Chenfei Song

(934) 444-2262 | chenfeisong.cs@gmail.com | www.linkedin.com/in/chenfei-song

SKILLS

PROGRAMMING	Python (TensorFlow, PyTorch, scikit-learn, Pandas, NumPy), OOP, Spark, Scala, R, SQL
SOFTWARE & PLATFORM	AWS, Hadoop, Databricks, Airflow, MLflow, GitHub, MySQL, Hive, Presto, Tableau
MACHINE LEARNING	Time Series, Classification, Regression, Anomaly Detection, Clustering, NLP, LLM, MLOps

PROFESSIONAL EXPERIENCE

Big Data Federation Inc *San Jose, CA*
Data Scientist Full-time *Mar 2022 – Nov 2023*

- Built Python package for inflation number forecast with Linear Regression models that accurately predicted unexpected spikes or drops in 67% of cases during H2 2022, protected portfolio against a single day drop of 3.7% (1.5M USD).
- Developed sector-level trading strategy with classification model (SVM, XGBoost, Random Forest) using Python, protected the portfolio against 8% loss in 2023 (4M USD).
- Develop prompt-based GPT data pipeline in collaboration with software team and defined metrics for job monitoring.
- Created company recommendation tool for both internal stock forecasting and external client utilization, generated 6% more revenue.

Condé Nast *New York, NY*
Data Scientist Full-time *Feb 2021 – Mar 2022*

- Developed a Python package of multi-touch and Markov attribution models to characterize user behavior patterns over 17M users.
- Implemented cross-channel features into user conversion model with PySpark over Databricks, improved offline AUC score by 8%; maintained and monitored model pipeline with MLflow and Airflow.
- Built Tableau dashboard with TabPy integration to automate and visualize attribution model estimates.
- Optimized budget allocation and marketing strategies with user journey analysis in collaboration with marketing team.
- Led 5-member team to develop a prototype platform connecting donors to students lacking tech resources (Hackathon - Top5 Pitch).

Apple Inc *Beijing, CN*
Data Scientist Intern *May 2019 – Aug 2019*

- Engaged 26,000 users through a 5-week fitness app campaign to boost Apple Watch user engagement and customer sales.
- Enhanced customer perception and product engagement by topic modeling (LDA) using user-generated content from social media.
- Developed an automated price-tracking pipeline using web-scraping, slashing vendor workload by 50%.

Amazon.com *Beijing, CN*
Data Analyst Intern *Nov 2018 – Apr 2019*

- Designed an end-to-end return-to-factory logistics workflow and built dashboard for performance monitoring with Looker.
- Automated business reports using ETL to monitor KPI metrics related to package delivery, weight distribution, and customer returns.

EDUCATION

Columbia University, M.S. in Operations Research, GPA 4/4 *New York, NY*
- Courses: Machine Learning, Stochastic Models, Analyzing on AWS, Recommendation Systems *Aug 2019 – Feb 2021*
- Teaching Assistant for Social Network Analysis & Natural Language Processing (NLP)

Tsinghua University, M.S. in Business Analytics, GPA 3.7/4 *Beijing, CN*
- Tsinghua-Columbia Dual Master's Degree Program *Sep 2018 – Apr 2021*
- Courses: Big Data Analysis, Business Intelligence, Statistical Inference

Renmin University of China, B.S. in Statistics, GPA 3.8/4 *Beijing, CN*
- Courses: Time Series Analysis, Operations Research, Regression Analysis, Data Mining, Database *Sep 2014 – Jul 2018*
- Exchange program in **University of British Columbia** *Sep 2016 – Dec 2016*

ACADEMIC PROJECTS

Twitter Sentiment Analysis on AWS (Columbia University) *Jan 2020 – Mar 2020*

- *Goal*: Build a public tweet sentiment prediction API.
- *Data*: 1.6 million labeled user tweets provided in class.
- *Process & Result*: Deployed Twitter Sentiment Analysis API through AWS pipeline (S3, Glue, SageMaker, Lambda, API Gateway), prediction accuracy reached 82.8% (Top1).

Unfollowing Mechanism Study, Data Extraction and Machine Learning (Tsinghua University) *Sep 2018 – Nov 2019*

- *Goal*: Improve understanding and prediction accuracy of users' unfollow behavior.
- *Data*: 2.2 million follow-up actions from Kuaishou (TOP2 live-streaming app in China) users in Sep 2018.
- *Process & Result*: Extracted data and explored features of user-following data with Hive, built machine learning models (Logistic Regression, SVM, Decision Tree and Neural Network) with Python, achieved prediction accuracy of 72.3%.