CHENFENG LI

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

Summary

Data Scientist / Data Engineer with 2+ years of experience in building **end-to-end data solutions across data engineering, machine learning, and business intelligence**. MS in Statistics from the University of Chicago. Proficient in **Python**, **SQL**, **R**, **Power BI**, and **Excel**. Certified **AWS Cloud Practitioner**, **Azure Data Engineer Associate**, and **Microsoft Certified Power BI Data Analyst Associate**. Hands-on expertise in designing **cloud-based ETL pipelines** and **data warehouse architectures** using **AWS**, **Azure** and **Snowflake**. Experienced in implementing **data modeling**, **data quality validation**, and **security controls**. Skilled in **machine learning** and **deep learning**, applying techniques such as forecasting, NLP-based sentiment analysis, and feature engineering, with experience using **Scikit-learn**, **PyTorch**, and **Hugging Face**. Strong background in **statistical analysis** (GLM, Bayesian methods, time series) and **algorithmic problem-solving**. Adept at creating interactive **Power BI** dashboards with advanced **Power Query** and **DAX** for actionable business insights.

Skills & Certifications

- **Programming Languages:** Python, SQL, R, DAX, C, C++
- **Database/Data Warehouse:** MySQL, PostgreSQL, Microsoft SQL Server, AWS RDS, Snowflake, Azure Synapse Analytics, Google BigQuery
- **Cloud Platform:** AWS (S3, Glue, RDS, DMS, CloudWatch, CloudFormation, IAM, SNS), Azure (Data Factory, Databricks, Synapse Analytics), GCP (BigQuery, Cloud Storage)
- **Data Processing/Orchestration:** Apache Spark, PySpark, Apache Airflow, Apache Kafka, AWS Glue, Azure Data Factory
- **Data Visualization/Report:** Power BI, Tableau, Excel, Power Query, DAX, Python (Matplotlib, Seaborn), R
- **Libraries:** NumPy, Pandas, SciPy, Matplotlib, Seaborn, PySpark, Boto3, Scikit-Learn, PyTorch, TensorFlow, Hugging Face, NLP toolkits, GenAI APIs
- **ML/AI Techniques:** Regression (Linear, Logistic, Ridge), Decision Tree, Random Forest, KNN, SVM, KMeans, Ensemble Model (XGBoost, LightGBM), Neural Network (DNN, CNN), NLP (text preprocessing, sentiment analysis), Feature Engineering
- **Statistical Analysis:** GLM, Bayesian Inference, Time Series, Non-Parametric statistics
- **Development/Version Control:** Jupyter Notebooks, RStudio, Visual Studio Code, Git, GitHub

Certifications

- **Power BI Data Analyst Associate** (Microsoft) – Data modeling, visualization, and interactive dashboard design
- **AWS Certified Cloud Practitioner** (AWS) - Cloud computing and infrastructure with AWS services
- **Azure Data Engineer Associate** (Microsoft) – Data integration, transformation, analysis with Azure services
- **Microsoft Office Specialist: Excel 2019 Associate** (Microsoft) – Data management and spreadsheet design
- **Google Advanced Data Analytics** (Google, Coursera) – Large datasets, data analytics, machine learning
- **Deep Learning Specialization** (DeepLearning.AI, Coursera) – Neural networks, Transformers and application in industry

Employment History

SynergisticIT, Fremont, CA
Data Analyst / Data Engineer
Project: Cloud-Based Clinical Workload Monitoring System

January 2025 – Present

Designed and implemented a **cloud-based data pipeline and analytics platform** for a hospital to consolidate clinical, HR, and scheduling data for daily reporting on provider workload, documentation timing, and staffing efficiency. The system integrated **AWS-based ETL pipelines**, a **Snowflake** data warehouse, and **Power BI** dashboards to deliver near real-time insights for HR and clinical operations, enabling proactive workload balancing and reducing reporting latency by over 80%. presenting key metrics in an interactive and user-friendly format.

Roles and Responsibilities:

- Designed and maintained a multi-stage **ETL pipeline** leveraging **AWS DMS** to capture daily CDC updates from data from multiple systems in **Amazon RDS** into a partitioned **raw S3** zone, then using **AWS Glue (PySpark)** to transform and standardize datasets into a curated **processed S3** zone for efficient loading into **Snowflake**.
- Implemented **data standardization**, **null handling**, **duplicate** resolution, and **timestamp alignment** to ensure cross-system consistency.
- Performed **SQL-based data quality check** and **remodeled** the data into **star schema** with **SCD Type 2 dimensions** in **Snowflake** to support business-critical reporting.
- Applied **column-level masking** in **Snowflake** and aligned all **data access controls** with HIPAA compliance.
- Automated pipeline orchestration using **AWS Managed Workflows for Apache Airflow** (**MWAA**). Monitored workflows using **AWS CloudWatch** and configured SNS alerts for pipeline failures, or data anomalies.
- Collaborated with BI analysts to define clinical workload metrics, built curated **Snowflake views**, and implemented **row-level security** for department-specific access.
- Developed **Power BI dashboards** for HR and operations teams, visualizing appointment volumes, documentation delays, off-shift activity, and workload distribution by clinic and provider role.
- Enabled HR to proactively identify staff overloads, reduce burnout risk, and improve resource allocation; cut reporting time from 12 hours to under 3 hours, supporting data-driven staffing decisions across multiple clinics.
- Trained users on pipeline and dashboard functionality and interpretation, facilitating adoption for HR team.

Technologies Used: AWS (Glue, S3, RDS, DMS, CloudWatch, SNS, IAM, KMS, CloudFormation, MWAA/Airflow), Snowflake, SQL, PySpark, Power BI, Python,

SynergisticIT, Fremont, CA
Data Scientist
Project: Social Media Sentiment Analysis and Reporting

July 2024 – December 2024

Partnered with a data scientist to develop an Azure-based NLP system for sentiment analysis on 16,000 post-sale reviews for an online clothing store. Built an **Azure Data Factory** ingestion pipeline and implemented an **Azure Databricks (PySpark)** workflow to clean, process, and classify review text using a **BERT-based model**. Visualized sentiment trends and generated automated weekly reports to support product feedback analysis and marketing decisions.

Roles and Responsibilities:

- Configured an **Azure Data Factory** pipeline to ingest data from an on-premise database into **Azure Data Lake Storage**, scheduled for weekly updates.
- Designed an **Azure Databricks** notebook using **PySpark** to perform data cleaning, deduplication and null handling.
- Implemented **NLP** models for text cleaning, lemmatization, stop-word removal, and perform sentiment intensity analysis with **BERT** based model from **Hugging Face** Transformers. Trained a **Random Forest classifier** to predict sentiment labels.
- Evaluated model performance and optimized preprocessing parameters to improve classification accuracy.
- Visualized results by time and item, producing comprehensive sentiment trend reports. Concluded an overall satisfactory rate of 96% and trend upward over time.
- Automated the entire workflow from ingestion to reporting using **ADF triggers** and Databricks job scheduling.

Technologies Used: Python, PySpark, Azure Data Factory, Azure Databricks, Azure Data Lake Storage, Hugging Face Transformers, BERT, Random Forest, NLP

SynergisticIT, Fremont, CA
Data Scientist

Project: Predictive Sales Analytics Platform

January 2024 – June 2024

Developed a **machine learning forecasting model** to predict total sales for each product and store for the upcoming months using daily historical sales data. This involved data preprocessing, feature engineering, and applying models including **Ridge, XGBoost**, and **LightGBM**. The project optimized model accuracy and provided valuable forecasts for inventory management and strategic planning.

Roles and Responsibilities:

- Imported and merged multiple datasets into **pandas** DataFrames, removed duplicates and imputed missing value.
- Engineered numerical, categorical, and text-based features, including **TF-IDF** embeddings and **matrix factorization** of product and store names, lag features, and trend-based **time series indicators**.
- Conducted **Exploratory Data Analysis (EDA)**, including visualization of target distribution and time trends. Used multivariate heatmaps to analyze numerical and categorical pairings.
- Applied **mean encoding** for categorical variables and constructed ML pipelines with **Ridge**, **XGBoost**, and **LightGBM** regressors.
- Performed **feature selection** using **Recursive Feature Elimination with Cross-Validation (RFECV)** and optimized hyperparameters using Bayesian optimization to minimize **Root Mean Square Error (RMSE).**
- Evaluated models with cross-validation and deployed the best-performing pipeline for ongoing forecasting.
- Predicted future outcomes and compiled comprehensive reports. Improved forecasting accuracy by 20% over baseline, reducing inventory mismatches and supporting proactive stocking decisions across multiple stores.

Technologies Used: Python, Scikit-Learn, Machine Learning Pipeline, NLP, TF-IDF, mean encoding, matrix factorization, Ridge Regressor, LightGBM, XGBoost, feature selection, hyperparameter optimization.

SynergisticIT, Fremont, CA
Data Analyst / Business Intelligence Analyst

Project: Sport Corporation Sales Analysis

September 2023 – December 2023

Developed an advanced **Power BI** dashboard for an international sports corporation to analyze sales performance, discount trends, and regional success. Integrated **SQL-based** data extraction, **Power Query** transformations, and **DAX** calculations to deliver real-time, interactive insights. The dashboard was designed to facilitate data-driven decision-making by presenting key metrics in an interactive and user-friendly format.

Roles and Responsibilities:

- Queried and filtered on-premise data using **SQL**, leveraging a **star schema** data model with the Sales table at the center for optimized reporting performance.
- Cleaned and transformed datasets in **Power Query**, ensuring data accuracy and consistency.
- Developed advanced **DAX** measures for fiscal year insights, discount analysis, and update time display.
- Designed a one-page interactive **Power BI** dashboard with key metrics, including total sales, customer counts, product sales, and discount breakdown.
- Enabled **scheduled refresh** for near real-time data updates.
- Published the dashboard to the Power BI service and implemented user access controls.
- Collaborated with stakeholders to understand business requirements and tailor the dashboard to meet their needs.
- Conducted user training sessions to ensure effective use and interpretation of the dashboard.

Technologies Used: Microsoft SQL Server, Power BI services, Power Query, DAX

SynergisticIT, Fremont, CA
Data Analyst / BI Analyst
Project: Retail Chain Transaction Analysis

June 2023 – August 2023

Built a multi-page **Power BI** dashboard to analyze product sales, customer behavior, seasonal trends, and promotion effectiveness for a retail chain. Leveraged Power Query transformations, advanced DAX measures, and interactive navigation features to deliver actionable insights, enabling marketing and sales teams to optimize promotional strategies and inventory planning.

Roles and Responsibilities:

- Utilized **Power Query** to clean and transform data, including splitting and unpivoting the Product column.
- Developed advanced **DAX formulas** to create calculated columns and measures for performance tracking and comparative analysis
- Created dedicated **Product Analysis** and **Customer Analysis** pages with interactive visuals (line charts, treemaps, ribbon charts) and combined them into a **Retail Analysis** page using **bookmarks** for seamless navigation.
- Ensured alignment and consistency across all dashboard pages for a cohesive user experience.
- Presented key findings to stakeholders through **PowerPoint** summaries, driving data-informed decision-making.
- Conducted stakeholder meetings to gather requirements and incorporate feedback into the dashboard design.
- Provided training and support to end-users for effective utilization and interpretation of the dashboard.

Technologies Used: Power BI services, Power Query, DAX, Bookmarks, Microsoft Excel, Microsoft PowerPoint

Department of Statistics, UChicago, Chicago, IL
Statistical Consultant
September 2022 – December 2022

Worked in a five-member consulting team to address analytical challenges for university-affiliated medical and research clients. Gathered requirements, validated details with stakeholders, and delivered statistical recommendations in written reports and presentations.

Roles and Responsibilities:

- UChicago Medicine – Proposed the use of logistic regression and patient grouping methodology for evaluating the impact of a COVID medication on ventilation outcomes.
- UChicago Biological Sciences Division (BSD) – Recommended applying logistic regression without propensity score weighting to assess the effect of Home-based Community Services (HBCS) on Post-Acute Care (PAC) outcomes.
- UChicago Hospital – Identified high collinearity in CPR-related measurements; advised removal of highly correlated covariates, determined required sample size, and developed appropriate linear regression models.

Skills Used: Team collaboration, client communication, statistical modeling, data analysis review, requirements gathering

Education

MS in Statistics | University of Chicago (GPA: 3.73/4) September 2022 – June 2024

BS in Mathematics | Chinese University of Hong Kong (CUHK) September 2018 – July 2022