Li Sun

Cambridge, MA • 585-957-4337 • lsun@g.harvard.edu • LinkedIn • Website

EDUCATION

Harvard University — **Cambridge, Massachusetts**
Master of Science in Data Science — *Anticipated December 2022*
- GPA: 3.96/4.0
- Courses: Machine Learning, Systems Development, Linear & Non-linear Models, Data Visualization, NLP

University of Rochester — **Rochester, New York**
Bachelor of Science in Data Science & Bachelor of Arts in Financial Economics — *May 2021*
- GPA: 3.96/4.0; Phi Beta Kappa Honor Society; Dean's List All Eligible Semesters
- Courses: Data Mining, Database Systems, Computational Statistics, Data Structures & Algorithms, Time Series
- Leadership Experience: Producer of a theatre group (Led a group of 45 students, oversaw and orchestrated four productions)

PROFESSIONAL EXPERIENCE

Red Ventures — **Fort Mill, South Carolina**
Data Scientist Intern (Python, Spark, SQL) — *June 2022 – August 2022*
- Developed a machine learning pipeline to predict credit card revenue by querying and cleaning 6 million data, applying oversampling strategy, implementing LGBM, leveraging hyperparameter tuning and calibrating the model.
- Improved the AUC from 81% to 91% and paid search error rate from 20% to 3% compared to the current BAU model.
- Inferred an explainable conclusion by analyzing model outcomes and feature importance; effectively communicated the model design and performance results to stakeholders and internal leadership.
- Researched on the latest test-to-image generation technologies (DALL-E2 & Imagen) and presented to the AI product team.

University of Rochester Medical Center — **Rochester, New York**
Research Data Analyst (Python, R, EDA, NLTK, Face++, Ethnicolr) — *December 2019 – August 2021*
- Deployed Latent Dirichlet Allocation (LDA) and sentiment analysis to monitor 7TB Twitter data including users' discussions surrounding mental health concerns during the COVID-19 pandemic and the ban on flavored e-cigarettes.
- Employed deep learning algorithms (Face++ & Ethnicolr) to infer demographic composition of the 3000 target Twitter users.
- Performed time-series analysis to explore the trends of mental health conditions under pandemics in the US.
- Used Beautiful Soup to collect disposable e-cigarette brands and flavors from online stores, conducted EDA to detect the nature and the promotion techniques on social media after FDA banned e-cigarette usage of certain types.
- Published three papers as first author and three papers as second author to make recommendations on effective mechanisms for future communication about the danger of e-cigarette use.

University of Rochester Visual Intelligence & Social Multimedia Analytics Lab — **Rochester, New York**
Undergraduate Researcher (Python, PyTorch) — *May 2020 – August 2020*
- Quantitatively analyzed short video content-based cultural differences between Douyin and Tiktok, discovered different user behaviors and preferences, recommended differentiated commercialization strategies and application feature designs.
- Employed Faster R-CNN and I3D model to perform object and human action detections in 10k videos.
- Conducted statistical analysis, compared the distributional results through data visualizations along the content dimensions of object quantity, object categories, and human action categories.
- Published a paper as the first author and gave a presentation at IEEE Big Data Conference with 500 audiences.

RELEVANT PROJECT

Mortality Prediction and Interpretation, Harvard University — *October 2021 – December 2021*
- Applied PCA to deal with multicollinearity problem and reduce runtime and complexity for the later training; treated data imbalance and sample bias to prevent overfit with SMOTE & 5-fold cross-validation.
- Implemented KNN, Logit Regression, Random Forest, AdaBoost, and Neural Network, leveraged hyperparameter tuning and model ensemble for an optimal robust model, determined Random Forest to be the best model based on F1 score (0.843) and AUC (0.861); interpreted feature importance using SHAP value.

SKILLS

- **Programming:** Python (PyTorch, Scikit-Learn, NLTK, Pandas, Numpy), SQL, R, Java; Basic: JavaScript, HTML
- **Tools & Frameworks:** Spark, Git, AWS Certified Cloud Practitioner, Linux, Excel, Tableau
- **Product Skills and Others:** Hypothesis testing, A/B testing, Data modeling, Data visualization, EDA, ETL