CHENFENG LI

Chicago, IL 60637 | (872)-215-0270 | cfli@chenfengli.com | https://chenfengli.com

EDUCATION

MS **Statistics** | **University of Chicago** (GPA: 3.73/4) Sep 2022 – Jun 2024

Relevant Courses: Reinforcement Learning, Trustworthy Machine Learning, Deep Learning Systems
Scholarships: Tuition Scholarship of Statistics Master Program (2022, 2023)

BS **Mathematics** | **Chinese University of Hong Kong (CUHK)** Sep 2018 - Jul 2022

Major Concentration: Computational Big Data Analytics; Minor: Statistics
Scholarships and Honors: BS degree with First Class Honor (2022), Undergraduate Mathematics Scholarship (2021), College Scholarships (2019, 2022)

SKILLS and CERTIFICATIONS

Programming: Python, R, SQL, DAX, C, C++
Technical Skills: Machine Learning (TensorFlow, PyTorch, scikit-learn), LLM, Data Processing (Excel, Pandas), Visualization (Matplotlib, Power BI, Tableau), Databases and Data warehousing, Cloud Platform (AWS, Azure), Data pipeline and processing tools (Airflow, Kafka, Spark), Statistical Analysis, Algorithms
Language: English (Fluent), Mandarin Chinese (Native), Cantonese (Native)
Certifications: AWS Certified Cloud Practitioner (AWS, 2024), Azure Data Engineer Associate (Microsoft, 2024), Power BI Data Analyst Associate (Microsoft, 2024), Excel 2019 Associate (Microsoft, 2024), Google Advanced Data Analytics (Google, Coursera, 2023), Deep Learning Specialization (DeepLearning.AI, Coursera, 2023)

WORKING and PROJECT EXPERIENCE

Data Scientist | Synergistic IT, CA, Remote Jun 2023 – Present

Project: Cloud-Based Clinical Workload Monitoring System

Designed and implemented an AWS-based reporting pipeline data pipeline for a Hospital to support daily reporting on provider workload, documentation timing, and shift analysis across multiple clinics.

- Ingested data from Amazon RDS using DMS (full + CDC loads) and stored partitioned csv files in S3. Automated ETL with AWS Glue and PySpark with data aggregation, and alerted via CloudWatch and SNS.
- Loaded transformed data into Snowflake from S3 bucket and enforced data quality through SQL checks. Remodeled the data into star schema with SCD-enabled dimensions.
- Applied column-level masking for PHI to meet HIPAA compliance. Collaborated with DevOps on IAM and KMS integration for secure access and encryption.
- Supported BI analysts in delivering Power BI dashboards with row-level security, enabling HR to monitor workload trends, reduce burnout risk, and inform staffing expansion.
- Reduced reporting latency by over 80% and helped inform staffing reallocation and expansion.

Project: Social Media Sentiment Analysis and Reporting

Partnered with a data scientist to develop an Azure-based system for analyzing 16,000 post-sale reviews sentiment for an online clothing store on social medias and reporting insights.

- Constructed and configured an Azure Data Factory pipeline to ingest data from an on-premises database to Azure Data Lake Storage. Cleaned and filtered data using SQL scripts in Azure Synapse Analytics.
- Designed a notebook in Azure Databricks for data processing, analysis, and visualization. Implemented NLP models for text cleaning, lemmatization, stop-word removal, and performed sentiment intensity analysis a BERT-based model. Trained a random forest model to identify significant words for different sentiments.
- Visualized results by time and item and created comprehensive reports. Concluded a result of 96% of non-negative reviews and uptrend over time.
- Updated the dataset with a scheduled trigger in Data Factory on a weekly basis for long-term use.

Project: Predictive Sales Analytics Platform

Developed a machine learning model to predict total sales for each product and store of a retail chain for the upcoming month using daily historical sales data, to provide valuable forecasts for strategic management.

- Implemented feature engineering on the dataset. Created text-based features by performing TF-IDF and matrix factorization on item and shop names, and lagged and trend-based features for time series analysis.
- Conducted Exploratory Data Analysis, including visualization of target distribution and time trend.
- Constructed and trained pipelines with Ridge, XGBoost and LightGBM regressors. Applied feature selection using RFECV and optimized hyperparameters using Bayesian optimization. Evaluated the above models.
- Predicted the future outcomes and compiled comprehensive reports.

Project: Sport Corporation Sales Analysis

Developed an advanced Power BI dashboard to analyze sales data for an international sports corporation, providing real-time insights into sales performance, discount analysis, and regional success.

- Created and filtered data using SQL on an on-premises dataset, leveraging a star schema data model with the Sales table at the center for optimized performance.
- Cleaned and transformed data using Power Query, ensuring data accuracy and consistency. Implemented advanced DAX formulas to create columns and measures for in-depth analysis,
- Designed a one-page interactive dashboard with key metrics, incorporating various visualizations for comprehensive sales insights and using bookmark to toggle between views.
- Published the dashboard to the Power BI service and enabled scheduled refresh to ensure real-time data updates. Provided training and support to users for effective utilization and interpretation.

Statistical Consultant | Department of Statistics, UChicago Sep - Dec 2022

Worked in a team of five consultants. Analyzed requirements from clients about data issue. Communicated with clients to verified details. Provided recommendation in data analysis and delivered consulting report.

- Suggested logistic regression application and method of grouping the patient data for a study from UChicago Medicine about the impact of a COVID medication on ventilation.
- Recommended a study from UChicago BSD about the effect of Home-based Community Services (HBCS) on Post-Acute Care (PAC) to use logistic regression without propensity score weighting.
- Advised a study from UChicago Hospital about significant of chest-to-left ventricle distance on CPR to drop highly correlated covariates. Helped determine the required sample size and linear regression models.

RESEARCH EXPERIENCE

Independent Researcher | Master's Thesis, Department of Statistics, UChicago Sep 2023 - May 2024

Examining the Interplay Between Politicians' Facial Expressions in Media Images and News Corporation Bias.

- Constructed a name recognition model with politicians' image from Wikimedia to identify news photos.
- Analyzed facial expression logits of influential politicians from various media outlets.
- Implemented classification models and visualized results through dimensionality reduction on the logits.
- Concluded no significant effect of media orientation on facial expression selection.

Research Team Member | Department of Computer Science, UChicago Oct - Dec 2023

Modified Attention with Non-Linear Kernels and its Impact on Few-Shot Learning.

- Collaborated within a three-fellows team. Trained GPT-2 models on nanoGPT using OpenWebText with replacing the dot product kernel in attention mechanism by Gaussian, polynomial and periodic kernels.
- Evaluated the models with MMLU, ARC and Translation tests. Determined that traditional dot product kernels performed best overall, with some non-linear kernels excelling in specific tests.

Project Leader | Department of Statistics, UChicago Apr - May 2023

Robustness to Spurious Correlations via Distributionally Robust Optimization (DRO).

- Led a team of three researchers, coordinating tasks and communication.
- Reviewed and analyzed the theory of DRO.
- Applied DRO and empirical models on an MNIST dataset with spurious correlations. Made comparison of the performance. Concluded DRO model effectively eliminates the influence of spurious correlations.