Prediction of raining in Australia using GLM and GLMM based on data Australia weather data from 2007 to 2017

Ruike Xu 1006562550

29/08/2021

Introduction

There is no doubt that raining prediction is an essential technique for people around the world. In ancient times, perceptive people with such insight to detect nature's signs has lead people to do the right tasks at the right time, such as agricultural and migration planning. In contemporary society, countries that heavily depend on the agricultural industry are closely bonded upon crop productivity and rainfall, like Australia. According to the Department of Agriculture, Water and the Environment in Australia, Agriculture and its closely related sectors earn approximately $155 billion a year for a 12% share of GDP. (Department of Agriculture, Water and the Environment, 2021) The growing condition of the agricultural industry has a great impact on the economic and social aspects, so it is important to precisely predict rain given different weather conditions for improvement of crop productivity and efficient use of water resources. We will apply modeling techniques (mainly GLM and GLMM) to a dataset of weather conditions of 22 cities in Australia.

Data analysis

Data cleaning

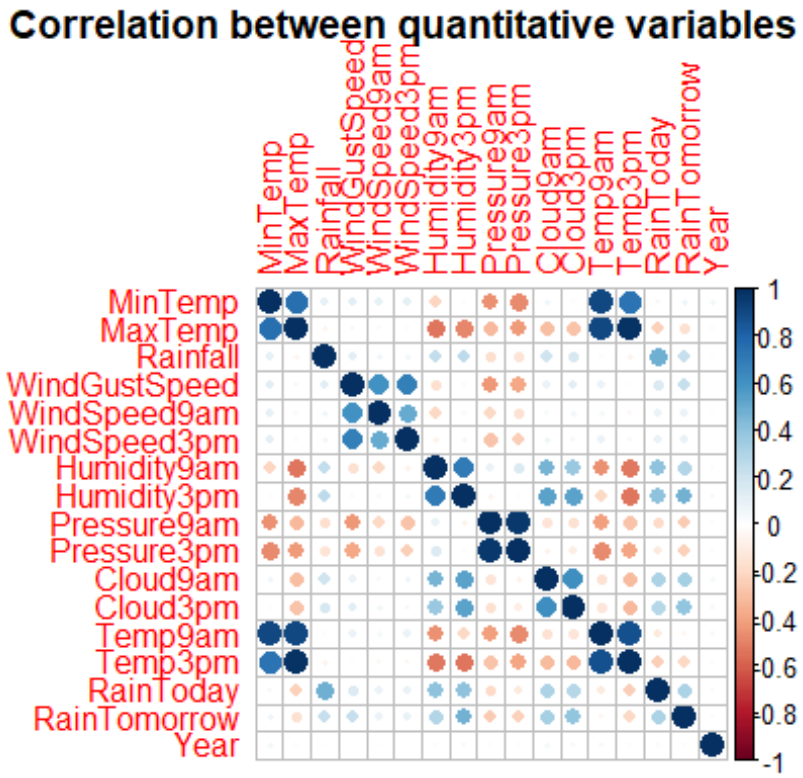
Our chosen dataset for this study contains weather data for 22 cities in Australia, which are measured and recorded from 2007 to 2017. There are in total 145460 observations in this dataset, which consists of details of weather-related measurements. By calculating the percentage of non-missing values in each variable for the whole dataset, Evaporation and Sunshine have a tiny amount of real recorded data comparing with other variables, Evaporation has only 1.254% of non-missing values whereas Sunshine has only 1.849%. On the other hand, these two variables are binary categorical variables, which makes us hard to interpret these missing values. Simply removing these missing values would lead to a large reduction of observations. Thus, we would remove these two variables from the dataset. The rest of the variables have much fewer missing values, after removing all the missing values, we preserve 71045 observations. We also convert RainToday and RainTomorrow to be dummy binary variables (Yes = 1; No = 0) We extract the year of observations that are recorded and split the observations that are measured in 2017 to be testing data. The rest of the dataset then becomes the training data of our model.

Essential data summaries

The overall rain percentage of the training dataset from 2007 to 2016 was 0.2379532, which indicates that only about a quarter of the observations in training data was recorded as raining. We could observe that some of the cities have a relatively large sample size than the others in the dataset, for instance, the city that contains most of the observations is Darwin(2941 data points) while the fewest city Uluru has only 202 observations. Such a huge gap in the number of observations between different cities could result from the missing values, some cities might be unable to record data properly due to technical issues, which would cause a huge statistical bias in our study.

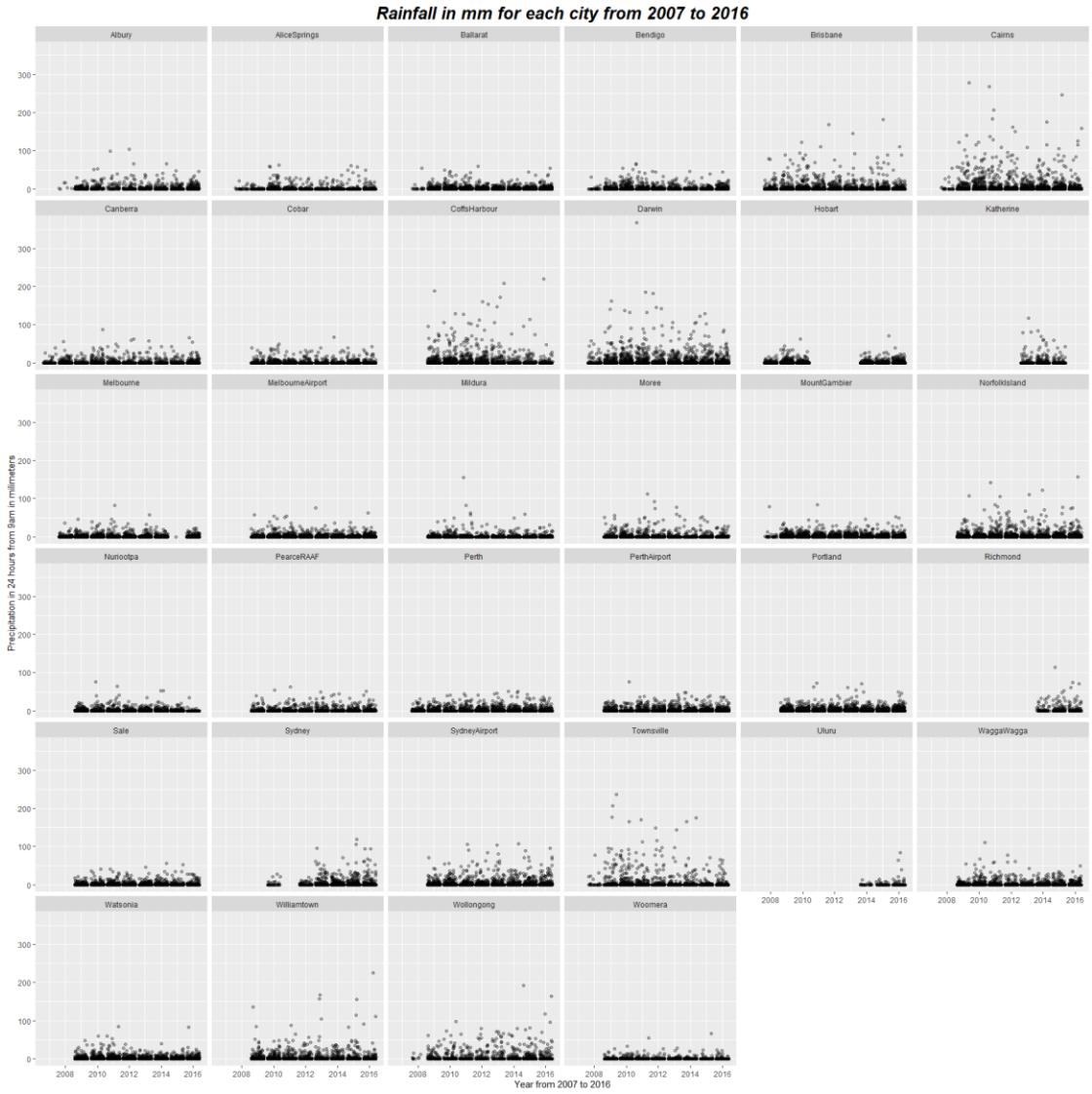
Data visualization

We could observe the distribution of quantitative variables of the dataset using histogram. For the histogram of quantitative variables in Appendix, most of the variables are approximately normally distributed, however, the fraction of sky obscured by cloud at 9 am and 3 pm have fewer observations at the center, indicating the weather of the given observations are mostly either a bright sky or overcloud. The distribution of rainfall level, on the other hand, is seriously right-skewed. Most of the data points are centered at 0, which means that the given observations are recorded in mostly dry conditions.



From the above correlation plot among quantitative variables(Giorgio Garziano, 2020), we could observe that Temp9am is strongly positively correlated to MinTemp, MaxTemp, and Temp3pm, which means that these predictors are not independent in the dataset. In addition, Temp3pm is also strongly positively correlated to MaxTemp and MinTemp. The Pressure at 9 am is strongly positively correlated to the pressure measured at 3 pm. Humidity3pm has a moderate positive

correlation with Humidity9am while Temp3pm has a moderate negative correlation with Humidity at 9 am and 3 pm. Since these variables are not independent of each other, we need to take into account the interaction effect when we construct our models.



From the above points plots for each city, we can see some of the cities have lost a relatively large number of observations comparing to other cities, for instance, Hobart, Katherine, Richmond, and Uluru. Brisbane, Cairns, CoffsHarbour, Darwin, Townsville, and Williamtown have a higher rainfall level on average than the other cities in our dataset.

Methods

GLM is generalized linear regression that allows the linear model to be related to the response variable by a link function. There are several assumptions need to be preserved for GLM: the independence of each observation, homogeneity of variance, normality of the residuals, and linearity between the response variable and the linear predictor. GLMM is an extension of GLM

where the linear predictors contain fixed effects with random effects. There are a few assumptions to maintain for GLMM, the intercepts and slopes of the random effects are normally distributed, the use of link function is appropriate for the model, homogeneity of variance.

Choice of model

The first model we choose to fit was a full logistic GLM without date variable and interaction terms. This model contains 21 predictors, by simply looking at the summary table, several predictors are insignificant in predicting whether tomorrow will rain. The AIC and BIC values of the model indicate that the model is poorly fitted (AIC = 23014.786; BIC = 23102.814) and the huge amount of regression coefficients make the model very hard to interpret. We employ a likelihood ratio test to compare between the fitted model and nested model with each independent term removed. The modified model makes the individual predictors significant from the previous model, both AIC and BIC values for this model decrease (AIC = 22697.851; BIC = 22780.403), which indicates the modified model is a better fitted one.

For the second and third models we constructed, we take into account the correlation between predictors in our dataset. We first include interaction terms of strongly correlated predictors, then we proceed stepwise AIC and BIC regression in both directions. Both stepwise regression models have improved accuracy in the case of AIC and BIC comparing to the previous model.

As we assumed in the previous GLM models, all the observations in our training dataset are independent, however, we only record 22 cities' weather conditions, which means that the observations from the same city are correlated to each other. We would like to measure the random effect of locations in order to measure the potential population effect of the cities. We first use stepwise BIC regression to get appropriate predictors from the full model, then we add the location as a random effect to the model. We first employ the PQL method, which can be flexible and widely implemented, but it's less accurate than Laplace or Gauss_Hermite approximation due to bias for large variance and small means. This model has a huge AIC and BIC value, so we would like to model using maximum likelihood. However, due to the large sample size and predictors, the model failed to converge with predictors from the previous model, so we removed some of the predictors that are highly correlated with each other. The GLMMs still fail to converge given several attempts of variables selection.

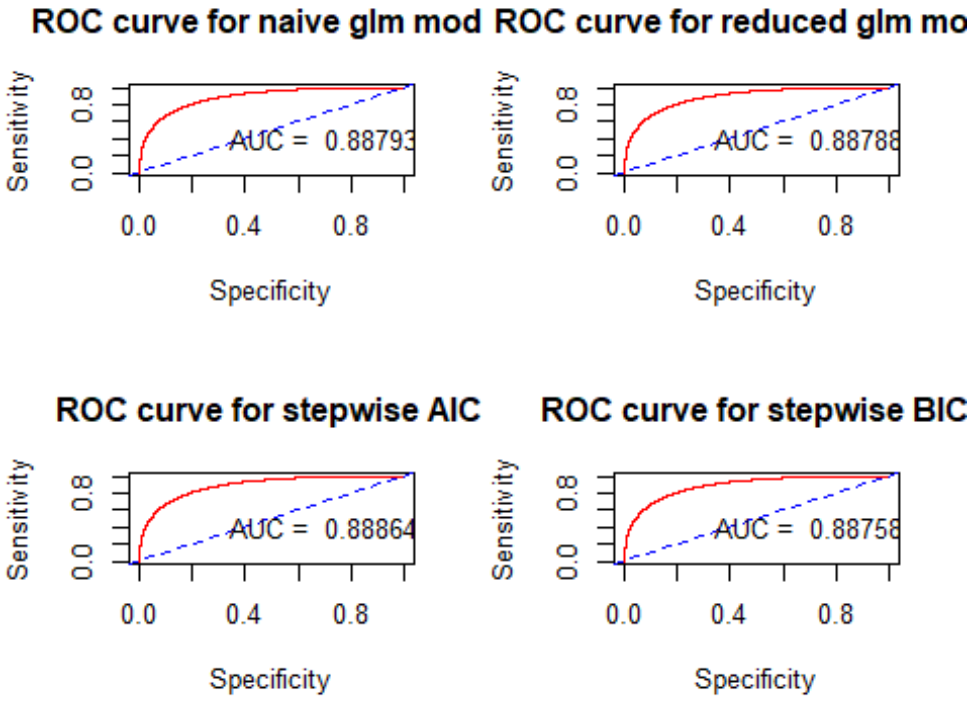
Result

Final model selection

Models	AIC	BIC	Prediction correction percentage
Full GLM model	23014.786	23102.814	0.8192128
Reduced GLM model with likelihood ratio test	22697.851	22780.403	0.8188793
Stepwise AIC with interactions in both directions	22622.405	22714.998	0.8198799
Stepwise BIC with interactions in both directions	22632.596	22695.066	0.817545

The above table demonstrates the properties of the models we constructed based on the training dataset. The prediction correction percentage is calculated by comparing the predicted values with observed values given a 0.5 threshold for predicted response. I would choose the stepwise AIC with interactions model as our final model since it has the highest prediction correction percentage with adequate AIC and BIC values.

Goodness of Final model



ROC is a curve showing the performance of a classification model at all classification thresholds. AUC is a measure of the separation ability of the model. The higher the AUC, the better the model is in distinguishing positive and negative classes. The final model Stepwise AIC has the

highest AUC value of 0.8887, which means there is a 88.9% probability that this model will be able to distinguish between positive and negative classes.

From the Appendix, the plots of deviance residuals v.s linear predictors and predicted values for stepwise AIC regression model show an unevenly distributed variation, but the problem of homogeneity of variance is improved than the previous model. By checking the QQ plot and half-normal plot, the normality of residuals is mostly preserved and there are only a few unusual observations given such a huge dataset. The unconstant variance could be caused by highly biased data and correlation among the observations.

Discussion

For our stepwise AIC regression model, others hold constant, an increase of 1% in the relative humidity at 3 pm is associated with an increase of 5.29% (exp(0.05157) - 1) in the odds of raining tomorrow. The city Alice Springs is 27.9% (1- exp(-0.327)) less likely to rain compared with the city Albury. Else hold constant, if both minimum and maximum temperature of the day increase by 1 degree Celsius, the odds of raining tomorrow decreases by 11.50% (1-exp(-0.02347-0.09338-0.005269)).

Our final model has accomplished the goal of our study: predict tomorrow's raining status given today's weather conditions. We have tried to reduce the predictors in the model to improve interpretability. The final model improves the discrimination ability and prediction accuracy comparing to the naïve glm model. Our study would help people in Australia to get an accurate prediction of raining status, improve their living quality and agriculture production.

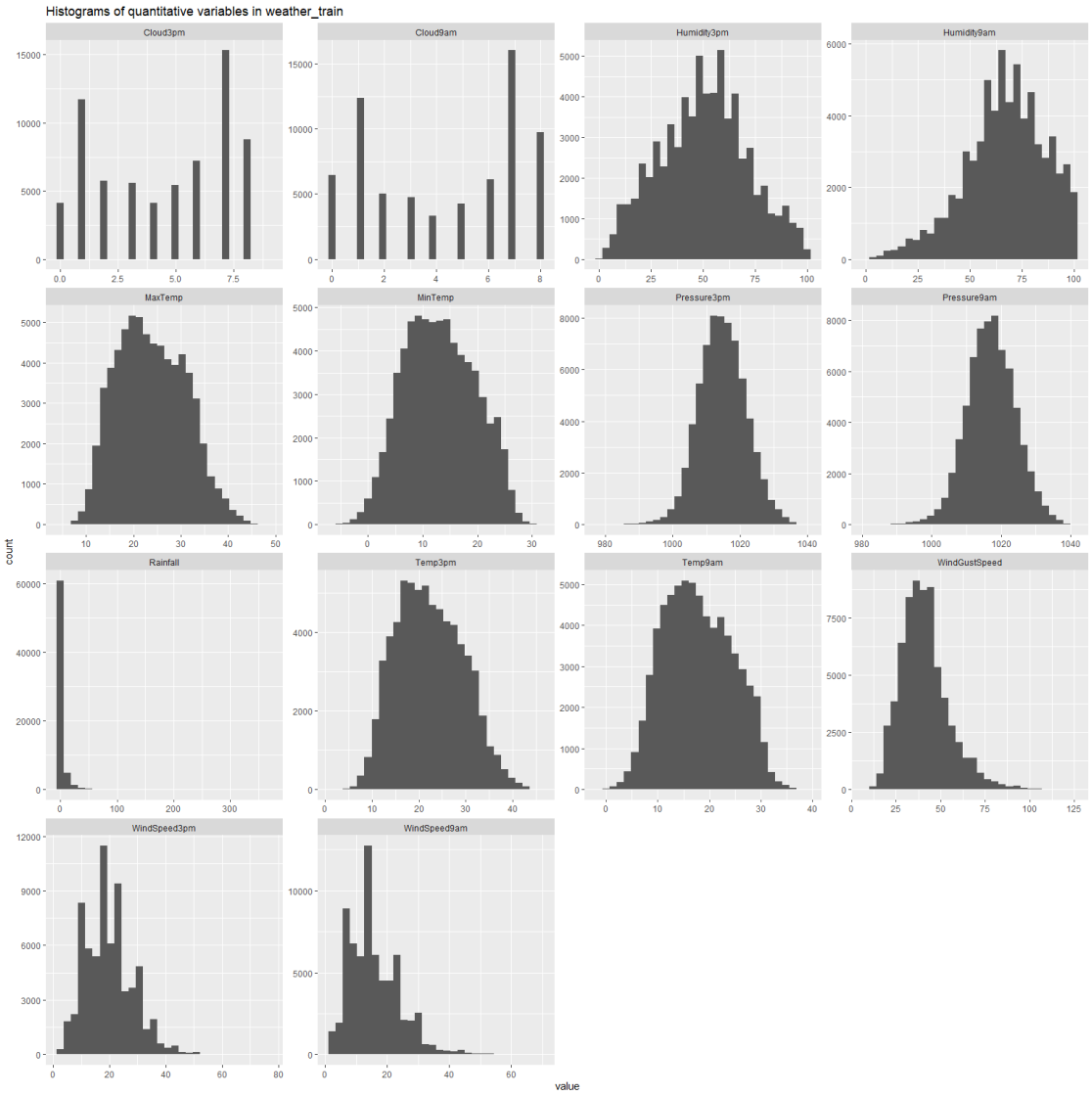
There are several limitations in our model that could have an impact on our study. First, the dataset was reduced to half of the original size, and two of the variables are removed due to missing values, which would increase the bias of the sample and variation. Second, the observations are treated as independent objects while many of them are correlated with each other, which violates the assumptions of GLM. Third, there exists heterogeneity of variance in our final model, we could improve that by employing a model transformation and reducing the bias of the sample. Finally, although we have reduced a certain amount of predictors, it is still not easy to interpret our model due to interaction terms that are used to balance the correlation among variables.

Reference

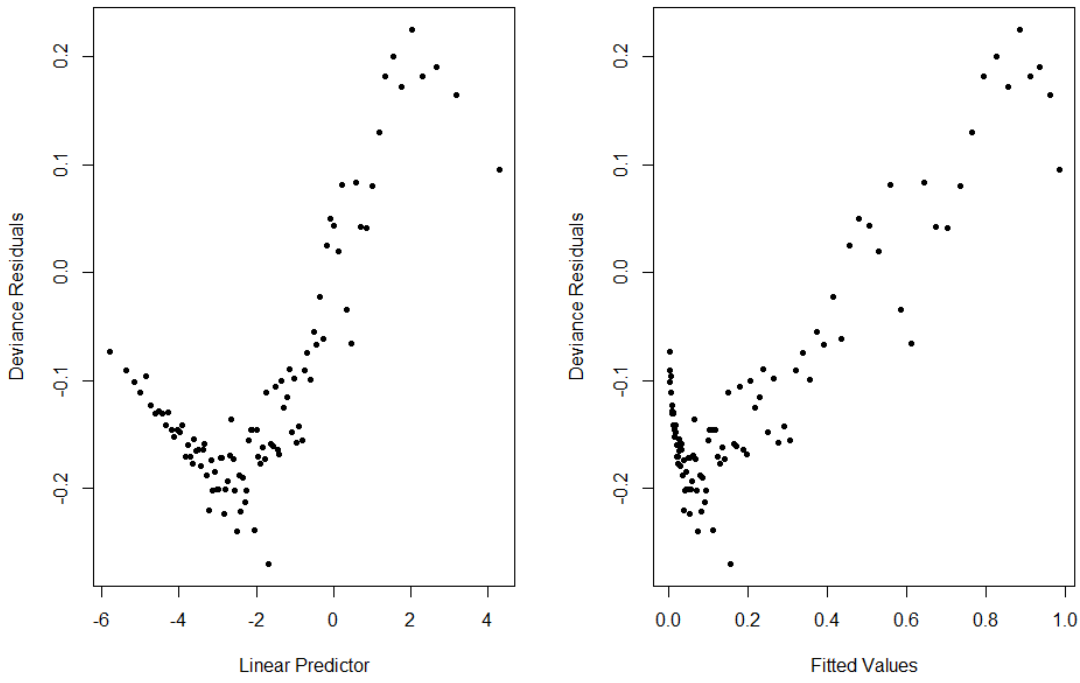
Department of Agriculture, Water and the Environment Data - Department of Agriculture. (n.d.). https://www.agriculture.gov.au/abares/data. (Last Accessed: August 23, 2021)

Garziano, G. (2020, July 10) *Regression Models in R* Datascience+. https://datascienceplus.com/ (Last Accessed: August 20, 2021)

Appendix


Histograms of quantitative variables in weather_train
Cloud3pm
Cloud9am
Humidity3pm
Humidity9am
MaxTemp
MinTemp
Pressure3pm
Pressure9am
Rainfall
Temp3pm
Temp9am
WindGustSpeed
WindSpeed3pm
WindSpeed9am
count
value

Residuals v.s. linear predictors and fitted value for stepwise AIC model



Plot for stepwise AIC model