

YUANPEI CAO

215-720-6285 | yuanpei.cao@airbnb.com

WORKING EXPERIENCE

Airbnb - Data Scientist

Apr. 2018 -

- Smart Routing: lead end-to-end statistics project for optimizing payment routes between multiple processing providers
 - Developed heuristic tree segmentation algorithm (using statistical inference and offline reply estimate) to find the opportunity route segments.
 - Optimized the first attempt acceptance rate without violating contractual obligations from different processor companies (Chase, WorldPay, etc).
 - Increased first attempt acceptance rate by x% compared to current system according to offline reply estimate, which can be translated to xx millions dollars annual revenue increase.
 - Filed US Patent (Ref 26887-41473): Tree Structure-Based Smart Inter-Computing Routing Model.
- Provided analytics support (anomaly detection / new payment service monitoring / data pipeline & dashboard) for multiple high-visibility strategic partnerships, resulting in millions dollars in bottom-line savings.
- Designed and analyzed A/B tests for Guest & Host Commerce (Pay Less Upfront, Alipay Payout, Quickpay, etc).

AdColony (Opera Software) - Data Scientist

Jan. 2017 - Mar. 2018

Ad Install Rate Prediction: lead end-to-end machine learning project for IR prediction on new ads (cold start)

- Handled with raw data cleaning and feature engineering from more than 100TB historical data sources, including the query, the text and video of the ad creative, and various ad-related metadata.
- Developed the ensemble learning framework (k-NN, gradient boosting regression tree, a variety of content-based similarity models using deep learning and NLP techniques, etc) to find the target users.
- Automated pipeline using disparate tools/sources like Shell, Python, BigQuery, Redshift, s3, ec2, Cron.
- Increased install rate and eCPM by more than 300% compared to previous models.
- Presented the work to CTO and other technical leaders; Also presented to non-technical audience from business team.

Department of Biostatistics, University of Pennsylvania - Postdoc

Sep. 2016 - Jan. 2017

- Developed statistical models (hypothesis testing, multivariate analysis, network estimation, missing value recovery) for high-dimensional compositional data with applications to human microbiome study.
- Handled with raw data cleaning from cross-section study and clinical trials, and implemented the statistical algorithm and data visualization techniques by R/MATLAB/Python/Cytoscape.
- Collaborated with experts from Children Hospital of Philadelphia (CHOP) to explore the relationships between microbiome and human health.

KAGGLE COMPETITION

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

Languages: Python (Scipy, Pandas, Scikit-learn, XGBoost, Tensorflow, etc), R, Java, Shell, SQL, Matlab
Cloud: AWS (ec2, S3, EBS, Redshift), Google Cloud (BigQuery), MySQL, Presto, Hive, Airflow, Aerospike
Miscellaneous: Tableau, Linux, Cron, scripting automation, git, tmux, LaTeX, experienced in Hadoop, Apache Spark
Quantitative Analysis: Ph.D. level knowledge in statistics, working experience in machine learning and A/B testing

EDUCATION

University of Pennsylvania	Sep. 2011 - Aug. 2016
Ph.D. in Applied Mathematics and Computational Science	Overall GPA: 3.96/4.00
The Wharton School, University of Pennsylvania	Sep. 2011 - May 2016
M.A. in Statistics	Overall GPA: 4.00/4.00
Fudan University, Shanghai, China	Sep. 2007 - Jun. 2011
B.S. in Applied Mathematics (with honors)	Overall GPA: 3.72/4.00 (Top 5%)

PREPRINTS AND WORKING PAPERS

- **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)*.
- **Cao, Y.**, Zhang, A. and Li, H. (2019): *Multi-sample Estimation of Bacterial Composition Matrix in Metagenomics Data*, *Biometrika*, accepted, arXiv:1706.02380.
- Feng, X., Wang, S., Gao, S., **Cao, Y.**, and Murray, A.T. (2019): *MOTO: A Multi-Objective Trajectory Optimization Method for Finding Sequential Activity Locations for Multiple Moving Objects along Road Networks*, submitted to *Environment and Planning B: Urban Analytics and City Science*, acceptable with minor revision.
- **Cao, Y.**, Lin, W. and Li, H. (2018): *Two-sample Mean Tests for High Dimensional Compositional Data*, *Biometrika*, 105(1): 115-132.
- Ma, R., **Cao, Y.**, and Li, H.: *Sparse High-Dimensional Precision Matrix Estimation for Compositional Data*, submitted.
- Jie, C., **Cao, Y.**: *GARCH Modeling and Extreme Value Theory-based Fund Risk Measurement and Performance Evaluation*, *Modern Business* (2011) (in Chinese).

RESEARCH COMMUNITY SERVICES

- Australian & New Zealand Journal of Statistics, Reviewer
- Microbiome, Reviewer

CERTIFICATION

Deep Learning Nanodegree on Udacity	Oct. 2018
Convolutional Neural Networks by deeplearning.ai on Coursera	Jan. 2018
Structuring Machine Learning Projects by deeplearning.ai on Coursera	Jan. 2018
Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization by deeplearning.ai on Coursera	Jan. 2018
Neural Networks and Deep Learning by deeplearning.ai on Coursera	Jan. 2018
Shell Scripting: Discover How to Automate Command Line Tasks on Udemy	May 2017
Hadoop Platform and Application Framework on Coursera	Sep. 2016
Introduction to Apache Spark on EdX	Aug. 2016