Projet Scoring : la cyber sécurité

Thomas Schuler
Pierre Deneux
Nicolas Moine

Master 2 ISF
Université Panthéon-Assas, Paris II
1 juillet 2022

Introduction

1. **Présentation des données**

1.1 Analyse qualitative de la base de données

1.2 Création de la base de données

1.3 Analyse quantitative du jeu de données

2. Présentation des modèles de Scoring

2.1 Le modèle de régression Logistique

2.2 Le modèle de Classification Naïve Bayésienne Gaussienne

2.3 Le modèle Support Vector Classifier

3. Résultats obtenus

3.1 Le critère de performance des modèles

3.2 Choix du modèle préférentiel

3.3 Présentation de la carte de score

4. Conclusion

Introduction

Dans le cadre de notre projet en scoring, nous avons voulu nous attarder aux sujets de cyber sécurité. Dans l'environnement actuel, tendu également par les conflits internationaux, l'intégralité des acteurs privés et publics se protègent des menaces de cyber sécurité.

En ce sens il y a deux défis majeurs auquels il faut répondre, apprendre à détecter ces menaces avant qu'elles ne fassent des dégâts et leur empêcher l'accès aux hardwares ainsi qu'aux softwares et évidemment éliminer les menaces au cas où elles s'infiltrent.

Nous allons ici nous attacher à la première partie du problème, c'est à dire la détection des menaces.
Sur ce sujet, de nombreuses recherches ont été menés, car lorsqu'on cherche à éliminer une menace il faut être capable d'anticiper les menaces existantes, mais aussi celles qui ne seraient pas connu. C'est donc pourquoi les bases se doivent d'être exhaustives et de lister toutes les caractéristiques communes à une cyber attaque.

Pour rappel, les sujets de cyber intrusion ont d'abord été abordé au début des années 1990 qui ont amené à une première base intitulée KDD-Cup 1999 qui a fait foi dans le domaine de la lutte contre la cyber intrusion pendant une dizaine d'année. Avec le développement rapide des outils et d'internet, il a fallu concevoir de nouvelles bases, si cette base a été ainsi challengée, c'est en 2015 qu'on trouve un réel successeur à cette base avec la UNSW-NB15.

C'est sur cette base que nous allons travailler afin d'établir un modèle de scoring qui nous permettrait de détecter facilement les menaces avec une lecture facile des caractéristiques d'une intrusion peu importe sa nature.

Si la base initiale fait 4M5 de ligne nous allons nous attacher à un extrait équilibré de cette base proposée par l'université de Sidney qui distingue déjà une base d'entraînement et une base de test et qui équilibre déjà le poids de trafic normal et de tentative d'intrusion.

Mots clefs : Cyber sécurité, Scoring, Modèle logistique, Modèle Gaussian Naive Bayes, Modèle SVM, Courbe de ROC, Critère d'AUC, Matrice de confusion, Carte de score…

1. Présentation des données

Pour nettoyer la base de données, nous commençons par regarder le nombre de valeur manquante. Comme attendues, les données étant générées artificiellement, sont complètes.

1.1 Analyse qualitative du jeu de données

Comme dit précédemment, notre base de données est tirée d'un jeu de données initiale de 4M5 lignes provenant de l'université de Sidney. Pour notre projet, utilisons une partie du jeu initial composée de 170 000 lignes. Pour la suite de notre étude, nous utilisons naturellement une base d'entraînement composée de 80 % de la base de données et une base de test, composée des 20 % restant. Elle est composée de plusieurs activités dites normales ainsi que des comportements d'attaque. Pour chaque opération, nous relevons plusieurs indicateurs mesurant l'activité sur le web. Nous constatons que la plupart de ces indicateurs sont des variables catégorielles.

1.2 Création de la base finale

Dans chaque étude, il est primordial de travailler sur un jeu de données homogène, complet et cohérent. Pour nettoyer la base de données, nous commençons par regarder le nombre de valeur manquante. Comme attendue, les données étant générées artificiellement, sont complètes. De ce fait, nous pouvons, dans un premier temps, étudier nos données et observer quelques corrélations potentielles entre les variables.

1.3 Analyse quantitative du jeu de données

La première étape consiste à établir une sélection puis d'étudier les corrélations entre les variables. Nous retirons toutes les variables ayant une corrélation avec d'autres variables supérieures à 80 %. L'objectif est d'éviter des biais d'endogénéité dans notre modèle dont la cause proviendrait de la corrélation entre les variables. Pour cela, nous allons utiliser différentes méthodes. Pour les variables continues ou discrètes, nous utilisons la corrélation classique de

Spearman. Pour les variables catégorielles, nous utilisons un test de chi-square ainsi qu'un test ANOVA. De plus, nous créons des groupes de valeurs pour les variables continues dans l'idée d'éviter une disparité entre les données et de rassembler les variables par classes.
En regardant un peu plus en profondeur les données, nous remarquons qu'il n'y a pas de variables discrimantes seules.

Dans la section suivante, nous expliciterons les modèles utiliser pour répondre au problème de classification établi dans ce projet.

2. Présentation des modèles de Scoring

Lorsque nous sommes confrontés à des problèmes de scoring, nous avons à notre disposition une large palette de modèle. Dans les sections suivantes, nous optons pour les modèles de régression logistique, le modèle de classification naïve bayésienne et le modèle Support Vector Classifier (SVC). Avant de commencer la modélisation, il est nécessaire de séparer notre jeu de données de la manière suivante :
- 80 % de la base finale représentant notre base d'entraînement,
- 20 % de la base finale représentant notre base de test.

Dans un premier temps, commençons, dans la section suivante, par la présentation des modèles.

2.1 Le modèle de Régression Logistique

L'objectif de ce projet de répondre à une problématique de classification. Ainsi, nous décidons d'entamer notre étude par une méthode de classification binaire classique : la régression logistique. Dans ce modèle, si la probabilité estimée est supérieure à 50 %, alors le modèle prédit que l'observation appartient à la classe d'étiquette « 1 », appelée classe positive. En revanche, si la probabilité estimée est inférieure à 50 %, alors le modèle prédira que l'observation appartient à la classe d'étiquette « 0 », appelée classe négative.

Dans un modèle de régression linéaire, on calcule la somme pondérée des caractéristiques des entrées, à laquelle on ajoute un terme constant. Cette quantité est reliée à $\mathbb{E}[y|x]$ par une fonction de lien g. Ainsi, nous obtenons la relation suivante :

$$g(\mathbb{E}[y|x]) = f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

Dans le cadre de la régression logistique, g est choisie comme étant l'inverse de la fonction *sigmoïde* :

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Cette fonction permet de renvoyer une valeur comprise entre 0 et 1 pour que la probabilité y soit égal à 1, et ce à partir de $\beta_0 + \sum_{j=1}^{p} \beta_j x_j$.

Ainsi, l'équation ci-dessus devient :

$$\mathbb{E}[y|x] = \mathbb{P}(y = 1) = \sigma(x^T \beta) = \sigma(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)$$

Le modèle de régression logistique va alors estimer la probabilité $\hat{p} = \sigma(x^T \hat{\beta})$. Si $\hat{p} < 0{,}5$ alors $\hat{y} = 0$ et notre observation x appartient à la classe négative. Inversement, Si $\hat{p} > 0{,}5$ alors $\hat{y} = 1$ et notre observation x appartient à la classe positive.

A ce stade nous savons donc comment notre modèle de régression logistique va estimer les probabilités et effectuer ses prédictions. Mais nous ne savons comment ce dernier est entraîné. Pour entraîner le modèle, il est commun d'utiliser les régressions de Lasso et de Ridge qui rajoute des contraintes supplémentaires aux coefficients de pondération du modèle afin d'éviter les problèmes de surajustement. Pour cela, les méthodes évoquées ci-dessus ont pour objectif d'améliorer la qualité de la prédiction en diminuant significativement la variance du modèle tout en acceptant une légère augmentation du biais.

2.2 Le modèle de Classification Naïve Bayésienne gaussienne

Cette méthode de classification est un algorithme d'apprentissage supervisé qui, d'un point de vue historique, était utilisée pour la classification de documents et l'élaboration de filtres anti-spam. Parmi ces atouts les plus significatifs, cette méthode possède un apprentissage rapide qui ne nécessite pas un grand volume de données et son exécution est extrêmement rapide au regard d'autre méthodes plus complexes. Dans notre cas présent, nous utilisons cette méthode en postulant que nos données sont distribuées normalement. Malgré la forte hypothèse de variables indépendantes, cette méthode obtient des résultats performants dans de nombreux domaines.

Dans un premier temps, nous développons le modèle probabiliste pour le classifieur via un modèle conditionnel tel que :

$$p(C|F_1, \dots, F_n)$$

Où C est une variable de classe dépendante dont les instances ou classes sont peu nombreuses, conditionnées par plusieurs variables caractéristiques $F_1, \dots, F_n$.
Ce modèle étant développée par le théorème de Bayses, nous obtenons :

$$p(C|F_1, \dots, F_n) = \frac{p(C)p(F_1, \dots, F_n|C)}{p(F_1, \dots, F_n)}$$

Le classificateur bayésien naïf couple ce modèle avec une règle de décision. Elle consiste à choisir l'hypothèse la plus probable, la règle du maximum a posteriori. Le classificateur correspondant à cette règle est la fonction classifieur suivante :

$$classifieur(f_1, \dots, f_n) = arg_c \max p\,(C = c) = \prod_i^n p(F_i = f_i | C = c)$$

2.3 Le modèle Support Vector Classifier (SVM)

Le SVM appartient à la catégorie des classifieurs linéaires. Il a pour objectif de trouver la frontière entre les différentes catégories. Il s'agit d'un des modèles les plus prisés en matière d'apprentissage automatique, du fait de sa puissance, de sa polyvalence et de son adaptabilité à des jeux de données complexes.

Dans notre situation, notre prédiction est binaire. Ainsi, chaque observation sera représentée par un point et placée dans le plan, selon la région à laquelle elle appartient. Le but du SVM est de déterminer la frontière qui délimite notre plan en deux régions : l'une contenant les sinistres frauduleux, et l'autre contenant les sinistres légitimes. La difficulté est de trouver la forme optimale de notre frontière qui devra maximiser sa distance avec les points du jeu d'entraînement (x_j). C'est à dire que notre frontière devra se situer aussi loin que possible dès que $y_i = 1$ dès que $y_i = 0$. Ainsi, on appelle "vecteurs supports" les points de chaque catégorie qui sont les plus proches de la frontière.

L'objectif est de déterminer la position optimale d'un hyperplan affine le séparateur. Pour cela, il est nécessaire de maximiser la marge entre deux catégories de points. Le problème se formalise de la manière suivante :

$$(\hat{w}, \hat{b}, \hat{\xi_\iota}) = arg_{w,b,\xi} \min \frac{1}{2} = w^T w + C \sum_{i=1}^{n} \xi_i$$

Sous les contraintes $f(x) = \begin{cases} y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$

où ξ représente le vecteur des distances à lesquelles on "permet" aux observations de dépasser la frontière de la marge correcte, et C est un hyperparamètre de régularisation que l'on pourra calibrer.

3. Résultats obtenus

Dans cette dernière partie, l'objectif sera de développer nos critères de performance des modèles pour converger vers le choix du modèle optimal et ainsi, inférer sur une carte de score dans l'objectif de répondre à la problématique posée.

3.1 Le critère de performance des modèles

Afin de pouvoir comparer les différents classifieurs obtenus par ces méthodes, nous allons le faire à l'aide d'une matrice de confusion : permet de mesurer la qualité d'un système de classification. Notre variable à prédire pouvant prendre la valeur 0 ou 1, cette matrice est tout à fait adaptée à notre problème. Elle prendra la forme suivante :

	Prédicted 0	**Predicted 1**
Actual 0	True -	False +
Actual 1	False -	True +

Le cas *True Positive* (TP) correspond au cas où on l'on prédit 1 pour notre variable à prédire Label, traduisant que l'envoi est dangereux lorsque c'est réellement le cas. Le cas *True Negative* (TN) correspond au cas où l'on prédirait 0 pour cette variable, également à raison. Pour le cas de *False Negative* (FN), cela correspond au cas où l'on prédirait 0, l'envoie n'est pas dangereux, alors qu'il l'est. Enfin, le cas *False Positive* (FP) représente le cas où l'on prédit 1, à tort.

Le but à travers nos méthodes est d'obtenir un classifieur maximisant le nombre de valeurs bien prédites, c'est à dire True Positive et True Negative.

Pour chaque classifieur obtenu, nous allons calculer sa matrice de confusion, et à partir de cette dernière nous allons mesurer la performance de notre modèle à l'aide des métriques suivantes :

- $Accuracy : \frac{TN+TP}{TN+FP+FN+TP}$
- $Precision : \frac{TP}{TP+FP}$
- $Accuracy : \frac{TP}{FP+FN}$
- $F1\ Score : 2.\frac{Precision.Recall}{Precision+Recall}$

Un autre outil souvent utilisé pour l'analyse de classifieur binaire est la courbe du ROC : Receiver Operating Characteristic. Cette courbe mesure l'efficacité du récepteur : on représente cette courbe avec en abscisse le taux de FP, et en ordonnée le taux de TP. De cette manière, afin de comparer les classifieurs entre eux, on pourra le faire en comparant l'aire sous cette courbe : plus le classifieur est bon en ce sens, plus son aire sous la courbe se rapprochera de 1.

Nous utiliserons cette courbe dans l'optique de conclure sur le meilleur modèle dans la section suivante.

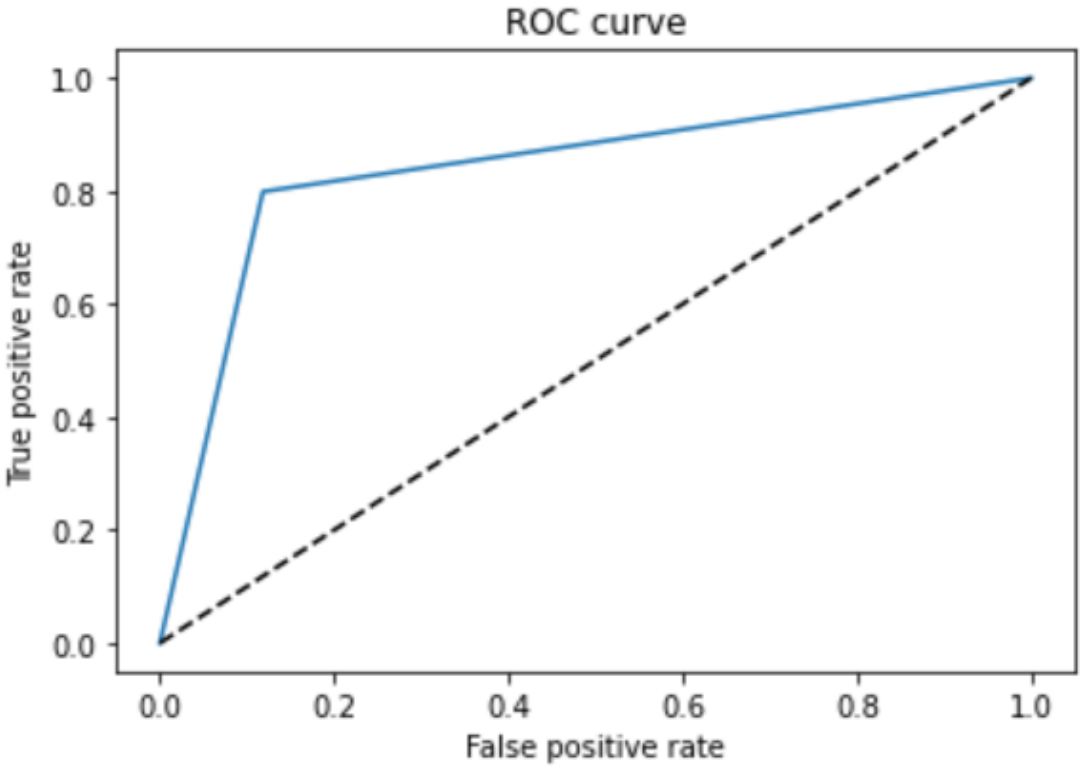
3.2 Choix du modèle préférentiel

Comme évoqué précédemment, il existe différents critères dans le choix du modèle. Nous choisissons de nous baser sur l'AUC (Area Under Curve) qui traduit l'air sous les différentes

courbes de ROC émanent des méthodologies sous-jacentes. Nous obtenons les résultats suivants :

Modèles	Logistique	Classification Naïve Bayésienne	SCV
AUC	0,84	0,80	0,79

Ainsi, nous définissons notre meilleur modèle, dans le cadre de notre étude, comme le modèle de régression logistique dont nous pouvons représenter sa courbe de ROC.



Nous pouvons ainsi construire une carte de score, ce que nous verrons dans la section suivante.

3.3 Présentation de la carte de score

Après avoir défini le meilleur modèle pour classifier nos observations, nous pouvons construire une carte de score. Elle nous permet d'ordonner nos observations en fonction de ses critères et du modèle utilisée et ainsi sensibiliser les utilisateurs aux menaces et ainsi se prémunir d'une cyber attaque. Nous pouvons représenter un échantillon de la carte de score obtenu :

Nom initial	Nom	Score - Final
is_sm_ips_ports	is_sm_ips_ports	-15
state	state : CON	-7
state	state : FIN	2
state	state : INT	21
state	state : REQ	16

4. Conclusion

Finalement, nous retenons deux choses, lors du travail de détection de cyber attaque, il est évident, que le traitement ne se fera pas à l'aide d'une carte de score, mais bien par ordinateur pour des soucis d'efficacité, il faut comprendre que l'enjeu de la cyber sécurité dans le monde est un enjeu majeur qui recense des millions de tentative sur les réseaux de toute nature et avec une multitude de cible, et donc qu'un traitement manuel n'est pas envisageable.

Néanmoins, dans un contexte d'entreprise à faibles budgets, il serait envisageable de proposer l'entreprise de bloquer certains paramètres observés dans la carte de score qui garantissent une nette attitude propice à un danger de cyber attaque.

On note également que si le modèle n'est pas énormément avancé en SVC (support de vecteur machine) ou en Gaussian Naive bayes les performances ne sont pas nécessairement meilleur, (84 % pour notre régression logistique contre 75 % pour les deux autres modèles). Nous conseillons donc très sérieusement aux professionnels modélisant des enjeux de nature similaire, pas par contexte, mais structure de donnée, de s'attacher à définir des modèles plus simples, plus lisibles et pouvant garantir des résultats plus avancés. Pour rappel, on parle par exemple, en risque de crédit de bons modèles, au-dessus de 80 % (Source : Deloitte - *Credit scoring Case study in data analytics*).