Coding the Climate:

Modeling the temporal relationship between Greenhouse Gases and Sea Ice

Abstract: The objective of this project was to test the feasibility of modeling sea ice extent and greenhouse gases. The more specific goal was to select a regression model best suited for modeling a potential relationship between publicly available greenhouse gas data and the complete loss of our Arctic sea ice. The data consisted of 381 rows spanning the dates October 1986 to October 2018. The features used in the final model are carbon dioxide, methane, and nitrous oxide. A simple linear regression model proved unsatisfactory with a R^2 score of -3.4. Cross validation methods resulted in the best model to pursue being a second degree polynomial regression. By applying a polynomial degree to both nitrous oxide and carbon dioxide the R^2 score increased to 0.034. The model struggled to account for the cycliclic/seasonal ice extent data. Based on the baseline evaluation and modeling I recommend a Seasonal Auto Regressive Integrated Moving Average (SARIMA) model for further use of these datasets.

Design: The data for this climate change project is publicly available through the National Ocean and Atmospheric Administration (NOAA), National Center for Atmospheric Research, Massachusetts Institute of Technology (MIT), and the National Snow and Ice Data Center. The dataset included nitrous oxide, carbon dioxide, and methane, in the metric of parts per billion averaged over each month between 1986 and 2018. Model training data used was between 1986 and 2020. Test data consisted of time series data between 2010 and 2018. October 26, 2018. Sea ice extent data is in units of millions of square kilometers. El Nino and La Nina climate data was scraped using BeautifulSoup4 and used as categorical variables in the model.

Data & Algorithms: The data consisted of 381 rows spanning the dates October 1986 to October 2018. Original data included 8 greenhouse gases but were reduced to 3 after baseline linear regression modeling. The features used in the final model are carbon dioxide, methane, and nitrous oxide. 301 rows were selected for the training data and 80 for test data. Cross validation was used to determine the direction to go for feature engineering of a model.

Tools: Tools used for this project included:

1. BeautifulSoup4 was used to scrape climate data based on year to input into the model.
2. Ingesting raw data into a SQL database via web scraping
    a. SQL was used to clean and remove duplicate data
3. Python Pandas was used to merge categorical dummies, greenhouse gases, and
4. Seaborn was used to map the pair plots of the features.
5. SKLearne was used to build regression models