Joseph Serigano, Ph.D.
Data Scientist

New York, NY | 1 (631) 943 - 8573 | jserigano4@gmail.com
LinkedIn | Github | Portfolio

About Me

Data scientist with 8+ years of data analytics experience. I use my experience with large data sets in combination with Python, SQL, MATLAB, and various machine learning techniques to explore data, understand trends, predict future outcomes, and provide tangible results.

Technical Skills: Python | Matlab | SQL | Tableau | AWS | Jupyter Notebooks | Github | Excel | Word | PowerPoint
Modeling Techniques: Web Scraping, Data Cleaning, Supervised and Unsupervised Machine Learning (ML), Predictive Modeling, Data Visualization, Data Analysis, Natural Language Processing, Big Data
Python libraries: Pandas | NumPy | SciPy | Matplotlib | Seaborn | Scikit-learn | Natural Language Toolkit| Beautiful Soup | Tweepy | Praw | Keras | TensorFlow

Data Projects

Predicting New York Times Cooking Recipe Popularity

- Built multiple classification models to predict whether a NYT Cooking recipe will create higher than average site traffic based on several recipe features.
- Scraped, cleaned, and processed data from 10,000 NYT Cooking recipes for analysis.
- Developed a model to determine whether or not a recipe is considered highly trafficked with an accuracy of 0.79.
- Determined characteristics of a recipe that are most likely to increase the overall popularity, and thus lead to higher site traffic and provided recommendations on how to modify a recipe to be more popular before posting.

Combatting Air Pollution and Inequality in the US

- Created multiple regression models to predict ground PM2.5 air quality measurements at current EPA monitoring sites using publicly available satellite data.
- Developed a model that was able to account for 71% of the variability in our test data.
- Determined regions in the US that suffer from poor air quality and recommenedd locations in the US that could benefit from hyperlocal air quality monitoring.

NLP Reddit Classification

- Collected 20,000 Reddit posts from two subreddits using Pushshift's Reddit API to predict from which subreddit a given post originated.
- Performed data cleaning and feature engineering on Reddit posts to prepare for modeling.
- Built multiple natural language processing (NLP) classification models with hyperparameter fine-tuning to predict the origin of Reddit posts.
- Developed a model that accurately predicted the subreddit classification of 75.7% of posts based on input features.

Regression Model to Predict Home Sale Prices in Ames, IA

- Explored a Kaggle dataset of home sales from 2006 to 2010 provided by the Ames' Assessor's office.
- Built a regression model using LASSO and Ridge regularization to determine what characteristics of a home are most likely to increase the sale price and what changes can be made to increase sale price before putting a home on the market.
- Developed a model that was able to predict housing prices with an R^2 score of 0.94.

Experience

Johns Hopkins University - Postdoctoral Research Scientist — JUL 2021 - MAY 2022

- Collaborated with scientists to further develop MLR model from Ph.D. thesis for use on new data sets, including laboratory experiments and other planetary atmospheres, leading to three peer reviewed papers.
- Redeveloped material for introductory planetary science course, leading to a more accessible and interactive course and increasing student enrollment by ~110% over two semesters.
- Facilitated writing of NASA grants to fund research and development, resulting in two successful grants totaling $610,502 over three years.

Johns Hopkins University - Graduate Research Scientist — AUG 2015 - JUL 2021

- Spearheaded research using NASA data from the Cassini spacecraft to investigate the compositional interactions between Saturn's inner rings and upper atmosphere, resulting in seven peer reviewed publications.
- Performed extensive cleaning and feature engineering of raw spacecraft data to produce a data set fit for analysis.

- Developed a novel multiple linear regression model to analyze mass spectral data, allowing for a more in-depth statistical analysis of unit resolution spaceflight mass spectrometry than previously existed.

NASA Goddard Spaceflight Center - Research Scientist JUN 2014 - AUG 2015

- Developed new techniques to utilize low-signal data from the Atacama Large Millimeter Array (ALMA) for cutting-edge research on Titan's atmosphere, leading to multiple discoveries of new chemical species in Titan's atmosphere through atmospheric modeling.
- Trained scientists on new techniques in order to establish ongoing collaborations with multiple NASA facilities and universities, resulting in eight peer-reviewed publications and counting.

Education

General Assembly MAY 2022 - AUG 2022
Data Science Immersive

Johns Hopkins University AUG 2015 - JUL 2021
Ph.D. and M.A., Earth and Planetary Sciences
Dissertation: Compositional Measurements of Saturn's Upper Atmosphere and Rings from Cassini Ion and Neutral Mass Spectrometer

Boston University AUG 2010 - MAY 2014
B.A., Astronomy and Physics