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 severity of disease on a medical image dataset.
- Used Python, TensorFlow, and Keras to process over 80,000 images, and train and test 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 tutorial on training deep neural networks which incorporates experiences learned.

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

- Cleaned, visualized, and analyzed engine emission data to quantify pollution impacts of biofuels.
- Performed statistical hypothesis tests such as ANOVA and identified outliers.
- Processed SEM images of particulate matter including noise filtering and background segmentation.
- Presented relevant data analysis to peers and wrote data analysis methods sections for publication.
- Developed graphical tools in MATLAB to facilitate data cleaning and manipulation.

Projects

Bayesian HMM

- Implemented fitting Bayesian Hidden Markov Models from scratch in Python, numpy, and scipy.
- Produced a Python package, wrote documentation, and provided examples with Jupyter Notebook.

Text Analysis of Job Descriptions

- Web-scraped job descriptions from Indeed.com and transformed the corpus using R.
- 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

- Placed first in an in-class Kaggle competition, using xgboost to predict size of insurance claims.
- Used feature engineering, ensembling, and custom objective functions to improve performance.
- Wrote reproducible training scripts in R to transform data and iterate on models.

Skills

- Proficient in Python and R. Currently learning C/C++. Some experience with SQL and Spark.
- Experienced with Python data science stack including scikit-learn, matplotlib, and pandas.
- Experienced with R data science stack including RStudio, R markdown, dplyr, ggplot, and caret.
- Other software skills include git / Github, LaTeX, Linux, and bash shell.