YEWEN ZHOU

https://www.linkedin.com/in/yewen-zhou/ (626) 492-9028
https://hegelim.github.io/personalwebsite/ yz4175@columbia.edu

EDUCATION

Columbia University New York, NY
M.S. in Data Science Dec 2022

- GPA: 3.80 / 4.0
- Coursework: Algorithms, Big Data, Causal Inference, Machine Learning, Applied Deep Learning (fall 2022), Finance for DS

University of California, Berkeley Berkeley, CA
B.A. in Data Science, Business Analytics Concentration May 2021

- GPA: 3.93 / 4.0, Phi Beta Kappa Society
- Coursework: Data Structures, Time Series, Artificial Intelligence, Probability and Statistics, Decision Analytics, Intro to Finance

SKILLS & TECHNOLOGIES

Programming:	Python, Jupyter, Linux, SQL, R, Java, HTML5, CSS, JavaScript
Python Packages:	pandas, pytorch, tensorflow, keras, pyspark, numpy, scipy, scikit-learn, matplotlib
Frontend Frameworks & Cloud Services:	Django, Bootstrap, Plotly, AWS, Google Cloud Platform
Development Tools:	Git, Docker, VSCode, PyCharm, RStudio
Writeup:	Markdown, reStructuredText, LaTeX

WORK EXPERIENCE

Scry Analytics, Inc San Jose, CA
Data Science and Engineering Intern May 2022 – Aug 2022

- Benchmarked 30 text recognition models from 5 open-source repositories using PyTorch, Docker, AWS
- Generated synthetic dataset from 1,791 images with existing tags for chart detection model training
- Reduced ABINet recognition model inference time by half, significantly making the current product more competitive
- Trained detectron2 deep learning model for chart detection with image augmentation, achieving 82 AP in test set
- Contributed to a million-dollar worth project in extracting key-value pairs from bar charts and finished the base version

SAFE Lab, Columbia University New York, NY
Data Scientist Oct 2021 – May 2022

- Matched 200 medical notes based on cosine similarities; trained logistic regression classifier on Bag of Words (BOW) and TF-IDF matrices with hyper-parameter search (sklearn, google cloud platform), achieving cross-validation recall 0.99
- Combined 3 tables with more than 2,000 rows and grouped with datetime intervals for each medical record number (MRN), allowing convenient table lookup for team members (pandas, numpy)

iQIYI, Inc Beijing, CN
Ads Algorithm Backend Intern May 2021 – Aug 2021

- Developed a testing framework for ads allocation emulator with more than 10,000 records; deployed in the server launched overseas in more than 5 countries (pandas, logging, numpy)
- Created a SARIMA time series module for ads inventory prediction, achieving a cross-validation RMSE less than 0.2 (statsmodels)
- Implemented High Water Mark (HWM) algorithm from scratch (logging, numpy, pandas) based on Yahoo research paper for compact allocation; used as the 1st version by algorithm and product teams of more than 10 people

PROJECTS

Columbia University, Realtime Twitter Sentiment Analysis Nov 2021 – Dec 2021

- Developed 6 ML models including Linear Regression, Ridge Regression, Gradient Boosting, AdaBoost, Random Forest, and SVR for aggregated twitter sentiment prediction, attaining test RMSEs less than 0.1 (sklearn)
- Leveraged Virtual Machine (VM) on Google Cloud Platform (GCP) to decrease model training time by 16x
- Created a dashboard using Bootstrap, Django, HTML5/CSS/JavaScript/Plotly, displaying real-time Twitter sentiment prediction

Columbia University, Stock Price Prediction Nov 2021 – Dec 2021

- Utilized Airflow Scheduler to collect stock prices from 5 tech companies automatically daily at 7 am
- Trained and updated 5 linear regression models for stock price prediction; obtained relative errors less than 0.01

Open Source, The solveminmax Python Package Jul 2021 – Sep 2021

- Implemented an object-oriented, open-source Python module to solve a sum of min and max equations applying regular expressions, numpy, sympy, and matplotlib
- Designed unit tests using pytest to validate module extensively with more than 30 testing cases
- Distributed on the Python Package Index (PyPI) with documentation hosted on GitHub written in reStructuredText and Markdown