ANGELINE PROTACIO

DATA SCIENTIST

email@angelineprotacio.com www.angelineprotacio.com (917) 408-3266 New York, NY
dataangeline https://www.linkedin.com/in/angeline-protacio angelinepro

SUMMARY

Leverages 7+ years of data analysis and research experience to translate insights into business outcomes. Adept at facilitating data communication and reproducibility, and streamlining data science processes.

EMPLOYMENT

Metis · New York, NY — March 2020 to June 2020
Data Scientist

- Completed an accredited full-time immersive data science bootcamp that provided training in machine learning, statistics, Python programming, natural language processing, data visualization, and database systems, through project-oriented learning. Selected work includes:
- Create The Album Discoverer, an album recommendation system that leverages Spotify track data and employs principal components analysis and aggregation at the album level with cosine similarity to identify similar albums, presented as a user-friendly interactive Flask web app
- Explore differences in communication between New York State and City leadership during Covid-19, using web scraping to obtain speech transcripts for natural language processing including parts of speech analysis, sentiment analysis, and non-negative matrix factorization for topic modeling to understand which topics are more frequently addressed, which populations they serve, and how their speech sentiments vary in time with the pandemic's progression
- Build a model to classify flight arrival delays using six million rows of flights data, leveraging Google Cloud Platform to handle imbalanced classes and run multiple machine learning models including logistic regression, random forest, and gradient boosted trees in order to optimize recall and identify the features most important for predicting flight arrival delays
- Predict baseball pitcher quality starts using web scraping to obtain prior season and future projection baseball data, employing linear regression to predict future quality starts and lasso regularization for feature selection, with further exploration to identify and understand cases where the model performed well, and where it failed to predict well

Protacio Analytics, LLC · New York, NY — November 2018 to Present
Data Scientist

- Operationalize client-defined metrics to detect ongoing fraud in real-time administrative financial data
- Test and troubleshoot R packages and Shiny dashboards for bioinformatics to ensure data visualizations satisfy client specifications
- Collaborate on data science projects and Shiny application development with an international and fully remote team using GitHub and collaborative project management software
- Provide clients with clear, reproducible R Markdown reports demonstrating analysis logic and data visualizations to aid communication
- Provide technical aid and supporting reference documentation to facilitate data scientists and analysts learning R
- Present and communicate analytic findings to data science community at meetups and conferences

New York City Department of Health and Mental Hygiene · New York, NY — October 2015 to October 2018
Surveillance Data Lead, Research Data Analyst

- Lead the analysis of mental health surveillance and survey data to assess suicidal behavior, prevalence of mental illness and mental health service use, and needs after psychiatric hospitalization
- Create, implement, and manage surveillance protocols for detecting and monitoring mental health syndromes in New York City emergency departments to facilitate real-time response to public mental health crises
- Supervise members of research team to guide data analyses from exploration to dissemination of findings
- Develop, organize, and teach R programming and data visualization workshops for agency data analysts
- Collaborate across bureaus, divisions, and agencies to promote data-driven decision making at all stages of data analysis
- Communicate research findings through conference talks, posters, and publications

Columbia University Mailman School of Public Health · New York, NY — July 2013 to July 2015
Data Analyst

- Perform statistical analyses to investigate relationships between lifecourse and socioeconomic factors, and intermediate markers of breast cancer risk in cross sectional and longitudinal studies
- Interpret and present results of analyses to colleagues and collaborators
- Manage data, including integrating data from varied sources ,data cleaning, dataset distribution, and variable creation
- Collaborate with research teams at multiple sites to implement changes to data processing and distribution
- Supervise members of research team to guide data collection, cleaning, and analysis

PRESENTED WORK

"Using R and the Tidyverse to Play Fantasy Baseball" — July 2019
useR! Lightning Talk

"Using R and the Tidyverse to Play Fantasy Baseball" — May 2019
R-Ladies New York City Meetup Oral Presentation

"New York City Mental Health Syndromic Surveillance System" — September 2018
New York State Suicide Prevention Conference Oral Presentation

"Suicides in New York City, 2000 to 2014" — September 2017
New York State Suicide Prevention Conference Poster Presentation

"Timing of Female Reproductive Factors and General and Abdominal Obesity in Midlife" — January 2014
New York City Epidemiology Forum Poster Presentation

SKILLS

PROGRAMMING

Python (Pandas, scikit-learn, matplotlib, seaborn, spaCy, VADER, CorEx, Flask, Beautiful Soup, Selenium)

R (Tidyverse, Shiny, R Markdown, package development)

SQL

SAS (Base)

STATA

SUDAAN

Bash

STATISTICAL AND MACHINE LEARNING MODELS

Natural Language Processing

Dimensionality Reduction (Non-Negative Matrix Factorization, Principal Components Analysis)

Clustering Algorithms

Random Forest and Gradient Boosted Trees

Logistic Regression (Binary and Multinomial)

Linear Regression with Regularization

Cox Proportional Hazards Regression

Generalized Estimating Equations

LANGUAGES

English - Native

French - Conversational (DELF B2 Diploma 2019)

Tagalog - Conversational

Japanese - Basic

Korean - Basic

OTHER

Git

Google Cloud Platform

ArcGIS

EXPERIENCE

Columbia University - Mailman School of Public Health
2013
MPH Epidemiology

University of California Berkeley - College of Natural Resources
2008
BS Genetics and Plant Biology