed a user segmentation model for rejected users on our App, labeling each group and pushing customized landing page lists for third-party subprime lending platforms. Through precise segmentation, personalized targeting, and A/B testing optimization, the registration rate on third-party platforms increased by approximately 8% per month, and the transaction initiation rate improved by about 5%, significantly boosting project revenue and profit growth.

Big Data Cluster Maintenance

- Responsible for daily maintenance and inspection of the big data cluster, ensuring stable operation under high concurrency and large data volumes.
- Conducted in-depth analysis and optimization of slow HiveSQL queries, significantly improving HiveSQL execution efficiency.
- Addressed storage issues and processing bottlenecks caused by small Hive files, merging them to reduce I/O overhead and enhance data read/write performance, notably improved overall cluster performance and resource utilization.
- Developed and maintained Bulkload task scripts for importing risk control input data from Hive to HBase. This significantly improved the efficiency of data ingestion and retrieval under high concurrency.

China Electronics Corporation (CEC) Industrial Internet Co., Ltd.

Hunan, China

Big Data Engineer — *Apr. 2019 - Jan. 2021*

- Built, deployed, configured, and maintained distributed big data clusters, including Hadoop, HBase, Zookeeper, Kafka, and Spark clusters on KylinOS, including testing and ongoing maintenance.
- Enhanced the Expectation-Maximization (EM) algorithm by integrating Markov Chain Monte Carlo (MCMC) methods with random variables, efficiently imputing and correcting missing data for concrete pump truck operations. This approach ensured sample estimates converged to true values, effectively addressing missing data issues.

Hunan Province Mobile Business Intelligence Lab

Hunan, China

Assistant Researcher — *Oct. 2018 - June. 2019*

- Responsible for implementing the Random Walk with Restart (RWR) recommendation model within a complex heterogeneous network, analyzing node similarity within the network and delivering high-quality collaboration recommendations.
- Used Matlab and Python to conduct exploratory data analysis, visualization, simulation, and modeling with complex network data of 1,3000 researchers, 350,000 articles, and 2,333,636 citations from ResearchGate.

Honors & Awards

Domestic

Nov. 2017 **National First Prize**, Contemporary Undergraduate Mathematical Contest in Modeling (CUMCM) — *Hunan, China*

Jun. 2019 **Outstanding Graduates**, Hunan University of Technology and Business — *Hunan, China*

International

Apr. 2018 **Honorable Mention**, Mathematical Contest in Modeling (MCM) — *MA, USA*

Papers

Social Recommendation With Large-Scale Group Decision-Making for Cyber-Enabled Online Service

IEEE Transactions on Computational Social Systems — *Oct. 2019*

- vol. 6, no. 5, pp. 1073-1082, Oct. 2019, doi: 10.1109/TCSS.2019.2932288