Weather pridiction

Introduction:

The goal of this study is to try to predict whether it will rain tomorrow or not with the training set of data from 22 cities in Australia from 2007 to 2016 using variables like location, minimum/maximum temperature of the day, humidity etc. The purpose of this report is to find a relatively good model using selected explanatory variables to help predict whether it will rain tomorrow or not. The importance of the study is reflected on two degrees. For individuals, it help us plan our schedule and avoid getting drenched from the rain. For outdoor service industries, knowing the weather tomorrow can also help them prepare stuff like canopy for the customers.

Preprocess the data:

Before we can get into our data analysis, we need to deal with all the missing values to help with our model. For this dataset, I divided all the missing values into 3 categories. First, all the numerical data, second, categorical data that we don't know if its important or not (to be discovered in this report) and last the important categorical data.

For the numeric missing values, based on property of this dataset (all weather related), I decided to first divide the data into regions and then months, then fill in the missing value with the mean of its location and month. (This conclusion is drown after I compared the variance of the data before and after I did the separation and it leads to significantly lower variance with the change.), Also, Cloud9am and Cloud3pm is removed because more than 30,000 values are still missing after this method is applied which means there are cities that have no record of this data for whole months which makes predicting weather using these variables not reliable.

For the most important categorical variables like RainToday and RainTomorrow, I decided to delete all rows with either of these two variables missing because RainTomorrow is what we are predicting, imputing anything in it doesn't make sense.

For other categorical variables other than Evaporation and Sunshine. I decided to impute them with the mode of the factors at the given region and month. And for Evaporation and Sunshine, since 1/3 or the values for these two are missing, after investigation, I found that for regions like Albury Cobar and Newcastle, there is no data for Evaporation and Sunshine across the whole dataset. Which makes these two variables not reliable and I decide to remove this two variables directly.

Model & variable selection:

The model I choose to use before variable selection is logistic regressions. Since our response variable is binary which suggest using a logic link is more appropriate, and the reason for not using generalized mixed model is that it takes forever to run selection algorithms using location as a mixed effect (since weather is considered to be a regional thing, therefore, using location as a mixed effect is reasonable)

For our dataset, all three AIC method (backwards, forwards and both direction) gives the same final model with the same AIC score of 93884 with 16 independent

variables. For BIC, compared with the model after AIC, BIC suggest the effect of Temp9am is 0. However, while checking the summary statistics of the AIC models, it shows that Temp9am is statistically significant. Consider the natural of BIC that tend to underfitting and AIC overfitting, in terms of weather prediction, I think keeping Temp9am to not underfit my model is better.

Analyze the data:

Model selected: glm with response variable RainTomorrow and all other remaining column as explanatory variable and using binomial family with logit link function.

Below is the QQ plot and residual vs fitted value plot of the selected model.

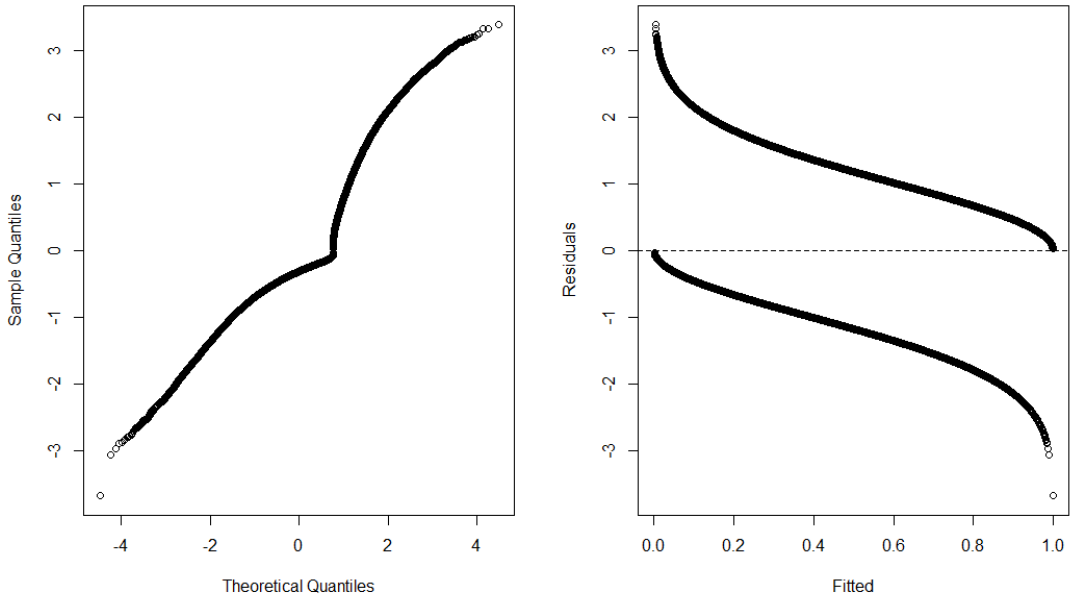

Figure 1 QQ plot and residual-fitted value plot for model assumptions.

because of the nature of logistic regression is curvilinear, therefore the shape of the fitted vs residual plot looks fine. And the normal QQ plot looks approximated like a straight line so we are fine. And according to the QQ plot, there doesn't seem to be any extreme outliers too.

Now lets try to find the ROC curve for this model

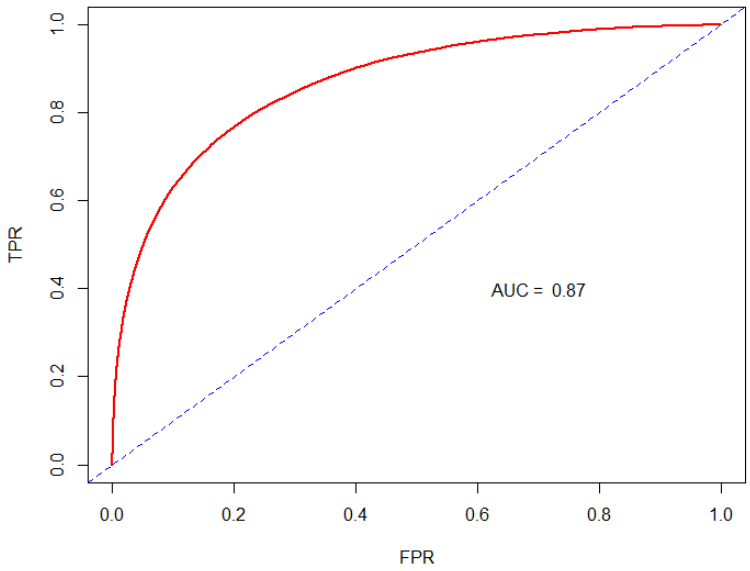

Figure 2 ROC curve and its AUC score for the model.

We can see that the AUC is 0.87 which means that the model will be distinguish between rain tomorrow and not rain tomorrow 87% of the time.

For cross-validation on this model as shown below, we can see that the bias corrected line almost fit the 45 degree line, thus this model does perform pretty well in predicting the responses from the training dataset.

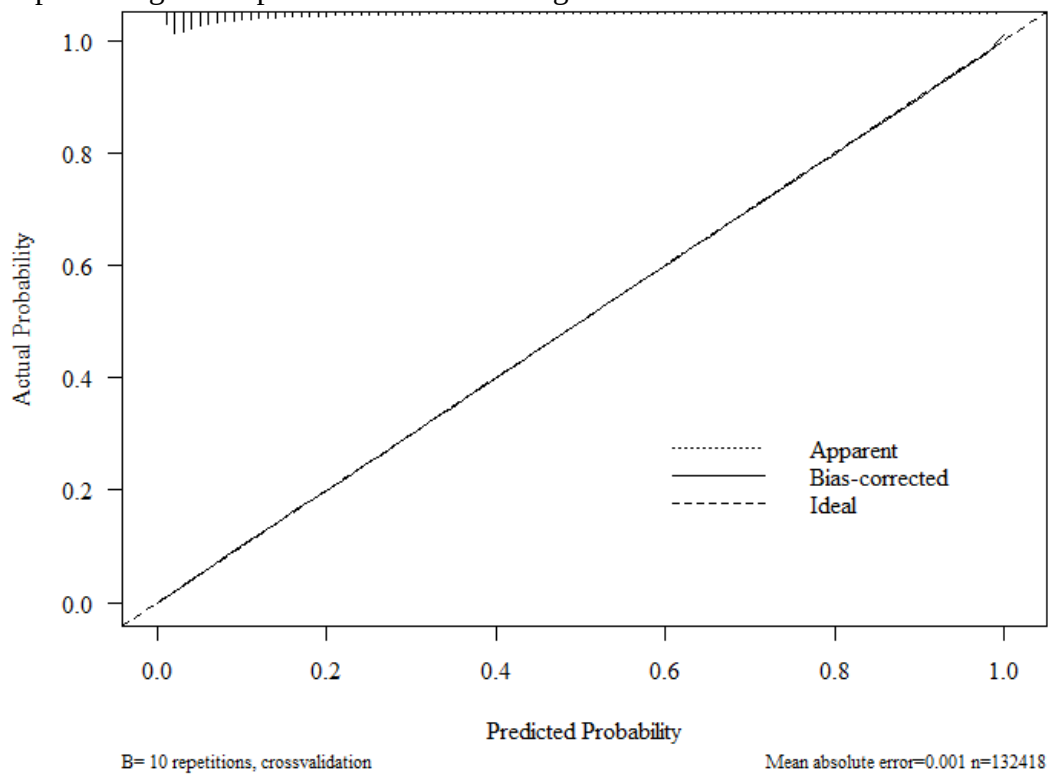

Figure 3 cross-validation plot

Below is the mapping of the predicted vs observed graph using the test data.

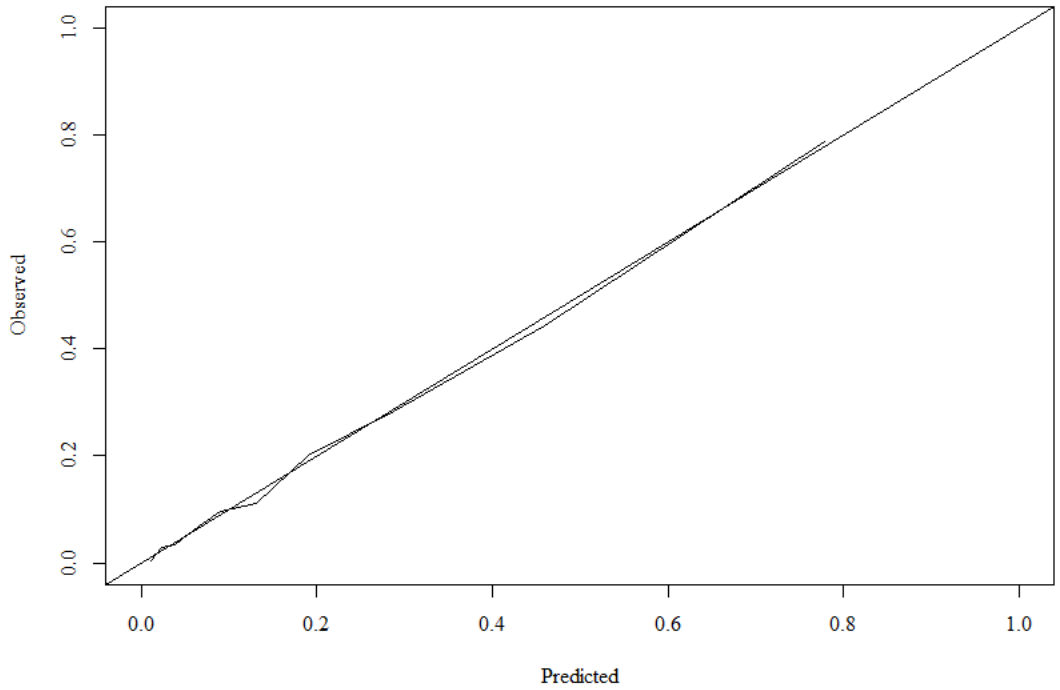

Figure 4 use the test data against the model to see its performance

From here, we can see that the only place that shows a little drop in observed probability is from decile 1 to somewhere near decile 2. This shows that the model does have a relatively strong predictive power.