STA303H1S/STA1002HS Final Project - Weather Prediction Study

Jeff Xu
1003342545

Introduction Section

The goal of the study is to construct a model based on the **weather** dataset, that can be used to acceptably predict if it will rain tomorrow, or equivalently the value of **RainTomorrow.** The model will include various weather characteristics as predictors that are influential. Precipitation prediction is important in many fields, for example, it can help to improve efficiency of planting and harvesting; logistics companies will be interested in knowing the weather to make sure shipments arrive on time.

Methods Section

Choice of Methods:

The response variable **RainTomorrow** is categorical and is a binary variable, its values can be indicated by "1" for "Yes" and "0" for "No". Then assuming that the probability of success(π) is the probability of **RainTomorrow** equaling "Yes" and is identical for any given day, we have independent trials, and thus Bernoulli trials. So, we can say that **RainTomorrow** has Bernoulli distribution. Then one can fit a GLM using logistic regression, given by

$$\log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{X\beta}$$

assuming evenly distributed variance and correct distribution of residuals. GLM extends linear regression model for response variables that do not follow a Normal distribution and thus will be suitable for **RainTomorrow**. Consider a GLMM,

$$\log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{X\beta} + \gamma_i, \qquad \gamma_i \sim Normal(0, \sigma_\gamma^2)$$

which is a GLM that also accounts for random effects. Consider the covariates in the dataset, it is possible that some of these weather characteristics are correlated, such as **Rainfall** and **MaxTemp** and **MinTemp**, etc. Therefore, a GLMM can also be suitable. To check if a GLMM is more preferable, need to determine the ICC=$\frac{\sigma_\gamma^2}{\sigma_\gamma^2+\sigma_\varepsilon^2}$ if it is close to zero then a GLM without the random effect is sufficient.

Variable Selection:

Given that there might exists multicollinearity among the predictors, want to select a subset of predictors that are best for the purpose of the study. AIC and BIC are criterions commonly used for variable selection; they both penalizes for the number of parameters. One would want to choose the model that gives the minimum AIC/BIC. BIC applies more penalty than AIC. It prefers simpler model which might result underfit. AIC on the other hand might result overfitting when sample size is small or when the number of parameters estimated is not small enough relative to the sample size. The model selection method that I prefer in this scenario is the AIC $\propto nlog\left(\frac{RSS}{n}\right) + 2p$, as BIC might overly narrow down the number of predictors. Considering our purpose is to make predictions, it is preferable to include some extra variables that can help to improve prediction accuracy.

Model Violations / Diagnostics:

After obtaining the model selected by AIC, one will need to first check if it fulfills the assumptions for a GLM/GLMM such as independency of outcomes, variance spread and correct distribution of residuals. Consider drawing two residual plots, with predictors/fitted values plotted against deviance residuals. These residual plots can be used to diagnostic the variance assumption. The assumptions are satisfied when the points vary evenly as the linear predictors/ fitted values vary. A Normal QQ-plot can be used to diagnostic the distribution of residuals. Data with underlying Normal distribution will have QQ-plot as straight line. Given that the response variable is categorical and binary, its QQ-plot cannot resemble the shape of Normal distribution data. Lastly a Half-Normal plot can be used to identify unusual observations or outliers, removing the outliers can potentially improve the model. To examine the goodness of fit. One can use cross validation to perform an internal validation. If the model is of good fit, the plot should resemble the shape of a straight line. Another method is to draw the ROC curve and obtain the AUC value. This gives us an insight on how well the model can successfully discriminate between events, and if the coefficient estimates are in good agreements with the true values of the coefficients. Lastly to check the predictive strength, use data points from year 2017 as the test set and obtain prediction error.

Results Section

Description of Data:

Looking at the dataset, notice that the values of covariates **Evaporation** and **Sunshine** are mostly "NA", so we do not have enough data for these two covariates, it is better to omit them in the model. Also notice that there are many data points collected from the same location, and many data points from the same date. From Table 1 by reading the variable **RainTomorrow,** one can see that there are 31877 days that rained tomorrow which is less than days that did not rain the next day. And reading the **RainToday** variable there are 31880 days with rain and is less than days without rain. The counts for the two variables are very similar, this is expected since every day with rain is the tomorrow of the previous day. For other categorical covariates such as **WindGustDir, WindDir9am, WindDir3pm,** the number of data points from each category varies. This means that the direction of strongest gust varies, and the wind direction varies and are different for prior to 9am and prior to 3pm on the same day.

Table 1 Data Summary

	2013-03-01	2013-03-02	2013-03-03	2013-03-04	2013-03-05	2013-03-06	(Other)
Date	49	49	49	49	49	49	145166
	Canberra	Sydney	Adelaide	Brisbane	Darwin	Hobart	(Other)
Location	3436	3344	3193	3193	3193	3193	125908
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
MinTemp	-8.50	7.60	12.00	12.19	16.90	33.90	1485
MaxTemp	-4.80	17.90	22.60	23.22	28.20	48.10	1261
Rainfall	0.000	0.000	0.000	2.361	0.800	371.000	3261

	W	SE	N	SSE	E	Other	NA's
WindGustDir	9915	9418	9313	9216	9181	88091	10326
	N	SE	E	SSE	NW	Other	NA's
WindDir9am	11758	9287	9176	9112	8749	86812	10566
	SE	W	S	WSW	SSE	Other	NA's
WindDir3pm	10838	10110	9926	9518	9399	91441	4228
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
WindGustSpeed	6.00	31.00	39.00	40.03	48.00	135.00	10263
WindSpeed9am	0.00	7.00	7.00	14.04	19.00	130.00	1767
WindSpeed3pm	0.00	13.00	19.00	18.66	24.00	87.00	3062
Humidity9am	0.00	57.00	70.00	68.88	83.00	100.00	2654
Humidity3pm	0.00	37.00	52.00	51.54	66.00	100.00	4507
Pressure9am	980.5	1012.9	1017.6	1017.6	1022.4	1041.0	15065
Pressure3pm	977.1	1010.4	1015.2	1015.3	1020.0	1039.6	15028
Cloud9am	0.00	1.00	5.00	4.45	7.00	9.00	55888
Cloud3pm	0.00	2.00	5.00	4.51	7.00	9.00	59358
Temp9am	-7.20	12.30	16.70	16.99	21.60	40.20	1767
Temp3pm	-5.40	16.60	21.10	21.68	26.40	46.70	3609
	No	Yes	NA's				
RainToday	110319	31880	3261				
RainTomorrow	110316	31877	3267				

For the numerical covariates they are examined further through graphical summaries. From Figure 2-3 in Appendix, the plots tell us that the distribution of the covariates are similar on days with rain tomorrow and days without rain tomorrow. And there are more days without rain tomorrow than days that rained the next day. Note that here did not plot **Rainfall, Cloud9am,** and **Cloud3pm** since the range for these variables is relatively small so we know they vary little on different days. And for all weather characteristics collected on 9am and 3pm, I picked the one with fewer "NA".

Process of Obtaining Final Model:

First fit a logistic regression using all predictors except **Location** and **Date** since as previously said there are many data points with the same location and date. It is reasonable to expect the weather in the same location or on the same day to be very similar, so they might cause some mixed effect which will be later checked using a GLMM. Then use AIC to do variable selection, the selected predictors are **Rainfall, WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am, Temp3pm, RainToday**. Now need to check if **Location** should be added as random effect. Obtained from R code result, after adding **Location** as random effect, the ICC is given by 0.332705/ 0.9754563 = 0.3410763. ICC is close to 0 so can conclude that GLM is sufficient. After adding **Date** as random effect ICC is 0.7307457/0.9058693 = 0.806679, which is not quite close to 0, but the model has many predictors adding random effect will significantly complicate model fitting. Therefore, the model selected by AIC is the final model.

Goodness of Final Model:

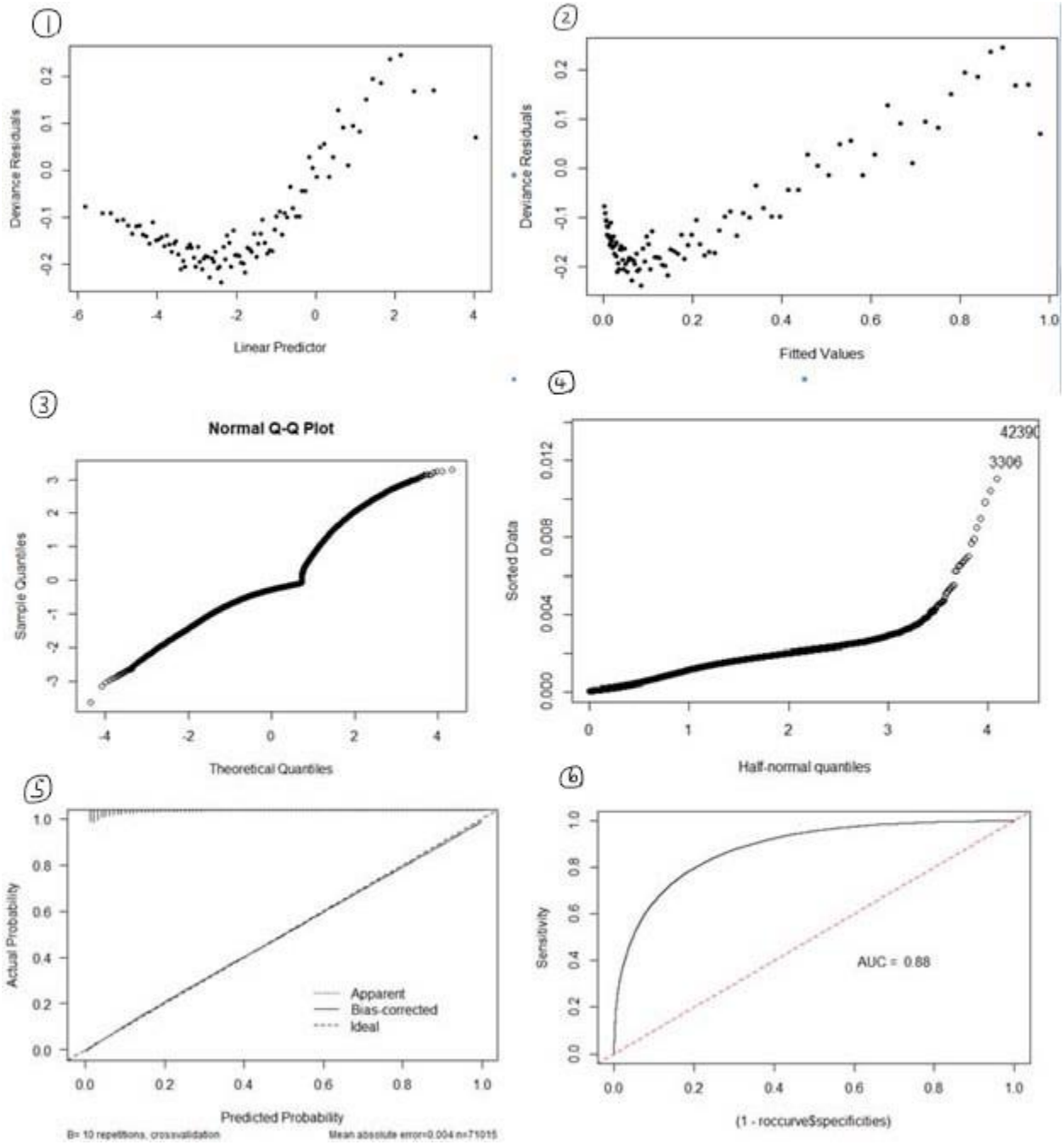

Figure 4 Residual plots, Normal QQ-plot, Half-Normal plot, Internal validation, ROC curve.

To check the model assumption, looking at Figure 4 plot 1-2 given above, we can observe an even variation of deviance residuals as linear predictor/fitted values increase. So, the variance assumption is satisfied. Looking at the plot 3, the Normal QQ-plot is not a straight line, so variance distribution assumption satisfied. And the Half-Normal plot (plot 4) suggests that there two outliers, however, since the sample size is very large and the two points are not very extreme, it's acceptable to keep them in the model. After performing the internal validation, plot 5 shows

that actual probability overlaps predicted probability fairly well, so the model is of good fit. The ROC curve in plot 6 gives the AUC value of 0.88, which implies that the model can correctly discriminate between the events 88% of the times. This can be judged to be an acceptable accuracy. Therefore, from the above analysis, it is safe to conclude that the model assumptions are satisfied, and the final model is of good fit for the dataset.

Discussion Section

Final Model Interpretation and Importance:

The following predictors are significant, tested at the 90% confidence level. For categorical predictors, their exp(coefficient) is the odds ratio of raining tomorrow given its category against other categories. For example, if **RainToday = Yes,** the odds of raining tomorrow is exp(0.490359)= 1.632902 times more likely than when **RainToday = No**. For numerical predictors, their exp(coefficient) is the odds ratio of raining tomorrow when predictor increased by 1. For example, when **WindGustSpeed** increase by 1, the odds of raining tomorrow is exp(0.059493)= 1.061298 times more likely than original.

Table 5 Significant Predictors Coefficient Estimates and Odds Ratio

Predictor	**Odds Ratio**
Rainfall	exp(0.007568)
WindGustDirS	exp(0.209643)
WindGustDirSE	exp(0.235682)
WindGustDirSSE	exp(0.285398)
WindGustDirSSW	exp(0.185539)
WindGustDirSW	exp(0.266169)
WindGustDirWSW	exp(0.247735)
WindGustSpeed	exp(0.059493)
WindDir9amENE	exp(0.295715)
WindDir9amESE	exp(-0.175927)
WindDir9amN	exp(0.223178)
WindDir9amNE	exp(0.346472)
WindDir9amNNE	exp(0.542990)
WindDir9amS	exp(-0.274270)
WindDir9amSE	exp(-0.230342)
WindDir9amSSE	exp(-0.191017)
WindDir9amSSW	exp(-0.192991)
WindDir3pmNNW	exp(0.293144)
WindDir3pmNW	exp(0.250084)
WindDir3pmSW	exp(-0.206883)
WindDir3pmWNW	exp(0.229539)
WindSpeed9am	exp(-0.014514)
WindSpeed3pm	exp(-0.031958)
Humidity3pm	exp(0.059801)

Pressure9am	exp(0.167522)
Pressure3pm	exp(-0.221208)
Cloud9am	exp(0.058178)
Cloud3pm	exp(0.173021)
Temp9am	exp(0.026259)
Temp3pm	exp(-0.027993)
RainTodayYes	exp(0.490359)

Fitting predictors from the test set into the final model to obtain the predicted probabilities. These probabilities will range from 0 to 1, the convention being used is treat any predicted probability below 0.5 as 0, otherwise, treat as 1. Compare these predicted values with the observed values in the test set, the prediction error is 0.2333333. This prediction error is quite small; thus, the model can be used to predict the value for **RainTomorrow** with fair accuracy. If the predicted value is 1, the model predicts there will be rain tomorrow, otherwise, the model predicts there is no rain tomorrow.

Limitations of Analysis:

The GLM also has an independent outcome assumption. However, if there's rain on one day it is likely there's rain the next day, so the outcomes are not guaranteed to be independent, this potentially impacts the accuracy of the model. But since we also included **RainToday** as a predictor, this problem should be dealt with already. The model omitted the variables **Location** and **Date**. However, the ICC value suggests **Date** should be added as random effect, and a GLMM is complex but more ideal. Therefore, a GLMM might potentially improve prediction accuracy.

Appendix

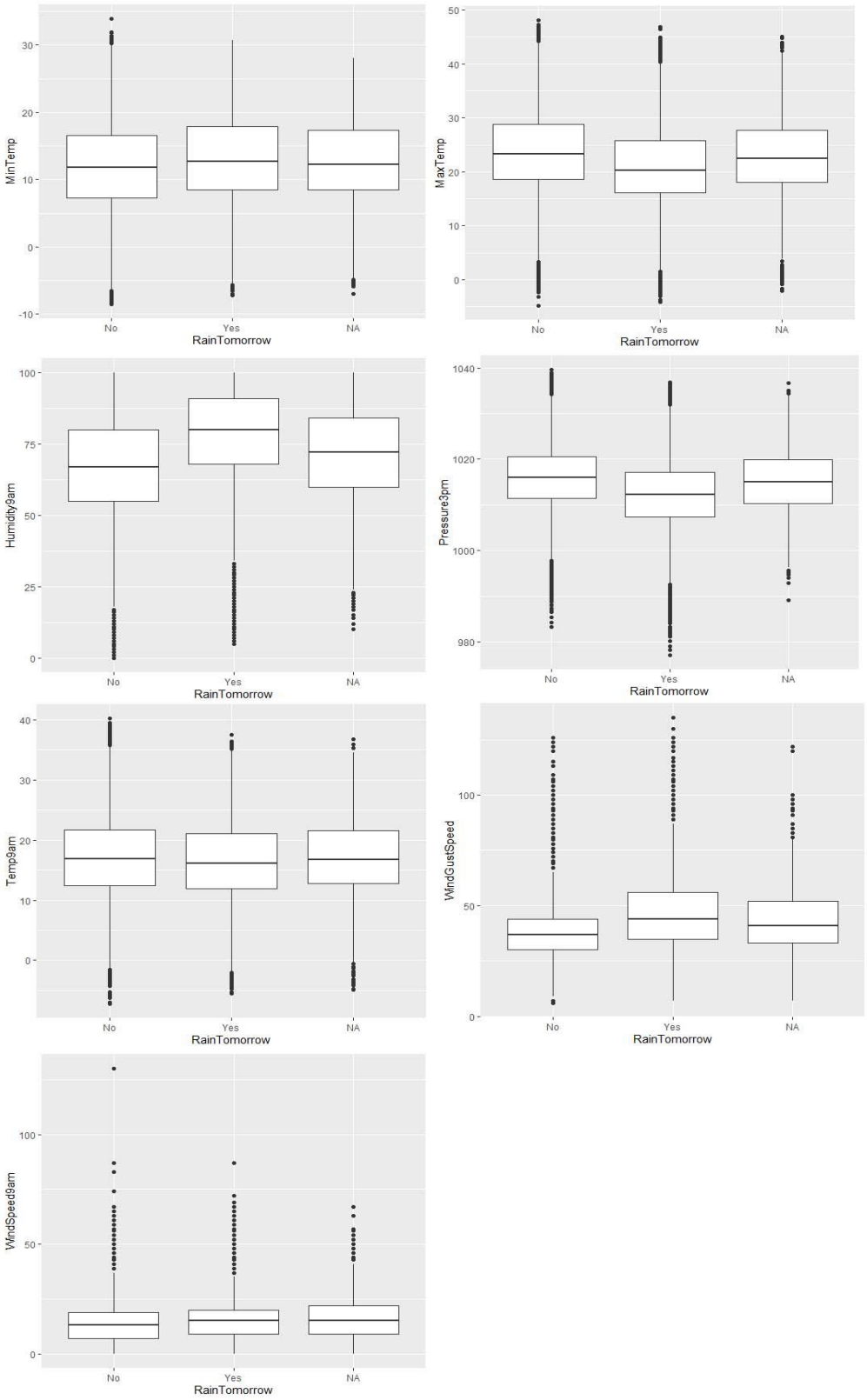

Figure 1 boxplot for numerical covariates

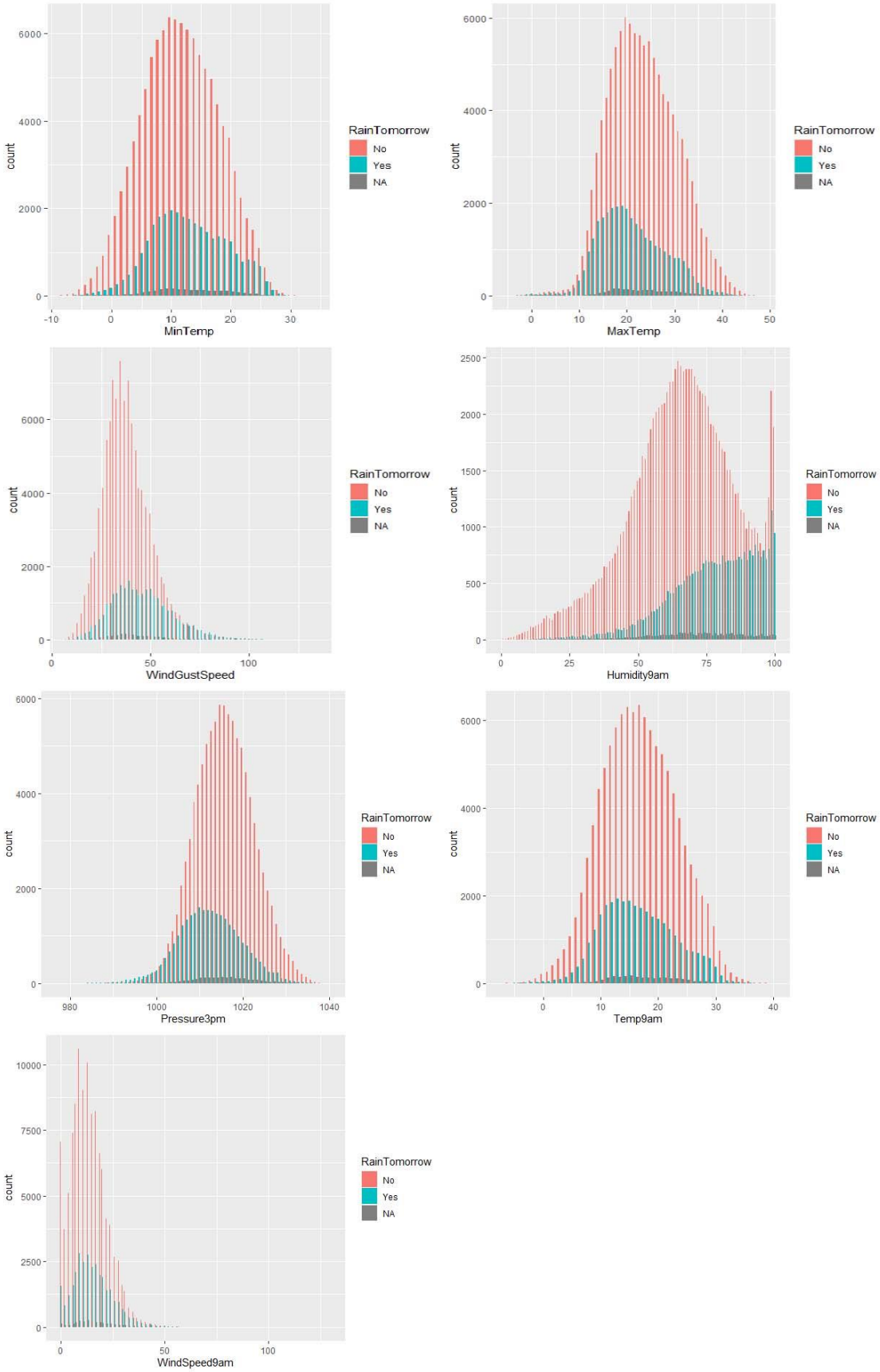

Figure 3 Histogram for numerical covariates