Johnny Yu

Johnnyyu14@gmail.com | 415-312-2772 | johnnyyu14 | johnnyyu14.github.io

Education

University of California Santa Barbara Santa Barbara, CA

Bachelor of Science in Statistics and Data Science *September 2018 – June 2022*

Major GPA: 3.64/4.00

Awards: Dean's Honors L&S (Spring 2020, Winter 2021, Spring 2021, Summer 2021, Winter 2022, Spring 2022), Awarded $5000 to participate in UCSB's NSF-funded Central Coast Data Science Fellowship for the 2021-2022 academic year

Technical Skills

Programming Languages: Python, SQL, R, SAS, HTML, CSS

Visualizations/Libraries/Analysis: Tableau, Power BI, matplotlib, ggplot2, pyTorch, TensorFlow, NumPy, pandas, tidyverse, scikit-learn

Experience

ENVENT Labs Santa Barbara, CA

Data Analyst *September 2021 – June 2022*

- Collaborated under the supervision of Dr. Matto Mildenberger to explore how demographic attributes (age, income, race, religion) are differentially predictive of climate opinion in different countries, continents, and time periods
- Analyzed and cleaned covariate data from 250+ surveys, merged demographic information into existing data processing routines
- Developed complex supervised and unsupervised ML models using SVM and PCA to make predictions of an individual's climate score based on their demographic info
- Presented the results of analyses to the National Science Foundation to promote Data Science education at California State University and the University of California

Projects

Forecasting Short-Term Future COVID-19 Cases *September 2021 - December 2021*

- Created two time-series forecasting models that predicts the number of new daily COVID cases trained on the estimated percentage of COVID related outpatient doctor visits
- Identified Decision Tree Regressor performed 52% better than SVR model with the accuracy score being able to capture the overall trend in the rise and fall of new cases
- Evaluated that neither model accurately predicted the number of new daily COVID cases using historical data due to wide range of values of the dataset and the features not being good predictors of new COVID cases

2016 Presidential Election Prediction Model *January 2021 - June 2021*

- Explored various models to answer whether or not Donald Trump would win a county based on demographic variables
- Examined that Random Forest had a 0.8% higher accuracy score compared to Logistic Regression. Using the optimal threshold for Logistic Regression, the true negative rate increased but total misclassification rate increased to 19%
- Detected that Trump won around 87% of the counties in the test data. Which implies if the misclassification rate in a model was above 13%, there is no improvement using a model to predict election results

UFO Sightings 1914-2010 reported by US Gov. *March 2021 - June 2021*

- Examined spatial density of sightings before and after 2000 to identify diurnal differences in UFO sightings throughout the years
- Explored the pairwise relationships between all variables and developed a correlation matrix to assess the relationship between the quantitative variables in the data set
- Utilized KDE curves, histograms, and scatter plots to visualize the density changes specific to latitude or to longitude