Rapport

Dieynaba Dieng

2023-07-11

```
tinytex:::is_tinytex()
```

```
## [1] TRUE
```

Présentation des variables

Nous commençons par présenter nos données pour une meilleure compréhension. Le jeu de données meteo.train comporte 47 variables avec 1180 observations fournit selon les informations suivantes constituant les conditions météorologiques et caractérisant les pluies:

Temperature.daily.mean..2.m.above.gnd. = Température moyenne quotidienne 2.m au-dessus de la terre

Relative.Humidity.daily.mean..2.m.above.gnd. = Humidité relative moyenne quotidienne 2.m au-dessus de la terre.

Mean.Sea.Level.Pressure.daily.mean..MSL. = Pression au niveau moyen de la mer moyenne quotidienne MSL

Total.Precipitation.daily.sum..sfc. = Somme totale des précipitations quotidiennes.sfc.

Snowfall.amount.raw.daily.sum..sfc. = somme quotienne de Chute.de.neige.montant.brut.sfc.

Total.Cloud.Cover.daily.mean..sfc. = Couverture totale des nuages moyenne quotienne .sfc.

High.Cloud.Cover.daily.mean..high.cld.lay. = Une couverture nuageuse élevée moyenne quotienne de.cld.élevée . Medium.Cloud.Cover.daily.mean..mid.cld.lay. = Moyenne quotidienne de la couverture nuageuse moyenne..mid.cld.lay.

Low.Cloud.Cover.daily.mean..low.cld.lay. = Faible Couverture nuageuse quotidienne moyenne..faible.clay

Sunshine.Duration.daily.sum..sfc. = somme de la durée quotidienne du Soleil.sfc.

Shortwave.Radiation.daily.sum..sfc. =somme quotidienne de Rayonnement à ondes courtes.sfc.

Wind.Speed.daily.mean..10.m.above.gnd. = Vitesse du vent moyenne quotidienne 10.m au-dessus de la terre.

Wind.Direction.daily.mean..10.m.above.gnd. = Direction du vent moyenne quotidienne 10.m au-dessus de la terre.

Wind.Speed.daily.mean..80.m.above.gnd. = Vitesse du vent moyenne quotidienne 80.m au-dessus de la terre.

Wind.Direction.daily.mean..80.m.above.gnd. = Direction du vent moyenne quotidienne 80.m au-dessus de la terre

Wind.Speed.daily.mean..900.mb. = Vitesse du vent moyenne quotidienne 900.mb.

Wind.Direction.daily.mean..900.mb.= Direction du vent moyenne quotidienne 900.mb.

Wind.Gust.daily.mean..sfc. = Moyenne quotidienne de la rafale d'air.sfcVent.Gust.daily.mean..sfc.

Temperature.daily.max..2.m.above.gnd. = Température quotidienne max 2.m au-dessus de la masse.

Temperature.daily.min..2.m.above.gnd. = Température quotidienne min 2.m au-dessus de la terre.

Relative.Humidity.daily.max..2.m.above.gnd. = Humidité Relative quotidienne max 2.m.au-dessus de la terre.

Relative.Humidity.daily.min..2.m.above.gnd. = Humidité.relative.quotidienne.min..2.m.au-dessus.de.terre.

Mean.Sea.Level.Pressure.daily.max..MSL.= Pression au Niveau moyen de la mer quotidienne max..MSL.

Mean.Sea.Level.Pressure.daily.min..MSL. = Pression au niveau moyen de la mer quotidienne min.MSL.

Total.Cloud.Cover.daily.max..sfc. = Couverture nuageuse totale quotidienne max.sfc.

Total.Cloud.Cover.daily.min..sfc. = Couverture nuageuse totale quotidienne min.sfc.

High.Cloud.Cover.daily.max..high.cld.lay.= Couverture nuageuse élevée quotidienne max.haute.cld.lay

High.Cloud.Cover.daily.min..high.cld.lay.= Couverture nuageuse élevée quotidienne min haute.cld.lay

Medium.Cloud.Cover.daily.max..mid.cld.lay. = Nébulosité moyenne quotidienne max mid.cld.lay..

Medium.Cloud.Cover.daily.min..mid.cld.lay. = Couverture nuageuse moyenne quotidienne min..mid.cld.lay.

Low.Cloud.Cover.daily.max..low.cld.lay. = Couverture nuageuse basse quotidienne. max.

Low.Cloud.Cover.daily.min..low.cld.lay. = Couverture nuageuse basse quotidienne min.low.cld.lay.

Wind.Speed.daily.max..10.m.above.gnd. = Vitesse du vent.quotidienne max 10.m.au-dessus de la terre.

Wind.Speed.daily.min..10.m.above.gnd. = Vitesse du vent quotidienne min 10.m.au-dessusde la terre.

Wind.Speed.daily.max..80.m.above.gnd. = Vitesse du vent quotidienne max 80.m au-dessus de la terre.

Wind.Speed.daily.min..80.m.above.gnd. = Vitesse du vent quotidienne min. 80.m au-dessus de la terre.

Wind.Speed.daily.max..900.mb. = Vitesse du vent quotidienne max 900.mb.

Wind.Speed.daily.min..900.mb. = Vitesse du vent quotidienne min 900.mb.

Wind.Gust.daily.max..sfc. = Rafale de vent quotidienne max.sfc.

Wind.Gust.daily.min..sfc. = Rafale de vent quotidienne min.sfc.

Notre variable d'intérêt est la variable pluie.demain : c'est elle que nous chercherons à prédire pour le lendemain en construisant des modèles.

Exploration des variables

L'analyse la plus évidente à réaliser sur nos données est la détection d'éventuelles dépendances entre les variables exogènes. Après vérification, on trouve que la base de données n'a pas de valeurs manquantes. On retrouve une multicolinearité entre certaines variables. Le graphique généré nous montre que certaines variables sont distribuées de la même façon. Parmi celles-ci nous avons Hour et minute,

Low.Cloud.Cover.daily.min..low.cld.lay., Medium.Cloud.Cover.daily.min..mid.cld.lay., High.Cloud.Cover.daily.min..high.cld.la et Total.Cloud.Cover.daily.min..sfc.

Wind.Speed.daily.min..900.mb., Wind.Gust.daily.min..sfc., Wind.Speed.daily.min..80.m.above.gnd. et Wind.Speed.daily.min..10.m.above.gnd.

De plus notre tableau de correlation nous montre que l'année est très fortement corrélée avec X, Low.Cloud.Cover.daily.mean..low.cld.lay. est corrélée Total.Cloud.Cover.daily.mean..sfc. La variable Medium.Cloud.Cover.daily.mean..mid.cld.lay. est aussi très corrélée avec Total.Cloud.Cover.daily.mean..sfc

Shortwave.Radiation.daily.sum..sfc., corrélée avec Sunshine.Duration.daily.sum..sfc. une agmentation de la somme quotidienne des Rayonnement à ondes courtes entraine une elevation de la somme de la durée quotidienne du soleil.

Wind.Speed.daily.mean..10.m.above.gnd., corrélée avec Wind.Speed.daily.mean..80.m.above.gnd. et Wind.Gust.daily.mean..sfc. Une augmentation de la vitesse moyenne quotidienne du vent à 10m au dessus

de la terre entraine une augmentation de la vitesse moyenne quotidienne du vent à 80m au dessus de la terre.

Wind.Speed.daily.mean..80.m.above.gnd. est corrélée avec Wind.Gust.daily.mean..sfc. Une augmentation de la vitesse moyenne quotidienne du vent à 80 m au dessus de la terre entraine une augmentation de la rafale de vent moyenne quotidienne.

Wind.Direction.daily.mean..80.m.above.gnd. est corrélée avec Wind.Direction.daily.mean..10.m.above.gnd.

Wind.speed.daily.mean..900.mb. avec Wind.Gust.daily.mean..sfc.

Mean.Sea.Level.Pressure.daily.max..MSL. est très fortement corrélée avec Mean.Sea.Level.Pressure.daily.min..MSL.

Wind.Speed.daily.max..10.m.above.gnd. est corrélée avec les variables, Wind.Speed.daily.min..10.m.above.gnd., Wind.Speed.daily.max..900.mb., Wind.Speed.daily.max..80.m.above.gnd., Wind.Speed.daily.min..80.m.above.gnd., Wind.Speed.daily.min..900.mb., Wind.Gust.daily.max..sfc., Wind.Gust.daily.min..sfc.

Ces variables sont corrélées positivement entre eux. une augmentation de l'une entraine également une agmentation de l'autre. On a une multicolinearité entre les variables indépendantes; ce qui posera problème sur la regression. On voit dans certains cas que la corrélation entre les variables indépendantes et la variable cible est faible ou nulle. Ce qui peut conduire au modèle de régression à avoir du mal à capturer une relation significative entre ces variables. D'où certaines performances de prédiction médiocres.

Pour une meilleure construction du modèle il serait nécessaire de faire une selection entre les variables. Ce qui permettrait de supprimer certaines qui sont fortement corrrélées au risque de ne pas biaiser la modélisation de conséquentes erreurs.

La variable à prédire à bien 2 modalités : TRUE et FALSE dont 601 TRUE et 579 FALSE.

Constructions des modèles

La variable à prédire ("pluie.demain") est une variable catégorielle donc la regression lineaire ne peut être appliquer. Cette dernière ne s'applique que si la variable est continue. Nous procéderons à la regression logistique.

Modèle complet

La variable d'intérêt, pluie.demain est de type binaire et prend deux modalités. Les deux modèles que nous allons priorisés sont logit et probit qui se différencient par leur fonction de lien : logistique pour la première et normale pour la seconde.

Le modèle logistique s'écrit comme suit:

$\ln(\frac{pi}{1-pi}) = \beta 0 + \beta 1 xi$ avec Yi $\sim$ Bernoulli(pi)

Le modèle probit s'écrit comme suit:

$\pi = \phi(\beta 0 + \beta 1 xi)$

En premier lieu, nous allons faire une regression logistique en prenant en compte toutes variables constituant le modèle complet. Ce qui montre qu'il ya un effet significative de certaines variables avec la pluie. La pression au niveau moyen de la mer moyenne quotidienne .MSL, la direction du vent moyenne quotidienne à 900.mb., la pression au niveau moyen de la mer quotidienne max.MSL., la Pression au niveau moyen de la mer quotidienne min.MSL., la Vitesse du vent quotidienne min à 10.m.au-dessus de la terre., la couverture nuageuse moyenne quotidienne max mid.cld.lay. ont influence significative sur la pluie.

une augmention d'une journée de ces variaables va entrainer une augmentation de la probabilité de pleuvoir. Nous allons développer cela pour les variables du modèle qui sera retenu.

On peut constater une correlation lineaire entre les heures et les minutes du fait des NA.

le modèle probit nous montre que la pression au niveau moyen de la mer moyenn .quotidienne.MSL, la direction du vent moyenne quotidienne 900.mb, la pression au niveau moyen de la mer quotidienne max .MSL, la pression au niveau moyen de la mer quotidienne min .MSL, la couverture nuageuse moyenne

quotidienne max mid.cld.lay, la vitesse du vent quotidienne min à 10.m.au-dessus de la terre, la direction du vent quotidienne moyenne à 80m au-dessus de la terre, la vitesse du vent quotidienne max à 10.m au-dessus de la terre ont une influence significative avec la pluie.

Dans le tableau suivant, nous avons consigné une caractéristique du modèle selon les 2 critère suivants que nous avons jugé pertinents : l'AIC et l'erreur de prediction. On a plusieurs autres critères pour faire le choix du modèles tels que le BIC, le cp de malows, l'AICc...

AIC : critre d' nformation d' Akaike, mesure de la qualité statistique d'un modèle

Erreur de prédiction : proportion de prédictions fausses du modèle sur le sous-échantillon de test

Pour déterminer ces prédictions, nous avons fixé un seuil alpha = 0,5.

Modèle complet		
	AIC	Erreur de prediction
Logit	1322.4	0.4789157
Probit	1325.8	0.5241379

Dans le modèle complet, l'aic et l'erreur de prédiction de la regression logistique sont inférieure à ceux de la regression probit.

Modèle par rétention des variables significatives

Dans la suite nous allons exclure les variables qui ne sont pas significatives. Nous repartons sur le même modèle vu précédemment dans le cadre avec intéraction. Le processus est maintenant de retirer manuellement les variables qui ne seraient pas significatives.

Modèle par rétention		
	AIC	Erreur de prediction
Logit	1329.8	0.51379314
Probit	1333.9	0.5103448

En ne gardant que les variables significatives on n'obtient un AIC de 1329.8 pour le modèle logit réduit et 1333.9 le modèle probit. On a toujours l'aic du modèle logistique qui reste plus faible. Mais avec une légère hausse de l'erreur de prédiction.

Modèle par sélection basée sur l'AIC

Dans cette section, nous cherchons à déterminer le meilleur modèle au sens de l'AIC avec la fonction step. Ce critère étant un critère de parcimonie, il favorise les modèles avec moins de variables explicatives et il convient de chercher à le minimiser. Le processus est de partir du modèle complet, puis d'éliminer ou de rajouter successivement des variables tant que cela permet de diminuer l'AIC et de s'arrêter dès lors qu'il n'est plus possible de minimiser ce critère. Dans le cas des modèles logit on a retenu les variables suivantes: X, 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. qui constitue le dernier modèle de la selection. Pour le modèle probit on a retenu X, Temperature.daily.mean..2.m.above.gnd.,Mean.Sea.Level.Pressure.daily.mean..M Snowfall.amount.raw.daily.sum..sfc., Total.Cloud.Cover.daily.mean..sfc., 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.min..sfc., High.Cloud.Cover.daily.max..high.cld.lay., Medium.Cloud.Cover.daily.max..mid.cld.lay., Low.Cloud.Cover.daily.max..low.cld. Wind.Speed.daily.max..10.m.above.gnd., Wind.Speed.daily.min..10.m.above.gnd., Wind.Gust.daily.max..sfc.

Modèle AIC		
	AIC	Erreur de prediction
Logit	1283.3	0.5103448
Probit	1287.3	0.4965517

L'aic du modèle logit avec la fonction step est de 1283.3 celui du modèle probit 1287.3. Et une lègère hausse de l'erreur de prédiction pour le modèle logit.

En interprétant le résultat de la regression logit, on peut conclure: quand la temperature quotidienne moyenne, de la Pression moyenne au niveau de la mer moyenne quotidienne, de la Moyenne quotidienne de la couverture nuageuse moyenne..mid.cld.lay, la Direction du vent moyenne quotidienne 900.mb, la Couverture nuageuse totale quotidienne max.sfc., Couverture nuageuse totale quotidienne min.sfc., Nébulosité moyenne quotidienne max mid.cld.lay.., Vitesse du vent.quotidienne max 10.m.au-dessus de la terre,Vitesse du vent.quotidienne min 10.m.au-dessus de la terre, Rafale de vent quotidienne max.sfc. augmente d'une unité la probabilité qu'il pleut le lendemain est de 1.160557,1.619634, 1.010919, 0.8913661, 0.9973755, 0.9973755, 1.004575, 0.8995146,0.7848207, 0.7360921, 1.00837, 1.007843, 1.007843, 1.006234, 1.062378, 1.118401, 1.023584 et négativement pour les autres variables.

Modèles avec interactions

Les dépendances étudiées plus haut nous permettent d'envisager de différentes manières le modèle réduit par une sélection basée sur l'AIC. Parmi les variables retenues dans ce modèle, nous pouvons étudier l'effet de certaines variables sur d'autres dans la régression. Nous avons fait l'interaction entre la pression et la nebulosité ainsi que la vitesse et la direction du vent. Mais dans l'étude nous prenons en compte le modèle logistique retenons l'interaction entre la pression et la nebulosité du vent.

Modèle avec interaction		
	AIC	Erreur de prediction
Logit	1283.3	0.5034483

On trouve un aic de 1283.3

Modèle multinomiale et modèle de poisson

Sachant que notre variable d'intérêt est binaire, les modèles logit et probit pourrait être les meilleurs pour réaliser notre étude. Mais nous allons ajouté les 2 modèles àa savoir le modèle multinomiale et le modèle de poisson. Et que la regression ordinaire n'utile quand l'ensemble prend des valeurs finies ordonnées. Nous avons réaliser une regréssion avec les variables retenus comme significative du modèle logit (dans le code R). Nous allons choisir de présenter les résultats obtenus par la fonction de selection de variable step.

Modèle AIC multinomiale et poisson	
	AIC
Multinomial	1283.294
Poisson	1863.9

Au critère de l'AIC, on trouve que le modèle multinomiale est meilleur que le modèle de poisson

Choix du meilleur modèle

L'étude de nos critères retenus nous indique que le choix du meilleur modèle doit être celui du modèle logit qui minimise l'AIC. Ceci en se focalisant sur le modèle obtenu par la fonction step Sachant l'erreur de prediction est presque la même pour les mêmes pour le modèle logit et probit..

Prédictions

le résultat de la prédiction retenu nous donne les jours dans le jeu de données test pour les quels la probabilité qu'il pleut qui doit être predite par le modèle. La table des prédictions comparées aux vraies valeurs nous montre qu'il y a 64 faux negatives et 84 faux positives. en calulant l'odds ration, on voit qu'il ya une association positive enntre le nombre de jours et et la pluie qui est la variable à expliquer.

En cherchant le seuil optimal on trouve le résultat de 0. Nous allons mesurer la performation des modèles (logit et probit) via l'AUC et la validation croisée. En moyenne les deux modèles modèles ont la même qualité de représentativité du modèle. mean du modèle logit est de 0.7371803 et celle du modèle probit est de 0.7344299.