Sijia Yue

100 Haven Ave, Apt 5H, New York, NY, 10032 | 1-(646)-226-2982 | sy2824@columbia.edu

LinkedIn: [https://www.linkedin.com/in/sijia-yue](https://www.linkedin.com/in/sijia-yue-4100b7158/) | GitHub: <https://github.com/ysjbluemoon>

**EDUCATION**

**Columbia University** **Mailman School of Public Health** | New York, NY Sep 2018 – May 2020

***Master of Science in Biostatistics***

Relevant coursework: Data Science, Statistical Learning and Data Mining, Biostatistical Methods I & II, Statistical Inference, Relational Databases and SQL Programming for Research and Data Science, Machine Learning, Applied Deep Learning

**Beijing University of Posts and Telecommunications** | Beijing, China Sep 2014 – May 2018

***Bachelor of Management of Information System***, joint degree with Queen Mary University of London

Relevant coursework: Database and Data Mining, Data Structure, Java Programming, Information System Management, Enterprise Management, Enterprise Strategy, Supply Chain Management

**SKILLS**

**Business:** Microsoft Office (Excel, PowerPoint, Word, Access, Visio), Axure RP

**Technical:** R (shiny, ggplot2, tidyverse, caret, CRAN), SAS (Macro, Proc SQL), Python (Scikit-Learn, Keras, TensorFlow, pandas, Numpy, Scipy, matplotlib, seaborn), MySQL, Git, HTML, CSS, Java, JavaScript, XML, C, Bash, LaTeX

**EXPERIENCES**

**Columbia University Mailman School of Public Health** | New York, NY May 2019 – Present

***Research Assistant***

***Optimizing Cutoff Points of SCI-CG Score to Improve Complicated Grief Classification***

* Fitted logistic regression and support vector machine with R to model the probability of complicated grief using ICG score
* Implemented algorithms to optimize the cutoff points based on 10 statistical criteria including sensitivity, specificity and Kappa
* Evaluated algorithm performance by visualizing distribution of predicted classes, specificity, sensitivity and ROC, reaching 0.92 AUC
* Run leave-one-out cross validation with R to compare sensitivity, specificity, Kappa and F1 score

***Functional Data Analysis of Acute Kidney Injury after Surgery***

* Use SQL to implement ETL to build data map and workflow on a large dataset stores more than 1,000,000 observations
* Implemented loess smoother, linear spline smoother and moving median method to reduce noise in blood pressure measures
* Visualized trends of mean atrial pressure during the surgery taking critical surgery events into account
* Utilized Principal Components Analysis with Python to define the patterns of change in blood pressure, reaching 0.93 AUC

**Columbia University Mailman School of Public Health** | New York, NY Sep 2019 – Present

***Teaching Assistant: Relational Databases and SQL Programming for Research and Data Science***

* Held office hour and attended in class to answer questions from students
* Graded students’ weekly assignments, midterm and final exams and final projects in MySQL and Access

**Accenture |** Beijing Dec 2017 – May 2018

***IT Consultant Internship***

* Cooperated with four departments to optimize interfaces of Adama company’s SAP system; increased the processing speed by 60 percent
* Analyzed clients’ requirements, designed business rules, wrote software specification documents of online E-commerce platform development for 5 companies and delivered reports to clients
* Assisted to train the 10 new interns as the internship team lead and set up meetings for weekly working process reports

**PROJECTS**

**The Image Classification of Landmarks in Columbia University**| Columbia University Sep 2019 – Oct 2019

* Preprocessed the videos of different landmarks in Columbia University into 30,000 images with ffmpeg in bash
* Fitted a CNN model to classify the images on Google Drive including transfer learning from pre-trained model MobelNetV2 and data augmentation with Python in TensorFlow2 and Keras, reaching 0.98 test accuracy
* Converted the model into JavaScript file in tensorflowjs; posted the model on a html webpage that runs in browser

**The Prediction of Prostate Cancer in Machine Learning Methods** | Columbia University Sep 2019 – Sep 2019

* Fitted Ridge regression, Lasso regression, partial least squares, best subset regression and principle component regression with Python to predict the probability of prostate cancer
* Implemented algorithms to find the best model based on one-standard-error-rule of 10-fold cross validation test error and BIC
* Wrote functions to visualize the relationship between cross validation error and degree of freedom in each model, displaying the best model after implementing one-standard-error-rule
* Generated statistical summary reports including illustrating model selection, interpreting the model coefficients and results

**The Prediction of Movie Rating in Machine Learning Methods |** Columbia University May 2019 – Jun 2019

* Scraped data of 15,000 movie information with Python using BeautifulSoup3 from IMDB website
* Preprocessed the dataset by classified missing values and outliers with Pandas and Numpy; performed exploratory data analysis with Seaborn and Matplotlib to summarize main characteristics of the dataset
* Fitted RandomForest Model with Python in Scikit-learn to predict audience ratings
* Evaluated model performance with k-fold cross validation and reached the highest F1 score at 0.97

**The Comparison of Two Models Predicting of Cardiovascular Disease |** Columbia University May 2019 – Jun 2019

* Implemented data cleaning and missing data imputation with SAS using Macro
* Fitted logistic regression and proportional odds model with ordinal responses to predict the diagnosis in SAS with stepwise model selection
* Compared the goodness-of-fit based on the Hosmer and Lemeshow Goodness-of-Fit test statistics between two models
* Generated statistical summary reports including illustrating model selection, interpreting the model coefficients

**The Analysis of Organ Donor Enrollment In New York State |** Columbia University Nov 2018 – Dec 2018

* Visualized the relationship between demographic elements and donation registration rate with R in ggplot
* Implemented a mix-effect linear model with a random intercept to analyze the donation registration rate affected by the policy change in the New York State
* Utilized ggmap and Shinny in R to plot the map of registration rate throughout the New York State in years

**The Analysis of Efficacy of Weight-Loss Drug WL001A |** Columbia University Nov 2018 – Dec 2018

* Set up an Access database for the RCT examining the efficacy of WL001A including 8 tables and over 500 records, built Access Forms and Access Reports for data entry and summary
* Wrote SQL queries and generated reports that provide summaries of weight change by treatment

**The E-Commerce Website Development for a Souvenir Company** | Beijing, China May 2016 – Aug 2016

* Designed the layout of the website and set up the front-end development of the website in HTML5.0, CSS, JavaScipt and JQuery
* Set up the back-end database that stores the customer data and purchase records in MySQL

**LICENSES**

* **SAS Certified Base Programmer for SAS 9** (Certificate Verification Number: HZRJVMH11EVE1S9F)
* **SAS Certified Advanced Programmer for SAS 9** (Certificate Verification Number: Z8J3EYB2CNFE1MCY)

**PUBLICATIONS**

Xiaomei Zhuang, Tianhong Zhang, Sijia Yue**: Allosteric Activation of Midazolam Cyp3a5 Hydroxylase Activity by Icotinib - Enhancement by Ketoconazole.** *Biochemical Pharmacology* 2016, 121:67-77.