YUANPEI CAO

215-720-6285 | https://yuanpeicao.github.io/ | caoyuanpei@gmail.com

WORKING EXPERIENCE

Airbnb - AI Lab, Data Scientist and Machine Learning Engineer Apr. 2019 - Present

- Image Quality Model: Led end-to-end computer vision pipeline for empowering Airbnb search rank and marketing.
  - Increased Airbnb bookings by incorporating the image quality model into Airbnb's search recommender system.
  - Increased Airbnb ads CTR by optimizing displayed ads image quality score on Toutiao and Tencent.
- Designed CDN-based online experiment architecture to avoid experiment imbalance caused by latency differences between treatment groups.

Airbnb - Payments, Data Scientist Apr. 2018 - Apr. 2019

- Payment Routing: Led end-to-end project for optimizing payment routes between payment processors.
  - Developed tree segmentation algorithm to find optimal route segments without violating contractual obligations.

AdColony (Opera Software) - Data Scientist Jan. 2017 - Mar. 2018

- Ad Install Rate (IR) Prediction: Led end-to-end machine learning project for IR prediction on new ads (cold start).
  - Developed ensemble learning framework to find target users.

EDUCATION

University of Pennsylvania Sep. 2011 - Aug. 2016
Ph.D. in Applied Mathematics and Computational Science

The Wharton School, University of Pennsylvania Sep. 2011 - May 2016
M.A. in Statistics

Fudan University Sep. 2007 - Jun. 2011
B.S. in Applied Mathematics (with Honors)

SELECTED PUBLICATIONS

- **Cao, Y.**, Zhang, A, and Li, H. (2020): Multi-sample Estimation of Bacterial Composition Matrix in Metagenomics Data, Biometrika, 107, 75-92.
- **Cao, Y.**, Lin, W. and Li, H. (2019): Large Covariance Estimation for Compositional Data via Composition Adjusted Thresholding, Journal of the American Statistical Association (JASA), 114:526, 759-772.
- **Cao, Y.**, Lin, W. and Li, H. (2018): Two-sample Mean Tests for High Dimensional Compositional Data, Biometrika, 105(1): 115-132.

PATENTS

Tree Structure-Based Smart Inter-Computing Routing Model, Ref 26887-41473

KAGGLE COMPETITIONS

New York City Taxi Fare Prediction Sept. 2018
Hosted by Google Cloud and Coursera Final Leaderboard Rank: 9/1488 (Top 1%)

Google Analytics Customer Revenue Prediction Feb. 2019
Hosted by Google and R Studio Final Leaderboard Rank: 11/1089 (Top 1%, Gold Medal)

KNOWLEDGE AND SKILLS

Python (Scipy, Pandas, Scikit-learn, XGBoost, Tensorflow), R, Java, Shell, Matlab
AWS (ec2, S3, EBS, Redshift, Sagemaker), Google Cloud, MySQL, Presto, Hive, Airflow, Hadoop, Spark