Data Challenge 2023 Report: Temperature Predictions in USA (Jan 2021 – May 2022)

Principle Investigator: ChadGPT

Introduction

As we plan for the tomorrows of our lives, we often turn to our phones to check the temperature forecasts. While usually accurate, there are times when the actual temperature is hotter or colder than projected. Thus, our questions of interest for this analysis are as follows: **1)** When do weathermen make more accurate predictions? **2)** If the predictions differ a lot from the actual readings, what factors cause overestimations/ underestimations? We believe that, among the many factors, geographical characteristics such as latitude and longitude and time variables such as month and season affect the temperature forecast accuracy.

Background Information and Intuition

Forecasters collect data from sources such as satellites and radar and then use prediction models based on these data to predict temperature[1]. Despite such technological advancements, temperature forecasts have yet to reach 100% accuracy due to the chaotic nature of the atmosphere. We will use several of the 18 variables in the weather dataset to understand the pattern of the errors made.

We start with a two-sided t-test to examine whether the mean of the observed temperature is the same as the mean of the forecast temperature. We obtain a p-value of ~0.0005, which leads us to reject the null hypothesis that the mean temperature observed is the same as the mean temperature prediction.

Now that we are confident about the presence of errors, we analyze the prediction inaccuracies, which are calculated by forecast temperature - observed temperature. We use a histogram to map out the error distribution and a boxplot to see if there are any outliers in the prediction error. We can see some outliers in this data, which signifies that the weather forecasters sometimes overestimate and underestimate the temperature.

Analysis Part 1: Overestimations vs Underestimations

We view this as a binary classification problem, where we set 0 = Underestimation and 1 = Overestimation. Looking at the outliers to answer this question, we build and compare **two** models:

1) Logistic Regression with LASSO for feature selection and bi-directional stepwise selection. These methods optimize our model and help prevent multicollinearity.
2) Extreme Gradient Boosting Classifier (a decision-tree algorithm) with LASSO for feature selection. This algorithm builds upon weaker models to obtain the best model.

To compare these models, we look at the accuracy and AUC (area under the curve) metrics.

Model 1 (Logistic Regression)		**Model 2 (Extreme Gradient Boosting)**	
Accuracy	AUC	Accuracy	AUC
67.5%	0.708	70.4%	0.768

It is clear that **Model 2** has better accuracy and AUC scores, thus we will identify the features leading to overestimations/ underestimation based on Model 2. By plotting a graph of feature importance, we gather that the **type of temperature forecast** (high or low), **longitude**, and **latitude** are the three most significant variables. We extend this analysis further by evaluating how the top 10 most important features contribute to temperature overestimations/ underestimations. For this, we use a Partial Dependence Plot (PDP), which can be found in the .rmd code file.

[1] National Weather Service. (n.d.). Forecast Process. https://www.weather.gov/about/forecast-process

Factors that contribute to temperature overestimations/ underestimations.

Overestimations	Higher longitude, lower latitude, forecast for high temperature, extremely high or lower ‘elevation_change_four’, medium to high elevation, and higher or extremely low wind speeds
Underestimations	Lower longitude, forecast for low temperature, high ‘elevation_change_four’, extremely high elevation, forecast made on the month of January, February, or March, and Sunny forecast outlook

Analysis Part 2: Factors Contributing to More Accurate Temperature Predictions

We now shift our attention towards evaluating the accurate temperature predictions, which we define as the prediction error data points within the boxplot. Here, we use the **Select KBest algorithm based on linear regression** to perform feature selection with the absolute prediction error being our y-variable. We utilize the same 14 initial predictors as we did when answering our first question. Though there is no real significance, we choose K=7 as it is half of the total predictors. We further inspect each of the seven features to determine what factors yield a more accurate prediction. Note: we are relaxing correlation assumptions as we are not fitting a model.

Important Variables	**What Makes a More Accurate Prediction?**
1) Month	Readings taken in June, July, and August. They make up the summer season. This agrees with the NOAA’s observation[2].
2) Forecast Outlook	Readings taken during SMOKE, BLZZRD, and TSTRMS. Coincidentally, these outlooks mainly make up the summer season.
3) Type of Forecast (high or low)	Regardless of the type of temperature forecast, both have low average prediction error.
4) State	Readings taken for FL, IN, and GA.
5) City	Readings taken for Daytona Beach (FL), Atlanta (GA), and Detroit (MI, adjacent to IN). We can almost see an apparent pattern here.
6) Region	Readings taken for South and Midwest USA. Notice that FL and GA are in the South and IN/MI in Midwest.
7) Longitude	Readings taken for locations at lower longitude. We can see that all previous location variables concur to the fact that the East coast of the USA make more accurate predictions.

Conclusion

To sum up, we discovered that multiple factors, including geographical attributes and time features lead to more accurate predictions, overestimations, and underestimations. We have genuinely garnered a great deal of new knowledge from our 3-week analysis. Nonetheless, some areas could be improved, such as finding a model to predict the errors. In the future, we hope that our readers can use our report to learn more about the intricacies of weather forecasting and why it is hard to predict temperatures precisely.

[2] NOAA. (2022). Average Seasonal Temperatures in the Contiguous 48 States, 1896-2021. https://shorturl.at/ckoF8