# Irene Yang

xyang68@dons.usfca.edu | 415-624-4393 | San Francisco, CA | LinkedIn | Github | Portfolio

## 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.*

- Built a review sentiment classifier to improve reputation management using a deep learning model with LSTM and self-attention. Improved model accuracy and developed a sentiment visualization to interpret model results. (Python, PyTorch)
- Identified top customer concerns by building a multi-gram keyword extraction tool using syntactic dependency analysis. Improved accuracy of previous tool. (Blog)
- Built an automated operational insight reporting tool to assess client strengths & weaknesses of user experiences by optimizing SQL query. 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 & analysis (Python, SQL, Tableau) to support real-estate investment decision making.
- Built a web scraping tool to collect companies' public geographical and financial data. Improved data collection efficiency.
- Generated reports and visualizations on geographical and financial patterns to support office building design and investment.

**Research Assistant in Econometrics | Xiamen University | Fujian, China** Oct. 2017 – June 2018

- Analyzed social media effect on Initial Coin Offering (ICO) funding using regression for over 1000 companies.
- Implemented causal inference with tree-based and LASSO-based machine learning methods on high dimensional data.

## 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 on 20 GB user session 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. (Presentation)

**Newsfeed Product Development** *[Python, AWS]*

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

## Technical 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