

Colin Cui

Princeton, New Jersey • (415)518-3959 • colstat@gmail.com

Skills

Languages: R, Python, Matlab, Stata, C, C++, Java, SQL, AWS

Data Science: Machine Learning, AI, Statistical Modeling, Optimization, Bayesian Inference.

Packages: NumPy, SciPy, Pandas, Scikit-learn, dplyr, glmnet, ggplot2, requests

Paper

Colin Cui. On Statistical Learning Theory, Oracle inequality, and the Lasso. [link] January, 2021

Projects

Predict credit card default with XGBoost with respect to other models Princeton, NJ

- Performed exploratory data analysis and feature engineering on imbalanced data
- Implemented classification and hyperparameter tuning on class scale weights.
- Evaluated XGBoost performance score with respect to logistic regression, decision tree, random forest, k-nearest neighbor, and naïve bayes, tested different precision and recall

Sentiment analysis using NLP on customer reviews from the web Princeton, NJ

- Scraped and parsed 1GB of online review keywords from web pages of multiple sources
- Analyzed text keywords with Word2vec representation in NLTK and gensim
- Identified sentiments and classified keywords with k-means, NMF, LDA in scikit-learn

Classification of frontal face images using convolutional neural network Princeton, NJ

- Implemented CNN variations in Keras on frontal face images
- without maxpooling, 20% dropout, without dropout and maxpooling, standard out-of-the-box
- Trained and plotted different CNN models to improve final validation accuracy

Experience

Machine Learning Engineer, Ideally, Feb. 2018 - Present Princeton, NJ

- Collected healthcare data from three public sources in feature formats. Merged data in R pipeline to create markdown summary statistics report for insight analysis
- Built logistic regression, decision tree, and random forest to classify mortality after emergency visits, interpreted machine learning model coefficients and feature importance
- Migrated R data pipeline and model to Python and improved running time by 15%
- Optimized model and identified feature importance to understand the relationship between mortality and explanatory variables, outperformed R^2 by 20%

Research Assistant, Rutgers University, July 2014 – Jan 2018 Piscataway, NJ

- Developed Bayesian model to predict the uncertainty of cancer tissue in biomedical imaging
- Built a Gaussian Process (GP) meta-model to estimate tissue viscosity uncertainty
- Reduced variance of uncertainty even after added noise, comparing to finite element model

Education

Rutgers University

MS., Statistics

University of California, Davis

B.S. in Statistics