Bonjour, je vais vous présenter le projet 8 communiquer vos résultats.

J’ai travaillé sur des données qui m’ont été fourni par la bio-informaticienne de la start up arraytec.

C’est une toute nouvelle start-up qui travail à la création de test diagnostique, sur des maladies respiratoires à l’aide de puce à ADN. Son équipe est composée de six personnes un PDG, une qualiticienne, une biologiste, un technicien de laboratoire, un chimiste et la bio informaticienne avec laquelle j’ai travaillé.

Actuellement, le diagnostique est très souvent réalisé grâce aux méthodes microbiologiques traditionnelles (avec cultures bactériennes). L’émergence de nouvelles résistances aux antibiotiques et le temps important nécessaire à la culture bactérienne sont deux obstacles pour ce type de diagnostic. Ainsi, de plus en plus d’entreprises cherchent à développer des tests faciles d’utilisation, rapides, sensibles et spécifiques, applicables aux échantillons habituels tout en gardant un prix raisonnable.

La construction des tests Arraytec se fait en plusieurs étapes, une de ces étape consiste a utilisé un modèle pour prédire la présence de bactéries dans un échantillon à partir de l’intensité de fluorescence de sonde. Mon but était de construire un modèle plus performant que le modèle actuel.

Le dispositif d’arraytec possède de très gros atouts :

* Il permet de tester jusqu’à 10 pathogènes en même temps c’est ce qui fait ça force.
* Il est rapide et peu cher

Mais aussi des point faible :

* Une technique pas assez automatisé
* un manque de précision par rapport au méthode microbiologique

Les 10 pathogènes à détecter sont les suivants :

La bordetella, la moraxella cataralis , haemosphilus influenza , Klebsiella pneumoniae, Legionellapneumophila, Staphilococcus spp, streptococus pneumoniae, pseudomonas aeruginosa, neisseria meningitidis et mycoplasma pneumoniae

Ces bactéries sont les principales responsablent des otites, kératite, conjonctivites, septicémies, endocardites, méningite, pneumoni, pneumopathies, arthrites etc….

elles peuvent aussi être responsable de la coqueluche, de la maladie du légionnaire, de la fièvre pontiac et de bien d’autre encore.

Pour détecter la présence des bactéri arraytec utilise des puces à ADN

Une fois les puces fabriquées, un échantillon d’ADN préalablement amplifié, marqué par PCR et purifié peut être déposé sur la lame. Il s’hybridera ou non avec les sondes, suivant le matériel génétique qu’il contient. Lorsqu’il y a hybridation entre la sonde et l’oligonucléotide cible complémentaire à celle-ci, un spot sera visible. Dans le cas contraire, le spot sera éteint.

L’ADN utilisé peut provenir de souches qui sont achetées auprès de fournisseurs ou bien d’échantillons cliniques de patients atteints d’une infection respiratoire bactérienne et qui sont transmis à Arraytec par des laboratoires d’analyses médicales partenaires ou hôpitaux.

Un scanner lit ensuite la lame puis on obtient les résultats de fluorescence grace au logiciel mapix. Enfin les donnés sont analysé sur R.

Problématique :

Arraytec utilisent la regression logistique comme méthode de classification mais ils n’en ont testé aucune autres.

Il est possible que de meilleurs méthodes plus adapté existent.

Arraytec souhaite aussi réduire le nombre de sonde utilisé afin de pouvoir tester plusieurs échantillons en même temps mais ils ne veulent néanmoins pas perdre de précision.

Actuellement aucune méthode automatique n’a été mise en place pour sélectionner les sondes les plus importantes.

Mes missions :

* Rechercher un modèle de prédiction plus efficace que le modèle utilisé actuellement. Le but étant d’avoir une précision et un indice kappa maximal avec un nombre de sondes moins important (40 ou moins) alors que le modèle actuelle en compte 66.
* Développer des outils pour accélérer la sélection des sondes et orienter la construction des nouvelles sondes.

Les données

66 variables explicatives quantitaive continue:

* Elles prennent comme valeurs les indices d’intensités de fluorescence des sondes ( 1sonde 1 variable)

1 variable expliquée :

* Variable qualitatives à 11 modalités
* Ces modalités sont les 10 souches bactériennes dont on a parlé précédemment

et la modalité « Xneg » qui signifie négatif aucune bactérie n’a été détectée.

Dans ma base de donné j’ai 236 échantillons :

* 100 proviennent de souches pures et 136 d’échantillons cliniques
* Les échantillons proviennent d’une base d’apprentissage et sont vérifiés

J’ai construit mon plan en suivant la démarche d’analyse que j’ai suivi à savoir :

-analyse univari

-analyse bivarié

-analyse multivarié

-choix et optimisation du modèle

-conclusion

-bilan personel

Ma démarche :

Analyse univarié :

J’ai construit les boxplot de chacune de mes variables pour étudier leur dispersion. J’ai observé que j’avais à faire à une distribution exponentiel et qu’une transformation logarithmique devrait permettre une meilleur lecture des données. J’ai refait mes boxplot après transformation et en effet les données était bien plus lisibles. Mes valeurs étaient comprise entre 0 et 65000 maintenant elles sont entre 0 et 4,81.

J’ai ensuite essayé de comprendre comment été construites les sondes. J’ai appris qu’il y en avait 2 types les spécifique qui ne s’allument que pour une seule souche et les aspécifique qui s’allument pour plusieurs souches. Chaque sonde pouvant être de qualité variable. J’ai alors créer des tableaux pour résumer l’information afin de pouvoir identifier le type d’une sonde et sa qualité.

Le premier tableau nous donne le taux d’allumage des sondes pour chaque souche. Cela nous permet de connaitre très facilement le type des sondes et la qualités.

On voit très vite par exemple que les sondes X1 et X2 sont de mauvaise qualité elles ne s’allument presque jamais, que X24 est spécifique à la souche Mpn et que c’est une souche de très grande qualité elles ne s’allume presque que pour la souche mpn et s’allume 100 % du temps dans ce cas la etc….

Ce tableau en revanche ne nous donne aucune indication sur l’intensité des sondes.

Le deuxieme tableau nous donne les valeur médiane des sonde par souche et le troisième les moyenne des valeurs des sondes par souche. Grâce à ces deux tableaux j’ai une bonne indication sur l’intensité des sondes. Cela nous permet par exemple de savoir si une sonde aspécifique s’allume de manière différente pour chaque souche et donc d’en connaitre la qualité.

Du coup grace à ces trois tableaux j’ai trier mes sonde. Je les ai trié dans un 4em tableau avec 2 colonne, d’un coté les spécifique de l’autre les aspécifique et 4 lignes pour les différentes qualités ++ très bonne qualité + bonne qualité – mauvaise qualité –- très mauvaise qualité

Analyse bivarié :

J’ai étudié les corrélations entre mes variables. J’ai vu que beaucoup de mes sondes étaient fortement corrélé. J’ai alors construit la matrice de corrélation de ces dernières, afin d’identifier les groupes de corrélations.

A priori Réduire à une variable chacun de ces groupes n’aura que très peu d’impact sur les prédictions de mes futurs modèles. Je testerai d’éliminer ces variable en tant voulu. Je décide pour l’instant de garder toute l’information.

Analyse graphique + kmeans :

Je décide de faire une PCA, d’abord pour mieux évaluer mes variables grace au cercle de corrélation puis pour visualiser mes échantillons. J’ai utilisé une forme et une couleur différente pour chaque souche pour voir à quel point les échantillons d’une même souche se regroupaient.

Et même si mon premier plan n’explique que 30.7% de la variance, on observe une assez bonne répartition par souche cela est plutôt de bonne augure, les résultat des prédictions devraient être très bons. Je vais tenter de faire un kmeans clustering. Même si les résultat sont souvent moins bon qu’en apprentissage supervisé, l’avantage qu’on aura si les résultat sont satisfaisant c’est qu’on pourrait par exemple identifier de nouvelles bactéries non présente dans notre étude ou même qui n’existe pas encore. Malheureusement 79.7% de précision sur un mélange de souche pure et d’échantillon très propre ce n’est pas suffisant.

Apprentissage supervisé :

C’est l’étape principal de mon étude, j’ai d’abord passé beaucoup de temps a étudié de nombreux modèle d’apprentissage (Dans des livre et sur des site spécialisé dans le machine learning) puis j’ai sélectionné les 5 modèles qui me semblait les plus pertinent pour mon étude . J’ai donné dans mon rapport une explication rapide pour chaque modèle utilisé.

Les modèle que J’ai choisis sont :

la lda analyse discriminante linéaire, elle est simple à comprendre et efficace dans de nombreux cas mais cette méthode est assez rigide et émet des hypothèses très forte sur les données qui ne sont ici pas respecté

LDA suppose que les prédicteurs sont normalement distribués (distribution gaussienne) et que les différentes classes ont des moyennes spécifiques à la classe et une variance / covariance égale.

La rda analyse discriminante régularisé qui est plus difficil à interpréter que la premiere mais bien plus flexible.

La régression logistique multinomiale qui est la méthode utilisé par l’entreprise actuellement. C’est extension de la regression logistique a un modèle multiclasse

Le réseau de neurone qui est sans doute la méthode la plus performante d’apprantissagesupervisé. C’est une méthode extrêmement flexible en revanche c’est une « boite noir » elle ne nous apportera aucune information sur les vairables

Random forest c’est une méthode très flexible et très puissante et qui permet de connaitre l’importance des variables pour le modèle.

Phase de Tunning :

Avant de pouvoir utiliser cetain modèle il m’a fallu trouver les paramètres qui leur permettrai d’être les plus efficace possible. J’ai fait cela pour le réseau de neuronne et random forest.

Alors concrètement, j’ai testé plein de combinaison de paramètre j’ai affiché les résultat de chaque combinaison dans un tableau puis j’ai sélectionner la plus précise.

Pour le réseau de neurone on utilisera 50 neurone dans la couche caché et on utilsera un poid de decay de 5\*10 exposant -10. J’ai choisis des valeur très faible pour éviter le surapprentissage.

pour random forest on va utiliser 1500 arbre et on utilisera 5 variable a chaque division.

Sélection du modèle :

J’ai ensuite testé mes 5 modèle via deux méthode d’abord en cross-validation sur la base d’apprentissage avec une répartion 70/30 classique j’ai afficher les résultat sous forme de matrice de confusion, j’ai ainsi accés au faux positif, faux négatif vrai positif vrai négatif pour chaque souche par exemple j’ai ici pour les hi 20 faux négatif dont 2 prédit Nm et 18 prédit spn. Et 51 faux positif qui sont en réalité des spn.

J’ai aussi la précision et l’indice kappa du modèle, j’avais aussi accés au spécificité et sensibilté de chaque souche je ne les pas affiché dans mon rapport car arraytec n’avait pas d’objectif précis à atteindre concernant ces 2 paramètres .

J’ai ensuite analysé mes résultat dans un tableau qui donne la précision et l’indice kappa de mes modèle pour ce premier test. J’ai alors éliminé 2 modèle

La lda qui obtient un résultat trop faible en cross validation

Et Le réseau neurone car il obtient un résultat similaire aux autre alors qu’il ne permet pas la sélection de variable

Ensuite j’ai testé les modèles sur de nouveaux échanillons moins propre que les premiers

(en effet dans ces nouveaux échantillon on peut souvent noter la présence de bactéri en portage ce sont des bactérie présente dans la salive mais qui ne sont pas pathogène, qui viennent s’hybrider avec nos sonde et donc brouiller un peu notre étude, cela est tout à fait naturel mais les échantillons de la base d’apprentissage sont eux presque pur ils ont été sélectionné pour ça) .

Ce test est aussi d’une grande importance car il est plus proche d’un test en condition réel.

Ici aussi j’ai analysé les résultats dans un tableau qui compare la précision et l’indice de kappa des trois modèle restant. Et j’ai éliminé 2 modèles

La regression logistique et l’analyse discriminante régularisé qui obtiennent un score trop faible sur les nouveaux échantillons par rapport a random forest le modèle que j’ai choisit

Etude approfondie du modèle choisi :

J’ai trouve des fonctionnalité très intéressante qui me permettent de connaitre l’importance globale de mes variable ainsi que leur importance pour chaque souche. J’ai reporté ces importance dans deux tableaux.

Le premier donne l’importance par souche de chaque sonde et le deuxieme l’importance globale

*La diminution moyenne de Gini est la moyenne (moyenne) de la diminution totale d'une variable de l'impureté du nœud, pondérée par la proportion d'échantillons atteignant ce nœud dans chaque arbre de décision individuel dans la forêt aléatoire. Il s'agit en fait d'une mesure de l'importance d'une variable pour estimer la valeur de la variable cible pour tous les arbres qui composent la forêt. Une diminution moyenne plus élevée de Gini indique une importance variable plus élevée.*

J’ai pu construire à partir du premier tableau des barplot qui donne les 10 sondes les plus importantes pour chaque souche cela permet de bien les visualiser et nous donne de bonnes indication sur les futurs sondes à créer.

On peut voir par exemple qu’il faudrait créer de nouvelles sondes spécifiques pour la souche Spn.

La souche Spn est déterminée presque exclusivement par une seule sonde la sonde X32 et il se trouve que c’est une sonde spécifique de mauvaise qualité. En effet elle s’allume plus d’une fois sur deux pour une autre souche, la souche Hi. C’est ce qui explique les nombreux Hi qui sont classés Spn et vice et versa par nos modèles.

Optimisation :

Grâce à l’algorithme rfe (recursive featur elimination). J’ai tracé la courbe qui donne la précision de mon modèle en cross-validation selon mon nombre de variable.

Je vois que pour 40 sondes mon modèle est encore très efficace + de 94% de précision

Je décide alors de récupérer les 40 sondes les plus utiles selon l’algo rfe et de les utilisé pour créer mon modèle final.

Dans les sonde éliminer je retrouve les sondes que j’avais classé de mauvaise qualité au départ ainsi que certaine variable très corrélés. Cela valide la cohérence de l’algorithme rfe.

J’ai ensuite testé mon modèle random forest à 40 variable sur mes nouveaux échantillons et j’ai obtenu une précision de 82% ce qui est très satisfaisant.

Conclusion : dans la conclusion j’ai réexpliqué toute ma démarche et j’ai tenté de montré la supériorité de mon nouveau modèle par rapport à l’ancien. J’ai pour cela construit un tableau qui récapitule les donnés des 2 modèles.

On voit que Notre nouveau modèle est légèrement moins bon en cross-validation mais bien meilleur sur les échantillons test. Il a aussi un nombre de sondes plus faible.

Le modèle a ensutie été validé par la bio informaticienne d’arraytec au vue des avantages qu’il confère.

Bilan personel

J’ai remercié la bio informaticienne d’arraytec puis ou j’ai donné mes impression sur ce projet