JINGJING LIN

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

EDUCATION

Georgetown University, USA	– Master of Science, Data Science and Analytics	08/2018 – 05/2020
University of Manchester, UK	– Master of Science, Management and Information Systems	09/2015 – 12/2016
Tianjin Polytechnic University, CN	– Bachelor of Engineering, Software Engineering	09/2011 – 06/2015
Tianjin Polytechnic University, CN	– Bachelor of Economics, Finance	09/2011 – 06/2015

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, DBSCAN), Bayesian, Ensemble(Random Forest, Boosting), Deep Learning(CNN, RNN), Natural Language Processing
Statistics	Probability, Distribution, Sampling, Hypothesis Testing, Bayes Theorem, Correlation
Visualization	Tableau, Plotly, Matplotlib, ggplot2 and R-markdown
Cloud Computing	AWS (EMR, S3, Hadoop, MapReduce, Spark, git); Google Cloud (BigQuery, storage buckets)
Database & Tools	RDBMS: MySQL (JDBC) and Access; Command Line, Jupyter notebooks

EXPERIENCE

Data Science Development Engineer – Georgetown University, *Washington D.C.* 08/2020 – Present

- Developing methods to track news and scientific papers related to COVID-19 using APIs (web-scraping)
- Building data-oriented features (e.g. visualizations) to explain the scientific progress in the fight against COVID-19

Data Science Research Assistant – The Center for Security and Emerging Technology of Georgetown University, *Washington D.C.* 09/2019 – 12/2019

- Performed exploratory data analysis (EDA) on academic publication datasets to characterize tech fields in Artificial Intelligence through BigQuery, storage buckets, and virtual machines in Google Cloud Console
- Conducted textual analysis, including converting bags-of-words, vectorizing tf-idf and running text similarity algorithms, to increase matching rates across academic publication databases

Marketing Technology (MarTech) Intern – Dollar Shave Club (Unilever), *Los Angeles CA* 06/2019 – 09/2019

- Developed an Urchin Tracking Module (UTM) parameters generator tool independently to manage Ads campaign information using VBA(macros) and SQL; designed a plan for long term maintenance and operations across the company
- Implemented marketing integrations in tag management systems from Google Analytics to Adobe Analytics
- Created a business proposal for 'DSC x Military' to build connections with military communities

Research Analyst – Wall Street Tequila Consulting Inc., *Shanghai, China* 09/2017 – 04/2018

- Investigated the trend on target firms' recruitment plans and strategies to generate guides and periodic reports
- Created writing materials by restructuring resources to support marketing team (yielded 50% growth in average view count of 15 articles on WeChat platform) and consulting team (developing speech drafts and slides)

Software Development Engineer Intern – ChinaSoft International Ltd., *Tianjin, China* Summers, 2012 – 2015

- Designed and built UI, database and prototype for 4 systems: [1] 'Dieting Assistant' Fitness System (2015), [2] Veterinary center management system (2014), [3] Online shopping website (2013), [4] Static social website (2012) with Java, HTML, CSS and MySQL (JDBC) for 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

NLP: IMDB Rating Prediction by Modeling Movie Scripts (*Click*) 03/2019 – 04/2019

- Collected ~1300 film scripts from 22 genres and their IMDB ratings, performed text normalization e.g. case uniform
- Calculated and vectorized numerical and categorized features, including tf-idf, the mean number of words per sentence, and the frequency of parts of speech with "pos tag" using NLTK
- Trained linear regression and Random Forest models with feature combinations with sklearn ; compared the two models using Pearson's r with SciPy and demonstrated the performance of Random Forest reaching an accuracy of ~85%

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 them by Tableau and Plotly
- Applied hypothesis testing on cancer using linear regression and classifiers including KNN, Naïve Bayes, SVM