Yang Wu

Chicago, IL | Princeton, NJ

6984541877 | yangwu@gmail.com | github.com/YangWu1227 | linkedin.com/in/yangwu1227

Personal Profile

I am a University of Chicago Master's student in the Data Science program. I specialize in statistical learning and machine learning and have two years of expertise in building data science products. I am searching for machine learning, data science, and Python or R development positions.

Education

University of Chicago

Chicago, IL

M.S. Applied Data Science

Sept 2022 - Current

- **Cumulative GPA:** 4.0
- **Machine Learning and Predictive Analytics:** Multi-objective recommender systems (collaborative and content-based filtering, matrix-factorization & co-visitation matrix approaches), GLM and statistical models, regularized & robust regression (Lasso, Elastic Net, IRLS), SVM, decision trees, gradient boosted machines and random forests
- **Deep Learning:** Theory, paper review, and toy implementations of deep learning architectures for computer vision (image classification, object detection, segmentation), NLP (translation, summarization, classification), and structured data problems, Model (distributed) training & deployment practice with Tensorflow/Keras and PyTorch on AWS Cloud
- **Data Mining Principles:** Association and sequence rules discovery, dimensionality reduction techniques (T-SNE, PCA, MSD, UMAP, truncated & randomized SVD), graph models (Bayesian Network & graph models), clustering (partition-based, density-based, hierarchical, K-prototypes for mixed data), NLP vectorization (TF-IDF, LSA, word embeddings with Word2Vec & GloVe), statistical density-estimation and probability distribution fitting
- **Big Data Platforms:** MapReduce fundamentals (examples in matrix-vector multiplication, relational-algebra operations, computing projections), data engineering with PySpark (implementing relational operations in SparkSQL and custom SparkML transformers for machine learning problems), similarity & duplication detection (sim-hashing, min-hashing, LSH) with PySpark, cloud platforms (learning to set up & configure AWS (EMR) and GCP (DataProc) for cluster computing workflows, and managing auto-scaling, storage, and identity)
- **Time Series Analysis and Forecasting:** Decomposition techniques (seasonal-trend decomposition using LOESS, Holt-Winters, singular spectrum analysis), statistical approaches (exponential smoothing, ARIMA & (S)ARIMA(X), VAR(MAX), GARCH, bootstrapping & bagging), deep-learning based approaches (dilated CNN, RNN (LSTM & GRU))
- **Bayesian Methods:** Conjugate distributions, Bayesian Networks (directed acyclic graphs, structure learning, inference & hypothesis testing), Non-parametric Bayesian models (robust regression, mixed-effects models, hierarchical models), Markov Chain Monte Carlo (Metropolis-Hastings, Gibbs, Hamiltonian sampling)

Kenyon College

Gambier, OH

B.A. in Statistics & Economics

Sept 2017 - May 2021

- **Cumulative GPA:** 3.76 *Magna Cum Laude* and Merit Lists (2018, 2019, 2021)
- **Math courses:** Probability Theory, Combinatorics, Calculus (I, II, III), Linear Algebra
- **Statistics courses:** Mathematical Statistics (limit theorems, exponential families, sampling distribution, order statistics, interval estimation, point estimation, maximum likelihood estimation, moment generating functions, likelihood ratio tests, and hypothesis testing), Data Analysis (ANOVA variants, mixed-effects models, and non-parametric models with applications in medical studies and data), Advance Regression Model (count data regression models for frequency of events in health sciences and social science, quantile regression models for distributional effects of treatments and outlier detection, duration models for modeling social science phenomena like duration of unemployment)
- **Economics courses:** Econometrics (regression discontinuity design for causal-inference and panel data analysis for modeling within-individual and between-individual variation), Mathematical Economics, Financial Accounting, Portfolio Analysis

Projects

Twitter Data Analysis

Chicago, IL

University of Chicago

2022

Business Objective

- Identify whether Twitter can be considered a credible source of information for the emergence of important trends or topics in education

Technical Implementation

- Extracted and processed ~ 100 million Tweets (~ 500GB) using **PySpark** and related NLP libraries (e.g., NLTK, spaCy) on **Google Cloud Platform**
- Identified the most prolific influential Twitterers using metrics like tweet volume and retweets and classified them into groups (government entities, universities, non-profit, news outlets, influencers)
- Conducted similarity analysis for tweet content using pre-trained sentence-transformers on **hugging face**
- Conducted sentiment analysis to classify education-related tweets as positive, negative, and neutral using fine-tuned pre-trained models on hugging face

Outcome

- **Insights:** Presented findings and recommendations (Twitter's score on trustworthiness for education-related content) to stakeholders using interactive data visualizations in a dashboard (**Plotly Dash**)
- **Application:** An education-tweet sentiment analyzer deployed using **Gradio**

Deciphering Micro-Business Density Growth

Chicago, IL

University of Chicago

2022

Business Objective

- Help policymakers identify US county-level microbusiness patterns and improve resource allocation
- Facilitate investment decisions of institutional investors such as loan providers and venture capitalists

Technical Implementation

- Used unsupervised nearest neighbors learning to identify similar counties in high micro-business density clusters and identified their shared characteristics and trends
- Built possessing and feature selection pipelines using cudf, cuml and scikit-learn utilities and **AWS Sagemaker processing jobs**
- Trained ETS & (S)ARIMA(X) models as baselines and deep-learning based (RNN with GRU & LSTM cells) models on **AWS Sagemaker** for forecasting; tuned and evaluated models using time-series cross-validation

Outcome

- **Business Insights**: A dashboard (**plotly Dash**) containing visualizations, tables, insights on factors conducive to micro-business density growth, and recommendations for different types of stakeholders (government entities, private investors, service providers, NGOs); deployed application using **AWS Elastic Beanstalk**
- **Forecast Service**: A web API with an HTTP endpoint (**AWS Lambda** and **API Gateway**) for API-based inference returning json-structured forecasts

Work Experience

Home Partners of America

Chicago, IL USA

Data Science Intern

June 2023 - Present

- Developed computer vision models, trained on real estate images, for classification and object detection tasks used to automatically identify features (e.g., pool, stainless fridge, flooring materials, etc.) pivotal to property valuation and validation, investment decision, and rent pricing for the leasing business.
- Deployed trained model with **AWS Lambda** & **API Gateway** utilizing batch inference jobs for large amounts of images.
- Leveraged **AWS GroundTruth** & **Roboflow** to set up a streaming data annotation task in preparation for continuous training; independently implemented and oversaw labeling task UI design and labeling workforce management.
- **Technical Skills:** Deep-learning model implementations with Keras, Model training & deployment with **AWS SageMaker**, Docker images for model training and serving
- **Soft Skills:** End-to-end project execution, Cross-functional collaboration, Bridging technical-business gaps

Citizen Data

Washington DC, USA

Jr. Data Scientist

Jan 2022 - July 2022

- Built, tested, & deployed end-to-end models (cluster, anomaly detection, social network, text classification models) on US voter and survey data underpinning our data science products— voter segmentation, social network analysis, campaign advertisement A/B testing and resource allocation.
- Implemented monitoring systems (**AWS CloudWatch** & **Lambda**) for data drifts (label & covariates) based on statistical tests such as the Kolmogorov-Smirnov test (continuous) & the Chi-square test of independence (categorical).
- Proposed, built, and deployed (Docker & Shinyproxy) our first interactive dashboard in R shiny, which later became an official product offering as a front-end to our data science products (model predictions, visualizations, tables, insights and interpretations on policy and elections) refactored with Python Dash.
- **Primary tasks:** Data mining & feature engineering, Scikit-learn & hugging face pipelines, In-house package development
- **Technical Skills:** Model training, tuning, validation, inference, and monitoring, dashboard development for model operations
- **Soft Skills:** Client Presentation Q & A, Time Management, UX & UI

Citizen Data

Washington DC, USA

Engineering Intern

Oct 2022 - Dec 2022

- Developed a three-module Python package (with **Poetry**) to support in-house data engineering & feature engineering tasks. Two modules include custom transformers & pipeline utilities for processing jobs, using packages & libraries such as pyspark, KerasNLP (for text data), cudf, and pandas (for structured data). The third module encapsulates cloud storage (AWS S3) & data warehouse (Redshift) related tasks, using mainly **psycopg2**, **boto3** & the AWS SDK.
- Developed an R package to facilitate fast proprietary report generation for our political scientist team.
- Developed an internal self-service drag-drop UI tool (flask-app) for non-technical team members to produce query results (**Amazon Redshift**) for ad hoc data requests; achieved cost-saving by avoiding purchasing a paid albeit more featureful SQL builder tool.
- **Primary tasks:** Features & unit tests for internal packages, build docker images for packaging and deploying the custom packages for the cloud, Data pipelining
- **Technical Skills:** Git & Github, Unit Testing with PyTest, CI-CD with GH actions, Docker, SQL, Python & R development, AWS SDK
- **Soft Skills:** Proactivity, Attention-to-detail, Problem-solving, Creativity

Skills

Programming	Python (PySpark, Polars, Cudf, Dask, Cuml, Tensorflow, Pytorch, Scikit-learn), R (Shiny, Data.table, Tidyverse), Reshift/Postgre/MySQL
Miscellaneous	Git, Docker, Shell (Bash/Zsh), LaTeX(Overleaf/ Markdown), Power BI, Tableau, AWS Quicksight, Microsoft Office
Personal Traits	Quick learner, Humility, Self-motivated, Curiosity, Conscientiousness

References available upon request