# 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**



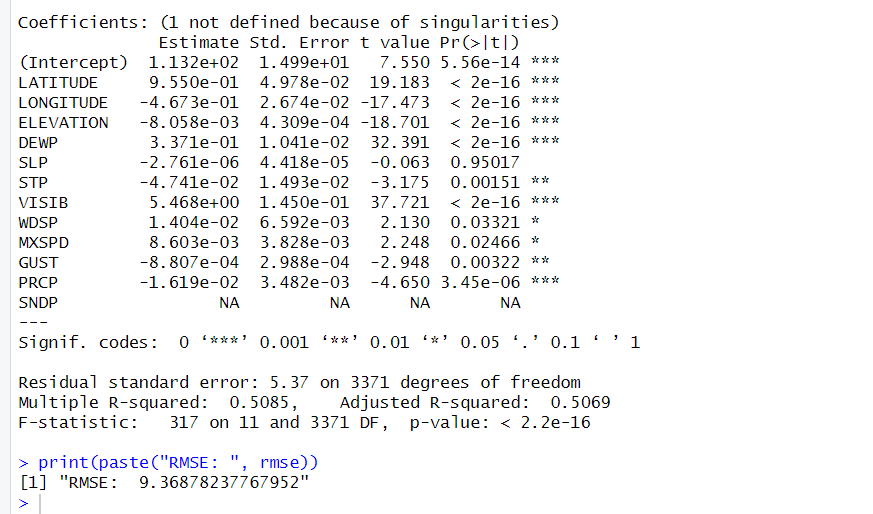
**b. Variables:**

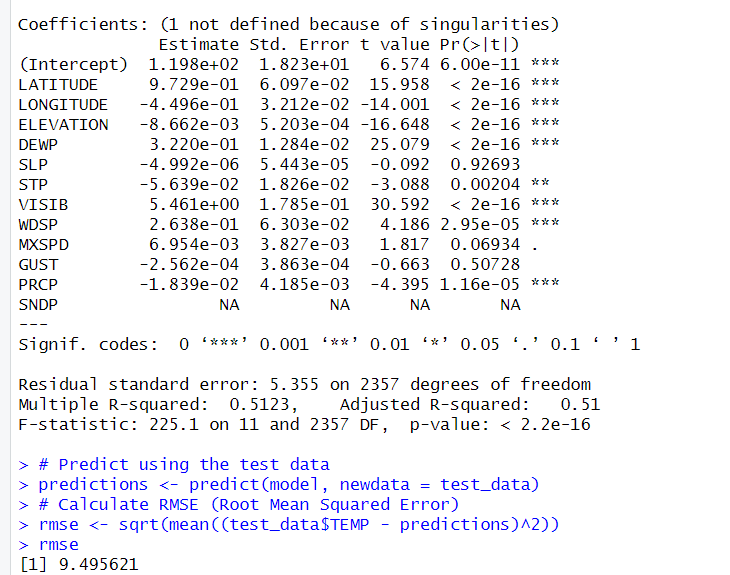
* The variable you're trying to forecast or explain is known as the target variable. It's probably TEMP (temperature) in your situation.
* Predictor variables are those that have the potential to affect or clarify changes in the target variable. Potential predictor variables based on your dataset couldinclude: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

# Multilinear regression output



Splitting of data 70% training 30 % testing output 

# 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.

**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.

.