**Devoir final de cours de l’introduction à l’exploitation de données**

**Niloofar Sokhandan Asl**

**11278920**

**Le 26 Juin 2020**

**Partie 1 :**

Toujours au sein de la compagnie, vous décidez de bâtir un modèle prédictif afin de mieux cibler les personnes qui ont le plus de chance d’aller à Vinales. Vous construisez alors 3 modèles : un modèle d’arbre de classification, un modèle de forêt aléatoire et un modèle de réseaux de neurones.

1. Pour cet arbre de classification, veuillez décrire les règles du modèle qui conduisent à la prédiction d’un voyage à Vinales. Avec ces résultats, veuillez décrire brièvement en quoi consisterait la campagne pour attirer de nouveaux clients vers cette destination.

* Règle 1 :

27% des clients **âgé de 35 ans et plus dont revenue est supérieure à 60k** vont acheter le billet d’avion pour aller à Vinales. (300 personnes ; 300/1560 = 19.2% de tous les clients)

* Règle 2 :

89% des clients **âgé de 24 ans et moins dont revenue est supérieure à 60k** vont acheter le billet d’avion pour aller à Vinales. (100 personnes ; 100/1560 = 6.4% de tous les clients)

* Règle 3 :

13% des clients **dont revenue est entre 20k et 60k** vont acheter le billet d’avion pour aller à Vinales. (200 personnes ; 200/1560 = 12.8% de tous les clients)

* Règle 4 :

95% des clients **dont revenue est inférieure à 20k et qui ont un diplôme l'universitaire et qui ont voyagé plus que 2 fois par année** vont acheter le billet d’avion pour aller à Vinales. (560 personnes ; 560/1560 = 36% de tous les clients)

* Règle 5 :

18% des clients **dont revenue est inférieure à 20k et qui ont un diplôme l'universitaire et qui ont voyagé 2 fois et moins par année** vont acheter le billet d’avion pour aller à Vinales. (100 personnes ; 100/1560 = 6.4% de tous les clients)

* Règle 6 :

15% des clients **dont revenue est inférieure à 20k et qui n’ont pas un diplôme l'universitaire** vont acheter le billet d’avion pour aller à Vinales. (300 personnes ; 300/1560 = 19.2% de tous les clients)

Selon les résultats de cet arbre, afin de fructifier la vente de billet à Vinales, la compagne veut mieux préparer des promotions pour les clients de la règle 4 qui construisent 36% de la clientèle de la compagnie.

1. À l’aide des informations contenues dans la sortie de l’arbre de classification, veuillez déterminer le taux de mauvaise classification. En comparant ce taux de mauvaise classification avec une autre mesure, veuillez-vous prononcer sur l’utilité du modèle pour la compagnie Traveligo.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Règle 1 | Règle 2 | Règle 3 | Règle 4 | Règle 5 | Règle 6 | Somme |
| % des gens qui achètent le billet | 27% | 89% | 13% | 95% | 18% | 15% |  |
| # de l'individus de la feuille | 300 | 100 | 200 | 560 | 100 | 300 | 1560 |
| # des gens qui ont voyagé | 81 | 89 | 26 | 532 | 18 | 45 |  |
| # des gens qui n’ont pas voyagé | 219 | 11 | 174 | 28 | 82 | 255 |  |
| # des gens mal-placés | 81 | 89 | 26 | 28 | 18 | 45 | 209 |

Taux de mauvaise classification (TMC)= (81+89+26+28+18+45) / 1560 = 209/1560= 13.4%

Taux naturel d’erreur (TNE) = 22%

Le TMC < TNE 🡪 avec l’arbre on va se tromper sur 13.4% des gens alors qu’avant le modèle on se trompait sur 22% des gens. Donc, avec l’arbre on a un gain de 8.6% (22%-13.4%-8.6%) sur l’erreur c.-à-d. que l’arbre apprend des données.

1. Sachant que vous avez une plus grande importance à prédire les personnes qui ont voyagé, quel est le meilleur modèle parmi les 3 modèles (justifiez votre réponse) ?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Apprentissage (75 000 observations)*** | TMC | Lift cumulé 4ème décile | ROC | Sensitivité |
| Réseaux de neurones | 15% | 3,2 | 0.81 | 51% |
| Arbre de classification | 18% | 2,1 | 0.86 | 63% |
| Forêts aléatoires | 21% | 2,7 | 0.8 | 76% |
| ***Apprentissage (40 000 observations)*** | TMC | Lift cumulé 4ème décile | ROC | Sensitivité |
| Réseaux de neurones | 18% | 2.7 | 0.77 | 45% |
| Arbre de classification | 18% | 2,0 | 0.85 | 63% |
| Forêts aléatoires | 21% | 2,7 | 0.8 | 75% |

Dans les deux cas (75k observations et 40k observations), le taux de mauvaise classification (TMC) des trois modèles sont plus petits que le taux naturel d’erreur (TNE). Donc, les trois modèles font un bon travail en prédiction, mais cela ne veut pas dire qu’un modèle dont TMC est beaucoup plus petit que le TNE est nécessairement le meilleur modèle. Ici on cherche de bien prédire les 1 de target (les personnes qui vont voyager) ; donc, il faut porter une attention particulière a la sensibilité qui est un indicateur de la puissance du modèle en termes de prédiction des 1.

Les réseaux neurones (RN) a un TMC plus petit que les autres modèles dans des deux cas (75k observations et 40k observations), et son lift accumule est aussi plus grande qu’aux autre modèles mais ce modèle a une sensitivité assez basse, surtout pour les 40k observation ou sa sensibilité est plus base de 50%. Cela veut dire que le RN n’est pas très performant en prédisant les 1 de target. Donc, on rejette à utiliser le modèle RN pour la prédiction des personnes qui vont voyager cause de son mal performance en prédiction de sensibilité.

De l’autre cote, la sensibilité du foret aléatoire c’est le meilleur des autres dans les deux cas (75% et 76%), aussi, sa valeur de lift cumulé est plus grande que l’arbre de décision. Le ROC plus enlevé de l’arbre de décision montre que l’arbre est plus puissant en prédiction des 0 (pas des 1).

Aussi, il est important à noter qu’en tous les modèles, on est toujours exposé a sur-apprentissage mais en modèles d’ensemble on est moins exposé a sur-apprentissage. Foret aléatoire est un modèle d’ensemble.

Pour des raisons mentionnées ci-haut, je choisis le foret aléatoire pour prévoir les gens qui vont voyager.

**Partie 2 :**

Pour cet exercice, nous vous demandons de construire une forêt aléatoire comportant 2 arbres de classification. Chaque arbre aura pour complexité une profondeur maximale de 2 ainsi qu’un indice d’impureté de Gini. Pour la construction des arbres, nous vous demandons de choisir aléatoirement les variables à chaque embranchement.

On va construire 2 arbres. Pour construire chaque arbre je tirer aléatoirement 2 variables explicatives parmi les 5 variables explicative de fichier d’apprentissage. Pour la 2ère profondeur, je tirer encore une fois aléatoirement 2 variables explicatives parmi les 5 variables explicative de fichier d’apprentissage

**Arbre 1 :**

Étape 1 : Variables explicatives aléatoirement choisies de la 1ère profondeur : Provenance, Nombre de produit

|  |  |
| --- | --- |
| Variable | Gain informationnel |
| Provenance | 0.009 |
| Nombre de produit | 0.164 |

Gain informationnel de "Nombre de produit" est plus élevé que celui de "Provenance" 🡪 On commence la construction de notre arbre avec la variable "Nombre de produit".

Étape 2 : La moyenne de la variable "Nombre de produit" est 6.17, alors je divise cette variable en deux sous-catégories :

* "Nombre de produit" > 6
* "Nombre de produit" <= 6

Variables explicatives de la 2ère profondeur pour "Nombre de produit" > 6 : Provenance, Sexe

|  |  |
| --- | --- |
| Variable | Gain informationnel |
| Provenance | 0.019 |
| Sexe | 0.056 |

Gain informationnel de "Sexe" est plus élevé que celui de "Provenance" 🡪 On développe cette branche avec la variable "Sexe".

Étape 3 : Variables explicatives de la 2ère profondeur pour "Nombre de produit" <= 6 : Revenue, Sexe

La moyenne de la variable "Revenue" quand « Nombre de produit" <= 6 est 47812.375 , alors je divise cette variable en deux sous-catégories :

* "Revenue" > 47,812
* "Revenue" <= 47,812

|  |  |
| --- | --- |
| Variable | Gain informationnel |
| Revenue | 0.075 |
| Sexe | 0.008 |

Gain informationnel de " Revenue " est plus élevé que celui de "Sexe" 🡪 On développe cette branche avec la variable " Revenue".

Structure de l’arbre 1 :

A screen shot of a computer monitor

Description automatically generated

**Arbre 2 :**

Étape 1 : Variables explicatives aléatoirement choisies de la 1ère profondeur : Sexe, Revenue

La moyenne de la variable "Revenue" est 44939.5, alors je divise cette variable en deux sous-catégories :

* " Revenue" > 44939
* " Revenue" <= 44939

|  |  |
| --- | --- |
| Variable | Gain informationnel |
| Sexe | 0.061 |
| Revenue | 0.023 |

Gain informationnel de "Sexe" est plus élevé que celui de "Revenue" 🡪 On commence la construction de l’arbre avec la variable "Sexe".

Étape 2 : Variables explicatives de la 2ère profondeur pour "Sexe"= F : Provenance, Age

La moyenne de la variable "Age" quand Sexe=F est 25.7, alors je divise cette variable en deux sous-catégories :

* "Age" > 25
* "Age" <= 25

|  |  |
| --- | --- |
| Variable | Gain informationnel |
| Provenance | 0.078 |
| Age | 0.006 |

Gain informationnel de "Provenance" est plus élevé que celui de "Age" 🡪 On développe cette branche avec la variable "Age".

Étape 3 : Variables explicatives de la 2ère profondeur pour "Sexe"= H : Provenance, Revenue

La moyenne de la variable "Revenue" quand Sexe=H est 46309.7, alors je divise cette variable en deux sous-catégories :

* "Revenue" > 46309
* "Revenue" <= 46309

|  |  |
| --- | --- |
| Variable | Gain informationnel |
| Provenance | 0.138 |
| Revenue | 0.021 |

Gain informationnel de "Provenance" est plus élevé que celui de "Revenue" 🡪 On développe cette branche avec la variable "Provenance".

Structure de l’arbre 2 :

A screen shot of a computer monitor

Description automatically generated

Après la construction de la forêt aléatoire, veuillez calculer le taux moyen d’erreur OOB grâce aux fichier EXCEL data\_examen\_final\_foret\_alea sur Zone- Cours.

|  |  |
| --- | --- |
| Arbre | Taux de mauvaise classification |
| Arbre 1 | (4 sur 9) 44% |
| Arbre 2 | (6 sur 8) 75% |

OBB moyen = (44% + 75%) / 2 = 59.7%

Enfin, pour Marc-André, un jeune homme montréalais de 26 ayant 6 produits à son actif et des dépenses de 250 $ pour un revenu annuel de 64 000$, veuillez déterminer son niveau de fidélité avec l’organisation. Veuillez expliquer chacune de vos étapes pour la construction et la prédiction de Marc-André.

|  |  |  |  |
| --- | --- | --- | --- |
| Nom | Prédiction Arbre 1 | Prédiction Arbre 2 | Prédiction finale |
| Sexe= H  Provenance = Montréal  Nombre de produit = 6  Revenue = 64000 | Moyen | Moyen | Moyen |

Évaluation de l’arbre 1 :

Nombre de produit = 6 et Revenue = 64 000 > 47 812 🡪 prévision de l’arbre 1 = moyen

Évaluation de l’arbre 2 :

Sexe = H et Provenance = Montreal 🡪 prévision de l’arbre 2 = moyen

Évaluation finale est fait par vote majoritaire comme dans le modèle Bagging :

Évaluation de l’arbre 1= moyen et l’évaluation de l’arbre 2 = moyen 🡪 évaluation finale = moyen

**Partie 3 :**

Toujours étant analyste de données pour la même compagnie, l’organisation vous demande d’être dans les discussions en ce qui concerne les différentes stratégies d’acquisition. En effet, vous avez produit un modèle de réseau de neurones afin de prédire si une personne à une forte propension de devenir un client fidèle ou pas. Vous avez décidé de faire un réseau de neurones comprenant une couche d’entrée, une couche cachée qui elle contient deux neurones puis une couche de sortie.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nom | Taille | Nombre d’enfants | Nombre d’année d’expérience | Nombre de carte de crédit |
| Claire | 5.7 | 0 | 5 | 1 |
| Marc | 5.9 | 2 | 4 | 1 |
| Sébastien | 6.2 | 1 | 6 | 3 |

1. Veuillez représenter graphiquement l’architecture du réseau de neurones.

A screen shot of a computer

Description automatically generated

PV = poids assignés de neurone vert (neurone 1 de la couche cachée)

PO = poids assignés au neurone orange (neurone 2 de la couche cachée)

PR = Poids assignés au neurone rose (neurone 1 de la couche sortie)

1. Un de vos collègues de travail vous demandent pourquoi avoir utiliser dans ce modèle une fonction logistique et non linéaire. Veuillez argumenter en quelques lignes sur pourquoi la fonction d’activation logistique et la plus appropriée pour ce genre de problème.

Ici, le problème est un problème de classification, car on cherche à savoir si un client va devenir un client fidèle oui ou non. Donc, on cherche que le résultat de la fonction d’activation soit l’équivalent de probabilité (donc entre 0 et 1). Le résultat de la fonction logistique est toujours entre 0 et 1, tout comme la probabilité. Alors, le résultat de cette fonction va représenter la probabilité de notre événement.

Afin de donner une probabilité pour chacune de nos personnes, nous utilisons les poids suivants :

Couche d’entrée vers la couche cachée

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Taille | Nombre d’enfants | Nombre d’année d’expérience | Nombre de carte de crédit |
| 1er neurone | -0.2 | 0.55 | 0.2 | -0.02 |
| 2ème neurone | 0.15 | 0.57 | -0.1 | 0.4 |

Couche cachée vers la couche de sortie

|  |  |
| --- | --- |
| 1er neurone | -0.21 |
| 2ème neurone | 0.27 |

En prenant une fonction d’activation logistique et un point de coupure de 50% (0.5), veuillez donner pour Claire, Marc et Sébastien une prédiction si oui ou non la personne deviendra membre de l’organisation. En ce qui concerne l’explication des calculs, veuillez seulement ceux en rapport à la prédiction de Claire.

On normalise les données d’entre tout d’abord pour enlever l’effet d’échelle :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nom | Taille | Nombre d’enfants | Nombre d’année d’expérience | Nombre de carte de crédit |
| Claire | -0.93 | -1.00 | 0.00 | -0.58 |
| Marc | -0.13 | 1.00 | -1.00 | -0.58 |
| Sébastien | 1.06 | 0.00 | 1.00 | 1.15 |

Résultats de réseaux neurones pour Claire, Marc et Sébastien

|  |  |  |  |
| --- | --- | --- | --- |
| Nom | Claire | Marc | Sébastien |
| La probabilité de l’événement | 0.495 | 0.512 | 0.517 |
| Prédiction du RN | Non | Oui | Oui |

Explication des calculs pour Claire :

|  |  |
| --- | --- |
| Neurone 1 : | Claire |
| Fonction de combinaison | -0.93\*-0.2 + -1\*0.55+ 0\*0.2 + -0.58\*-0.02 = -0.35 |
| Fonction d'activation | 1/(1+EXP(-1\*-0.35)) = 0.41 |

|  |  |
| --- | --- |
| Neurone 2 : | Claire |
| Fonction de combinaison | -0.93\*0.15 + -1\*0.75+ 0\*-0.1 + -0.58\*0.4 = -1.1 |
| Fonction d'activation | 1/(1+EXP(-1\*-1.1)) = 0.25 |

|  |  |
| --- | --- |
| Neurone de couche de sortie : | Clair |
| Fonction de combinaison | 0.41\*-0.21 + 0.25\*0.27 = -0.02 |
| Fonction d'activation | 1/(1+EXP(-1\*0.02)) = 0.495 |
| Résultat de réseaux neurones | 0.495 < 0.5 🡪 résultat = Non |

**Partie 4 :**

À l’aide et seulement à l’aide du logiciel R, veuillez répondre aux questions suivantes.

1. Pouvez-vous déterminer l’âge moyen ainsi que la médiane concernant la balance du compte de notre clientèle ?

Age médian = 39 ans

Age moyen = 40.93 ans

1. Dans la base de données se trouve la variable y représentant si la personne à souscrit au dépôt direct. Pouvez-vous changer le nom de cette variable en ‘deposit’ ?

SVP, referez-vous au code R

1. Pour le reste de l’analyse, nous nous intéresserons seulement aux clients qui possèdent une balance strictement positive. Veuillez créer la table de données data\_bank\_deposit qui ne possèdera que les clients qui ont une balance strictement supérieure à 0.

SVP, referez-vous au code R

1. Avant de rentrer dans la prédiction, vous vous demandez à quoi ressemblerait un arbre de classification sur la table de données data\_bank\_deposit en prenant toutes les variables explicatives. Cet arbre de classification sera un arbre à base de la mesure de Gini, possédant une profondeur maximale de 5 ainsi qu’un nombre minimal de 50 observations dans les feuilles terminales.

Veuillez effecteur un tel arbre de classification et afficher le graphique de l’arbre en question.

A screenshot of a social media post

Description automatically generatedA picture containing clock

Description automatically generated

1. Pouvez interpréter la règle menant à la première feuille terminale ?

Caractéristiques du première feuille terminale :

poutcome=failure, other, unknown 32226 2072 no (0.93570409 0.06429591) \*

La feuille parente ce cette feuille terminale est une feuille de 2 branches :

* poutcome = failure, other, unknown (la feuille terminale qu’on etudie ici dans cette question)
* poutcome = success

Cette feuille terminale contient 32226 lignes dont 2072 lignes ont une valeur de deposit de « yes » et 32226-2072 (30152) lignes ont une valeur de deposit de « no ». C.-à-d. 2072 valeurs de deposit sont mal-classifiées.

Donc, la majorité dans cette feuille est avec « no » qui construit (30152/32226) 93.57% des observation de cette feuille.

Ainsi,

Donc, le jugement est « yes » avec une probabilité de 93.6% et le 6.4% (2072/32226) des valeurs deposit de cette feuille sont mal-classifiées.

1. À partir de cet arbre, pouvez citer les 3 variables explicatives ayant la plus grande importance dans la construction de l’arbre.

|  |  |  |
| --- | --- | --- |
| Variable | Rank | Importance de variable |
| duration | 1 | 1180.13 |
| poutcome | 2 | 790.26 |
| contact | 3 | 2.30 |

1. Vous décidez maintenant de passer à la construction d’un modèle de prédiction.

Veuillez déterminer le taux naturel d’erreur que nous possédons dans notre table de données data\_bank\_deposit.

Taux naturel d’erreur =12.6%

1. Veuillez construire trois modèles, boosting, bagging, et forêt aléatoire tous de 50 arbres et ayant une profondeur maximale de 4. Pour ce qui est de la forêt aléatoire, veuillez choisir un tirage aléatoire de variables de 5.

SVP, referez-vous au code R

1. En comparant le taux de mauvaise classification de tous les modèles, lequel est le plus performant ?

|  |  |
| --- | --- |
| Modèle | TMC |
| Bagging | 10.56% |
| Boosting | 9.8% |
| Forêt aléatoire | 10.28% |

Le taux de mauvaise classification des trois modèles sont plus petits que le taux naturel d’erreur. Cela veut dire que les trois modèles sont performants. Mais le taux de mauvaise classification du modèle Boosting est plus petit par rapport aux autres ; alors le modèle de Boosting est le plus performant des 3 modèles d’ensemble.

1. Vous avez un très fort intérêt à bien prédire les personnes qui vont prendre le dépôt direct car des frais seront associés lors de la prise de contact. Avec cette information, quel est le modèle que vous retenez.

Je dois donc m’assurer qu’il n’y a pas de sur-apprentissage dans mes modèles. Mais, vue que je n’ai pas de données de validation, je ne peux pas valider mon modèle. On sait qu’avec la méthode de foret aléatoire, on n’pas besoin de fichier d’apprentissage. Donc, bien de le TMC de foret aléatoire est plus élevé que celui de Boosting, je choisirais le modèle de foret aléatoire pour prédire (ici, on perde en performance mais on gagne en stabilité).

1. Dans ce projet, il y a un concept qui n’a pas été appliqué pour s’assurer que les modèles ne contiennent pas de sur-apprentissage. Veuillez expliquer en quoi consisterait ce concept.

Pour s’assurer qu’il n’y a pas de sur-apprentissage dans le modèle, il faut toujours comparer différentes mesures de performance du fichier apprentissage avec ceux du fichier validation. Ici, il manque le fichier de validation pour faire l’étape de validation (normalement dans notre table de données, on doit avoir une colonne indicatrice qui nous aide à séparer le fichier en deux groupes de l’apprentissage et validation). Le format de cette colonne est facteur et ces valeurs sont « validation » et « apprentissage ».

Aussi, deux points qui n’ont rien à voir avec le sur-apprentissage : ici le nombre des arbres sont paires pour tous les modèles. Alors que due à la nature de vote majoritaire de modèles de foret aléatoire et Bagging, il veut mieux choisir un nombre impair pour qu’on n’ait pas le choix aléatoire lors de la l’étape de validation. Aussi, on doit refaire le modèle avec plusieurs points de coupure pour trouver le meilleur modèle.