JINGJING LIN

isjingjing.lin@gmail.com | Arlington, VA 22202
Portfolio jingjingl.georgetown.domains imjingjinglin

EDUCATION

GEORGETOWN UNIVERSITY, Washington D.C. 08/2018 – 05/2020

- Master of Science: Data Science and Analytics, GPA 3.5/4.0, GRE (Q: 166) 91st percentile
- Peer lead mentor, supervised and collaborated with other mentors to support ~70 first-year graduate students

UNIVERSITY OF MANCHESTER, Manchester, UK 09/2015 – 12/2016

- Master of Science: Management and Information Systems, GPA 3.3/4.0

TIANJIN POLYTECHNIC UNIVERSITY, Tianjin, China 09/2011 – 06/2015

- B.Eng.: Software Engineering, GPA 85/100 and B.Econ.: Finance(2nd), GPA 88/100
- 2014 Presidential First-Class (top 3%), and 2013 Second-Class (top 5%) Scholarship; Outstanding Graduate (top 2%)

SKILLSET

Programming	**Python**(scikit-learn, pandas), **R**(dplyr, glmnet), **SQL**, VBA(Excel-Macros), JAVA, HTML, CSS
Machine Learning	Regression(Linear/Logistic), Decision Tree, Clustering(K-Means, Hierarchical), Natural Language Processing(NLP), Deep Learning(CNN, RNN), Bayesian, Ensemble(Random Forest, Boosting)
Statistics	Probability, Distribution, Sampling, Hypothesis Testing, Bayes Theorem, Correlation
Cloud Computing	AWS (EMR, S3, Hadoop, MapReduce, Spark, git); Google Cloud (BigQuery, storage buckets)
Visualization & Tools	Tableau, Plotly, Matplotlib, ggplot2 and R-markdown; MySQL, Command Line, Jupyter notebook

EXPERIENCE

GEORGETOWN UNIVERSITY, *Data Science Development Engineer*, Washington D.C. 08/2020 – Present

- Developing website for tracking 100+ coronavirus vaccines research process from scientific papers and news
- Converting and unifying textual vaccines development process into data-oriented interactive visualizations using Tableau

CENTER FOR SECURITY & EMERGING TECHNOLOGY OF GEORGETOWN UNIVERSITY, *Data Science Research Assistant*, Washington D.C. 09/2019 – 12/2019

- Performed exploratory data analysis (EDA) on 3 academic publication datasets (130+ million rows of 14+ GB data) to characterize tech fields in Artificial Intelligence through BigQuery, storage buckets, and virtual machines in GCP
- Increased matching rates to ~20% in non-matched records from filtered databases by conducting textual analysis(NLP), including converting bags-of-words, vectorizing tf-idf and running text similarity algorithms

UNILEVER - DOLLAR SHAVE CLUB, *Marketing Technology Intern*, Los Angeles, CA 06/2019 – 09/2019

- Sole analyst responsible for analyzing and optimizing the current manual-operated 20+ spreadsheets with 1000+ records of Urchin Tracking Module (UTM) tags information
- Reduced UTM parameters setting time by 90% for the marketing-acquisition team, and implementing time by 33% for data systems team by independently developing the new tags management tool using VBA and SQL embedded in Excel
- Delivered the UTM tool to Acquisition and Data systems teams (20+ users) independently, designed a plan for long term maintenance and operations across the company
- Created a business proposal for 'DSC x Military' to build connections with military communities

WALL STREET TEQUILA CONSULTING (Startup), ***Research Analyst***, Shanghai, China 09/2017 – 04/2018

- Gathered and analyzed data on target firms' finance and development strategy; created periodic reports and guidebooks (4 chap.) on recruitment programs of 4 fields (finance, consulting, data and technology) across global markets
- Led resource management effort to create and restructure marketing materials that yielded 50% increase (from ~3000 to ~4500) in average view count of over 15 supported articles on Wechat platform
- Supervised an intern and 2 junior colleagues on document research methods and writing materials revision

CHINASOFT INTERNATIONAL LTD. (Gartner 2019 Top 100 global IT service providers) **,** Summers, 2012 – 2015
Software Development Engineer Co-op, Tianjin, China

- Designed and developed 4 types of systems: 'Dieting' fitness system ('15), Veterinary center management system ('14), Online shopping website ('13), Static social website ('12) with Java and MySQL during 3 consecutive summers
- Documented feasibility analysis reports and project development plans; delivered final presentations

PROJECTS (More projects can be found on GitHub)

Massive Data: Top Comment Identification in Reddit (Click) 04/2019 – 05/2019

- Accessed and loaded large datasets of Reddit comments(~500GB) in JSON from S3 and preprocessed data, including handling missing values, inconsistent values using PySpark in EMR
- Performed EDA with Spark SQL; created features in numeric (text-length) and categorized (scores) variables
- Conducted features encoding through MLlib; built a "pinned" comment identifier by applying features to logistic regression through Machine Learning pipeline. Average score of AUC for the testing data was higher than 0.90

Data Analytics: Where Should You Live for Your Health (Click) 09/2018 – 12/2018

- Acquired datasets through API and performed data wrangling (~20k rows) to classify water quality data with pandas
- Implemented clustering (e.g. k-means) and association rule mining analysis, visualized results by Tableau and Plotly
- Applied hypothesis testing on cancer using linear regression and classifiers including KNN, Naïve Bayes, SVM