Contact

www.linkedin.com/in/connorcapitolo (LinkedIn)
github.com/connorcapitolo (Other)

Top Skills

Machine Learning

C++

Python (Programming Language)

Certifications

Taming Big Data with Apache Spark and Python

Connor Capitolo

Looking to utilize data to guide business decisions. Interested in the intersection of data science, machine learning, and software engineering.

United States

Experience

Day Zero Diagnostics, Inc

Associate Data Scientist

July 2022 - Present (10 months)

Boston, Massachusetts, United States

• Co-lead quarterly production run; use pandas and matplotlib to analyze model performance and present to CTO

• Extend production environment's modeling pipeline to incorporate held-out test sets in order to automate model performance comparisons over time; process automation now provides real-time results for senior leadership and improves team efficiency

• Create Python modules that build PostgreSQL queries for data ingestion, utilize the Prefect workflow automation package in conjunction with Docker and GCP to allow for parallel computing across models, log experiment results to MLFlow, and develop appropriate unit and integrations tests that are incorporated into a CircleCI job

• Manage versioning, documentation, and package updates for modeling codebase; incorporated support for M1 chip

Harvard University

Teaching Fellow for CS107: Systems Development for Computational Science

August 2022 - December 2022 (5 months)

• led weekly office hours and pair programming sessions on Python, Git, Bash, Linux, DevOps basics, containers and databases; graded biweekly homework assignments; team lead for final projects

Harvard University

1 year 8 months

Teaching Fellow for CS109b: Advanced Topics in Data Science

January 2022 - May 2022 (5 months)

Cambridge, Massachusetts, United States

• led weekly office hours on GAMs, clustering, Bayesian statistics, CNNs, RNNs, Transformers, autoencoders and reinforcement learning; graded biweekly homework assignments; team lead for final projects

Harvard Advanced Practical Data Science: Crypto Forecasting App Developer

August 2021 - January 2022 (6 months)

• Built a web app that accepts real-time data from Binance and produces minute-by-minute predictions using a LSTM
• Utilized Google Cloud Platform (GCP) for storing the data queried from Binance (Cloud SQL), the Docker containers (Google Container Registry), and the models (Google Bucket)
• Utilized a GCP virtual machine and Kubernetes cluster for running the React frontend to display history and predictions

Harvard Data Science Capstone Project; Spotify Disagreement Detection Developer

August 2021 - December 2021 (5 months)

Cambridge, Massachusetts, United States

• Studied the detection of disagreement in Spotify podcast episodes based on audio data and automatic transcriptions
• Created an annotated podcast disagreement dataset, established a data processing pipeline, and utilized AWS for ML predictions
• Presented findings to Spotify partners along with a research paper; created GitHub that can be used by future researchers

Teaching Fellow for CS107: Systems Development for Computational Science

August 2021 - December 2021 (5 months)

• led weekly office hours and pair programming sessions on Python, Git, Bash, Linux, DevOps basics, containers and databases; graded biweekly homework assignments; team lead for final projects

Data Engineer, Economics and Computer Science Research Group

October 2020 - July 2021 (10 months)

Cambridge, Massachusetts, United States

• Collaborated with Tata Communications leadership to predict if sales opportunities are won based on employee email correspondence using BERT’s deep learning NLP classification model
• Managed storage, resources, and security for 1TB database of sensitive email data on dedicated Harvard server

• Maintained and documented GitHub repository of 10 contributors; created standardized Git, GitHub, and Anaconda practices
• Wrote Python and Bash scripts utilizing Linux commands to ingest emails from ElasticSearch and perform preprocessing in an embarrassingly parallel manner; expanded experience with Python multiprocessing and Harvard’s supercomputer
• Created executive summary for senior management to outline the data preprocessing steps from ingestion up to NLP classification modelling

Seagate Technology
Data Science Intern
May 2021 - August 2021 (4 months)
Longmont, Colorado, United States

• Collaborated with six senior data scientists as only intern to build imbalanced classification model based on customer data that can predict hard drive failures utilizing ~5M observations and 270 features; project goal was to identify patterns and factors that lead to hard drive failures, thereby improving performance of product and reducing customer replacement costs
• Led development of novel deep learning time-to-event model using Cox Proportional Hazards and Neural Networks that improved precision by 25%; created foundation for model that will be used in production
• Created Python scripts for ingestion of data using PostgreSQL, preprocessing and feature engineering using Pandas, dimensionality reduction using R, modelling using time-to-event package, Bayesian hyperparameter optimization using hyperopt, and logging data using MLFlow
• Built and presented proof-of-concept to senior leadership showing how model would perform in production setting

Seagate Technology
Data Science Quality Intern
June 2020 - August 2020 (3 months)
Longmont, Colorado, United States

• Built a machine learning model based on an imbalanced classification problem using supplier data composed of ~1.8M observations and 166 variables to identify patterns and factors that lead to hard drive failure during certification and reduce manufacturing costs; presented findings to Seagate data science team
• Presented the benefits of using multivariate statistical processing control to management at Seagate’s Minnesota wafer factory in order to better predict and understand the variables that lead to wafer failure

• Recreated an autoencoder model for anomaly detection to identify heads that should be removed from the factory line in Thailand in order to improve quality and reliability of final product
• Visualized trends in missing value data from Seagate's head factory in Thailand to ensure quality of the dataset

Vanderbilt University

Research Assistant, Vanderbilt Neuroimaging & Brain Dynamics Lab
October 2019 - August 2020 (11 months)
Nashville, Tennessee, United States

• Collaborated with professor and classmate to provide more insights into understanding brain function, specifically neural connectivity, to help analyze and diagnose neurological diseases and disorders like Parkinson's and ADHD
• Analyzed fMRI data collected from the Human Connectome Project's 100 Unrelated Subjects to determine if the cortex, thalamus, or a combination of the two contribute most to determining an individuals' cognitive scores
• Created cortex, thalamus, and cortical-thalamus matrices using functional connectivity with ridge, lasso, elastic net, and K-nearest neighbors regressions to predict cognitive scores like an individual's processing speed
• Wrote preliminary report on findings (link: https://github.com/connorcapitolo/Neuroimaging-Research)

Booz Allen Hamilton

Booz Allen Summer Games Intern
June 2019 - August 2019 (3 months)
Washington D.C. Metro Area

• Created proprietary recruiting system to better attract and retain top talent; presented project to Booz Allen Hamilton senior management and pitched product to major government agency
• Led the implementation of Pega Infinity, a workflow automation software, to build a proof-of-concept workflow for improved recruiting and processing of potential candidates
• Utilized Agile (software development lifecycle methodology) and Jira (task management and work allocation tool) to improve communication and efficiency across intern project team
• Used logistic regression to calculate the probability of a candidate receiving an interview based on skill set and work experience in order to expedite candidate pre-screening process

• Developed a text mining matching algorithm in Python to match candidates and interviewers based on skill sets, location, commuting preferences, and academic background in order to lower recruiting administrative costs
• Built dynamic data fields that combined Boolean logic with complex mathematical functions to automatically clean and prepare incoming applicant data for use in machine learning algorithms and statistical analysis

STRIVR

Data Analyst Intern
May 2018 - August 2018 (4 months)
San Francisco Bay Area

• Transformed complex data using Excel features like VLookup, SumIf, and Pivot Tables to create meaningful summaries that identified cost savings and improved efficiency
• Forecasted 2018 travel spend (representing ~11% of STRIVR's annual expenses) for employees and projected hires using data from NetSuite ERP software to help CFO determine when company will need to raise capital
• Led the implementation of STRIVR's first 401(k), which involved evaluating and hiring a financial advisor and financial provider; selected investment funds with more than 1.2 million in assets while considering the needs of key employee stakeholders
• Managed relationship with NexTravel, STRIVR's corporate travel partner; identified and reconciled discrepancies between actual employee travel data and travel reports resulting in 7% savings due to improved travel policies; recommended NexTravel contract renewal
• Audited employee ride-share spend using Lyft data; analysis was used to cut Lyft spend by 9% by identifying excessive business rides and to evaluate the cost-benefit of a commuter benefits program
• Created document summarizing all client contracts for relevant financial milestones to assist CFO in cash-flow revenue recognition projections

STRIVR

Operations Intern
June 2017 - July 2017 (2 months)
Menlo Park, CA

• Assisted in day-to-day running of the office as STRIVR grew 12% to 50 persons
• Worked cross-functionally to solve problems and ensure smooth operations while interfacing with all departments; reported directly to Director of Finance and COO

• Performed product and quality assurance checks of STRIVR hardware and software
• Created a comprehensive Travel and Expense Policy tailored to STRIVR's culture, needs and goals
• Analyzed expense reports and customer acquisition costs; developed margin analysis on STRIVR projects
• Developed a SWOT analysis of STRIVR based on write-ups by Stanford GSB students which identified significant opportunities. Presented actionable recommendations to CEO and VP of Strategy
• Reviewed all client contracts for accuracy, consistency, and correct pricing; reorganized contract database into format that is easily searchable and retrievable by senior management

Education

Harvard University

Master of Science in Data Science, Data Science · (September 2020 - May 2022)

Vanderbilt University

Bachelor of Arts, Magna Cum Laude, Economics, Mathematics and Computer Science · (2016 - 2020)

Williams College

Mathematics · (2015 - 2016)

The Branson School

High school · (2011 - 2015)