Helen (Mengxin) Ji

mengxin.ji1992@gmail.com • +510-502-3296 • https://mengxinji.github.io/

Summary

I am a third year Ph.D. student majored in Resource Economics and second year master student in Statistics. My research interests focus on machine learning method and its application to Economics.

Programming: R, Python (e.g. numpy, scipy, pySpark, scikit-learn, Pandas), Stata.

Applied Machine Learning: Supervised/unsupervised Learning

Statistics and Econometrics: Generalized Linear Model, Causal Inference.

Education

University of California, Davis

PhD in Resource Economics *Jun. 2018(expected)*
GPA: 3.90/4.0

Master of Science in Statistic *Jun. 2017(expected)*
GPA: 4.0/4.0
Related courses: Probability and Statistic, Sampling, Statistical Learning, Machine Learning, Data Mining.

Projects

Bay Area Rapid Transit (BART) Throughput Pattern Analysis *Jan 2016 - present*

- Visualize the throughput of the BART among 45 stations, from 2011 to 2015, with more than 40 million ridership record. The motion chart could be found on my **github**.
- Investigate the key factors for the daily total throughput with generalized linear model. Study a wide range of factors, like weather information (temperature, humidity) and demographic data (crime rate, income). The top 3 statistically significant factors are: visibility, humidity, and wind speed.
- Applied random forest to assess the variable importance of each factor, which helps capture the non-linear impact from each factor. The most important factors are consistent with the linear model.

Policy Causal Impact on California Soda Market Consumption Pattern *Jul 2016 - present*

- Inference and mining the causality between the policy and the soft drink market consumption and policy impact on teenager obesity rate with state-mandated banning high school student consume soft drink policy;
- Used datatable, dplyr, and ggplot2 package in R for exploratory analysis visualization, automatically characterization of California soft drink market by product descriptions. The data size is 32GB;
- Apply causal inference method (triple difference approach) based on the all California supermarkets household level consumptions with transaction details. Results show the impact of this policy on the compensation for soda market consumption outside school is not significant.

Energy Prediction of Peaking Power Plant in California *Sep 2016 - Oct 2016*

- Aimed to accurately predict the demand of power, and thus reduce the cost of building unnecessary power plants; data consisted about 10,000 observations, with outcome variable XX, and 6 environmental attributes (e.g. AA, BB).
- Applied and investigated multiple statistical and machine learning algorithms (e.g. OLS, ridge regression, Lasso, decision tree, random forest model) in pySpark for the prediction of power usage;
- Used cross-validation for model selection, and achieved 78% prediction accuracy on the testing set with Random Froest model;

California Burbank Water Study *Jul-Sep 2015*
Assisted in data cleaning and data analysis for California Burbank Water project; mainly engaged in cleaning and analyzing large data set;
Used STATA to analyze the dataset of over 10,000 households which records every hour per day during three months.