FANFEI (FAUSTINE) LI

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

Education

Duke University, Durham NC   **Aug 2016 - May 2018**
  MS Statistical Science

California Institute of Technology, Pasadena CA   **Sep 2011 - Jun 2015**
  BS Chemical Engineering

Work Experience

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

- Trained convolutional neural networks to classify disease severity in medical images.
- Used Python, TensorFlow, and Keras to process over 80,000 images and build models.
- Diagnosed and resolved model performance issues such as overfitting and unbalanced classes.
- Created a web dashboard using Plotly Dash that allows users to interactively receive predictions.
- Produced a written tutorial on training neural networks which incorporates experiences learned.

Research Intern, Oak Ridge National Laboratory   **Jun 2015 – Jul 2016**

- Cleaned and visualized engine emission data to quantify pollution impacts of biofuels.
- Performed statistical hypothesis tests such as ANOVA and identified potential outliers.
- Processed hundreds of SEM images including noise filtering and background segmentation.
- Presented statistical results to peers and wrote data analysis methods sections for publication.

Projects

Bayesian Hidden Markov Models

- Implemented inference for Bayesian HMMs from scratch with Python, numpy, and scipy.
- Developed a Python package, wrote documentation, and provided examples in Jupyter Notebook.
- Automated testing with Github and CircleCI build tools.

Text Analysis of Job Description

- Web-scraped job descriptions from Indeed.com and transformed the corpus using R.
- Clustered similar jobs based on descriptions using Latent Dirichlet Allocation.
- Developed a Shiny interface to interact with job data which includes a map and word cloud.

Duke Kaggle Competition

- Placed first in an in-class Kaggle competition, predicting the size of insurance claims.
- Wrote scripts to reproducibly train and test gradient boosted models using xgboost.
- Used feature engineering, ensembling, and custom objective functions to improve performance.

Skills

- Proficient in Python and R. Currently learning C/C++. Some experience with SQL and Spark.
- Experienced with Python and R data science stack including scikit-learn, pandas, dplyr, and ggplot.
- Other software skills include git, Github, Tableau, LateX, Linux, and bash shell.
- Coursework includes Bayesian statistics, causal inference, and categorical data analysis.