Stat 220 Lab 7

1 Introduction

You are a part of a multidisciplinary team of scientists, statisticians, and oceanographers stationed at an international marine research facility. Your team has been tasked with understanding the mechanisms behind the El Niño phenomenon. El Niño is a warming of the ocean waters that affects weather patterns as well as agriculture in regions that in turn affect the rest of the world. Accurate prediction of El Niño is an area of much interest. El Niño happens on average once every 5 years. However, some anomalies have been observed, such as 2016 and 2017, where back to back El Niño's were observed. You can learn more about it here.

Your team has managed to collect a trove of valuable data during various expeditions, risking turbulent seas and unpredictable weather conditions. But data, as raw as the ocean's tides, needs to be refined to extract the wisdom within. As a data-savvy member of this team, you're equipped with the tools of statistics and data science to dig deeper. Our goal in this lab is to build the best model we can based on the data.

Data Acquisition and Initial Exploration

Before diving into the models, understanding the data is crucial.

1. Download the dataset from the following link: `https://richardson.byu.edu/220/sst.csv`
2. Load the dataset into python and perform a preliminary exploration of the data. This initial glimpse will guide your next steps.

Variable	**Description**
water_temperature	Target variable (degrees Celsius)
depth	Depth in meters
month	January, June, or December
salinity	Salinity in PSU
wildlife_seen	Number of wildlife seen
wind_speed	Wind speed in km/h
cloud_cover	Cloud cover in percentage
wave_height	Wave height in meters
oxygen_levels	Dissolved oxygen levels in mg/L

Table 1: Data dictionary for the El Niño dataset.

Regression Analysis

1. Fit an initial linear regression model using all predictors. While this may lead to overfitting, it serves as a starting point.
2. Evaluate the initial model using metrics like MSE and R^2. This will quantify the model's performance.
3. Fine-tune the model to avoid overfitting or underfitting by finding a good subset of predictors. Justify your reasoning as to why you are removing or keeping certain variables.

Regression Tree Analysis

Another approach to tackle this problem is using a regression tree model.

1. Build an initial regression tree model.
2. Tune the model parameters like tree depth and minimum samples per leaf. This ensures that the model is neither too complex nor too simple.
3. Evaluate the tuned model's performance using MSE and R^2.

Dicussion

Be sure to discuss the following in your report

1. State which linear regression model you found to be the best model, reporting appropriate metrics to support your results.
2. State which regression tree model you found to be the best model, reporting appropriate metrics to support your results.
3. State and justify the best overall model between the best regression tree and the best linear regression model
4. Present your models' findings with respect to predicting sea surface temperature, meaning interpret the model in the context of the problem, including interpreting all key
5. State which variables you found to be significant or not significant in your model.

Deliverables

Your final submission will consist of a Python Notebook or some other form of a report. Your report should be well-organized and formatted as a report, results, and visualizations. Make sure to answer the questions directly in the report. For easier grading, and to make sure we don't miss anything when grading, label the questions answered directly i.e. have dedicated sections to the four parts above and be clear what question you are answering in each section.