Irene Yang

xyang68@dons.usfca.edu | 415-624-4393 | San Francisco, CA | LinkedIn: xi-irene-yang | GitHub: IreneYang218

Education

MS in Data Science - University of San Francisco July 2018 – July 2019

- **Courses**: Machine Learning, Natural Language Processing, Relational Databases (SQL), Experimental Design (AB Testing), Deep Learning, Linear Regression, Distributed Computing

BA in Economics - Xiamen University Sep. 2014 – June 2018

- **Courses**: Statistics, Econometrics, Microeconometrics

Experience

Data Science Intern, Reputation.com Oct. 2018 - Present

Sentiment Analysis via Attention *[NLP, PyTorch]*

- Classified review sentiment using Random Forest and XGBoost. Encoded reviews using BoW and TF-IDF
- Improved the accuracy to 93% with a deep learning sentiment classifier using LSTM and self-attention
- Developed a sentiment visualization for trigger words, to identify the keywords related to the sentiment

Keyword Extraction Tool **(Blog)** *[NLP, Python]*

- Constructed a multi-gram keyword extraction tool to identify 100 customer concern keywords from reviews
- Improved previous n-gram keyword extraction tool by 10% through syntactic dependency analysis

Operational Insights Report Generator *[Python, SQL]*

- Automated insights report including client's review volume, sources, sentiment by topics and sentiment trends, to enhance market team's productivity
- Improved data selection and sorting rules of the insights dashboard, to deliver the actionable insights

Research Assistant in Econometrics, Xiamen University Oct. 2017 - June 2018

Causal Effect Estimation using Machine Learning *[R]*

- Implemented causal inference with tree-based and LASSO-based machine learning methods on high dimensional data. Identified and visualized heterogeneous treatment effect on empirical data

Data Analyst Intern, Vanke Mar. 2018 - June 2018

Office Building Prospective Customer Analysis *[Python]*

- Developed web scraping tool to collect prospective customers data including public geographical and financial data, improved data collection efficiency. Analyzed customers features to support office building design.

Projects

Newsfeed Product Development *[Python, AWS]*

- Designed topic level controversy score and authors' impact score to measure the credibility of news
- Fetched news and authors Twitter information in daily basis. Built an ETL pipeline from APIs to AWS RDS.
- Extracted news topics using LDA and conducted sentiment analysis. Deployed model on AWS EC2

Mobile In-App Purchase Prediction *[Python, GCP]*

- Predicted user purchase within next 7 days with recall 0.88 using a stack of tree-based models (Top 5 Team)
- Analyzed over 20 GB user session data. Captured users behavior with time series feature engineering

Distributed NYC Parking Tickets Clustering Analysis (Paper Accepted) *[Spark, MongoDB, AWS]*

- Preprocessed and clustered 8G parking tickets based on vehicle characteristics and travel time using Spark-ML and Spark-SQL. Visualized parking violations in different clusters
- Built an ETL pipeline to load data from S3 to MongoDB on AWS EMR and compared the performance between different EMR configures. Reduced the time of data preprocessing and model training by 60%

Programming Skills

Languages: Python, PyTorch, SQL(PostgreSQL, Redshift), NoSQL(MongoDB), R

Big Data Techniques & Tools: AWS(S3, EC2, EMR), GCP, Git, Spark, Hive, Tableau

Statistics: Experimental Design (AB Testing), Hypothesis Testing, Time Series Analysis, Regression

Machine Learning: NLP, Random Forest, Gradient Boosting, Clustering, Neural Network