Project3: Classfication

the BRICS

3/19/2019

1. Data loading and processing

1.1 Dataset introduction

This dataset contains daily weather observations from numerous Australian weather stations. There are 142k observations and 24 different variables in the dataset. The dependent variable is whether tomorrow will rain or not. In order to get the prediction, we have both categorical features and numerical features, including temperature, direction and the speed of the strongest wind, humidity, evaporation, pressure, cloud, sunshine and whether the previous day rains or not. The data is from kaggle.

1.2 Target variable

Our target variable is a dummy variable indicating whether tomorrow rains or not.

1.3 Data Description

	n	min	mean	median	sd	max
MinTemp	58117	-6.7	13.34	13.1	6.47	31.4
MaxTemp	58117	4.1	24.13	23.8	6.97	48.1
Rainfall	58117	0.0	2.12	0.0	7.00	206.2
Evaporation	58117	0.0	5.44	4.8	3.69	81.2
Sunshine	58117	0.0	7.70	8.6	3.77	14.5
WindGustDir*	58117	1.0	8.48	9.0	4.79	16.0
WindGustSpeed	58117	9.0	40.55	39.0	13.39	124.0
Humidity9am	58117	0.0	66.23	67.0	18.63	100.0
Humidity3pm	58117	0.0	49.70	51.0	20.22	100.0
Pressure9am	58117	980.5	1017.33	1017.3	6.94	1040.4
Pressure3pm	58117	977.1	1014.88	1014.8	6.90	1038.9
Cloud9am	58117	0.0	4.25	5.0	2.80	8.0
Cloud3pm	58117	0.0	4.33	5.0	2.65	9.0
Temp9am	58117	-0.9	18.08	17.7	6.61	39.4
Temp3pm	58117	3.7	22.63	22.3	6.83	46.1
RainToday*	58117	1.0	1.22	1.0	0.41	2.0
RainTomorrow*	58117	1.0	1.22	1.0	0.41	2.0
month*	58117	Inf	NaN	NA	NA	-Inf

Variable name	Description	Reason
MinTemp/MaxTemp/ Temp9am/ Temp3pm	The min/max temperature in degrees Celsius; Temperature at 9am/3pm	The temperature of the observation taken day would influence the possibility of raining of the day
Rainfall	The amount of rainfall recorded for the day in mm	The amount of rain recorded today would cast influences on the possibility of raining of tomorrow
Evaporation	The so-called Class A pan evaporation (mm) in the 24 hours to 9am	When evaporation is high, the possibility of raining may go down
Sunshine	The number of hours of bright sunshine in the day.	The longer the sunshine duration of the observation day, the lower the possibility of raining
WindGustDir/WindGustSpeed/ WindDir9am/WindDir3pm/ WindSpeed3pm/WindSpeed3pm	The direction/speed of the strongest wind gust in the 24 hours to midnight; The direction/speed of the wind at 9am/3pm	The direction/speed of the wind of the observation day, determines the direction of the clouds, which is a necessity for raining
Humidity9am/ Humidity3pm	Humidity (percent) at 9am/3pm	The higher the percent of humidity, the gaseous state of water would turn into liquid state and fall down as rain
Pressure9am/Pressure3pm	Atmospheric pressure (hpa) reduced to mean sea level at 9am/3pm	Before the rain, the humidity in the air increases and the water vapor content is high, which would lead to a low atmospheric pressure
Cloud9am/ Cloud3pm	Fraction of sky obscured by cloud at 9am.	The cloud is a necessity of raining, so when the fraction of sky abscured by cloud is high, the possibility of raining would increase with the increase in the fraction of cloud in the sky.
RainToday	1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0	Whether today rains or not would cast an influence on tomorrow's raining possibility

Figure 1: Table1

1.4 Data preprocessing

After having a general observation of the dataset, we performed the following steps to process the data before we start the modeling process.

a. Delete the variable *RISK_MM*
We deleted the variable *RISK_MM8* because this variable measures the data related with tomorrow and contains information of the dependent variable.

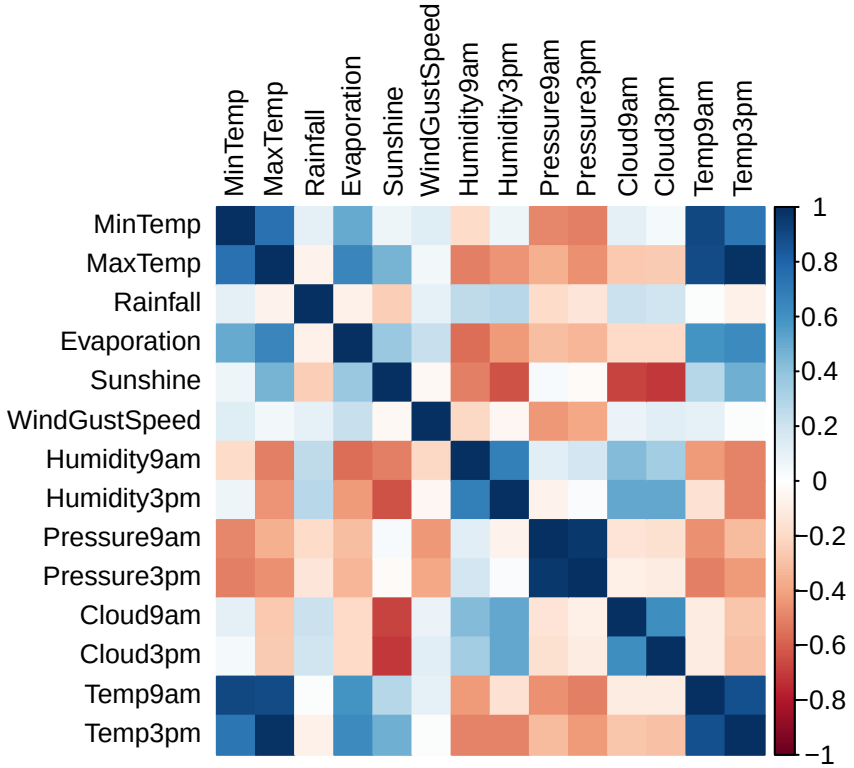
b. Delete the variable *Location*
We deleted the variable *Location* because under this variable, each observation has little variance.

c. Delete the specific wind condition variable
We deleted all the variables related with the specific wind conditions since these variables have little relation with possibility of raining. They are *WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm*

d. Treatment of missing values
We deleted all rows with NA values, there are still enough data (58117) remaining. Model might be biased if the data are not randomly missed.

e.Create new variable *month*
We created a now variable named *month* which is the date of the observation day in numerical version.

f. Correlation matrix plot
After looking at the Correlation matrix, we deleted highly correlated predictors (>0.8) including *Temp9am, Temp3pm, Pressure9am* to avoid the problem of colinearility.



2. Model Selection

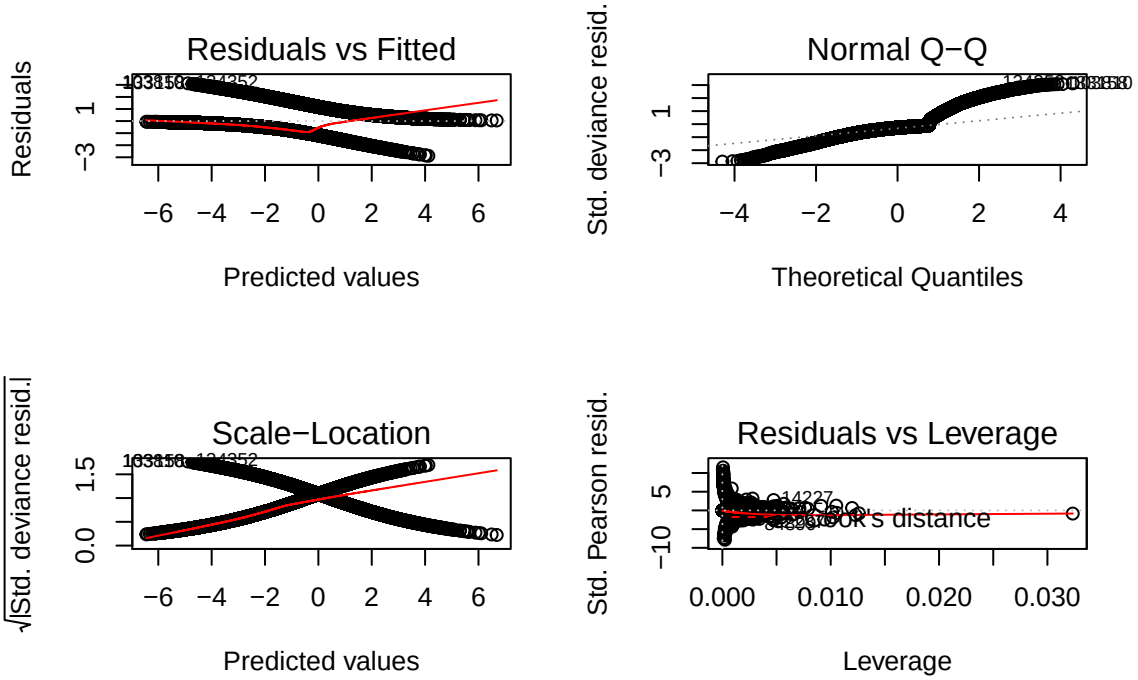
2.0 Method Introduction

For all of the three models, the code we define control is `control <- trainControl(method='cv',number=5)` .In the trainControl() function, we have set the argument `method` equals to `cv`, which stands for cross-validation. Also, `number=5` means 5-fold cross-validation. By doing so, we go through a resampling procedure to evaluate the model on unseen data. To be more specific, the original dataset is shuffled and randomly partitioned into 5 equal sized subsamples. Of the 5 subsamples, 4 subsamples are used as training data while the remaining 1 subsample is retained as the validation data for testing the model. The cross-validation process is then repeated 5 times, with each of the 5 subsamples used exactly once as the validation data. The 5 results can then be averaged to produce a single estimation. In this way, we can avoid getting an overfitting model. Also, we scale and center our data to make all features at similar magnitude.

2.1 Logistic Regression

2.1.1 First trail

As mentioned before, we use RainTomorrow as the target variable. After the basic data preprocess, we then conduct our first Logistic model (model 1). We run the model against all other variables and the results are shown below. Based on the regression result, we find that WindGustDir, a categorical variable, is not statistically significant for most of its levels, and there are some months that do not have statistically significant coefficients. As a result, we decide to get rid of the variables that are not statistically significant in an effort to improve the predictability of the model. Moreover, we check for the outlier. By looking at the residual plots of the model, we find 5 outliers with observation number of 134352, 133158, 103810, 82287, 84899 and delete them from our dataset.




```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.8767  -0.5148  -0.2763  -0.1248   3.1288
##
## Coefficients:
##                 Estimate Std. Error  z value             Pr(>|z|)
## (Intercept)    -1.970635   0.016768 -117.525 < 0.0000000000000002 ***
## MinTemp        -0.327268   0.031272  -10.465 < 0.0000000000000002 ***
## MaxTemp         0.589775   0.035511   16.608 < 0.0000000000000002 ***
## Rainfall        0.074654   0.014328    5.210      0.0000001883794 ***
## Evaporation    -0.029820   0.021816   -1.367             0.171670
## Sunshine       -0.630378   0.022699  -27.771 < 0.0000000000000002 ***
## WindGustDirENE -0.015548   0.019089   -0.814             0.415382
## WindGustDirESE  0.005048   0.018171    0.278             0.781171
## WindGustDirN    0.014852   0.018231    0.815             0.415259
## WindGustDirNE   0.028175   0.017627    1.598             0.109966
## WindGustDirNNE  0.045008   0.016252    2.769             0.005617 **
## WindGustDirNNW  0.046739   0.015804    2.957             0.003102 **
## WindGustDirNW   0.045777   0.016251    2.817             0.004850 **
## WindGustDirS   -0.012249   0.017985   -0.681             0.495811
## WindGustDirSE   0.007484   0.018336    0.408             0.683156
## WindGustDirSSE  0.017070   0.017735    0.962             0.335807
## WindGustDirSSW  0.003709   0.018284    0.203             0.839271
## WindGustDirSW   0.013459   0.018661    0.721             0.470782
## WindGustDirW    0.020989   0.018175    1.155             0.248144
## WindGustDirWNW  0.048177   0.016445    2.930             0.003395 **
## WindGustDirWSW  0.026077   0.018006    1.448             0.147545
## WindGustSpeed   0.527664   0.014656   36.005 < 0.0000000000000002 ***
## Humidity9am     0.049122   0.022146    2.218             0.026550 *
## Humidity3pm     1.165264   0.024595   47.378 < 0.0000000000000002 ***
## Pressure3pm    -0.471906   0.016993  -27.771 < 0.0000000000000002 ***
## Cloud9am       -0.072632   0.020909   -3.474             0.000513 ***
## Cloud3pm        0.305100   0.021468   14.212 < 0.0000000000000002 ***
## RainTodayYes    0.163375   0.014632   11.166 < 0.0000000000000002 ***
## monthAug        0.045989   0.018396    2.500             0.012420 *
## monthDec       -0.061975   0.017979   -3.447             0.000567 ***
## monthFeb       -0.089378   0.018119   -4.933      0.0000008107735 ***
## monthJan       -0.129778   0.019569   -6.632      0.0000000000331 ***
## monthJul        0.012947   0.018280    0.708             0.478776
## monthJun       -0.023517   0.018117   -1.298             0.194247
## monthMar       -0.052296   0.018255   -2.865             0.004173 **
## monthMay        0.009469   0.017874    0.530             0.596281
## monthNov       -0.019254   0.017987   -1.070             0.284430
## monthOct        0.033489   0.018216    1.838             0.065998 .
## monthSep        0.016002   0.017987    0.890             0.373676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61116  on 58116  degrees of freedom
## Residual deviance: 39035  on 58078  degrees of freedom
## AIC: 39113
##
## Number of Fisher Scoring iterations: 6
```

2.1.2 Second trail

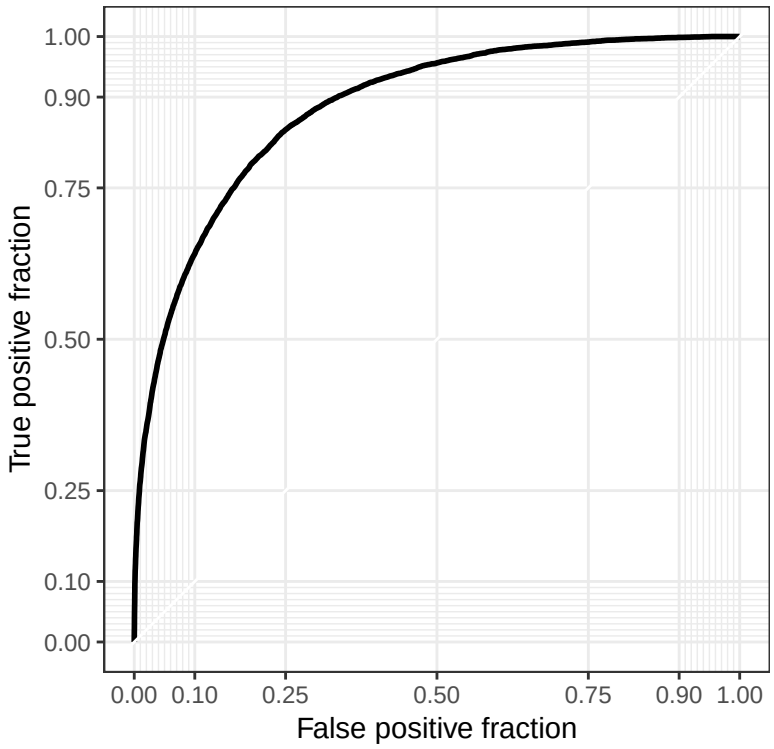
We then run a second Logistic regression with the new dataset and all statistically significant variables. The result for model 2 is shown below.

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8851  -0.5149  -0.2784  -0.1262   3.1181
##
## Coefficients:
##               Estimate Std. Error  z value             Pr(>|z|)
## (Intercept)   -1.96478    0.01668 -117.801 < 0.0000000000000002 ***
## MinTemp       -0.34640    0.03077  -11.256 < 0.0000000000000002 ***
## MaxTemp        0.59388    0.03476   17.087 < 0.0000000000000002 ***
## Rainfall       0.07231    0.01429    5.060      0.0000004197312 ***
## Evaporation   -0.04088    0.02108   -1.940             0.052429 .
## Sunshine      -0.62528    0.02217  -28.205 < 0.0000000000000002 ***
## WindGustSpeed  0.53266    0.01443   36.921 < 0.0000000000000002 ***
## Humidity9am    0.04636    0.02199    2.108             0.035021 *
## Humidity3pm    1.16126    0.02439   47.608 < 0.0000000000000002 ***
## Pressure3pm   -0.49686    0.01603  -30.989 < 0.0000000000000002 ***
## Cloud9am      -0.07284    0.02053   -3.547             0.000389 ***
## Cloud3pm       0.30877    0.02127   14.516 < 0.0000000000000002 ***
## RainTodayYes   0.16180    0.01450   11.161 < 0.0000000000000002 ***
## monthAug       0.04426    0.01334    3.317             0.000910 ***
## monthDec      -0.06861    0.01410   -4.867      0.0000011314759 ***
## monthFeb      -0.09731    0.01453   -6.695      0.0000000000216 ***
## monthJan      -0.13870    0.01532   -9.053 < 0.0000000000000002 ***
## monthMar      -0.05941    0.01393   -4.265      0.0000200161175 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61116  on 58116  degrees of freedom
## Residual deviance: 39094  on 58099  degrees of freedom
## AIC: 39130
##
## Number of Fisher Scoring iterations: 6
```

2.1.3 Third trail

The third Logistic regression we run includes a quadratic term *WindGustSpeed^2*. For this model, we still scale and center our data and conduct a 5-fold cross validation. According to the regression result, this model (model 3) has all statistically significant variables. Also, model 3 has a relatively high sensitivity, relatively low specificity, and a highest ROC of 0.88257 among all models (ROC curve shown below). Therefore, we would conclude that model 3 is our final best Logistic model. The regression results together with the odd ratio (exp(B)) are shown below.

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7944  -0.5132  -0.2771  -0.1247   3.1006
##
## Coefficients:
##                       Estimate Std. Error  z value             Pr(>|z|)
## (Intercept)          -1.96949    0.01672 -117.824 < 0.0000000000000002 ***
## MinTemp              -0.36322    0.03088  -11.764 < 0.0000000000000002 ***
## MaxTemp               0.60529    0.03476   17.413 < 0.0000000000000002 ***
## Rainfall              0.07172    0.01430    5.015     0.00000052935377 ***
## Evaporation          -0.04355    0.02108   -2.066              0.03882 *
## Sunshine             -0.63789    0.02226  -28.661 < 0.0000000000000002 ***
## WindGustSpeed         0.89957    0.05718   15.733 < 0.0000000000000002 ***
## Humidity9am           0.05266    0.02199    2.395              0.01664 *
## Humidity3pm           1.15968    0.02439   47.539 < 0.0000000000000002 ***
## Pressure3pm          -0.50462    0.01609  -31.370 < 0.0000000000000002 ***
## Cloud9am             -0.07582    0.02052   -3.694              0.00022 ***
## Cloud3pm              0.30438    0.02126   14.315 < 0.0000000000000002 ***
## RainTodayYes          0.15958    0.01451   10.997 < 0.0000000000000002 ***
## monthAug              0.04383    0.01333    3.287              0.00101 **
## monthDec             -0.07069    0.01409   -5.017     0.00000052570048 ***
## monthFeb             -0.10028    0.01454   -6.898     0.00000000000528 ***
## monthJan             -0.14173    0.01532   -9.249 < 0.0000000000000002 ***
## monthMar             -0.06198    0.01394   -4.448     0.00000868627288 ***
## `I(WindGustSpeed^2)` -0.36100    0.05419   -6.661     0.00000000002714 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61116  on 58116  degrees of freedom
## Residual deviance: 39049  on 58098  degrees of freedom
## AIC: 39087
##
## Number of Fisher Scoring iterations: 6
```



The following is the coefficients interpretation of model 3 based on the odd ratio of each variable.

Variable	Odd Ratio	Interpretation
MinTemp	0.6954	With the minimum temperature increases by 1 degree Celsius, the odds of rain tomorrow will decrease by 30.46%.
MaxTemp	1.8317	With the maximum temperature increases by 1 degree Celsius, the odds of rain tomorrow will increase by 83.18%.
Rainfall	1.0744	With the increase of 1 mm in rainfall, the odd ratio of rain tomorrow would increase 7.44%.
Evaporation	0.9574	With the increase of 1 mm in class A pan evaporation in the 24 hours to 9am, the odds of rain tomorrow will decrease 4.26%.
Sunshine	0.5284	With 1 more hour of bright sunshine in the day, the odds of rain tomorrow will decrease by 47.16%.
WindGustSpeed	2.4585	With the increase of 1 km/h (speed) of the strongest wind gust in the 24 hours to midnight, the odd ratio of rain tomorrow would increase 145.86%.
Humidity9am	1.0540	With 1 percent increase of humidity at 9am, the odds of rain tomorrow will increase by 5.41%.
Humidity3pm	3.1889	With 1 percent increase of humidity at 3pm, the odds of rain tomorrow will increase by 218.9%.
Pressure3pm	0.6037	With the increase of the pressure (hpa) at 3pm, the odd ratio of rain tomorrow would increase 39.63%.
Cloud9am	0.9269	With 1 more “oktas” of sky obscured by the cloud at 9am, the odds of rain tomorrow will decrease by 7.3%.
Cloud3pm	1.3557	With 1 more “oktas” of sky obscured by the cloud at 3pm, the odds of rain tomorrow will increase by 35.58%.
RainTodayYes	1.1730	If today rains, the odds of rain tomorrow will increase by 17.30%.
Aug	1.0448	The odd ratio of rain tomorrow would increase 4.48% if in August compared to that of in April.
Dec	0.9317	The odd ratio of rain tomorrow would decrease 6.82% if in December compared to that of in April.
Feb	0.9045	The odd ratio of rain tomorrow would decrease 9.54% if in February compared to that of in April.
Jan	0.8678	The odd ratio of rain tomorrow would decrease 13.21% if in January compared to that of in April.
Mar	0.9401	The odd ratio of rain tomorrow would decrease 6% if in March compared to that of in April.
WindGustSpeed^2	0.6969	It is hard to interpret the exact relationship between WindGustSpeed^2 and the odds of rain tomorrow, but the smaller than 1 odd ratio indicates that with the increase of wind speed^2, the odds of rain tomorrow will decrease.

Figure 2: Table2

2.2 KNN

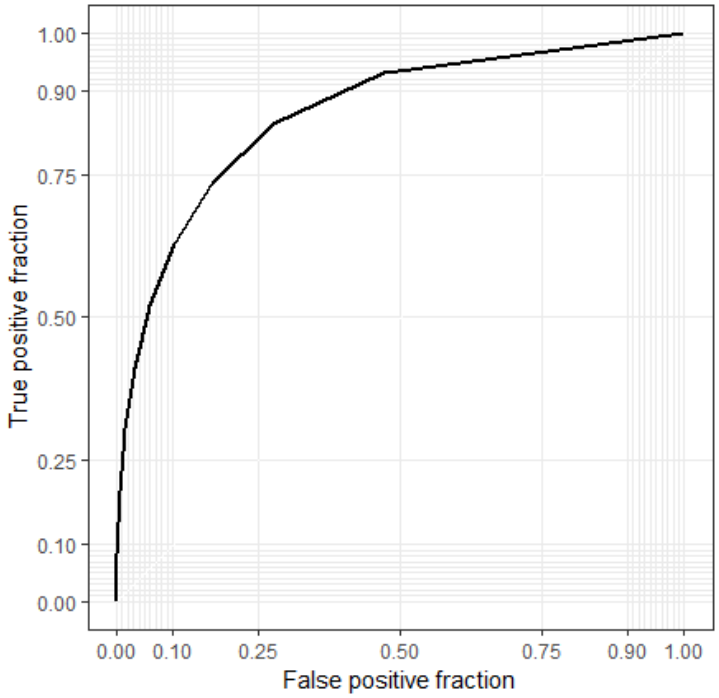
The k-nearest neighbors (k-NN) classification is instance-based learning. It first saves training set in memory and then compares unseen data to the training set. The unseen object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

In our k-nn model, when k is 5, the sensitivity is 0.9301 and the specificity is 0.5278. When k is 7, the sensitivity is 0.9359 while the specificity is 0.5227. When k is 9, the sensitivity is 0.9399 and the specificity is 0.5166. Although specificity is not that satisfactory, high sensitivity means the prediction of the k-nn model is much accurate. Therefore, we can conclude that the k-nn model is not overfitting.

Besides, the ROC Curve is another powerful performance measure for binary classification. To interpret the curve, we need to check whether the curve is close to the left upper corner, where the true positive rate is close to 1.0 while the false positive rate is close to 0. Another equivalent way is to see the Area Under the Curve or AUC. Typically, good classifiers tend to have a big area under the ROC curve.

From the summary, we can see that when k=9, the value of ROC becomes the largest, meaning the model is optimal at this time. Thus, the final k value used in the model is 9. The corresponding ROC curve is drawn.

```
## k-Nearest Neighbors
##
## 58117 samples
##    13 predictor
##     2 classes: 'No', 'Yes'
##
## Pre-processing: centered (18), scaled (18)
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 46494, 46494, 46494, 46493, 46493
## Resampling results across tuning parameters:
##
##   k  ROC        Sens       Spec
##   5  0.8382914  0.9301044  0.5277581
##   7  0.8521062  0.9358996  0.5226541
##   9  0.8593573  0.9399541  0.5166078
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was k = 9.
```



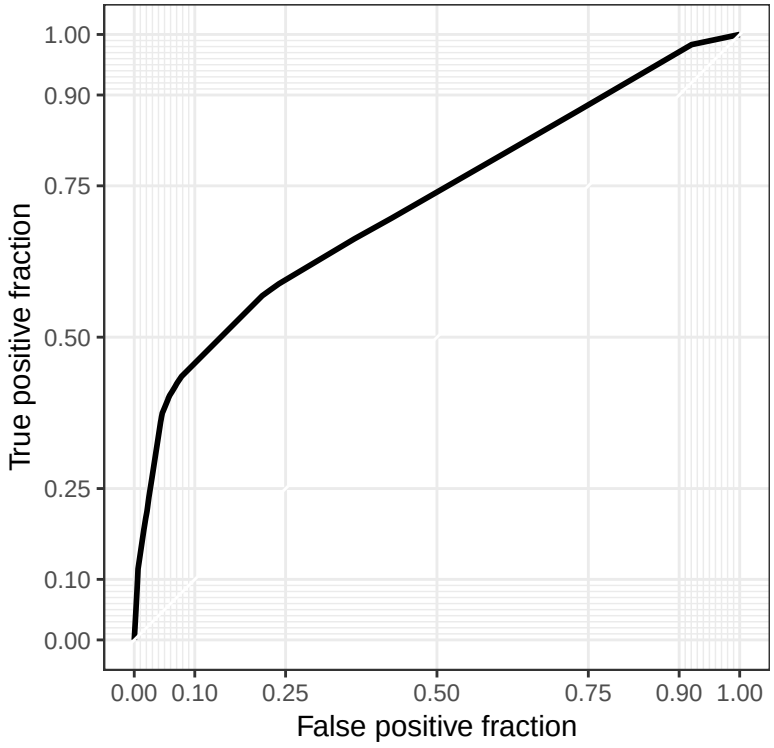
2.3 Classification Tree

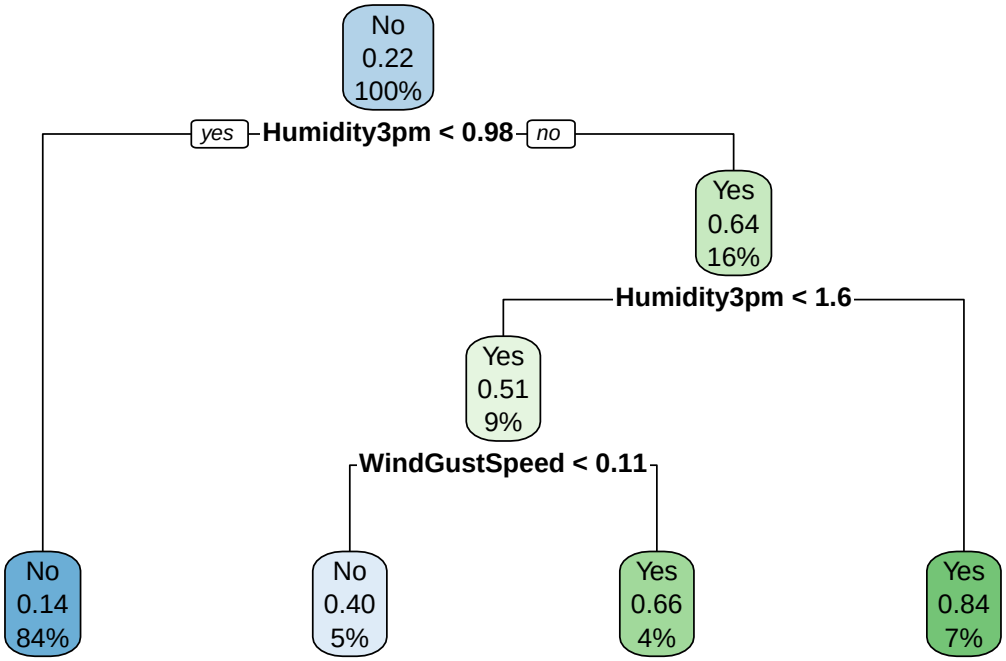
Classification tree automatically assigns a class to new observations with certain features. The classification process is finished by asking some questions step by step.

In order to find the best model for the classification tree, we need to compare the complexity parameter (cp) with the ROC. An optimal cp value can be estimated by testing different cp values and using cross-validation approaches to determine the corresponding prediction accuracy of the model. The best cp is defined as the one that maximizes the cross-validation accuracy.

In our model, when cp is 0.007067138, the ROC value is 0.7521, the sensitivity is 0.9578 and the specificity is 0.42449941. When cp is 0.025363172, the ROC value is 0.7017, the sensitivity is 0.9433 and the specificity is 0.42850412. When cp is 0.210051040, the ROC value is 0.5373, the sensitivity is 0.9850 and the specificity is 0.0896. Although the third one has the highest sensitivity and the lowest specificity, its ROC value is relatively low, meaning the accuracy is not guaranteed. Based on these, we choose the model with a cp of 0.007067138 as the final model because it produces the largest ROC while not overfitting the model.

```
## CART
##
## 58117 samples
##    13 predictor
##     2 classes: 'No', 'Yes'
##
## Pre-processing: centered (18), scaled (18)
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 46493, 46494, 46494, 46494, 46493
## Resampling results across tuning parameters:
##
##   cp           ROC        Sens       Spec
##   0.007067138  0.7520676  0.9578027  0.42449941
##   0.025363172  0.7017185  0.9433254  0.42850412
##   0.210051040  0.5373338  0.9849934  0.08967413
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.007067138.
```





If we take a closer look at the tree above, the classification process is done by asking questions about humidity and WindGustSpeed. Particularly, if humidity at 3 pm today is lower than 0.98, then it will not rain tomorrow. If humidity at 3 pm today is lower than 1.6 and the WindGustSpeed is lower than 0.11, then it will not rain tomorrow. If humidity at 3 pm today is lower than 1.6 but the WindGustSpeed is higher or equal to 0.11, then tomorrow will be a rainy day. If the humidity at 3 pm today is higher or equal to 1.6, then it will rain tomorrow.

3.Model Comparison

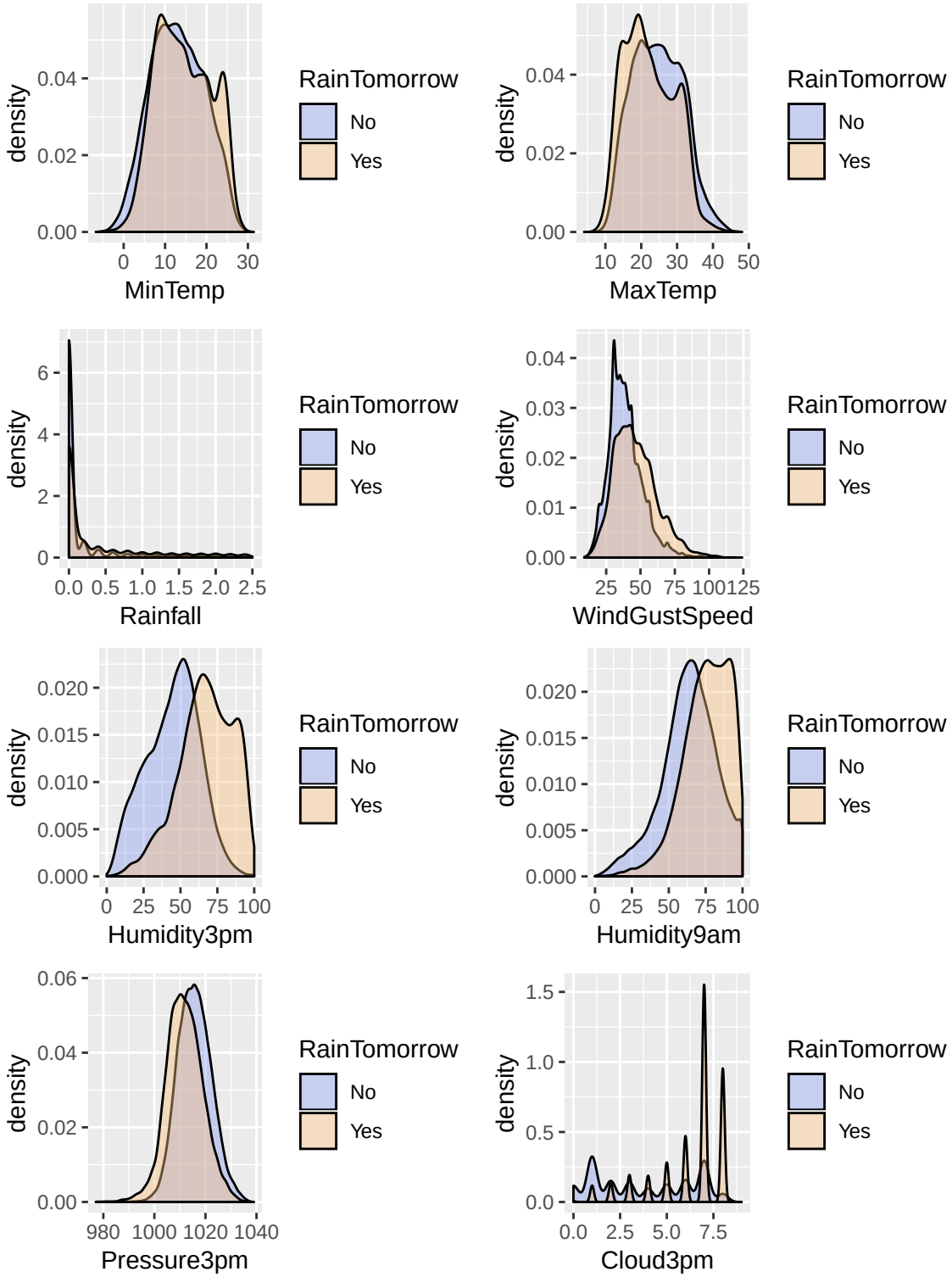
Before we compare and contrast the three different models we conducted, we have to determine the best model of each method.
For Logistic Model, like what we mentioned above, we choose the last model (model 3) as the best logstic model since it has the highest AUC value of 0.882.
For the KNN model, we choose the model with the largest k of 9 as the best KNN model.KNN model is non-parametric, it means that it does not make any assumptions on the underlying data distribution. Therefore, KNN could and probably should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution data.
For the classification tree model, both AUC and accuracy increase when cp decreases. So, we choose the decision tree model with the lowest cp value of 0.007 as our final decision tree model, which means that we have chosen to use the simplest tree among the three.
After the determination of the best model for each method, we then compare them altogether and try to find the FINAL BEST model for our project. We chose AUC and accuracy as our measure for choosing from 3 different models.

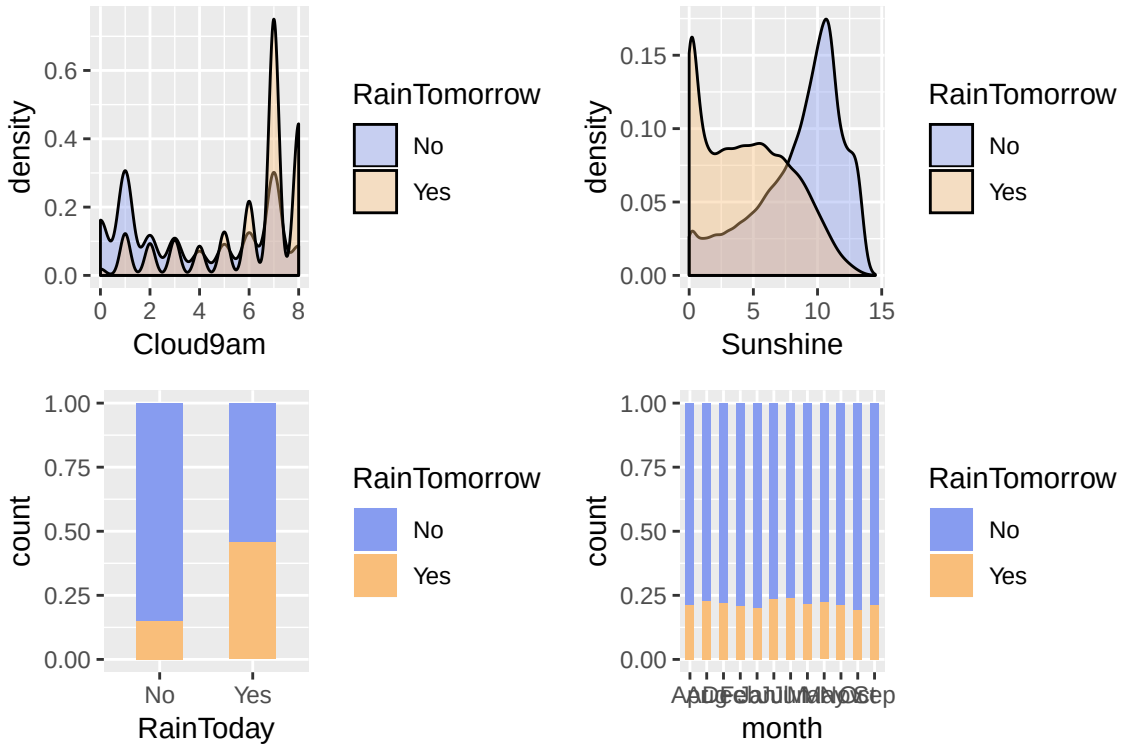
	Accuracy	AUC	Sens	Spec
logit	0.853	0.882	0.943	0.530
knn	0.848	0.858	0.941	0.520
rpart	0.843	0.777	0.959	0.430

As we can see in this form, the AUC, accuracy and specificity of the logistic model are all the highest among 3 models. But decision tree model has the highest sensitivity, which contradicts with other measurements.

In this specific case, we treat sensitivity more important than specificity because false prediction of not rain will bring larger trouble. If the model tell us it won't rain tomorrow,we will not consider bringing umbrellas, which will cause a lot of inconvenience if tomorrow actully rains. However, the sensitivity of the 3 models are quite close, leaving only 1.6% difference in logistic model and the decision tree model. Considering the fact that the specificity of the decision tree is 10% lower than that of the logistic model, we still choose the logistic model as our best model for its high AUC and accuracy.

4.Visulization




density
0.6
0.4
0.2
0.0
0
2
4
6
8
Cloud9am
RainTomorrow
No
Yes
density
0.15
0.10
0.05
0.00
0
5
10
15
Sunshine
RainTomorrow
No
Yes
count
1.00
0.75
0.50
0.25
0.00
No
Yes
RainToday
RainTomorrow
No
Yes
count
1.00
0.75
0.50
0.25
0.00
month
RainTomorrow
No
Yes