FANFEI (FAUSTINE) LI

Email: faustineli12@gmail.com **Cell:** 678-704-6395 **Website:** faustineli.github.io

Profile

A Master's student in the Duke Statistical Science program with a passion for solving problems, especially in the fields of energy, environment, health, and technology. Current interests include machine learning and Bayesian methods. Seeking a summer internship in data science or analytics.

Education

Duke University, Durham NC **Expected May 2018**

MS Statistical Science

California Institute of Technology, Pasadena CA **June 2015**

BS Chemical Engineering

Work Experience

Research Fellow, Oak Ridge National Laboratory **2015 – 2016**

- Cleaned, analyzed, and visualized data collected on particulate matter from engine emissions.
- Wrote MATLAB code to perform outlier detection and statistical inference.
- Automated data cleaning steps including time-alignment, filtering, and error checking.
- Segmented SEM images of particulate aggregates using thresholding and edge-detection.
- Produced publication quality plots and wrote a set of experimental guidelines.
- Research presented at Health Effects Institute Symposium and abstract accepted to SAE.

Undergraduate Researcher, Caltech **Summer 2014**

- Simulated the kinetics of organic species in photochemical smog using MATLAB.
- Developed a set of rate equations to describe the mechanism of glyoxal production.
- Discovered a connection between reactive oxygen species and the production of acids.

Projects

Duke Kaggle Competition **November 2016**

- First place in Kaggle competition, predicting car insurance claim severity.
- Tuned parameters of gradient boosted trees to achieve the lowest mean absolute error.
- Used feature engineering, ensembling, and custom objective functions to improve performance.
- Set up a reproducible data cleaning, model training and validation procedure.

Text Analysis of Job Descriptions **December 2016**

- Worked with a group to implement an interface to explore data-related jobs.
- Web-scraped text from Indeed and transformed the corpus using the R package tm.
- Clustered similar jobs based on description using Latent Dirichlet Allocation.
- Created a Shiny interface to interact with job data, including a map and word cloud.

Skills

- Proficient in R and MATLAB. Familiar with Python, Spark. SQL, and HTML.
- Tools include git, LateX, Markdown, and Unix command line utilities.
- Relevant courses include Machine Learning, Bayesian Statistics and Statistical Computing.