Precipitation forecast

Code

Achir Oukelmoun

August 31, 2019

1 Data import and description

1.1 Data import and first adjustment

In this section, the training data is imported and irrelevant variables, such as hours and minutes, are removed.

Variables associated with wind directions use quantitative values from 0 - 360 but it is misleading in terms of regression since direction associated with the ends of the interval (0 and 360) are close while numerically they are not. Therefore, variable associated with wind direction are removed and 4 qualititative variables are created, each one associated with a given wind direction. An intermediate mean_direction is used. The mean_direction takes the mean direction taken at different heights. The hidden code below indicates how the wind direction variables are defined.

Code

For the same reason as wind directions, the variable associated with months is transformed to factors. Indeed, the variable month takes values ranging from 1 to 12, which is misleading since, for instance, the month 12 is closer to month 1 than month 10.

The names of the variables (columns) are modified so that they can be easily plotted. The actual columns names are stored in a list “t”. The key for short names is provided below:

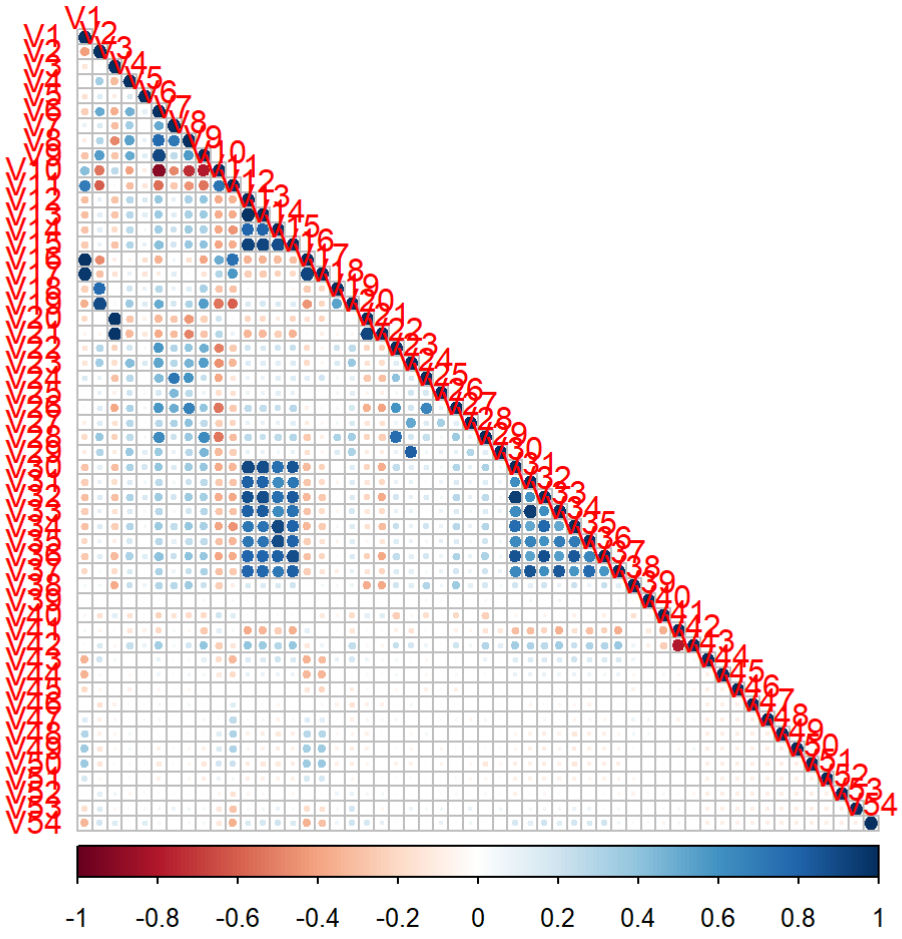
```
## V1: Temperature.daily.mean..2.m.above.gnd.
## V2: Relative.Humidity.daily.mean..2.m.above.gnd.
## V3: Mean.Sea.Level.Pressure.daily.mean..MSL.
## V4: Total.Precipitation.daily.sum..sfc.
## V5: Snowfall.amount.raw.daily.sum..sfc.
## V6: Total.Cloud.Cover.daily.mean..sfc.
## V7: High.Cloud.Cover.daily.mean..high.cld.lay.
## V8: Medium.Cloud.Cover.daily.mean..mid.cld.lay.
## V9: Low.Cloud.Cover.daily.mean..low.cld.lay.
## V10: Sunshine.Duration.daily.sum..sfc.
## V11: Shortwave.Radiation.daily.sum..sfc.
## V12: Wind.Speed.daily.mean..10.m.above.gnd.
## V13: Wind.Speed.daily.mean..80.m.above.gnd.
## V14: Wind.Speed.daily.mean..900.mb.
## V15: Wind.Gust.daily.mean..sfc.
## V16: Temperature.daily.max..2.m.above.gnd.
## V17: Temperature.daily.min..2.m.above.gnd.
## V18: Relative.Humidity.daily.max..2.m.above.gnd.
## V19: Relative.Humidity.daily.min..2.m.above.gnd.
## V20: Mean.Sea.Level.Pressure.daily.max..MSL.
## V21: Mean.Sea.Level.Pressure.daily.min..MSL.
## V22: Total.Cloud.Cover.daily.max..sfc.
## V23: Total.Cloud.Cover.daily.min..sfc.
## V24: High.Cloud.Cover.daily.max..high.cld.lay.
## V25: High.Cloud.Cover.daily.min..high.cld.lay.
## V26: Medium.Cloud.Cover.daily.max..mid.cld.lay.
## V27: Medium.Cloud.Cover.daily.min..mid.cld.lay.
## V28: Low.Cloud.Cover.daily.max..low.cld.lay.
## V29: Low.Cloud.Cover.daily.min..low.cld.lay.
## V30: Wind.Speed.daily.max..10.m.above.gnd.
## V31: Wind.Speed.daily.min..10.m.above.gnd.
## V32: Wind.Speed.daily.max..80.m.above.gnd.
## V33: Wind.Speed.daily.min..80.m.above.gnd.
## V34: Wind.Speed.daily.max..900.mb.
## V35: Wind.Speed.daily.min..900.mb.
## V36: Wind.Gust.daily.max..sfc.
## V37: Wind.Gust.daily.min..sfc.
## V38: pluie.demain
## V39: Wind_N
## V40: Wind_E
## V41: Wind_S
## V42: Wind_W
## V43: M1
## V44: M2
## V45: M3
## V46: M4
## V47: M5
## V48: M6
## V49: M7
## V50: M8
## V51: M9
## V52: M10
## V53: M11
## V54: M12
```

The variables from V43 (M1) to V54 (M12) result from the transformation of the variable Month into factors. In fact, each month x has been associated with an Mx variable taking TRUE or FALSE as values.

1.2 Correlation analysis

In this section, the correlation between the variables is studied in order to avoid a high collinearity between the variables selected for the next steps of the analysis.

The correlation matrix is plotted below:



A strong correlation between several variables is observed. This is not surprising because many variables refer to the same physical quantity but taken at different heights and times.

Given that:

- There is a high colinearity between variables
- The high number of variables

A simple algorithm is developed in the following section to avoid any high collinearity between two variables.

2 Model Selection

2.1 Description of the methodology

The approach to select the best model will consist of several steps:

- Remove variables with high correlation so as to ensure that two selected variables have a correlation which is less than a given threhold S. Indeed, for each variable X, the other variables highly correlated with the variable X will be removed. This step is sensitive to the ordre of apperance of variables in the data frame, this is why the varibales will be shuffled several times using pre-defined seeds. 10 shuffles will be considered and 5 correlation thresholds.

- Use stepAIC function to select the model with the lowest AIC
- Use several threshold probability to predict. 7 probability thresholds will be considered.

To summarize, 10 seeds to shuffle the order of appearance of variables in the dataframe, 5 threshold of correlation between variables and 7 probability thresholds will be used during a 8-fold cross-validation. Then, the set (seed, correlation threshold, and probability threshold) associated with the lowest error will be retained.

2.2 The R code used

The hidden R code below is the one used to automate the choice of model. Comments are added to better describe the purpose of each instruction.

Code

The retained model is therefore the following:

```
##   V5 + V7 + V9 + V20 + V26 + V28 + V34 + V40 + V41 + V44 + V45 + V46 + V50 + V51 + V52 + V
53
```

```
## Prediction threshold is: 0.47
```

```
## Mean error is:  0.29
```

```
## where :
```

```
## V5: Snowfall.amount.raw.daily.sum..sfc.
## V7: High.Cloud.Cover.daily.mean..high.cld.lay.
## V9: Low.Cloud.Cover.daily.mean..low.cld.lay.
## V20: Mean.Sea.Level.Pressure.daily.max..MSL.
## V26: Medium.Cloud.Cover.daily.max..mid.cld.lay.
## V28: Low.Cloud.Cover.daily.max..low.cld.lay.
## V34: Wind.Speed.daily.max..900.mb.
## V40: Wind_E
## V41: Wind_S
## V44: M2
## V45: M3
## V46: M4
## V50: M8
## V51: M9
## V52: M10
## V53: M11
```

2.3 Performance of the model

The perfomance of the model is estimated by a leave-one-out validation. The results are the following:

```
## The error estimate is therefore:  0.294
```

The variables used are consistent but the performance of the model seems to be relatively low. In addition, the testing of several combinations through cross-validation did not result in a significant improvement in model performance.

The summary of the model selected is as follows:

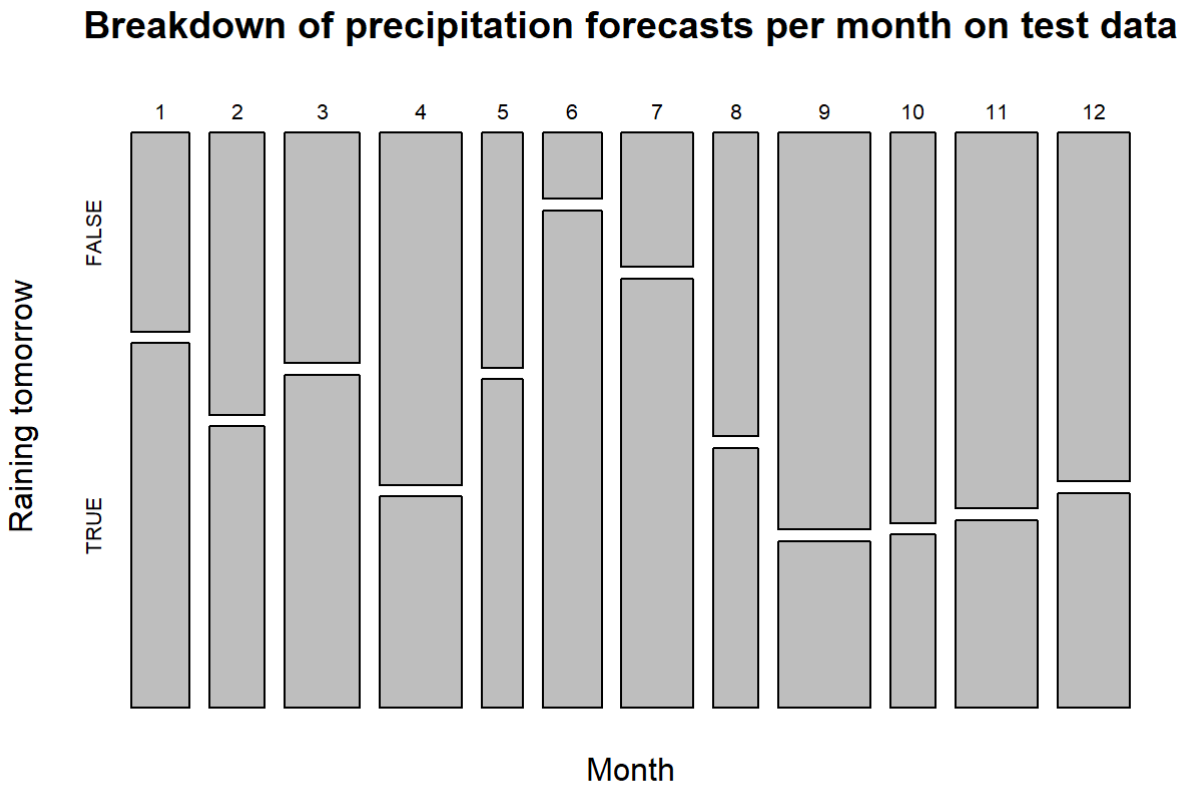
```
##
## Call:
## glm(formula = formule[[1]], family = binomial, data = d2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3766  -0.9088  -0.2736   0.9256   2.1778
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 87.326277  10.623006   8.220  < 2e-16 ***
## V5           0.366521   0.279084   1.313 0.189082
## V7           0.011129   0.003772   2.950 0.003175 **
## V9           0.004733   0.002940   1.610 0.107425
## V20         -0.086612   0.010353  -8.366  < 2e-16 ***
## V26          0.010013   0.002073   4.830 1.37e-06 ***
## V28         -0.003531   0.002386  -1.480 0.138917
## V34          0.011256   0.003443   3.269 0.001078 **
## V40         -0.554255   0.272544  -2.034 0.041988 *
## V41          0.262963   0.153201   1.716 0.086077 .
## V44TRUE     -0.823222   0.258506  -3.185 0.001450 **
## V45TRUE     -0.962438   0.266872  -3.606 0.000311 ***
## V46TRUE     -0.817596   0.256925  -3.182 0.001461 **
## V50TRUE     -0.484532   0.234868  -2.063 0.039113 *
## V51TRUE     -0.768195   0.255167  -3.011 0.002608 **
## V52TRUE     -0.883169   0.242943  -3.635 0.000278 ***
## V53TRUE     -0.898593   0.265577  -3.384 0.000716 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1724.5  on 1243  degrees of freedom
## Residual deviance: 1385.3  on 1227  degrees of freedom
## AIC: 1419.3
##
## Number of Fisher Scoring iterations: 5
```

All the variables selected are significant and the selected model is the one that minimizes the AIC. For example, increased cloud cover (V7) increases the probability of precipitation, while higher wind speed decreases the probability of precipitation.

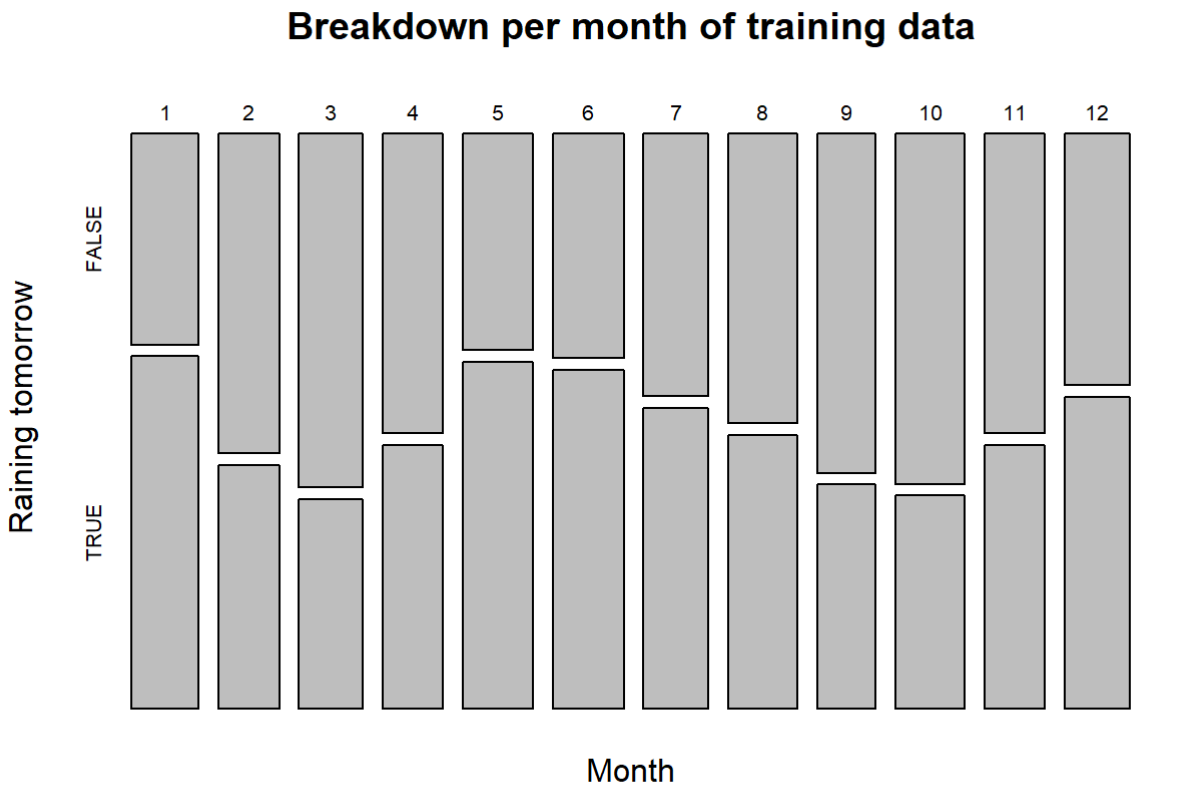
3 Testing the model

In this section, the test data set is imported and the same transformation on the variable are applied so as to make the predictions. The predictions are stored in the variable “Pluie.demain” of the csv file “meteo.test.predictions.csv”.

Therefore, to check beforehand that the prediction are somehow consistent with the previous training data, the following mosaic plots provide the beakdown of the predictions by month for test data and training data respectively.



While the breakdown of the prediction by month of training data is provided below:



By comparing the plots above, about the same behavior is observed, but not exactly the same since June is predicted to be the rainiest month while in the training data June is the third rainiest month. Finally, the model selected does not have a high level of accuracy, but provides globally acceptable results.