XIAOYU SUN

10 River Rd, New York, NY, 10044
xysun@bu.edu 857-400-4158

seeking data scientist position available on February 1st

EDUCATION

gre score: 324/340

Boston University, Boston, MA M.A. in Statistics Sep 2014 – now
Related Coursework: Linear Model, Generalized Linear Model, Biological DataBase, Machine Learning, Time Series, Statistical Software(SAS), Probability Theory, Estimation Theory, Hypothesis Testing
GPA in Major: 3.8/4.0

Peking University, Beijing, China B.S. in Mathematical Science Aug 2010 – Jul 2014
Related Coursework: Mathematical Analysis, Advanced Algebra, Data Mining, Computational Statistics, Actuarial Science, Mathematical Modeling, Data Structure

SAS Base/Advanced Certification Jan 2015

INTERNSHIP EXPERIENCE

BioPIER Clinical Workbench, Boston Jul–Aug 2015

- Intern as a sas programmer for an extension team of the company which specializes in providing biostatistics and statistical programming services to the pharmaceutical and biotechnology industry.
- Produced CRF including summary tables, data listings using SAS for a local pharmaceutical company.
- Independently validated SDTM datasets and summary tables from other programmers.
- Some of the micros I wrote for common tables have been generally used for the whole team to save time of qc work.

The USDA Human Nutrition Research Center on Aging (HNRCA), Tufts University Jan-May 2015

- Built a database and a website open for lab members that incorporates several public resources and can be queried to produce gene-centric reports about the genetic basis of the response to diet and human cardiometabolic phenotypes from their recent research.
- Designed ER diagram and managed the front end part including webpage design, data visualization, result file download and user-friendly image map.
- Using front end technologies: html, javascript, bootstrap, Ajax, google charts;
  using back end technologies: python, cgi; Database: MySQL under Unix system.
- It's now put into use in HNRCA assisting researchers and being updated regularly with the advances in research. It will be open to public along with the results of human nutrition research.

ACADEMIC AND EXTRACURRICULAR PROJECTS

Data Mining Project in Department of Math and Statistics, Peking University

- Collected movie data of size about 10 thousand mb with twitter crawler using python.
- Conducted Sentiment Analysis to the terms from raw data using python, compared to sentiments dictionary and calculated the score of peoples' opinions about them.
- Set up a website presenting the results of the top 5 popular movies and most popular descriptive words about them.

Biostatistical Big Data Project in Center for Statistical Science, Peking University

- Comparison of approaches for detecting tumor driver genes targeted by copy number variation.
- Cleaned data from 1901 breast cancer samples measured through different platforms such as Affymetrix Genome-Wide Human SNP 6.0 platform of 1.8 million gene markers for each sample.
- Simulated data using Gaussian model for the same gene makers with real data for 100 samples.
- Applied models on simulated and real data with total size of 2000 million bytes and compare with known cancer genes census and analysis using R and SAS.
- Compared the accuracy and FDR control of 5 models and evaluated the feature of their performance.

Machine Learning Project, Kaggle Competition — Walmart Trip Type Classification Nov 2015

- Manipulated train and test datasets of 2 GB using numpy and pandas in python.
- Conducted machine learning classification models like random forests and XGBoost and tuned models manually to robust model with better classification results.
- Applying stacking with logistic regression method to ensemble models.
- Current results have beaten the traditional random forests benchmark.

Programming Experience

Database: MySQL Framework: CGI, client-server, Hadoop, PIG, Hive
Programming Languages: R, SAS, Python(Numpy, Pandas, scikit-learn), Java, Javascript, C