

Email: datascisteven@gmail.com
Portfolio: datascisteven.github.io
Mobile: (646) 991-4222

GitHub: [@datascisteven](https://github.com/datascisteven)
LinkedIn: [@datascisteven](https://www.linkedin.com/in/datascisteven)
Location: New York Metro

STEVEN YAN

I am a collaborative team player with excellent communication skills able to translate technical findings into understandable terms. With a passion for continuous learning, I am eager to contribute my expertise to solve challenging data-driven problems and drive innovation within your organization.

WORK EXPERIENCE

Data Scientist, *StartOut*, Remote (1/2022—7/2023)

- Performed Sparse PCA for dimensionality reduction of non-policy data into principal components to pass into the Difference-in-Differences model as features to demonstrate causality between policy measures and economic indicators to generate recommendations for each state and 120 metro areas
- Developed algorithms for imputing missing or unavailable data (i.e. gender/race guessing, estimation of LGBTQ+ population) and resolving or merging multiple data sources for main economic indicators
- Developed code for automating data collection process from sources
- Analyzed data to identify top insights on LGBTQ+ entrepreneurship for publication on website and reporting to key stakeholders
- Served as technical lead for team of college interns and tracked progress of assigned data research, entry, and cleaning tasks

VOLUNTEER EXPERIENCE

Machine Learning Engineer, *Omdena*, Remote

- 10/2023-Current: Team Lead on project to develop mock interview chatbot by fine-tuning an LLM model with custom dataset of Q&A's
- 7/2023-9/2023: Led team to develop deep learning model for identifying TB from chest X-ray images for deployment on Streamlit app
- 11/2021-1/2022: Developed image segmentation model with 95% accuracy for identifying schools in Sudan from high-res satellite images

PROJECTS

Tuberculosis Detector Using DenseNet Model ([Github](#), 9/23)

- Implemented Transfer Learning DenseNet-121 Model on dataset of 12200 Chest X-Ray images to achieve 95% accuracy and 96% F1-score

Hate Tweet Detector Using NLTK with Flask ([Github](#), 5/21)

- Implemented SVM algorithm on dataset of 24783 tweets to achieve 92% accuracy and 96.5% ROC-AUC

OTHER EXPERIENCE

Educational Consultant (Freelance), New York Metro (8/2009—10/2020)

- **MCAT Content Fellow**, *Khan Academy* (2014-2015): Created MCAT practice content for the Khan Academy learning platform
- **MCAT Bootcamp Manager**, *Mentoring In Medicine* (2009-2011): Developed and delivered multiple rounds of intensive MCAT bootcamp
- **Teacher Trainer/Curriculum Developer**, *Kaplan*, New York, (2/2006-5/2009)
- Managed team of 15 instructors in Manhattan/Bronx region

PUBLICATIONS

- [StartOut State of LGBTQ+ Entrepreneurship Report](#) (2022)
- Published multiple articles for Medium (datascisteven.medium.com)

EDUCATION

Flatiron School
(10/2020—3/2021)
*Data Science Online
Immersive Bootcamp*

University of Chicago
(9/1996—6/2000)
*BA Biology, 3.7 GPA
College Honor Scholarship
Dean's List*

TECHNICAL SKILLS



Programming Skills
Python, SQL, HTML/CSS

Machine Learning
*Predictive Modeling
Scikit-learn, Tensorflow*

Data Cleaning / Munging
Numpy, Pandas

Statistical Analysis
Excel, SciPy

Data Mining & Interpretation
*Data Visualization
Matplotlib, Seaborn, Tableau*

Version Control
Git, Github

Cloud Platform
Google Cloud Platform

CERTIFICATIONS

**DeepLearning.AI Tensorflow
Developer Specialization**

**DeepLearning.AI Deep
Learning Specialization**

REFERENCES

Available Upon Request