

# Matthew Dalton

## Data Scientist



(210)-219-4852



matthew.w.dalton@gmail.com



/matthew-dalton-09324794/



/mdalton87

### Professional Summary

I am a proud Army Veteran with a diverse background in medicine, engineering, and aviation. My advanced degree in biomedical engineering and love for problem-solving has guided me to see the truth in data. I am fascinated by the endless possibilities data science can make our lives better.

### Technical Skills

Python - Pandas - Matplotlib - Git - Applied Statistics - Seaborn - Machine Learning - SQL - Scikit-learn SciPy Statsmodels - Beautiful Soup - NLP - Apache Spark

### Military Experience

#### Clearance:

Inactive - Secret

Oct 2006 - Oct 2016

#### Clinical Laboratory Scientist

U.S. Army

Jul 2010 - Oct 2014

Responsible for managing night shift operations within the hospital lab. Maintained \$1.8M of laboratory equipment with no losses or damages. Trained incoming personnel for certification testing with an 80% pass rate while also managing day-to-day operations.

#### Drone Pilot

U.S. Army.

Oct 2006 - Jul 2010

Responsible for providing over 800 hours of vigilant aerial coverage in support of Operation Enduring Freedom in Afghanistan. Maintained and utilized advanced communications to direct and coordinate the movement of aircraft and ground personnel.

### Education

#### The University of Texas at San Antonio

Fall 2018 - Spring 2020

Master's of Science in Biomedical Engineering  
GPA: 3.55 / 4.00

Fall 2014 - Spring 2018

Bachelor's of Science in Biomedical Engineering  
GPA: 3.25 / 4.00

#### Codeup

Dec 2020 - Jun 2021

Certificate of Completion

Fully immersive, project-based 22-week career accelerator that provides students with 670+ hours of expert instruction in applied data science. Students develop expertise across the full data science pipeline (planning, acquisition, preparation, exploration, modeling, delivery), and become comfortable working with real, messy data to deliver actionable insights to diverse stakeholders.

### Data Science Projects

#### America's Blues: An Analysis of Fatal Police Encounters

*Classification Modeling*

Our four-member capstone team acquired a data set from MappingPoliceViolence.org, which has gathered over 9000 police killings from 2013 to May 2021. We collaborated to clean and prep the Excel spreadsheet. After developing our initial hypotheses, we will use Pandas, Seaborn, Matplotlib, and Tableau to explore our ideas. Top drivers of civilian fatalities identified will shed light on changes to be made to save more lives. We then used Scikit-Learn on multiple encoded features to apply to our classification models predicting whether the victim was indeed an attacker. Our final predictive model had an accuracy percent improvement of 39%

#### Predicting Tax Value of Real-estate:

*Regression Modeling*

Used regression modeling to target the tax value of properties in the Zillow data set. I used a SQL query to bring the data from MySQL into JupyterLabs. There I used pandas, seaborn, and Matplotlib to explore the data set for drivers of tax value. I ran statistical tests: correlations and t-tests, to verify the features are in fact drivers. I then used Scikit-learn to create a Quadratic Regression model that reduced the RMSE by nearly 22%.

#### Predicting Microsoft's Programming Language of GitHub Repos Using the README File.

*Natural Language Processing with Classification Modeling*

I acquired 2200 readme files and the repository's primary coding language from Microsoft's organizational GitHub using requests and BeautifulSoup. After I normalized, tokenized, lemmatized, and removed the stopwords from the readme text, I removed all null values leaving me with 1800 records. I established the baseline of 28% accuracy by selecting C# as the base prediction and created an SGD Classifier model that more than doubled the baseline accuracy at 58%.

### Personal Side Projects

#### Predicting Stroke:

*Classification Modeling*

The goal of this project was for me to learn and add new tools for approaching imbalanced datasets. This data contained only 5% of patients with a stroke. I used the imblearn library to rebalance the dataset allowing me to create a KNN model with a recall of 29%.