The Next-Day Rain Prediction Model Using Various Wheather Factors

Eun Lim

28/08/2021

Introduction

Background

In this manuscript, we built a next-day rain prediction model by studying whether various weather characteristics measured in Australia on a daily basis influence the next-day rain.

Australia has a varied climates, where it varies from location to location. According to the Bureau of Meteorology of the Australian Government, for simplicity, locations with approximately similar climates have been combined into eight climate zones (Bureau of Meteorology-Australian Government, n.d.).

- Climate zone 1 - high humidity summer, warm winter
- Climate zone 2 - warm humid summer, mild winter
- Climate zone 3 - hot dry summer, warm winter
- Climate zone 4 - hot dry summer, cool winter
- Climate zone 5 - warm temperate
- Climate zone 6 - mild temperate
- Climate zone 7 - cool temperate
- Climate zone 8 - alpine

Since Australian weather differs greatly by location, it is inevitable that eventhough weather characteristics are measured on the same day, the characteristics may differ.

Methods

Data

Background

The data is composed of weather data for 22 cities in Australia, measured from 2007 to 2017. The data were collected at the day level. It includes 23 features, representing various weather characteristics, and it contained 145,460 total observations. After the data cleaning process, the dataset is divided into a training set and a test set, where the training set contains data from 2007 to 2016 and the test set contains data for 2017. The training set is used throughout the process of model building and the test set is used to test the model accuracy.

Missing Values

The treatment of missing values is done in three steps. The first step was dropping the entire variable. These variables are the ones that proportion of missing values is over 35%. Such a large proportion made the variable less representative. The second step was replacing the missing values by the mean of the location and if the location's mean is missing, then the missing values were replaced by the mean of the 22 cities.

Such a method was only applied to continuous variables. The last step was replacing the missing values with the mode, the most frequent response. Such a method was applied to categorical variables and if the data for a location is completely missing, then the missing values were replaced by the mode among the 22 cities.

Choice of Methods

To find the relationship between various weather characteristics and the next-day rain, the logistic generalized linear mixed model, GLMM, is used.

A logistic GLMM is typically useful for a nested data with binary responses since it accounts for dependencies within hierarchical groups by defining random effects terms and it effectively analyses a non-normal response.

For our dataset, we assumed that the location variable behaves like a grouping factor, and so it is a random effects term. The assumption is made under the fact that the weather characteristics in Australia differ by location and tested statistically by comparing the values of AIC and BIC.

A suitable mixed effects model for these purposes can be constructed by defining a random-effect for location into the standard logistic regression model.

Let p_{ij} be the probability of next-day rain on day j in location i, β be the a vector of model coefficients; X_{ij} be a vector of predictor variables for day j in location i, and u_i be the random intercept, such that $u_i \sim N(0, \sigma^2)$. Then, the model expression of the mixed effects model is given by,

$$\ln(\frac{p_{ij}}{1-p_{ij}}) = \beta^T X_{ij} + u_i$$

Variable Selection

In advance of building a GLMM with the weather data and selecting variables, we had to consider the possible risk of having the collinearity between predictor variables since it can be an issue in interpreting coefficient estimates because they, perhaps, do not have independent effects and result in high standard errors.

To examine the collinearity, we looked at the pairwise correlations between predictor variables. Instead of the Pearson correlation coefficient, we use the Spearman rank correlation coefficient, since it does not assume linearity between variables.

To deal with the high collinearity problem, we adopt the method of removing variables and using a linear combination of variables instead of using the variable directly to the model. Then, we checked the variance inflation factors, VIFs, of the variables to assess the collinearity.

After fixing the multicollinearity problem, we used the bidirectional stepwise selection method using the Akaike Information Criterion, AIC, and Bayesian Information Criterion, BIC. By the method, we tested at each step for variables to be included or excluded by comparing the trade-off between the goodness of fit of the model and the simplicity of the model using the value of AIC and BIC. If we let k be the number of parameters and $\hat{L}$ be the maximum likelihood estimate, then the AIC and BIC are expressed by,

$$\text{AIC} = 2k - 2\ln(\hat{L})$$
$$\text{BIC} = k\ln(n) - 2\ln(\hat{L})$$

Model Diagnostics

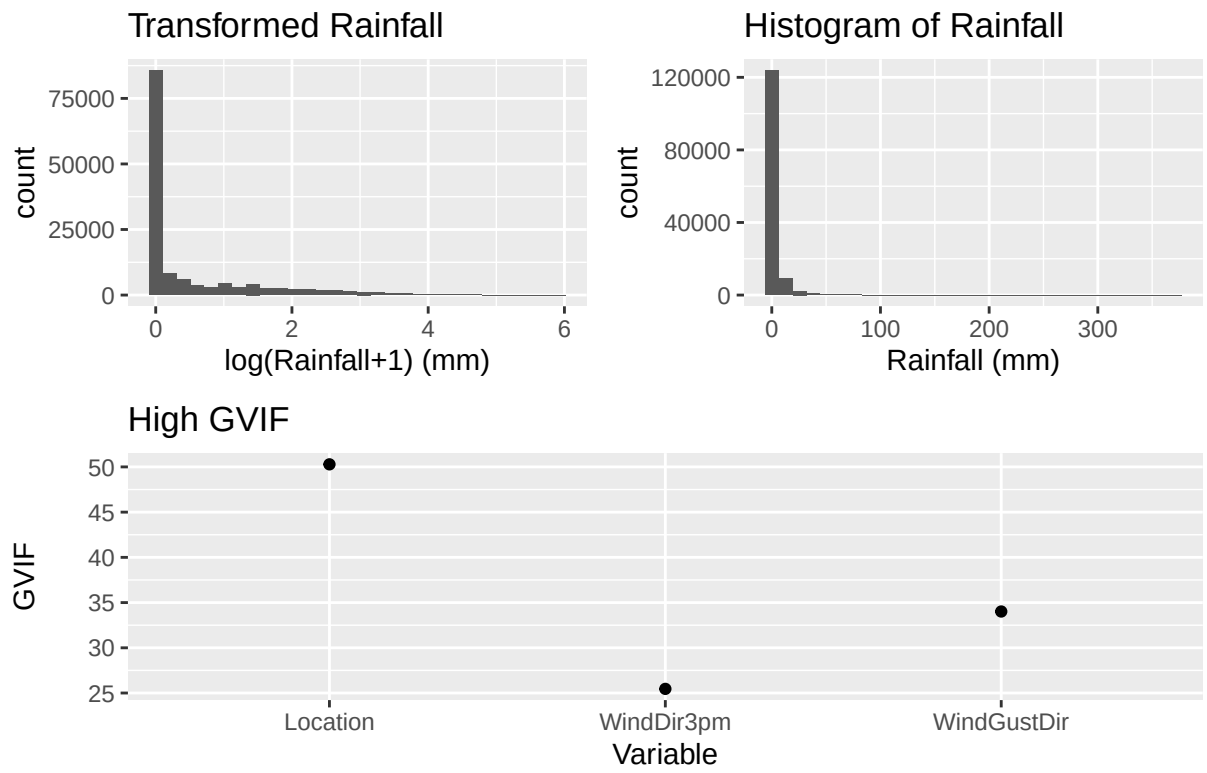
In this section, we checked the model assumptions by diagnosing the variance of data, the linearity in response with respect to continuous predictors, outliers, and the normality assumption of the random effects. The diagnostics were done by plotting a binned residual plot, partial residual plot, and the QQplot. The binned residual plot shows the averages of fitted values versus averages of residuals, instead of raw residuals. A partial residual plot essentially attempts to model the residuals of one predictor variable against the response variable.

Also, the model's goodness of fit will be tested based on the ability to discrimination the Receiver-Operating Characteristics, ROC which shows the trade-off between sensitivity, TPR, and specificity, $1 - FPR$.

Results

Description of Data

Fig 1: Exploratory Data Analysis



The variable denoting today's rain is redundant since the variable rainfall variable explains the same thing. For example, if the rainfall variable has a value greater than 0, it simultaneously indicates a "Yes" for today's rain variable. Thus, the today's rain variable is dropped.

From the correlation plot in appendix 2, we can find that the temperature related variables are all higly correlated with each other.

The wind direction related variables are all highly correlated with each other and the wind speed related variables show the same pattern.

The humidity and pressure variables are also highly correlated with each of them measured at different imes.

To deal with the high collinearity problem, we adopt the method of removing variables and using a linear combination of variables instead of using the variable directly to the model.

In a day, the highest temperature is measured around 2-3pm and the lowest temperature is measured right before sunrise, around 6am to 7am. Therefore, the maximum temperature we assume that the variables denoting maximum temperature and minimum temperature partly or almost fully explain the temperature at 9am and 3pm, therefore we rather use a new variable about the mean temperature during the day and the difference between the maximum temperature and the minimum temperature, instead.

Also, since the characteristics measured at 9 am and 3pm is all highly correlated, so we will rather use the mean of the two variables, mean humidity and mean pressure.

According to the second plot in appendix 2, we removed the multicolinearity problem substantially by the steps above.

The VIFs of WindGustDir, WindDir3pm, and Location are very high. We will get the average wind direction by the unit vector approach of average wind direction (Grange, 2014). We do not consider the VIF of Location is problematic since we treat it as a random effect. After the treatment, none of the VIFs, except location, was significantly high.

The histogram of continuous predictor variables shows that the Rainfall variable is extremely skewed, so we will transform it by $\log(1 + Rainfall)$ to make it less skewed and to avoid the log(0) problem.

Model

Table 1. Model Summary

	Logistic GLM			**Logistic GLMM**			**Logistic GLMM (Final Model)**		
Parameter	**OR**[1]	**95% CI**[1]	**p-value**	**OR**[1]	**95% CI**[1]	**p-value**	**OR**[1]	**95% CI**[1]	**p-value**
TempMean	1.02	1.01, 1.02	<0.001	1.02	1.01, 1.02	<0.001	1.01	1.01, 1.02	<0.001
TempDiff	0.95	0.95, 0.96	<0.001	0.95	0.95, 0.96	<0.001	0.95	0.94, 0.95	<0.001
logRainfall	1.25	1.23, 1.27	<0.001	1.20	1.18, 1.22	<0.001	1.18	1.16, 1.20	<0.001
WindDir									
E	—	—		—	—		—	—	
ENE	1.00	0.91, 1.11	>0.9	1.06	0.96, 1.17	0.3	1.07	0.97, 1.18	0.2
ESE	0.97	0.88, 1.08	0.6	0.96	0.87, 1.07	0.5	0.97	0.87, 1.07	0.5
N	1.75	1.60, 1.92	<0.001	1.81	1.64, 1.99	<0.001	1.79	1.63, 1.97	<0.001
NE	1.07	0.97, 1.18	0.2	1.08	0.98, 1.20	0.12	1.10	1.00, 1.22	0.059
NNE	1.35	1.22, 1.49	<0.001	1.34	1.21, 1.48	<0.001	1.36	1.23, 1.51	<0.001
NNW	1.79	1.63, 1.96	<0.001	1.86	1.69, 2.05	<0.001	1.85	1.68, 2.04	<0.001
NW	1.67	1.52, 1.82	<0.001	1.62	1.47, 1.78	<0.001	1.62	1.48, 1.78	<0.001
S	0.93	0.85, 1.02	0.13	0.92	0.84, 1.01	0.079	0.93	0.84, 1.02	0.13
SE	0.95	0.87, 1.05	0.3	0.92	0.84, 1.02	0.11	0.94	0.86, 1.04	0.2
SSE	0.96	0.87, 1.05	0.4	0.93	0.84, 1.02	0.11	0.94	0.85, 1.03	0.2
SSW	1.01	0.92, 1.11	0.8	0.92	0.84, 1.01	0.093	0.93	0.85, 1.02	0.14
SW	1.01	0.92, 1.11	0.8	0.87	0.79, 0.96	0.004	0.88	0.80, 0.97	0.011
W	1.17	1.07, 1.27	<0.001	1.02	0.93, 1.11	0.7	1.03	0.94, 1.12	0.5
WNW	1.41	1.29, 1.54	<0.001	1.28	1.17, 1.40	<0.001	1.29	1.18, 1.41	<0.001
WSW	1.07	0.98, 1.17	0.2	0.92	0.84, 1.01	0.093	0.94	0.86, 1.03	0.2
WindSpeed	1.04	1.04, 1.04	<0.001	1.05	1.05, 1.06	<0.001	1.05	1.05, 1.06	<0.001
Humidity	1.07	1.07, 1.07	<0.001	1.08	1.07, 1.08	<0.001			
Pressure	0.93	0.93, 0.94	<0.001	0.93	0.93, 0.93	<0.001	0.93	0.93, 0.93	<0.001
SquaredHumidity							1.00	1.00, 1.00	<0.001

[1]OR = Odds Ratio, CI = Confidence Interval

AIC = 104,496; BIC = 104,722; Log-likelihood = -52,225; Deviance = 104,210

We firstly set a logistic GLM, such that the variable denoting the next-day rain is a response variable, and the variables denoting the mean temperature, the difference in maximum and minimum temperature, mean wind direction, mean wind speed, mean humidity and mean pressure as predictor variables. Having set the model as a null model, we performed a bidirectional stepwise selection using the AIC and BIC. As a result, the null model, is selected as the best model with AIC = 107075.7 and BIC = 107291.9. Also, all of the predictor variables are significant in the model.

Then, we set a logistic GLMM, introducing the location variable as the random effects term. The AIC and

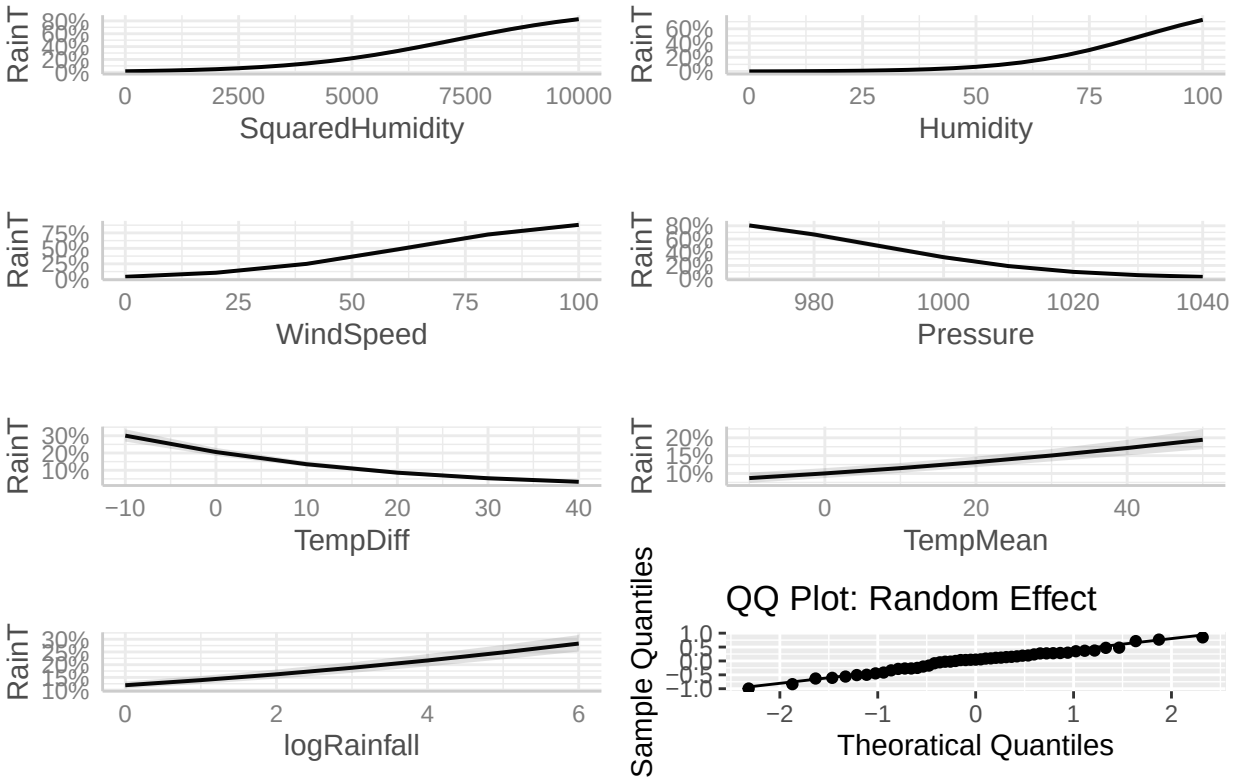
the BIC of the GLMM are 105169.6 and 105395.6 respectively. Since the AIC and BIC are a lot smaller for the GLMM than the GLM, we can state that the GLMM is a better model and the inclusion of the random effects term, Location, makes the model significantly better. Additionally, all of the variables are tested to be significant.

For the GLMM, the number of points per axis for evaluating the adaptive Gauss-Hermite approximation to the log-likelihood, i.e.,nAGQ in the `glmer` command, is set to be zero. The choice of zero is made under the consideration that values greater than 1 produces greater accuracy in the evaluation of the log-likelihood at the expense of speed. According to the vignette of the command, a value of zero uses a faster but less exact form of parameter estimation for GLMMs by optimizing the random effects and the fixed-effects coefficients in the penalized iteratively reweighted least-squares step.

As a result, we chose the logistic GLMM with a random intercept as the best model and the more specific numbers can be found in table 1.

Goodness of Model

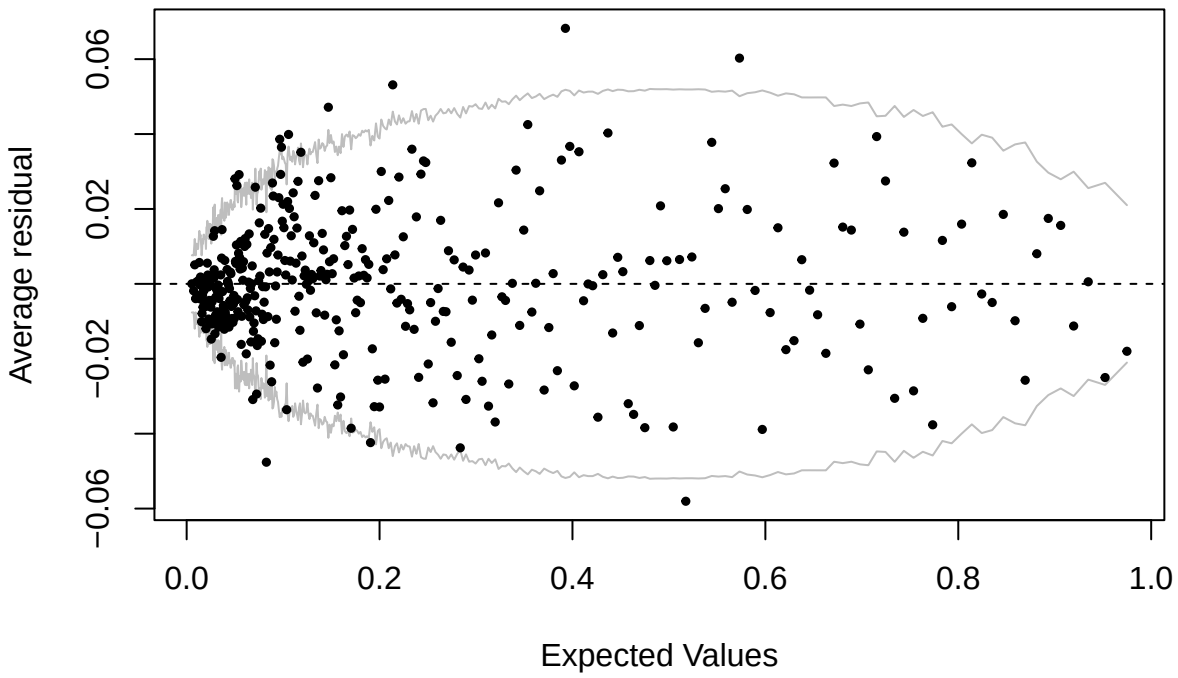
Fig 2: Partial Residual vs. Predictors



The marginal effect plots of each predictor variable show that except for the humidity variable, variables show very little deviation from being perfectly near. The humidity variable, on the other hand, shows a non-linear shape, therefore we transformed the variable by squaring it. The figure 2 shows that the transformed humidity variable no longer shows a non-linear structure. So, our final model now has the squared humidity as one of the predictor variables instead of the initial humidity variable.

The QQ plot of the random effects term shows that the location variable is approximately normally distributed.

Fig 3: Binned Residual vs. Expected Values



Most of the binned residual fitted values lie within the $\pm 2SE$ confidence bands. However, we see more outliers when the expected values are less than 0.4. We do not see a pattern indicating the violation of constant variance assumption. A bunching of values appears on the left-hand side but mostly the values are scattered around 0. We cannot really see any functional or systemic patterns. There are some outliers, the values outside the confidence bands, but the number of outliers is not that big nor they do not lie too far away from the band. Also, it seems that more than 95% of them are witnin the bands.

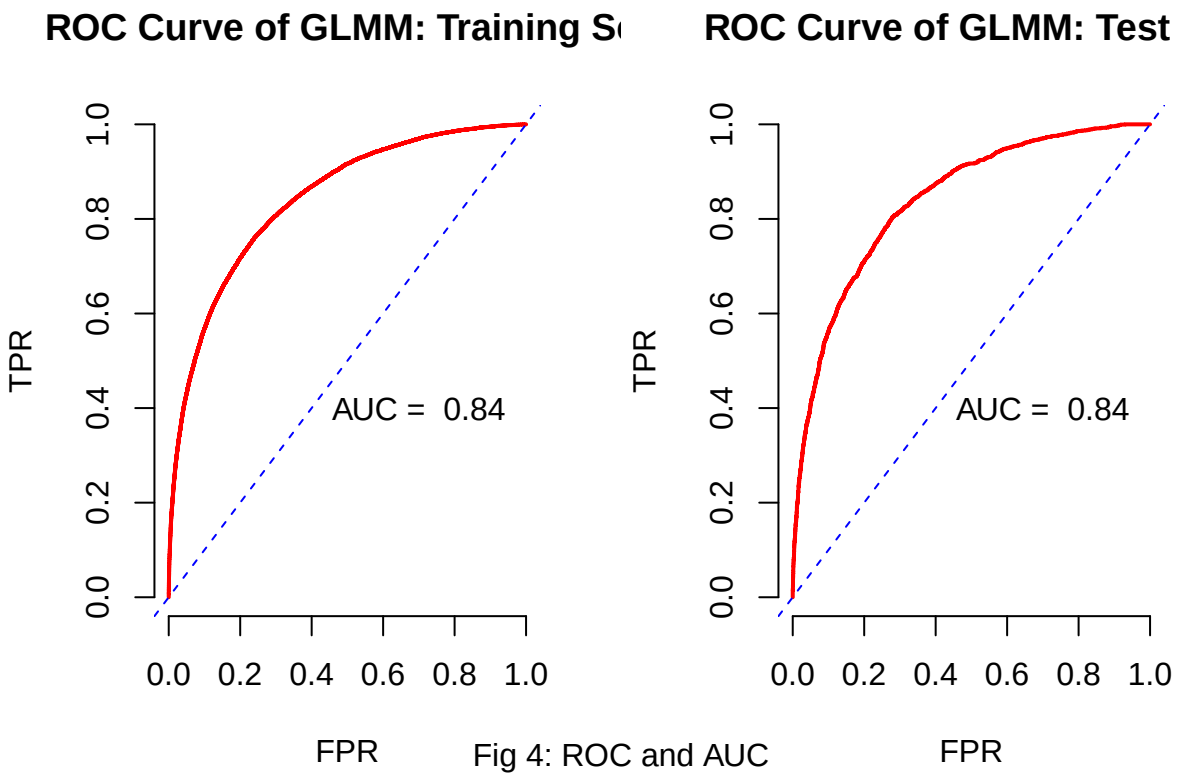


Fig 4: ROC and AUC

The area under the ROC curve, AUC, is 0.84, which implies that there is an 84% chance that the model will be able to distinguish between positives and negatives. Also, the AUC of 0.84 is pretty close to 1, so we can say the model has a pretty good discrimination ability. The ROC curve of the test model has the same result, so the interpretation is the same. Also, it shows that the model is pretty good since the discrimination ability is good with the test set, as well.

Discussion

Interpretation of Model

Our final model is a logistic GLMM such that the response variable is the next-day rain, the predictor variables are the variables denoting mean temperature, the difference between the maximum and minimum temperatures, rainfall, mean wind direction, mean wind speed, humidity, and mean pressure, and the random intercept term of Location. The rainfall variable is a log-transformed variable and the humidity variable is a power transformed variable.

The reference level of the model is when the average wind direction is East and all the other variables are equal to zero. The AIC of the final model is less than the AIC of the candidate model GLMM model-one with the humidity variable that is not transformed.

The odds of the next-day rain increase by 1.01 and 0.95 for a 1-degree increase in the average temperature and difference between the maximum and minimum temperature respectively. The odds of the next-day rain increases by 1.18 for a 1mm increase in $\log(Rainfall + 1)$. It implies that for a 1mm increase in today's rainfall, the odds of tomorrow's rainfall increases by 2.25. Similarly, the odds of tomorrow's rainfall increase by the numbers that appear on the Table. 1, under the OR column, for each average wind direction of today. However, if the average wind direction of today is the same as those with p-values greater than 0.025, for example, NE, W, SSE, etc., their impacts are negligible. For 1km/h increase in the average wind speed of

today, the odds of tomorrow's rain increases by 1.05. For a 1% increase in the average humidity today and 1 hectopascal increase in the average pressure today, respectively, increase the odds of tomorrow's rain by 1 and 0.93. See Table.1 for more detail.

The random effect of the location variable has a variance of 0.1558 and a standard deviation of 0.3947. This tells us how much, on average, the next-day rain factor bounces around as we move from location to location. On average, the next-day rain increases or decreases by 0.3947 from one location to another.

We can say that the predictor variables, the weather characteristics, are useful information to predict whether it will be raining tomorrow and the model predicts pretty well based on the discrimination ability of 84%.

Limitation

The limitation of the model is that it completely ignores that the data is time-series data, and so it completely ignores the possible auto-correlation. Autocorrelation is the correlation of lagged values of a time series. The data is measured on daily basis, which indicates that it is time-series data. In addition, according to the Australia State of Environment 2016 produced by the Department of Agriculture, Water and the Environment of the Australian Government, broadly, Australia can be divided into 2 seasonal rainfall regimes: the north and the south (Australian Government, 2016). Therefore, the seasonal or cyclical pattern may appear in the weather data.

Thus, not considering the temporal autocorrelation leads to the violation of the assumption that each data points are independent of each other, which, in turn, leads to the underestimation of standard errors and increases the risk of type I errors.

Bibliography

Bureau of Meteorology-Australian Government. (n.d.). *Regional Weather and Climate Guides*. Bureau of Meteorology. Retrieved August 29, 2021, from http://www.bom.gov.au/climate/climate-guides/

Grange, G. S. (2014). *Technical note: Averaging wind speeds and directions. Technical Note: Averaging Wind Speeds and Directions*, 2–6. https://doi.org/10.13140/RG.2.1.3349.200

Australian Government. (2016). *State of Environment 2016*. State of Environment 2016. https://soe.environment.gov.au/theme/overview

Appendix

Appendix 1: Code

```
knitr::opts_chunk$set(echo = TRUE)
library(mgcv)
library(tidyverse)
library(pROC)
library(MASS)
library(glmnet)
library(lme4)
library(glmmML)
library(car)
library(GGally)
library(ggpubr)
library(gtsummary)
library(gt)
library(gridExtra)
setwd("C:\\Users\\lsy76\\Desktop\\20-21\\summer\\final")
```

```
dat <- read.csv("weather.csv")
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

df<- dat %>%
  dplyr::select(-c(Evaporation, Sunshine, Cloud9am, Cloud3pm, RainToday))%>%
  group_by(Location)

df <- df%>%
  mutate_at(vars(c("MinTemp" ,"MaxTemp","Rainfall",
                   "WindGustSpeed","WindSpeed9am",
                   "WindSpeed3pm", "Humidity9am",
                   "Humidity3pm","Pressure9am",
                   "Pressure3pm","Temp9am","Temp3pm")),
    funs(ifelse(is.na(.), mean(., na.rm = TRUE),.))) %>%
  mutate_at(
    vars(c("WindGustDir","WindDir9am","WindDir3pm","RainTomorrow")),
    funs(dplyr::if_else(is.na(.), getmode(.),as.character(.))))

d <- ungroup(df)

df <- df%>%
  mutate_at(vars(c("WindGustSpeed", "Pressure9am", "Pressure3pm")),
    funs(ifelse(is.na(.), mean(d$., na.rm = TRUE),.)))%>%
  mutate_at(vars(c("RainTomorrow","WindGustDir", "WindDir3pm", "WindDir9am")),
    funs(ifelse(is.na(.),getmode(d$.),as.character(.))))%>%
  mutate(Date=as.Date(Date))

dir <- (setNames( seq(0, 337.5 , by=22.5),
                  c("N","NNE","NE", "ENE", "E", "ESE", "SE", "SSE",
                    "S", "SSW", "SW", "WSW", "W", "WNW", "NW", "NNW")))

df<-cbind(df,D=dir[df$WindGustDir])%>%
  cbind(.,D9=dir[df$WindDir9am])%>%
  cbind(.,D3=dir[df$WindDir3pm])

d2c <- function(x) {
  upper <- seq(from = 11.25, by = 22.5, length.out = 17)
  card1 <- c('N', 'NNE', 'NE', 'ENE', 'E', 'ESE', 'SE', 'SSE', 'S', 'SSW', 'SW',
             'WSW', 'W', 'WNW', 'NW', 'NNW', 'N')
  ifelse(x>360 | x<0,NA,card1[findInterval(x,upper,rightmost.closed = T)+1])
}

wind<-ungroup(df)%>%dplyr::select(c(D,D9,D3,WindGustSpeed,WindSpeed9am,
                                    WindSpeed3pm))%>%
  rename(W=WindGustSpeed,
         W9 = WindSpeed9am,
         W3 = WindSpeed3pm)%>%
  mutate_all(funs(as.numeric(.)))
```

```
wind$u.wind <- subset(wind,select=c(W,W9,W3))*
  sin(2*pi*subset(wind,select=c(D,D9,D3))/360)

wind$v.wind <- subset(wind,select=c(W,W9,W3))*
  cos(2*pi*subset(wind,select=c(D,D9,D3))/360)

mean.u <- rowMeans(wind$u.wind[,], na.rm = T)
mean.v <- rowMeans(wind$v.wind[,], na.rm = T)

wd.average <- ((atan2(mean.u, mean.v) *360/2/pi)+360)%%360
wd.average <- d2c(wd.average)
df <- cbind(df,WindDir=wd.average)
df$RainT <- ifelse(df$RainTomorrow=="Yes",1,0)
df <- ungroup(df) %>%
  mutate_at(vars(c(WindDir,WindGustDir,WindDir9am,WindDir3pm)),
            funs(as.factor(.)))
test <- df%>%
  filter(Date>="2017-01-01"&Date <= "2017-12-31")

train <- df%>%filter(Date<"2017-01-01")
cont <- train%>%
  ungroup()%>%
  dplyr::select(-c(Date,Location,RainTomorrow,WindDir3pm,WindDir9am,
                   WindGustDir,WindDir,D,D9,D3,RainT))

corr1 <- ggcorr(cont, method = c("everything", "spearman"),label=T,
                hjust = 0.9, size = 5, color = "grey50", layout.exp =2.5) +
  ggtitle("Correlogram: Before Fixing Variables")

train <- train %>%
  mutate(Humidity = (Humidity9am+Humidity3pm)/2,
         WindSpeed = (WindSpeed3pm+WindSpeed9am+WindGustSpeed)/3,
         Pressure = (Pressure9am+Pressure3pm)/2,
         TempDiff = (MaxTemp-MinTemp),
         TempMean = (MinTemp+MaxTemp+Temp3pm+Temp9am)/4)
test <- test %>%
  mutate(Humidity = (Humidity9am+Humidity3pm)/2,
         WindSpeed = (WindSpeed3pm+WindSpeed9am+WindGustSpeed)/3,
         Pressure = (Pressure9am+Pressure3pm)/2,
         TempDiff = (MaxTemp-MinTemp),
         TempMean = (MinTemp+MaxTemp+Temp3pm+Temp9am)/4) %>%
  dplyr::select(c(Date,Location,Rainfall,WindDir,WindSpeed,RainT,
                  Humidity,Pressure,TempDiff,TempMean))

cont <- train%>%
  dplyr::select(c(Rainfall,WindSpeed,Humidity,Pressure,TempDiff,TempMean))

corr2 <- ggcorr(cont, method = c("everything", "spearman"),label=T, hjust = 0.9,
                size = 5, color = "grey50", layout.exp =2.5) +
  ggtitle("Correlogram: After Fixing Variables")

glm<-glm(RainT~TempMean+TempDiff+Rainfall+WindGustDir+WindDir3pm+ WindDir9am+
           WindSpeed+ Humidity+ Pressure+Location,family=binomial,data=train)
```

```
v1 <-car::vif(glm)%>%
  as.data.frame(.)%>%
  filter(GVIF>=10)

v1_plot <- ggplot(data=v1,aes(y=GVIF,x=rownames(v1)))+
  geom_point() +
  ylab("GVIF") +
  xlab("Variable") +
  ggtitle("High GVIF")

train <- train %>%
  mutate(Humidity = (Humidity9am+Humidity3pm)/2,
         WindSpeed = (WindSpeed3pm+WindSpeed9am+WindGustSpeed)/3,
         Pressure = (Pressure9am+Pressure3pm)/2,
         TempDiff = (MaxTemp-MinTemp),
         TempMean = (MinTemp+MaxTemp+Temp3pm+Temp9am)/4)%>%
  dplyr::select(c(Date,Location,Rainfall,WindDir,WindSpeed,RainT,Humidity,
                  Pressure,TempDiff,TempMean))

library(cowplot)

vifs <- plot_grid(v1_plot)

cont <- cont%>%
  mutate(logRainfall = log(Rainfall+1))
train<-train %>% mutate(logRainfall = log(Rainfall+1))
test<-test %>% mutate(logRainfall = log(Rainfall+1))
rainfall_hist <- ggplot(data=train,aes(x=Rainfall)) +
  geom_histogram()+
  xlab("Rainfall (mm)") +
  ggtitle("Histogram of Rainfall")

lograinfall_hist <- ggplot(data=train,aes(x=logRainfall)) +
  geom_histogram()+
  xlab("log(Rainfall+1) (mm)") +
  ggtitle("Transformed Rainfall")
hist <- plot_grid(rainfall_hist,
                  lograinfall_hist, ncol=1, nrow=2)


library(patchwork)
patchwork<-(lograinfall_hist+rainfall_hist)/v1_plot

eda<-patchwork + plot_annotation(
  title = 'Fig 1: Exploratory Data Analysis'
)

glm <- glm(RainT~TempMean+TempDiff+logRainfall+WindDir+WindSpeed+Humidity+
             Pressure,family=binomial,data=train)

glmer <- glmer(RainT~TempMean+TempDiff+logRainfall+WindDir+WindSpeed+Humidity+
                 Pressure+(1|Location),family=binomial,data=train,nAGQ=0)
stepAIC(glm,direction="both")
```

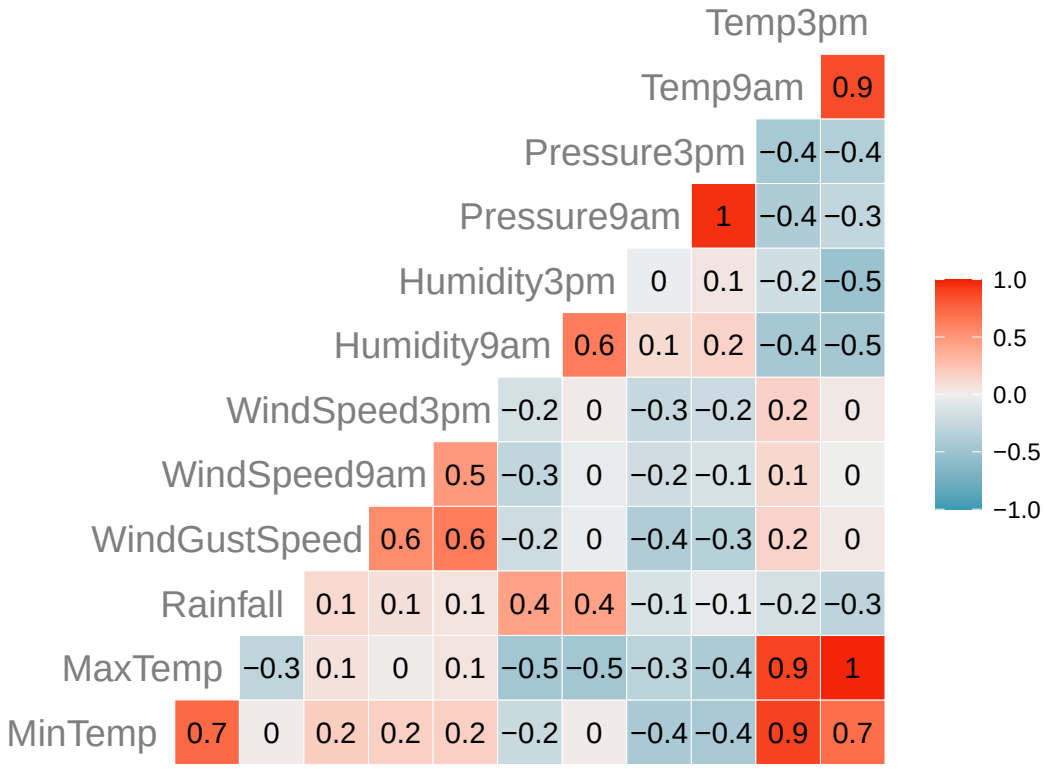
```
stepAIC(glm,direction="both",k=log(nrow(train)))
library(ggeffects)
windspeed<-ggpredict(glmer, terms = c("WindSpeed"))
humidity <- ggpredict(glmer, terms = c("Humidity"))
pressure <- ggpredict(glmer, terms = c("Pressure"))
tempd <- ggpredict(glmer, terms = c("TempDiff"))
tempm <- ggpredict(glmer, terms = c("TempMean"))
rainfall <- ggpredict(glmer, terms = c("logRainfall"))

train <- train %>% mutate(SquaredHumidity=Humidity^2)

glmer2 <- glmer(RainT ~ TempMean + TempDiff + logRainfall + WindDir +
                  WindSpeed +  SquaredHumidity + Pressure + (1 | Location),
                data=train,family=binomial,nAGQ=0)

windspeed2<-ggpredict(glmer2, terms = c("WindSpeed"))
humidity2 <- ggpredict(glmer2, terms = c("SquaredHumidity"))
pressure2 <- ggpredict(glmer2, terms = c("Pressure"))
tempd2 <- ggpredict(glmer2, terms = c("TempDiff"))
tempm2 <- ggpredict(glmer2, terms = c("TempMean"))
rainfall2 <- ggpredict(glmer2, terms = c("logRainfall"))

tib<-tibble(ranef = ranef(glmer2)$Location[[1]])
reqq<- ggplot(tib, aes(sample = ranef)) +
  stat_qq() +
  stat_qq_line() +
  xlab("Theoratical Quantiles") +
  ylab("Sample Quantiles")+
  ggtitle("QQ Plot: Random Effect")

partial <- plot_grid(plot(humidity2)+ggtitle(""),#+ggtitle("Average Wind Speed"),
plot(humidity)+ggtitle(""),
plot(windspeed)+ggtitle(""),#+ggtitle("Squared Average Humidity"),
plot(pressure)+ggtitle(""),#+ggtitle("Average Pressure"),
plot(tempd)+ggtitle(""),#+ggtitle("Difference in Temperatures"),
plot(tempm)+ggtitle(""),#+ggtitle("Average Temperature"),
plot(rainfall)+ggtitle(""),
reqq,ncol=2, nrow=4)#+ggtitle("Average Rainfall"),ncol=3, nrow=2)

title <- ggdraw() +
  draw_label(
    "Fig 2: Partial Residual vs. Predictors",
    fontface = 'bold',
    x = 0,
    hjust = 0
  ) +
  theme(
    # add margin on the left of the drawing canvas,
    # so title is aligned with left edge of first plot
    plot.margin = margin(0, 0, 0, 7)
  )
partial <- plot_grid(
  title, partial,
```

```
  ncol = 1,
  # rel_heights values control vertical title margins
  rel_heights = c(0.05, 1)
)
glmer_test <- glmer(RainT~TempMean+TempDiff+logRainfall+WindDir+WindSpeed+
                      I(Humidity^2)+Pressure+(1|Location),family=binomial,
                    data=test,nAGQ=0)

p <- fitted(glmer2)
roc_train <- roc(train$RainT ~ p)
TPR_train <- roc_train$sensitivities
FPR_train <- 1 - roc_train$specificities

p2 <- fitted(glmer_test)
roc_test <- roc(test$RainT ~ p2)
TPR_test <- roc_test$sensitivities
FPR_test <- 1 - roc_test$specificities
library(gt)

tglmer <- gtsummary::tbl_regression(glmer, exponentiate = TRUE)
tglm <-gtsummary::tbl_regression(glm, exponentiate = TRUE)

tglmer2 <- gtsummary::tbl_regression(glmer2, exponentiate = TRUE) %>%
  add_glance_source_note(include= c(AIC, BIC, logLik, deviance))

t <- tbl_merge(tbls= list(tglm, tglmer, tglmer2),
          tab_spanner= c("**Logistic GLM**",
                         "**Logistic GLMM**",
                         "**Logistic GLMM (Final Model)**"))%>%
  modify_header(label= "**Parameter**") %>%
  modify_caption("**Table 1. Model Comparison**")  %>%
  bold_labels()%>%
  as_gt() %>%
  tab_header(title = md("Table 1. Model Summary"),
             subtitle = md("")) %>%
  gt::as_latex()

eda
t
partial
binned <-arm::binnedplot(x = fitted(glmer2, type='response'),
                y = resid(glmer2,type='response'),
                xlab = "Expected Values",
                main = "Fig 3: Binned Residual vs. Expected Values", cex.pts=.5,
                col.int="grey")

par(mfrow=c(1,2))
plot(FPR_train, TPR_train, xlim = c(0,1), ylim = c(0,1), type = 'l', lty = 1,
     lwd = 2,col = 'red', bty = "n", main = "ROC Curve of GLMM: Training Set",
     xlab="FPR",ylab = "TPR", cex=0.4)
abline(a = 0, b = 1, lty = 2, col = 'blue')
text(0.7,0.4,label = paste("AUC = ", round(auc(roc_train),2)))
```

```
plot(FPR_test, TPR_test, xlim = c(0,1), ylim = c(0,1), type = 'l', lty = 1,
     lwd = 2,col = 'red', bty = "n", main = "ROC Curve of GLMM: Test Set",
     xlab="FPR",ylab = "TPR", cex=0.4)
abline(a = 0, b = 1, lty = 2, col = 'blue')
text(0.7,0.4,label = paste("AUC = ", round(auc(roc_test),2)))

mtext("Fig 4: ROC and AUC", side = 1, line = -2, outer = TRUE)
corr1;
corr2
```

Appendix 2: Correlogram

Correlogram: Before Fixing Variables



Correlogram: After Fixing Variables