Yan Yin

Data Scientist

San Mateo, CA, 94403

(607) 319-9282 | yinyan5988@gmail.com | yinyan5988.github.io/ | yan-yin-588461168

Education

Cornell University *Ithaca, NY, USA*

Master in Applied Statistics **GPA: 4.0/4.0** *Aug. 2018 - May. 2019*

- Courses: Supervised Learning, Unsupervised Learning, Database Designing and SQL, A/B testing by Google (Udacity)

Renmin University of China *Beijing, China*

B.S in Statistics **Minor in Economic Statistics** **GPA: 3.78 /4.0** *Sep. 2014 - June. 2018*

- Courses: Statistical Inference, Time Series, Statistical Software, Probability Theory, Econometrics

Skills

- Data Science Skills: Data Analysis, Data Visualization, Data Profiling, Data Modeling, AB testing
- Machine Learning: Recommender System, Natural Language Processing, Neural Network
- Language: Python(Numpy,Pandas,Tensorflow, Pytorch), SQL, R(ggplot2), MATLAB, C++, SPSS, LaTex, SAS, Eviews

Work Experience

Ten-x Commercial *Sep 2019 - present*

Data Scientist, *San Mateo, CA*

- Analyzed website's page views data using Random Forest to rank viewers in their inclination to bid on the listed assets. Made it available as a tool for sales team to refine their strategy and targets of potential bidders.
- Developed and maintained an email recommender system to recommend relevant assets for users using logistics regression.
- Monitored click-to-open rate on recommendation email and conducted AB testing to further test the recommender system's efficiency.
- Trained the business entity deduplication model using Random Forest as the business entities data volume increased. Improved model's performance with hierarchical clustering.

Schlumberger Ltd. *Oct. 2017 - Mar. 2018*

Data Scientist, *Beijing China*

- Conducted data cleaning on historical oilfield trajectory data, collaborated with product managers and engineers to achieve feature engineering.
- Built a recommendation system using Bayesian method to suggest input parameters for clients when they are requesting trajectory design in the company's system.
- Implemented and optimized data pipeline using python to realize ETL on 3.5 million data from database.

Selected Projects

AB Test Experiment Design, Architecture and Inference, Data Scientist, *Nations Info Corp* *Jan 2019- June 2019*

Adivisor: Prof. Yang Ning, *Associate Professor of Statistics and Data Science Department*

- Led the designing of the A/B test workflow to compare multi-variants' impact on conversion rate of an e-commerce company.
- Designed randomization strategy, validated assumptions, calculated sample size and confidence interval in Python.
- Presented the workflow and results to manager. The script successfully deployed in production and improved future hypothesis testing processes.

Recommender System Design Research using Amazon Transaction data *Feb 2018- Jun 2018*

Adivisor: Prof. Yanyun Zhao, *Dean of Statistics Department*

- Completed data cleaning on Amazon's transaction data. Conducted literature review on matrix factorization and on Hao Wang et al.'s collaborative deep learning method for recommender system.
- Replicated Hao Wang et al.'s research result in integrating Stacked Denoising Autoencoder in matrix factorization. Using features such as users' demographic information and product descriptions to learned items' latent representations from Autoencoder.
- Received highest score on my thesis report for this project among my class, and put forward ideas of designing hierarchical neural network for future improvements on the model.

User Activity Modeling and Exploration with NLP, Analyst, *CVSC TNS Research* *Jul 2017- May 2018*

Adivisor: Prof.Yang Li, *Professor of Statistics Department*

- In a team of five, responsible for data cleaning on 50 million text data of social network with 200 million Daily Active Users. Dealt with sparse text data through including hashtag information and aggregating data on users' level.
- Conducted preliminary analysis and examined users' demographic information from the cleaned data, presented using data visualization of association rules, along with bar plot and pivot tables to present the results to our stakeholders.
- Used light-LDA to find out the hot topics among users' text data. Gave specific suggestions on most popular topics for next year's Spring Festival Gala.