Sean Crofut

sean.crofut@gmail.com | (408) 489-8289 | linkedin.com/in/crofut | seancrofut.github.io

EDUCATION

University of California, Berkeley **Berkeley, CA** | May 2023
B.A. in Data Science with a Concentration in Economics

Relevant Coursework:
Computational Structures in Data Science, Data Mining and Analytics, Data Engineering, Data Inference and Decisions
Data Structures, Algorithms, Machine Learning, Databases

Certificates:
Programming Foundations with Javascript, HTML, and CSS **Online course - Duke University** | August 2021

SKILLS

Computer Languages:	Java, Python, C++, R
Python Packages:	Seaborn, Matplotlib, Pandas, Numpy, Scikit-learn, Tensorflow, Keras, Scipy, PyTorch
Databases and Visualizations:	MySQL, PostgreSQL, MS SQL, MongoDB, Tableau
Other Software:	AzureML, Excel, Microsoft Office Suite, Google Suite, Git

WORK EXPERIENCE

Firstly **Berkeley, CA** | January 2023 - March 2023
ML Engineer Intern

- Implemented a recommendation system between users using a clustering algorithm for collaborative filtering
- Migrated data from Airtable to Postgres and used data pipeline to reduce data processing time by 50%
- Integrated search engine into mentorship app to enhance user search filtering and sorting

PROJECTS

Spam E-mail Classifier

- Created a logistic regression model that predicted email types with over 90% accuracy using Scikit-learn
- Used feature engineering on keywords to classify emails as spam or regular based on their appearance
- Reduced the rate of falsely flagged spam to 1% by prioritizing the handling of false positives

Cook County Housing Prices Predictor

- Implemented a data processing pipeline to ingest housing data from the Cook County Assessor's office
- Reduced overfitting using a LASSO regression for model regularization to better apply the model on new data
- Fit features from CCAO dataset to a linear regression model to predict housing prices with 92% accuracy

NLP Song Genre Identifier

- Implemented a natural language processing model to identify song genres based on English lyrics
- Cleaned lyrical data using word vectorization, stemming, lemmatization, and word embedding methods
- Identified song genres with different lyrical content with 90% accuracy

COVID Infections Analysis

- Scraped global data on COVID containing from the web and filtered data in Excel
- Imported data into SSMS for data wrangling and exploratory data analysis using SQL
- Visualized cleaned data in Tableau to gather insights on COVID infections and deaths by location