Ocean Oracles: Twins of the Winds

El Niño-Southern Oscillation (ENSO) Sea Surface Temperature Prediction

Team Members
- Nidhish Doshi
- Mallireddy Rupesh
- Harshit Paras Babariya

Introduction

El Niño and La Niña are integral parts of the El Niño-Southern Oscillation (ENSO) cycle, which significantly impacts global weather patterns, ecosystems, and economies. Accurate prediction of Sea Surface Temperature (SST) is critical for understanding and forecasting these phenomena. Our project aims to develop a machine learning model to predict SST using various atmospheric and oceanographic data.

Problem Statement

Objective

Develop a Machine Learning (ML) model that predicts Sea Surface Temperature (SST) using sequential data. The model will be trained on labelled data and used to predict SST for unlabelled data.

Data Description

- **Date**: Day, Month and Year of observation
- **Latitude and Longitude**: Location of the buoy
- **Wind Data**: Zonal Wind and Meridional Wind
- **Humidity**: Relative humidity during observation
- **Air Temperature**
- **Sea Surface Temperature**: Target Variable

Datasets

- **train.csv**: Contains labelled data for training the model
- **evaluation.csv** and **data_1997_1998.csv**: Contains unlabelled data for prediction

Methodology

Data Preprocessing

1. **Handling Missing Values**: Imputed missing values using the method of multiple imputation and median of the respective feature.

2. **Normalization**: Scaled features using Min-Max normalization to bring all values within the range [0, 1].
3. **Date Handling**: Extracted day, month, and year from the date for feature engineering.

Feature Engineering

1. **Wind Components**: Calculated wind speed and direction from zonal and meridional wind components.
2. **Time Features**: Created cyclical features for month to capture seasonal patterns.
3. **Geographical Features**: Used latitude and longitude to calculate the distance from the equator and central Pacific Ocean.

Model Implementation

We experimented with various models including:

1. **Linear Regression**
2. **Random Forest Regressor**
3. **Gradient Boosting Regressor**
4. **Logistic Regression**
5. **K-Nearest Neighbours**
6. **Support Vector Regressors**
7. **LSTM Neural Network**

After comparing the performance, we selected the LSTM Neural Network for its ability to capture sequential dependencies in the data for data_1997_1998.csv and Random Forest Regressor for the data in evaluation.csv.
We took this decision by splitting the train.csv into train and test datasets in a ratio of 80:20. We took into account various metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE) and R2-Score.

Thought Process

Initial Approach

Our initial approach involved understanding the data and the domain of the problem. We researched the ENSO cycle and the importance of SST in predicting El Niño and La Niña events.

Data Exploration

We performed exploratory data analysis to understand the distribution and relationships between features. Visualization techniques helped us identify patterns and correlations crucial for feature engineering.

Model Selection

Given the sequential nature of the data, we hypothesized that recurrent neural networks, specifically LSTM, would effectively capture temporal dependencies for predicting future data and Random Forest Regressor, would effectively capture trends and would be useful for predicting data in evaluation.csv. We compared it with traditional regression models to ensure robustness.

Challenges

1. **Data Imputation**: Handling missing values without introducing bias.
2. **Feature Scaling**: Ensuring different scales of features did not adversely affect model training.

3. **Model Tuning**: Hyperparameter tuning of the LSTM network was computationally intensive.

Conclusion

The project successfully developed an LSTM Neural Network and Random Forest Regressor model to predict Sea Surface Temperature, demonstrating better performance over traditional models. The insights gained from this project underscore the importance of sequential data handling and feature engineering in time-series forecasting.