Yanan Cao

(+1) 415-605-3722 | [yanancao34@yahoo.com](mailto:yanancao34@yahoo.com) | San Francisco, CA | linkedin.com/in/yanancao21 | github.com/LLan-ovoy

# EDUCATION

**UNIVERSITY OF SAN FRANCISCO, San Francisco, CA**

***Master of Science in Data Science*** Jul. 2022 (expected)

Courses: Advanced Machine Learning, Deep Learning, Data Structure and Algorithm, Experiment in Data Science, Distributed Data System (Spark), Relational Databases (SQL), Linear Regression Analysis, Time Series Analysis, Data Acquisition with Python, Product Analytics and Communication

# BEIJING NORMAL UNIVERSITY, Beijing, China

***Bachelor of Art in Management, minoring in Statistics*** Jun. 2021

Courses: Statistical Learning, Multivariate Statistical, Stochastic Process, Probability Theory, Mathematical Statistics, Scientific Computing (Python), Statistics Software (R), Database Principles (MySQL)

# WORK EXPERIENCE

**Walmart Global Tech** Dec. 2021 - Present

***Data Science Intern, Repurchase, Personalization*** San Francisco Bay Area, CA

- Repurchase Prediction

* Pre-processed data using Apache Spark, and developed TIFU-KNN model for predicting items in customer's next order.
* Built TIFU-KNN model from scratch using Apache Spark, achieved 5% improvement in precision compared to Walmart’s initial model.

- Personalized Seasonal Item Recommendation

* Analyze customers’ seasonal purchase behavior for Easter products using Apache Spark and Seaborn library.
* Developed an item reranking algorithm to evaluate the efficiency of boosting repurchased items during Easter, and presented at the meeting with product managers.

**DeGroote School of Business, McMaster University** Jul. 2020 – Jan. 2021

***Research Intern*** Ontario, Canada

* Worked on building an economic model based on price discrimination to study mobile Apps pricing strategy; Responded for statistical model selection and interpretation for 3 projects; formed a manual book for statistical instruction purpose.
* Developed ETL processes for different price version information of more than 25,000 apps; collected data from websites with Python; designed a mobile Apps database and updated periodically collected data into it.

**Lenovo Group** Feb. 2020 – Jul. 2020

***Data Analyst Intern, Lenovo Voice, Software Development Team*** Beijing, China

* Built a pipeline for multi-language testing of NLP based software product and suggested algorithm improvement by analyzing the confusion matrix; the codes had been reused by the team.
* Embedded sentences with pre-trained BERT, and compared several models including Logistic Regression, OneVsRest Random Forest and Neural Network model for multi-classification on unstructured sentences, reaching a recall of 70%

# SKILLS

**Programming** Python (Scikit-Learn, PyTorch, Seaborn), Spark SQL, AWS, GCP, Databricks, NoSQL (MongoDB), Bash, R

**Skills** Machine Learning, Deep Learning, ETL, Data Analysis, Time Series, A/B Testing, Git/GitHub, MS Office