Dillon Dugan

COMP 4448 Data Science Tools 2

Final Project (Written)

**Data Source and Variables**

The bomregion dataset from the Data Analysis and Graphics Data and Functions contains information about Australian regional temperature, regional rainfall, Southern Oscillation Index, CO2, and average annual sunspot count from 1900 to 2012. The regional temperature and rainfall data are area-weighted averages for the given regions: East, Southeast, South, Southwest, West, and North Australia, as well as data from the Murray-Darling Basin and average for all regions of Australia. The Southern Oscillation Index is the difference between barometric pressure at sea level in French Polynesia and Darwin, North Territory, Australia. Both the temperature and rain data as well as the SOI data was observed by the Australian Bureau of Meteorology. CO2 concentrations are collected from the Global Monitoring Laboratory in Mauna Loa, Hawaii. Sunspot observation comes from the Solar Influences Data analysis Center. All these separate datasets were combined by the Data Analysis and Graphics Data and Functions to create a comprehensive data source of climate data in Australia over the 20th and 21st century. Being from 1900 to 2012, there is 113 rows of data with 22 columns of features from the combined datasets. However, the regional temperature data has missing information from 1900 to 1909.

For the use of this project, the model will be built with the average temperature and rainfall for all of Australia and complete CO2 series bringing the total features from 22 to 6: Year, Rain, SOI, CO2, sunspot, and Temperature. This allows the model to be build more simply and Rain/Temperature can be swapped out with one of the regional data to analysis the severity of risk by territories. The input features are Year, Rain, SOI, CO2, and sunspot which will predict the outcome of Temperature. Looking at scatterplots with fitted lines, seems Year and CO2 will have a positive relation with Temperature, neutral relations between sunspot and Temperature, negative relation with Rain and SOI between Temperature. The negative relation between SOI and Temperature seems to defy laws of physics, as temperature increases, so does pressure. However, SOI is the difference between barometric pressure therefore an increase in SOI implies a decrease in French Polynesia or an increase in Darwin barometric pressure.

**Research Question**

The effects of Climate Change during the 21st century has shown that humankind need to be able to focus resources to minimize damages to the ecosystems and to the civilizations. One of the most recent and most devasting events was the wild bushfire of 2019-2020 in Australia that resulted in the death of 479 humans, 2.8 billion wildlife affects, and 46 million acres of affected land. Creating accurate predictors of environmental conditions could allow emergency services and wildlife protectors to allocate resources to the most at-risk locations and effectively minimize disasters caused by future wildfires. The question is, which regression algorithm has the best predictive power for temperature with previous Australian climate data?

**Data Preprocessing**

The obtained dataset was mostly clean with very few issues. First, the temperature data from Australia Bureau of Meteorology has missing values from 1900 to 1909. These records were dropped using .dropna(), resulting in 103 rows of data. Removing unnecessary columns to result with desired 6 columns for the model was also performed by running .drop(). Then the data was partitioned using train\_test\_split() with test\_size=0.3, which created a train set of 72 rows and a test set of 31 rows.

**Algorithms**

For this project, Linear Regression, Lasso, Ridge, Bayesian Ridge, Decision Trees, and Random Forests were used to predict temperature from the selected features. Each one was implemented by the scikit-learn package available online for them. Then the training set was fitted, and the mean squared error was tested for both the training and the testing set on each regressor. The Random Forest was selected for having the best mean squared error value. It seems that for all the Regressors were significantly effective at predicting temperature from the feature set as well as performs equally between training and testing sets. The Random Forest was the best of all the algorithms and had a similar effect on unseen data with a MSE value of 0.0776 for new data.