BENJAMIN BATORSKY

Cambridge, MA
github.com/bpben
benbatorsky@gmail.com
benbatorsky.com
www.linkedin.com/in/bbatorsky

SKILLS

- Machine Learning (Ensemble and linear models, clustering, imbalanced class analysis)
- Neural Network architectures (RNN, CNN, Encoder-decoder)
- Natural Language Processing (LLMs, Transformers, Embeddings)
- Causal Inference (mixed-effect models, instrumental variables, simulations)
- Probabilistic modeling

TOOLS

- Python (scikit-learn, SciPy stack, statsmodels, gensim, spaCy, PyTorch, PyMC, TensorFlow, transformers)
- Interactive web applications (Flask, FastAPI, Streamlit)
- Cloud Architecture (AWS, GCP, Azure)
- R (tm, psych, cluster, lmer, ggplot)
- SQL (MySQL, MSSQL, PostgreSQL, SparkSQL)
- QGIS, ArcGIS

RELEVANT EXPERIENCE

Data Science Lead (03/2023 - present)
Senior Data Scientist (02/2022 - 03/2023)
Institute for Experiential AI at Northeastern University, Boston, MA

- Developing technical consulting proposals in collaboration with researchers, business development and external stakeholders for industry clients
- Designing demand forecasting pipelines for hundreds of products, reducing prediction error by 30%-50% and enabling more accurate production targets
- Designing pricing models and integrating results into a data dashboard for improving sales processes
- Design of Responsible AI (RAI) technical evaluation strategy for the Institute and delivery of RAI products to external clients in telecommunications and insurance
- Building and managing a six-person technical team of data scientists, engineers and project managers

Biomedical Data Scientist (11/2020 - 02/2022)
Ciox Health Real World Data, Cambridge, MA

- Designing/implementing entity (e.g. medication and diagnosis) recognition and linking systems for electronic health records with performance comparable to state-of-the-art systems (e.g. Amazon Comprehend Medical)
- Leading development of a graph-based medical concept representation system for recognition model improvement and high-level classification (e.g. RxNORM parent class)
- “Population builder” application enabling large-scale search of medical records
- Managing clinical expert annotation initiatives and ensuring consistency with inter-rater reliability at near 100%

Contract Data Scientist (08/2020 - 11/2020)
bPrescient, Cambridge, MA

- Unsupervised learning on large-scale medical record data to extract patterns in patient populations to direct exploration of genetic data
- Leveraging structured medical taxonomies (UMLS) to create informative representations of billing codes recorded in patient narratives

Associate Director, Data Science (06/2019 - 06/2020)
MIT Sloan Food Supply Analysis and Sensing Group, Cambridge, MA

- Designing and implementing a robust data pipeline for parallel processing of structured and unstructured data, allowing researchers near-live access to food supply chain data
- Design and implementation of a named-entity recognition model to identify products and agencies, allowing for greater flexibility than previous inventory-based methods
- Collaborating with researchers to understand user needs and develop product requirements
- Managing a multi-site international team of engineers to build robust data pipelines

Data Scientist (05/2017 - 06/2019)
ThriveHive, Boston, MA

- Developing models to identify customers at different points in their lifecycle (i.e. conversion, upsell, churn), enhancing sales and marketing initiatives
- Developing a pipeline for segmenting business customers based on text data, social media categories and local population characteristics in order to improve performance of predictive models and personalize content delivery
- Using daily results data from hundreds of thousands of marketing campaigns (e.g. Adwords, Facebook) to predict outcomes based on customer and campaign characteristics, empowering sales targeting and marketing messaging
- Developing and managing the Data Science team
- Delivering quarterly business reports on outcomes and planned products
- Collaborating with engineering team for deployment of interactive apps and model APIs in the AWS production environment

Data Science Fellow (06/2015-05/2017)
Department of Innovation and Technology, City of Boston, Boston, MA

- Combining social media data (e.g. Twitter, Craigslist, Yelp) and city data to predict city code violation risk and more efficiently utilize inspector resources
- Developing partnerships with city stakeholders to integrate data insights into operations (e.g. enforcement, prioritization)
- Developing routing algorithm (TSP solver) for inspectors to maximize coverage of high-risk areas
- Identifying authorship based on syntactic similarity using clustering methodologies
- Development of interactive labor force estimation tool (RShiny)

Assistant Policy Analyst (09/2012-12/2016)
Pardee RAND Graduate School, RAND Corporation, Santa Monica, CA

- Performing parametric and non-parametric modelling to answer research questions from a variety of subject areas (e.g. Health, Infrastructure)
- Managed Twitter Research team, which aids researchers in accessing and analysing large Twitter corpora
- Developed interactive web application for analyzing workforce data

EDUCATION

Ph.D. in Policy Analysis (2016) Pardee Rand Graduate School, Santa Monica, CA
Dissertation topic: Effective design and evaluation of workplace wellness programs

MPH (2012) Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Capstone Project: *Focus on Dysfunction: An examination of barriers to receiving mental health care among OEF/OIF veterans*

BA English and History (2007) Rutgers University, New Brunswick, NJ
Honors Thesis: *Malefici*: A Study of Medieval Belief and Superstition

VOLUNTEER WORK AND PRESENTATIONS

Bagging to BERT: A tour of applied Natural Language Processing (November 2022): Workshop on applied NLP given at ODSC West.
(https://odsc.com/speakers/bagging-to-bert-a-tour-of-applied-nlp/)

New methods, old problems (October 2021): Talk on ethics and bias in Natural Language Processing at the Spark NLP Summit.
(https://www.nlpsummit.org/new-methods-old-problems-ethics-and-bias-in-modern-natural-language-processing/)

Named-entity Recognition from Scratch (March 2020)**:** Talk on training a custom neural named-entity recognition model on a large non-English text corpus.
(https://conferences.oreilly.com/strata-data-ai/stai-ca/public/schedule/detail/80050)

Vision Zero Crash Modelling Project (2017 - Present): Lead on volunteer project for Data for Democracy focused on providing cities with tools for assessing vehicle crash risk based on available data.
(https://github.com/Data4Democracy/boston-crash-modeling)

Computing Customer Similarity with Text Data (February 2019): Workshop on Natural Language Processing-based approach to computing similarity between customer websites
(https://datascience.salon/austin/)

Estimating Customer Budget with Hierarchical Probabilistic Models (September 2018): Conference talk on using probabilistic models to estimate unknown customer budgets
(https://odsc.com/training/portfolio/estimating-customer-budget-with-hierarchical-probabilistic-models)

Building Data Science Teams (February 2018): Discussion of the considerations of building Data Science teams and pipelines (https://generalassemb.ly/education/hiring-building-data-science-teams)