Predicting Maximum Temperatures in Southern California with Machine Learning

Marianna Shand
UCLA Atmospheric and Oceanic Science C111
December 8, 2023

INTRODUCTION

Weather predictions play a pivotal role in numerous facets of our lives. They give us a forecasted glimpse into atmospheric conditions that shape our daily activities and long-term planning. The significance lies in its ability to preemptively inform us about potential hazards, enabling proactive measures to safeguard lives, property, and infrastructure against severe weather events like storms or extreme temperatures. In the agriculture sector, temperature predictions help farmers in crop planning. The predictions allow for better harvests and minimize losses. Moreover, people often plan events, travel, and outdoor activities based on expected weather conditions. Ultimately, weather predictions serve as an invaluable tool, empowering us to make informed decisions and adapt to the dynamic nature of our environment.

We used the ERA5 fifth-generation ECMWF reanalysis dataset spanning over eight decades from 1940 onwards to make temperature predictions for November 2023. The ERA5 dataset is a comprehensive global dataset that amalgamates model data and worldwide observations through data assimilation. It combines prior forecasts with up-to-date observations, creating an accurate depiction of atmospheric states like temperature (Hersbach et al. 2023). Unlike real-time forecasting, reanalysis operates at reduced resolution, ensuring a consistent dataset over decades by incorporating improved observations retrospectively. With hourly estimates for diverse atmospheric, ocean-wave, and land-surface parameters, ERA5 includes uncertainty estimates derived from a 10-member ensemble, aiding in assessing the data's information content and sensitive areas (Hersbach et al. 2023).

Using the temperature and humidity data from the ERA5 reanalysis dataset, we predicted the maximum temperature of each hour for the last few days of November 2023. In the context of solving this problem and because this is a direct weather prediction requiring pertinent parameters, supervised machine learning was required. Ridge regression, a linear regression technique, was the supervised learning technique used for the predictions. This type of regression is suitable with labeled data, meaning the output must already be known. Linear regression specifically deals with predicting a continuous target variable based on one or more input features. In supervised learning, a clear understanding of predictions, along with labeled data (inputs paired with corresponding outputs) is integral.

Our approach tackled the challenge of predicting temperatures by leveraging Ridge regression and incorporating a range of features to refine our model's accuracy and successfully predicted the temperature for the last two days of November 2023. Through meticulous feature engineering, including hourly and daily averages alongside derived metrics, we captured nuanced patterns within the data. Our conclusions from this comprehensive analysis showcase promising results: a notably reduced mean absolute error, stronger correlations between predictors and the target variable, and insightful visualizations that highlight the model's ability to capture trends. We've learned that through diligent feature engineering and model refinement we can enhance the accuracy of temperature predictions.

DATA

The dataset used was obtained from the Copernicus Climate Change Service (C3S). The European Centre contracted the data for Medium-Range Weather Forecasts, operator of C3S on behalf of the European Union (Delegation Agreement signed on 11/11/2014 and Contribution Agreement signed on 22/07/2021) (Hersbach et al. 2023). The dataset was downloaded from the public Climate Data Store and contains over two thousand rows of data, accompanied by many columns of features, capturing various aspects such as time, temperature, relative humidity, maximum temperatures, and minimum temperatures. (Hersbach et al. 2023)

Specifically, the data description is as follows:

DATA DESCRIPTION	
Data type	Gridded
Projection	Regular latitude-longitude grid.
Horizontal coverage	Global
Horizontal resolution	Reanalysis: 0.25° x 0.25° Mean, spread and members: 0.5° x 0.5°
Vertical coverage	1000 hPa to 1 hPa
Vertical resolution	37 pressure levels
Temporal coverage	1940 to present
Temporal resolution	Hourly
File format	NetCDF

Relative humidity	%	This parameter is the water vapour pressure as a percentage of the value at which the air becomes saturated (the point at which water vapour begins to condense into liquid water or deposition into ice). For temperatures over 0°C (273.15 K) it is calculated for saturation over water. At temperatures below -23°C it is calculated for saturation over ice. Between -23°C and 0°C this parameter is calculated by interpolating between the ice and water values using a quadratic function.

Temperature	K	This parameter is the temperature in the atmosphere. It has units of kelvin (K). Temperature measured in kelvin can be converted to degrees Celsius (°C) by subtracting 273.15. This parameter is available on multiple levels through the atmosphere.

Table 1: The data source and variable description from the Copernicus Climate Change Service. (Hersbach et al. 2023)

From this dataset, we pulled temperature and relative humidity data from 1000 hPa (~ surface level) spanning hourly from September to November of 2023 of the sub-region North

35°, West -120°, South 32°, and East -115°. Before diving into the analysis, we performed several preprocessing steps to ensure data quality and suitability for modeling. We calculated the average temperature per hour over the defined area. We did this to reduce the dataset to a 1D time series so a more accurate forecast could be made.

Figure 1

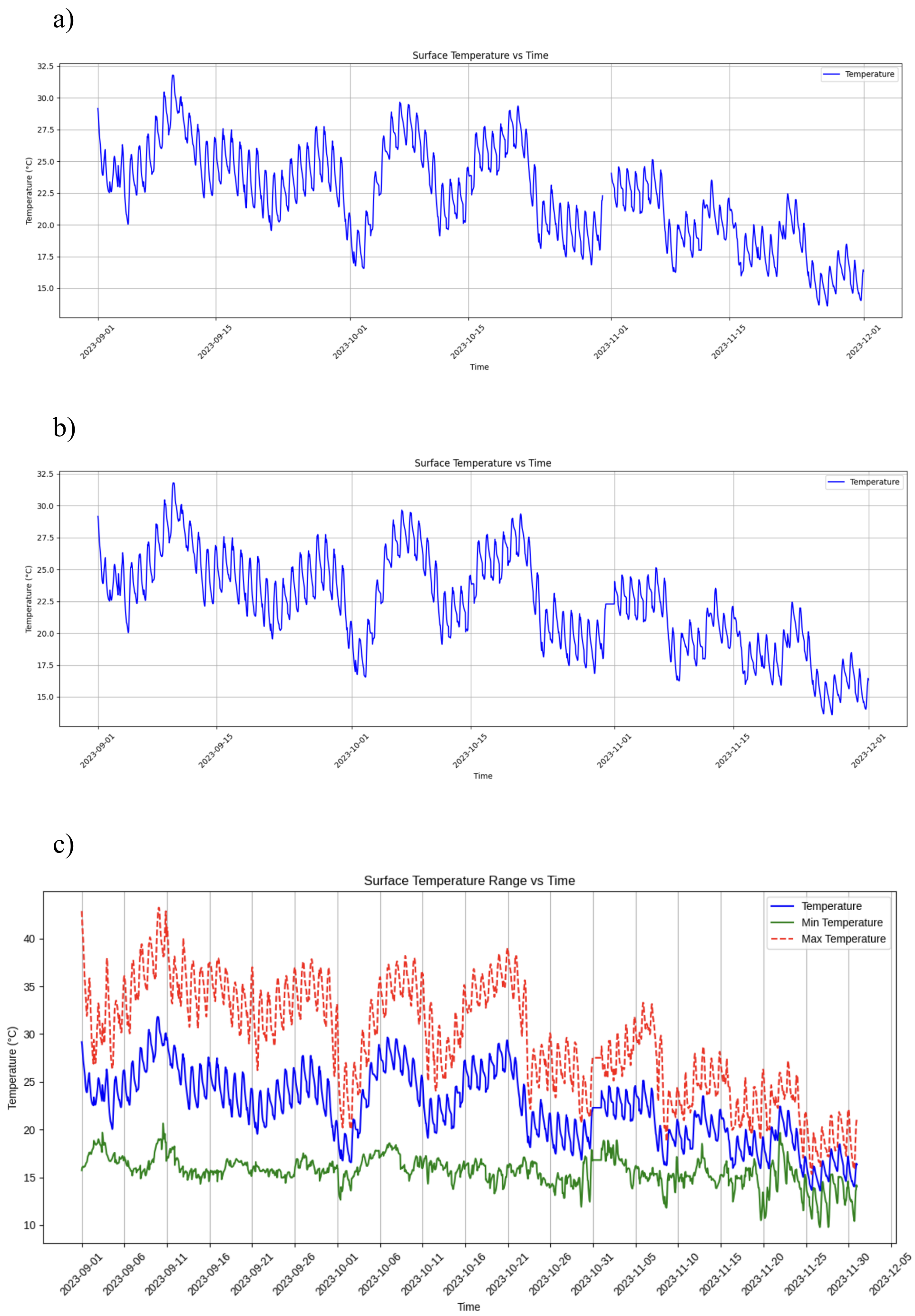

Figure 1a, 1b, 1c: Plot 1a displays the variation of average temperature over time, measured in degrees Celsius. The graph illustrates the hourly fluctuations in temperature recorded across the observed time period. Note the gap in the data on Oct. 31 [1a]. Plot 1b displays the variation of *average temperature (updated variable) over time, measured in degrees Celsius Note the removal of the gap on Oct. 31 [1b]. Plot 1c displays the variation of average temperature, minimum temperature, and maximum temperature (updated variable) over time [1c].*

The Temperature variable is the main variable we are working with to forecast the maximum temperature for the last few days of November, but there is a gap in the time series data on October 31st that hinders the running of the regression. Machine learning models do not do well with null values, so we had to replace these NaN values with the most recent non-NaN value observed in the column. Taking this approach is very useful when handling time series data and carrying the most recent value forward does not upset the analysis or modeling.

Modeling

Ridge Regression is a linear regression technique that incorporates regularization to address overfitting, which can be seen in standard linear regression models. Despite being a variant of linear regression, Ridge Regression extends the basic linear regression by adding a penalty term to the regression coefficients. Ridge Regression is particularly well-suited for datasets where multicollinearity among predictors is anticipated. It helps address this issue effectively. In scenarios where features (predictors) are correlated, standard linear regression might become sensitive to the variations or noise present in the data and can lead to unstable coefficient estimates.

The addition of features such as the Max Temperature and Min Temperature gives the dataset predictors of different atmospheric measures "temperature," "max_temperature," "min_temperature" and "humidity." In weather-related datasets, these variables are correlated, where one variable can be highly related to another (e.g., temperature variables might be strongly correlated). This correlation can cause numerical instability in standard linear regression, affecting the model's predictive performance. By penalizing large coefficients, Ridge Regression controls the influence of correlated predictors and prevents them from dominating the model while also stabilizing the coefficient estimates. This regularization reduces the model's sensitivity to variations in the data because it improves its generalization capability.

The code instantiates a Ridge model with an alpha value of 0.1, and we chose key predictors of temperature metrics and humidity to influence the target variable. The function, 'create_predictions,' trains the Ridge Regression model on the training data from '2023-09-01 00:00:00' to '2023-11-28 00:00:00' and generates predictions for the subsequent two days of hourly test data. It then calculates the mean absolute error between the predicted and actual target values on the test set, offering insights into the model's predictive accuracy.

```
from sklearn.linear_model import Ridge

reg = Ridge(alpha=0.1)

predictors = ["temperature", "max_temperature", "min_temperature", "humidity"]
```

```
train = df.loc[:"2023-11-28 00:00:00"]
test = df.loc["2023-11-28 01:00:00":]

def create_predictions(predictors, df, reg):
   train = df.loc[:"2023-11-28 00:00:00"]
   test = df.loc["2023-11-28 01:00:00":]
   reg.fit(train[predictors], train["target"])
   predictions = reg.predict(test[predictors])
   error = mean_absolute_error(test["target"], predictions)
   combined = pd.concat([test["target"], pd.Series(predictions, index=test.index)], axis=1)
   combined.columns = ["actual", "predictions"]
   return error, combined
```

This method was developed by splitting the dataset into training and test datasets. The model was trained on the training set using our chosen predictors and the target variable. The performance of the model was then evaluated using the mean squared error. The Ridge Regression hyperparameter (alpha = .1) controls the strength of the of the regularization. Additionally, the use of a grid search optimization technique was used to find the optimal alpha value that minimized the error.

```
from sklearn.model_selection import GridSearchCV
import numpy as np

alpha_range = np.linspace(0.1, 10, 20)
param_grid = {'alpha': alpha_range}

ridge = Ridge()
grid = GridSearchCV(estimator=ridge, param_grid=param_grid, scoring='neg_mean_absolute_error', cv=5)
X = df[predictors]
y = df["target"]
grid.fit(X, y)
```

```
best_alpha = grid.best_params_['alpha']
best_error = -grid.best_score_

print(f"Best alpha value: {best_alpha}")
print(f"Mean Absolute Error with best alpha: {best_error}")
```

Results

Figure 2

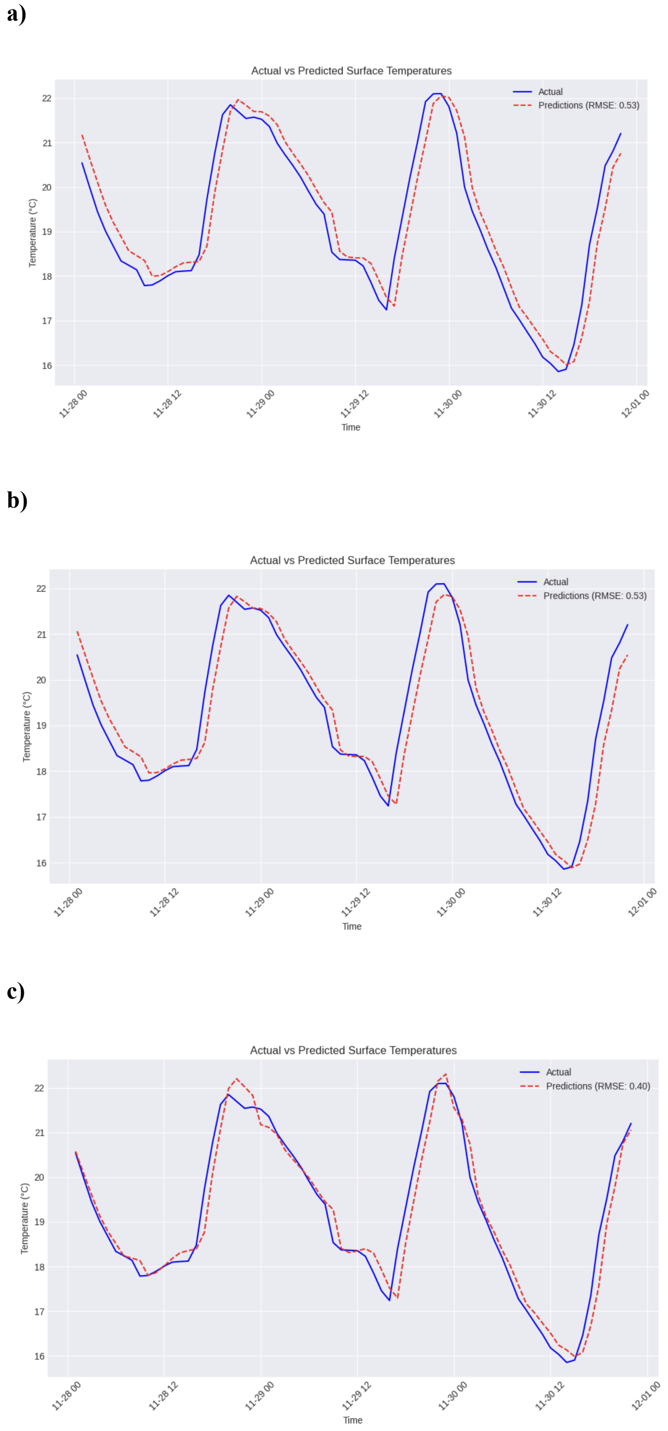

Figure 2a, 2b, 2c: Comparison between the Actual and Predicted Temperature Values over the test period of November 28th through November 30th. The blue line represents the actual temperature values, while the red dashed line represents the predicted values obtained from the model. The plot showcases the model's performance in approximating the true temperature trend over the given time period. Plot 'a' depicts the initial prediction with predictors "temperature", "max_temperature", *"min_temperature", and "humidity" [2a]. Plot 'b' depicts the second interaction of the prediction with predictors "temperature", "max_temperature", "min_temperature", "day_max", "month_day_max", "humidity", and "max_min" [2b]. Plot 'c' depicts the final prediction with predictors "temperature", "max_temperature", "min_temperature", "day_max", "month_day_max", "hourly_avg", "day_of_the_month", and "max_min"[2c].*

Upon visual inspection, the model seems to capture the overall trend of temperature fluctuations; however, there are instances where it slightly diverges from the actual values [4a]. The three different runs (4a, 4b, 4c) show 3 subsequent iterations of the model after additional predictors. Upon the addition of further predictors, the model aligns with the 'actual' data more accurately [4c]. As the predicted values align closer, the RMSE is decreased from 0.53 to 0.4. This is a successful minimization of the RMSE. By looking at the feature correlation with the target variable we can also see quantitatively the strength of their relationship with the target and predictor variables.

```
          Correlation with Target:
temperature                      0.946845
humidity                        -0.024183
min_temperature                  0.556135
max_temperature                  0.994025
target                           1.000000
day_max                          0.927157
month_day_max                   -0.363440
max_min                          0.887094
hourly_avg                       0.664312
day_of_the_month                 0.938830
```

Figure 3

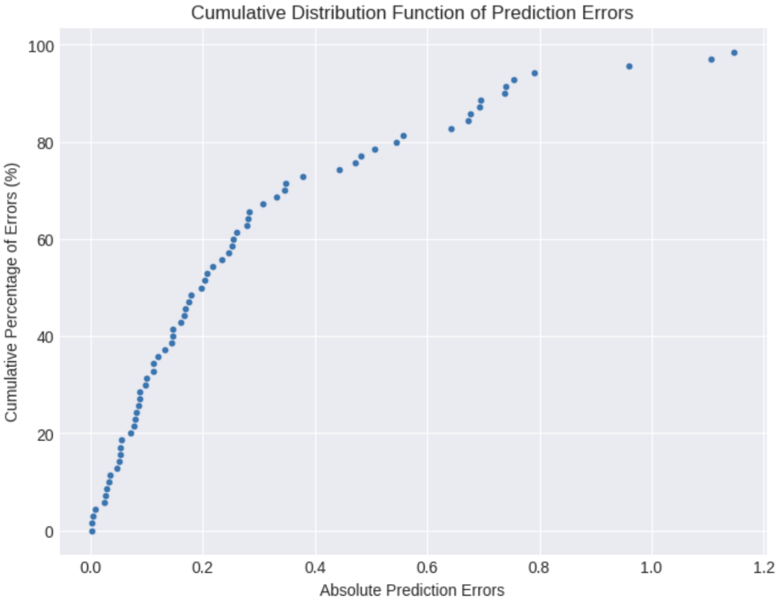

Figure 3: Plot showing the Cumulative Distribution Function (CDF) of prediction errors generated by the model. The x-axis represents the magnitude of prediction errors, while the y-axis displays the cumulative percentage of errors. [3].

The cumulative distribution function (CDF) shows the prediction of errors obtained from the model's forecasts against actual values [3] (Brownlee, J. 2019). Each point on the plot represents a prediction error, sorted in ascending order from the smallest to the largest error. The x-axis represents the magnitude of prediction errors, while the y-axis indicates the cumulative percentage of errors. The steeper the curve, the more accurate the predictions. This graph gives insight into the overall performance of the model. As the curve is in the middle of the graph, we can see that the model's prediction accuracy was not perfect, but it had minimal errors, and the majority of prediction errors were minimal [3].

Figure 4

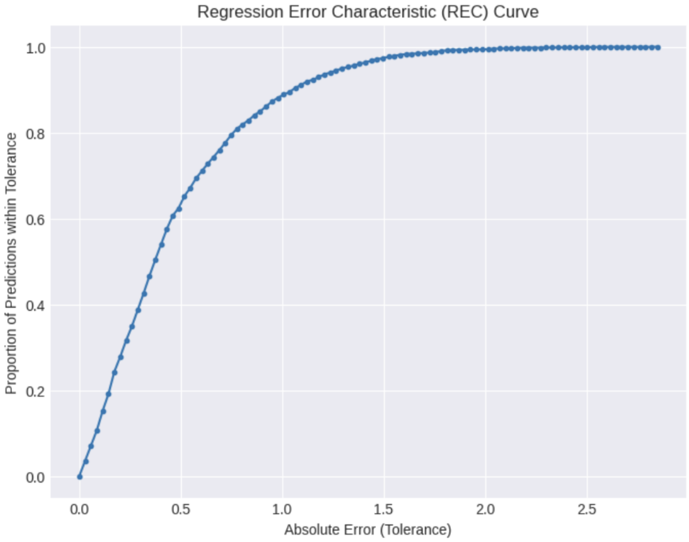

Figure 4: The Regression Error Characteristic (REC) Curve illustrates the proportion of predictions within varying absolute error tolerances for a Ridge Regression model trained on 'temperature', 'max_temperature', 'min_temperature', and 'humidity' predictors. The x-axis displays absolute error (tolerance), while the y-axis represents the proportion of predictions within the corresponding tolerance. [4].

This curve visualizes the model's predictive performance across different error thresholds, providing insights into its accuracy and tolerance levels [4]. The Regression Error Characteristic (REC) Curve is a visualization that assesses the predictive performance of the Ridge Regression model based on 'temperature', 'max_temperature', 'min_temperature', and 'humidity' predictors [4]. This curve is crucial as it showcases the model's accuracy at varying absolute error tolerances. The REC Curve aids in setting practical expectations for the model's performance in real-world applications and lets users make informed decisions based on the model's accuracy under different error constraints.

Discussion

The use of Ridge Regression for temperature predictions alongside comprehensive feature engineering gave us substantial insights. To begin, Figure 1 illustrated the temporal trends in average temperature and reflected the hourly fluctuations in the data. The presence of gaps in the time series data, notably on October 31st, showed the necessity of careful data handling strategies to mitigate the impact of missing values in predictive modeling, as highlighted in Figures 1a, 1b, and 1c. By substituting null values with the most recent non-null observation, the analysis ensured continuity which was a crucial step in preprocessing the time series.

Modeling approaches like Ridge Regression, as elaborated in Figure 2, play a fundamental role in temperature prediction endeavors. The deployment of predictors such as "temperature," "max_temperature," "min_temperature," and "humidity" enabled initial predictions, indicating the model's capability to capture temperature fluctuations. The next iterations, incorporating additional predictors such as "day_max," "month_day_max," "hourly_avg," and "day_of_the_month," reflected an improvement in alignment between predicted and actual values, effectively minimizing the Root Mean Squared Error (RMSE). The addition of features in the predictors variable improves the model adaptability and refined temperature forecasts, as evident in Figures 2a, 2b, and 2c.

The Cumulative Distribution Function (CDF) in Figure 3 provided a comprehensive overview of prediction errors, showcasing the model's accuracy distribution and emphasizing its proficiency in minimizing errors. Meanwhile, the Regression Error Characteristic (REC) Curve in Figure 4 delineated the model's predictive accuracy across varying absolute error thresholds and gave us insights into its performance under different tolerance levels. These figures highlighted the importance of visually seeing the model’s performance.

Conclusion

This study delved into the wide realm of weather forecasting, specifically focusing on predicting hourly maximum temperatures using the ERA5 reanalysis dataset and Ridge Regression. Key conclusions were found through the exploration of it features. Our application of Ridge Regression, fortified by feature engineering and regularization techniques, performed well and showed a tangible reduction in mean absolute error. This affirmed the role of precise predictors and temperature predictions and highlighted the significance of leveraging supervised learning methodologies to refine the temperature forecast. Thus, although the model performed adequately, it could have been improved with further variables.

Our analyses also emphasized the pivotal role of visualizations, specifically the Cumulative Distribution Function (CDF) and Regression Error Characteristic (REC) Curve, in assessing model accuracy. These graphical representations gave us visual insights into prediction errors, tolerance levels, and model performance across varying thresholds, empowering stakeholders with actionable insights for decision-making in weather-sensitive domains.

To advance this work further, future projects could explore advanced machine learning algorithms beyond Ridge Regression, evaluating ensemble methods or neural networks to discern their efficacy in weather prediction. The incorporation of additional meteorological parameters could also enrich the model's predictive capability.

In summary, this project showed the potential of machine learning, exemplified by Ridge Regression, in enhancing weather forecasting precision. It taught us the importance of practicing organized coding and always exploring other ways to improve our model. In addition to instilling machine learning-like thinking, the project outlined avenues for future research. It emphasized the importance of continuous innovation to improve predictive accuracy and showed us that we can always do more to improve our model.

Works Cited

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. (2023): ERA5 hourly data on pressure levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.bd0915c6 (Accessed on 11-29-2023)

Brownlee, J. (2019, September 24). *Continuous probability distributions for machine learning*. MachineLearningMastery.com. https://machinelearningmastery.com/continuous-probability-distributions-for-machine-learning/#:~:text=CDF%3A%20Cumulative%20Distribution%20Function%2C%20returns,equal%20to%20the%20given%20probability.