YI MING

Phone: +1 480-881-4410|**Email:** yming897@usc.edu
LinkedIn: linkedin.com/in/yiming123 **; GitHub:** https://github.com/yming2-bu/Yi_Ming-Projects-Papers

EDUCATION

M.S. in Applied Data Science (STEM) January 2022 - December 2023
Viterbi School of Engineering, University of Southern California, Los Angeles, CA
GPA: 3.73/4.00
M.S. in Business Analytics (STEM) August 2020 - December 2021
Questrom School of Business, Boston University, Boston, MA
B.S. in Business Data Analytics (STEM) January 2017 - May 2020
W.P Carey School of Business, Arizona State University, Tempe, AZ
GPA: 3.81/4.00; Summa Cum Laude Graduate; Dean's List 2017-2020

RESEARCH EXPERIENCES & RELEVANT COURSEWORK

Recommendation System Competition (Top 5 Students) January 2023 – May 2023
University of Southern California (DSCI 553 – Foundation of Data Mining)
Instructor: Prof. Wei-min Shen

- I am the winner of the Recommendation System Competition, ranked among the top 5 out of a total of 247 participants in the competition based on accuracy metrics (RMSE) of test dataset.
- Build a hybrid recommender system with content-based regression model (XGBRegressor) and collaborative filtering in Python.
- Performing feature engineering on the Yelp dataset comprising over 300 features for both businesses and users involves utilizing PCA and heuristic methods to extract and select the most relevant features.
- Fine-tuned hyperparameters of XGBRegressor with cross-validation and Gridsearch

User Study of Emotions, Budgets, and Restaurant Types on Food Choice Preferences January 2023 – May 2023
University of Southern California (CSCI 599 – Research Method and Analysis for User Study)
Advisor: Prof. Gale Lucas

- Designed and executed a Mixed-Factorial Experimental Design involving 85 participants to investigate the impact of emotions, budgets, and restaurant types on food choice preferences.
- Use rigorous statistical test (repeated measure ANOVA with between-subject factor added) on the experiment results, revealing statistically significant findings that shed light on the effect of emotions, budgets, restaurant types on food choice preferences.
- Conducted power analysis before and after the experiment to assess and ensure sufficient statistical power.

Representation Learning of user roles in Online Health Community (Ongoing) March 2023 – Present
Central University of Finance and Economics (Work Remotely)
Supervisor: Prof. Xi Wang

- Build an automated web-scraping program to collect 4,312,580 records of data from the Online Health Community
- Train a model by topic modeling techniques (LDA2Vec) to vectorize the users' posts (text data) in OHC
- Model interactions between users and threads' topics by dynamically embedding them into low-dimensional space
- Capture 9 users' roles by clustering users' dynamic embeddings via Gaussian Mixture Model
- *Skills: PyTorch, Recurrent Neural Network, LDA, Clustering.*

Supply Chain and Innovation Networks Research (Ongoing) March 2022 - Present
City University of Hong Kong (Work Remotely)
Supervisor: Prof. Zhiya Zuo

- Create an automated web-scraping program to collect patent data from Google Patents and Derwent API
- Parse the text data programmatically using Python BeautifulSoup package & Regular Expression
- Clean and manipulate raw data by using Python Pandas package
- Create a heuristic method to differentiate individual assignees and company assignees

Predictive Analysis for Flight Delays January 2021 - March 2021

Boston University, Boston, Massachusetts

Supervisor: Prof. Georgios Zervas

- Designed and completed a project on flight delay prediction as a team under the supervision of Prof. Georgios Zervas
- Gathered and organized flight data from the US department of Transportation for March 2019 & 2020 related to three target airports: LA, Atlanta, and Boston
- Trained and tested different models: Linear Regression, Ridge/Lasso Regression, Decision Tree, and Random Forest, and evaluated the performance of models by K-means fold cross validation

Text Data Analysis for Google Store Reviews January 2021 - March 2021

Boston University, Boston, Massachusetts

Supervisor: Prof. Brock Tibert

- Established a text analysis project to analyze reviews of mobile apps in Google Apps Store and decompose the attributes of the feedbacks as a team under the supervision of Prof. Brock Tibert
- Tokenized raw text data and split the data into positive and negative group based on the result of sentiment analysis
- Conducted sentiment analysis and clustering analysis (included K-Means Clustering and Hierarchical Clustering) by splitting and tokenizing the raw data and segmenting users into distinct groups according to tokened reviews
- Utilized LDA Topic Modeling to identify 6 groups of key attributes affecting the nature of reviews on Google Store

Spatial Data Science Research October 2019 – December 2019

University of Houston, Houston, Texas

Supervisor: Dr. Gangbing Song

- Built road network including road junction and segment tables from the open street map (OSM) road spatial dataset
- Detected Crash Hotspot Intersections through geographic information system visualization of the intersection hotspots shapefile

Spatial Data Mining Research May 2019 – August 2019

University of Houston, Houston, Texas

Supervisor: Dr. Gangbing Song

- Processed traffic data (road network, crashes, and region boundary) in spatial database by using PostgreSQL and PostGIS spatial data engine
- Built conceptualization of spatial relationships included two types: linear road relationships and road-crash relationships
- Identified road clusters with high-frequency crashes (RCHC) by using the cluster and outlier analysis (local Moran's I)

PUBLICATIONS

- Zhonggui Zhang, Yi Ming, Gangbing Song (Feb. 2020) "A New Approach to Identifying Crash Hotspot Intersections (CHIs) Using Spatial Weights Matrices", Applied Sciences https://doi.org/10.3390/app10051625
- Zhonggui Zhang, Yi Ming, Gangbing Song (Dec. 2019). "Identify Road Clusters with High-Frequency Crashes Using Spatial Data Mining Approach", Applied Sciences https://doi.org/10.3390/app9245282

PROFESSIONAL EXPERIENCES

Entity Data Analyst (Intern) September 2021 – December 2021

Connected2fiber, Milford, MA

- Mining valuable information from point of interest (POI) of OpenStreetMap (OSM) data.
- Validated and enriched our own database with OSM data through PostgreSQL
- Analyzed and correct entities' Domain Name System (DNS), soft deleted unreliable entities.

Official Facilitator for Boston University March 2021 – August 2021

Google Cloud, Boston, MA

- Gaining strong Cloud Computing skills with weekly hands-on labs
- Facilitating student at Boston University who are on Big Data & Machine Learning track.
- Designed my own Cloud Computing train program for students at Boston University which includes Machine Learning & Data Analysis on BigQuery; Data Processing on Trifacta; Data Engineering Workflow.

Data Analyst: Predicting the Fire Loss Price (Co-op) January 2021 - July 2021

Data Science Team of Bunker Hill Insurance, Boston, Massachusetts

- Preprocessed and Cleaned 4 different sources of data with more than 1500 variables
- Presented an exploratory data analysis summary included key measurements (claim rate, severity, and pure premium) by executing one & two variable analysis, which successfully pointed out the worst performance states in New England area.
- Conducted data integrity checking and sample design.
- Completed feature selection: features in top 90% feature importance by Random Forest
- Model evaluation: employed lift charts to evaluate model performances. Random Forest can capture the increasing pattern and identify the important top 20% risky loss.

Senior Capstone Project Team Lead (Co-op) January 2020 - May 2020

USAA (United Services Automobile Association), Tempe, Arizona

- Mined latest text data weekly from Twitter API to predict upcoming customer service priorities
- Presented a data analytics report weekly to help USAA employees anticipate any issues
- Directed text data mining for about 10,000 customer feedback/comments on Twitter to generate a tag cloud of negative customer service keywords
- Assessed USAA's new policy performance by conducting a sentiment analysis on Twitter (keyword COVID-19)
- Applied topic modeling to analyze customers' top 10 concerns and locations

RELEVANT SKILLS

Technical Skills: Spark/Hadoop, Web Scraping, Text Analytics, Machine Learning, Deep Learning

Programming Languages: Python, SQL