# XIAOMENG "LAURA" YU

Los Angeles, CA 90007 | (213) 234-8588 | yuxiaome@marshall.usc.edu | https://laurayu0916.github.io

## EDUCATION

**University of Southern California** 08/2021 – 05/2023
Master of Science in Business Analytics GPA: 3.84/4.00
Courses: Data Manipulation, Supervised/Unsupervised Learning Method, NLP, Dynamic Optimization, Database Management

**University of Illinois at Urbana-Champaign** 08/2016 – 12/2019
Bachelor of Science in Actuarial Science (Minor in Business), summa cum laude GPA: 4.00/4.00
Courses: Statistics and Probability, Life and Annuities, Finance and Investments, Loss Reserve, Risk Management Practices

## INTERNSHIP EXPERIENCE

**Fashom** ***Miami, FL***
***Data Scientist Intern*** 05/2022 – Present
- Trained multi-label and multi-class sequence classification models on reviews using BERT and PyTorch, reached 92% accuracy
- Constructed a pipeline to deployed the NLP models to the recommendation engine to decrease the return rate
- Rectified the shirt sizes in customers' profiles based on their purchase history using Matrix Factorization in TensorRec
- Designed an body shape algorithm to classify customers into different body type using K-means clustering
- Improved the recommendation system by adding a content-based filtering algorithm based on body shape features using PyTorch

**Keck Medicine of USC** ***Los Angeles, CA***
***Business Analyst Intern*** https://github.com/laurayu0916/Patient-Intake-Workflow 01/2022 – 05/2022
- Headed Process Improvement Project to reduce patients' waiting time and improve inquiry-to-patient conversion rate
- Scrutinized 250 patient cases from 2020 to 2022 and summarized each case into standard patient intake process in PostgreSQL
- Calculated average, standard deviation, correlation with conversion for each intake step and visualized the result in Dashboard
- Scraped visa waiting days from travel.state.gov using Selenium Driver in Python

**Tencent** ***Shenzhen, China***
***Entertainment Market Analyst Intern*** 09/2020 – 01/2021
- Scraped monthly KPIs of 2,000 mobile games from App Annie and analyzed underlying reasons of performance fluctuations
- Fetched monthly Share of Voice Rankings from 15+ platforms on SensorTower and researched apps' advertising strategies
- Studied the difference in "mentions" on social media of *Genshin Impact* in U.S. (200k+ /month) *vs.* Japan (90k+ /month)
- Presented on *State of Survival* and *Free Fire*, including advertising costs, content marketing, operation objectives, target audience

## PROJECTS

**NLP Analysis for Dating App Reviews** https://github.com/laurayu0916/Dating-App-Reviews-NLP
- Investigated the underlying structure of *Bumble* app reviews and deep dived into *Bumble*'s business profile
- Tokenized reviews with BERT embeddings, clustered similar reviews and derived topics with C-TF-IDF
- Trained RNN and LSTM models to predict important reviews, *i.e.* reviews that would receive lots of Thumb-ups later on

**Vehicle Insurance Sales Prediction** https://github.com/laurayu0916/Vehicle-Insurance-Sales-Prediction
- Analyzed Health Insurance policyholders' profiles to predict whether they would also be interested in Vehicle Insurance
- Refactored Exploratory Data Analysis code into 5 functions and preprocessed data with PySpark DataFrame in Databricks
- Trained and tuned hyperparameters of Random Forest and Gradient Boosted Trees using PySpark MLlib with highest recall 0.92

**Anomaly Detection for Financial Transaction Data** https://github.com/laurayu0916/Fraud-Detection
- Researched transaction data from an e-commerce company to predict fraud transactions
- Extracted highly predictive time-related features and utilized SMOTE sampling to adjust imbalanced dataset
- Built a scoring system based on predict_prob to decide whether an transaction should be passed, declined or manually inspected

**LA Crime Data Analysis and Modeling** https://github.com/laurayu0916/LA-Crime-Analysis
- Studied the LA crime data to provide hints on when and how the number of crimes would change in this city via Spark SQL
- Identified top-3 danger areas and visualized the crime events w.r.t category and time to give insights on how to distribute the police

## SKILLS

**Programming:** Python, SQL, PySpark, MongoDB, Neo4j, Tableau, Excel
**Machine Learning:** Random Forest, XGBoost (LightGBM), Neural Network, TF-IDF, RNN, LSTM, Transformers and BERT
**Statistical Analysis:** Descriptive Statistics, Probability Methodology, Hypothesis Testing(A/B Testing), Time Series Analysis
**Certificate:** SOA Financial Mathematics (FM/2), SOA Probability (P/1), AWS Certified Cloud Practitioner