TODD WARCZAK, Ph.D.

Data Scientist, Bioinformatician

CONTACT

✉ twarczak@gmail.com

☎ (206) 999-6478

📍 Pleasanton, CA

🌐 toddwarczak.netlify.app

LinkedIn

Github

EDUCATION

Ph.D.
Molecular & Cellular Biology, Dartmouth College

B.S.
Biology, University of Utah

SKILLS

R ★★★★★

Python ★★★

SQL ★★★★

RShiny Apps & Dashboards

Bioinformatics, Genomics

Github, Jira, Agile, Scrum

ETL Process, Linux, Bash, CLI

Quarto, Markdown, Jupyter

AWS, EC2, S3, SageMaker

Docker

Snowflake

Tidyverse, Pandas, NumPy

Data Visualization

AI, Machine Learning

Tidymodels, scikit-learn

TensorFlow, Keras

Time-Series Forecasting

Probability, Statistics

Regression, Classification

Natural Language Processing

Data Management

Analytical Thinking

Database Administration

Clinical Research

WORK EXPERIENCE

Data Scientist II

Bio-Rad Laboratories, January 2022 - March 2023 / Pleasanton, CA

- Developed “ddTrackR”, a highly modularized R-Shiny application for team of scientists to upload experimental ddPCR results to database on AWS cloud, track metadata from clinical samples, build custom ggplots, and perform statistical calculations in browser, without code.
- Built ddPCR and NGS data analysis pipelines for 10+ scientists. Markdown reports delivered and/or Shiny dashboards hosted for team to visualize/explore results.
- Scrum master for team of 6 software developers building Bio-Rad back-end software in Python and Java for the new ddPCR QX600 machine. Utilized Jira Software for sprints.
- Collaborated w/ sales, marketing, and software teams to develop the front-end, customer-facing data analysis tools for a Non-Invasive Prenatal Test (NIPT) kit on the QX600.
- Hosted weekly R & Python data science workshops for 15+ scientists. Topics included RStudio/ VS Code setup, data wrangling w/ base-R/Tidyverse/Pandas/NumPy, statistics, mastering ggplots, exploratory data analysis, custom & dynamic reactable tables, and ddPCR/NGS analysis.

Molecular Biologist

Dartmouth College, September 2012 - September 2020 / Hanover, NH

- Engineered novel genome-wide association study (GWAS) that identified genes controlling arsenic tolerance in plant roots. Utilized expert data cleaning, wrangling, and analytic skills to summarize findings from millions of observations across thousands of unique genomes.
- Determined plant gene AtNIP1;1 is the major genetic factor for tolerating arsenic in root cells.
- Built lab RNA-seq pipeline for 25000+ plant genes and wrote R scripts for gene clustering (PCA, hierarchical clustering), regression (GLMs/ANOVA), and exploratory data analysis.

SIDE PROJECTS OF NOTE

SageMaker + RStudio to Predict Home Prices w/ Multi-class XGBoost; Explaining Model Behavior with Geospatial Plots and SHAP

- Explored Austin dataset to predict home price in Kaggle competition. (Blog, Github)
- Built static and interactive geospatial plots overlayed with feature data.
- Feature engineered high/low important words that associate w/ price using NLP.
- Trained/tuned/evaluated/deployed SageMaker Multi-class XGBoost on holdout data & submitted predicted binned price to Kaggle competition. Submission scored 0.8876 (mLogLoss), which would have placed 6th (out of 90 entries) in live competition.
- Modified {SHAPforxgboost} package to generate multi-class SHapley Additive exPlanations (SHAP) values/plots that explain how XGBoost model made predictions.

Forecasting Daily Sales with {modeltime}

- Performed EDA and generated 3 month forecasts of daily sales for Kaggle dataset selling furniture, technology, and office supplies. (Blog, Github)
- Tested multiple {tidymodels} workflows with {modeltime} for time-series forecasting.
- Best individual models combined to generate single weighted ensemble forecast.