FINAL ASSIGNMENT

Ilaria Crippa

2024-03-03

Contents

0.1 The dataset

The dataset has been generated by the Korean Meteorological Administration over the city of Seoul (SK), in the time period between 2013 and 2017, and uploaded on the UC Irvine Machine Learning Repository. The data consists in the forecasts of the LDAPS (Local Data Assimilation and Prediction System) model, the present-day minimum and maximum temperatures and other auxiliary variables, while the two output variables are the next-day maximum and minimum air temperatures.

The data has been previously studied to select the optimal methodology for the bias correction of the future minimum and maximum air temperature forecasts of the Korean Meteorological Administration LDAPS model, in order to precisely anticipate and prevent the possible damages caused by extreme meteorological events, such as heat waves and cold spells. Bias errors usually arise with extreme, but unusually occurring, rainfall intensities or temperatures, that are underestimated by numerical weather prediction models, due to a variety of factors (e.g. a large grid size - which leads to a less detailed representation of the weather, due to a reduced spacial resolution - and the simplification of thermodynamic processes).

0.2 Applied goals in studying the data

The objectives of this research consist in the prediction of the probability of summer heat waves occurring and the classification of hot days into two categories: those posing a risk in this regard and those with lower hazard levels. This research is crucial, since heat waves constitute significant threats to human health, infrastructure, ecosystems, agriculture and the economy, which underscores the importance of adaptation measures to reduce their impacts.

0.3 Representation of the data

The study area is the metropolitan city of Seoul (SK), which is located in a geographical region surrounded by four mountains and split in two by the Han River. Precipitation is particularly abundant in summer, due to the East Asian monsoon, and the period is characterized by temperatures above 30°C.

The dataset is composed of data collected from the 30th of June 2013 to the 30th of August 2017 and it consists of the following variables: the ones produced by the LDAPS model for the forecasts are the next-day maximum and minimum air temperatures ("Tmax_lapse", "Tmin_lapse", °C), next-day maximum and minimum relative humidity ("RHmax" and "RHmin", %), next-day average wind speed ("WS", m/s), next-day average latent heat flux ("LH", W/m^2), next-day cloud covers between 0-5h, 6-11h, 12-17h and 18-23h ("CC1", "CC2", "CC3", "CC4", %) and next-day precipitation in the same time splits ("PPT1", "PPT2", "PPT3", "PPT4", %). Additionally, the data collected in-situ comprises the observed next-day maximum and minimum air temperatures ("Next_Tmax" and "Next_Tmin", °C) - used as target variables -, as well as the present max and min air temperatures ("Present_Tmin", "Present_Tmax"), obtained from a grid with 25 automatic weather stations scattered around the city. Other auxiliary variables are the latitude and longitude ("lat", "lon", in coordinates), elevation ("DEM", m) and slope ("Slope", °) and daily incoming solar radiation ("Solar.radiation", Wh/m^2).

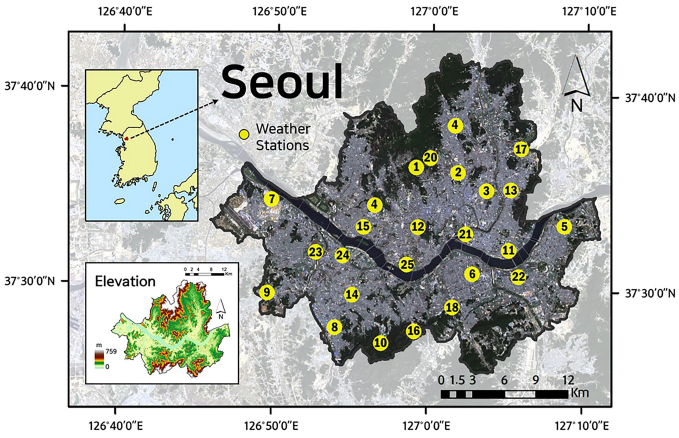


Figure 1: Location of automatic weather stations (source: Cho, D., Yoo, C., Im, J., & Cha, D.-H., 2020)

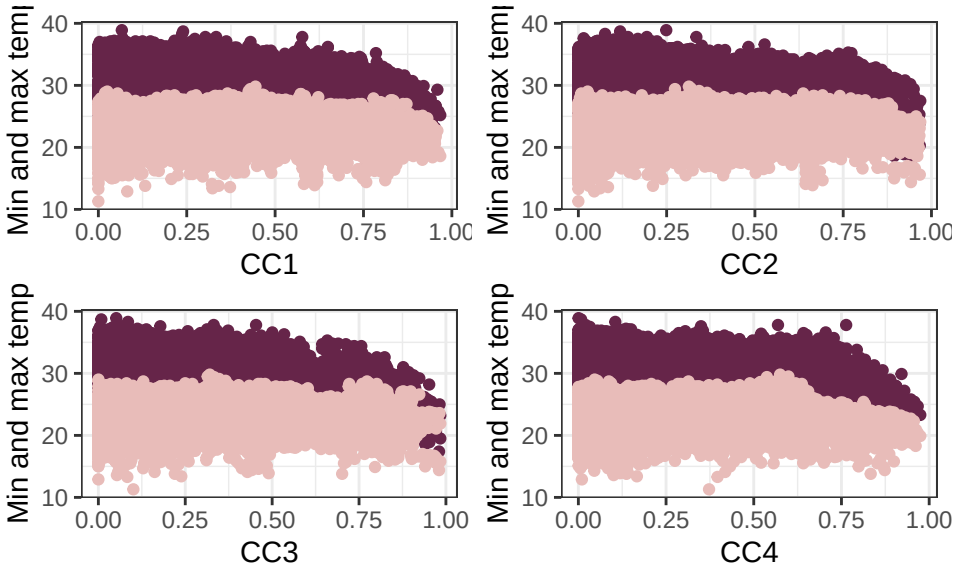
We now read the data, remove the 164 rows with missing values and provide a summary of the data set. The number of observations is 7588 and there are 25 variables in total.

```
data = read.csv("Bias_correction.csv", header = T)
data = na.omit(data)
data$Date = as.Date(data$Date, format = "%d/%m/%Y")
station = as.factor(data$station)
summary_output = summary(data)
print(summary_output)
```

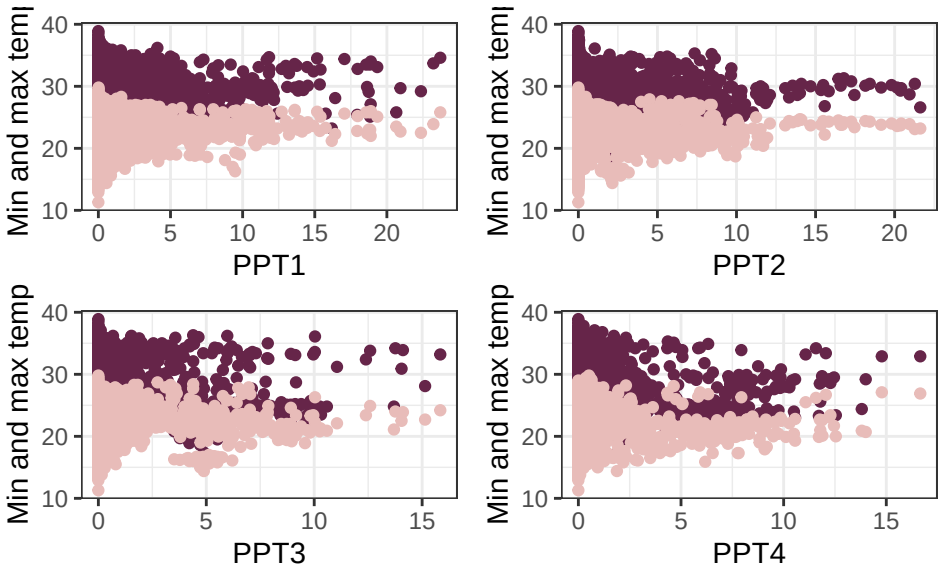
```
##     station           Date              Present_Tmax    Present_Tmin
##  Min.   : 1.00   Min.   :2013-06-30   Min.   :20.00   Min.   :11.3
##  1st Qu.: 7.00   1st Qu.:2014-07-15   1st Qu.:27.80   1st Qu.:21.6
##  Median :13.00   Median :2015-07-29   Median :29.90   Median :23.4
##  Mean   :13.01   Mean   :2015-07-27   Mean   :29.75   Mean   :23.2
##  3rd Qu.:19.00   3rd Qu.:2016-08-14   3rd Qu.:32.00   3rd Qu.:24.8
##  Max.   :25.00   Max.   :2017-08-30   Max.   :37.60   Max.   :29.9
##      RHmin           RHmax          Tmax_lapse      Tmin_lapse
##  Min.   :19.79   Min.   : 58.94   Min.   :17.62   Min.   :14.27
##  1st Qu.:45.96   1st Qu.: 84.20   1st Qu.:27.67   1st Qu.:22.09
```

```
##  Median :55.02   Median : 89.78   Median :29.71   Median :23.76
##  Mean   :56.72   Mean   : 88.36   Mean   :29.62   Mean   :23.51
##  3rd Qu.:67.12   3rd Qu.: 93.74   3rd Qu.:31.71   3rd Qu.:25.16
##  Max.   :98.52   Max.   :100.00   Max.   :38.54   Max.   :29.62
##        WS               LH              CC1              CC2
##  Min.   : 2.883   Min.   :-13.60   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 5.675   1st Qu.: 37.21   1st Qu.:0.1465   1st Qu.:0.1403
##  Median : 6.548   Median : 56.90   Median :0.3157   Median :0.3117
##  Mean   : 7.094   Mean   : 62.49   Mean   :0.3685   Mean   :0.3555
##  3rd Qu.: 8.029   3rd Qu.: 84.24   3rd Qu.:0.5742   3rd Qu.:0.5572
##  Max.   :21.858   Max.   :213.41   Max.   :0.9673   Max.   :0.9684
##       CC3              CC4               PPT1               PPT2
##  Min.   :0.0000   Min.   :0.00000   Min.   : 0.00000   Min.   : 0.00000
##  1st Qu.:0.1009   1st Qu.:0.08149   1st Qu.: 0.00000   1st Qu.: 0.00000
##  Median :0.2618   Median :0.22746   Median : 0.00000   Median : 0.00000
##  Mean   :0.3175   Mean   :0.29827   Mean   : 0.58901   Mean   : 0.48074
##  3rd Qu.:0.4964   3rd Qu.:0.49813   3rd Qu.: 0.05259   3rd Qu.: 0.01774
##  Max.   :0.9838   Max.   :0.97471   Max.   :23.70154   Max.   :21.62166
##       PPT3                PPT4                lat             lon
##  Min.   : 0.000000   Min.   : 0.000000   Min.   :37.46   Min.   :126.8
##  1st Qu.: 0.000000   1st Qu.: 0.000000   1st Qu.:37.51   1st Qu.:126.9
##  Median : 0.000000   Median : 0.000000   Median :37.55   Median :127.0
##  Mean   : 0.275007   Mean   : 0.265373   Mean   :37.54   Mean   :127.0
##  3rd Qu.: 0.007855   3rd Qu.: 0.000017   3rd Qu.:37.58   3rd Qu.:127.0
##  Max.   :15.841235   Max.   :16.655469   Max.   :37.65   Max.   :127.1
##       DEM              Slope         Solar.radiation   Next_Tmax
##  Min.   : 12.37   Min.   :0.0985   Min.   :4330    Min.   :17.40
##  1st Qu.: 28.70   1st Qu.:0.2713   1st Qu.:5001    1st Qu.:28.20
##  Median : 45.72   Median :0.6180   Median :5442    Median :30.40
##  Mean   : 61.92   Mean   :1.2598   Mean   :5344    Mean   :30.24
##  3rd Qu.: 59.83   3rd Qu.:1.7678   3rd Qu.:5729    3rd Qu.:32.60
##  Max.   :212.34   Max.   :5.1782   Max.   :5993    Max.   :38.90
##    Next_Tmin
##  Min.   :11.30
##  1st Qu.:21.30
##  Median :23.10
##  Mean   :22.91
##  3rd Qu.:24.60
##  Max.   :29.80
```

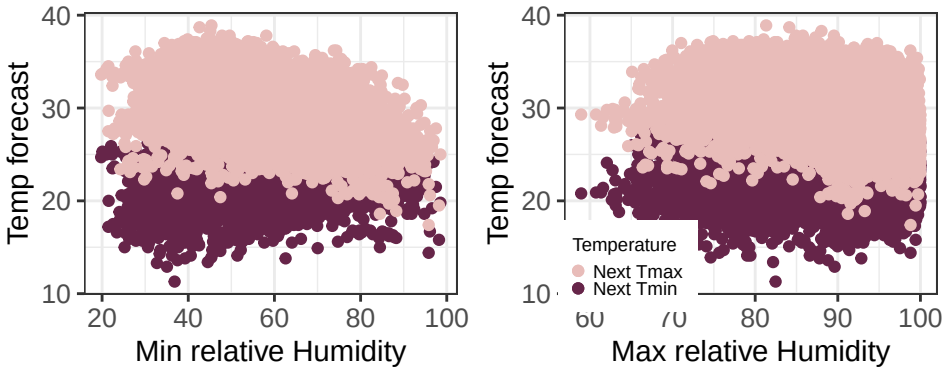
Some of the most relevant predictors are plotted against the next-day maximum and minimum air temperatures. There seems to be a slight negative correlation with the cloud coverage variables (CC1, CC2, CC3, CC3).



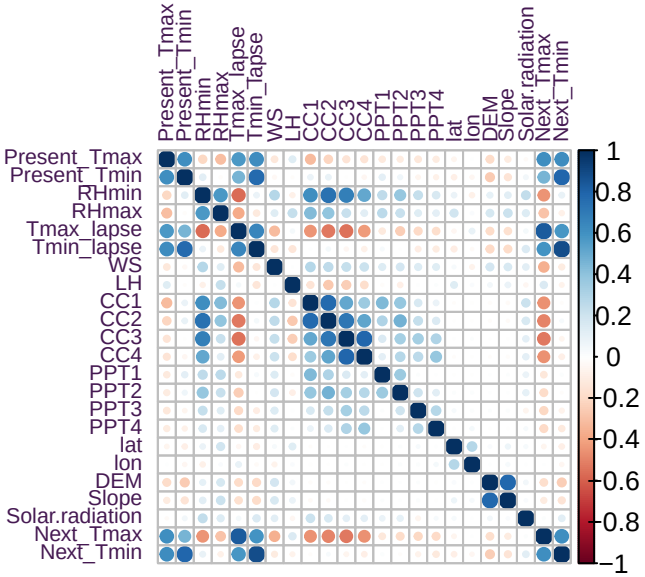
The correlation among next day temperatures and precipitation in each time split (PPT1, PPT2, PPT3, PPT4) is very weak, but tends to be negative with the next day maximum temperature.



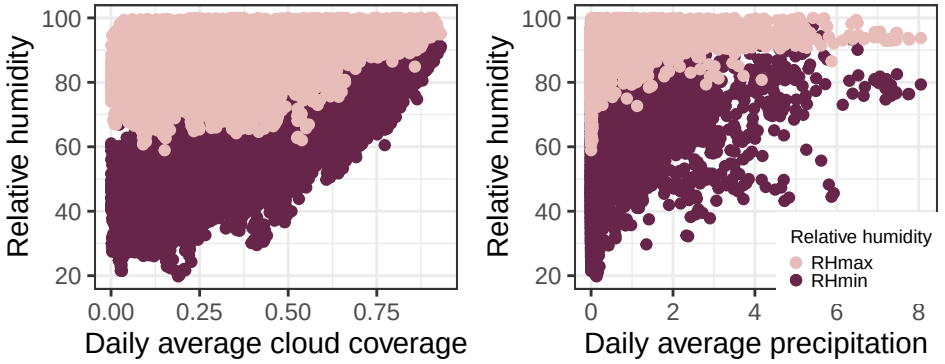
The relationship between the relative humidity forecasts and next day temperatures are also strongly negative, apart from that between minimum humidity and minimum temperature.



It is also crucial to inspect a correlation matrix including all the variables. We can notice that collinearity is present between precipitation, cloud coverage and relative humidity.



A scatter plot of the aforementioned variables can help in visualizing more clearly the positive correlation between them.



0.4 Logit model

In order to achieve the previously described goals, the focus is shifted to the continuous response "Next_Tmax", which is split into two groups to create a binary response, called "risk_temp". The latter is a binary variable

that takes the value 1, when the next day maximum temperature is higher or equal to 29.4°C, and the value 0, when it is below that same threshold. 29.4°C is selected as cutoff value, as it is the 98th percentile warm season daily mean temperature for the city of Seoul and it characterizes the intensity of heat waves, as defined in the paper "The Impact of Heat Waves on Mortality in Seven Major Cities in Korea" (Son et al., 2012). Although we are dealing with daily maximum temperatures and not daily average temperatures, the variable "risk_temp" could still be a valuable insight in assessing whether the risk of having a heat wave is substantial.

```
data$risk_temp = as.numeric(data$Next_Tmax >= 29.4)
table(data$risk_temp)
```

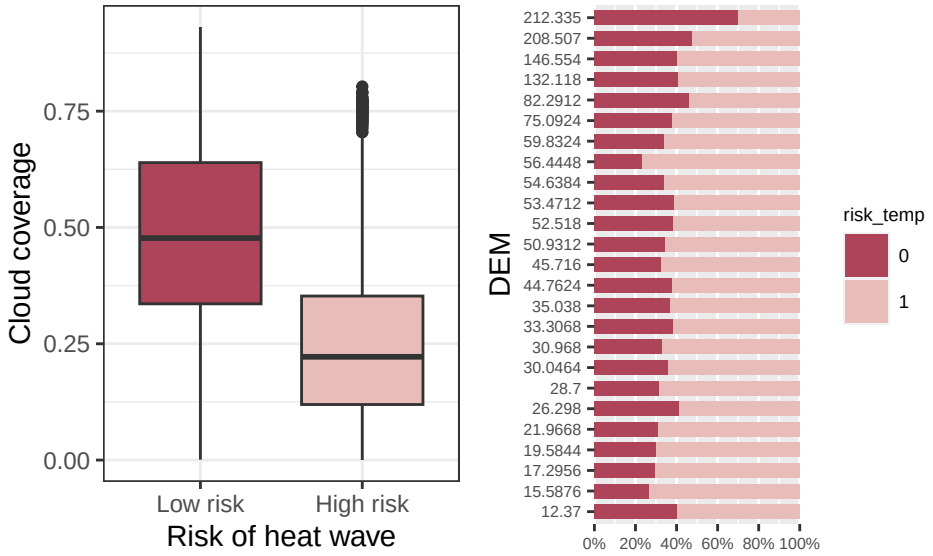
```
##
##    0    1
## 2822 4766
```

The binary variable identifies 2822 observations as "not risky", while the remainder 4766 (i.e. 62.8% of the total observations) as temperatures that could potentially induce a heat wave.

In the following plots we can observe the relationship between the binary variable and the level of cloud coverage: three quarters of the risky observed temperatures have been recorded when the estimated next day average cloud coverage is below 35%, while three quarters of the non-risky temperatures have been recorded with higher values of cloud coverage, suggesting that this independent variable might be relevant in future analysis. The second plot shows the relationship between risky temperatures and elevation: it seems that lower elevation could be associated to riskier temperatures.

```
x_labs = c("Low risk", "High risk")
plot3 = ggplot(data, aes(x = factor(risk_temp), y = CC)) +
geom_boxplot( fill=c(pal[4], pal[6])) +
labs(x = "Risk of heat wave", y = "Cloud coverage") +
scale_x_discrete(labels = x_labs) +
theme_bw()

plot4 = plot_xtab(x   = data$DEM, grp = data$risk_temp, margin  = "row", bar.pos = "stack",
show.summary = F, coord.flip   = T, show.values = F, geom.spacing = 0.9,
geom.colors = c(pal[4], pal[6]))
```

```
grid.arrange(plot3, plot4, nrow = 1, ncol=2)
```



We consider all the covariates present in the dataset to explain the response and assume a binomial regression

model with logit link. To explore all the possible models, we perform both the forward and backward step-wise selections and employ the AIC and BIC criteria to identify the best models. In order to decrease the number of categories of the variable "station", its levels are grouped to create a binary variable named "grouped_riskystation", such that those whose percentage of observed risky temperatures is higher than 65% (i.e. 11 stations) get assigned the value 1, while the remaining 14 stations that have detected a lower number of risky temperatures are assigned the value 0. The threshold is close to the percentage of risky temperatures observed in the entire dataset and allows the groups to be almost evenly distributed.

```
risk_stations = c(4, 6, 9, 11, 13, 14, 18, 22, 23, 24, 25)
data$grouped_riskystation = ifelse(data$station %in% risk_stations, 1, 0)
```

The "stepAIC()" function with direction forward analyses all the possible models - starting from the null model - generated by adding one predictor at each iteration and finds the best group of predictors based on the selected criterion. The process stops when adding any new variable increases the AIC/BIC of the previous model instead of improving it. The backward step-wise method starts from the full model instead and consists in the removal of a predictor at each iteration.

```
null_model = glm (risk_temp  ~ 1, family = binomial(link = logit), data = data)

full_model = glm (
  risk_temp  ~ data$grouped_riskystation + RHmin + RHmax + Tmax_lapse + Tmin_lapse + WS
+ LH + CC1 + CC2 + CC3 + CC4 + PPT1 + PPT2 + PPT3 + PPT4 + lat + lon + DEM + Slope +
Solar.radiation, family = binomial(link = logit), data = data)

forward_aic = stepAIC (null_model, scope=list(lower=null_model, upper=full_model),
direction='forward', k = 2, trace = F, data = data)
aic_values = forward_aic$anova$AIC

forward_bic = stepAIC (null_model, scope=list(lower=null_model, upper=full_model),
direction='forward', k = log(nrow(data)), trace = F, data = data)
bic_values = forward_bic$anova$AIC
```

```
backward_bic = stepAIC (full_model, scope=list(lower=null_model, upper=full_model),
direction='backward', k = log(nrow(data)), trace = F, data = data)
bic_back = backward_bic$anova$AIC

backward_aic = stepAIC (full_model, scope=list(lower=null_model, upper=full_model),
direction='backward', k = 2, trace = F, data = data)
aic_back = backward_aic$anova$AIC
```

```
summary(backward_aic)
```

```
##
## Call:
## glm(formula = risk_temp ~ data$grouped_riskystation + RHmax +
##     Tmax_lapse + Tmin_lapse + WS + LH + CC1 + CC2 + CC4 + PPT2 +
##     PPT4 + lon + DEM + Slope + Solar.radiation, family = binomial(link = logit),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9269  -0.3485   0.1587   0.4657   2.5458
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.724e+02  5.984e+01   4.552 5.32e-06 ***
```

```
## data$grouped_riskystation  6.412e-01  8.170e-02   7.848 4.24e-15 ***
## RHmax                      1.092e-02  6.156e-03   1.774 0.076084 .
## Tmax_lapse                 5.645e-01  2.854e-02  19.779  < 2e-16 ***
## Tmin_lapse                 4.016e-01  2.841e-02  14.136  < 2e-16 ***
## WS                        -2.150e-01  1.868e-02 -11.508  < 2e-16 ***
## LH                         9.368e-03  1.307e-03   7.169 7.57e-13 ***
## CC1                       -1.815e+00  2.148e-01  -8.452  < 2e-16 ***
## CC2                       -1.269e+00  2.926e-01  -4.336 1.45e-05 ***
## CC4                       -2.169e+00  1.834e-01 -11.823  < 2e-16 ***
## PPT2                       2.151e-01  2.187e-02   9.835  < 2e-16 ***
## PPT4                      -1.033e-01  4.006e-02  -2.579 0.009898 **
## lon                       -2.343e+00  4.720e-01  -4.965 6.86e-07 ***
## DEM                       -4.326e-03  1.117e-03  -3.874 0.000107 ***
## Slope                      2.547e-01  4.296e-02   5.928 3.06e-09 ***
## Solar.radiation            2.626e-04  8.605e-05   3.052 0.002274 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10015.6  on 7587  degrees of freedom
## Residual deviance:  4875.1  on 7572  degrees of freedom
## AIC: 4907.1
##
## Number of Fisher Scoring iterations: 6
```

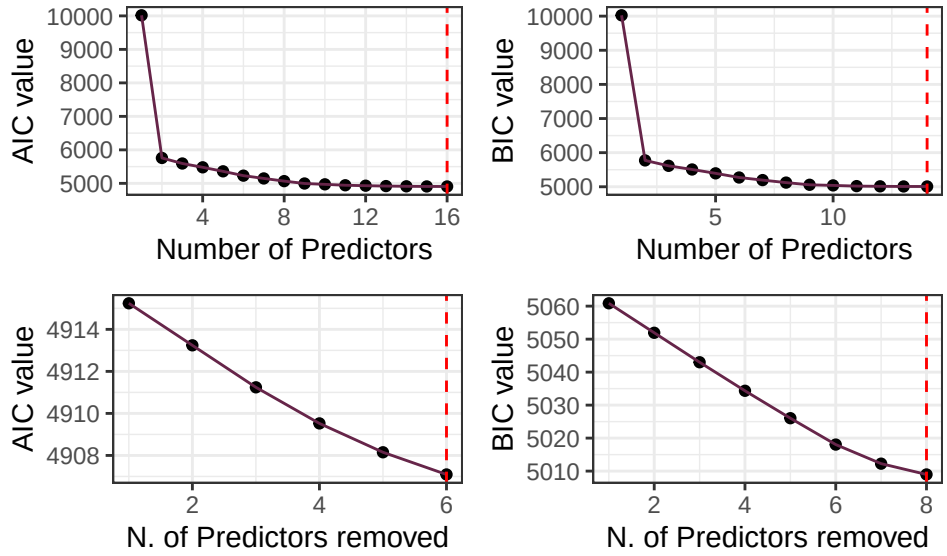
```
summary(backward_bic)
```

```
##
## Call:
## glm(formula = risk_temp ~ data$grouped_riskystation + Tmax_lapse +
##     Tmin_lapse + WS + LH + CC1 + CC2 + CC4 + PPT2 + lon + DEM +
##     Slope + Solar.radiation, family = binomial(link = logit),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9086  -0.3506   0.1577   0.4690   2.5718
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                2.757e+02  5.977e+01   4.614 3.96e-06 ***
## data$grouped_riskystation  6.332e-01  8.155e-02   7.764 8.21e-15 ***
## Tmax_lapse                 5.541e-01  2.812e-02  19.707  < 2e-16 ***
## Tmin_lapse                 4.090e-01  2.817e-02  14.518  < 2e-16 ***
## WS                        -2.223e-01  1.846e-02 -12.040  < 2e-16 ***
## LH                         9.908e-03  1.238e-03   8.003 1.22e-15 ***
## CC1                       -1.724e+00  2.100e-01  -8.212  < 2e-16 ***
## CC2                       -1.162e+00  2.898e-01  -4.010 6.08e-05 ***
## CC4                       -2.354e+00  1.727e-01 -13.630  < 2e-16 ***
## PPT2                       2.117e-01  2.192e-02   9.659  < 2e-16 ***
## lon                       -2.362e+00  4.714e-01  -5.010 5.45e-07 ***
## DEM                       -4.262e-03  1.115e-03  -3.821 0.000133 ***
## Slope                      2.650e-01  4.254e-02   6.230 4.66e-10 ***
## Solar.radiation            2.588e-04  8.591e-05   3.012 0.002593 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10015.6  on 7587  degrees of freedom
## Residual deviance:  4883.9  on 7574  degrees of freedom
## AIC: 4911.9
##
## Number of Fisher Scoring iterations: 6
```

0.5 Best overall model

The best model selected by using the AIC criterion (both with the forward and backward step-wise methods) has 5 predictors less compared to the full model: RHmin, CC3, PPT1, PPT3 and lat are dropped; on the other hand, the outcomes obtained with the BIC criterion are mode parsimonious, as they also exclude the variables PPT4 and RHmax.



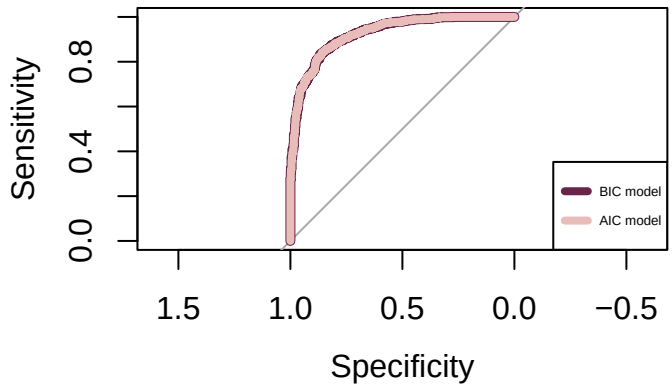
Finally, by comparing the ROC curves of the two models obtained with the AIC and BIC criteria, we can observe that they overlap, suggesting that their predictive power is comparable. In conclusion, it is better to select the model with less predictors.

```
set.seed(123)
train.idx = sample(nrow(data), 0.8*nrow(data))
train = data[train.idx, ]
test = data[-train.idx, ]

logitmod_bic = glm(risk_temp ~ grouped_riskystation + Tmax_lapse + LH + lon + WS
+ Tmin_lapse + CC1 + CC2 + CC4 + PPT2 + DEM + Slope + Solar.radiation,
family = binomial(link = "logit"), data = train)

predicted_bic = predict(logitmod_bic, test, type = "response")
```

```
logitmod_aic = glm(risk_temp ~ grouped_riskystation + Tmax_lapse + LH + lon + WS +
Tmin_lapse + CC1 + CC2 + CC4 + PPT2 + PPT4 + DEM + Slope + Solar.radiation +
RHmax, family = binomial(link = "logit"), data = train)
```

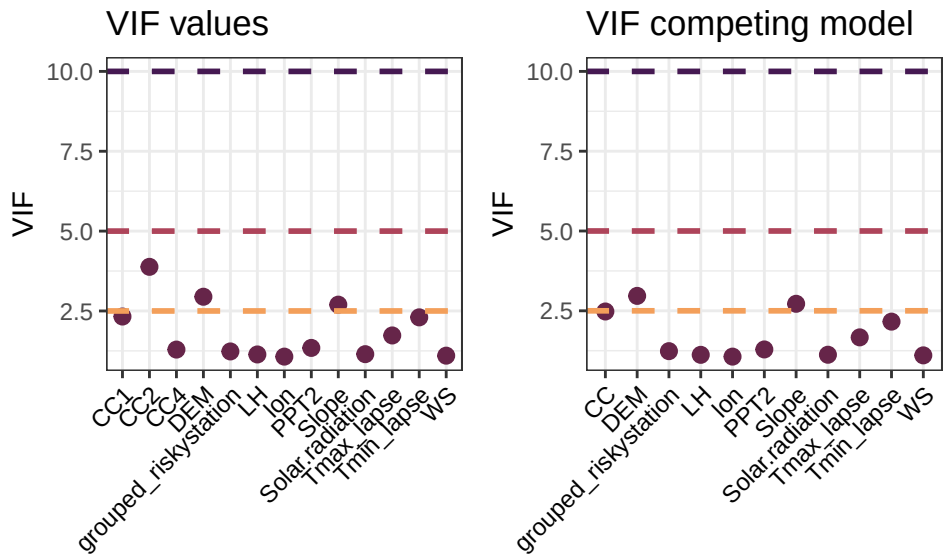
```
predicted_aic = predict(logitmod_aic, test, type = "response")
```

```
roc_bic <- pROC::roc(test$risk_temp, predicted_bic)
roc_aic <- pROC::roc(test$risk_temp, predicted_aic)

plot(roc_bic, col = pal[3], lwd = 5)
plot(roc_aic, col = pal[6], add = TRUE, lwd = 4)
legend("bottomright", legend = c("BIC model", "AIC model"), col = c(pal[3], pal[6]),
lty = 1, lwd = 4, cex = 0.4)
```



0.6 Potential collinearity issues

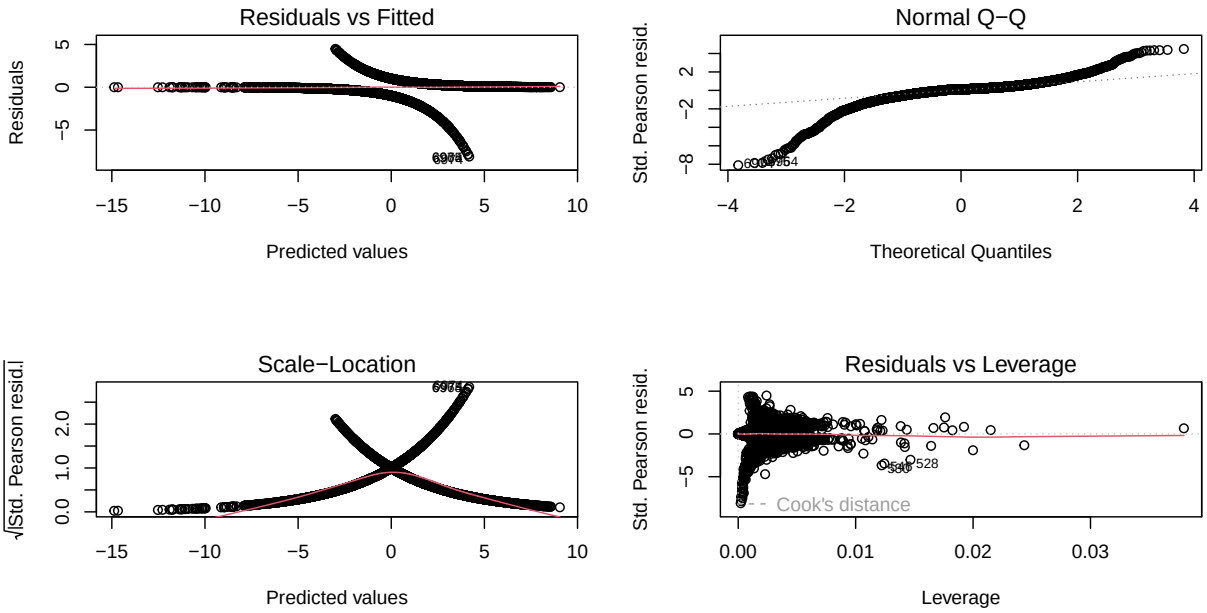
The VIF values of the predictors of the selected model are not worryingly high, however a potential competing model could be taken into consideration: we calculate the daily average values of cloud coverage, since the variable with the highest VIF value is CC2 (average CC between 6.00 and 11.00 a.m.); by doing so, the collinearity among the variables corresponding to each time split is reduced, while the other parameters remain almost unchanged.



0.7 Regression diagnostics

The generic plot function is employed to conduct regression diagnostics. Since it is normal to see patterns in the residual plots for logistic regression models, the first and third plots are not indicative of any particular issue with the model; additionally, the Q-Q plot is not relevant because we do not assume normality. Finally, the last plot shows that there are neither high leverage points nor outliers, based on Cook's distance. No improvement is needed.

```
par(mfrow = c(2, 2))
plot(logitmod)
```



0.8 The coefficients

We print the exponentiated coefficients and their confidence intervals, as well as a plot showing their values and significance.

- The extremely large coefficient value of the intercept suggests an extremely large odds ratio, when all the variables are set to 0.
- For a one degree Celsius increase in the prediction of the next-day maximum temperature, having a heat wave temperature is approximately 1.67 times more likely than not, holding all other predictors constant.
- For a one percent increase in forecasted daily average cloud coverage, the outcome a risky temperature occurring is approximately 200 times less likely than that not, holding all other predictors constant.
- For a one-unit (m/s) increase in forecasted wind speed, the odds of the outcome are approximately equal to 0.80, holding all other predictors constant, meaning that it is less likely to have a heat wave than not.
- For a one degree Celsius increase in the forecasted next-day minimum temperature, having a heat wave temperature is approximately 1.57 times more likely than not, holding all other predictors constant.
- For a one-unit increase in forecasted latent heat flux, the probability of the outcome occurring is almost equal to the probability of it not occurring, holding all other predictors constant.
- For a one percent increase in forecasted precipitation between 6.00 and 11.00, the outcome is approximately 1.24 times more likely to occur than not, holding all other predictors constant.

- When considering a station located in an area where the observed temperature is risky 65% of the times or more, the outcome is approximately 1.87 times more likely to occur than not, holding all other predictors constant.
- For a one degree increase in slope, the outcome is approximately 1.31 times more likely to occur than not, holding all other predictors constant.
- For a one degree increase in longitude, the odds of the outcome are approximately equal to 0.12, holding all other predictors constant, meaning that it is much less likely to have a heat wave than not.
- For a one meter increase in elevation, the probability of the outcome occurring is almost equal to the probability of it not occurring, holding all other predictors constant.
- For a one-unit (Wh/m^2) increase in Solar radiation, the probability of the outcome occurring is almost equal to the probability of it not occurring, holding all other predictors constant.

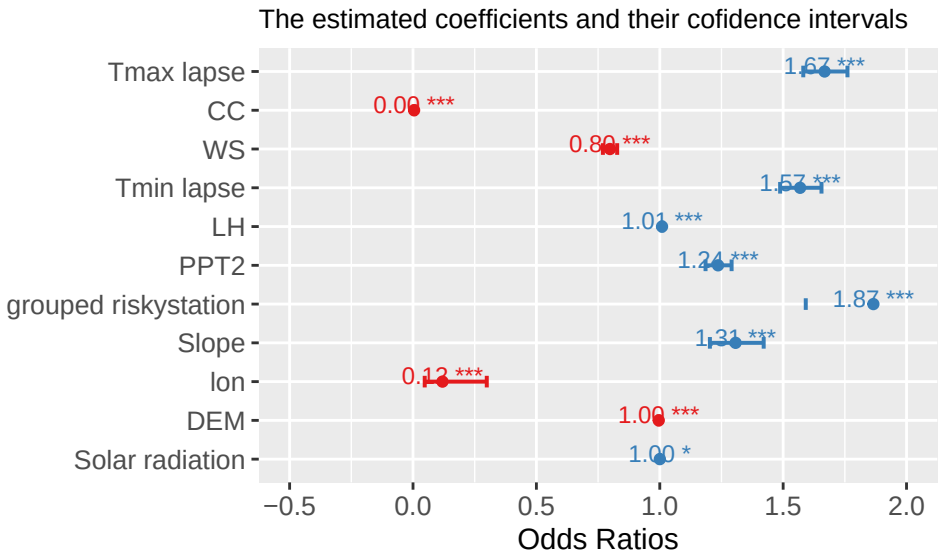
```
exp(logitmod$coefficients)
```

```
##          (Intercept)           Tmax_lapse                   CC
##        1.064778e+107         1.667914e+00         4.451985e-03
##                   WS           Tmin_lapse                   LH
##         7.981821e-01         1.568707e+00         1.009143e+00
##                 PPT2 grouped_riskystation                Slope
##         1.236188e+00         1.865623e+00         1.307233e+00
##                  lon                  DEM      Solar.radiation
##         1.194059e-01         9.957656e-01         1.000205e+00
```

```
confint(logitmod)
```

```
##                              2.5 %        97.5 %
## (Intercept)           1.300528e+02  3.633378e+02
## Tmax_lapse            4.584331e-01  5.656590e-01
## CC                   -5.992298e+00 -4.843957e+00
## WS                   -2.618989e-01 -1.895006e-01
## Tmin_lapse            3.972739e-01  5.039975e-01
## LH                    6.720746e-03  1.150110e-02
## PPT2                  1.701121e-01  2.554379e-01
## grouped_riskystation  4.647246e-01  7.833526e-01
## Slope                 1.848835e-01  3.514242e-01
## lon                  -3.047315e+00 -1.207469e+00
## DEM                  -6.430290e-03 -2.060497e-03
## Solar.radiation       3.857130e-05  3.720798e-04
```

```
plot_model(logitmod, width = 0.3, show.intercept = F, show.p = T, show.values = T,
value.size = 3, dot.size = 1.5) + ylim(-0.5, 2) + labs(title = "The estimated coefficients and their co
axis.title.x = element_text(size = 11), axis.title.y = element_text(size = 11),
axis.text = element_text(size = 10))
```



0.9 Hypothesis testing of each coefficient

To test each of the coefficient to be 0 against two-sided alternatives, we use the anova function with X^2 test. We can conclude that each coefficient is significantly different from 0, because their p values are lower than the significance level of 0.05 (the null hypothesis is rejected).

```
drop1(logitmod, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## risk_temp ~ Tmax_lapse + CC + WS + Tmin_lapse + LH + PPT2 + grouped_riskystation +
##     Slope + lon + DEM + Solar.radiation
##                       Df Deviance    AIC    LRT  Pr(>Chi)
## <none>                     4905.1 4929.1
## Tmax_lapse             1   5308.9 5330.9 403.85 < 2.2e-16 ***
## CC                     1   5279.1 5301.1 373.99 < 2.2e-16 ***
## WS                     1   5066.3 5088.3 161.21 < 2.2e-16 ***
## Tmin_lapse             1   5206.1 5228.1 301.02 < 2.2e-16 ***
## LH                     1   4962.2 4984.2  57.08 4.184e-14 ***
## PPT2                   1   5012.2 5034.2 107.14 < 2.2e-16 ***
## grouped_riskystation   1   4964.9 4986.9  59.83 1.033e-14 ***
## Slope                  1   4945.5 4967.5  40.36 2.110e-10 ***
## lon                    1   4925.8 4947.8  20.69 5.400e-06 ***
## DEM                    1   4919.6 4941.6  14.52 0.0001385 ***
## Solar.radiation        1   4910.9 4932.9   5.83 0.0157686 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

0.10 Hypothesis testing on a group of regressors

Now we test a group of regressors, after removing the variables Solar.radiation, DEM, LH, because their coefficients are close to 1. The null hypothesis of the test states that the reduced model fits the data as well as the bigger model. Since the p-value is close to 0, we can reject the null hypothesis and conclude that the larger model is significantly better than the reduced model.

```r
mod_reduced = glm(risk_temp ~ Tmax_lapse + CC + WS + Tmin_lapse + PPT2 +
grouped_riskystation + Slope + lon, family = binomial(link = logit),
data = data)

anova(logitmod, mod_reduced, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: risk_temp ~ Tmax_lapse + CC + WS + Tmin_lapse + LH + PPT2 + grouped_riskystation +
##     Slope + lon + DEM + Solar.radiation
## Model 2: risk_temp ~ Tmax_lapse + CC + WS + Tmin_lapse + PPT2 + grouped_riskystation +
##     Slope + lon
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      7576     4905.1
## 2      7579     4985.5 -3  -80.459 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

0.11 Goodness of fit

To check whether the model suits the data, the Hosmer-Lemeshow test - with null hypothesis that the model suits - is used. Since $p < 0.05$, the null hypothesis is rejected, as the fit is not adequate.

```r
hoslem.test(data$risk_temp, predict(logitmod, type = "response"))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  data$risk_temp, predict(logitmod, type = "response")
## X-squared = 18.706, df = 8, p-value = 0.01652
```

0.12 Prediction with new observations

We use the same split of the data as before to train the model on a random sample composed of 80% of the data entries and then produce the predictions based on the test set. We then randomly select one of the predictions and compare it to the corresponding observed value, after classifying it based on the optimal cut-point given by Youden's index (equal to 68.3%). Since the predicted probability is equal to 35.4%, the prediction is classified as "not risky", which corresponds to the actual observed value. The standard error is low, suggesting that the predicted probability is relatively precise, in other words, close to the true underlying probability.

```r
logitmod_prediction = glm(risk_temp ~ Tmax_lapse + CC + WS + Tmin_lapse  + LH + PPT2 +
grouped_riskystation + Slope + lon  + DEM + Solar.radiation,
family = binomial(link = "logit"), data = train)

prediction = predict(logitmod_prediction, newdata = test, type = "response", se.fit = T)
prediction$fit[1518]
```

```
##      7748
## 0.3538155
```

```r
prediction$se.fit[1518]
```

```
##       7748
## 0.03437121
```

```
youden_ind = cutpointr::cutpointr(prediction$fit, test$risk_temp,
method = maximize_metric, metric = sum_sens_spec)
youden_ind$optimal_cutpoint
```

```
## [1] 0.6832449
```

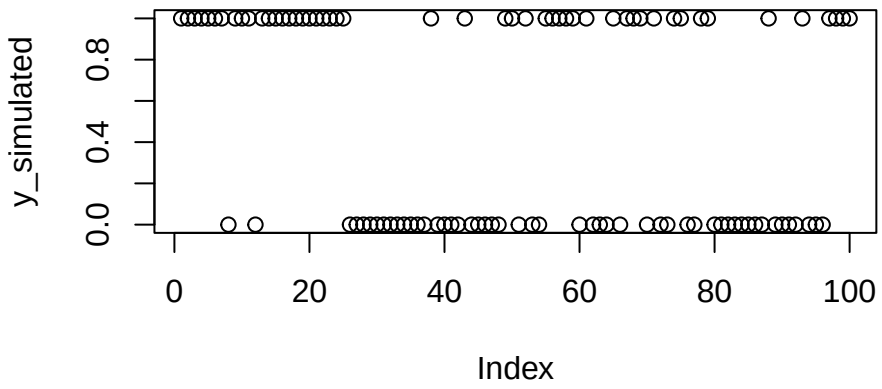
```
test$risk_temp[1518]
```

```
## [1] 0
```

0.13 Simulation of n data points from the fitted regression model

We simulate 100 datapoints, assuming that the estimated parameters are the true parameters.

```
coefficients = coef(logitmod)
n = 100

predictor_list = list(data$Tmax_lapse, data$CC, data$WS, data$Tmin_lapse, data$LH,
data$PPT2, data$grouped_riskystation, data$Slope, data$lon, data$DEM, data$Solar.radiation)

terms = list()

for (i in 1:length(predictor_list)) {
  term = coefficients[i+1] * predictor_list[[i]]
  terms = c(terms, list(term))}

eta_simulated = numeric(n)
prob_success = numeric(n)

for (i in 1:n) {
  eta_simulated[i] = coefficients["(Intercept)"] + sum(unlist(lapply(terms, `[`, i)))
  prob_success[i] = exp(eta_simulated[i]) / (1 + exp(eta_simulated[i]))}

y_simulated = rbinom(n, size = 1, prob = prob_success)
plot(y_simulated)
```



0.14 References

Cho, D., Yoo, C., Im, J., & Cha, D.-H., (2020). Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. Earth and Space Science, 7, e2019EA000740. https://doi.org/10.1029/2019EA000740

Ji-Young Son, Jong-Tae Lee, G. Brooke Anderson, and Michelle L. Bell, (2012). The Impact of Heat Waves on Mortality in Seven Major Cities in Korea. Environmental Health Perspectives, vol. 120, n. 4.