

# Wen Su (she/her/hers)

Email : [ws482@cornell.edu](mailto:ws482@cornell.edu)

Mobile : +1 (607)327-7407

Linkedin: <https://www.linkedin.com/in/wennnsu9>

## EDUCATION

### Cornell University

Ithaca, NY

*Master of Engineering in Operations Research and Information Engineering; GPA: 4.15/4.3*

*Expected Dec 2024*

Courses: Introduction to Machine Learning, Deep Learning, Optimization I, Case Study.

### Tongji University

Shanghai, China

*Bachelor of Engineering in Data Science and Big Data Technology; GPA: 4.66 (91.6/100)*

*Sep 2019 - July 2023*

Courses: High-Level Language Programming(C++), Cloud Computing and Virtualization Technology, Data Mining.

## CORE TECHNICAL SKILLS

- **Languages:** Python, SQL, C++, Scala
- **Software& Platform:** MySQL, PostgreSQL, Spark, Hive, Hadoop, PowerBI

## EXPERIENCE

- **Tesla(Shanghai) Co., Ltd** Shanghai, China  
*Data Science Intern* Dec 2022 - Apr 2023
  - Built ETL pipeline using **SQLAlchemy** to streamline data processing for 8 million+ customer text feedback from Hive to MySQL and deployed a classification model for detecting customer demands based on Natural Language Processing algorithms **textCNN** and **Bert**, AUC of which reached 0.82
  - Automated the monthly model training and upgrading process through **Python** and **Airflow** and displayed the model results by building a product search engine through **Flask** and **ElasticSearch**, supporting R&D teams to identify customer demands and enhance product localization
  - Visualized the accuracy of filling feedback templates by service staff through leveraging business metrics in **PowerBI**, which saved 8 hrs manual time per week and transformed template context into SQL tables, designing dashboards and helping service team to faster figure out product deficiencies
- **Bytedance Co., Ltd** Shanghai, China  
*Big Data Engineering Intern* Jun 2022 - Nov 2022
  - Designed data warehouse models to store billion-level data in **Hive** from TikTok moderation team and optimized ETL data pipelines to guarantee stable data workflow in the stage of business expansion
  - Appraised **Spark Web** and **log** to figure out blocking points and reallocated computing resources, reducing running time of daily TB-level data pipelines to average 30 minutes
  - Collaborated with workflow management group to automate multi-platform data extraction with **HiveSQL** queries and design verifier management ETL pipelines and data warehouse models, saving 10 hrs/week for the operation team
- **iCKey Internet Technology Co., Ltd** Shanghai, China  
*Data Analysis Intern* Aug 2021
  - Extracted data from the data warehouse with **SQL** and conducted **RFM analysis** on the customer purchasing behaviors
  - Cleaned raw data from company website, conducted a comparative analysis to get the relationship between product search and product price fluctuation, and streamlined website visiting reports for operation reference weekly
  - Established a **logistic regression model** to detect abnormal user login and filtered out anomaly IP address, preventing malicious data crawling

## ACADEMIC PROJECTS

- **Long-term Traffic Flow Forecasting with Graph Neural Network** Mar 2021 - Apr 2022
  - Applied geometrics algebra, which improved feature extraction, to enhance neural network convolution, resulting in a more accurate prediction with historical data inputs
  - Put forward GAConvLSTM-GAT Model, which turned out 10% more accurate than the baseline models GAT and GAConvLSTM under the condition of whole-day speed prediction
  - **Publication:** [Chenglin Miao, Wen Su, Yanqing Fu, Xihao Chen, Di Zang \(2022\). Long-Term Traffic Speed Prediction on Geometric Algebra ConvLSTM and Graph Attention Model. Paper Presented at 2022 IEEE International Conference on Smart Internet of Things \(SmartIoT 2022\): 108-115](#)
- **Prediction for Loan Default based on Machine Learning** Sep 2021 - Nov 2021
  - Used Exploratory Data Analysis and designed new features to utilize domain knowledge and boost model performance
  - Constructed a tow-phase model based on Xgboost and GBDT; enlarged the scale of training data by figuring the threshold in the first phase and predicted loan default in the second phase
  - Generated 0.896 in AUC of the model and ranked 131 out of 4113 (top 5%) in 2021 China Computer Federation Big Data & Computing Intelligence Contest

## LEAERSHIP AND ACTIVITIES

- **Student Clubs Management Center in Tongji University, Jiading Campus** Shanghai, China  
*Vice President* Aug 2021 - Aug 2022
  - Taken in charge of a team of 50 people in event planning department and the media center for planning and outreaching campus club activites
  - Operated WeChat official account by designing original posts, increasing followers by 50% with average reading of 2000