Projet

de Data Science 2A

Prédiction de maladies cardiaques

RAPPORT DU PROJET

Elève 1 :
Aaron MAMANN

Elève 2:
Edgar JOUISSE

Enseignant:
M. Xavier DUPRE

1 Introduction : Problématique et Aperçu de la base de données

Les maladies, qu'elles soient de nature physiologiques ou psychologiques, ont toujours fait interroger le milieu scientifique quant à leurs origines. Durant ce projet, nous avons décidé de nous pencher sur la question des maladies cardiaques : quels critères bien précis (présence de douleurs au niveau de la poitrine, débit du sang...) nous permettent de conclure que le patient est en effet atteint d'une maladie cardiaque ? Nous avons pour cela étudié une base de données intitulée "*Heart Disease UCI*" que l'on peut retrouver sur le site Kaggle. Nous nous rendons vite compte que la variable qui représente le plus grand intérêt est "*target*", une variable binaire qui indique si le patient est atteint d'une maladie cardiaque ou non. D'autres variables, pertinentes dans la détermination d'une maladie cardiaque, sont présentes comme le type de douleur au niveau de la poitrine, le taux de cholestérol, ou bien des résultats de l'électrocardiographie (liste exhaustive de la base de données sur le notebook). Notons que la base de données est relativement petite étant donné qu'elle comporte que 303 observations et 14 variables. Nous avons donc choisi des modèles d'apprentissage en conséquence de cette particularité. Notre but a été de voir, avec des modèles d'apprentissage choisis, dans quelle mesure il est possible de prédire la présence d'une maladie cardiaque (autrement dit la variable d'intérêt est la variable "target") avec l'aide des autres variables de la base de données (variables explicatives).

Implications dans le projet de chaque collaborateur :

Chapitre 2 : Présentation brève des modèles et de leur pertinence pour cette base de données : Chapitre Fait par nous deux : **Edgar**

Chapitre 3 : Comparaison des deux modèles selon différents critères : concernant le premier modèle (Extremely Randomized Trees) : **Aaron**, Concernant le deuxième modèle (Régression Logistique) : **Edgar**

Chapitre 4 : Amélioration de la résistance à l'overfitting de l'Extremely Randomized Trees avec sélection optimal des hyperparamètres : **Aaron**

Chapitre 5 : Réduction des coûts par réduction du nombre de variables explicatives : **Aaron**

Chapitre 6 : Comparaison de la vitesse d'apprentissage des deux modèles choisis : **Edgar**

2 Présentation brève des modèles et de leur pertinence pour cette base de données

Le premier modèle choisi est l'Extremely Randomized Trees. Ce modèle est proche de Random Forest mais a néanmoins quelques divergences. Il consiste à construire un grand nombre d'arbres de décision. Ses arbres de décision sont différents de ceux établis par le Random Forest. Dans les arbres de décision de l'Extremely Randomized Trees, chaque séparation ("split") est faite de manière aléatoire alors que les séparations dans les arbres de décision du Random Forest sont faites notamment dans le but de discriminer le plus possible les observations de la base d'apprentissage.

Par conséquent, les arbres de décision de l'Extremely Randomized Trees sont beaucoup moins soumis au problème d'overfitting que ceux du Random Forest. Or, étant donné que la base de données ne comporte que 300 observations et donc que la base d'apprentissage est encore plus petite, le risque d'overfitting est relativement problématique. Ainsi, il est très profitable d'avoir un modèle qui est peu affecté par ce problème d'overfitting. Le deuxième critère de choix pour ce modèle est sa performance très satisfaisante (quant à ce problème) que l'on va évaluer de différentes manières par la suite.

La régression logistique consiste à approximer la probabilité p qu'une des deux classes de la variable d'intérêt (ici la classe '$target = 1$') soit choisie étant donné les observations sur les variables explicatives. L'approximation de cette probabilité est donnée par la fonction suivante dans laquelle X représente les variables explicatives :

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

La classe attribuée ($target = 1$ ou $target = 0$) sera évaluée en fonction du seuil donné à p (par défaut 0.5). L'estimation des coefficients β_0 et β_1 pour estimer la probabilité p est effectuée par le biais de la méthode du maximum de vraisemblance.

Nous avons choisi la régression logistique d'abord pour sa résistance à l'overfitting et ensuite pour ses bonnes performances pour ce problème, que l'on va également évaluer de différentes manières par la suite.

3 Comparaison des deux modèles selon différents critères

3.1 Score d'Exactitude sur la base d'apprentissage/test et Problème d'Overfitting

Après avoir entrainé les deux modèles sur le même set d'apprentissage, on a évalué leur exactitude (précision) sur le set d'apprentissage et sur le set de test. Concernant le set d'apprentissage, le classifieur Extremely Randomized Trees performe plus que le classifieur issu de la régression logistique : leur score d'exactitude est respectivement de 100% contre 84.6%. Néanmoins, les tendances s'inversent quand il s'agit du set de test, bien que l'écart de score entre les deux modèles ne soit pas très grand non plus. En effet, pour le set de test, le classifieur Extremely Randomized Trees est moins performant que le classifieur issu de la régression logistique : leur score d'exactitude est respectivement de 81.6% contre 86.8%.

On peut donc émettre de ces scores des déductions quant aux deux modèles. Pour le premier modèle (Extremely Randomized Trees), on remarque un score d'exactitude très élevé sur le training set et un écart significatif avec ce score calculé sur le set de test. Bien que le score calculé sur le set de test (81.6%) reste bon, cela souligne le fait que le modèle a légèrement du mal à généraliser car trop spécialisé sur le training set. Il est important de souligner le caractère léger de l'overfitting car le modèle de l'Extremely Randomized Trees restent tout de même résistant à ce sur-apprentissage. Pour le second modèle (Régression Logistique) bien que son score d'exactitude sur le training set n'est pas aussi élevé que celui du premier modèle, bien que bon (84.6%), il se généralise très bien puisqu'il conserve (voir améliore légèrement) son score sur le set de test.

3.2 Validation Croisée et Problème d'Overfitting

Il existe plusieurs types de validations croisées. Celui utilisé ici s'appelle le k-fold cross-validation (par défaut on a $'k = 5'$ mais on a choisi $'k = 15'$ pour améliorer le test). Le k-fold cross-validation consiste à diviser l'échantillon

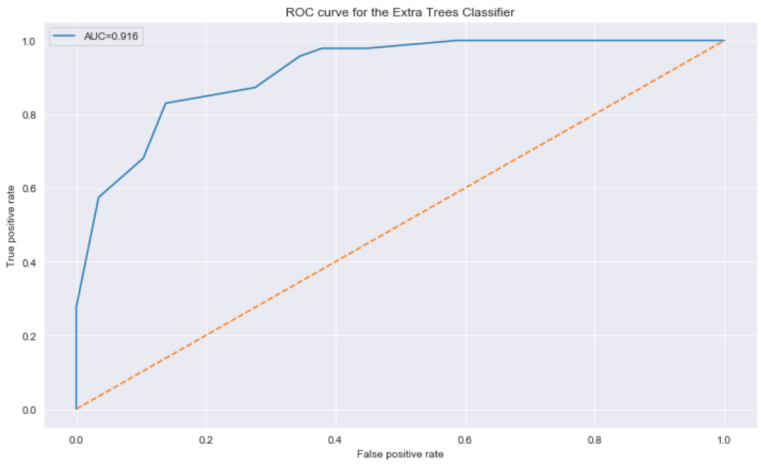
original en k échantillons, puis on sélectionne un des k échantillons comme ensemble de validation et les $k-1$ autres échantillons constitueront l'ensemble d'apprentissage. On calcule comme dans la première méthode le score de performance, puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les k-1 échantillons qui n'ont pas encore été utilisés pour la validation du modèle. L'opération se répète ainsi k fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation.

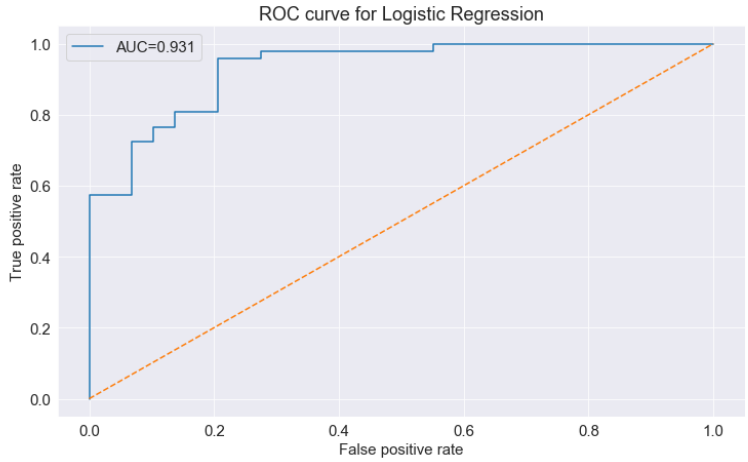
Là encore le test montre que le modèle de la régression logistique généralise mieux que celui de l'Extremely Randomized Trees. Etant donné que la validation croisée prend chaque fois des ensembles de tests différents elle met encore plus en lumière le problème d'overfitting. En effet, les moyennes du vecteur des scores (d'exactitude) de validation croisée est égal à 0.789 pour l'Extremely Randomized Trees et égal à 0.825 pour la régression logistique. Cela montre encore une fois que la résistance à l'overfitting du premier modèle est plus faible que le second.

3.3 Evolution du taux de "vrais positifs" en fonction du taux de "faux positifs"

Le taux de "vrais positifs" mesure la proportion d'observations qui vérifient $'target' = 1$ et qui sont correctement identifiées comme telles. Autrement dit, le taux de "vrais positifs" est égal au nombre de malades qui sont correctement identifiés sur le nombre total de malades. En terme de vocabulaire, les "positifs" peuvent correspondre à n'importe quoi, ici ce sont les malades.

Voici donc la courbe ROC des deux modèles qui montre, pour les deux modèles, l'évolution du taux de "vrais positifs" en fonction du taux de "faux positifs".





L'aire sous les courbes ROC est égale à l'indice AUC. Quand cet indice est par exemple supérieur à 0.90 cela signifie que les prédictions du modèle sont très fiables et donc que le modèle est performant. C'est le cas pour nos deux modèles ce qui veut dire qu'ils performent très bien. La principale différence entre les deux courbes ROC ci-dessus est que leur forme diffèrent. La courbe ROC de la régression logistique segmentée en paliers montre que quand le taux de faux postifs augmente le taux de vrai positif stagne ou augmente d'un coup alors que cette augmentation est plus progressive dans le cadre de l'Extremely Randomized Trees.

4 Amélioration de la résistance à l'overfitting de l'Extremely Randomized Trees avec sélection optimal des hyperparamètres

On va choisir les hyperparamètres de l'Extremely Randomized Trees pour qu'il se généralise mieux, quitte à ce qu'il soit moins performant sur la base d'apprentissage. Pour la sélection optimal des hyperparamètres, on fait appel sur Python aux fonctions d'hyperopt. En effet, on sélectionne les hyperparamètres de manière à faire augmenter la moyenne des scores (d'exactitude) de validation croisée qu'on a donc pris comme fonction objectif. En effet, on a vu que la moyenne des scores de validation croisée était un indicateur fiable de l'overfitting. Les hyperparamètres qu l'on a mis en jeu pour optimiser le modèle sont la profondeur maximale de chaque arbre de décision impliqué dans l'Extremely Randomized Trees et le nombre d'estimateurs.

Grâce à cette méthode, on a pu faire remonter la moyenne des scores d'exactitude de validation croisée de l'Extremely Randomized Trees au même niveau que celle de la Régression logistique c'est à dire à 0.826, ce qui est un signe de l'amélioration de la résistance à l'overfitting de l'Extremely Randomized Trees grâce à la sélection optimale des hyperparamètres.

5 Réduction des coûts par réduction du nombre de variables explicatives

Chaque variable explicative a un certain coût pour l'étude. Dans le cadre d'une réduction des coûts, on souhaite réduire le nombre de variables explicatives tout en gardant les meilleures performances possibles.

Pour réduire de manière optimale les variables explicatives, on utilise la pénalisation avec la norme L1 dans les deux modèles. En effet, à la différence de la pénalisation par la norme L2, la pénalisation par la norme L1 ne réduit pas simplement la valeur de toutes les coordonnées d'un vecteur mais en annule ce qui permet de sélectionner les variables explicatives.

Dans le cadre de la régression logistique, on rappelle que l'on a :
$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$

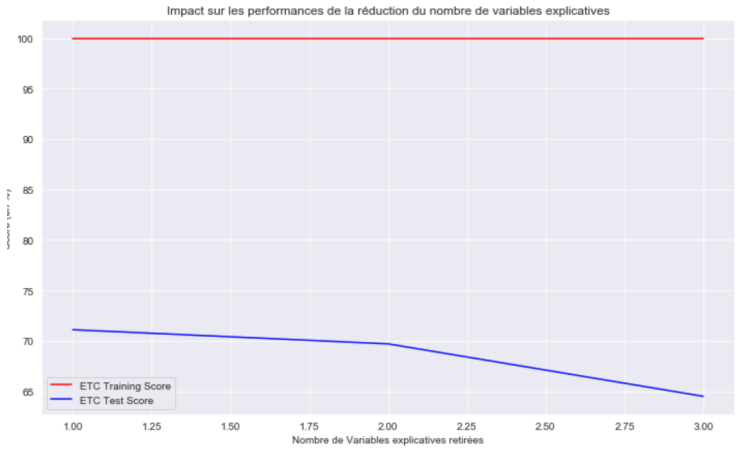
Avec la pénalisation par la norme L1 on arrive à annuler les coordonnées β_1 qui sont devant les variables explicatives, ce qui permet donc de les sélectionner. Par cette procédure, on voit bien dans notre notebook, en appliquant la régression logistique pénalisée en L1 avec une tolérance $tol = 1e\text{-}6$ que 4 variables explicatives ont été retirées et qu'on maintient pour de très bonnes performances puisque le score d'exactitude sur le set d'apprentissage est de 86.842%. Les variables explicatives retirées sont l'âge, le taux de cholestérol, la tension au repos, le rythme cardiaque maximal au repos.

Dans le cadre de l'Extremely Randomized Trees, pour la réduction du nombre de variables, on a utilisé la méthode du Sparse ACP qui permet de maximiser la variance total des observations en réduisant leurs nombres de features/variables explicatives. Ces variables explicatives ne sont plus les mêmes mais sont des combinaisons linéaires des précédentes variables explicatives.

Or, comme on a pénalisé en norme L1, les coefficients dans les combinaisons linéaires des précédentes variables explicatives s'annulent, ce qui fait que les nouvelles variables ne dépendent plus des anciennes et donc on n'a plus besoin de mesurer certaines variables, ce qui réduit les coûts.

En augmentant le coefficient devant la pénalisation, on a remarqué que toutes les nouvelles variables explicatives (trouvées par le Sparse ACP) ne dépendent plus de 1 puis de 2 puis de 3 variables explicatives d'origine. En effet, en regardant les coefficients de chaque nouvelle variable, on retrouve les coefficients nuls devant les mêmes variables d'origine.

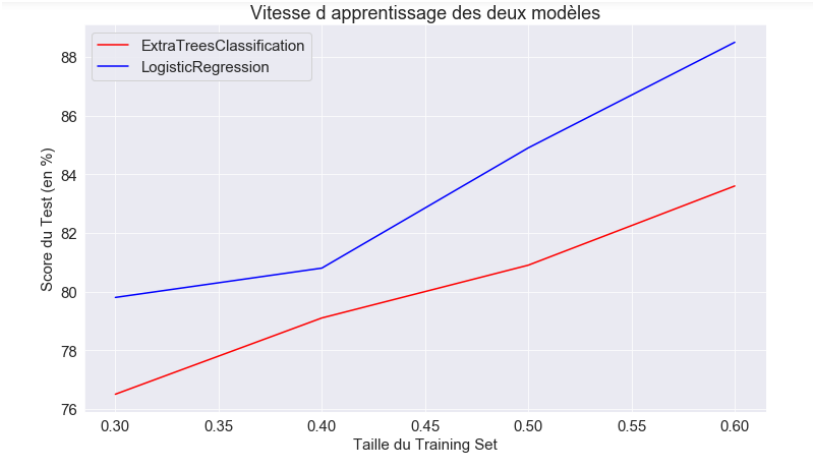
On mesure chaque fois les performances de l'Extremely Randomized Forest entraîné puis testé sur la nouvelle base qui ne dépend plus progressivement d'1, de 2 puis de 3 variables explicatives d'orgine.



On remarque que le score sur le set de test de l'Extremely Randomized Forest reste très correct puisque pour une variable d'origine retirée, il est de 71.1% , puis pour deux variables d'origine retirées il est de 69.7% et enfin pour trois variables d'origine retirées il est de 64.5%.

6 Comparaison de la vitesse d'apprentissage des deux modèles choisis

Nous présentons ci-dessous le graphique montrant l'évolution du score du test en fonction de la taille du training set pour chacun des 2 modèles.



Premièrement, nous remarquons que quelque soit la taille du training set, la régression logistique présente un meilleur score de test que celui du Extremely Randomized Forest. De plus, à partir d'une certaine taille (0.40 de la base de données totale), le score de test croît plus vite pour la régression logistique que pour l'autre modèle, ce qui peut se voir à travers la pente des deux droites respectives. Néanmoins, les 2 modèles présentent des performances satisfaisantes en terme de vitesse d'apprentissage car elles sont linéaires avec des pentes différentes selon les intervalles.

7 Conclusion

En définitive, nous avons étudié les performances d'apprentissage de deux modèles différents, la régression logistique et l'Extremely Randomized Trees, sur une base de données répertoriant plusieurs indices, comme la douleur au niveau de la poitrine, centrés sur une variable binaire stipulant si un individu est atteint d'une maladie cardiaque ou non. Il s'avère que ces deux modèles ont été choisis en raison de leur résistance à l'overfitting et à leur adaptabilité aux variables d'intérêt binaires. Il se trouve que les deux performances sont satisfaisantes, malgré le fait que la régression logistique semble légèrement mieux en ce qui concerne certains critères, comme la validation croisée, le score d'exactitude sur la base d'apprentissage/test, ou encore la vitesse d'apprentissage. En revanche, en ce qui concerne l'aire sous la courbe ROC, le modèle Extremely Randomized Trees semble légèrement préfèrable par sa stabilité. Néanmoins, la performance de ce dernier peut être relevée par l'introduction d'hyperparamètres qui égalise les performances des 2 modèles. La réduction des coûts était par ailleurs plutôt satisfaisante.