Generalized linear model:

A generalized linear model (GLM) is used to forecast and model the RainTomorrow variable in the "weatherAUS" data set. There are five eigenvalues for the data sets: "MaxTemp," "MinTemp," "Rainfall," "RainToday," and "RainTomorrow." A GLM is a statistical model that simulates the relationship between response variables and predictor variables by combining linear regression with an appropriate link function and a predetermined distribution family.

1. Data preparation: Choose the "weatherAUS" data set's relevant feature variables and target variables, then perform data preprocessing such as handling missing values and changing data types.

2. Splitting the data set: First of all, we adopted a random partition method, dividing the original data set into training set, verification set and test set according to a certain proportion. Here, we divide the data set into 70% training set, 15% validation set, and 15% test set. When dividing data sets, it is necessary to maintain the consistency of sample category distribution to avoid category imbalance in the training and evaluation process.

3. Model construction: A generalized linear model is created using the 'glm()' function. Here, the other feature variables are utilized as the prediction variables, while the goal variable "RainTomorrow" is used as the response variable. By indicating the proper connection function and distribution family, the model type is established.

4. Model fitting: Model fitting and parameter estimates were performed using the GLM model that was built using the training set of data.

5. Model evaluation: Calculate the model's accuracy, sensitivity, and specificity as well as other evaluation indicators using the test set data.

6. Model prediction: To forecast fresh observation data and get prediction results, the trained model is used.

Here, we employ a binary generalized linear model to forecast the "RainTomorrow" variable, which accepts either "Yes" or "No". To resolve this binary issue, we employ the logit link function and the binomial family of distributions. Depending on the data collection and the particular challenge, the precise application of GLM can be modified and expanded. In order to anticipate the outcomes of the visualization, a mixed matrix was created and used to build the ROC curve.

Discuss and analyze the findings of the experiment:

The model's classification performance can be deemed to be good based on the AUC value of 0.766. The model can better distinguish between positive and negative data since AUC is near to 1.

Accuracy: 0.8044, indicating that the prediction accuracy of the model is about 80.44%.

Sensitivity: 0.2685, also known as recall rate or true positive rate, indicating that the model has the ability to correctly identify positive samples of about 26.85%.

Specificity: 0.9568, indicating that the ability of the model to correctly identify negative samples was about 95.68%.

The overall performance of the model is measured according to the accuracy rate. The accuracy rate of the model is about 80.44%, indicating that the model is relatively accurate for the classification prediction of samples on the whole. However, in terms of sensitivity and specificity, the ability of the model to identify positive samples was weak, while the ability to identify negative samples was strong.

Taken together, the model does a good job of predicting whether it will rain or not, but it leaves something to be desired when it comes to correctly detecting positive samples. At the same time, we also concluded that the temperature difference of the day will have some effect on the rainfall.



