Jiajun Song

jiajun.song@duke.edu | jiajunsong629.github.io | github.com/jiajunsong629

EDUCATION

Duke University Graduating May 2021
M.Sc. Statistics **GPA: 4.0/4.0**
Relevant Courses: Predictive Modeling (A) | Data Analysis at Scale in Cloud (A) | Bayesian Statistics (A) | Applied Natural Language Processing (A)

Peking University September 2014 - June 2018
B.Sc. Mathematics **GPA: 3.7/4.0**
Relevant Courses: Applied Time Series Analysis (A) | Applied Stochastic Process (A) | Statistical Inference (A) | Introduction to Deep Learning (A)
Clubs: Research Assistant @ Machine Learning Seminar, Competition Manager @ Badminton Association of Peking University

Udacity May 2020
Data Engineering Nanodegree
Relevant Courses: Data Lakes with Spark | Data Modeling with PostgreSQL and Cassandra | Data Pipeline with Airflow

PROFESSIONAL EXPERIENCE

Duke University | Durham, NC January 2020 - Present
Research Statistician under the guidance of Lawrence David

- Developing an Automatic Smart Toilet for tracking patient fluid output data and its application in hospitals and **IoT Analytics**
- Assisting with inferences of user events based on processing **High-Frequency data** of 11 sensors and more than 100 user events, including load scale, infrared camera, and microphone, as well as providing analyses that guide sensor inclusion and design
- Researching on the breakpoint detection of time series w/ **VGG deep neural network** using **TensorFlow** on the spectrogram of the audio dataset; Precision rate and Recall rate of Human Activity Recognition has been improved to above 85%

Iqunxing | Shanghai December 2018 - May 2019
Data Analyst Intern

- Collaborated with Marketing Team on credit rating and fraud detection based on seasonal E-commerce transaction record
- Individually implemented **ETL pipeline** using **Amazon Redshift** and **Airflow** on more than 1M transaction records of 10k SKUs over 18 months
- Presented weekly data analytics reports w/ **Tableau** to colleagues with limited statistics knowledge, topics including the detection of sales patterns with Zero-Inflated Poisson model, large scale simulation with Items Clustering and Association Rule

Course Mathematical Stats at Duke University | Durham, NC August 2019 – January 2020
Teaching Assistant

- Facilitated discussion on lectures and problems sets on **Hypotheses Testing** and **Large Random Samples Property** for over 80 students
- Managed course content weekly through Sakai, Piazza and Slack together with one-on-one tutoring and regular out of class assistance

TECHNICAL PROJECTS

Driving Forces Behind Art Valuation | In-Class Competition November 2019 – January 2020

- Analyzed the 18th-century auction pricing data w/ model stacking based on **Random Forest**, **Bayesian Linear Regression**, and **Gradient Boosting**
- Engineered customized features by text mining on the auction information which defined latent factors that account for the driving forces, i.e. famous, significantly decreasing the RMSE by 15%
- Won the 1st place among 11 teams in multiple model evaluations including RMSE and Coverage Rate

Assessing the Funniness of Edited News Headlines | CodaLab Competition January 2020 – May 2020

- Applied multiple models such as **Conv+LSTM** as our benchmark scored 0.626, **Transformer** models on the pre-trained **BERT** embeddings scored 0.552, top one scored 0.512 and median scored 0.562 on the leaderboard for RMSE.
- Extracted **Contextualized Word Embeddings** by **Fine-Tuning** on humor headlines datasets and/or researched on linguistic elements of humor based on Error Analysis to improve previous results

OCR Application with Serverless Data Engineering Pipeline | Individual Project April 2020

- Built an Automatic OCR application on **AWS Lambda** with Rekognition APIs to detect text in images from S3 Bucket and store detected labels into **DynamoDB** triggered by **Cloud Watch**
- Set up the Docker environment and infrastructure using **Cloud Formation** and **AWS SAM CLI**

Stochastic Gradient Hamiltonian Monte Carlo Python Package Development | Team Project March 2020 – May 2020

- Reproduced experiments of SGHMC sampler, implemented and optimized the algorithm into Python Package with **Cython** and **C++**

Spotify Music Analytics Dashboard | Team Project December 2019

- Actively communicating and strategized project with 3 teammates on collecting data with Spotify **API** and designing dashboard using **R Shiny**
- Implemented multiple functionalities including lyrics **Sentimental Analysis**, **Recommender System** and unique gadgets for Searching

SKILLS & INTERESTS

Languages/ Frameworks: Python, R, JavaScript, C++, TensorFlow, PyTorch, Spark MLlib
Big Data/ Cloud Experience: Amazon Redshift, Amazon EMR, Amazon SageMaker, PostgreSQL, Cassandra, Airflow
Presentation: Tableau, R Shiny, R Markdown, LaTeX, Notion, Microsoft Suites
DevOps: Docker, Git, Linux, Bash, Vim, CI/CD, IaC
Interests: Badminton, Movie, Cooking, Computer Hardware