JACK ETHEREDGE, PHD

DATA SCIENTIST

jack.etheredge@gmail.com
jacketheredge.com
(803) 367-1152
New York City, 10016
jack-etheredge-phd/
Jack-Etheredge

SKILLS

PROGRAMMING: Python, R, bash/shell, Matlab, SQL
BIG DATA: AWS, Hadoop, Hive, Spark (and PySpark), PostgreSQL, MongoDB (and PyMongo)
MACHINE LEARNING: linear regression, supervised learning classification, unsupervised learning, natural language processing, deep learning, tree methods
GENERAL LIBRARIES/FRAMEWORKS: pandas, NumPy, Scikit-learn
VISUALIZATION AND WEB TOOLS: Flask, dash, Jekyll, plotly, D3.js, seaborn, matplotlib
DEEP LEARNING: Keras, TensorFlow, PyTorch, CNNs, RNNs, fastai
WEBSCRAPING: Beautiful Soup, Scrapy
VIDEO ANALYSIS/COMPUTER VISION:
NATURAL LANGUAGE PROCESSING: Topic Modeling, Sentiment Analysis, Word2Vec

RELEVANT COURSES

Fast.ai Advanced Deep Learning · fast.ai
Jeremy Howard, University of San Francisco

Deeplearning.ai Deep Learning Specialization · Coursera
Andrew Ng, Stanford University

Mathematics for Machine Learning Specialization · Coursera
David Dye, Samuel Cooper, Marc Deisenroth; Imperial College London

Using Python to Access Web Data · Coursera
Charles Severance, University of Michigan

Machine Learning · Coursera
Andrew Ng, Stanford University

Scientists Teaching Science ·
Barbara Houtz, STEM Education Solutions
Pedagogy course for science educators

EDUCATION

University of Cambridge
PhD Neuroscience 2018

Emory University
BS/MS Biology 2011

EXPERIENCE

American Express — New York, NY
Senior Machine Learning Engineer — Aug. 2018 to Current

I perform the role of a "true" machine learning engineer: Data to Deployment. General workflow: SQL/Hive queries against an internal data lake > modeling in Python > deployment as a Dockerized intranet microservice to a kubernetes cluster.
Our small group is uniquely tasked with democratizing data science across the enterprise by creating reusable machine learning microservices.
Outside of contributing to the model deployment framework proper, I've developed and deployed models for forecasting and anomaly detection as well as personalization for customer-facing websites for multiple teams spanning multiple organizations/departments within the enterprise. I create containerized machine learning models and deploy (CI/CD: Jenkins and XL Release) them for production use as RESTful APIs using Docker and kubernetes. In other cases, I have helped many different groups take their siloed machine learning models and deploy them into production after reviewing their models. I was the first hire for the team and I've helped to train and onboard new members with the missing skills not typically gained from a traditional quantitative background such as Docker, CI/CD, and cloud deployment of containerized models.
Also ran (and judged) an internal data science competition modeled after Kaggle competitions along with one other data scientist to increase ML awareness and adoption within American Express using data for a real use case within the company that is now implemented as a production service.

SharpestMinds — New York, NY
Data Science and Machine Learning Mentor — Mar. 2019 to Current

At YC-backed SharpestMinds, we help strong candidates get past the finish line. I help build their real-world ML/DS skills and land jobs as data scientists and machine learning engineers. I work one-on-one with the mentees, guiding them through at least one real-world ML project, introductions, interview prep, negotiations, and help rebrand themselves to be able to transition to industry more smoothly.

Metis (Kaplan) — New York, NY
Data science teaching assistant — July 2018 to Aug. 2018

Support senior data scientists in teaching a twelve-week data science bootcamp.
Assist students with issues relating to project design, web scraping, machine learning, data visualization and interpretation of results.
Perform code reviews and give personalized feedback for students to improve understanding, technique, and efficiency.

New York, NY
Data scientist — Apr. 2018 to June 2018

Full-time ACCET-accredited immersive data science bootcamp including full-cycle projects from inception to execution, including data acquisition, modeling, and communicating results. Machine learning and statistical modeling methods: linear regression, supervised learning and classification, unsupervised learning and clustering, natural language processing, deep learning.

Completed several end-to-end data science projects working primarily in Python (code for each project is available on my GitHub):
Python libraries/frameworks used consistently for all projects: numpy, pandas, SciPy, scikit-learn, matplotlib, seaborn, Jekyll.

1) ***Linear regression of web-scraped Steam video game data***: Predicting number of users and determining the most important predictive features.
Web-scraping using Beautiful Soup and Selenium to extract values for every game (~23,000 games) in the Steam store. Modeling using scikit-learn and linear regression with Lasso and Ridge regularization. Determined important features for predicting game popularity from coefficients.

2) ***Predicting early hospital readmissions among diabetic patients***: Evaluated many different supervised machine learning models and created a cost function to optimize a random forest predictor implemented as a web app.
Used scikit-learn to compare the performance of models including KNN, SVM, Decision tree, Boosted trees, Random Forest, and Naive Bayes. Modified hyper-parameters to balance precision and recall. Created a cost function to determine a threshold value that minimizes the number of early patient readmissions before incurring additional insurance fees. Used Flask to make a predictor app deployed to AWS.

3) ***Project Guten-bag-of-words***: Automatic genre grouping, similar book recommendation, and visualization of narrative patterns in Project Gutenberg books using unsupervised machine learning and natural language processing (NLP).
Used topic modeling on all sci-fi and fantasy books (~2000 books) to cluster books into sub-genres and visualized narrative patterns within each book using sentiment analysis and topic modeling. Tools used: textblob, NLTK, CountVectorizer, TF-IDF, NMF, LDA, AWS, dash.
Used plotly and dash to make a web app deployed to AWS: http://flask-env.svfvjqygqf.us-east-1.elasticbeanstalk.com/

4) ***Radiology-3D-CNN***: Used convolutional neural networks to locate and classify brain tumors, and predict length of patient survival from 3D MRI images of brain tumors (high- and low-grade gliomas).
Tools used: AWS, Keras, TensorFlow.

Howard Hughes Medical Institute — Ashburn, VA
Research Specialist — 2017 to 2018

Worked on a team to support multiple labs in various capacities to meet their specific needs. * Aided in decision-making for experimental design. * Performed data analysis for high-throughput behavioral assays and RNA-Sequencing experiments. * Updated and revived deprecated analysis pipeline code. * Created documentation for behavioral assays and analysis pipelines. * Created new analysis pipelines for genomic data. * Trained new junior members of the team (scientific research technicians). * Programming languages used: bash, R, Matlab.
(Supervisor: Gudrun Ihrke, PhD)

Cambridge, UK & Ashburn, VA, USA
PhD Graduate Student — 2011 to 2018

Research question: How are diverse neurons generated from similar stems cells?
Performed RNA-Seq analysis (using bash, SQLite, and R) on self-generated data for 64 neuronal stem cell lineage samples to determine transcriptional identity of neuronal stem cell lineages in nerve cord of fruit fly (*Drosophila*, an organism with ~16000 genes). Visualized data with principal component analysis and heatmaps with hierarchical clustering. Performed genomic alignments with command line software using a on-premises compute cluster. Used multiple analysis pipelines to determine the concordance in statistically significant differential gene expression.
Joint PhD program with funding from HHMI (University of Cambridge Advisor: Andrea Brand, PhD; HHMI Advisor: Jim Truman, PhD)