FANFEI (FAUSTINE) LI

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

Profile

A Master's student in Statistical Science at Duke University with a passion for solving problems, especially in the fields of energy, environment, health, and technology. Seeking a full-time position in Data Science.

Education

Duke University, Durham NC **Expected May 2018**
MS Statistical Science

California Institute of Technology, Pasadena CA **June 2015**
BS Chemical Engineering

Work Experience

Statistics Intern, Eli Lilly and Company **May 2017 – August 2017**

- Trained deep neural networks to automatically classify severity of disease from medical images.
- Used TensorFlow, Keras and Python to iterate on various convolutional neural network architectures.
- Wrote scripts to automate the processing of over 80,000 images and to test and validate models.
- Created a web dashboard that allows users to interactively receive predictions from images.
- Produced a deep learning tutorial that includes experiences and practical guidelines.

Research Intern, Oak Ridge National Laboratory **June 2015 – July 2016**

- Cleaned, analyzed, and visualized data collected on particulate matter from engine emissions.
- Independently wrote MATLAB code to test for outliers and perform statistical inference.
- Segmented SEM images of particulate aggregates using thresholding and edge-detection.
- Engineered user interfaces to align time series data and visualize particulate images.

Projects

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.com and transformed the corpus using the R package tm.
- Clustered similar jobs based on descriptions using Latent Dirichlet Allocation.
- Created a Shiny interface to interact with job data, including a map and word cloud.

Duke Kaggle Competition **November 2016**

- Placed first out of 34 in an in-class 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 using R.

Skills

- Proficient in R and Python. Familiar with TensorFlow, Keras, SQL, and Spark.
- R data science tools used include ggplot, dplyr, tidyr, rvest, Shiny, and caret.
- Python data science tools used include pandas, numpy, matplotlib, plotly, keras, and scikit-learn.
- Other software used include Git / Github, LateX, Markdown, Jupyter Notebook, and Unix utilities.