

Irene Yang

xyang68@dons.usfca.edu | 415-624-4393 | San Francisco, CA | [linkedin.com/in/xi-irene-yang](https://www.linkedin.com/in/xi-irene-yang) | [ireneyang218.github.io](https://github.com/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 (Spark), Data Structures and Algorithms

BA in Economics - Xiamen University

Sep. 2014 – June 2018

- **Courses:** Statistics, Econometrics, Linear Algebra, Probability and Calculus

Experience

Data Scientist Intern | Reputation.com | Redwood City, CA

Oct. 2018 – Present

Used ML & NLP techniques to assist company in managing online reputation.

- Developed a review sentiment classifier using a deep learning model with LSTM and self-attention to improve reputation assessment. Achieved 8% increase in model accuracy and developed a visualization to interpret results (Python, PyTorch).
- Extracted customer concerns by building a multi-gram keyword extraction tool using syntactic dependency analysis. Improved accuracy of previous tool (Python). [Blog: <https://bit.ly/2lwNTqD>]
- Built an automated operational insight reporting tool (SQL, Python) to assess strengths & weaknesses of client's user experiences. Increased marketing team's efficiency and supported product quality assurance.

Data Analyst Intern | Vanke | Fujian, China

Mar. 2018 – June 2018

A Fortune Global 500 company with \$44 Billion market cap.

- Conducted data acquisition & customer segmentation analysis (Python, SQL) to drive office building design investment decision.
- Constructed a web scraping tool to collect business data on over 3,000 companies. Improved data collection efficiency.
- Optimized building design investment decisions by predicting customers' space needs using customer segmentation (kmeans).

Research Assistant in Econometrics | Xiamen University | Fujian, China

Oct. 2017 – June 2018

- Implemented a proof of concept model on observational data to estimate underlying causal effects (LASSO, tree-based).
- Analyzed social media effect on Initial Coin Offering (ICO) funding using regression for over 1000 companies.

Projects

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

- Clustered 8 GB parking tickets data based on vehicle characteristics and travel time using Spark on AWS EMR. (Github)
- Built an ETL pipeline to load data from S3 to MongoDB. Optimized time efficiency via experiments on different configurations.

Mobile In-App Purchase Prediction [Python, GCP]

- Predicted user purchase in 7days on 20 GB user data using a stack of tree-based models with recall 0.88 (Top 5 Team).
- Captured user recency and monetization value with feature engineering and presented business insights. (Keynotes)

Newsfeed Product Development [Python, AWS]

- Developed an analytic news feed product with topic controversy and author impact analysis using Python. [Github]
- Deployed an end-to-end machine learning pipeline from data collection to topic modeling and sentiment analysis.

Technical Skills

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

Machine Learning: NLP, Random Forest, Boosting, Clustering, Neural Network, Recommendation System

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

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