D. Makowski, Paris, 2021

Travaux pratique : prédiction du rendement à l'aide de différentes méthodes de machine learning

Exercice 1 : Prédiction du rendement du maïs

Objectif

Le rendement agricole annuel d'une culture représente la quantité de produits récoltée par unité de surface une année donnée. Dans le cas du maïs, le rendement correspond à la quantité récoltée de grains, souvent exprimée en tonnes par hectare. Le rendement dépend des caractéristiques de la région où le maïs est cultivé et des conditions climatiques de l'année dans cette région (températures, rayonnements, précipitations etc.). La valeur du rendement est susceptible de varier fortement entre régions et entre années. Par exemple, le rendement pourra être anormalement faible une année présentant un déficit hydrique important à un stade clé du développement de la culture ou, au contraire, très élevé une année présentant des conditions climatiques optimales tout au long de la saison.

Il est important de prédire précisément le rendement avant la récolte. Celle-ci a généralement lieu en à l'automne pour le maïs. Des prédictions fiables avant la récolte offrent la possibilité aux opérateurs économiques régionaux de planifier leur récolte et constituent également une information stratégique utilisée par les acteurs opérants sur les marchés internationaux. Des prédictions de récoltes abondantes ou, au contraire, des prédictions de pertes importantes, peuvent fortement impacter les cours des marchés agricoles mondiaux.

L'objectif est de développer des outils permettant de prédire aussi précisément que possible les rendements du maïs en France.

Données

Deux fichiers sont disponibles :

- **Un fichier de données « entrainement »** pour le maïs (TrainingDataSet_Maize.txt) incluant (i) les anomalies rendements annuels de maïs en tonnes par ha pour différents départements français et pour 38 années tirées au hasard et (ii) les valeurs correspondantes de variables climatiques pour les mêmes départements et les mêmes années qui pourront être utilisées pour prédire les rendements. Une variable liée au pourcentage de surface agricole irriguée dans chaque département qui pourra également être prise en compte pour prédire le rendement. Dans ce fichier, les anomalies de rendement correspondent aux écarts entre les rendements observés et les tendances temporelles ajustées département par département. Une anomalie positive représente un gain de rendement par rapport au rendement attendu, et une anomalie négative correspond à une perte de rendement.
- **Un fichier « test »** maïs (TestDataSet_Maize.txt) incluant les valeurs des variables climatiques pour les mêmes départements mais pour 19 années non incluses dans les fichiers « entrainement ».

Questions

1. Entrainer un modèle LASSO sur le fichier d'entrainement en considérant différents niveaux de pénalisation. Présenter les coefficients de régression soit sous forme de tableau, soit sous forme de figure.
2. Identifier le niveau de pénalisation optimale par validation croisée.
3. Sélectionner le meilleur modèle d'après la validation croisée et déterminer les valeurs de ses coefficients de régression.
4. Utiliser le modèle sélectionné pour prédire les rendements du fichier test, en vous basant sur les coefficients estimés avec le fichier d'entrainement.
5. Calculer le RMSE de ces prédictions à partir du fichier test et réaliser un graphique comparant les valeurs prédites aux valeurs observées.
6. Répéter la procédure avec un modèle random forest : entrainer un modèle random forest avec les valeurs par défaut de `ntree` et `mtry`.
7. Entrainer deux autres modèles random forest avec en ajoutant +1 à `mtry` puis en retranchant -1 à `mtry`.
8. Sélectionner le meilleur modèle random forest d'après le % de variance expliqué. Identifier les variables les plus influentes en traçant un graphique d'importance avec ce modèle (fonction `varImpPlot`). Quelles sont les deux variables les plus influentes sur le rendement ? Est-ce logique ?
9. Utiliser le modèle random forest sélectionné pour prédire les rendements du fichier test, calculer son RMSE.
10. Choisir le meilleur modèle entre LASSO et random forest. Justifier.

Liste des variables

yield_anomaly : variable à prédire représentant l'anomalie de rendement de maïs (une valeur positive indique un rendement plus élevé qu'attendu, une valeur négative indique une valeur perte de rendement par rapport à la valeur attendue), exprimée en tonne par ha.

year_harvest : année (anonyme) de récolte (1 à 57)

NUMD : numéro (anonyme) indiquant le département (de 1 à 94).

La variable yield_anomaly doit être prédite **uniquement** à l'aide des variables suivantes (ou d'une partie de ces variables) :

- ETP_1… ETP_9 : Evapotranspiration potentielle moyenne mensuelle par année et par département (1= janvier, 9=septembre)
- PR_1… PR_9 : Précipitation cumulée mensuelle par année et par département (1= janvier, 9=septembre)
- RV_1… RV_9 : Rayonnement moyen mensuel par année et par département (1= janvier, 9=septembre)
- SeqPR1…SeqPR9 : Nombre de jours de pluie mensuel par année et par département (1= janvier, 9=septembre)
- Tn_1…Tn_9 : Température minimale journalière moyenne mensuelle par année et par département (1= janvier, 9=septembre)
- Tx_1…Tx_9 : Température maximale journalière moyenne mensuelle par année et par département (1= janvier, 9=septembre)

- IRR : variable comprise entre 1 et 5 liée à la fraction de surface agricole irriguée dans chaque département. La valeur 1 indique une fraction faible, la valeur 5 indique une fraction élevée de surface irriguée. Ces valeurs sont indicatives car établies sur la base d'information collectée pendant une seule année.

Important : Le maïs est généralement semé au printemps et est récolté à l'automne. Les valeurs des variables climatiques pour les mois 1 à 9 correspondent aux valeurs obtenues l'année de récolte. Elles sont disponibles avant la récolte et peuvent donc être utilisées directement pour prédire le rendement. Les valeurs des variables climatiques des mois 10 à 12 sont absentes des fichiers.

Exercice 2 : Test de deux méthodes supplémentaires

Répéter la procédure ci-dessus avec deux autres modèles :

- La régression ridge (argument alpha=0 dans `glmnet`)
- La régression pls

Pour ridge, optimiser le coefficient de pénalisation par validation croisée (fonction `cv.glmnet`) sur le jeu de données d'entrainement, sélectionner le meilleur modèle, puis utiliser ce modèle pour prédire les rendements du fichier test.

Pour pls, identifier le nombre optimal de composantes principales à l'aide de la validation croisée leave-one-out (LOO) de la fonction `plsr`. Utiliser le modèle correspondant pour prédire les rendements du fichier test.

Identifier le meilleur des quatre modèles testés.