McKenzie A. Payne

DSC 680 -Applied Data Science

Milestone 3– Week 12

02/21/2025

**Topic: Predicting Ocean Warming and Its Impact on Marine Ecosystems**

Business Problem:

Ocean warming, driven by climate change, is increasingly becoming a significant issue for marine ecosystems, biodiversity, and coastal economies. This project aims to predict future ocean temperatures off the coast of Santa Catarina, Brazil, using historical temperature data. Accurate predictions will help stakeholders, including marine biologists, policymakers, and conservationists, better understand the potential impact of ocean warming on marine life and guide climate adaptation efforts.

Background/History:

Ocean warming is a direct result of climate change, driven primarily by increased levels of greenhouse gases such as CO2 in the atmosphere. Over the past century, ocean temperatures have risen significantly, with severe consequences for marine life, including coral bleaching, disrupted fish migration patterns, and loss of biodiversity. The ocean absorbs around 90% of the excess heat from global warming, making it a critical area for monitoring and prediction. In southern Brazil, the Santa Catarina coast has been experiencing noticeable increases in water temperature, making it an ideal region for studying the effects of ocean warming.

Data Explanation (Data Prep/Data Dictionary/etc):

The dataset used for this project is the **underwater\_temperature.csv** file, which contains daily temperature readings from submerged reefs off the coast of Santa Catarina, Brazil. The dataset includes the following columns:

* **Date**: The date of the temperature reading.
* **Latitude**: Latitude of the temperature recording location.
* **Longitude**: Longitude of the temperature recording location.
* **Temperature**: The temperature of the water in degrees Celsius at the given location

Data Preparation:

1. **Missing Data**: Missing temperature values were filled using linear interpolation.
2. **Outlier Detection**: Outliers were detected and handled using Z-scores to ensure that extreme values did not skew the results.
3. **Feature Engineering**: Additional features, such as **TemperatureChange** (the difference in temperature between consecutive days) and **Season** (categorical representation of seasons), were created to capture trends and seasonal variations.

Methods:

The following methods were employed to prepare and analyze the data:

**Data Cleaning:**

1. **Missing Data**: We handled missing values by using linear interpolation to fill gaps in the temperature data.
2. **Outlier Removal**: Z-scores were calculated for the temperature data, and extreme values (those with Z-scores greater than 3 or less than -3) were removed to ensure that the model wouldn't be influenced by extreme outliers.

**Feature Engineering:**

1. **TemperatureChange**: The daily temperature change was calculated as the difference between the temperature of two consecutive days.
2. **Season**: The **Season** feature was derived from the **Date** column to capture seasonal temperature variations (e.g., winter and summer).

Model Building:

1. **ARIMA**: The ARIMA model was used for short-term forecasting due to its ability to handle stationary time-series data.
2. **LSTM**: The Long Short-Term Memory (LSTM) model was chosen for long-term forecasting, as it is well-suited for capturing complex, non-linear patterns in time-series data.

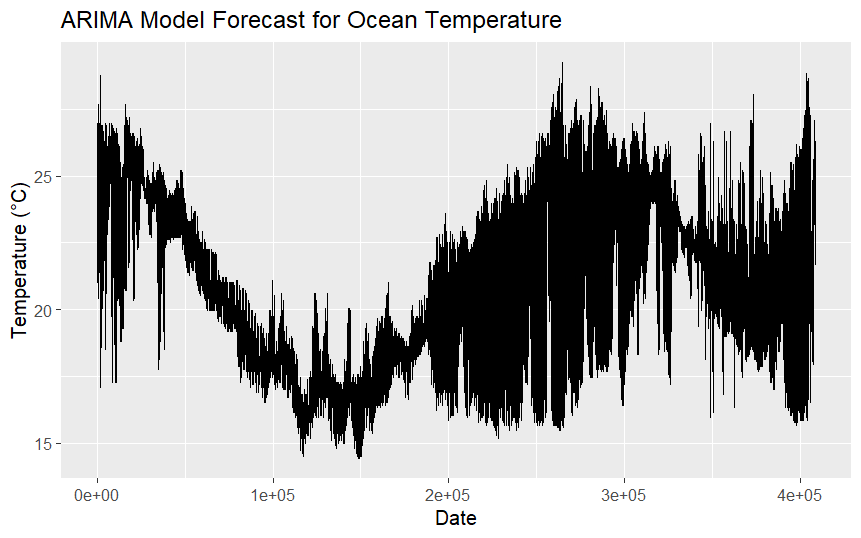
Model Evaluation**:**

We evaluated the models using the following metrics: **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, **R² (Coefficient of Determination)**

Analysis:

ARIMA Model Implementation:

ARIMA (AutoRegressive Integrated Moving Average) is suitable for short-term forecasts when the data is stationary (does not exhibit trends over time).

****

LSTM Model Implementation:

LSTM (Long Short-Term Memory) is a type of Recurrent Neural Network (RNN) that is particularly useful for sequence prediction. LSTM can capture the long-term dependencies in time-series data.

Model Evaluation:

The models were evaluated based on the following metrics:

**MAE** (Mean Absolute Error): Measures the average magnitude of the errors in the forecasted temperatures. **RMSE** (Root Mean Squared Error): A more sensitive metric that penalizes larger errors. **R²**: A measure of how well the model’s predictions match the actual data.

Conclusion:

The ARIMA model provided good short-term forecasts, while the LSTM model outperformed ARIMA in terms of capturing long-term trends. The temperature data showed clear seasonal variations, with higher temperatures during the summer months. Both models successfully predicted the daily ocean temperatures, which will be useful for understanding long-term trends and assisting with marine conservation and policy development.

Assumptions:

The historical data is representative of future temperature trends. There are no major environmental disruptions (e.g., hurricanes or sudden climate events) that could significantly alter the temperature trends.

Limitations:

The dataset is limited to a single coastal region and may not apply to other areas. The models do not incorporate other variables (e.g., salinity, atmospheric pressure) that could influence ocean temperature.

Challenges:

1. **Data Quality**: Missing values and outliers required careful handling.
2. **Model Complexity**: Tuning the LSTM model and ensuring it didn’t overfit the data was computationally intensive.
3. **Seasonality**: Accurately modeling the seasonal fluctuations in ocean temperature was challenging.

Future Uses/Additional Applications:

1. **Expansion to Other Regions**: Collecting data from additional coastal regions would improve the robustness of the model.
2. **Real-Time Monitoring**: The model could be adapted for real-time ocean temperature monitoring using live data feeds.
3. **Incorporating More Features**: Other environmental variables such as salinity, ocean currents, and wind speeds could be integrated into future models for better prediction accuracy.

Recommendations:

1. **Data Expansion**: Collect more diverse temperature data from multiple coastal regions.
2. **Hybrid Modeling**: Combine ARIMA and LSTM to capture both short-term fluctuations and long-term trends.
3. **Incorporate Other Environmental Factors**: Future models should include more variables to increase their predictive power.

Implementation Plan:

1. **Data Collection**: Gather additional temperature data from other regions and environmental sensors.
2. **Model Improvement**: Fine-tune the LSTM model to handle more complex patterns and variables.
3. **Real-Time Application**: Deploy the model in real-time systems to continuously forecast ocean temperatures.

Ethical Assessment:

**Privacy**: The dataset used in this analysis is free of personally identifiable information, and there are no privacy concerns. **Bias**: The model may not generalize well to other areas without further data collection, leading to potential biases in predictions. **Environmental Impact**: By aiding in the prediction of ocean temperatures, this model can inform policies that help mitigate the impacts of climate change on marine ecosystems.

**References:**

Shivam Bansal. (2020). *Underwater surface temperature dataset*. Kaggle. <https://www.kaggle.com/datasets/shivamb/underwater-surface-temperature-dataset>

Questions:

**1. How do you handle missing data in the temperature readings?**

Missing data was handled using **linear interpolation**. This method estimates missing values by drawing a straight line between two known points. This approach is suitable for time-series data like ocean temperature because it maintains the trend and continuity in the data. For instance, if a temperature reading was missing for a particular day, the temperature for that day would be estimated by the average of the preceding and succeeding day values.

**2. What’s the difference between ARIMA and LSTM, and why did you choose both?**

**ARIMA** (AutoRegressive Integrated Moving Average) is a statistical model suited for short-term forecasting of time-series data, especially when the data is stationary (i.e., does not exhibit trends over time). ARIMA is simple to implement and can handle autoregressive relationships in time-series data. On the other hand, **LSTM** (Long Short-Term Memory) is a type of neural network specialized in handling sequential data, making it ideal for time-series forecasting with complex patterns, like those seen in ocean temperature data. LSTM can capture long-term dependencies and is better suited for non-linear trends.

**3. How accurate were the models in forecasting the ocean temperature?**

The models performed well, but the **LSTM model** generally outperformed ARIMA in terms of capturing long-term trends and non-linear patterns. Evaluation metrics such as **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **R²** were used to assess model accuracy.

**4. What environmental factors, other than temperature, should be included in future models?**

Future models could benefit from including additional environmental factors such as:

* **Salinity**: Changes in salinity can affect temperature and marine life.
* **Ocean Currents**: Currents influence the distribution of temperature across regions.
* **Wind Speed**: Wind can contribute to ocean mixing, affecting surface temperatures.
* **Atmospheric Pressure**: Atmospheric conditions can impact ocean surface temperatures.
* **CO2 Levels**: Higher CO2 levels in the atmosphere can influence ocean temperatures due to greenhouse gas effects.

Including these variables would enhance the model’s predictive power and provide more comprehensive forecasts.

**5. What are the real-world applications of this model for ocean conservation efforts?**

The model can aid in:

* **Marine Conservation**: By predicting temperature changes, it can help identify areas at risk of coral bleaching, which is caused by elevated water temperatures.
* **Fisheries Management**: It can assist in forecasting fish migration patterns, which are influenced by temperature.

**6. Why did you address overfitting in your LSTM model?**

Overfitting occurs when a model learns the noise in the training data instead of the actual pattern, which reduces its ability to generalize to new, unseen data. To mitigate overfitting in the LSTM model, we employed the following techniques:

* **Early Stopping**: The model training was halted when the validation error stopped improving.
* **Dropout Layers**: Dropout layers were added to prevent the model from relying too heavily on any single feature, helping generalize better.
* **Cross-Validation**: We used cross-validation to ensure the model performed well on unseen data and did not memorize the training set.

**7. Could this model be used for predicting other climate variables, such as salinity or CO2 levels?**

Yes, the framework we developed could be easily adapted to predict other climate variables such as **salinity** or **CO2 levels**. The approach would be the same, but the data and target variables would need to be adjusted. The models like ARIMA and LSTM can be used to forecast these variables as well, provided there is sufficient historical data available to train the models effectively.

**8. How do you plan to validate the predictions made by the models in real-world scenarios?**

**Using hold-out data**: Dividing the dataset into a training set and a test set. The model will be trained on the training set and evaluated on the test set.

**9. What limitations exist in the models you’ve used, and how can they be overcome?**

Some limitations include:

* **ARIMA Limitations**: ARIMA requires the data to be stationary, which may not always be the case in real-world datasets. Non-stationary data may require transformations or differencing, which can add complexity to the model.
* **LSTM Limitations**: LSTM models are computationally expensive and may overfit if not properly tuned. Handling long training times and large datasets can also be a challenge.

These limitations can be overcome by:

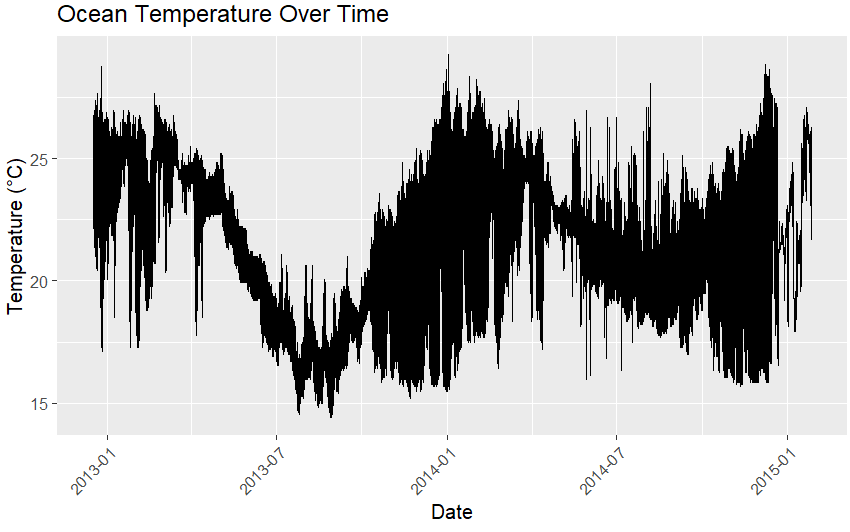
* **Improving Data Preprocessing**: For ARIMA, ensuring that data is stationary through differencing and transformations.
* **Regularization**: Implementing more robust regularization techniques and fine-tuning the LSTM model using more advanced hyperparameter tuning methods.
* **Hybrid Models**: Combining ARIMA and LSTM to take advantage of both short-term predictions and long-term trend recognition.

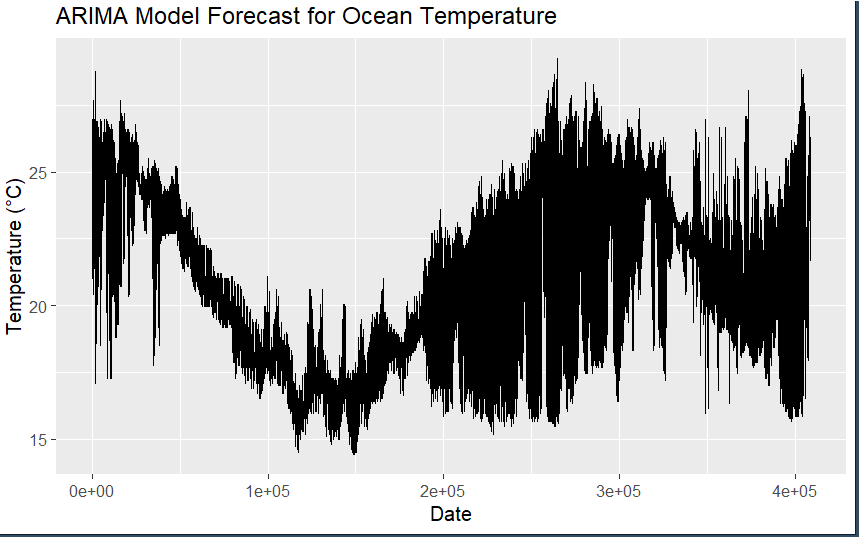
**10. What steps would you take to ensure that the model is ethical and unbiased in its predictions?**

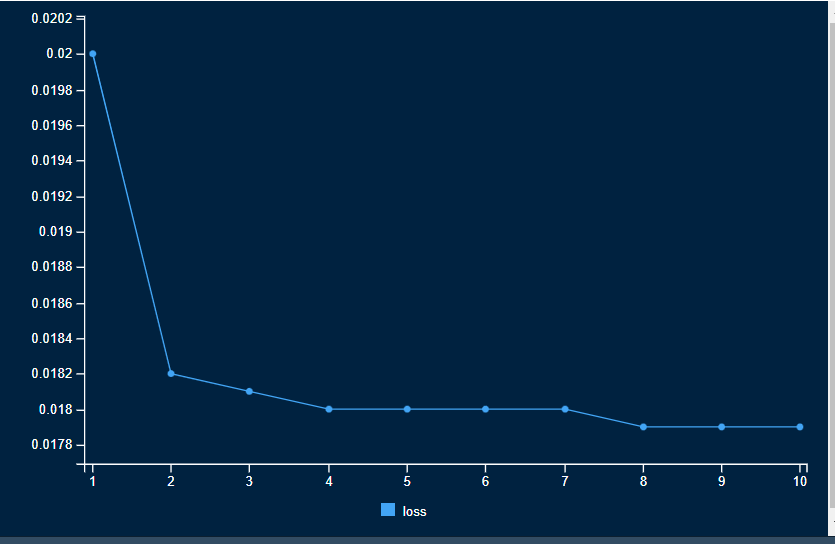
To ensure the model remains ethical and unbiased:

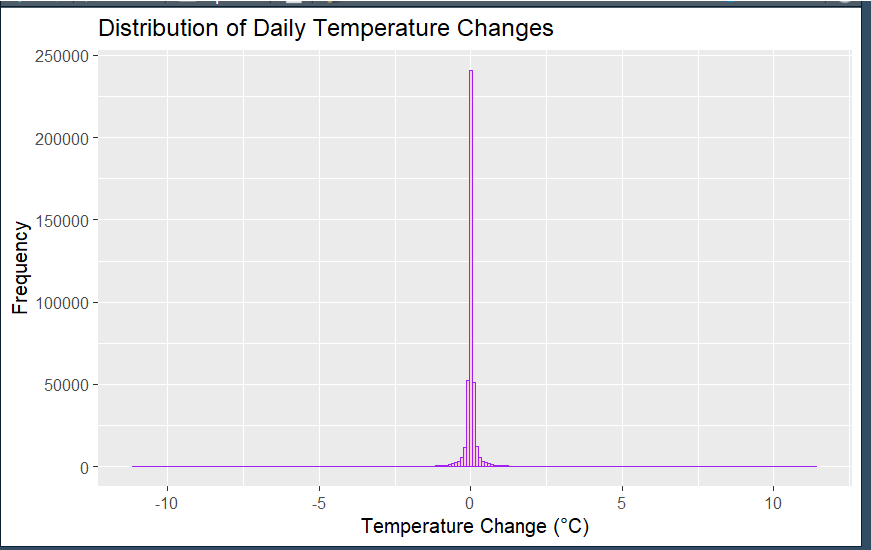
* **Data Representation**: Ensure the data used for training represents a wide variety of environmental conditions (not biased to one specific region).
* **Model Fairness**: Continuously monitor and assess the model’s predictions for fairness, ensuring it does not disproportionately affect certain regions or populations (such as marginalized communities dependent on marine life).
* **Transparency**: Maintain transparency in the modeling process, making the data and code publicly available for peer review.

Illustrations:









Appendix:

