Larry (Jianfeng) Yan
https://github.com/Larryjianfeng
526 West, 123 street. Apt 1W. NYC, N.Y
347-497-1148, jy2641@columbia.edu

EDUCATION:

Columbia University, Graduate School of Arts and Sciences — N.Y.C, New York
M.S in Statistics, **Data Science Focus**, GPA 3.6/4 — Dec, 2015
Core Courses: machine learning, Bayesian, algorithms, generalized linear regression, time series.

Wuhan University (Top 10 in China) — Wuhan, China
B.S. in Mathematics, GPA 3.6/4.0 — Jun, 2014
B.S. in Economics, GPA 3.51/4.0 — Jun, 2014

INTERNSHIP AND WORKING EXPERIENCE:

Columbia University — N.Y.C, New York
Teaching Assistant: grading homework of undergraduates and assisting professor in class. — Sep. 2015 to Dec. 2015

Columbia University — N.Y.C, New York
Research Assistant of Prof. Lucius J. Riccio. — Jul. 2015 to Sep.2015

- Data clean and manipulation. Database management with MySQL. Data scraping using python from website.
- For more than 1G data, conduct feature selection, modeling with non-parametric model and make prediction.
- Present and visualize the result with shiny, an online interactive data visualizing tool with R.

China Life Insurance (#37 in Forbes 500 Global) — Wuhan, China
Data Analyst Intern — Sep, 2013 to Nov, 2013

- Identify meaningful insight in private insurance sales strategies and operations by combining and analyzing data from various sources, including financial institutions and medical institutions.
- Using Poisson regression and classification method to detecting fraud claims.
- Find new way to optimize resources configuration by segmentation analysis of clients. Communicate with sales teams and insurance agents to identify problems. Present the result of data analysis to them with PPT.

Industrial and Commercial Bank of China — Wuhan, China
Operations Analysis Intern — June, 2013 to Sep, 2013

- Participate in the quantitative credit rating of medium-sized and small enterprise based on analyzing balance sheet, cash flow statements, debt structure, inventory turnover rate, management performance.
- Identifying key factors or key metrics leading to bad loans by using logistic regression model to over 10,0000 rows of loans data with SAS.

DATA SCIENCE EXPERIENCE:

Webpages classification. Classification of whether a webpage contains paid advertisement, including 300,000 html files with size over 30 GB. Method used: feature extraction, text processing, NLP model construction, feature selection, dimension reduction, random forest.

Music Recommendation System: using data from Million Song Database, create a LDA model based collaborative filtering recommendation system with python (more specific description in my Github).

Risk classification for insurance company: Ensemble random forest, adaboosting, gradient boosting regression, nueral network to make classification for risk, from level 1 to level 8 for the personal life insurance. (codes here)

Online interactive data analysis: Online R based step-wise regression application. Part of the visualization projects of Professor Zhen from Columbia University. (The application link is here, description is here)

Visualization of golfers' performance: Visualization of comparison of golfers based on historical data. (Project with Prof. Riccio, the link is here)

My Programming language: Python, R, SAS, SQL, MySQL, HTML, CSS, Java, SQLite3, MySQL.
Data visualization tools: Plotly in python, qqplot2 in R, shiny in R. Tableau.