Rain in Australia. Classification Prediction Model

by Sumaira Afzal, Viraja Ketkar, Murlidhar Loka, Vadim Spirkov

Abstract An abstract of less than 150 words.

Introduction

Background

Objective

Data Analisys

The data set we are going to use for our research contains daily weather observations from numerous Australian weather stations from 2007 till 2017. There are over 142000 records. It has been sourced from Kaggle

Data Dictionary

We exclude the variable Risk-MM when training your binary classification model. If we don't exclude it, you will leak the answers to our model and reduce its predictability

Column Name	Column Description
Date	Date of observation
Location	Common name of the location of the weather station
MinTemp	Minimum temperature in degrees Celsius
MaxTemp	Maximum temperature in degrees Celsius
Rainfall	Amount of rainfall recorded for the day in mm
Evaporation	So-called Class A pan evaporation (mm) in the 24 hours to 9am
Sunshine	Number of hours of bright sunshine in the day
WindGustDir	Direction of the strongest wind gust in the 24 hours to midnight
WindGustSpeed	Speed (km/h) of the strongest wind gust in the 24 hours to midnight
WindDir9amDirection	Of the wind at 9am
WindDir3pmDirection	Of the wind at 3pm
WindSpeed9amWind	Wind speed (km/hr) averaged over 10 minutes prior to 9am
WindSpeed3pmWind	Wind Speed (km/hr) averaged over 10 minutes prior to 3pm
Humidity9amHumidity	Humidity (percent) at 9am
Humidity3pmHumidity	Humidity (percent) at 3pm
Pressure9amAtmospheric	Pressure (hpa) reduced to mean sea level at 9am
Pressure3pmAtmospheric	Pressure (hpa) reduced to mean sea level at 3pm
Cloud9amFraction	Area of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eights. It records how many eights of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast
Cloud3pmFraction	Area of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cloud9am for a description of the values

Column Name	Column Description
Temp9amTemperature	Temperature (degrees C) at 9am
Temp3pmTemperature	Temperature (degrees C) at 3pm
RainTodayBoolean	Rainy today. 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
RISK_MM	Amount of rain. A kind of measure of the "risk". This column is redundant and will be dropped
RainTomorrowThe	**Target variable. Will it rain tomorrow?**

Data Exploration

Let's take a close look at the data set. We start with loading weather observations from the file into a data frame. We remove RISK_MM as explained and convert Date column to *date* format

```
weatherData = read.csv("../data/weatherAUS.csv", header = TRUE, na.strings = c("NA","","#NA"),sep=",")
weatherData = subset(weatherData, select = -RISK_MM)
weatherData$Date = as.Date(as.character(weatherData$Date),"%Y-%m-%d")
```

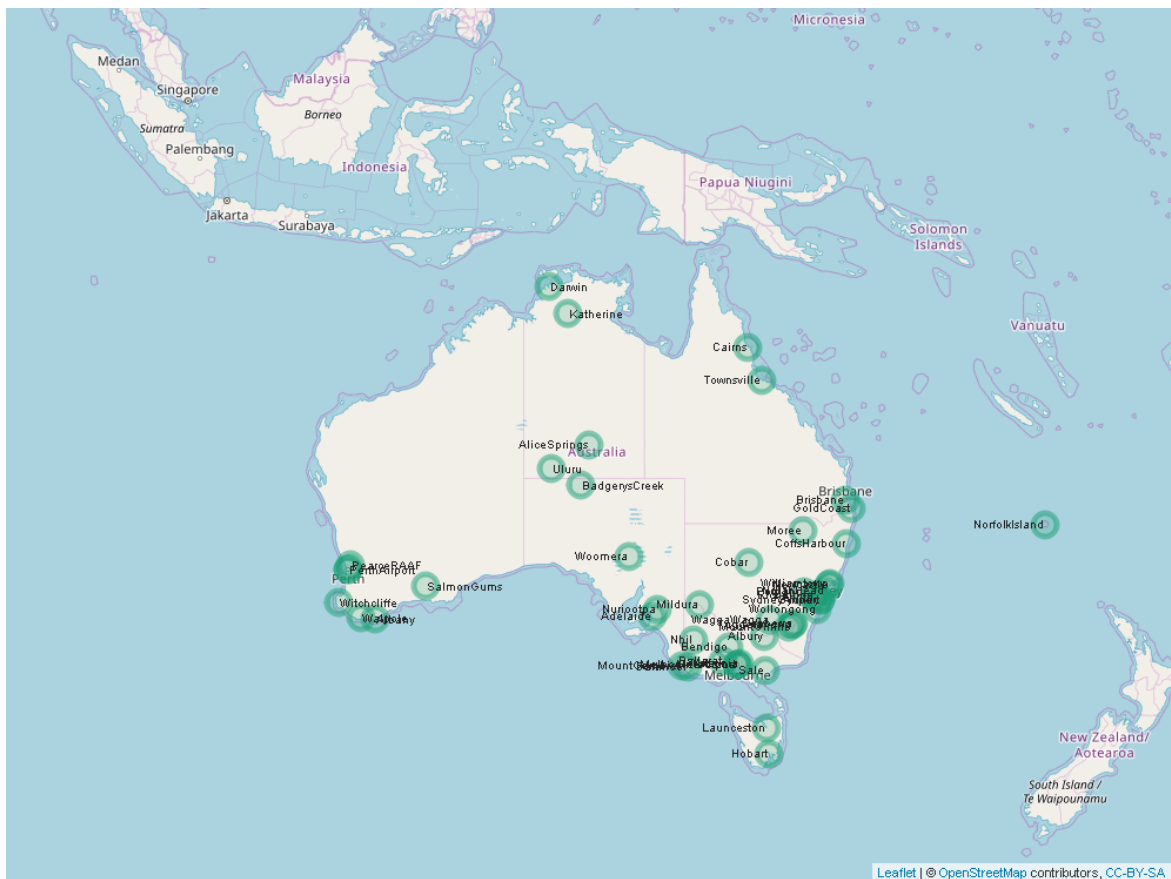Now let's load coordinates of the weather stations and have a bird-eye view of the weather station locations

Figure 1: Australian Weather Stations

Let's review data summary

```
summary(weatherData)
```

```
#>      Date               Location        MinTemp          MaxTemp
#> Min.   :2007-11-01   Canberra: 3418   Min.   :-8.50   Min.   :-4.80
#> 1st Qu.:2011-01-06   Sydney  : 3337   1st Qu.: 7.60   1st Qu.:17.90
#> Median :2013-05-27   Perth   : 3193   Median :12.00   Median :22.60
```

```
#>   Mean   :2013-04-01   Darwin  : 3192   Mean   :12.19   Mean   :23.23
#>   3rd Qu.:2015-06-12   Hobart  : 3188   3rd Qu.:16.80   3rd Qu.:28.20
#>   Max.   :2017-06-25   Brisbane: 3161   Max.   :33.90   Max.   :48.10
#>                        (Other) :122704  NA's   :637     NA's   :322
#>    Rainfall        Evaporation       Sunshine       WindGustDir
#>  Min.   :  0.00   Min.   :  0.00   Min.   : 0.00   W      : 9780
#>  1st Qu.:  0.00   1st Qu.:  2.60   1st Qu.: 4.90   SE     : 9309
#>  Median :  0.00   Median :  4.80   Median : 8.50   E      : 9071
#>  Mean   :  2.35   Mean   :  5.47   Mean   : 7.62   N      : 9033
#>  3rd Qu.:  0.80   3rd Qu.:  7.40   3rd Qu.:10.60   SSE    : 8993
#>  Max.   :371.00   Max.   :145.00   Max.   :14.50   (Other):86677
#>  NA's   :1406     NA's   :60843    NA's   :67816   NA's   : 9330
#>  WindGustSpeed     WindDir9am       WindDir3pm      WindSpeed9am
#>  Min.   :  6.00   N      :11393   SE     :10663   Min.   :  0
#>  1st Qu.: 31.00   SE     : 9162   W      : 9911   1st Qu.:  7
#>  Median : 39.00   E      : 9024   S      : 9598   Median : 13
#>  Mean   : 39.98   SSE    : 8966   WSW    : 9329   Mean   : 14
#>  3rd Qu.: 48.00   NW     : 8552   SW     : 9182   3rd Qu.: 19
#>  Max.   :135.00   (Other):85083   (Other):89732   Max.   :130
#>  NA's   :9270     NA's   :10013   NA's   : 3778   NA's   :1348
#>   WindSpeed3pm     Humidity9am      Humidity3pm      Pressure9am
#>  Min.   : 0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 980.5
#>  1st Qu.:13.00   1st Qu.: 57.00   1st Qu.: 37.00   1st Qu.:1012.9
#>  Median :19.00   Median : 70.00   Median : 52.00   Median :1017.6
#>  Mean   :18.64   Mean   : 68.84   Mean   : 51.48   Mean   :1017.7
#>  3rd Qu.:24.00   3rd Qu.: 83.00   3rd Qu.: 66.00   3rd Qu.:1022.4
#>  Max.   :87.00   Max.   :100.00   Max.   :100.00   Max.   :1041.0
#>  NA's   :2630    NA's   :1774     NA's   :3610     NA's   :14014
#>   Pressure3pm      Cloud9am        Cloud3pm         Temp9am
#>  Min.   : 977.1   Min.   :0.00   Min.   :0.0    Min.   :-7.20
#>  1st Qu.:1010.4   1st Qu.:1.00   1st Qu.:2.0    1st Qu.:12.30
#>  Median :1015.2   Median :5.00   Median :5.0    Median :16.70
#>  Mean   :1015.3   Mean   :4.44   Mean   :4.5    Mean   :16.99
#>  3rd Qu.:1020.0   3rd Qu.:7.00   3rd Qu.:7.0    3rd Qu.:21.60
#>  Max.   :1039.6   Max.   :9.00   Max.   :9.0    Max.   :40.20
#>  NA's   :13981    NA's   :53657  NA's   :57094  NA's   :904
#>    Temp3pm       RainToday     RainTomorrow
#>  Min.   :-5.40   No :109332   No :110316
#>  1st Qu.:16.60   Yes : 31455  Yes: 31877
#>  Median :21.10   NA's:  1406
#>  Mean   :21.69
#>  3rd Qu.:26.40
#>  Max.   :46.70
#>  NA's   :2726
```

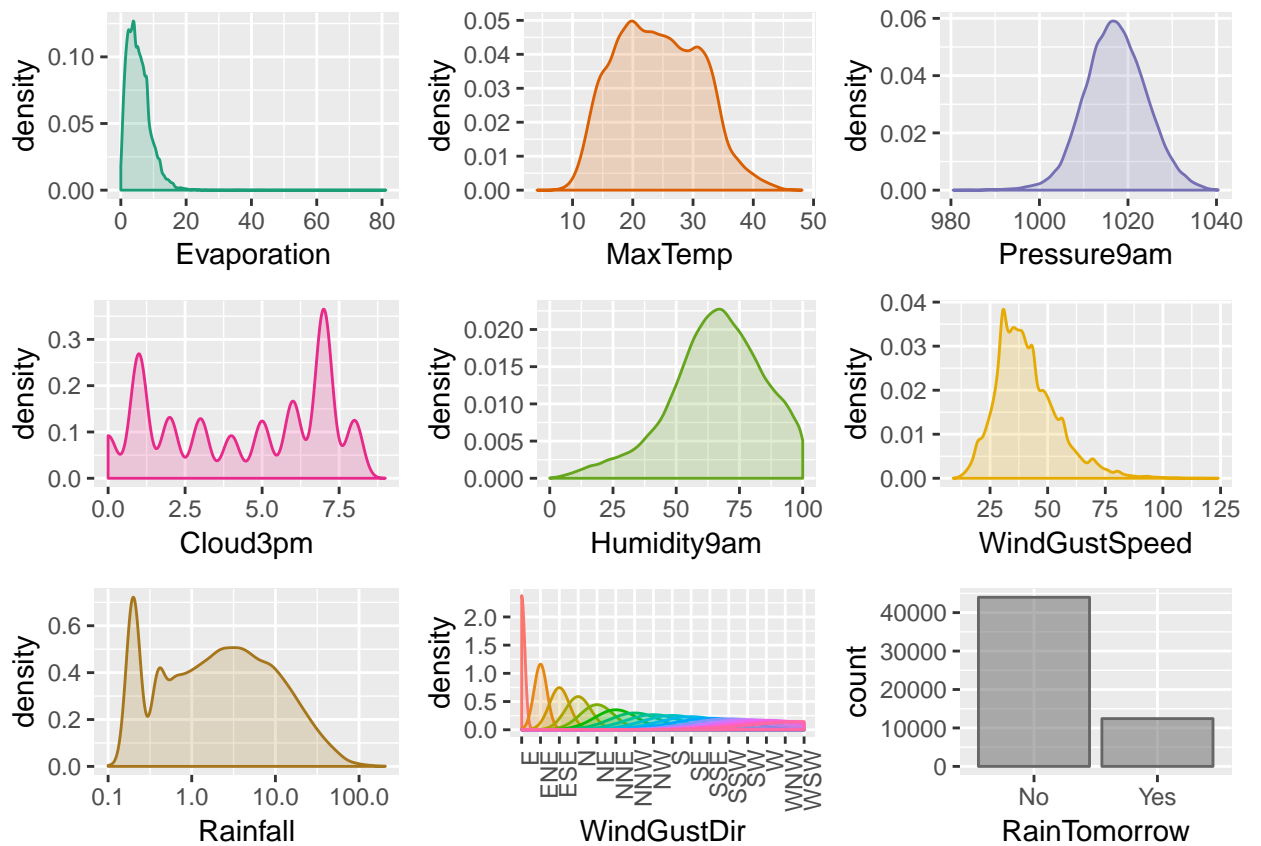Next set of plots renders distribution of a few selected features.

Figure 2: Observations Distribution

Missing Data

Further analysis of data shows that many features are missing. Some data losses are very significant. We are going to identify what data is missing and if it is feasible to recover the data.

```
print(sort(colSums(is.na(weatherData)), decreasing = T))
```

```
#>     Sunshine   Evaporation      Cloud3pm      Cloud9am    Pressure9am
#>        67816         60843         57094         53657          14014
#>    Pressure3pm    WindDir9am    WindGustDir  WindGustSpeed    WindDir3pm
#>         13981         10013          9330          9270           3778
#>    Humidity3pm       Temp3pm    WindSpeed3pm    Humidity9am      Rainfall
#>          3610          2726          2630          1774          1406
#>      RainToday    WindSpeed9am       Temp9am       MinTemp       MaxTemp
#>          1406          1348           904           637           322
#>           Date      Location   RainTomorrow
#>              0             0             0
```

To speed up data processing and plot rendering we are going to use a data sample. For population of 142K observations, 20K sample size would be sufficient for 99% confidence level with the confidence interval 1.

```
weatherSample = sample_n(weatherData, SampleSize)
aggr(weatherSample, numbers = F, prop = T, col = mainPalette, sortVars = T, bars = F, varheight = T)
```
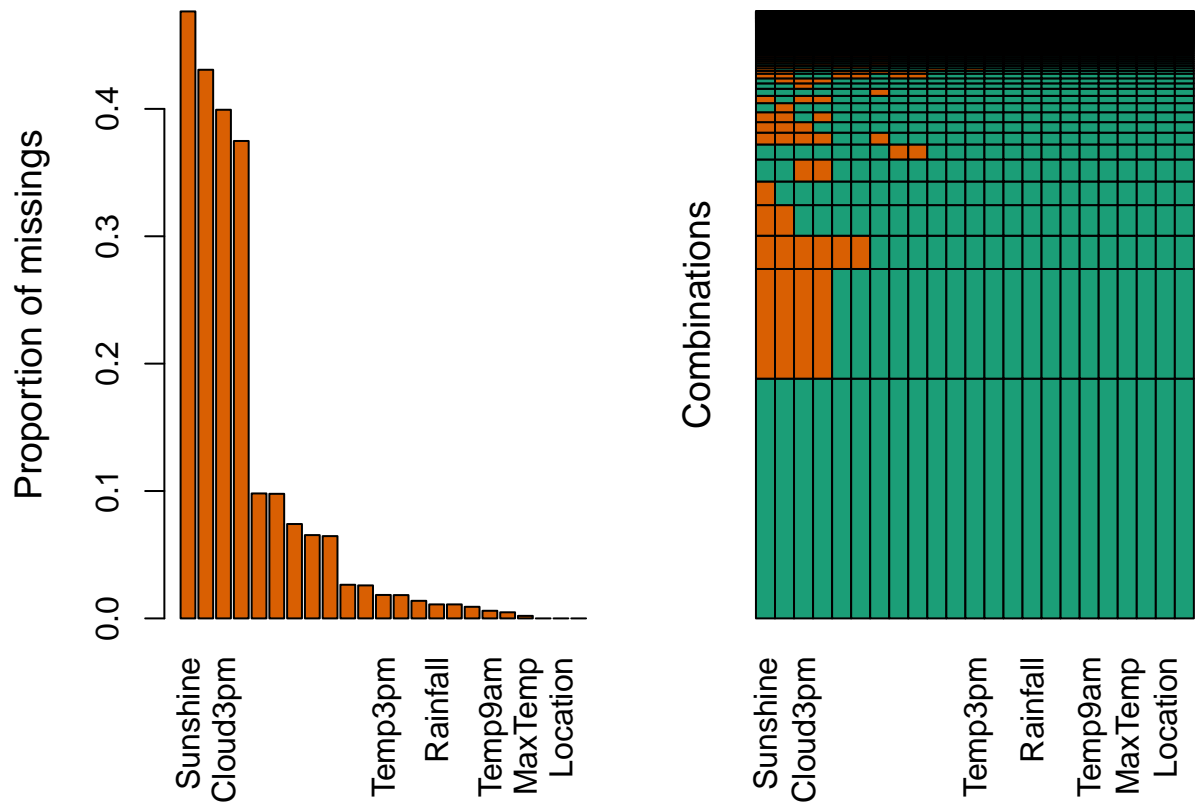
Figure 3: Missing Data Summary

```
#>
#>  Variables sorted by number of missings:
#>        Variable  Count
#>        Sunshine 0.4765
#>     Evaporation 0.4307
#>        Cloud3pm 0.3993
#>        Cloud9am 0.3748
#>      Pressure3pm 0.0981
#>      Pressure9am 0.0978
#>       WindDir9am 0.0741
#>      WindGustDir 0.0654
#>    WindGustSpeed 0.0646
#>      Humidity3pm 0.0264
#>       WindDir3pm 0.0259
#>          Temp3pm 0.0184
#>      WindSpeed3pm 0.0183
#>      Humidity9am 0.0138
#>         Rainfall 0.0110
#>        RainToday 0.0110
#>      WindSpeed9am 0.0091
#>          Temp9am 0.0060
#>          MinTemp 0.0048
#>          MaxTemp 0.0020
#>             Date 0.0000
#>         Location 0.0000
#>      RainTomorrow 0.0000
```
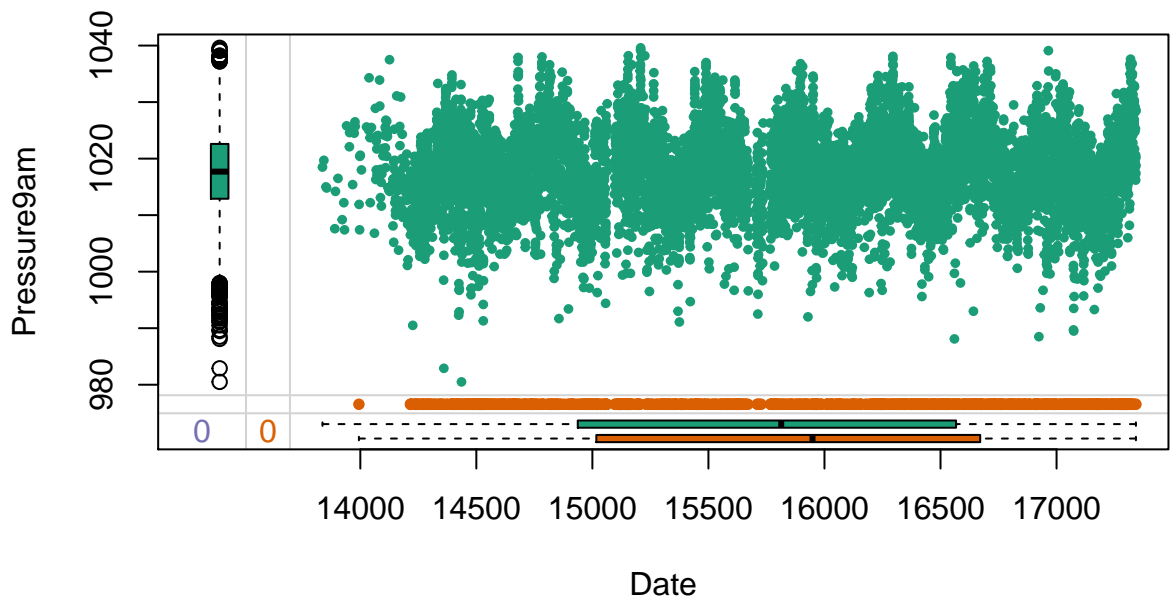
As demonstrated in Figure 3 *Sunshine*, *Evaporation* and *Clouds* columns safer the loss of data between **48%** and **38%**. This is significant! Since we are dealing with the weather patterns we should be observing cyclical data patterns. Let's review data distribution of features that damaged the most.
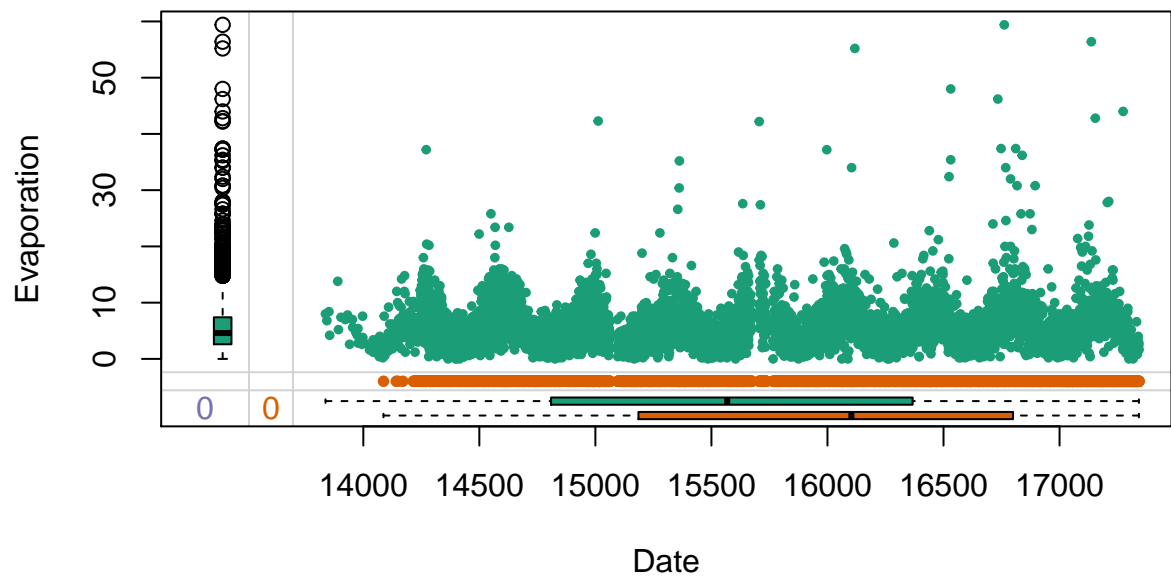
Figure 4: Date/Evaporation Margin Plot
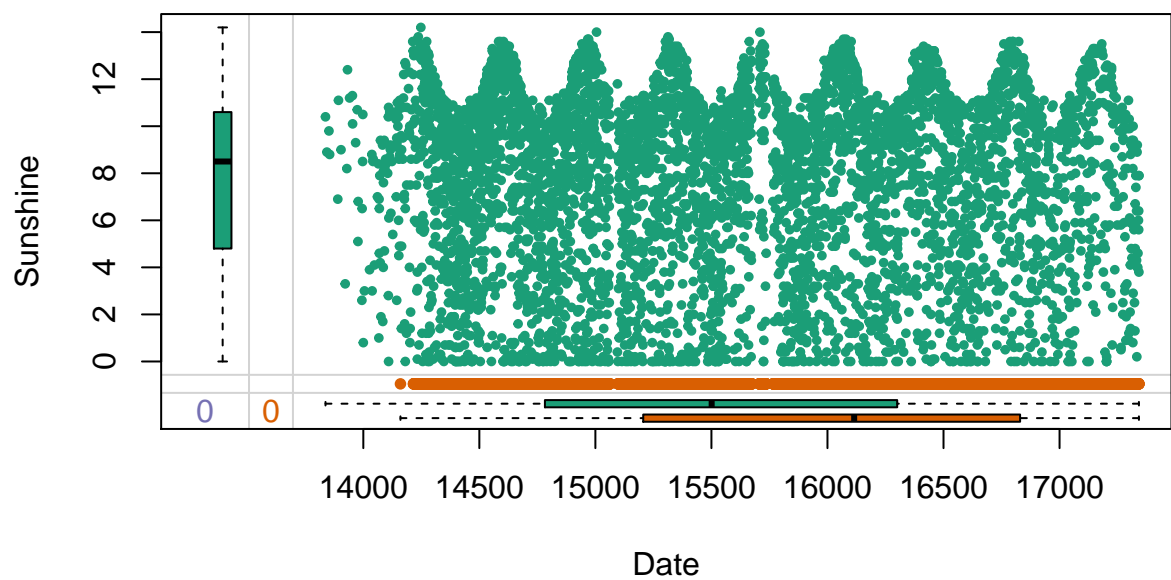
Figure 5: Date/ Sunshine Margin Plot

Figure 6: Date/ Pressure3pm Margin Plot

So what do the margin plots tell us? First of all let's take a look at *Date* axis. The *Date* has been converted to number to ensure continuous flow of the data . All features we picked exhibit cyclical pattern as expected. Along the vertical axis we observe the box plot of the respective feature. *Evaporaton* data is quite remarkable (Figure 4); it has very narrow distribution and a lot of so-called outliers. Though forces of nature follow seasonal patters they often exhibit wide range of seasonal anomalies, which the plots highlight. The distribution of the missing data of a given feature is depicted along the horizontal axis. In all three cases the missing data is randomly distributed along observed date range. Along the horizontal axis we may see box plots of the date and a given feature. *Presure9am* ((Figure 6)) distributed evenly across the observed date frame. *Evaporation* and *Sunshine* exhibit more data losses towards the end of the observed period

Let's examine one more dimension of the missing data, namely features vs feature vs location
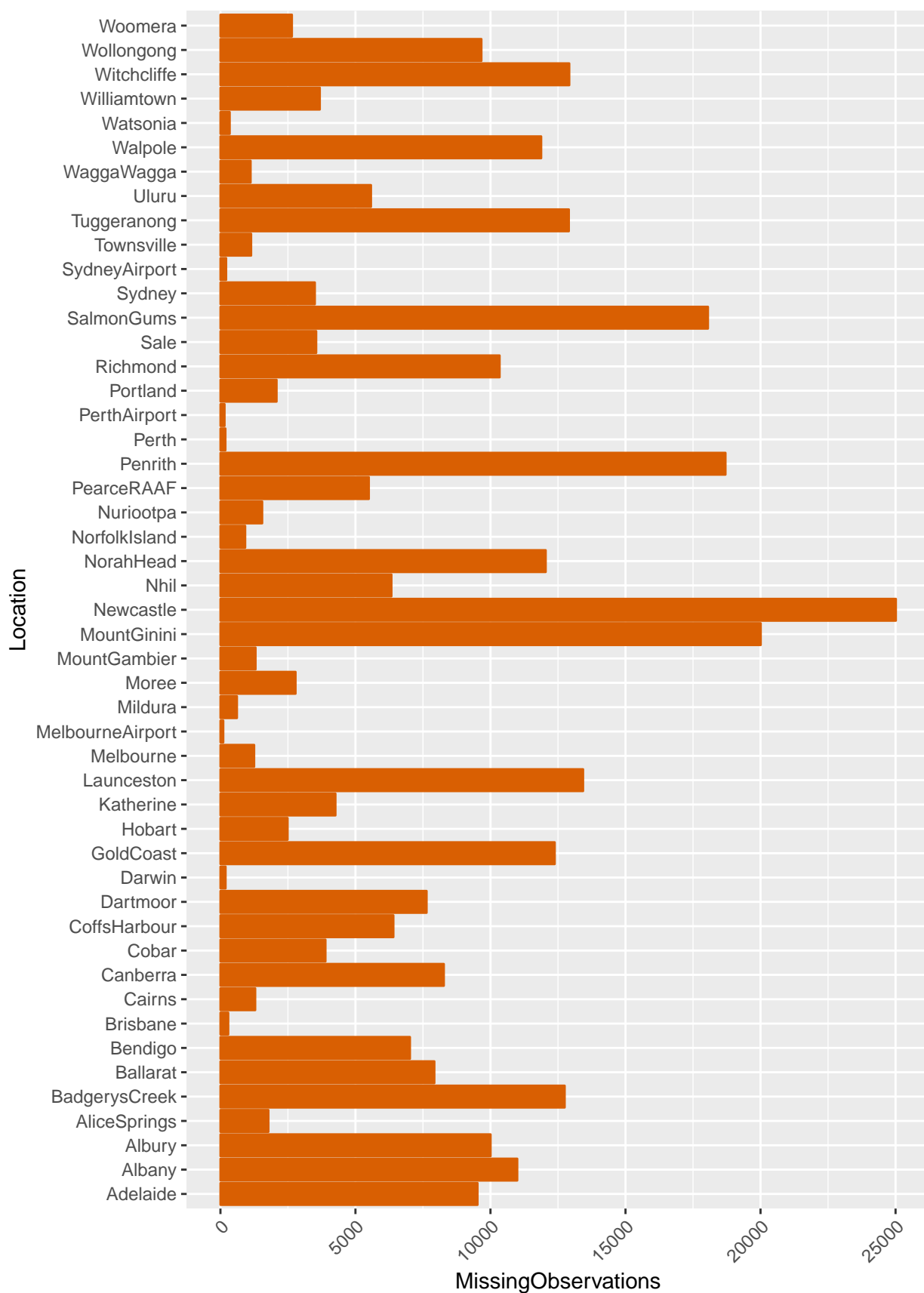
Figure 7: Missing Data By Location

Remarkably Figure 7 shows that **6460** observations are missing on average per location. Though if we take a second look at the weather station map 1 we would see that Mount Gini (the station that

miss the most data), Bendigo and Ballarat are close to Melbrun, where the staff has kept observing data on regular basis. Newcastle to Sydney and so on. . .

Data correlation and other observations

Let's examine how the features are correlated to each other. Knowing weather we can make an accurate prediction that the temperature features should be highly correlated, as well as pressure, wind speed, clouds and humidity groups
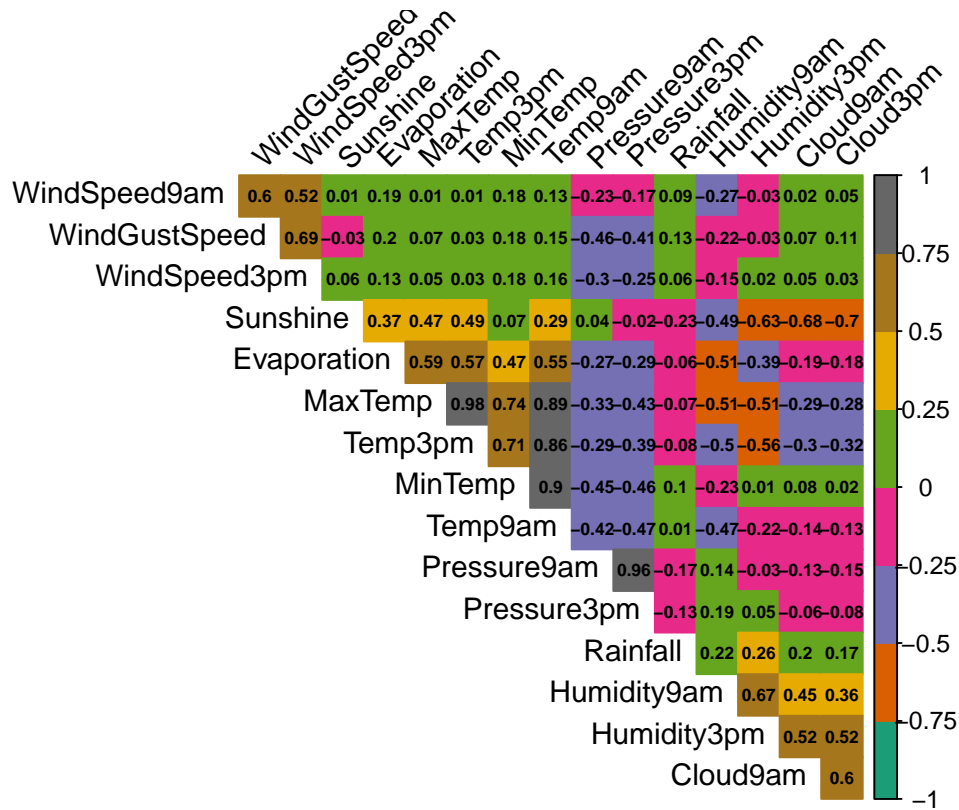

Figure 8: Data Correlation

Figure 8 confirms our initial guess. This observation will help us to eliminate redundant features later when we get to the point of selecting useful predictors for our model

Takeaways from Data Exploration Excersize

- The data we are dealing with suffers major observation losses (Figure 3)
- The least represented features are
- *Sunshine* **48%**
- *Evaporation* **43%**
- *Cloud* group (**40%** and **38%** respectively)
- The rest of the features exhibit medium to minor data losses, where *Pressure* group leads the way with 10%
- The missing data is distributed randomly over the observed time frame (Figures 4, 5, 6)
- We also witnessed that some weather stations recorded less data and some were almost prefect at record keeping (Figure 7). Luckily many majority weather stations situate relatively close to each other (see figure 1). Thus if a station has data gaps the neighboring station data could be used to approximate the missing data with plausible accuracy
- We have also noticed that many features are either positively or negatively correlated (Figure 8), where
- *MaxTemp*, *Temp3pm* and *Temp9am* exhibits correlation of **0.86** to **0.98**
- *Pressure9am* and *Pressure3pm* have correlation coefficient of **0.96**
- *Sunshine* and *Cloud* group correlated negatively with coefficient of **-0.7**
- *Rainfall* feature is of particular interest since this is what we are trying to predict. Unfortunately it does not demonstrate any strong correlations with any other feature

- Doing the data analysis we have also seen seasonal patterns and data that fall outside of the normal distribution range by far (outliers). Those are anomalies of nature.
- The last but not least the target feature (the value we are trying to predict) is unbalanced. so we are dealing with unbalanced data set. See Figure 2 *RainTomorrow* plot

Data Preparation

Data exploration confirmed that despite of significant data loss we should be able to impute data with high degree of plausibility

Datas Imputing

Before we start dealing with missing observations let's do some feature engineering, which will + improve imputation processing speed + improve model training performance and hopefully accuracy

First of all let's get rid of *Date* column. Outside of the presentation it does not carry too mach information. What would be useful indeed is a feature that captures seasonal observation fluctuations. That would bee *month* and *day* combined, giving us year-round (365) days of observations

Secondly we convert categorical features to numbers. But before we do so we would like to ponder about *Location*. We have couple options here. Either we convert the locations to the numbers or we can replace them with the real geographical coordinates. After some deliberation we can conclude that the coordinates will not add too much knowledge in the context of the model training. But they will certainly break this categorical feature (coordinates have 4,6 decimal places, which effectively make them continuous). So we stick with categories.

This is our original set:

```
str(weatherData)

#> 'data.frame':    142193 obs. of  23 variables:
#>  $ Date         : Date, format: "2008-12-01" "2008-12-02" ...
#>  $ Location     : Factor w/ 49 levels "Adelaide","Albany",..: 3 3 3 3 3 3 3 3 3 3 ...
#>  $ MinTemp      : num  13.4 7.4 12.9 9.2 17.5 14.6 14.3 7.7 9.7 13.1 ...
#>  $ MaxTemp      : num  22.9 25.1 25.7 28 32.3 29.7 25 26.7 31.9 30.1 ...
#>  $ Rainfall     : num  0.6 0 0 0 1 0.2 0 0 0 1.4 ...
#>  $ Evaporation  : num  NA NA NA NA NA NA NA NA NA NA ...
#>  $ Sunshine     : num  NA NA NA NA NA NA NA NA NA NA ...
#>  $ WindGustDir  : Factor w/ 16 levels "E","ENE","ESE",..: 14 15 16 5 14 15 14 14 7 14 ...
#>  $ WindGustSpeed: int  44 44 46 24 41 56 50 35 80 28 ...
#>  $ WindDir9am   : Factor w/ 16 levels "E","ENE","ESE",..: 14 7 14 10 2 14 13 11 10 9 ...
#>  $ WindDir3pm   : Factor w/ 16 levels "E","ENE","ESE",..: 15 16 16 1 8 14 14 14 8 11 ...
#>  $ WindSpeed9am : int  20 4 19 11 7 19 20 6 7 15 ...
#>  $ WindSpeed3pm : int  24 22 26 9 20 24 24 17 28 11 ...
#>  $ Humidity9am  : int  71 44 38 45 82 55 49 48 42 58 ...
#>  $ Humidity3pm  : int  22 25 30 16 33 23 19 19 9 27 ...
#>  $ Pressure9am  : num  1008 1011 1008 1018 1011 ...
#>  $ Pressure3pm  : num  1007 1008 1009 1013 1006 ...
#>  $ Cloud9am     : int  8 NA NA NA 7 NA 1 NA NA NA ...
#>  $ Cloud3pm     : int  NA NA 2 NA 8 NA NA NA NA NA ...
#>  $ Temp9am      : num  16.9 17.2 21 18.1 17.8 20.6 18.1 16.3 18.3 20.1 ...
#>  $ Temp3pm      : num  21.8 24.3 23.2 26.5 29.7 28.9 24.6 25.5 30.2 28.2 ...
#>  $ RainToday    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 2 ...
#>  $ RainTomorrow : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 2 1 ...
```

Transformation

```
data = mutate(weatherData,MMDD = as.numeric( format(Date, "%m%d")),Location = unclass(Location),
           WindGustDir = unclass(WindGustDir),
           WindDir9am = unclass(WindDir9am), WindDir3pm = unclass(WindDir3pm),
           RainToday = unclass(RainToday)-1, RainTomorrow = unclass(RainTomorrow)-1)
data =  subset(data, select = -Date)
```

Resulting data frame structure:

```
str(data)
```

```
#> 'data.frame':    142193 obs. of  23 variables:
#>  $ Location    : int  3 3 3 3 3 3 3 3 3 3 ...
#>  $ MinTemp     : num  13.4 7.4 12.9 9.2 17.5 14.6 14.3 7.7 9.7 13.1 ...
#>  $ MaxTemp     : num  22.9 25.1 25.7 28 32.3 29.7 25 26.7 31.9 30.1 ...
#>  $ Rainfall    : num  0.6 0 0 0 1 0.2 0 0 0 1.4 ...
#>  $ Evaporation : num  NA NA NA NA NA NA NA NA NA NA ...
#>  $ Sunshine    : num  NA NA NA NA NA NA NA NA NA NA ...
#>  $ WindGustDir : int  14 15 16 5 14 15 14 14 7 14 ...
#>  $ WindGustSpeed: int  44 44 46 24 41 56 50 35 80 28 ...
#>  $ WindDir9am  : int  14 7 14 10 2 14 13 11 10 9 ...
#>  $ WindDir3pm  : int  15 16 16 1 8 14 14 14 8 11 ...
#>  $ WindSpeed9am : int  20 4 19 11 7 19 20 6 7 15 ...
#>  $ WindSpeed3pm : int  24 22 26 9 20 24 24 17 28 11 ...
#>  $ Humidity9am : int  71 44 38 45 82 55 49 48 42 58 ...
#>  $ Humidity3pm : int  22 25 30 16 33 23 19 19 9 27 ...
#>  $ Pressure9am : num  1008 1011 1008 1018 1011 ...
#>  $ Pressure3pm : num  1007 1008 1009 1013 1006 ...
#>  $ Cloud9am    : int  8 NA NA NA 7 NA 1 NA NA NA ...
#>  $ Cloud3pm    : int  NA NA 2 NA 8 NA NA NA NA NA ...
#>  $ Temp9am     : num  16.9 17.2 21 18.1 17.8 20.6 18.1 16.3 18.3 20.1 ...
#>  $ Temp3pm     : num  21.8 24.3 23.2 26.5 29.7 28.9 24.6 25.5 30.2 28.2 ...
#>  $ RainToday   : num  0 0 0 0 0 0 0 0 0 1 ...
#>  $ RainTomorrow : num  0 0 0 0 0 0 0 0 1 0 ...
#>  $ MMDD        : num  1201 1202 1203 1204 1205 ...
```

To impute the missing data we employ **MICE** package. Our imputation strategy is to employ **Predictive mean matching** model which is a robust, fast imputation algorithm that works with numeric values (this is why we have converted all data to the numeric values) Lets do a dry run first to see what predictors and methods for each feature to cure *MICE* software chooses. As before we will be working with a 20K data sample. Imputation process on the whole set take about 3 hours and 20 minutes to complete! In addition we let *MICE* to choose predictors for us running **quickpred()** method

```
meta = mice(data, maxit = 0, print = FALSE)
weatherSample = sample_n(data, SampleSize)
methods = meta$method
predictors = quickpred(data)
```

Let's review the methods chosen by the software making sure that they meet our requirements highlighted prior in the imputation strategy paragraph

```
print(methods)
```

```
#>      Location      MinTemp      MaxTemp      Rainfall   Evaporation
#>            ""        "pmm"        "pmm"        "pmm"        "pmm"
#>      Sunshine  WindGustDir WindGustSpeed   WindDir9am   WindDir3pm
#>         "pmm"        "pmm"        "pmm"        "pmm"        "pmm"
#>  WindSpeed9am WindSpeed3pm  Humidity9am  Humidity3pm  Pressure9am
#>         "pmm"        "pmm"        "pmm"        "pmm"        "pmm"
#>   Pressure3pm     Cloud9am     Cloud3pm      Temp9am      Temp3pm
#>         "pmm"        "pmm"        "pmm"        "pmm"        "pmm"
#>     RainToday RainTomorrow         MMDD
#>         "pmm"           ""           ""
```

The code output above shows that 1 the features without missing data will not be imputed 2 The imputation targets will all be treated with *Predictive mean matching* algorithm ("pmm")

This is exactly what we need. Now let's review the predictors (*The command output is not included into report to save space*)

The matrix of predictors has the predictors in the columns and the features to be imputed in the rows. If the cell value equals has **1** the predictor will be employed in calculations for the respective imputation target. Surprisingly **MMDD** is not used widely to predict the missing data, nether do the **Location**.

Now we are going to start the imputation process. **Note: it might take about 4 - 5 minutes even for a smaple**. We have disabled the output of the function as we do not want to pollute the report with irrelevant messages

```
imputed = mice(weatherSample, pred = predictors, meth = methods, seed = 38019,
               nnet.MaxNWts = 2000, printFlag = F)
```

Now it is time to analyze the imputed values. In general, a good imputed value is a value that could have been observed had it not been missing. The MAR assumption can never be tested from the observed data. To check whether the imputations created by **MICE** algorithm are plausible we employ density charts and compare the distribution of the imputed values vs real observations. Let's do this (*again the plots take time, patience. . .*).
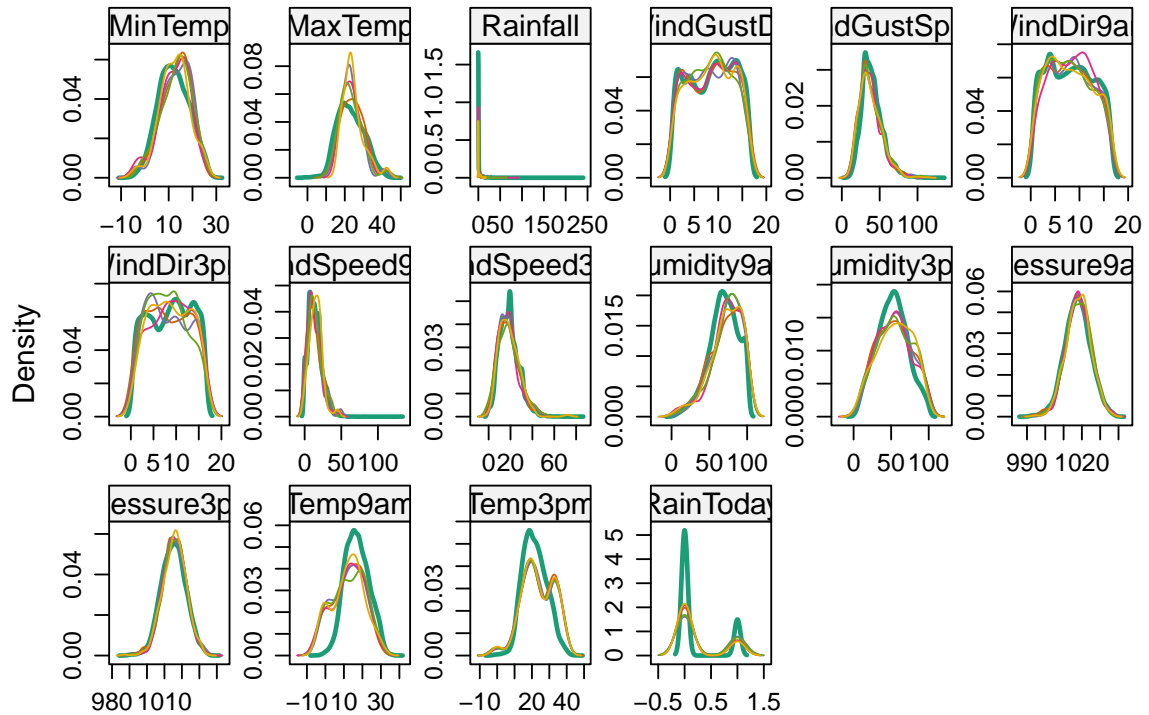

Figure 9: Imputed Values Distribution vs Real Observations

Figure 9 illustrates imputed value distribution for each imputed feature vs observed data. The fat green line renders the real data distribution and the thin lines of the other colors the distribution of imputed values after each imputation cycle (*there are five of them by default*). Where the last one is yellow. The yellow line should be shadowing the contour of the green one as close as possible, which give us an indication that the result of the imputation is plausible. Looking at the charts we can conclude that the imputation has been successful! Let's apply imputed values to our sample set and verify if there are any *NAs* left

```
weatherSample = complete(imputed)
print(colSums(is.na(weatherSample)))

#>       Location       MinTemp       MaxTemp      Rainfall    Evaporation
#>              0             0             0             0             0
#>       Sunshine   WindGustDir  WindGustSpeed     WindDir9am     WindDir3pm
#>              0             0             0             0             0
#>   WindSpeed9am  WindSpeed3pm    Humidity9am    Humidity3pm    Pressure9am
#>              0             0             0             0             0
#>     Pressure3pm       Cloud9am       Cloud3pm        Temp9am        Temp3pm
#>              0             0             0             0             0
#>       RainToday   RainTomorrow           MMDD
#>              0             0             0
```

Outstanding! There are no missing values. Now we move on to the next part - model training

Modeling and Evalutation

Finally we have reached the stage where we can start training and evaluating classification models. At this point we have clear understanding of our data. We have gotten rid of the features that did not present much value. We have filled the gaps in our data set employing sophisticated imputation technique.

Feature Selection

The weather observation data set originally had 24 features. We have removed *RISK_MM* and *Date* as explained earlier and added *MMDD*. Now the data set has 22 features and one label. Let's see if we can reduce the number of predictors without significant information loss. This would make our models faster and more interpretable for users. We shall keep in mind that at the data exploration phase we have discovered that many features are correlated (Figure 8). hopefully this knowledge will help us identify and remove redundant features.

Generally speaking feature evaluation methods can be separated into two groups: those that use the model information and those that do not. Clearly at this stage the models are not ready. Thus we will be exploring the methods that do not require model.

This group of the method could be spit further as follows:

- wrapper methods that evaluate multiple models adding and/or removing predictors. These are some examples:
- recursive feature elimination
- genetic algorithms
- simulated annealing
- filter methods which evaluate the relevance of the predictors outside of the predictive models.

The evaluation of various feature selection methods is not in the scope of this paper. Thus we opt for a recursive feature elimination method using accuracy as a target metric.

Before we precede any further let's ensure that all categorical values get converted to factors. This is useful for dimentiality reduction algorithms and model training.

```
weatherSample = mutate(weatherSample, Location = as.factor(unclass(Location)),
        WindGustDir = as.factor(unclass(WindGustDir)),
        WindDir9am = as.factor(unclass(WindDir9am)), WindDir3pm = as.factor(unclass(WindDir3pm)),
        RainToday = as.factor(unclass(RainToday)), RainTomorrow = as.factor(unclass(RainTomorrow)))
```

Let's run feature selection algorithm

```
predictors = subset(weatherSample,select = -RainTomorrow)
label = weatherSample[,22]

# run the RFE algorithm
rfePrediction = rfe(predictors, label, sizes=c(1:22),
                rfeControl = rfeControl(functions=rfFuncs, method="cv", number=3))
print(rfePrediction)

#>
#> Recursive feature selection
#>
#> Outer resampling method: Cross-Validated (3 fold)
#>
#> Resampling performance over subset size:
#>
#>  Variables Accuracy  Kappa AccuracySD  KappaSD Selected
#>          1   0.8296 0.3712   0.005503 0.015389
#>          2   0.8138 0.3809   0.005559 0.006866
#>          3   0.8300 0.4463   0.009185 0.017079
#>          4   0.8384 0.4945   0.007918 0.025167
#>          5   0.8481 0.5328   0.009928 0.028622
#>          6   0.8501 0.5392   0.008946 0.024177
#>          7   0.8532 0.5467   0.010644 0.031833
#>          8   0.8557 0.5565   0.007065 0.018593
```

```
#>        9   0.8563 0.5534   0.007582 0.013759        *
#>       10   0.8527 0.5468   0.007113 0.015276
#>       11   0.8518 0.5438   0.006777 0.018905
#>       12   0.8544 0.5492   0.008728 0.024885
#>       13   0.8531 0.5430   0.006871 0.017090
#>       14   0.8540 0.5446   0.008272 0.018118
#>       15   0.8511 0.5398   0.007560 0.017842
#>       16   0.8529 0.5454   0.007392 0.018842
#>       17   0.8524 0.5449   0.007487 0.024293
#>       18   0.8527 0.5461   0.007066 0.022966
#>       19   0.8533 0.5476   0.008760 0.025445
#>       20   0.8540 0.5483   0.008623 0.023708
#>       21   0.8542 0.5459   0.006517 0.017888
#>       22   0.8523 0.5394   0.007097 0.019002
#>
#> The top 5 variables (out of 9):
#>    Humidity3pm, Sunshine, WindGustSpeed, Location, Pressure3pm
```
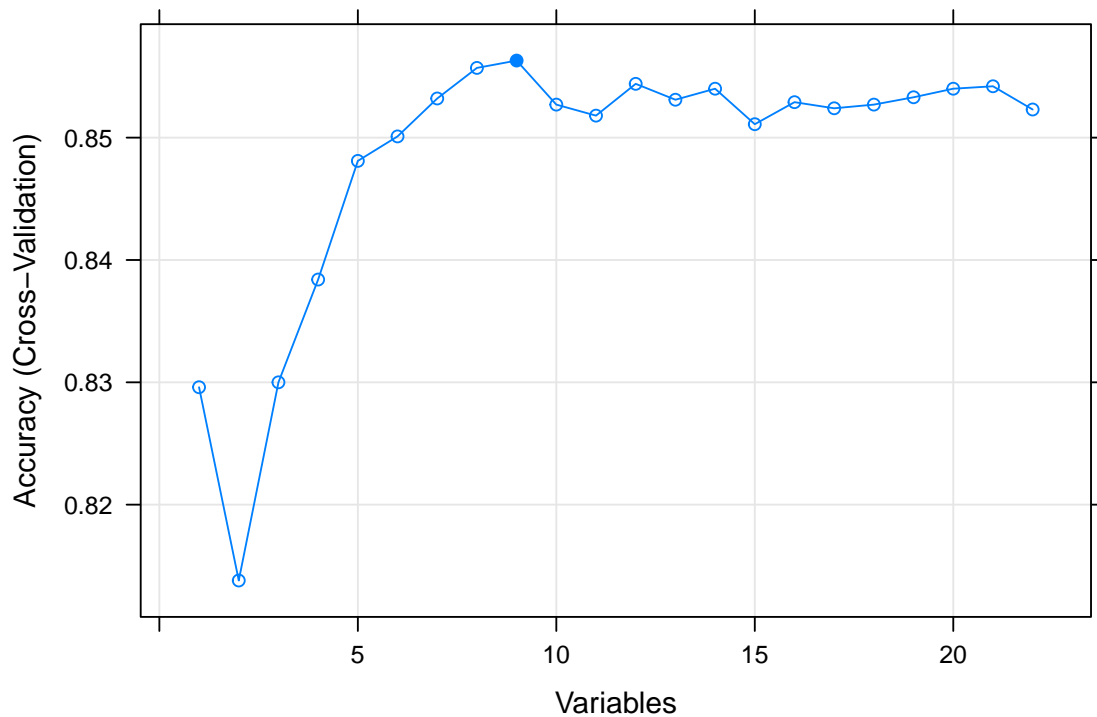

Figure 10: Number of Predictors vs Accuracy

Figure 10 shows that accuracy peaks a few times: with 9 predictors, 14 and tops at 22.The accuracy gain between 9 and 22 is negligible. Here is the list of features ordered by importance. We take first nine for model training.

```
print(predictors(rfePrediction))

#> [1] "Humidity3pm"   "Sunshine"      "WindGustSpeed" "Location"
#> [5] "Pressure3pm"   "Cloud3pm"      "Pressure9am"   "Rainfall"
#> [9] "Humidity9am"

l = length(predictors(rfePrediction))
selectedPredictors =  predictors(rfePrediction)[1:ifelse(l<9,l,9)]
# remove useless variables
rm(label,predictors,rfePrediction,l)
```

Data Upsampling

There is one more step before we get to the model training. As shown in Figure 2 our data set is unbalanced. This could cause model over-fitting. So let's split the data into the training and testing sets and up-sample the training set

```
#>
#>    0    1
#> 5473 5473
```

As we can see the training set is balanced.

Thus we have prepared our training and test data sets. We have identified the most important features. We are ready to work on the prediction models

Classification Tree Model

```
#> Conditional Inference Tree
#>
#> 10946 samples
#>     9 predictor
#>     2 classes: 'no', 'yes'
#>
#> No pre-processing
#> Resampling: Cross-Validated (5 fold)
#> Summary of sample sizes: 8757, 8756, 8758, 8756, 8757
#> Resampling results across tuning parameters:
#>
#>   mincriterion  ROC        Sens       Spec
#>   0.01          0.8832146  0.7833022  0.8298997
#>   0.50          0.8778256  0.7741586  0.8222255
#>   0.99          0.8609156  0.7615595  0.8013872
#>
#> ROC was used to select the optimal model using the largest value.
#> The final value used for the model was mincriterion = 0.01.

confusionMatrix(data = pred.classTreeModel.raw, testDataCopy$RainTomorrow)

#> Confusion Matrix and Statistics
#>
#>           Reference
#> Prediction   no  yes
#>        no  1779  189
#>        yes  566  465
#>
#>                Accuracy : 0.7482
#>                  95% CI : (0.7323, 0.7637)
#>     No Information Rate : 0.7819
#>     P-Value [Acc > NIR] : 1
#>
#>                   Kappa : 0.3888
#>  Mcnemar's Test P-Value : <2e-16
#>
#>             Sensitivity : 0.7586
#>             Specificity : 0.7110
#>          Pos Pred Value : 0.9040
#>          Neg Pred Value : 0.4510
#>              Prevalence : 0.7819
#>          Detection Rate : 0.5932
#>    Detection Prevalence : 0.6562
#>       Balanced Accuracy : 0.7348
#>
#>        'Positive' Class : no
#>
```
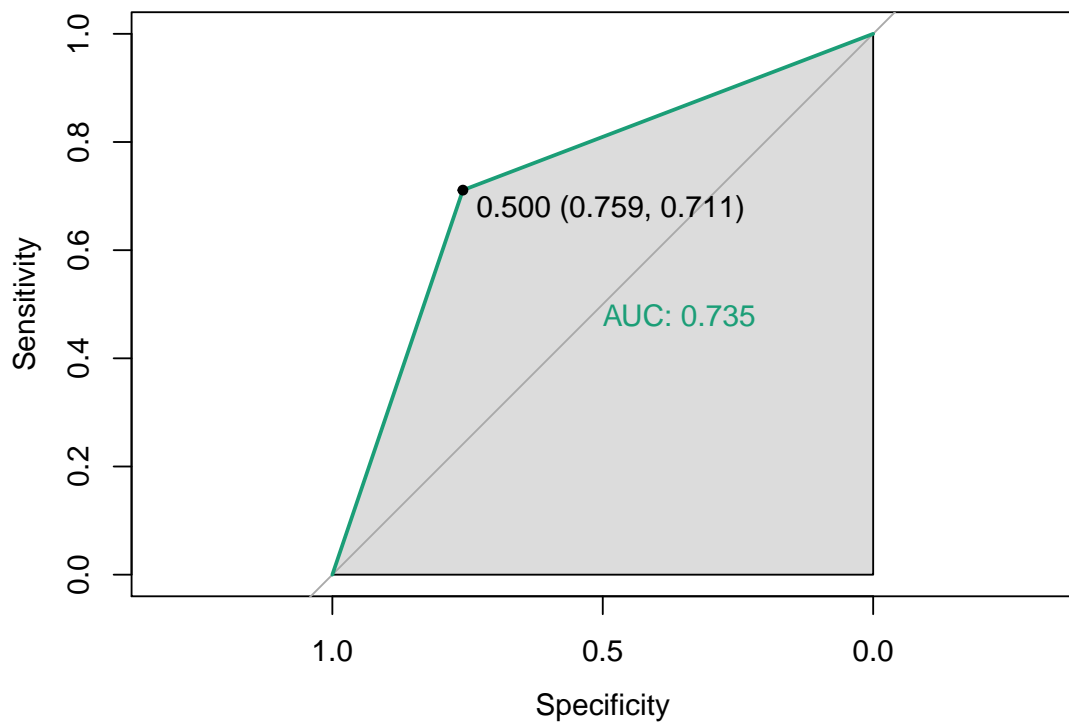
Figure 11: Classification Tree Model AUC and ROC Curve

Naive Bayes Model

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 24

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 86

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 109

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 132

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 136

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 154

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 155

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 371

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 422

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 558

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 574
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 589

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 651

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 702

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 721

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 789

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 826

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 876

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 908

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 974

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1002

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1035

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1048

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1134

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1156

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1184

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1235

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1248

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1316

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1369

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1388

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1401

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1423

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1443
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1547

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1736

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1773

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1879

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1891

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1957

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1984

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2024

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2050

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2054

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2097

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2169

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2182

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 24

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 86

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 109

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 132

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 136

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 154

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 155

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 371

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 422

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 558
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 574

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 589

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 651

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 702

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 721

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 789

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 826

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 876

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 908

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 974

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1002

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1035

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1048

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1134

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1156

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1184

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1235

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1248

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1316

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1369

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1388

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1401

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1423
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1443

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1547

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1736

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1773

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1879

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1891

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1957

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1984

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2024

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2050

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2054

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2097

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2169

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2182

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 10

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 47

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 72

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 79

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 94

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 127

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 144

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 190

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 211
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 236

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 319

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 335

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 343

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 347

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 350

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 351

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 401

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 408

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 443

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 450

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 518

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 525

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 532

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 533

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 601

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 651

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 661

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 682

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 700

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 738

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 757

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 792
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 808

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 820

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 862

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 871

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 887

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 941

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 949

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 960

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1107

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1156

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1179

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1238

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1278

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1310

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1315

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1316

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1340

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1493

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1622

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1682

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1695

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1708

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1942
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1979

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2002

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2066

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2169

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 10

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 47

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 72

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 79

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 94

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 127

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 144

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 190

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 211

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 236

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 319

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 335

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 343

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 347

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 350

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 351

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 401

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 408

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 443
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 450

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 518

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 525

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 532

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 533

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 601

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 651

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 661

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 682

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 700

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 738

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 757

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 792

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 808

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 820

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 862

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 871

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 887

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 941

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 949

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 960

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1107

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1156
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1179

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1238

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1278

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1310

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1315

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1316

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1340

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1493

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1622

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1682

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1695

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1708

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1942

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1979

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2002

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2066

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2169

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 39

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 69

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 79

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 223

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 260

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 261
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 276

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 282

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 296

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 385

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 438

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 493

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 531

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 559

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 811

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 902

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 984

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1088

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1172

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1180

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1301

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1379

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1431

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1433

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1522

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1558

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1637

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1762

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1852
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1861

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1911

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1912

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1988

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2135

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2152

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2178

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 39

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 69

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 79

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 223

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 260

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 261

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 276

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 282

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 296

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 385

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 438

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 493

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 531

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 559

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 811

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 902
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 984

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1088

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1172

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1180

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1301

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1379

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1431

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1433

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1522

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1558

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1637

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1762

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1852

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1861

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1911

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1912

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1988

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2135

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2152

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2178

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 26

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 143

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 246
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 301

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 571

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 772

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 789

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 814

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 836

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1143

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1164

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1179

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1195

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1313

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1360

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1464

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1484

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1624

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1817

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1869

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1880

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1939

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1951

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2006

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2092

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2171
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 26

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 143

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 246

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 301

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 571

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 772

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 789

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 814

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 836

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1143

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1164

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1179

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1195

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1313

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1360

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1464

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1484

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1624

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1817

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1869

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1880

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1939

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1951
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2006

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2092

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2171

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 110

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 195

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 237

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 252

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 324

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 355

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 377

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 460

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 520

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 525

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 540

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 544

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 687

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 737

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 759

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 765

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 825

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 837

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 930

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 991
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1065

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1077

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1086

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1174

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1241

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1259

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1269

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1321

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1367

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1549

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1655

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1770

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1775

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1838

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1876

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1930

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2075

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2110

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2186

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 110

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 195

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 237

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 252
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 324

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 355

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 377

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 460

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 520

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 525

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 540

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 544

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 687

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 737

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 759

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 765

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 825

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 837

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 930

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 991

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1065

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1077

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1086

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1174

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1241

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1259

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1269
```

```
#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1321

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1367

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1549

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1655

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1770

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1775

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1838

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1876

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 1930

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2075

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2110

#> Warning in FUN(X[[i]], ...): Numerical 0 probability for all classes with
#> observation 2186

#> Naive Bayes
#>
#> 10946 samples
#>     9 predictor
#>     2 classes: 'no', 'yes'
#>
#> No pre-processing
#> Resampling: Cross-Validated (5 fold)
#> Summary of sample sizes: 8757, 8756, 8758, 8756, 8757
#> Resampling results across tuning parameters:
#>
#>   usekernel  ROC        Sens       Spec
#>   FALSE      0.7429773  0.5130689  0.8101642
#>    TRUE      0.8546820  0.6846370  0.8375648
#>
#> Tuning parameter 'fL' was held constant at a value of 0
#> Tuning
#>  parameter 'adjust' was held constant at a value of 1
#> ROC was used to select the optimal model using the largest value.
#> The final values used for the model were fL = 0, usekernel = TRUE
#>  and adjust = 1.

confusionMatrix(data = pred.naiveBayesModel.raw, testDataCopy$RainTomorrow)

#> Confusion Matrix and Statistics
#>
#>           Reference
#> Prediction   no  yes
#>        no  1574  115
#>        yes  771  539
#>
```
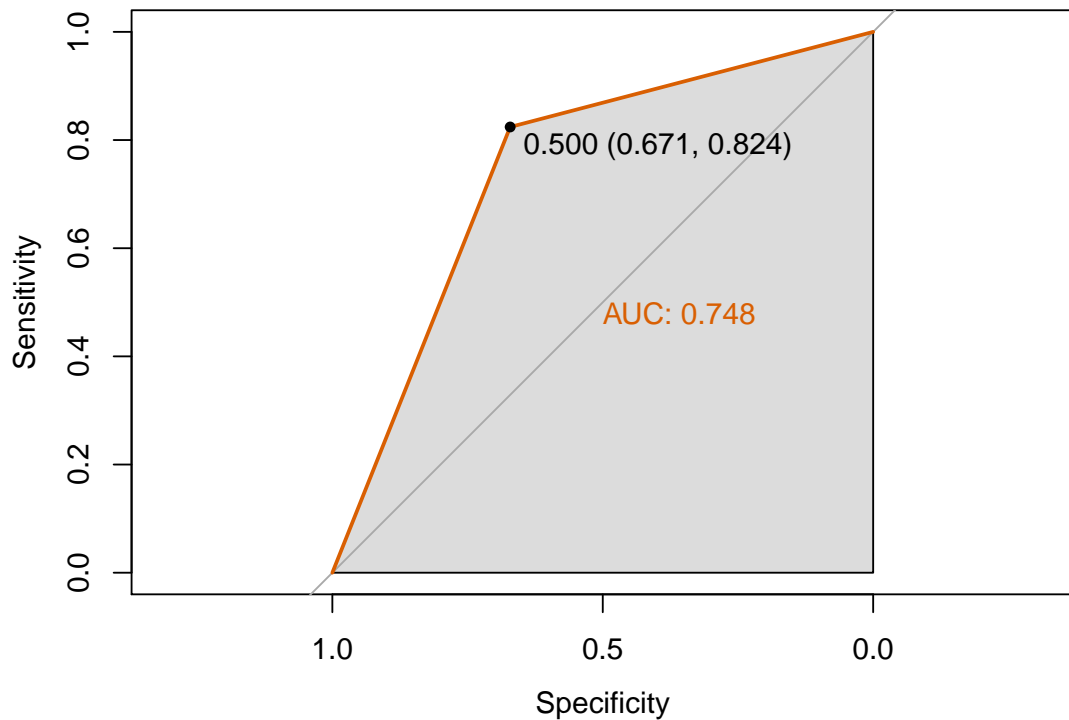
```
#>                 Accuracy : 0.7046
#>                   95% CI : (0.6879, 0.7209)
#>      No Information Rate : 0.7819
#>      P-Value [Acc > NIR] : 1
#>
#>                    Kappa : 0.3638
#>  Mcnemar's Test P-Value : <2e-16
#>
#>              Sensitivity : 0.6712
#>              Specificity : 0.8242
#>           Pos Pred Value : 0.9319
#>           Neg Pred Value : 0.4115
#>               Prevalence : 0.7819
#>           Detection Rate : 0.5248
#>     Detection Prevalence : 0.5632
#>        Balanced Accuracy : 0.7477
#>
#>         'Positive' Class : no
#>
```


Figure 12: Naive Bayes Model AUC and ROC Curve

Random Forest Model

```
#> Random Forest
#>
#> 10946 samples
#>     9 predictor
#>     2 classes: 'no', 'yes'
#>
#> No pre-processing
#> Resampling: Cross-Validated (2 fold)
#> Summary of sample sizes: 5473, 5473
#> Resampling results across tuning parameters:
#>
```

```
#>   mtry  ROC        Sens       Spec
#>    2    0.8913179  0.7949943  0.8238607
#>   29    0.9649512  0.8658873  0.9342221
#>   56    0.9626592  0.8594914  0.9347701
#>
#> ROC was used to select the optimal model using the largest value.
#> The final value used for the model was mtry = 29.

confusionMatrix(data = pred.randomForestModel.raw, testDataCopy$RainTomorrow)

#> Confusion Matrix and Statistics
#>
#>           Reference
#> Prediction  no   yes
#>        no  2143  271
#>        yes  202  383
#>
#>                Accuracy : 0.8423
#>                  95% CI : (0.8287, 0.8552)
#>     No Information Rate : 0.7819
#>     P-Value [Acc > NIR] : < 2.2e-16
#>
#>                   Kappa : 0.5192
#>  Mcnemar's Test P-Value : 0.001768
#>
#>             Sensitivity : 0.9139
#>             Specificity : 0.5856
#>          Pos Pred Value : 0.8877
#>          Neg Pred Value : 0.6547
#>              Prevalence : 0.7819
#>          Detection Rate : 0.7146
#>    Detection Prevalence : 0.8049
#>       Balanced Accuracy : 0.7497
#>
#>        'Positive' Class : no
#>
```
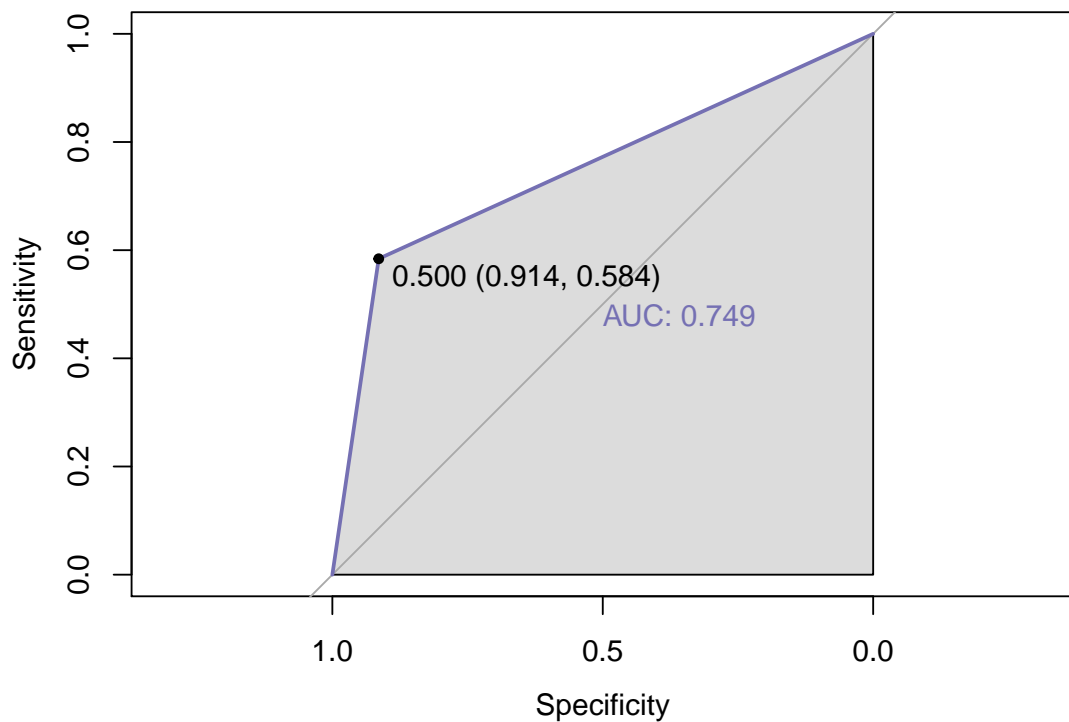
Figure 13: Random Forest Model AUC and ROC Curve

Logistic Regression Model

Logistic regression prediction accuracy will benefit if the data is normalized and are close to Gaussian distribution. Thus we apply addition transformation to the training data set. We will also be employing 5-fold cross validation resampling procedure to improve the model. In addition to the above we are going to convert categorical values (factors) to numeric data type

```
confusionMatrix(data = pred.logRegModel.raw, testDataCopy$RainTomorrow)

#> Confusion Matrix and Statistics
#>
#>           Reference
#> Prediction    0    1
#>          0 1861  150
#>          1  484  504
#>
#>                Accuracy : 0.7886
#>                  95% CI : (0.7735, 0.8031)
#>     No Information Rate : 0.7819
#>     P-Value [Acc > NIR] : 0.1946
#>
#>                   Kappa : 0.4765
#>  Mcnemar's Test P-Value : <2e-16
#>
#>             Sensitivity : 0.7936
#>             Specificity : 0.7706
#>          Pos Pred Value : 0.9254
#>          Neg Pred Value : 0.5101
#>              Prevalence : 0.7819
#>          Detection Rate : 0.6205
#>    Detection Prevalence : 0.6706
#>       Balanced Accuracy : 0.7821
#>
```

```
#>          'Positive' Class : 0
#>
```
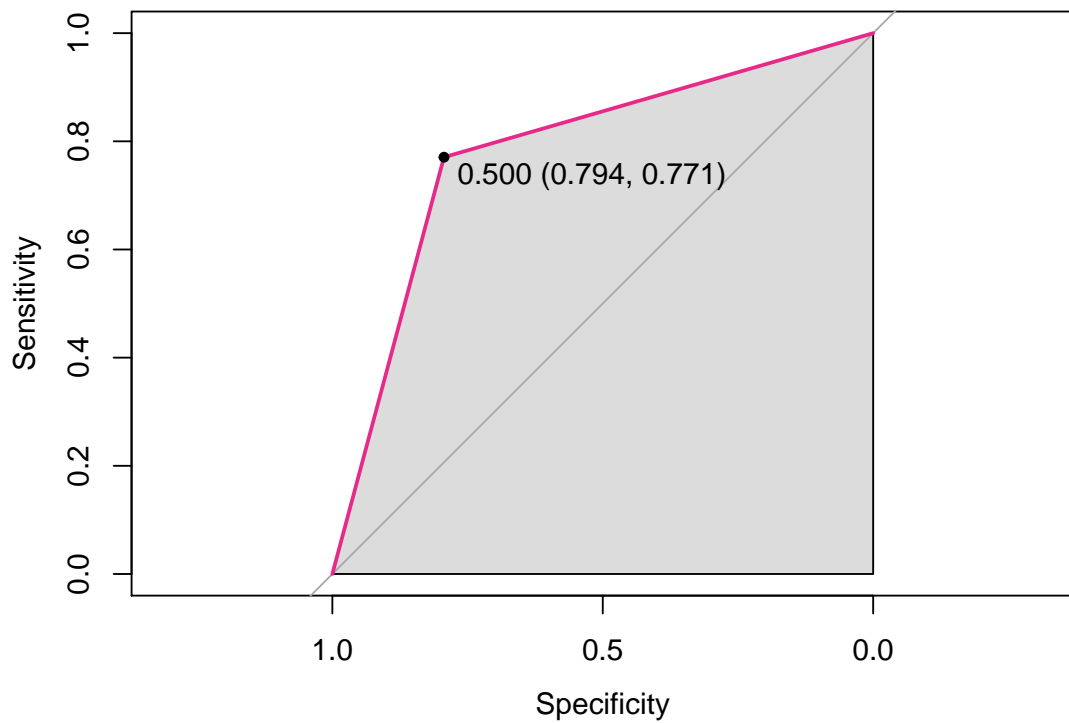

Figure 14: Logistic Regression Model AUC and ROC Curve

Confusion matrix and Figure 14 demonstrate the logistic model performance on the balanced data set. Using the proportion of positive data points that are correctly considered as positive (true positives) and the proportion of negative data points that are mistakenly considered as positive (false negative), we generated a graphic that shows the trade off between the rate at which the model correctly predicts the rain tomorrow with the rate of incorrectly predicting the rain. The value around 0.80 indicates that the model does a good job in discriminating between the two categories.
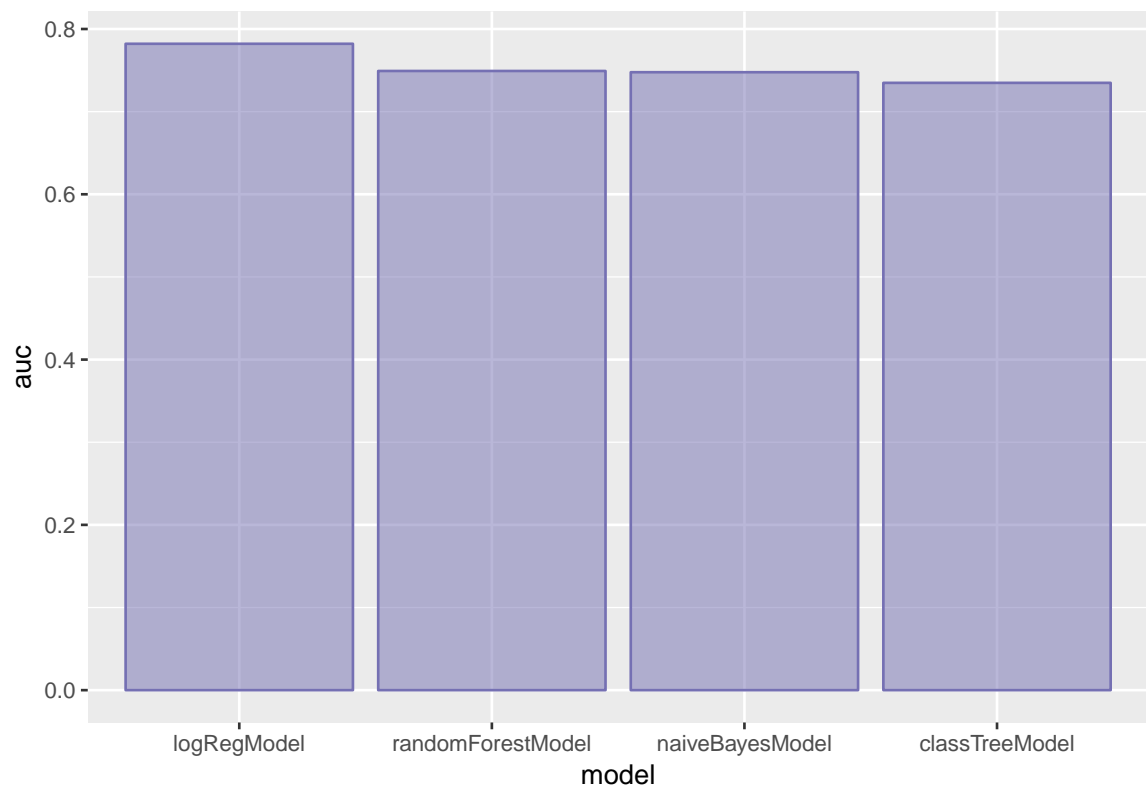
Model Comparison

Figure 15: Model Accuracy Comparison

Model Deployment

Conclusion

Bibliography

Note from the Authors

This file was generated using *The R Journal* style article template, additional information on how to prepare articles for submission is here - Instructions for Authors. The article itself is an executable R Markdown file that could be downloaded from Github with all the necessary artifacts (**?**).

Sumaira Afzal
York University School of Continuing Studies

Viraja Ketkar
York University School of Continuing Studies

Murlidhar Loka
York University School of Continuing Studies

Vadim Spirkov
York University School of Continuing Studies