KOMPALA SAI CHARAN
2327132

DOMAIN-SPECIFIC MODEL BUILDING-WEATHER

CIA-I:MLA

K.SAI CHARAN

2327132

1. Business Understanding

a. Problem Identification:

- **Problem Statement:**
  - How can we predict temperature based on geographical and meteorological variables?
  - What factors influence temperature variations in a specific region?\

Data dictionary

Variable Name	Description	Data Type	Additional Notes
STATION	Station code identifying the weather station	Text	
DATE	Date of the weather observation	Date	Format: YYYY-MM-DD
LATITUDE	Latitude of the weather station	Numeric	Decimal degrees
LONGITUDE	Longitude of the weather station	Numeric	Decimal degrees
ELEVATION	Elevation of the weather station	Numeric	Meters above sea level
NAME	Name of the weather station	Text	
TEMP	Temperature recorded	Numeric	Celsius
TEMP_ATTRIBUTES	Attributes related to temperature measurement	Text	Flags, quality indicators, etc.
DEWP	Dew point temperature	Numeric	Celsius
DEWP_ATTRIBUTES	Attributes related to dew point measurement	Text	Flags, quality indicators, etc.
SLP	Sea level pressure	Numeric	Millibars
SLP_ATTRIBUTES	Attributes related to sea level pressure measurement	Text	Flags, quality indicators, etc.
STP	Station pressure	Numeric	Millibars
STP_ATTRIBUTES	Attributes related to station pressure measurement	Text	Flags, quality indicators, etc.
VISIB	Visibility	Numeric	Miles
VISIB_ATTRIBUTES	Attributes related to visibility measurement	Text	Flags, quality indicators, etc.
WDSP	Wind speed	Numeric	Knots
WDSP_ATTRIBUTES	Attributes related to wind speed measurement	Text	Flags, quality indicators, etc.
MXSPD	Maximum sustained wind speed	Numeric	Knots
GUST	Wind gust speed	Numeric	Knots
MAX	Maximum temperature	Numeric	Celsius
MAX_ATTRIBUTES	Attributes related to maximum temperature	Text	Flags, quality indicators, etc.
MIN	Minimum temperature	Numeric	Celsius
MIN_ATTRIBUTES	Attributes related to minimum temperature	Text	Flags, quality indicators, etc.
PRCP	Precipitation	Numeric	Inches
PRCP_ATTRIBUTES	Attributes related to precipitation measurement	Text	Flags, quality indicators, etc.
SNDP	Snow depth	Numeric	Inches
FRSHTT	Weather event indicators	Integer	Flags indicating various weather phenomena

b. Variables:

- The variable you're trying to forecast or explain is known as the target variable. It's probably TEMP (temperature) in your situation.
-
- Predictor variables are those that have the potential to affect or clarify changes in the target variable. Potential predictor variables based on your dataset could include:

- LATITUDE: Latitude of the weather station
- LONGITUDE: Longitude of the weather station
- ELEVATION: Elevation of the weather station
- DEWP: Dew point temperature
- SLP: Sea level pressure
- STP: Station pressure
- VISIB: Visibility
- WDSP: Wind speed
- MXSPD: Maximum sustained wind speed
- GUST: Wind gust speed
- PRCP: Precipitation
- SNDP: Snow depth

output

```
> sum(is.na(weather_data))
[1] 3475
> # Summary statistics
> summary(weather_data)
    STATION             DATE                          LATITUDE       LONGITUDE
 Min.   :4.218e+10   Min.   :2023-01-01 00:00:00.00   Min.   :13.20   Min.   :72.87
 1st Qu.:4.251e+10   1st Qu.:2023-03-27 12:00:00.00   1st Qu.:17.45   1st Qu.:77.70
 Median :4.281e+10   Median :2023-06-25 00:00:00.00   Median :21.09   Median :79.05
 Mean   :4.278e+10   Mean   :2023-06-28 01:15:46.01   Mean   :20.88   Mean   :81.72
 3rd Qu.:4.300e+10   3rd Qu.:2023-09-28 00:00:00.00   3rd Qu.:25.70   3rd Qu.:88.45
 Max.   :4.313e+10   Max.   :2023-12-31 00:00:00.00   Max.   :28.57   Max.   :91.98
  ELEVATION          NAME                TEMP        TEMP_ATTRIBUTES      DEWP
 Min.   :  4.87   Length:3383        Min.   : 49.2   Min.   : 4.00   Min.   :27.60
 1st Qu.: 49.37   Class :character   1st Qu.: 74.7   1st Qu.:16.00   1st Qu.:58.60
 Median :314.85   Mode  :character   Median : 80.2   Median :24.00   Median :66.30
 Mean   :382.94                      Mean   : 79.7   Mean   :19.09   Mean   :65.53
 3rd Qu.:617.00                      3rd Qu.: 84.7   3rd Qu.:24.00   3rd Qu.:73.40
 Max.   :915.00                      Max.   :108.2   Max.   :24.00   Max.   :88.00
 DEWP_ATTRIBUTES      SLP           SLP_ATTRIBUTES        STP         STP_ATTRIBUTES
 Min.   : 4.00   Min.   :  996.6   Min.   :0.0000   Min.   :941.2   Min.   :0.0000
 1st Qu.:16.00   1st Qu.: 9999.9   1st Qu.:0.0000   1st Qu.:999.9   1st Qu.:0.0000
 Median :24.00   Median : 9999.9   Median :0.0000   Median :999.9   Median :0.0000
 Mean   :19.09   Mean   : 8955.5   Mean   :0.4851   Mean   :998.8   Mean   :0.1008
 3rd Qu.:24.00   3rd Qu.: 9999.9   3rd Qu.:0.0000   3rd Qu.:999.9   3rd Qu.:0.0000
 Max.   :24.00   Max.   : 9999.9   Max.   :8.0000   Max.   :999.9   Max.   :8.0000
     VISIB       VISIB_ATTRIBUTES      WDSP         WDSP_ATTRIBUTES     MXSPD
 Min.   :0.100   Min.   : 4.00    Min.   :  0.200   Min.   : 0.00   Min.   :  1.00
 1st Qu.:1.900   1st Qu.:16.00    1st Qu.:  3.300   1st Qu.:16.00   1st Qu.:  7.00
 Median :2.300   Median :24.00    Median :  4.800   Median :24.00   Median :  9.90
 Mean   :2.539   Mean   :19.09    Mean   :  5.358   Mean   :19.08   Mean   : 10.57
 3rd Qu.:3.300   3rd Qu.:24.00    3rd Qu.:  6.500   3rd Qu.:24.00   3rd Qu.: 12.00
 Max.   :5.400   Max.   :24.00    Max.   :999.900   Max.   :24.00   Max.   :999.90
      GUST            MAX          MAX_ATTRIBUTES          MIN
 Min.   : 15.0   Min.   :   54.50   Length:3383        Min.   :   34.20
 1st Qu.:999.9   1st Qu.:   84.20   Class :character   1st Qu.:   62.20
 Median :999.9   Median :   88.50   Mode  :character   Median :   69.80
 Mean   :858.3   Mean   :   91.53                      Mean   :   71.31
 3rd Qu.:999.9   3rd Qu.:   93.60                      3rd Qu.:   75.40
 Max.   :999.9   Max.   : 9999.90                      Max.   : 9999.90
 MIN_ATTRIBUTES          PRCP          PRCP_ATTRIBUTES          SNDP           FRSHTT
 Length:3383        Min.   : 0.000   Length:3383        Min.   :999.9   Min.   :     0
 Class :character   1st Qu.: 0.000   Class :character   1st Qu.:999.9   1st Qu.:     0
 Mode  :character   Median : 0.000   Mode  :character   Median :999.9   Median :     0
                    Mean   : 9.315                      Mean   :999.9   Mean   :  9588
                    3rd Qu.: 0.080                      3rd Qu.:999.9   3rd Qu.: 10000
                    Max.   :99.990                      Max.   :999.9   Max.   :110010

>
```

KOMPALA SAI CHARAN
2327132

Multilinear regression output

```
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.132e+02  1.499e+01   7.550 5.56e-14 ***
LATITUDE     9.550e-01  4.978e-02  19.183  < 2e-16 ***
LONGITUDE   -4.673e-01  2.674e-02 -17.473  < 2e-16 ***
ELEVATION   -8.058e-03  4.309e-04 -18.701  < 2e-16 ***
DEWP         3.371e-01  1.041e-02  32.391  < 2e-16 ***
SLP         -2.761e-06  4.418e-05  -0.063  0.95017
STP         -4.741e-02  1.493e-02  -3.175  0.00151 **
VISIB        5.468e+00  1.450e-01  37.721  < 2e-16 ***
WDSP         1.404e-02  6.592e-03   2.130  0.03321 *
MXSPD        8.603e-03  3.828e-03   2.248  0.02466 *
GUST        -8.807e-04  2.988e-04  -2.948  0.00322 **
PRCP        -1.619e-02  3.482e-03  -4.650 3.45e-06 ***
SNDP                NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.37 on 3371 degrees of freedom
Multiple R-squared:  0.5085,    Adjusted R-squared:  0.5069
F-statistic:   317 on 11 and 3371 DF,  p-value: < 2.2e-16

> print(paste("RMSE: ", rmse))
[1] "RMSE:  9.36878237767952"
>
```

Splitting of data 70% training 30 % testing output

```
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.198e+02  1.823e+01   6.574 6.00e-11 ***
LATITUDE     9.729e-01  6.097e-02  15.958  < 2e-16 ***
LONGITUDE   -4.496e-01  3.212e-02 -14.001  < 2e-16 ***
ELEVATION   -8.662e-03  5.203e-04 -16.648  < 2e-16 ***
DEWP         3.220e-01  1.284e-02  25.079  < 2e-16 ***
SLP         -4.992e-06  5.443e-05  -0.092  0.92693
STP         -5.639e-02  1.826e-02  -3.088  0.00204 **
VISIB        5.461e+00  1.785e-01  30.592  < 2e-16 ***
WDSP         2.638e-01  6.303e-02   4.186 2.95e-05 ***
MXSPD        6.954e-03  3.827e-03   1.817  0.06934 .
GUST        -2.562e-04  3.863e-04  -0.663  0.50728
PRCP        -1.839e-02  4.185e-03  -4.395 1.16e-05 ***
SNDP                NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.355 on 2357 degrees of freedom
Multiple R-squared:  0.5123,    Adjusted R-squared:   0.51
F-statistic: 225.1 on 11 and 2357 DF,  p-value: < 2.2e-16

> # Predict using the test data
> predictions <- predict(model, newdata = test_data)
> # Calculate RMSE (Root Mean Squared Error)
> rmse <- sqrt(mean((test_data$TEMP - predictions)^2))
> rmse
[1] 9.495621
```

KOMPALA SAI CHARAN
2327132

Lasso regression output

> lasso_model$lambda.min

[1] 0.007695422

> # Calculate RMSE (Root Mean Squared Error)

[1] 9.37286

Ridge regression output

> ridge_model$lambda.min

[1] 0.3254581

> rmse

[1] 9.368782

b. Model Output

Explanation of the Model Equation:

The model equation that results from your regression analysis shows how the predictor variables are used to estimate the target variable, or temperature, or TEMP. The equation in your situation is:

TEMP is defined as β0+β1×LATITUDE+β2×LONGITUDE+β3×ELEVATION+β4×DEWP+β5×SLP+β6×STP+β7×VISIB+β8×WDSP+β9×MXSPD+β10×GUST+β11×PRCP+ϵ\.text{TEMP} = 1 \times \beta_0 + \beta_1beta_2 \times \text{LONGITUDE} + \beta_3 \times \text{LATITUDE}beta_4 \times \text{DEWP} + \beta_5 \times \text{ELEVATION}text{SLP} + \times \beta_6text{STP} + **\beta_7** \times **VISIB} + **\beta_8** \times **WDSP} + **\beta_9** \times **MXSPD} + **\beta_{10} \times \text{GUST} + \times \beta_{11}\epsilon + text{PRCP}TEMP = β0 + β1 + β2 + LATITUDE + β3 + ELEVATION + β4 + DEWP + β5 + SLP + β6 + STP + β7 + VISIB + β8 + WDSP + β9 + MXSPD + β10 + GUST + β11 + PRCP + ϵ

- **Intercept (β0\beta_0β0)**: This is the estimated value of temperature when all predictor variables (LATITUDE,LONGITUDE,…,PRCP\text{LATITUDE}, \text{LONGITUDE}, \ldots, \text{PRCP}LATITUDE,LONGITUDE,…,PRCP) are zero. It represents the baseline temperature.
- **Coefficients (β1\beta_1β1 to β11\beta_{11}β11)**: These coefficients represent the estimated change in temperature for a one-unit change in each predictor variable, holding all other variables constant. For example:

  - β1×LATITUDE\beta_1 \times \text{LATITUDE}β1×LATITUDE: Indicates how much the temperature changes with each unit increase in latitude, assuming all other variables remain constant.
  - β2×LONGITUDE\beta_2 \times \text{LONGITUDE}β2×LONGITUDE: Shows the impact on temperature with changes in longitude.
  - Similarly, each coefficient (β3\beta_3β3 to β11\beta_{11}β11) corresponds to the respective predictor variable.
- **Error Term (ϵ\epsilonϵ)**: Represents the difference between the predicted temperature and the actual temperature, capturing factors not accounted for by the model.

Explanation of the Parameters:

Each parameter in the model equation serves a specific role:

- **Intercept**: Provides the baseline estimate for temperature, accounting for factors not included in the predictors.
- **Coefficients**: Indicate the magnitude and direction of the effect of each predictor variable on temperature. For instance:
  - Positive coefficients (e.g., LATITUDE, DEWP, VISIB) suggest an increase in temperature with an increase in these variables.
  - Negative coefficients (e.g., LONGITUDE, ELEVATION, STP, PRCP) indicate a decrease in temperature with an increase in these variables.

Explanation of the Coefficients:

Let's interpret the coefficients based on your model output:

- **LATITUDE (Coefficient: 0.955)**: A unit increase in latitude is associated with an average increase in temperature of approximately 0.955 degrees, holding other variables constant.
- **LONGITUDE (Coefficient: -0.467)**: Moving east or west by one unit of longitude is associated with an average decrease in temperature of approximately 0.467 degrees, assuming other factors remain constant.
- **ELEVATION (Coefficient: -0.00806)**: Increasing elevation by one unit results in an average decrease in temperature of approximately 0.00806 degrees, controlling for other variables.
- **DEWP (Coefficient: 0.337)**: An increase in dew point temperature by one unit leads to an average increase in temperature of approximately 0.337 degrees, all else being equal.
- **SLP (Coefficient: -2.761e-06)**: Sea level pressure shows no significant impact on temperature as indicated by its coefficient and p-value (0.95017).
- **STP (Coefficient: -0.0474)**: An increase in station pressure by one unit results in an average decrease in temperature of approximately 0.0474 degrees.
- **VISIB (Coefficient: 5.468)**: An increase in visibility by one unit corresponds to an average increase in temperature of approximately 5.468 degrees.

- **WDSP (Coefficient: 0.014)**: Wind speed shows a positive effect, where an increase in wind speed by one unit is associated with a slight increase in temperature (0.014 degrees).
- **MXSPD (Coefficient: 0.0086)**: Maximum sustained wind speed has a similar positive effect on temperature, with an increase of 0.0086 degrees per unit increase in maximum speed.
- **GUST (Coefficient: -0.0008807)**: Wind gust speed shows a negative impact on temperature, where an increase in gust speed by one unit leads to a decrease in temperature of approximately 0.0008807 degrees.
- **PRCP (Coefficient: -0.0162)**: Precipitation (PRCP) has a negative impact on temperature, where an increase in precipitation by one unit results in a decrease in temperature of approximately 0.0162 degrees.
- **SNDP (Coefficient: NA)**: Snow depth is marked as "NA", indicating missing data or insufficient observations for this variable in the model.

Model Fit Indices:

- **Multiple R-squared**: This is 0.5085, indicating that approximately 50.85% of the variability in temperature can be explained by the model's predictors.
- **Adjusted R-squared**: Adjusts the R-squared value for the number of predictors in the model, providing a more conservative estimate of model fit (Adjusted R-squared: 0.5069).
- **Residual Standard Error (RSE)**: This is 5.37, representing the average amount that actual temperatures deviate from the predicted values by the model.

Any Other Method of Model Identification/Development:

- **Regularization (Lasso and Ridge Regression)**: These techniques were employed to address multicollinearity and to select variables based on their impact on temperature. Lasso and Ridge regressions use penalty terms (lambda values) to shrink coefficients, with different lambda values influencing model selection and performance.

c. Model Interpretation from the Business Point of View

- **Geographical and Meteorological Insights**:
  - **Geographical Variables (LATITUDE, LONGITUDE, ELEVATION)**: Businesses can use these variables to predict temperature changes across different locations. For instance, higher latitudes generally correlate with cooler temperatures, while higher elevations typically lead to lower temperatures.
  - **Meteorological Factors (DEWP, PRCP, VISIB, WDSP, MXSPD, GUST)**: These factors are crucial for industries dependent on weather conditions. For example, construction companies can anticipate temperature variations affecting project timelines, while agricultural sectors can predict optimal planting and harvesting periods based on temperature forecasts.

KOMPALA SAI CHARAN
2327132

5. Model Evaluation and Diagnostics

- **Root Mean Squared Error (RMSE)**: This metric measures the average magnitude of the errors in predicting temperature. Your RMSE values are approximately 9.37 for the training data and 9.50 for the testing data, indicating how well the model predicts temperature variations.
- **F-statistic and p-value**: These statistics assess the overall significance of the model and its variables. In your case, the F-statistic is significant (p-value < 2.2e-16), indicating that the model as a whole explains a significant amount of variance in temperature.
- **Coefficient Significance**: Evaluates the individual significance of each predictor variable in relation to temperature prediction. Significant coefficients (marked with asterisks) provide insights into which variables strongly influence temperature changes.

6. A Short Note on Democratizing the Solution

- **Accessibility**: Deploying the model through user-friendly interfaces or APIs allows stakeholders to access temperature predictions easily. This accessibility supports decision-making across various sectors, from agriculture to transportation.
- **Interpretability**: Providing clear explanations of model outputs ensures that non-technical users understand temperature forecasts and their implications. This clarity promotes informed decision-making within organizations.
- **Scalability**: Ensuring the model can handle new data updates and maintain accuracy over time is crucial for sustaining its usefulness. Regular updates and monitoring ensure that the model remains reliable for ongoing operational and strategic planning.

Conclusion: Ridge Regression is best if you prioritize stable coefficient estimates and effective handling of multicollinearity. It maintains all variables in the model but penalizes large coefficients, making it suitable when interpretability of each predictor's impact is crucial.

- **Lasso Regression** is ideal if you need to perform feature selection and simplify the model by shrinking less relevant coefficients to zero. This approach helps in identifying the most influential variables but may overlook marginal predictors that still contribute to model accuracy.

For most applications focused on robust prediction and interpretability of all included variables, **Ridge Regression** is typically the preferred choice. It strikes a balance between model complexity and performance, making it well-suited for understanding temperature variations based on diverse geographical and meteorological inputs.

.