Will it rain tomorrow: Analysis on Australian Weather Data

Cong Liu

Professor George Stefan

STA303

28 August 2021

1 Introduction

Weather forecasting is the application of science and technology to predict the conditions of the atmosphere for a given location and time. Weather forecasts involve collecting quantitative data about the current state of the atmosphere, land, and ocean and using meteorology to project how the atmosphere will change at a given place (Wikipedia Contributors). The ultimate goal is to provide a reliable prediction about future weather within a given interval of time at a specific location (Cristani).

One of the most critical albeit basic weather question is “will it rain tomorrow”. The answer affects decisions range from the trivial ones: whether to dry clothes outside to the more influential decisions: airline flight schedules and agricultural production decisions. Rain affects many sectors of human life profoundly.

This report outlines an analysis conducted on a weather data set collected from 22 Australian cities over the time period of 2007-2017. The dataset offers measurements for 23 weather characteristics on a daily basis, including the essential ones like rainfall, wind speed and humidity. The response variable of interest is “RainTomorrow”, which takes on values of “Yes” if the rainfall next day is greater than one millimeter and “No”. The analysis involves identifying the statistically significant covariates that are associated with the response variable, determining how they affect the outcome and finally, establishing a model with a decent predictive ability that may answer the question “will it rain tomorrow”. The predictability of the fitted model is tested using the data from the entire year 2017.

2 Methods

2.1 Choice of Methods

Through the study in STA303 course, I have become equipped with theories of Generalized Linear Model (GLM) and Generalized Linear Mixed Model (GLMM) to answer questions. A GLM consists of two components: the response that belongs to the exponential family and a link function that describes how the mean of the response and a linear predictor are related (Faraway 126). Specifically for this weather dataset, the response variable "RainTomorrow" is a binary categorical variable and thus follows a Bernoulli distribution. Since Bernoulli is a member of the binomial family, the link function for GLM of the weather is $\log\left(\frac{\pi}{1-\pi}\right) = X\beta$, where π is the probability of raining tomorrow, X is the design matrix and β is the coefficient matrix.

However, given that many observations for this data are collected at the same location or same time, the observations are likely to be dependent from one another. The correlation in between the data points violates the assumption of GLM that the data is independent. To account for the correlation, I shall consider fitting a GLMM to the data. For a GLMM, the response variable still needs to be a member of the exponential family and a link function is required to connect the mean response with the linear predictor. The additional component of a GLMM link function, comparing to a GLM, is a random effect term μ_i that accounts for the randomness and follows a Normal distribution (i.e. ~ $N(0, \sigma_\alpha^2)$). Thus, the GLMM to be fitted for this dataset can be defined as $\log\left(\frac{\pi}{1-\pi}\right) = X\beta + \mu_i$. Although GLMM takes care of the potential random effect existing in the dataset, using this model could highly complex the computation and yields hard to interpret results. Thus, my final choice is to fit a GLM to the data and then compute the intraclass correlation coefficient (ICC)$= \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2}$ to decide whether a fitting a GLMM is necessary.

2.2 Variable Selection

Since the observations in the weather data are recorded at the same city or at the same date, there may exist a random effect. To prevent the accuracy of the fitted model from being affected by the potential random effect, the initial model is fitted using all variables in the training except "Location" and "Date" which are the major source of inter-correlation.

I choose to use the Bayesian Information Criterion (BIC) as criteria of the variable selection procedure due to its computational convenience and efficiency when handling large size data. BIC imposes a penalty term of $qlog(n)$ where q is the number of model parameters and n is the sample size. In a backward stepwise selection, variables are dropped one by one from the initial model as long as dropping results in a lower BIC.

2.3 Model Diagnostics

Since the response variable of interest follows a Bernoulli distribution, the Normal QQ plot would show residuals heavily deviating from the theoretical quantiles. Thus, a half-normal plot is used instead to check the distribution of the residuals and to evaluate the model validity. Cross validation technique is also used to check how the predicted probabilities are aligned with the expected value. ROC curve and area under the ROC curve (AUC) is calculated to further assess the discriminative ability of the fitted model.

3 Results

3.1 Description of Data

Table 1: Numerical summary of variables used in the final model from the weather data

	Rainfall	WindGust Speed	WindSpeed 9am	WindSpeed 3pm	Humidity3pm
Minimum	0.00	6.00	0.00	0.00	0.00
First Quantile	0.00	31.00	7.00	13.00	37.00
Median	0.00	39.00	13.00	19.00	52.00
Mean	2.36	40.03	14.05	18.66	51.54
Third Quantile	0.80	48.00	19.00	24.00	66.00
Maximum	371.00	135.00	130.00	87.00	100.00

Table 1 cont'd

	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm
Minimum	980.5	977.1	0.00	0.00	-7.20	-5.40
First Quantile	1012.9	1010.4	1.00	2.00	12.30	16.60
Median	1017.6	1015.2	5.00	5.00	16.70	21.10
Mean	1017.7	1015.3	4.45	4.51	16.99	21.68
Third Quantile	1022.4	1020.0	7.00	7.00	21.60	26.40
Maximum	1041.0	1039.6	9.00	9.00	40.20	46.70

Table 1 cont'd

	WindDir9am	RainToday
Count	145285	145285
Unique values	16	2
Top frequent value	N	No
Frequency	11744	110195

From life experience, high humidity often precedes a rainy weather. Thus, I use "Humidty3pm" (i.e. the relative humidity at 3 p.m. in percentage) to create plots against RainTomorrow.

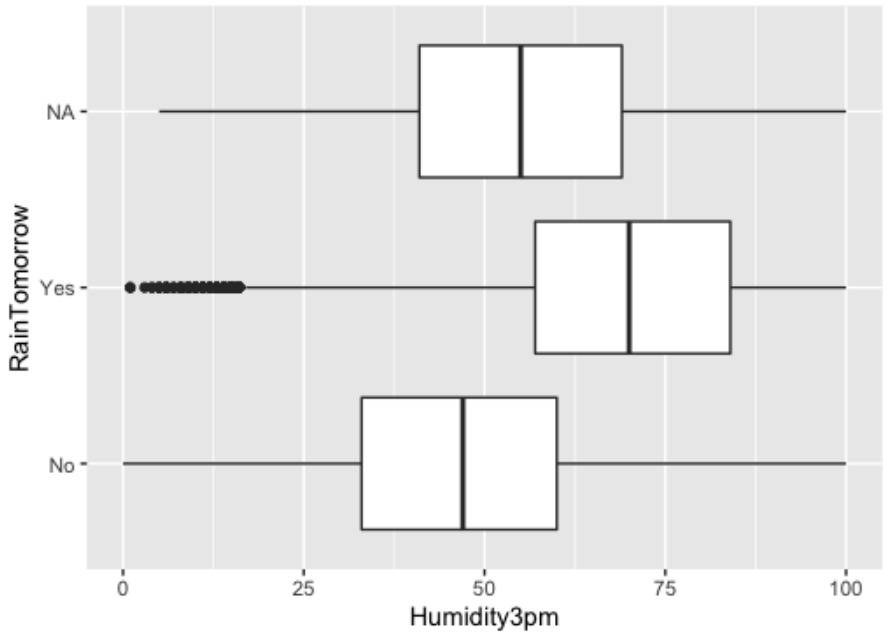

Figure 1: Boxplot of relative humidity at 3 p.m. across RainTomorrow

The boxplot above shows that the distribution of relative humidity for days that have a rain tomorrow is skewed to the left. In addition, the median relative humidity at 3 p.m. is clearly higher on days that would rain tomorrow. These two features in the plot confirm the common sense on forecasting a rain tomorrow.

3.2 Model Fitting Procedure

Since the majority of missing values occur in columns "Evaporation" and "Sunshine", I choose to drop the data entries with Evaporation and/or Sunshine missing. For better model building, the actual train set being used does not contain these two columns.

The initial model uses all variables in the training set, except location and date. The exclusion of location and date column is to prevent the random effect, as previously explained. The Akaike Information Criterion (AIC) value of the initial model fitted is 46978.

Attempting to reduce the model size, a variable selection procedure is conducted. The reduced and the final fitted model uses variables from a BIC based selection. The AIC value of the reduced model is 47070. The size of the model shrinks dramatically despite a slightly higher AIC.

Table 2: Coefficient estimates for significant variables in the reduced model

	Coefficient Estimate	Standard Error	z-value	p-value
Intercept	49.179447	2.192347	22.432	<2e-16
Rainfall	0.006704	0.001448	4.629	3.68e-06
WindGustSpeed	0.060374	0.001387	43.525	<2e-16
WindDir9amENE	0.278366	0.071622	3.887	0.000102
WindDir9amN	0.266122	0.063350	4.201	2.66e-05
WindDir9amNE	0.354497	0.071352	4.968	6.75e-07
WindDir9amNNE	0.572843	0.069198	8.278	<2e-16
WindSpeed9am	-0.015993	0.001810	-8.836	<2e-16

WindSpeed3pm	-0.032510	0.001898	-17.129	<2e-16
Humidty3pm	0.059246	0.001079	54.893	<2e-16
Pressure9am	0.162846	0.007183	22.672	<2e-16
Pressure3pm	-0.219137	0.007343	-29.842	<2e-16
Cloud9am	0.060101	0.005689	10.564	<2e-16
Cloud3pm	0.177413	0.006327	28.041	<2e-16
Temp9am	0.019937	0.005736	3.476	0.000510
Temp3pm	-0.023676	0.006407	-3.695	0.000220
RainToday	0.498298	0.029877	16.679	<2e-16

After deciding the set the variables to use in the final model, I move on to test for the need of fitting a GLMM. The R code output shows that the variance of the random effect term μ_i and the error term is 0.3313911 and 0.9778049, respectively. Thus the ICC value is approximately 0.2531257, which is small and indicates that no significant difference if including the random effect. Considering the size of the dataset, including random effect would largely complex the model or even may not converge. Hence the final decision is not to include the effect and keeps the reduced GLM model as the final model.

3.3 Goodness of Final Model

The half-normal plot for the final model shows that the majority of data lines up with the half-normal quantiles despite two outlying points. The two outliers are weather observations from Williamtown on March 3, 2013 and from Brisbane on May 2, 2015.

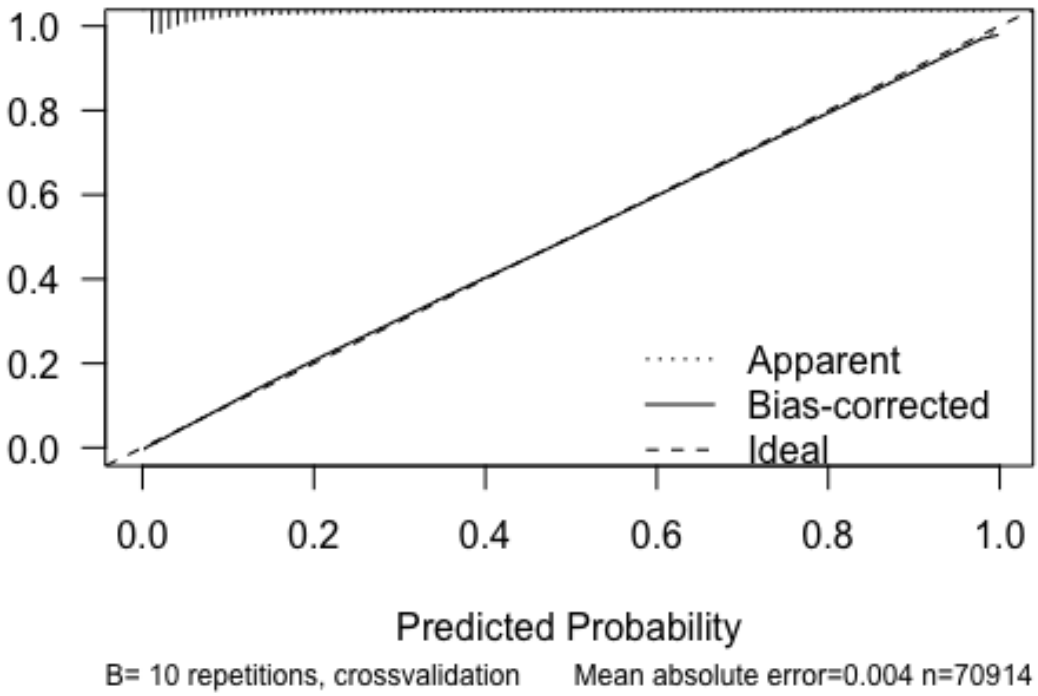

Figure 2: Cross validation plot for the final model.

The calibration plot from a 10-fold cross validation of the final model shows a satisfactory prediction accuracy.

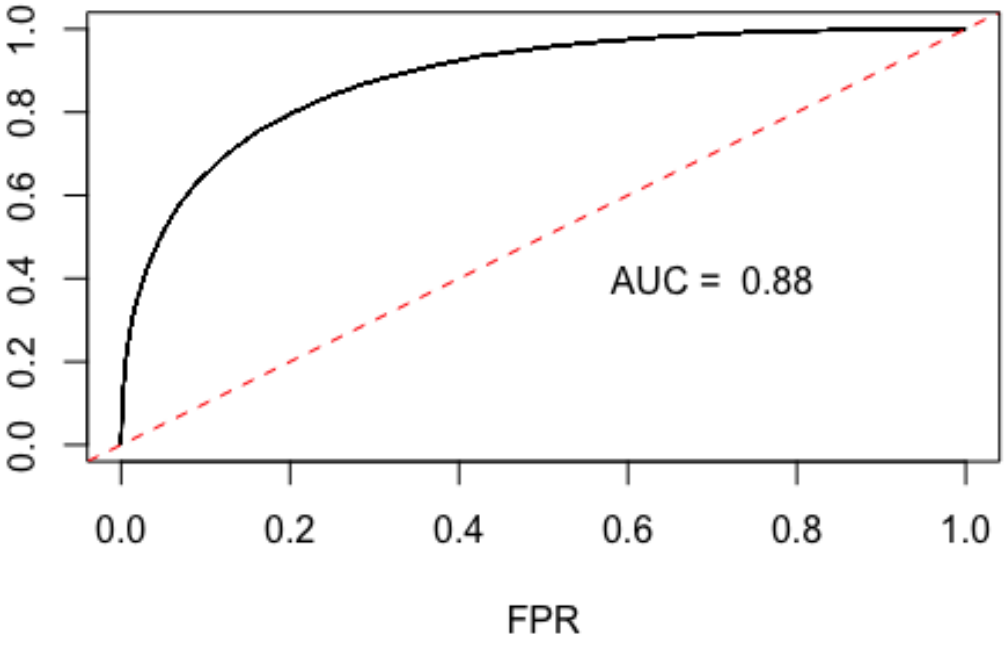

Figure 3: ROC curve of the final model

Figure 3 above shows that the area under the curve (AUC) is 0.88 which is quite close to 1. The AUC value of 0.88 suggests that the final model can correctly discriminate 88% of the times. The accuracy of predicting data from 2017 is approximately 84.82% when using 0.5 as the cutoff value for predicted probabilities. The calculated accuracy is closed to the area under the ROC curve.

4 Discussion

4.1 Final Model Interpretation

According to Table 2, the odds ratio of raining tomorrow for a rainy day is exp(0.498298), i.e. around 1.645918 times higher than a day not raining while controlling all other covariates in the final model. Also, controlling all other Xs in the final model, one millimetre increase in rainfall today increases the odds of having a rain tomorrow by around 0.6727%, which is a surprisingly low number. Another surprising finding from the final model is that the odds of raining tomorrow increase by a factor of 1.7333 for a day having wind direction coming from North-Northeast, controlling all other variables in the model. Wind coming from Northeast increases the odds of raining tomorrow by a factor of 1.4255.

4.2 Limitation

Using GLM as the predictive model for the weather data does not account for dependent observations, which may affect the model's prediction accuracy negatively. Despite the small ICC when treating location as the random effect, I may also try using date as the random effect to evaluate the need for a GLMM.

For outliers identified in section 3.3, they are still included in the train dataset despite being spotted on the half-normal plot. They should be more properly handled for their potential influence on model prediction.

The model's training dataset excludes observations that has missing value. The simple dropping of all entries with a missing value means an incomplete picture of the data which may also negatively affect the predictability of the fitted model.

References

Cristani, Matteo, et al. "It could rain: weather forecasting as a reasoning process." *Procedia Computer Science* 126 (2018): 850-859.

Faraway, Julian J. *Extending The Linear Model With R*. Chapman & Hall/CRC, 2006.

Wikipedia contributors. "Weather forecasting." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 15 Aug. 2021. Web.