

Junyue Wang

Email: starjunyue@gmail.com | Tel: (607)-216-7151 | [linkedin.com/in/Junyue-Wang](https://www.linkedin.com/in/Junyue-Wang)

SUMMARY

Highly motivated Master student looking for Data Science related job opportunities. Proficiency in MySQL, Machine Learning; Certified AWS cloud Practitioner, Tableau Specialist. Sufficient quantitative knowledge in **Statistics, Machine Learning, and Mathematical Modeling**.

EDUCATION

Cornell University, College of Computing and Information Science	Ithaca, NY
Master of Science, Applied Statistics GPA:3.93	January 2022 – December 2022
University of California Santa Barbara, College of Letters & Science	Santa Barbara, CA
Bachelor of Science, Statistical Science	September 2016 – July 2020

TECHNICAL SKILLS

- **Statistics:** Regression Analysis, Bayesian Inference, Stochastic Process, Genetics, Time Series, Survival Analysis, Quality Control, Probability Theory, A/B Testing, Experiment Design
- **Programming:** Python, SQL, R, SAS, C++, Javascript
- **Technology:** MySQL, PostgreSQL, pySpark, Tableau, SAS, PyTorch, Hadoop, Tensorflow, Keras, AWS, Github,
- **Machine Learning:** XGBoost, LSTM, PCA, LDA, Data Mining, Natural Language Processing

EXPERIENCE

Gravity Investment, Research Assistant	January 2022 – May 2022
<ul style="list-style-type: none">• Build automation pipeline to fetch daily financial news with asynchronous AJAX-loading using python scrapper and web design. Store in JSON format for next-level analysis.• Data Cleaning with regular expression to drop uninformed text, transform and synchronize news' date to pair up with stock market price in quarter time range.• Used tokenization and stemming with Natural Language Toolkit (NLTK) in Python to preprocess raw news text.• Extracted features and transformed them into numerical data by using Term Frequency - Inverse Document Frequency (TF-IDF) and N-gram to discover financial buzzwords for a holistic view of text data.• Calculate sentiment score with intensity analyzer using updated customized financial dictionary <i>NTUSD_fin</i>. Apply XGBoost, Lasso machine learning model to segmented news data together with stock market analysis on daily return rate of 110 stock code (RMSE 0.76)	
HealAI, Data Analyst	January 2021 – December 2021
<ul style="list-style-type: none">• Data extraction and database designation using SQL server. Compare ADE and MDCC with network to identify closeness of two scales and key factors with strength plot. Participate in Non-relational Database design with No-SQL.• Use Machine Learning Lasso to investigate BPSD and Autism in Toddlers with Elder's from Wuhan about Coronavirus psychological health questionnaire. Establish network among psychometrics and verify counseling result <i>bootnet</i>.• Apply recurrent neural network on massive EEG data to achieve classification on patients with accuracy of 0.83.• Boost 100K sales by intelligent business tool (Tableau/Excel) for insights toward local teenager's psychiatry market.• Create monthly report to customers in understanding the progression of cognitive screening among employees.	

SELECTED PROJECTS

Elders' Cognitive Function Analysis
<ul style="list-style-type: none">• Extracted cohort data of patients' records from authorized database. Stratified and grouped features referred to assembled study population methods.• Preprocessed the raw data set by data cleaning, transforming categorical features, and normalization. Missing values were imputed using <i>Missforest</i> – recursive Random Forest approach to ensure low variance.• Performed recursive feature elimination and Lasso regularization to pick most functional features. This step is intended for future application of simplified diagnosis tool.• Trained supervised learning models such as Logistic Regression, XGBoost, Neural Network with 5-fold Cross Validation and Regularization applied to find optimal parameters and prevent overfitting.• Built ensemble model out of base models for a comprehensive view of model results, top layered with logitBoost.• Evaluated models performance (Accuracy: 0.86, AUC: 0.85) and explored model calibration / Brier score to find out inadequate base model for future improvement.
Twitter Big Data Analytics with PySpark
<ul style="list-style-type: none">• Clustered unlabeled customer reviews into different groups and explored their latent semantic topics by using machine learning models in Python programming. Data cleaning and transformation with regular expression / NLTK.• Tokenized and mapped a sequence of terms to their term frequencies using the hashing trick for efficiency. Then calculate inverse document frequency to transform features into numerical data.• Multiclass Classification using pyspark machine learning Logistic Regression, Naives Bayes and Random Forest. Use GridSearchCV for optimal model parameters and returned best accuracy of 0.84.• Identified latent topics of each customer review by training unsupervised learning models of 5-cluster K-Means Clustering and Latent Dirichlet Allocation (LDA).
Stock Prices Prediction based on Deep Learning
<ul style="list-style-type: none">• Performed a time series prediction to predict future stock price using a Recurrent Neural Network (RNN) regressor.• Cut the time series into sequences and treat the time series prediction problem as a regression problem to apply RNN.• Deployed RNN model using the network architecture of Long Short Term Memory (LSTM)• Trained LSTM model using Adam Optimizer and MSE Loss function via PyTorch on GPU.• Deployed the built model to predict the variation of future stock price.