Jinghui (Elaine) Zhao

(+1)415-8232658 | jzhao54@dons.usfca.edu | Linkedin | github.io/Jinghui| San Francisco, CA 94111

EDUCATION

***MS in Data Science* - University of San Francisco** *July 2018 - Expected June 2019*

Courses: Machine Learning, Distributed Computing, SQL, NoSQL, Data Acquisition, Time Series, Data Visualization, Experimental Design, Objective Oriented Programming in Python, Product Analytics

***BS in Engineering Management* - Beijing Jiaotong University** *Sep 2014 - June 2018*

Courses: Operations Research, Engineering Economics, Applied Statistics, Probability Theory

WORK EXPERIENCE

Data Scientist Intern | Orange Silicon Valley | San Francisco, CA *Nov 2018 - Present*

- Identified devices potentially belonging to the same user by applying TFIDF on the devices' locations (Wrote functions to calculate TFIDF score of GPS coordinates)
- Worked with Spark Cluster and Spark Parquet files. Implemented data cleaning/extraction and feature engineering using Spark and Pandas.
- Predicted users' choices on grocery retailing brands with machine learning models (Generalized Linear Model, Random Forest, Xgboost, stacking and bagging models).
- Visualized the influence range of retailers by creating heatmaps in terms of the customers' home GPS.
- Visualized users' locations dynamically with R(Google API), R shiny, Plot.ly, geopy and folium.

Data Analyst Intern | China Telecom| Xi'an, China *Dec 2017 - Feb 2018*

- Categorized customers on a Recency-Frequency-Monetary basis with unsupervised learning(K-means). Designed strategies and customized services for each segment to increase customers' loyalty.
- Analyzed and visualized customer complaint records. Derived insights and generated a technical report.

Supply Chain Analyst Intern | JD.com | Xi'an, China *June 2017 – August 2017*

- Conducted daily operations in order management system and warehouse management system, including data wrangling and inputting.
- Optimized Inventory by identifying economic order quantity and optimal reorder point.
- Assisted in selecting suitable locations for building a distribution center to improve delivery efficiency.

SCHOOL PROJECTS

Air Quality Index Prediction *[EC2, S3, Spark SQL, Spark ML, MongoDB]* *[Link]*

- Created features and built ML regression models to predict nation-wide AQI. Achieved RMSE of 15.13
- Implemented distributed storage and computing of large-scale data with S3, MongoDB and Sagemaker Notebook backed by AWS EMR.

Canadian National Bankruptcy Rates Prediction *[R(tseries, forecast, vars)][Link]*

- Predicted bankruptcy rate in next 36 months with time series models: ARIMA/SARIMA, Exponential Smoothing, ARIMAX/SARIMAX, and VAR/VARX to achieve RMSE of 0.1523
- Improved model performance by hyper-parameters tuning, feature selections, and model ensembling.

Quora Insincere Question Classification *[Keras][Link]*

- Combined 2 pre-trained embeddings, implemented bidirectional LSTM and attention layer in the neural network, incorporated statistical features of the text, conducted stratified cross-validation and achieved an F1 score of 0.68
- Improved the model performance by fixing misspelling and OOV words in the text.

User In-App Purchase Prediction*[Python(sklearn, Pandas, Numpy)][Link]*

- Designed and created features based on users' in-app behaviors with Spark and Pandas.
- Built Xgboost and Random Forest models to predict the probability of users making future purchases.

SKILLS

- **Programming:** Python (Scikit-Learn, Pandas, Numpy, PyMC3, Matplotlib), R(tidyr, caret, dplyr, gglot, shiny, glmnet), Spark(Spark SQL, Spark ML), PyTorch, C
- **Database:** SQL (PostgreSQL), NoSQL (MongoDB)
- **Tools:** AWS (EC2, S3, EMR, Sagemaker), Latex, Git, HTML, Excel, Bash
- **Analysis Techniques:** Collaborative Filtering, Isolation Forest, Gradient Boosting, Bayesian Statistics, A/B testing, LSTM, GRU, CNN, Ordinary Least Squares, Ridge, Lasso Regression, CBOW, Skip-gram