STA 141C Project

Predicting Brazil's Temperature for Climate Change Analysis

Group: *Alexander Kapustin, Muhammad Reza, Samuel Pang Shao Herng, Zhanyi Qiu*

Abstract

Global warming has had a significant impact on Brazil's Amazon Basin, marked by an increase in wildfires and droughts. Brazil, as the largest greenhouse gas emitter in Latin America and the Caribbean, has been focused on reducing emissions and implementing strategies to mitigate climate change effects. The goal of our research is to build a model capable of forecasting future weather conditions across Brazil. This will aid in understanding climate change effects and laying the groundwork for better public policies and strategies to combat global warming. The data is formatted as a time series, in which "airTemp", or the air temperature, was the attribute to be predicted. Through preprocessing and preliminary analysis techniques, the data was cleaned and formatted appropriately, and the variables of interest were identified using a correlation threshold of 0.8. The linear regression, ridge regression, and lasso regression models had similar performance metrics (0.8939 MSE, 0.5987 MAE, and 0.9331 R-squared). The polynomial regression model had a multiple R-squared value of .9988, with an MSE of 0.016 and an MAE of 0.089. A random forest model with 6 features showed superior accuracy compared to other models, except the polynomial regression model. Some limitations of our work included the extensive size of our dataset, which led to longer computation times for certain model selection techniques. Additionally, there were instances of attribute omission before model training, and the initial state of our data might have influenced its formatting after preprocessing.

Introduction

Brazil, home to 60 percent of the Amazon Basin, boasts the largest remaining tropical rainforest, teeming with an unparalleled diversity of plant and animal species. This biodiversity not only sustains ecosystems crucial for the planet's health but also supports over 30 million people. However, Brazil faces significant challenges as the largest greenhouse gas emitter in Latin America and the Caribbean. Climate change threatens the Amazon Basin's delicate balance, impacting temperature-sensitive species, and freshwater ecosystems, and increasing the risk of devastating wildfires and extreme weather events. Recognizing the urgency, Brazil has committed to ambitious greenhouse gas emission reduction targets, underlining its dedication to combatting climate change. International organizations like USAID collaborate with the Brazilian government, civil society, and private sector to support biodiversity conservation efforts and sustainable development initiatives in the Amazon. Through strategic partnerships and innovative programs, significant progress has been made in reducing deforestation rates, promoting biodiversity-friendly enterprises, and empowering indigenous communities to adapt to climate change. These collaborative efforts not only mitigate climate change but also foster resilience and sustainable livelihoods among vulnerable populations in the region.

However, despite these efforts, the Brazilian Amazon experienced a surge in fires in 2023, worsened by a severe drought that led to historically low water levels in rivers and lakes. Amazonas, the country's largest state, witnessed a record 3,181 fires starting from October 1, 2023, doubling the number from the same period in 2022. The capital city of Manaus found itself shrouded in thick smoke, intensifying respiratory issues and leading to a rise in medical emergencies. Researchers warn that fires now constitute the primary threat to the Amazon, undermining decades of environmental protection efforts. Fueled by human activities and aggravated by severe drought conditions, these fires are projected to persist in the coming months. Alongside the environmental devastation, severe humanitarian crises have emerged, with

communities grappling with water and food shortages, and wildlife facing mass casualties. Climate change compounds these challenges, emphasizing the urgent need for coordinated action to mitigate risks to the Amazon Rainforest and its inhabitants.

Goal

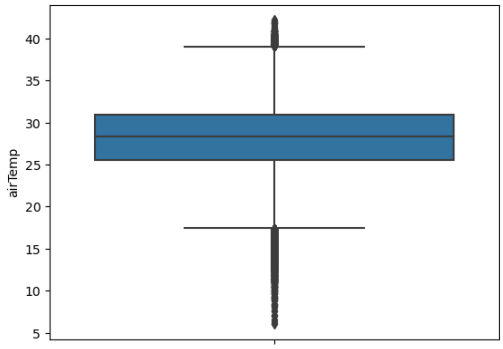
Our goal is to investigate the impact of global warming on weather patterns in various regions of Brazil over the years and develop a predictive model capable of forecasting future weather conditions. By accurately predicting temperature variations, we aim to track the progression of global warming and provide valuable insights to stakeholders, including large corporations, the general public, and policymakers.

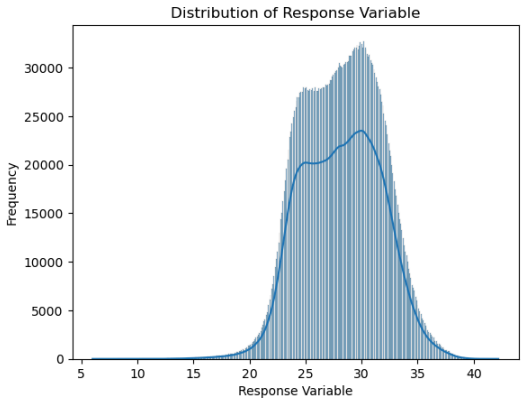
Our research findings will empower stakeholders to make informed decisions regarding policies, investments, and mitigation strategies to address the effects of climate change in Brazil. With access to reliable forecasts, stakeholders can implement proactive measures to minimize the adverse impacts of climate change, protecting both the environment and the well-being of communities across the country.

Data

Our data comes from the Kaggle website and it contains data on the weather patterns of several regions in Brazil. Each main dataset consists of 27 columns, encompassing various meteorological parameters, locational attributes, and time-based attributes. These parameters include daily dates, precipitation levels, atmospheric pressure at station level, minimum and maximum air pressure within the last hour, solar radiation, minimum and maximum air temperature, minimum and maximum dew point temperature, minimum and maximum relative humidity, as well as wind speed and direction. This data set is relatively large, as it has 8,392,319 rows of data. For our project, we will be using “airTemp” as the target variable, as we seek to predict future temperatures to provide further insights into the rate of climate change in Brazil.

In order to visualize the distribution of the target variable, “airTemp”, we used a histogram and a boxplot as shown here. We realize that a significant proportion of the temperatures lie between 25-32 degrees, and quite a number of points fall out of the 1.5 * Interquartile range (IQR) range, which can also be seen by the right skew in the histogram.



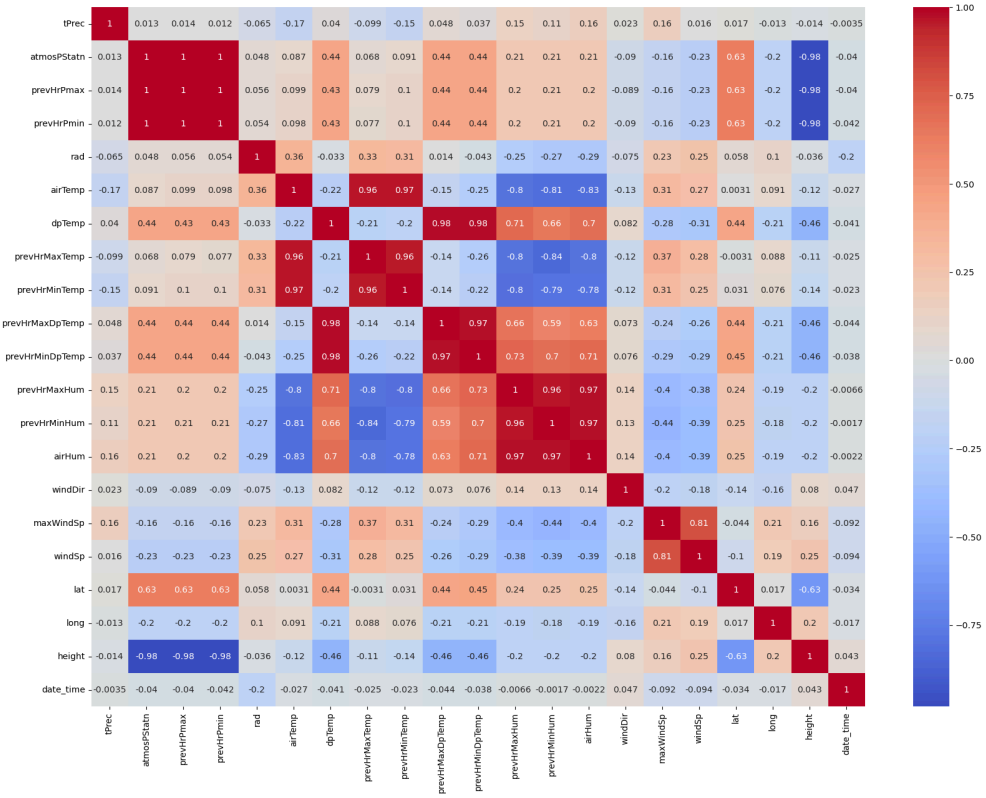


Data Preprocessing

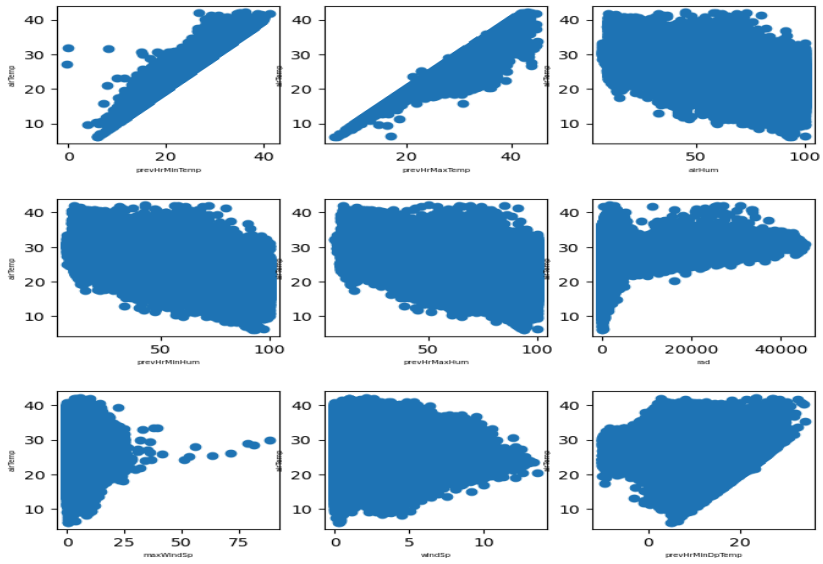
After visualizing our initial data, we proceeded with data preprocessing. Here are some of the issues with the data and how we tackled them.

1. Removal of NaNs: NaNs which were represented as -9999 in the dataset, had to be removed as it represented null data. To do so, we converted the -9999 values to NaNs and removed rows that had this value.
2. Removal of variables that meant the same thing: The combination of variables (latitude, longitude, height) are sufficient to identify a specific location or station, thus variables that represent similar location data such as station code, region, and state are removed.
3. Categorical data conversion to numerical: Data such as date and hour was considered to be categorical. To convert them to numerical values, we combined the variables date and time into a single column called date_time. We then mapped date_time to a Gregorian ordinal value which is numerical. This mapping is a categorical to numerical mapping specific for date times.
4. Column renaming: The columns had Portuguese names initially, thus we had to convert them to English names for comprehensibility.

We then constructed a correlation matrix to display the correlations between the variables. The target variable is also shown by an arrow. In general, for the correlation matrix, the more intense the color, the greater the correlation.



To visualize the significant correlations better, we generated scatterplots for the variables most correlated with “airTemp” by using data from the correlation matrix. In the plot, the magnitude of correlation decreases as we go down for each row or column. We can see that especially for the first two graphs, which are “prevHrMinTemp” and “prevHrMaxTemp” respectively, there is a relatively linear relationship between these variables and “airTemp”.



According to the correlation matrix, we realize that atmospheric pressure at the station level is perfectly correlated with “prevHrPmin” and “prevHrPmax”. This makes sense as any increases in the “prevHrpMin” or “prevHrPmax” would likely also mean corresponding changes in atmospheric pressure. Thus, we can proceed to omit “prevHrPmin” and “prevHrPmax” from our dataframe. Some other examples of high correlation are dewpoint temperature with “prevHrMaxDpTemp” and “prevHrMinDpTemp”, which is logical given that the min and max dewpoint temperature range in the previous would significantly influence the current dewpoint temperature, thus we can omit the “prevHrMax” and “Min DpTemp” from our dataframe. There are also many other highly correlated variables, with a correlation value > 0.8. For our project, we will seek to drop variables having a correlation value of > 0.8 with another, resulting in a cleaned dataset containing 11 predictor variables and 1 response variable.

Simple Model Selection

We first attempted three simple regression models(Linear, Ridge, and Lasso) to compare simple regression models with more complex ones effectively. We aimed to identify the best simple regression model for our dataset. We selected the best linear regression model using forward and backward selection methods. This method identified ten variables as significant, suggesting that each variable in our dataset might meaningfully contribute to predicting air temperature. However, since “prevHrMinTemp” is highly correlated to our target variable 'airTemp,' we will also remove it from the dataset. To optimize the performance of the Ridge and Lasso models, we used K-fold cross-validation and found that the best alpha for the Ridge model is 10.0, and for the Lasso model is 0.001.

Under the most optimized setting, the three models' performance was very similar: the values of MSE, MAE, and R2 were almost identical: they were approximately 0.8939, 0.5987, and 0.9331, respectively. Given that all three values (MSE, MAE, and R-squared) across the three different regression models (Linear, Ridge, and Lasso) are nearly the same, this indicates that for this particular dataset, all three models perform similarly in terms of prediction error and explained variance. Considering Lasso's ability to handle datasets with many features effectively, we select it as the best simple regression model to compare performance with more complex models.

Polynomial Regression

In our initial attempt to construct a polynomial regression model, we sought to determine the optimal polynomial degree for each predictor variable in the train set using the ANOVA approach. The analysis yielded that the optimal polynomial degrees varied across predictor variables. Degree 1 sufficed for variables like "lat", "long", and "prevHrMinTemp", while degree 2 was preferred for "windSp", "atmosPStatn", and "tPrec". Variables "dpTemp" and "windDir" performed best at degree 3, while "airHum" and "rad" showed optimal performance at degrees 4 and 5, respectively. It is important to note that we manually selected the degrees based on individual ANOVA results for each predictor variable. This decision was made because ANOVA results often showed similarities across multiple degrees. Thus, to avoid the risk of overfitting the model with high polynomial degree predictors, we opted for lower degrees, accepting a slight increase in RSS within a 10% threshold.

The final polynomial regression model incorporated these selected polynomial degrees for each variable. Evaluation on the test set demonstrated a low percentage error, indicating that the model's predictions closely matched the actual values. Additionally, the MSE and MAE confirmed the model's performance, showing relatively low errors.

The polynomial regression model showed highly promising predictive performance for estimating air temperature from environmental variables, almost seeming too good to be true. To refine the model further, we decided to omit a highly correlated variable, "prevHrMinTemp." Additionally, we attempted to conduct 10-fold cross-validation. However, due to the dataset's size and the limited computational power of our personal computer, we encountered difficulties. To address this challenge, we decided to take a small random sample (~10%) from the dataset and train the model on it. This approach enabled us to utilize 10-fold cross-validation for training our polynomial regression model.

The sampling approach proved invaluable in significantly reducing the processing time required for cross-validation, delivering results within a mere five minutes. Upon evaluation, the results closely mirrored those obtained previously. Notably, the initial model yielded a Multiple R-squared value of 99.87%, with an MSE of 0.016 and an MAE of 0.089. The model exhibited a mere 0.45% deviation in accuracy, further highlighting its strong performance. The updated model yielded a Multiple R-squared value of 99.88%, with an MSE of 0.017 and an MAE of 0.09, with only a slight 0.47% deviation in accuracy, reaffirming our previous results.

The highly promising results of our polynomial regression model demonstrate its effectiveness in estimating air temperature from environmental variables, marking it as the most robust model yet.

Random Forest

We attempted to utilize the Random Forest method as it differs from previous regression models since it is a non-parametric model. For our Random Forest model, we will generate 100 trees to fit the data, but we will limit the number of features to be taken into account to 6. We realize that the Random Forest model achieved a lower MSE of 0.036 as well as a lower MAE of 0.11 compared to linear, lasso, and ridge regression, but has a higher MSE and MAE than polynomial regression. We chose not to compare the R-squared values of the Random Forest model with the previous models given that comparing R-squared values of non-parametric models against those of parametric models may not be directly applicable or interpretable. In addition, similar to the above parametric models, it was mentioned during our presentation that prevHrMinTemp could be contributing to low MSE for the Random Forest model. Thus, we would drop this column and see how the models above perform. Unlike the parametric models, there was a decrease in the MSE and MAE values, which became 0.019 and 0.09 respectively, showing an improvement in model predictions.

Discussion

Although our methods led to the development of an effective model for forecasting the weather patterns in Brazil, there were some limitations to our approach.

Firstly, there were limitations to our feature selection methodology. Some variables that were omitted prior to training may have been relevant. Exploring techniques like forward, backward, or best subset selection on an expanded set, if not the full set, rather than the limited set of variables could have been beneficial. For future work, we suggest trying different correlation threshold values to determine which features to omit. Many of these methods were not feasible due to the dataset size and the lack of supercomputing resources at hand. Given the large size of the dataset, some of these feature selection methods may work for some models (e.g. OLS), while some of the methods (best subset selection) may be computationally less feasible for others (e.g. polynomial regression). The trade-off between dataset size and the suitability of feature selection methods must be considered.

Secondly, the data is not evenly distributed across the years due to disproportionate station data. Time series analyses are generally structured and proportioned evenly, where each "year" of data has an equivalent number of data points. This potentially inhibits the efficiency of an ordered train-test split in the context of time series analysis, where preceding data is used to forecast subsequent values. We accounted for this issue and were able to order the train-test split successfully, but it is an interesting perspective that is worth mentioning.

The last significant limitation occurred when trying to train certain models given the large dataset size. The random forest with a number of features set to 6 took approximately 25 minutes, while training a bagging model was not feasible as the duration extended to 30+ minutes without termination. In addition, it was not feasible to conduct multiple-fold CV on a full training set, as hyperparameter tuning via K-fold CV would not be feasible given the above large computational time. The limitations of applying polynomial models to large datasets include the challenge of selecting the optimal polynomial degree through K-fold CV. Due to computational constraints, random sampling is necessary to analyze a smaller sample size, which may affect the model's predictive performance on the entire dataset.

Conclusion

In summary, after evaluating various machine learning techniques on the dataset, Polynomial Regression and Random Forest emerged as the top-performing models. Polynomial Regression slightly outperforms Random Forest by a narrow margin. Linear, Ridge, and LASSO Regression yield less satisfactory results. Therefore, we would not recommend considering them for predicting temperature variations in Brazil on new datasets. However, it's worth noting that performing k-fold Cross-Validation on a large dataset like this requires substantial computing power, and random sampling is necessary. Additionally, the models presented may yield different results if supercomputing resources were utilized for the analysis. With regards to our top-performing models, we believe they could be effectively utilized to predict the future weather in Brazil and inform those in power of potential decisions they could make to combat climate change and preserve the environment within the Amazon Basin.

Moving forward, further research could explore alternative feature selection methods, such as forward, backward, or best subset selection, to optimize model performance. Additionally, investigating different correlation threshold values and deploying supercomputing resources could improve the robustness and scalability of predictive models.

In conclusion, our results emphasize the importance of utilizing advanced machine learning techniques to address climate change and protect the ecological balance of the Amazon Basin in Brazil.

References

https://education.nationalgeographic.org/resource/amazon-rainforest/

https://news.mongabay.com/2023/10/people-and-nature-suffer-as-historic-drought-fuels-calamitous-amazon-fires/

https://www.kaggle.com/datasets/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region/data