Modèle linéaire généralisé et Choix de modèles

Natacha BABALOLA.

2023-07-01

Table of Contents

Météo à Bâle

Le fichier meteo.train.csv contient des données sur les conditions météorologiques à Bâle (Suisse). Chaque ligne correspond à un jour entre 2010 et 2018. Les colonnes correspondent aux valeurs moyenne, minimale et maximale sur la journée de :

- Température (°C)
- Humidite relative (pourcentage)
- Pression (hPa)
- Nébulosité (pourcentage)
- Nébulosité forte, moyenne et faible
- Vitesse (en km/h) et direction (en degrés) du vent à 10 m d'altitude, 80 m d'altitude, et à l'altitude où la pression vaut 900 hPa
- Rafales de vent à 10 m

ainsi qu'aux valeurs totales sur la journée de :

- Précipitations (mm)
- Neige (cm)
- Minutes d'ensoleillement
- Rayonnement solaire (W/m2)

On cherche à prédire s'il pleuvra le lendemain (colonne pluie.demain). Pour cette variable d'intérêt :

- proposer et valider un modèle ;
- proposer une prédiction binaire pour les lendemains des journées incluses dans le fichier meteo.test.csv.

Source des données : MeteoBlue.

Import des données

Chargement des données Train et Test

Présentation des 1ères ligne de notre base d'entrainement

```
   # A tibble: 6 × 46
     pluie…¹  Year Month   Day  Hour Minute Tempe…² Relat…³ Mean.…⁴ Total…⁵
Snowf…⁶
     <lgl>   <dbl> <dbl> <dbl> <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
<dbl>
   1 FALSE    2010     6     2     0      0    15.0    76.5   1015.     1
0
   2 FALSE    2010     6     4     0      0    17.3    77.6   1017.     0
0
   3 TRUE     2010     6     6     0      0    21.6    69.5   1015.     3.7
0
```

```
   4 TRUE     2010     6     8     0      0    20.2    75.1   1008.     0.2
0
   5 TRUE     2010     6    10     0      0    22.6    73.5   1004.     0
0
   6 TRUE     2010     6    12     0      0    18.4    76.8   1012      2.2
0
   # … with 35 more variables: Total.Cloud.Cover.daily.mean..sfc. <dbl>,
   #   High.Cloud.Cover.daily.mean..high.cld.lay. <dbl>,
   #   Medium.Cloud.Cover.daily.mean..mid.cld.lay. <dbl>,
   #   Low.Cloud.Cover.daily.mean..low.cld.lay. <dbl>,
   #   Sunshine.Duration.daily.sum..sfc. <dbl>,
   #   Shortwave.Radiation.daily.sum..sfc. <dbl>,
   #   Wind.Speed.daily.mean..10.m.above.gnd. <dbl>, …

   tibble [6 × 46] (S3: tbl_df/tbl/data.frame)
    $ pluie.demain                                  : logi [1:6] FALSE FALSE
TRUE TRUE TRUE TRUE
    $ Year                                          : num [1:6] 2010 2010 2010
2010 2010 2010
    $ Month                                         : num [1:6] 6 6 6 6 6 6
    $ Day                                           : num [1:6] 2 4 6 8 10 12
    $ Hour                                          : num [1:6] 0 0 0 0 0 0
    $ Minute                                        : num [1:6] 0 0 0 0 0 0
    $ Temperature.daily.mean..2.m.above.gnd.        : num [1:6] 15 17.3 21.6
20.2 22.6 ...
    $ Relative.Humidity.daily.mean..2.m.above.gnd.: num [1:6] 76.5 77.6 69.5
75.1 73.5 ...
    $ Mean.Sea.Level.Pressure.daily.mean..MSL.      : num [1:6] 1015 1017 1015
1008 1004 ...
    $ Total.Precipitation.daily.sum..sfc.           : num [1:6] 1 0 3.7 0.2 0
2.2
    $ Snowfall.amount.raw.daily.sum..sfc.           : num [1:6] 0 0 0 0 0 0
    $ Total.Cloud.Cover.daily.mean..sfc.            : num [1:6] 79.8 4.7 42.1
67.5 56.3 ...
    $ High.Cloud.Cover.daily.mean..high.cld.lay.    : num [1:6] 3 0.67 21.21
54.71 50.25 ...
    $ Medium.Cloud.Cover.daily.mean..mid.cld.lay. : num [1:6] 31.6 0 25.9
65.8 55.3 ...
    $ Low.Cloud.Cover.daily.mean..low.cld.lay.      : num [1:6] 79.2 4.5 35.3
18.9 34.2 ...
    $ Sunshine.Duration.daily.sum..sfc.             : num [1:6] 287.2 821.4
441.3 41.9 473.2 ...
    $ Shortwave.Radiation.daily.sum..sfc.           : num [1:6] 6710 7974 4834
5390 7216 ...
    $ Wind.Speed.daily.mean..10.m.above.gnd.        : num [1:6] 11.64 6.34 8.4
5.4 9.16 ...
    $ Wind.Direction.daily.mean..10.m.above.gnd.    : num [1:6] 275 230 215 205
179 ...
    $ Wind.Speed.daily.mean..80.m.above.gnd.        : num [1:6] 14.99 8.92
10.38 6.53 11.91 ...
```

```
    $ Wind.Direction.daily.mean..80.m.above.gnd.  : num [1:6] 268 199 208 206
186 ...
    $ Wind.Speed.daily.mean..900.mb.              : num [1:6] 20.6 27.9 18.9
10.4 21.9 ...
    $ Wind.Direction.daily.mean..900.mb.          : num [1:6] 180.4 93.7
250.1 238.6 153 ...
    $ Wind.Gust.daily.mean..sfc.                  : num [1:6] 14.88 9.48 13.5
5.31 12.21 ...
    $ Temperature.daily.max..2.m.above.gnd.       : num [1:6] 18.5 25 26.2
24.2 30.7 ...
    $ Temperature.daily.min..2.m.above.gnd.       : num [1:6] 11.1 10.4 17.7
14.7 16.9 ...
    $ Relative.Humidity.daily.max..2.m.above.gnd. : num [1:6] 94 92 91 89 97
92
    $ Relative.Humidity.daily.min..2.m.above.gnd. : num [1:6] 59 54 57 62 39
65
    $ Mean.Sea.Level.Pressure.daily.max..MSL.     : num [1:6] 1017 1019 1016
1010 1006 ...
    $ Mean.Sea.Level.Pressure.daily.min..MSL.     : num [1:6] 1014 1016 1013
1006 1001 ...
    $ Total.Cloud.Cover.daily.max..sfc.           : num [1:6] 100 28 100 100
100 100
    $ Total.Cloud.Cover.daily.min..sfc.           : num [1:6] 0 0 0 0 0 0
    $ High.Cloud.Cover.daily.max..high.cld.lay.   : num [1:6] 16 11 100 100
100 28
    $ High.Cloud.Cover.daily.min..high.cld.lay.   : num [1:6] 0 0 0 0 0 0
    $ Medium.Cloud.Cover.daily.max..mid.cld.lay.  : num [1:6] 100 0 100 100
100 100
    $ Medium.Cloud.Cover.daily.min..mid.cld.lay.  : num [1:6] 0 0 0 0 0 0
    $ Low.Cloud.Cover.daily.max..low.cld.lay.     : num [1:6] 100 28 100 100
100 100
    $ Low.Cloud.Cover.daily.min..low.cld.lay.     : num [1:6] 0 0 0 0 0 0
    $ Wind.Speed.daily.max..10.m.above.gnd.       : num [1:6] 22 15.5 22.7
10.7 20.5 ...
    $ Wind.Speed.daily.min..10.m.above.gnd.       : num [1:6] 5.62 1.08 2.41
0 2.52 2.28
    $ Wind.Speed.daily.max..80.m.above.gnd.       : num [1:6] 23.8 18.7 32
10.2 23.4 ...
    $ Wind.Speed.daily.min..80.m.above.gnd.       : num [1:6] 8.65 0 0.51
1.44 2.97 3.1
    $ Wind.Speed.daily.max..900.mb.               : num [1:6] 32.1 48.1 44
22.2 40.8 ...
    $ Wind.Speed.daily.min..900.mb.               : num [1:6] 12.25 6.62 5.48
4.69 4.68 ...
    $ Wind.Gust.daily.max..sfc.                   : num [1:6] 25.2 20.2 41.8
11.2 24.1 ...
    $ Wind.Gust.daily.min..sfc.                   : num [1:6] 6.48 2.16 1.08
0.36 1.44 3.96
   NULL
```

Nous retrouvons pour chacune des covariables, le nombre de valeurs manquantes, la moyenne et d'autre données statistiques comme les différents quartiles.

Analyse descriptive des variables

Analyse de la varible réponse : Ici pluie.demain

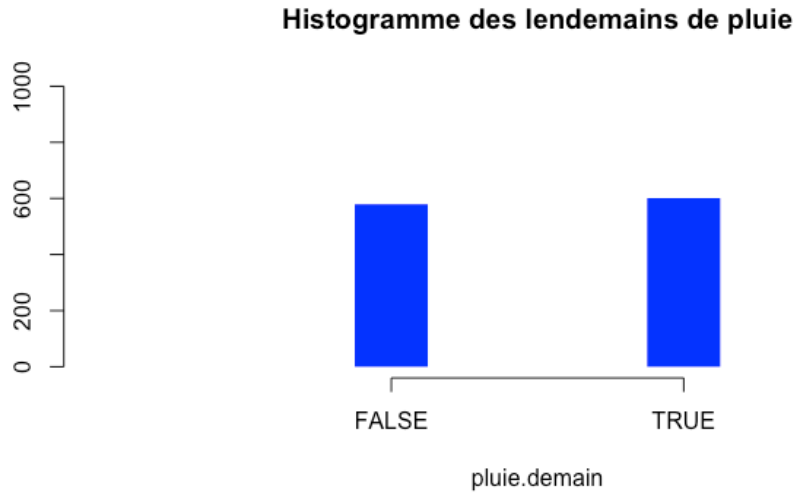
```
   pluie.demain
   FALSE  TRUE
     579   601
```

Recherche des Valeurs Manquantes

```
which(is.na(train), arr.ind = TRUE)

        row col
```

Nous n'avons aucune données manquantes dans notre base d'entrainement.

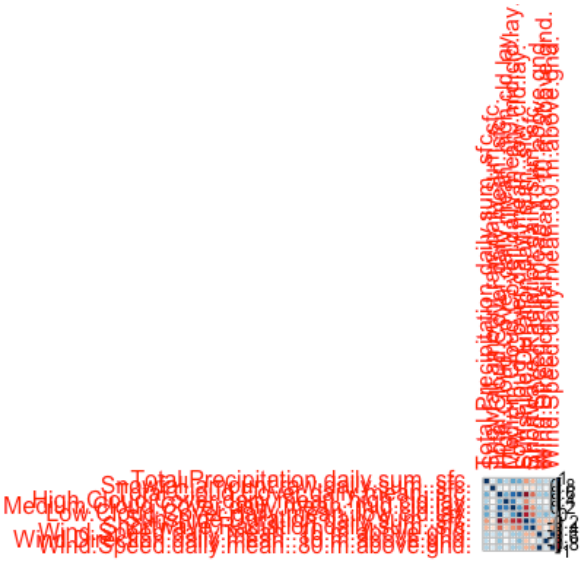


La variable réponse qui nous indique s'il pleut le lendemain dispose ici d'autant de résultats Vrai et Faux.

Le nombre de réalisation ou pas de pluie le lendemain en fonction de la température moyenne. Nous remarquons qu'avec des températures eleve la veille, les chances qu'ils pleuve le lendemain sont plus grande.

Test de corrélation entre nos covariables

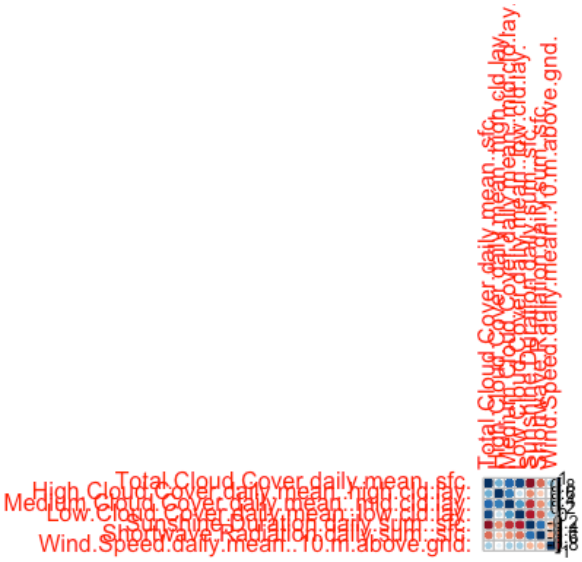
```
   Warning in corrplot(cor(train[, 10:20], use = "complete")): Not been able
to
   calculate text margin, please try again with a clean new empty window
using
   {plot.new(); dev.off()} or reduce tl.cex
```



Un leger zoom sur les données portant une part de colinéarité élevée.

```
   Warning in corrplot(cor(train[, 12:18], use = "complete")): Not been able
to
   calculate text margin, please try again with a clean new empty window
```

```
using
   {plot.new(); dev.off()} or reduce tl.cex
```



Nous observons que nos données sont fortement correlé pour la plupart d'entre elles.

Modelisation

GLM sur la base d'entrainement

Premier modèle

modele1 contenant toutes les variables explicatives

```
modele1 <- glm(pluie.demain ~ ., data = train, family = binomial)
summary(modele1)
```

```

Call:
glm(formula = pluie.demain ~ ., family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7911  -0.8301   0.2850   0.8297   2.9293

Coefficients: (2 not defined because of singularities)
                                              Estimate Std. Error z value
(Intercept)                                -7.646e+01  7.061e+01  -1.083
Year                                        6.921e-02  3.486e-02   1.985
Month                                      -1.860e-02  2.493e-02  -0.746
Day                                         1.179e-02  8.160e-03   1.445
Hour                                               NA         NA      NA
Minute                                             NA         NA      NA
Temperature.daily.mean..2.m.above.gnd.      1.830e-01  1.640e-01   1.116
Relative.Humidity.daily.mean..2.m.above.gnd. 1.986e-02  3.243e-02   0.612
Mean.Sea.Level.Pressure.daily.mean..MSL.    5.124e-01  1.394e-01   3.675
Total.Precipitation.daily.sum..sfc.         2.586e-02  2.805e-02   0.922
Snowfall.amount.raw.daily.sum..sfc.        -2.853e-01  2.339e-01  -1.220
Total.Cloud.Cover.daily.mean..sfc.          1.247e-02  1.200e-02   1.039
High.Cloud.Cover.daily.mean..high.cld.lay.  -3.253e-03  6.820e-03  -0.477
Medium.Cloud.Cover.daily.mean..mid.cld.lay.  5.590e-03  6.691e-03   0.835
Low.Cloud.Cover.daily.mean..low.cld.lay.    -4.340e-03  8.114e-03  -0.535
Sunshine.Duration.daily.sum..sfc.           4.908e-04  8.828e-04   0.556
Shortwave.Radiation.daily.sum..sfc.         2.938e-05  9.883e-05   0.297
Wind.Speed.daily.mean..10.m.above.gnd.     -4.640e-02  9.698e-02  -0.478
Wind.Direction.daily.mean..10.m.above.gnd.  5.637e-03  5.768e-03   0.977
Wind.Speed.daily.mean..80.m.above.gnd.     -9.430e-02  6.947e-02  -1.357
Wind.Direction.daily.mean..80.m.above.gnd. -9.489e-03  5.960e-03  -1.592
Wind.Speed.daily.mean..900.mb.              1.834e-02  2.594e-02   0.707
Wind.Direction.daily.mean..900.mb.          5.404e-03  1.451e-03   3.723
Wind.Gust.daily.mean..sfc.                  1.782e-02  3.689e-02   0.483
Temperature.daily.max..2.m.above.gnd.      -1.146e-02  9.593e-02  -0.119
Temperature.daily.min..2.m.above.gnd.      -1.302e-01  8.631e-02  -1.509
Relative.Humidity.daily.max..2.m.above.gnd.  6.722e-05  2.061e-02   0.003
Relative.Humidity.daily.min..2.m.above.gnd. -6.868e-03  1.856e-02  -0.370
Mean.Sea.Level.Pressure.daily.max..MSL.    -2.587e-01  7.502e-02  -3.449
```

```
Mean.Sea.Level.Pressure.daily.min..MSL.      -3.206e-01  7.572e-02  -4.234
Total.Cloud.Cover.daily.max..sfc.             3.412e-03  4.864e-03   0.701
Total.Cloud.Cover.daily.min..sfc.             7.789e-03  6.264e-03   1.243
High.Cloud.Cover.daily.max..high.cld.lay.     3.423e-03  2.886e-03   1.186
High.Cloud.Cover.daily.min..high.cld.lay.     6.148e-03  2.093e-02   0.294
Medium.Cloud.Cover.daily.max..mid.cld.lay.    6.159e-03  3.164e-03   1.946
Medium.Cloud.Cover.daily.min..mid.cld.lay.   -5.295e-03  9.463e-03  -0.560
Low.Cloud.Cover.daily.max..low.cld.lay.       2.944e-03  3.397e-03   0.867
Low.Cloud.Cover.daily.min..low.cld.lay.       1.197e-04  7.017e-03   0.017
Wind.Speed.daily.max..10.m.above.gnd.         5.588e-02  3.448e-02   1.620
Wind.Speed.daily.min..10.m.above.gnd.         1.690e-01  6.415e-02   2.635
Wind.Speed.daily.max..80.m.above.gnd.         3.933e-03  2.845e-02   0.138
Wind.Speed.daily.min..80.m.above.gnd.        -5.304e-02  4.219e-02  -1.257
Wind.Speed.daily.max..900.mb.                -1.342e-02  1.213e-02  -1.106
Wind.Speed.daily.min..900.mb.                -4.050e-03  1.911e-02  -0.212
Wind.Gust.daily.max..sfc.                     2.282e-02  1.730e-02   1.319
Wind.Gust.daily.min..sfc.                     5.101e-03  2.800e-02   0.182
                                             Pr(>|z|)
(Intercept)                                  0.278866
Year                                         0.047125 *
Month                                        0.455776
Day                                          0.148438
Hour                                               NA
Minute                                             NA
Temperature.daily.mean..2.m.above.gnd.       0.264488
Relative.Humidity.daily.mean..2.m.above.gnd. 0.540327
Mean.Sea.Level.Pressure.daily.mean..MSL.     0.000237 ***
Total.Precipitation.daily.sum..sfc.          0.356497
Snowfall.amount.raw.daily.sum..sfc.          0.222560
Total.Cloud.Cover.daily.mean..sfc.           0.298720
High.Cloud.Cover.daily.mean..high.cld.lay.   0.633421
Medium.Cloud.Cover.daily.mean..mid.cld.lay.  0.403515
Low.Cloud.Cover.daily.mean..low.cld.lay.     0.592768
Sunshine.Duration.daily.sum..sfc.            0.578242
Shortwave.Radiation.daily.sum..sfc.          0.766285
Wind.Speed.daily.mean..10.m.above.gnd.       0.632361
Wind.Direction.daily.mean..10.m.above.gnd.   0.328385
Wind.Speed.daily.mean..80.m.above.gnd.       0.174690
Wind.Direction.daily.mean..80.m.above.gnd.   0.111388
Wind.Speed.daily.mean..900.mb.               0.479457
Wind.Direction.daily.mean..900.mb.           0.000197 ***
Wind.Gust.daily.mean..sfc.                   0.629095
Temperature.daily.max..2.m.above.gnd.        0.904920
Temperature.daily.min..2.m.above.gnd.        0.131368
Relative.Humidity.daily.max..2.m.above.gnd.  0.997398
Relative.Humidity.daily.min..2.m.above.gnd.  0.711278
Mean.Sea.Level.Pressure.daily.max..MSL.      0.000564 ***
Mean.Sea.Level.Pressure.daily.min..MSL.      2.29e-05 ***
Total.Cloud.Cover.daily.max..sfc.            0.483090
Total.Cloud.Cover.daily.min..sfc.            0.213759
```

```
High.Cloud.Cover.daily.max..high.cld.lay.    0.235609
High.Cloud.Cover.daily.min..high.cld.lay.    0.768986
Medium.Cloud.Cover.daily.max..mid.cld.lay.   0.051598 .
Medium.Cloud.Cover.daily.min..mid.cld.lay.   0.575746
Low.Cloud.Cover.daily.max..low.cld.lay.      0.386126
Low.Cloud.Cover.daily.min..low.cld.lay.      0.986395
Wind.Speed.daily.max..10.m.above.gnd.        0.105153
Wind.Speed.daily.min..10.m.above.gnd.        0.008415 **
Wind.Speed.daily.max..80.m.above.gnd.        0.890059
Wind.Speed.daily.min..80.m.above.gnd.        0.208741
Wind.Speed.daily.max..900.mb.                0.268904
Wind.Speed.daily.min..900.mb.                0.832168
Wind.Gust.daily.max..sfc.                    0.187260
Wind.Gust.daily.min..sfc.                    0.855427
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1635.4  on 1179  degrees of freedom
Residual deviance: 1232.7  on 1136  degrees of freedom
AIC: 1320.7

Number of Fisher Scoring iterations: 5
```

Suite à cette première regression, notre modèle présente 7 variables significative à savoir :

- L'année d'observation
- Le niveau de Pression moyen de la mer
- La direction moyenne du vent où la pression vaut 900 hPa
- Le Max du niveau de Pression moyen de la mer
- Le min du niveau de Pression moyen de la mer
- Le Max du niveau de couverture nuageuse moyenne
- Le min du niveau de Vitesse quotidienne moyenne à 10M d'altitude.

Qualité d'ajustement du modèle complet

```
Generalized Linear Model

1180 samples
  45 predictor
   2 classes: 'FALSE', 'TRUE'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 944, 945, 943, 944, 944
Resampling results:
```

```
    Accuracy   Kappa
    0.7313436  0.4621462
```

Accuracy	Kappa
0.7313436	0.4621462

Le résultat obtenu est un résumé des performances de notre modèle de régression linéaire généralisée à partir de la validation croisée à 5 plis.

- Il y a 1180 observations (échantillons) dans votre ensemble de données d'entraînement.
- 46 variables explicatives (prédicteurs) dans votre modèle.
- 2 classes: $FALSE$, $TRUE$: La variable de sortie (pluie.demain) présente deux classes, "FALSE" (faux) et "TRUE" (vrai).

En ce qui concerne la section "Resampling results", les mesures de performance obtenues pour la validation croisée sont les suivantes :

- L'exactitude est la proportion de prédictions correctes par rapport à l'ensemble des prédictions. Dans notre cas, l'exactitude moyenne de votre modèle est de 0.7313436, soit environ **73%**.
- Le kappa est une mesure de concordance qui tient compte de l'exactitude due au hasard. Une valeur de kappa de 1 indique une concordance parfaite entre les prédictions et les vraies valeurs, tandis qu'une valeur de 0 indique une concordance due au hasard. Dans votre cas, la valeur de kappa moyenne de votre modèle est de **0.4621462**.

Ces mesures nous donnent une indication de la performance de notre modèle de régression linéaire généralisée lors de la validation croisée à 5 plis. Cependant, il est important de noter que ces résultats sont spécifiques à nos données d'entraînement et ne garantissent pas la performance sur de nouvelles données réelles.

Deuxième modèle

modele2 contenant uniquement les variables explicatives significatives

```
   Call:
   glm(formula = pluie.demain ~ Year +
Mean.Sea.Level.Pressure.daily.mean..MSL. +
       Wind.Direction.daily.mean..900.mb. +
Mean.Sea.Level.Pressure.daily.max..MSL. +
       Mean.Sea.Level.Pressure.daily.min..MSL. +
Medium.Cloud.Cover.daily.max..mid.cld.lay. +
       Wind.Speed.daily.min..10.m.above.gnd., family = binomial,
       data = train)

   Deviance Residuals:
       Min       1Q   Median       3Q      Max
```

```
   -2.2051  -0.8973   0.2789   0.9174   2.5520

   Coefficients:
                                                Estimate Std. Error z value
   (Intercept)                                -1.305e+01  5.784e+01  -0.226
   Year                                        5.069e-02  2.858e-02   1.774
   Mean.Sea.Level.Pressure.daily.mean..MSL.    4.673e-01  1.259e-01   3.712
   Wind.Direction.daily.mean..900.mb.          3.644e-03  9.701e-04   3.757
   Mean.Sea.Level.Pressure.daily.max..MSL.    -2.580e-01  6.602e-02  -3.908
   Mean.Sea.Level.Pressure.daily.min..MSL.    -2.988e-01  6.898e-02  -4.331
   Medium.Cloud.Cover.daily.max..mid.cld.lay.  1.380e-02  1.834e-03   7.527
   Wind.Speed.daily.min..10.m.above.gnd.       2.972e-02  2.086e-02   1.425
                                              Pr(>|z|)
   (Intercept)                                0.821529
   Year                                       0.076089 .
   Mean.Sea.Level.Pressure.daily.mean..MSL.   0.000206 ***
   Wind.Direction.daily.mean..900.mb.         0.000172 ***
   Mean.Sea.Level.Pressure.daily.max..MSL.    9.33e-05 ***
   Mean.Sea.Level.Pressure.daily.min..MSL.    1.48e-05 ***
   Medium.Cloud.Cover.daily.max..mid.cld.lay. 5.21e-14 ***
   Wind.Speed.daily.min..10.m.above.gnd.      0.154124
   ---
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   (Dispersion parameter for binomial family taken to be 1)

       Null deviance: 1635.4  on 1179  degrees of freedom
   Residual deviance: 1317.0  on 1172  degrees of freedom
   AIC: 1333

   Number of Fisher Scoring iterations: 4
```

Notre deuxième modèle nous présente les 6 variables significatives. Poursuivons les tests en gardant que ces variables.

Troisième modèle

modele3 contenant uniquement les 5 variables explicatives significatives

```
   Call:
   glm(formula = pluie.demain ~ Year +
Mean.Sea.Level.Pressure.daily.mean..MSL. +
       Wind.Direction.daily.mean..900.mb. +
Mean.Sea.Level.Pressure.daily.max..MSL. +
       Mean.Sea.Level.Pressure.daily.min..MSL. +
Medium.Cloud.Cover.daily.max..mid.cld.lay.,
       family = binomial, data = train)

   Deviance Residuals:
```

```
      Min       1Q   Median       3Q      Max
-2.1911  -0.9111   0.3211   0.9058   2.5541

Coefficients:
                                             Estimate Std. Error z value
(Intercept)                                -2.017e+01  5.761e+01  -0.350
Year                                        5.452e-02  2.845e-02   1.916
Mean.Sea.Level.Pressure.daily.mean..MSL.    4.440e-01  1.239e-01   3.584
Wind.Direction.daily.mean..900.mb.          3.867e-03  9.582e-04   4.035
Mean.Sea.Level.Pressure.daily.max..MSL.    -2.380e-01  6.378e-02  -3.731
Mean.Sea.Level.Pressure.daily.min..MSL.    -2.960e-01  6.867e-02  -4.311
Medium.Cloud.Cover.daily.max..mid.cld.lay.  1.395e-02  1.829e-03   7.627
                                           Pr(>|z|)
(Intercept)                                0.726211
Year                                       0.055364 .
Mean.Sea.Level.Pressure.daily.mean..MSL.   0.000338 ***
Wind.Direction.daily.mean..900.mb.         5.45e-05 ***
Mean.Sea.Level.Pressure.daily.max..MSL.    0.000191 ***
Mean.Sea.Level.Pressure.daily.min..MSL.    1.63e-05 ***
Medium.Cloud.Cover.daily.max..mid.cld.lay. 2.41e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1635.4  on 1179  degrees of freedom
Residual deviance: 1319.1  on 1173  degrees of freedom
AIC: 1333.1

Number of Fisher Scoring iterations: 4
```

Notre variable Année devient plus de plus en plus significative. Denière tentative avec un modèle sans l'année.

Quatrième modèle

modele4 contenant uniquement les 5 variables explicatives significatives

```
Call:
glm(formula = pluie.demain ~ Mean.Sea.Level.Pressure.daily.mean..MSL. +
    Wind.Direction.daily.mean..900.mb. +
Mean.Sea.Level.Pressure.daily.max..MSL. +
    Mean.Sea.Level.Pressure.daily.min..MSL. +
Medium.Cloud.Cover.daily.max..mid.cld.lay.,
    family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2412  -0.9073   0.3284   0.9060   2.5751
```

```
Coefficients:
                                          Estimate Std. Error z value
(Intercept)                             88.6214911 10.6206119   8.344
Mean.Sea.Level.Pressure.daily.mean..MSL.    0.4458289  0.1242212   3.589
Wind.Direction.daily.mean..900.mb.          0.0038555  0.0009575   4.027
Mean.Sea.Level.Pressure.daily.max..MSL.    -0.2356985  0.0638803  -3.690
Mean.Sea.Level.Pressure.daily.min..MSL.    -0.2991943  0.0688327  -4.347
Medium.Cloud.Cover.daily.max..mid.cld.lay.  0.0140413  0.0018249   7.694
                                          Pr(>|z|)
(Intercept)                                < 2e-16 ***
Mean.Sea.Level.Pressure.daily.mean..MSL.   0.000332 ***
Wind.Direction.daily.mean..900.mb.         5.66e-05 ***
Mean.Sea.Level.Pressure.daily.max..MSL.    0.000225 ***
Mean.Sea.Level.Pressure.daily.min..MSL.    1.38e-05 ***
Medium.Cloud.Cover.daily.max..mid.cld.lay. 1.42e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1635.4  on 1179  degrees of freedom
Residual deviance: 1322.8  on 1174  degrees of freedom
AIC: 1334.8

Number of Fisher Scoring iterations: 4
```

Qualité d'ajustement du modèle

```
Generalized Linear Model

1180 samples
   5 predictor
   2 classes: 'FALSE', 'TRUE'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 943, 944, 945, 944, 944
Resampling results:

  Accuracy   Kappa
  0.7271382  0.4528554
```

les mesures de performance obtenues pour la validation croisée sont les suivantes :

- L'exactitude est la proportion de prédictions correctes par rapport à l'ensemble des prédictions. Dans notre cas, l'exactitude moyenne de votre modèle est de 0.7271382, soit environ ** 73% **.
- Le kappa est une mesure de concordance qui tient compte de l'exactitude due au hasard. Une valeur de kappa de 1 indique une concordance parfaite entre les

prédictions et les vraies valeurs, tandis qu'une valeur de 0 indique une concordance due au hasard. Dans votre cas, la valeur de kappa moyenne de votre modèle est de ** 0.45**.

Ces mesures nous donnent une indication de la performance de notre modèle de régression linéaire généralisée lors de la validation croisée à 5 plis. Cependant, il est important de noter que ces résultats sont spécifiques à nos données d'entraînement et ne garantissent pas la performance sur de nouvelles données réelles.

Conclusion 1

Comparaison des AIC en fonction de nos différent modèles

- Pour le modèle complet, M1, nous avons un AIC = AIC: 1320.7
- Pour le modèle 2 , AIC = AIC: 1333
- Pour le modèle 3 , AIC = AIC: 1333.1
- Pour le modèle 4 , AIC = AIC: 1334.8

Nous remarquons un gain d'AIC lorsque notre modèle devient de plus en plus réduit.

A ce stade, notre choix se tourne vers le modèle 4.

Dernière comparaison en prenant en compte la regression réalisé par R et l'analyse de son meilleur modèle au sens du critère AIC

Regression selon R

```
   Call:
   glm(formula = pluie.demain ~ Year + Temperature.daily.mean..2.m.above.gnd.
+
       Mean.Sea.Level.Pressure.daily.mean..MSL. +
Snowfall.amount.raw.daily.sum..sfc. +
       Medium.Cloud.Cover.daily.mean..mid.cld.lay. +
Wind.Speed.daily.mean..80.m.above.gnd. +
       Wind.Direction.daily.mean..80.m.above.gnd. +
Wind.Direction.daily.mean..900.mb. +
       Temperature.daily.min..2.m.above.gnd. +
Mean.Sea.Level.Pressure.daily.max..MSL. +
       Mean.Sea.Level.Pressure.daily.min..MSL. +
Total.Cloud.Cover.daily.max..sfc. +
       Total.Cloud.Cover.daily.min..sfc. +
Medium.Cloud.Cover.daily.max..mid.cld.lay. +
       Wind.Speed.daily.max..10.m.above.gnd. +
Wind.Speed.daily.min..10.m.above.gnd. +
       Wind.Gust.daily.max..sfc., family = binomial, data = train)

   Deviance Residuals:
       Min       1Q   Median       3Q      Max
   -2.5754  -0.8336   0.2694   0.8517   2.8827
```

```
Coefficients:
                                          Estimate Std. Error z value
(Intercept)                             -68.942905  62.791813  -1.098
Year                                      0.065852   0.030853   2.134
Temperature.daily.mean..2.m.above.gnd.    0.146739   0.049863   2.943
Mean.Sea.Level.Pressure.daily.mean..MSL.  0.481900   0.131205   3.673
Snowfall.amount.raw.daily.sum..sfc.      -0.316432   0.215283  -1.470
Medium.Cloud.Cover.daily.mean..mid.cld.lay.  0.010810   0.004061   2.662
Wind.Speed.daily.mean..80.m.above.gnd.   -0.114538   0.029949  -3.824
Wind.Direction.daily.mean..80.m.above.gnd.  -0.002686   0.001524  -1.763
Wind.Direction.daily.mean..900.mb.        0.004585   0.001289   3.557
Temperature.daily.min..2.m.above.gnd.    -0.102830   0.054202  -1.897
Mean.Sea.Level.Pressure.daily.max..MSL.  -0.242121   0.070640  -3.428
Mean.Sea.Level.Pressure.daily.min..MSL.  -0.306000   0.071569  -4.276
Total.Cloud.Cover.daily.max..sfc.         0.008353   0.003504   2.383
Total.Cloud.Cover.daily.min..sfc.         0.007844   0.003863   2.031
Medium.Cloud.Cover.daily.max..mid.cld.lay.  0.006226   0.002671   2.331
Wind.Speed.daily.max..10.m.above.gnd.     0.059655   0.022770   2.620
Wind.Speed.daily.min..10.m.above.gnd.     0.111262   0.036110   3.081
Wind.Gust.daily.max..sfc.                 0.023669   0.010856   2.180
                                         Pr(>|z|)
(Intercept)                              0.272222
Year                                     0.032811 *
Temperature.daily.mean..2.m.above.gnd.   0.003252 **
Mean.Sea.Level.Pressure.daily.mean..MSL. 0.000240 ***
Snowfall.amount.raw.daily.sum..sfc.      0.141605
Medium.Cloud.Cover.daily.mean..mid.cld.lay. 0.007776 **
Wind.Speed.daily.mean..80.m.above.gnd.   0.000131 ***
Wind.Direction.daily.mean..80.m.above.gnd.  0.077965 .
Wind.Direction.daily.mean..900.mb.       0.000375 ***
Temperature.daily.min..2.m.above.gnd.    0.057806 .
Mean.Sea.Level.Pressure.daily.max..MSL.  0.000609 ***
Mean.Sea.Level.Pressure.daily.min..MSL.  1.91e-05 ***
Total.Cloud.Cover.daily.max..sfc.        0.017151 *
Total.Cloud.Cover.daily.min..sfc.        0.042272 *
Medium.Cloud.Cover.daily.max..mid.cld.lay.  0.019757 *
Wind.Speed.daily.max..10.m.above.gnd.    0.008795 **
Wind.Speed.daily.min..10.m.above.gnd.    0.002062 **
Wind.Gust.daily.max..sfc.                0.029238 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1635.4  on 1179  degrees of freedom
Residual deviance: 1246.8  on 1162  degrees of freedom
AIC: 1282.8

Number of Fisher Scoring iterations: 4
```

Le modèle qu'il propose au final Call:

glm(formula = pluie.demain ~ Year + Temperature.daily.mean..2.m.above.gnd. + Mean.Sea.Level.Pressure.daily.mean..MSL. + Snowfall.amount.raw.daily.sum..sfc. + Medium.Cloud.Cover.daily.mean..mid.cld.lay. + Wind.Speed.daily.mean..80.m.above.gnd. + Wind.Direction.daily.mean..80.m.above.gnd. + Wind.Direction.daily.mean..900.mb. + Temperature.daily.min..2.m.above.gnd. + Mean.Sea.Level.Pressure.daily.max..MSL. + Mean.Sea.Level.Pressure.daily.min..MSL. + Total.Cloud.Cover.daily.max..sfc. + Total.Cloud.Cover.daily.min..sfc. + Medium.Cloud.Cover.daily.max..mid.cld.lay. + Wind.Speed.daily.max..10.m.above.gnd. + Wind.Speed.daily.min..10.m.above.gnd. + Wind.Gust.daily.max..sfc., family = binomial, data = train)

retient plus de variables que le notre. Nous le gardons en mémoire pour la suite des tests.

Qualité d'ajustement du modèle complet

```
Generalized Linear Model

1180 samples
  17 predictor
   2 classes: 'FALSE', 'TRUE'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 943, 944, 944, 945, 944
Resampling results:

  Accuracy   Kappa
  0.7339179  0.4672321
```

les mesures de performance obtenues pour la validation croisée sont les suivantes :

- L'exactitude est la proportion de prédictions correctes par rapport à l'ensemble des prédictions. Dans notre cas, l'exactitude moyenne de votre modèle est de 0.7339179, soit environ **73 % **.
- Le kappa est une mesure de concordance qui tient compte de l'exactitude due au hasard. Une valeur de kappa de 1 indique une concordance parfaite entre les prédictions et les vraies valeurs, tandis qu'une valeur de 0 indique une concordance due au hasard. Dans votre cas, la valeur de kappa moyenne de votre modèle est de ** 0.46**.

Ces mesures nous donnent une indication de la performance de notre modèle de régression linéaire généralisée lors de la validation croisée à 5 plis. Cependant, il est important de noter que ces résultats sont spécifiques à nos données d'entraînement et ne garantissent pas la performance sur de nouvelles données réelles.

Anova

```r
#ANOVA
anova(modele4, modeleR, test = "LRT")
```

```
   Analysis of Deviance Table

   Model 1: pluie.demain ~ Mean.Sea.Level.Pressure.daily.mean..MSL. +
Wind.Direction.daily.mean..900.mb. +
       Mean.Sea.Level.Pressure.daily.max..MSL. +
Mean.Sea.Level.Pressure.daily.min..MSL. +
       Medium.Cloud.Cover.daily.max..mid.cld.lay.
   Model 2: pluie.demain ~ Year + Temperature.daily.mean..2.m.above.gnd. +
       Mean.Sea.Level.Pressure.daily.mean..MSL. +
Snowfall.amount.raw.daily.sum..sfc. +
       Medium.Cloud.Cover.daily.mean..mid.cld.lay. +
Wind.Speed.daily.mean..80.m.above.gnd. +
       Wind.Direction.daily.mean..80.m.above.gnd. +
Wind.Direction.daily.mean..900.mb. +
       Temperature.daily.min..2.m.above.gnd. +
Mean.Sea.Level.Pressure.daily.max..MSL. +
       Mean.Sea.Level.Pressure.daily.min..MSL. +
Total.Cloud.Cover.daily.max..sfc. +
       Total.Cloud.Cover.daily.min..sfc. +
Medium.Cloud.Cover.daily.max..mid.cld.lay. +
       Wind.Speed.daily.max..10.m.above.gnd. +
Wind.Speed.daily.min..10.m.above.gnd. +
       Wind.Gust.daily.max..sfc.
     Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
   1      1174     1322.8
   2      1162     1246.8 12    75.96 2.418e-11 ***
   ---
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le modèleR nous propose une meilleur déviance que celui obtenue par notre modele4, c'est à dire un écart entre les valeurs observées y_i et $n_i - y_i$ et les valeurs estimées $\hat{\mu}i$ et $n_i - \hat{\mu}i$ où où y_i est la valeur observée et $\hat{\mu}i$ la valeur prédite pour l'observation i.

Conclusion 2

Analysons l'algorithmique faite par R: Dans le modèle de base, AIC = 1334.8 Il rajoute des variables explicatives en plus de nos 4 variables clés et calcule l'AIC par ordre croissant des variables introduites: AIC= 1282.8; c'est le meilleur modèle selon R au sens de l'AIC mais pas forcément meilleur au notre. Au vu des résultats obtenues suites aux comparaisons des AIC, de la qualité de l'ajustement des différents modèles, notre choix de modèle est porté sur le modèle4 obtenu par nos soins. Nous réaliserons néanmoins les prédictions sur quelques modèles clés en plus et cela à titre comparatif.

Prédiction sur la base Test

Initialisation de notre base test

```
test = test[-1]
```

Notre base test contient 290 données et 46 covariables.

```
print(head(test))
```

```
   # A tibble: 6 × 45
      Year Month   Day  Hour Minute Tempe…¹ Relat…² Mean.…³ Total…⁴ Snowf…⁵
Total…⁶
     <dbl> <dbl> <dbl> <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
<dbl>
   1  2010     7     6     0      0    19.4    73.3   1022.     0.1       0
39.6
   2  2010     7    14     0      0    25.7    64.9   1008.     6.1       0
7.08
   3  2010     7    24     0      0    16.4    74.3   1021      2.3       0
55.1
   4  2010     7    28     0      0    18.0    76.1   1018.     4.4       0
57.0
   5  2010     8    13     0      0    17.9    73.9   1015.     0         0
75.7
   6  2010     8    25     0      0    18.0    68.4   1018.     0         0
31.7
   # … with 34 more variables: High.Cloud.Cover.daily.mean..high.cld.lay.
<dbl>,
   #   Medium.Cloud.Cover.daily.mean..mid.cld.lay. <dbl>,
   #   Low.Cloud.Cover.daily.mean..low.cld.lay. <dbl>,
   #   Sunshine.Duration.daily.sum..sfc. <dbl>,
   #   Shortwave.Radiation.daily.sum..sfc. <dbl>,
   #   Wind.Speed.daily.mean..10.m.above.gnd. <dbl>,
   #   Wind.Direction.daily.mean..10.m.above.gnd. <dbl>, …
```

Test de prédictions sur nos différents modèles

Modèle M1

$$modele1 = glm(pluie.demain\ ., data = train, family = binomial)$$

```
          1         2         3         4         5         6
  0.2273793 0.8074403 0.6181934 0.6887581 0.5374641 0.4494971

     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.01949 0.29266 0.55917 0.53009 0.75497 0.98674
```

Seuil de prédictions

α = 0.5 selon un ratio obtenue via les résultats de la base train $\frac{Vrai}{Vrai+Faux}$)

```
test_predictions1 = ifelse(pred1 >= 0.5, "Vrai", "Faux")

# Ajout des prédictions à la colonne "test" de l'ensemble de données de test
test$predictions1 = test_predictions1
head(test_predictions1)
```

```
     1      2      3      4      5      6
"Faux" "Vrai" "Vrai" "Vrai" "Vrai" "Faux"
```

```
test_predictions1
Faux Vrai
 126  164
```

ratio de prédiction “Vrai” et “Faux”

Matrice de confusion

```
test_predictions1 Faux Vrai
             Faux  126    0
             Vrai    0  164
```

Ne disposant pas de Vrai données sur notre base de test, la matrice de confusion de nous apporte pas d’informations supplémentaires à exploiter.

Calcul de l’exactitude

```
[1] "Exactitude: 1"
```

Modèle M4

$$modele4 < -glm(pluie.demain\ Mean.Sea.Level.Pressure.daily.mean..MSL.+Wind.Direction.daily.mean..9 = train, family = binomial)$$

Seuil de prédictions 1

```
test_predictions4 = ifelse(pred4 >= 0.5, "Vrai", "Faux")

# Ajout des prédictions à la colonne "test" de l'ensemble de données de test
test$predictions4 = test_predictions4
head(test_predictions4)
```

```
     1      2      3      4      5      6
"Faux" "Vrai" "Vrai" "Vrai" "Vrai" "Faux"
```

```
table(test_predictions4)
```

```
test_predictions4
Faux Vrai
 115  175
```

Seuil de prédictions 2

```
      1      2      3      4      5      6
 "Faux" "Vrai" "Vrai" "Vrai" "Faux" "Faux"

table(test_predictions4_2)

   test_predictions4_2
   Faux Vrai
    154  136
```

En faisant varier la valeur du seuil, nos prédictions de Vrai et Faux à la question de savoir s'il va pleuvoir demain ou pas j'ajuste considérablement.

Seuil de prédictions 3

```
      1      2      3      4      5      6
 "Vrai" "Vrai" "Vrai" "Vrai" "Vrai" "Faux"

table(test_predictions4_3)

   test_predictions4_3
   Faux Vrai
     95  195
```

En faisant varier la valeur du seuil, nos prédictions de Vrai et Faux à la question de savoir s'il va pleuvoir demain ou pas j'ajuste considérablement.

Calcul de l'exactitude

```
   [1] "Exactitude: 1"
```

Modèle MR

$$modeleR = stepAIC(modele1, data = train, family = binomial)$$

Seuil de prédictions

α = 0.5 selon un ratio obtenue via les résultats de la base train $\frac{Vrai}{Vrai+Faux}$)

```
      1      2      3      4      5      6
 "Faux" "Vrai" "Vrai" "Vrai" "Vrai" "Faux"

   test_predictionsR
   Faux Vrai
    125  165
```

Seuil de prédictions 2

```
      1      2      3      4      5      6
 "Faux" "Vrai" "Faux" "Vrai" "Faux" "Faux"

table(test_predictionsR_2)
```

```
   test_predictionsR_2
   Faux Vrai
    159  131
```

En faisant varier la valeur du seuil, nos prédictions de Vrai et Faux à la question de savoir s'il va pleuvoir demain ou pas j'ajuste considérablement.

Seuil de prédictions 3

```
        1      2      3      4      5      6
   "Vrai" "Vrai" "Vrai" "Vrai" "Vrai" "Faux"

table(test_predictionsR_3)

   test_predictionsR_3
   Faux Vrai
     95  195
```

En faisant varier la valeur du seuil, nos prédictions de Vrai et Faux à la question de savoir s'il va pleuvoir demain ou pas j'ajuste considérablement.

Calcul de l'exactitude

```
   [1] "Exactitude: 1"
```

Courbe de ROC

Coder la cible en 0 et 1

```
y = ifelse(test$predictions4 == "Vrai", 1, 0)
print(table(test$predictions4,y))
```

```
        y
           0   1
    Faux 115   0
    Vrai   0 175
```

On a : 1 si le résultat est Vrai et 0 si le résultat est Faux.

Nbre de positif Nbre de négatif

```
  [1] 175

  [1] 115
```

Création d'un dataframe avec y et les scores de prédictions

```
    y     pred4
  1 0 0.4994927
  2 1 0.7036494
  3 1 0.6639648
  4 1 0.6202121
  5 1 0.5854267
  6 0 0.3259313
```

Trier la dataframe avec les scores décroissants

```
      y     pred4
  161 1 0.9609882
  19  1 0.9558429
  206 1 0.9186428
  117 1 0.9127029
  278 1 0.9014303
  20  1 0.9010887
```

Ici les premières lignes de la dataframe

```
      y      pred4
  236 0 0.06039136
  240 0 0.05935753
  238 0 0.04627615
  57  0 0.04369653
  239 0 0.04313788
  163 0 0.03403608
```

Ici les dernières lignes de la dataframe crrée.

Taux de faux positifs

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

Taux de vrais positifs

```
   [1] 0.005714286 0.011428571 0.017142857 0.022857143 0.028571429
0.034285714

library(ROCR)
print(pred_roc)

   A prediction instance
     with 290 data points
```

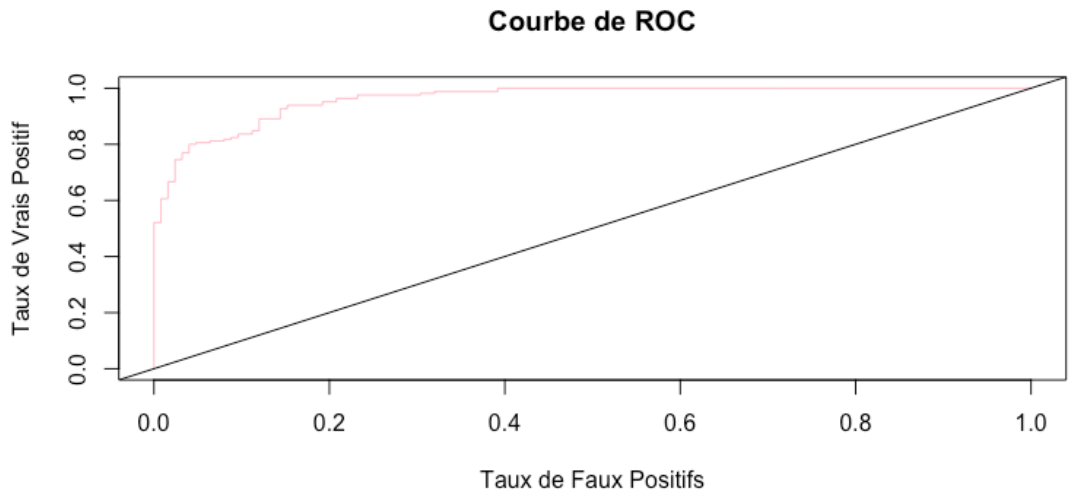
Mesure de la performance de l'objet

```
grph_roc = ROCR::performance(pred_roc, measure = "tpr", x.measure ="fpr")
print(grph_roc)

   A performance instance
     'False positive rate' vs. 'True positive rate' (alpha: 'Cutoff')
     with 291 data points
```

Graphique Courbe ROC

```
ROCR::plot(grph_roc, xlab = "Taux de Faux Positifs", ylab = "Taux de Vrais
Positif", col = "pink", main = "Courbe de ROC")
abline(a=0, b=1)
```



Notre courbe de ROC est assez satisfaisante, car elle nous montre le taux de vraix positifs en fonction du taux de faux positifs. Notre courbe est considérablement éloignée de la diagonale et à une allure correcte. Nous ne pouvons réaliser et mésurer la qualité de prédiction sur les données tests car nous ne disposons pas des Vrais données et celles obtenues dans la base de test sont celles prédites grâce à notre modèle M4.