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, Deep Learning, Design of Experiments (A/B testing), Distributed Computing, SQL, Spark, Time Series, Data Structures and Algorithms, Linear Regression, and Data Visualization.

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

WORK EXPERIENCE

Data Scientist Intern | Orange Silicon Valley | San Francisco, CA. *Nov 2018 – Expected June 2019*

- Identified similar users by applying TFIDF on location data from mobile devices. Implemented ETL and feature engineering with Spark and Pandas.
- Accurately predicted users' choices of grocery stores based on their location data and demographic data with random forest, Xgboost, and generalized linear models.
- Visualized the influence range of retailers by creating heatmaps of customers' GPS data.
- Visualized users' locations dynamically with R (Google API), R shiny 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 services for each segment to increase customers' loyalty.
- Analyzed and visualized customer complaint records with Tableau.

SCHOOL PROJECTS

Image Visual Search and Recommendation *[Python (BeautifulSoup, Selenium), Keras, Flask, AWS RDS]*

- Designed and developed a visual search web application and deployed it on AWS.
- Implemented back-end development including web scraping, server building, and database management. Fed images through neural nets and calculated image similarity to provide results and recommendations.

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

- Created features, built ML regression models to predict nation-wide AQI achieving RMSE of 15.13.
- Implemented distributed storage system with S3, MongoDB and Sagemaker.

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

- Built Xgboost and random forest models to predict the probability of users making future purchases. Employed Bayesian hyperparameter optimization to increase tuning speed.
- Designed and created features based on users' in-app behaviors with Spark and Pandas.

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

- Predicted bankruptcy rate for the next 36 months using 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 selection, and model ensembling.

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

- Using 2 pre-trained embeddings, implemented a bidirectional LSTM and attention layer in the neural network. Incorporated statistical features generated from the text and classified insincere questions with a 0.68 F1 score.
- Improved the model performance by fixing misspellings and non-standard words.

SKILLS

- **Programming:** Python, R, Spark, PyTorch, C
- **Database:** SQL (PostgreSQL), NoSQL (MongoDB)
- **Tools**: AWS (EMR, Sagemaker, EC2, S3, RDS), Flask, Latex, Git, HTML, Excel, Tableau
- **Analysis Techniques**: Collaborative Filtering, Isolation Forest, Gradient Boosting, Bayesian Statistics, LSTM, GAN, CNN