Jacob Bayer

LinkedIn • GitHub • PyPi
(914) 268-7131
jacobbenjaminbayer@gmail.com
Jersey City, NJ

Education

MS – Statistics | GPA: 3.71/4 | January 2022
City University of New York, Baruch College

BS – Economics | GPA: 3.87/4 (Summa cum Laude) | May 2020
State University of New York (SUNY) at New Paltz

Professional Experience

Data Science/Engineering Contractor at *Govini* • *Remote* **June 2022 – December 2022**

Govini is a software and consulting company serving government clients in the defense sector.

- Built 5 AWS Glue jobs to extract, transform, & load (ETL) large and complex data from Govini's data warehouse to an application database.
- Designed target database schemas to optimize storage space, API read efficiency, and ease-of-use.
- Collaborated with front-end, API, and data science teams to ensure output data meets product requirements.
- Orchestrated migrations and ETL jobs to synchronize with ingestion of new data into the data warehouse.

Data Analyst at *Phosphorus* • *New York, NY* **August 2020 – May 2022**

Phosphorus was a healthcare startup seeking to develop a vertically integrated genetic test.

- Created scripts, visualizations, and dashboards using Python/Pandas and SQL to measure production, forecast future production bottlenecks, make staffing decisions, and communicate with clients.
  - Presented insights based on these data to executives every week and made recommendations to improve operations, which resulted in a reduction in sample turnaround time from 80 days to 20 days over the course of 6 months while sample volume doubled in the same period.
  - Developed and maintained a web-based data visualization application using Plotly Dash, a Python web framework built on Flask, which was used by the laboratory manager, COO, and VP of Laboratory Operations on a daily basis.
- Developed a feature for our Ruby-on-Rails web application to allow superusers to view all observed mutations for a specific haplotype, disease, or disease group.
- Created Slack applications (a.k.a. bots) to provide updates when time-sensitive problems require action, eliminating the time that samples spend waiting for intervention, reducing turnaround time for those samples by several days.

Statistics Teaching Assistant at *SUNY New Paltz* • *New Paltz, NY* **August 2019 – May 2020**

- Taught statistics review lectures 4 times per semester with up to 40 students at a time.
- Tutored students in R programming for 10 hours per week, instructing groups of 1-6 students.
- Led transition to remote learning in response to COVID-19 by creating a series of 5 video lectures to explain statistical concepts as well as R programming.

Projects

Sunbelt **December 2022 – Present**

- Designed and built Sunbelt, an application that mines data from Reddit and makes it accessible via a GraphQL API.
- Sunbelt stores information about how posts, comments, users, and communities (subreddits) have changed over time, unlike other services that only provide point-in-time information about Reddit.
- These data can be accessed using the Sunbelt API Wrapper for Python (SAWP), which is available on PyPi.

Sentiment Analysis of Drug Reviews using NLTK in Python **December 2021**

- Performed data cleaning on a dataset of 10000 drug reviews from Drugs.com to prepare it for feature generation.
- Generated features from text using bag of words, word2vec, and TF-IDF.
- Evaluated performance of logistic regression, random forest, gradient boosting, and naïve bayes to determine the best model fit by comparing out-of-sample performance.
- Performed hyperparameter tuning to optimize performance.
- Selected the best performing model, logistic regression, to achieve 93% recall, 79% accuracy, 79% precision, 83% AUC, in 78.5 seconds of training time using the optimal hyperparameters.

Software Languages and Skills

Python (2 years), Pandas (2 years), R (2 years), SQL (2 years), Plotly Dash (1 year), PySpark, NumPy, Ruby-on-Rails, Bash, Git, Jira, Agile, Data Visualization, Machine Learning, Natural Language Processing, Data Analysis

Coursework

Machine Learning for Data Mining, Data Mining for Business Analytics, Foundations of Statistical Inference, Applied Probability, Applied Natural Language Processing, Multivariate Statistical Methods