# Li Sun

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

## EDUCATION

**Harvard University**                                                                                    **Cambridge, Massachusetts**
*Master of Science in Data Science*                                                                   *Anticipated December 2022*
- GPA: 3.96 out of 4.00
- Related Coursework: Machine Learning, Systems Development, Linear Models, Data Science, Data Visualization

**University of Rochester**                                                                                        **Rochester, New York**
*Bachelor of Science in Data Science & Bachelor of Arts in Financial Economics*                              *May 2021*
- GPA: 3.96 out of 4.00; Phi Beta Kappa Honor Society; Dean's List All Eligible Semesters
- Related Coursework: 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 with donation increases 10% every semester)

## 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, and leveraging hyperparameter tuning.
- Improved the AUC from 81% to 94% and paid search error rate from 20% to 1% compared to current BAU model.
- Summarized model design and performance results to stakeholders with data driven outcomes & actionable next steps.
- Researched on the latest test-to-image generation technologies (DALLE2 & Imagen) and gave a presentation to the data 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 pandemic in US.
- Used Beautiful Soup to collect available disposable e-cigarettes brands and flavors from online stores, detected the nature and the promotion techniques in disposable e-cigarettes related tweets 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 Tikok, 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 first author and gave a presentation on IEEE Big Data Conference with 500 audiences.

## RELEVANT PROJECT

**Mortality Prediction and Interpretation, Harvard University**                                **October 2021 – December 2021**
- Cleaned 15000+ data by changing data type, joining dataset, imputing missing data, and removing outliers.
- Applied PCA to deal with multicollinearity problem and reduce runtime and complexity for the later training.
- Performed KNN, Logit Regression, Random Forest, AdaBoost, and Neural Network on data, and determined Random Forest to be the best model based on F1 score (0.803) and AUC (0.861); Interpreted feature importance using SHAP value.

## SKILLS

- **Programming Language:** Python, R, SQL, Java, Linux; Basic: PHP
- **Tools & Frameworks:** PyTorch, Spark, NLTK, OpenCV, Pandas, Scikit-Learn, Git, Excel; Basic: AWS, STATA, Tableau
- **Product Skills:** Hypothesis testing, A/B testing, Data modeling/visualization