École Centrale Casablanca

Rapport

Devoir Machine Learning

Élèves :
Aissaoui Zineb
Amraoui Nouhaila
Boulaghla Mohammed Taha
El Fourati Ilyass
El Omari Safae
Essalmi Meryeme

28 janvier 2024

Table des matières

1 Introduction

2 Feature engineering

Dans le cadre de notre étude de cas, nous disposons de deux ensembles de données distincts. Le premier ensemble de données concerne les enchérisseurs et englobe diverses informations les concernant, telles que leur identifiant, leur compte de paiement, et leur adresse. Cet ensemble forme notre jeu de données d'entraînement, sur lequel nous concentrons nos efforts pour améliorer les performances de notre modèle. Nous aspirons à affiner et enrichir ces données afin d'obtenir des résultats optimaux dans la classification des enchères humaines ou automatisées. Le deuxième ensemble de données, quant à lui, concerne les enchères et compte pas moins de 7,6 millions d'entrées, toutes relatives à différentes enchères. Ces données représentent une dimension cruciale de notre étude, offrant une perspective détaillée sur les caractéristiques des enchères que nous cherchons à classifier.

Dans cette section cruciale de notre étude, nous abordons la création de caractéristiques, une étape essentielle pour enrichir notre ensemble de données d'entraînement. Notre approche repose sur l'exploitation du deuxième ensemble de données. En analysant ces données, nous avons généré de nouvelles colonnes spécifiques à chaque enchérisseur, lesquelles ont été judicieusement intégrées à notre ensemble d'entraînement initial. Cette démarche vise à renforcer la performance de notre modèle en lui fournissant des informations supplémentaires et pertinentes pour la classification des enchères.

2.1 Description des données

2.1.1 Ensemble de données d'entraînement

Cet ensemble de données est composé des colonnes suivantes :
— **bidder_id** - Identifiant unique d'un enchérisseur.
— **payment_account** - Compte de paiement associé à un enchérisseur. Ces informations sont obscurcies pour protéger la vie privée.
— **address** - Adresse postale d'un enchérisseur. Ces informations sont obscurcies pour protéger la vie privée.
— **outcome** - Étiquette d'un enchérisseur indiquant s'il s'agit ou non d'un robot. La valeur 1.0 indique un robot, tandis que la valeur 0.0 indique un être humain.

2.1.2 Ensemble de données des enchères

Cet ensemble de données est composé des colonnes suivantes :
— **bid_id** - Identifiant unique pour cette enchère
— **bidder_id** - Identifiant unique d'un enchérisseur.
— **auction** - Identifiant unique d'une enchère
— **merchandise** - La catégorie de la campagne du site d'enchères.
— **device** - Modèle de téléphone d'un visiteur
— **time** - Heure à laquelle l'enchère est effectuée (transformée pour protéger la vie privée).
— **country** - Le pays auquel l'adresse IP appartient
— **ip** - Adresse IP d'un enchérisseur (obscurcie pour protéger la vie privée).

— **url** - URL à partir de laquelle l'enchérisseur a été référé (obscurcie pour protéger la vie privée).

2.2 Création de features

Pour améliorer la représentation des enchérisseurs dans notre ensemble de données, nous avons introduit plusieurs colonnes, chacune apportant une dimension unique à l'analyse. Ces colonnes ont été créées en se basant sur des caractéristiques spécifiques des enchères effectuées par chaque utilisateur. Les sous-sections suivantes détaillent ces nouvelles colonnes et expliquent leur rôle dans le processus de classification.

2.2.1 Nombre d'éléments uniques

Nous avons enrichi notre ensemble de données en ajoutant des colonnes qui reflètent les caractéristiques uniques de chaque enchérisseur, en utilisant la fonction **groupby().nunique()** sur nos données d'enchères. Ces colonnes ont été fusionnées avec notre ensemble d'entraînement initial, offrant ainsi des informations supplémentaires et pertinentes pour notre modèle de classification.

Cependant, la fusion a révélé des valeurs manquantes dans la dataframe résultante. Ces valeurs manquantes concernent les enchérisseurs dont les identifiants (**bidder_id**) ne correspondent à aucun enregistrement dans les données d'enchères (**bids**). Nous avons interprété ces cas comme des enchérisseurs n'ayant jamais participé à une enchère, et avons remplacé les valeurs manquantes par des **0** pour maintenir l'intégrité des données.

Cette approche permet à notre modèle de saisir des tendances spécifiques à chaque enchérisseur, révélant des comportements atypiques ou automatisés. Par exemple, un enchérisseur présentant un nombre élevé d'actions uniques pourrait être identifié comme potentiellement robotique. En intégrant ces caractéristiques distinctives, notre modèle est mieux équipé pour discerner précisément entre les enchères humaines et robotiques, renforçant ainsi la fiabilité de notre système de classification.

2.2.2 Écart temporel

Une caractéristique cruciale que nous avons intégrée dans notre ensemble de données est l'écart temporel entre chaque paire d'enchères consécutives pour chaque enchérisseur, fournissant ainsi des indications sur la régularité et le rythme de leurs activités d'enchères. Toutefois, nous avons pris en compte le fait que le calcul de cette mesure pour la première enchère de chaque enchérisseur peut engendrer des valeurs manquantes. Consciemment, nous avons exclu ces valeurs dépourvues de signification interprétable afin de garantir la cohérence de nos résultats. Une fois l'écart temporel calculé, nous avons employé la fonction **groupby().describe()** sur la colonne dont nous avons calculé l'écart temporel pour dériver des caractéristiques essentielles. Ces caractéristiques, comprenant le **maximum**, le **minimum**, la **moyenne**, l'**écart-type**, et les **quartiles** de l'écart temporel, ont été extraites pour chaque enchérisseur. Elles offrent une compréhension approfondie de la variabilité des intervalles temporels entre les enchères, renforçant ainsi la capacité de notre modèle à capturer des schémas temporels significatifs dans le comportement des enchérisseurs. Cette approche contribue de manière significative à une classification plus précise des enchères humaines et robotiques. Ainsi que nous avons ajouté une colonne où nous calculons le **nombre de fois où la différence entre deux enchères est nulle**.

En effet, nous avons remarqué qu'il y a un grand nombre d'enchérisseur où la différence entre deux enchère est nulle, ce qui est presque impossible pour un être humain normal mais une chose qui est ordinaire pour un robot. Donc ça sera utile d'ajouter une colonne pour voir si un enchérisseur à un temps de difference nulle entre deux enchères ou une colonne combien de fois nous avons un temps de différence nulle.

Finalement, nous avons fusionné ces caractéristiques avec notre ensemble d'entraînement initial, enrichissant ainsi nos données. Pour gérer les valeurs manquantes, nous avons adopté une approche différente en remplaçant les valeurs manquantes dans notre ensemble d'entraînement par 0, tandis que les caractéristiques liées à l'écart temporel ont été remplacées par leurs médianes respectives. Cette stratégie vise à assurer la cohérence des données tout en préservant la signification des caractéristiques temporelles dans notre modèle. Dans l'ensemble, ces ajustements méthodologiques renforcent la robustesse de notre modèle en capturant efficacement les schémas temporels et les comportements spécifiques des enchérisseurs, contribuant ainsi à une classification plus précise des enchères humaines et robotiques

2.2.3 Premier et dernier enchère

Nous avons ajouté une colonne significative, qui analyse la fréquence à laquelle un enchérisseur se positionne en première ou en dernière place au cours des enchères. Cette caractéristique permet d'appréhender le comportement compétitif de chaque participant par rapport aux autres. En examinant combien de fois un enchérisseur se trouve en tête ou en queue d'une enchère, nous obtenons un aperçu immédiat de sa stratégie et de son niveau d'agressivité ou de réflexion. Cette dimension de l'analyse vise à enrichir notre ensemble de données d'entraînement, fournissant des informations spécifiques sur les habitudes de participation de chaque enchérisseur et renforçant ainsi la capacité de notre modèle à différencier avec précision entre les comportements humains et automatisés.

2.2.4 Nombre maximal d'enchères dans une enchère

Nous avons introduit une nouvelle colonne qui représente le nombre maximal d'enchères réalisées par chaque enchérisseur au cours d'une enchère donnée. Cette caractéristique revêt une importance particulière, car elle permet de visualiser la stratégie d'enchère de chaque participant. Étant donné que les robots aspirent à remporter l'enchère sans se retirer, ils tendent à réaliser un nombre accru d'enchères. Ainsi, pour chaque enchérisseur, nous sélectionnons l'enchère où il a effectué le plus grand nombre d'enchères, reflétant ainsi son niveau d'activité et sa persévérance dans le processus d'enchère. De plus, nous avons également inclus des colonnes spécifiques qui calculent le nombre d'enchères effectuées par chaque enchérisseur pendant la première et la deuxième moitié d'une enchère, offrant ainsi des insights précieux sur le timing de leurs participations au cours du processus d'enchère.

2.2.5 Nombre maximal d'enchères sur l'ensemble des appareils

Au cours de cette étape, nous avons déterminé le nombre maximal d'enchères effectuées par un enchérisseur à l'aide d'un appareil spécifique, ainsi que la fréquence maximale d'utilisation de cet appareil, c'est-à-dire le nombre maximum d'enchères où un appareil particulier a été employé.

2.2.6 Quelques caractéristiques intuitives

Nous avons élaboré des caractéristiques intuitives et significatives. Par exemple, la caractéristique "Nombre d'enchères par enchère" (bids_per_auct) a été créée de manière à offrir une perspective plus pertinente que la simple considération des variables "Nombre d'enchères" (num_bids) et "Nombre d'enchères" (num_auct) individuellement. Cette approche consiste à normaliser le nombre d'enchères effectuées par un enchérisseur par rapport au nombre total d'enchères auxquelles il a participé, fournissant ainsi une mesure plus nuancée de son activité et de son engagement. Une logique similaire a été appliquée à d'autres caractéristiques, renforçant ainsi la pertinence et la signification de notre ensemble de données dans le cadre de l'analyse des comportements d'enchérisseurs.

2.2.7 Les valeurs abérantes

Nous avons effectué une analyse des valeurs aberrantes et constaté **la présence de seulement 7 robots** ayant un nombre d'enchères inférieur à 100. Étant donné que notre ensemble de données ne compte pas un grand nombre d'enchérisseurs robots, nous avons décidé de supprimer uniquement les robots ayant un **num_bids égal à 1**.

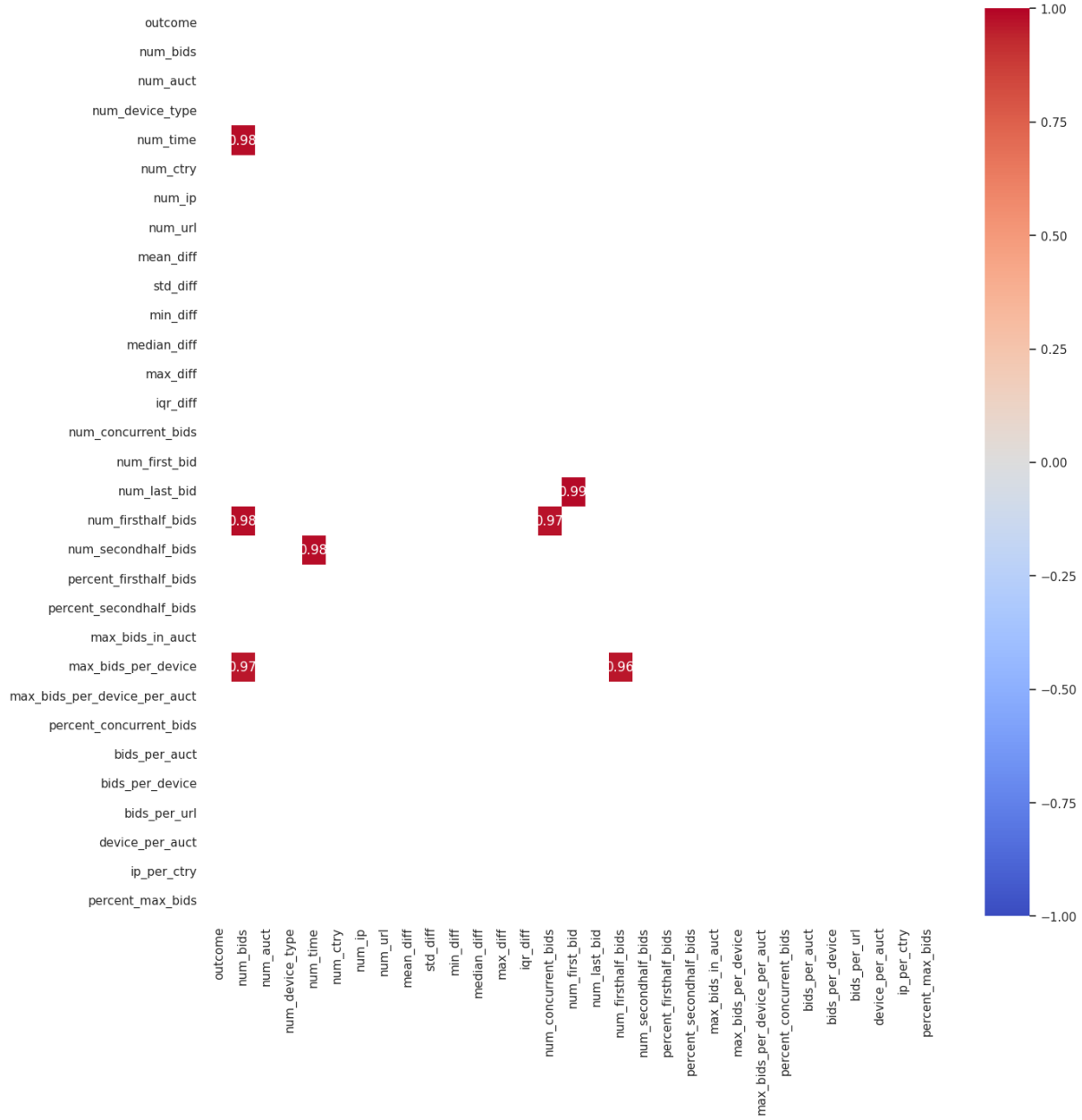
2.2.8 Conclusion

Afin de simplifier la réutilisation, nous avons élaboré des fonctions qui accomplissent toutes les étapes préalablement effectuées. Cette approche facilite grandement la tâche en cas de besoin de répéter ces mêmes étapes sur de nouvelles données.

2.3 Sélection des features signifiants

Dans le cadre de ce projet visant à détecter les enchérisseurs robots, la sélection des caractéristiques appropriées est essentielle pour garantir la précision et la robustesse du modèle. Pour ce faire, nous avons suivi une approche méthodique basée sur les critères suivants :
- Nous avons procédé à **l'élimination des caractéristiques redondantes**, celles qui fournissent des informations similaires, en calculant les corrélations entre les features extraites. Nous avons défini un seuil de 96% au-delà duquel les variables sont considérées comme très corrélées et doivent être éliminées.

Ci-dessous, vous trouverez la matrice de corrélation qui illustre les variables corrélées avec un seuil de plus de 96% :



Cette approche de suppression des caractéristiques redondantes nous permet de simplifier le modèle et d'éliminer tout bruit potentiel qui pourrait compromettre la performance de notre modèle de machine learning.

Après avoir identifié les caractéristiques fortement corrélées devant être supprimées, nous avons utilisé une méthode de densité pour vérifier qu'elles n'apportaient pas d'informations supplémentaires au modèle. Cette vérification nous a permis de confirmer les features redondantes. En conséquence, nous les avons éliminées du jeu de données. Ce processus garantit que seules les caractéristiques les plus informatives sont conservées, optimisant ainsi les performances du modèle tout en réduisant la complexité.

- La deuxième méthode vise à visualiser les différences de **distribution des caractéristiques** entre les enchérisseurs humains et les enchérisseurs robots. L'idée principale est d'utiliser des graphiques de densité pour représenter la répartition des valeurs de chaque

caractéristique, en distinguant les enchérisseurs humains des enchérisseurs robots. En comparant ces courbes, on peut observer visuellement les différences dans les distributions, ce qui peut aider à identifier les caractéristiques les plus discriminantes pour distinguer les enchérisseurs humains des enchérisseurs robots. Cette approche offre ainsi un aperçu intuitif des caractéristiques qui pourraient être importantes pour la modélisation et la détection des enchérisseurs robots.
Enfin, nous avons éliminé les caractéristiques suivantes : 'num_firsthalf_bids', 'num_time', 'num_first_bid'.

2.4 SMOTE

Après une analyse de la répartition de la donnée selon les deux classes de classification des enchères entre humains et automatisées, nous avons pris en considération le déséquilibre entre les deux classes, en particulier la rareté des enchères automatisées dans notre ensemble de données. Pour surmonter ce défi, nous avons opté pour l'utilisation de la technique SMOTE (Synthetic Minority Over-sampling Technique). Cette approche consiste à générer de manière synthétique des exemples supplémentaires de la classe minoritaire, dans notre cas, les enchères automatisées. En identifiant des instances individuelles de cette classe et en analysant leurs voisins les plus proches dans l'espace des caractéristiques, SMOTE crée de nouveaux exemples synthétiques en interpolant linéairement les valeurs des caractéristiques. Ces exemples synthétiques sont ensuite intégrés à notre ensemble de données d'entraînement, permettant ainsi d'équilibrer les proportions entre les classes. En utilisant SMOTE, notre objectif est d'améliorer la capacité du modèle à généraliser et à détecter efficacement les caractéristiques distinctives des enchères automatisées, contribuant ainsi à une meilleure performance de la classification.

3 Analyse exploratoire

3.1 Analyse Comparative des Moyennes des Caractéristiques entre les Enchérisseurs Humains et les Robots

Dans le cadre de l'analyse exploratoire des données, une étape cruciale consiste à comparer les moyennes des valeurs caractéristiques entre les deux classes distinctes, à savoir les enchères faites par des robots et celles effectuées par des humains. Cette comparaison vise à mettre en évidence les disparités ou tendances distinctes qui pourraient influencer la prédiction du modèle. Lors de cette étape, une observation notable a émergé, révélant un écart significatif entre les moyennes des caractéristiques pour les deux classes. Les valeurs moyennes divergentes suggèrent des différences intrinsèques dans les comportements d'enchères entre les enchérisseurs humains et les robots. Ces divergences pourraient potentiellement servir de signaux significatifs pour le modèle prédictif, soulignant l'importance de ces caractéristiques dans la distinction entre les deux catégories d'enchérisseurs. Cette constatation initiale fournit une base solide pour une analyse plus approfondie, mettant en lumière les aspects spécifiques des données qui peuvent être exploités pour une classification précise et fiable.

3.2 Analyse des Préférences d'Enchères par Catégorie et par pays entre les Enchérisseurs Humains et les Robots

Les occurrences des catégories Dans cette phase de l'analyse exploratoire des données, nous avons entrepris une comparaison approfondie des préférences d'enchères entre les enchérisseurs humains et les robots en examinant les catégories de marchandises auxquelles ils participent. En fusionnant les données d'enchères avec les informations d'enchérisseurs, nous avons pu identifier les principales marchandises pour lesquelles les deux groupes manifestent un intérêt. Un constat frappant a émergé de cette comparaison, révélant que les trois principales catégories d'articles suscitant l'intérêt des enchérisseurs humains et robots sont identiques, à savoir les articles de sport, les appareils mobiles et les bijoux. Par contre, des différences significatives ont été observées dans certaines catégories spécifiques, telles que les biens pour la maison et les pièces automobiles, qui ne sont pas proposées aux enchères par les robots. Bien que ces catégories représentent une petite proportion des enchères humaines, ces distinctions soulignent des comportements d'enchères distincts entre les deux groupes. Forts de ces constats, nous prenons la décision stratégique de ne pas inclure les variables catégorielles dans notre modèle, mettant en lumière l'importance de comprendre les préférences spécifiques à chaque groupe d'enchérisseurs pour une modélisation précise.

Les occurrences des pays

Dans cette phase de l'exploration des données, nous avons examiné de près la répartition géographique des enchères, en mettant en lumière les principaux pays où les enchérisseurs humains et les robots sont les plus actifs. Pour visualiser ces tendances, nous avons utilisé un graphique à barres divisé en deux parties : l'une représentant les principaux pays pour les enchérisseurs humains et l'autre pour les enchérisseurs robots. Les résultats ont révélé des similitudes dans les premières positions des principaux pays entre les deux groupes, avec une forte activité observée en Inde (in), au Bangladesh (bd), en Indonésie (id), aux États-Unis (us), en Afrique du Sud (za), au Kenya (ke), en Malaisie (my) et au Nigeria (ng). Cependant, des différences spécifiques ont également été notées, soulignant des préférences géographiques distinctes. Par exemple, les enchérisseurs robots ont également montré une forte activité en Thaïlande (th) et en Israël (il), tandis que les enchérisseurs humains étaient plus actifs au Bangladesh et au Kenya. Ces observations renforcent l'idée que la localisation géographique peut jouer un rôle crucial dans les comportements d'enchères, nécessitant une prise en compte minutieuse lors de la modélisation prédictive.

3.3 Comparaison d'autres Caractéristiques d'Enchères entre les Enchérisseurs Humains et les Robots

Dans cette étape de l'analyse, nous avons visualisé la différence entre les enchères réalisées par les enchérisseurs humains et les robots, en mettant en évidence plusieurs caractéristiques clés. Le graphique à barres présente les variations entre les deux groupes pour des caractéristiques spécifiques telles que le pourcentage d'enchères concurrentes, le nombre moyen d'enchères par auction, le nombre moyen d'enchères par appareil, le nombre moyen d'enchères par URL, le nombre moyen de types d'appareils par enchère, et le nombre moyen d'adresses IP par pays, et le pourcentage du nombre maximal d'enchères par appareil.

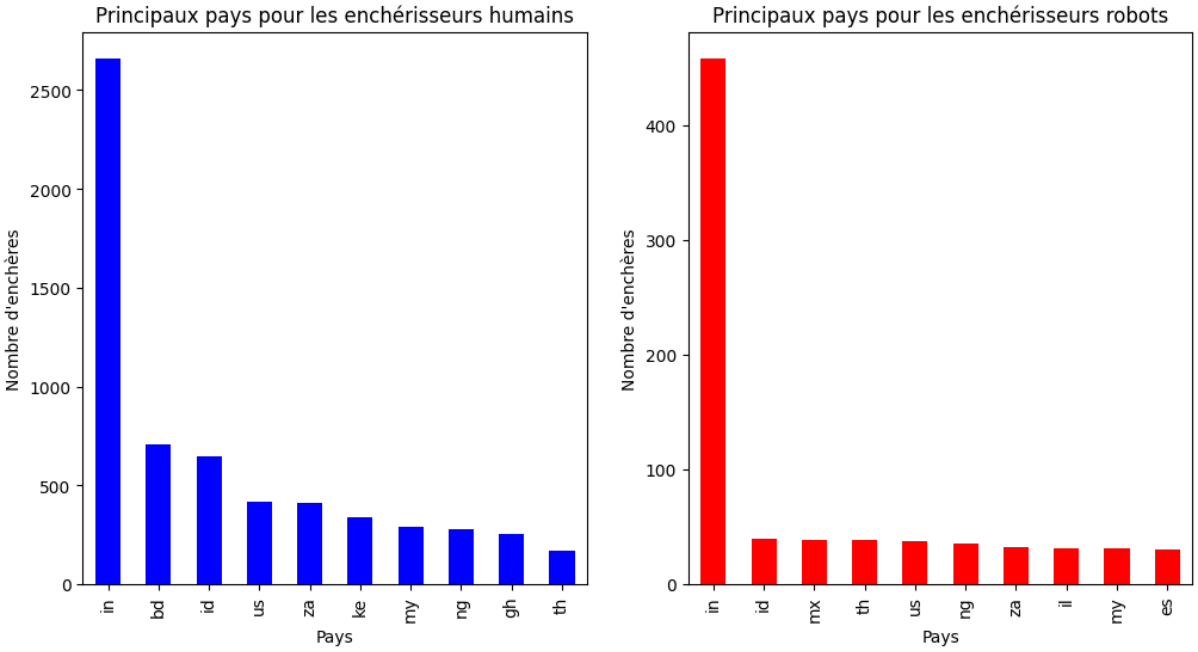


FIGURE 1 – Analyse Comparative des Principaux Pays d'Enchères entre les Enchérisseurs Humains et les Robots

Les résultats mettent en évidence des différences significatives entre les enchérisseurs humains et les robots dans ces caractéristiques, fournissant des insights cruciaux pour la modélisation prédictive. Par exemple, nous observons que les enchérisseurs robots ont tendance à avoir un pourcentage plus élevé d'enchères concurrentes,ainsi qu'ils ont plus tendances à faire plus d'enchères par auction, ils ont plus tendances à faire des enhères en utilisant des appareils différentes et des urls différentes. D'autres part, ils utilisent des adresses IP plus diversifié par chaque pays. Pour la feature 'device_per_auction' malgré que les valeurs moyennes sont presque égaux mais la variabilité de cette moyenne (avec un niveau de confiance de 95%) est trés grande pour les robot par rapport aux humains.

Finalement on clonclut que ces features d'avoir une compréhension plus approfondie des comportement des robots et des humain en enchère.

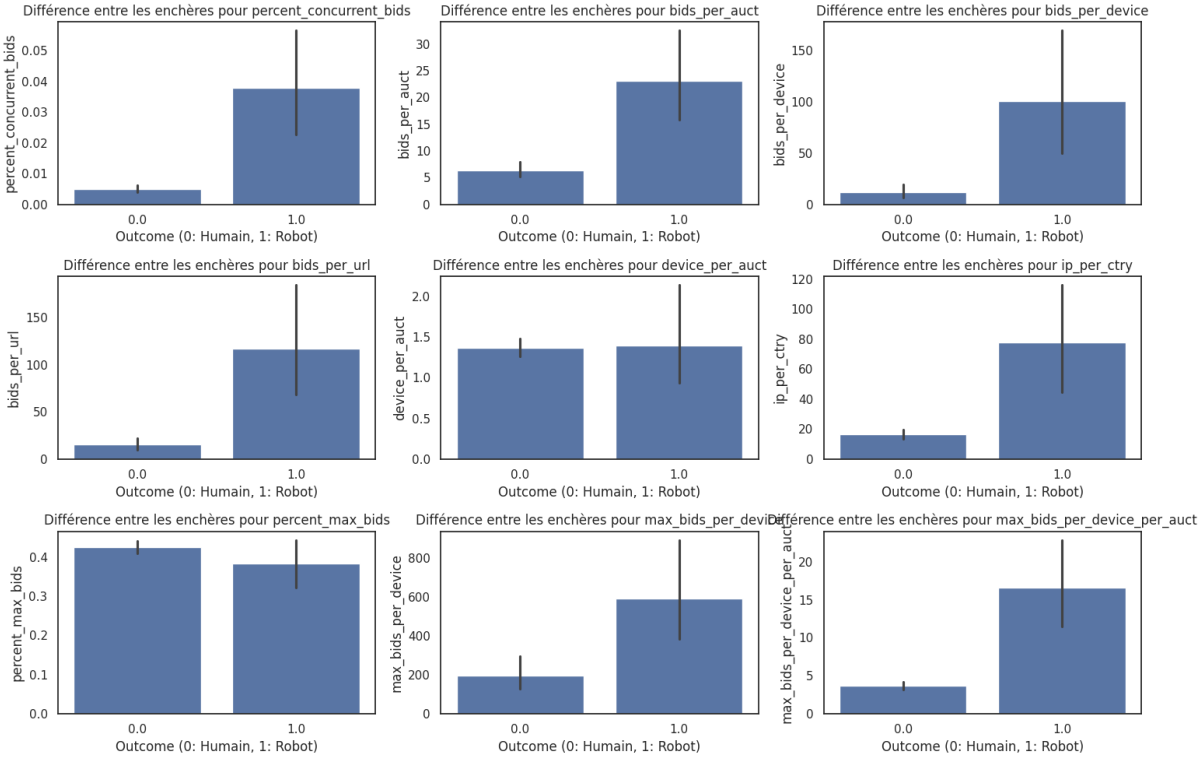


FIGURE 2 – Analyse Comparative de quelques features entre les Enchérisseurs Humains et les Robots

4 Modélisation Prédictive

4.1 Choix de métriques d'évaluation

Nous avons des données déséquilibrées, alors l'accuracy seule n'est pas une métrique signifiante. Nous allons nous baser sur les autres métriques comme :

1. **Matrice de confusion** : La matrice de confusion fournit un aperçu complet des résultats de la classification. Elle montre le nombre de vrais positifs, de vrais négatifs, de faux positifs et de faux négatifs.

2. **Précision** : La précision mesure le nombre de vrais positifs par rapport au nombre total de prédictions positives (vrais positifs + faux positifs). Une haute précision indique un faible nombre de faux positifs.

3. **Rappel (sensibilité) :** Le rappel mesure le nombre de vrais positifs par rapport au nombre total de vrais positifs et de faux négatifs (vrais positifs + faux négatifs). Il donne une indication sur la capacité du modèle à identifier tous les exemples positifs.

4. **F1-score :** Le F1-score est la moyenne harmonique de la précision et du rappel. Il est particulièrement utile lorsque vous voulez équilibrer la précision et le rappel.

5. **AUC-ROC (Courbe ROC et Aire sous la courbe) :**La courbe ROC (Receiver Operating Characteristic) est un graphique qui montre le taux de vrais positifs par rapport au taux de faux positifs à différents seuils de classification. L'AUC-ROC mesure l'aire sous cette courbe, fournissant une évaluation globale de la performance du modèle.

6. **Balanced Accuracy :** Il s'agit d'une version équilibrée de l'exactitude qui prend en compte le déséquilibre de classe en attribuant des poids égaux à chaque classe.

Puisque la classe des robots est plus petite que celle des humains, il est intéressant de

se concentrer sur des métriques qui tiennent compte du rappel et de la sensibilité pour nous assurer que le modèle identifie correctement la classe minoritaire.

4.2 Modèle AdaBoostClassifier

1ère partie : Nous avons entraîné le modèle AdaBoostClassifier sur les données déséquilibrées et ça nous a donnée les résultats suivant :

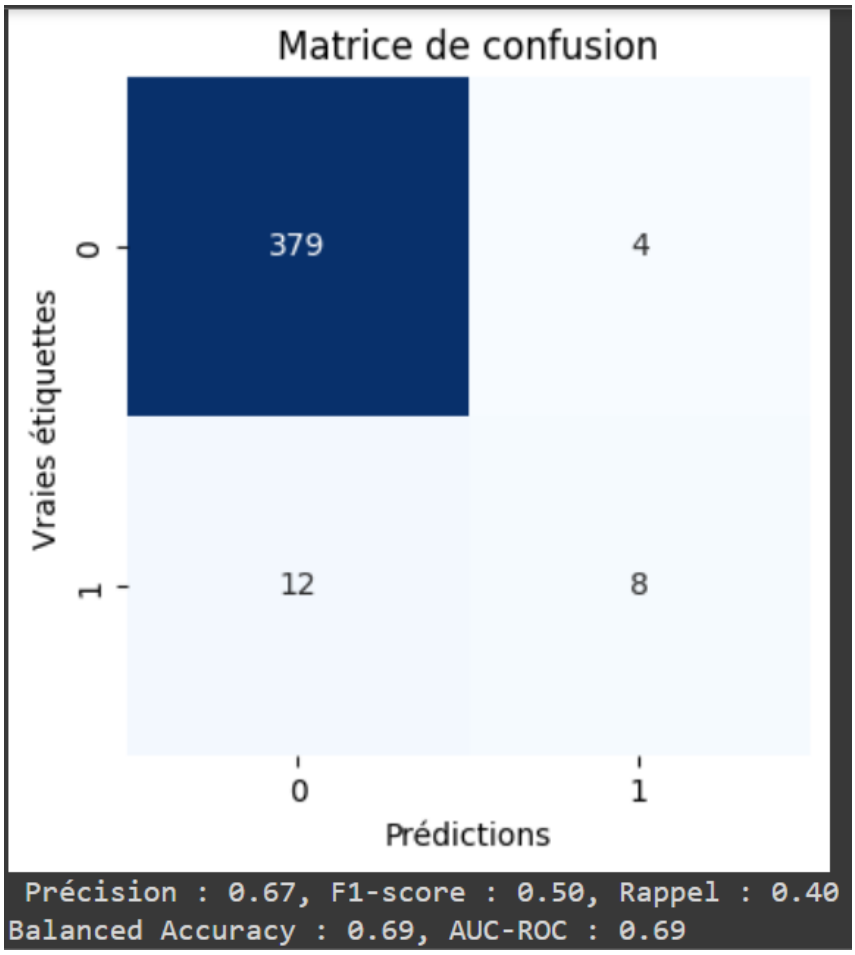


FIGURE 3 – Evaluation du modèle AdaBoostClassifier

2ème partie : Puis nous avons entraîné le modèle sur les données après SMOTE, et cela a amélioré les métriques : Balanced Accuracy, AUC-ROC et le Rappel. Par contre, nous pouvons remarquer que la précision et le F1-score ont diminué, ce qui est justifié par la dégradation des résultats dans la classe des humains

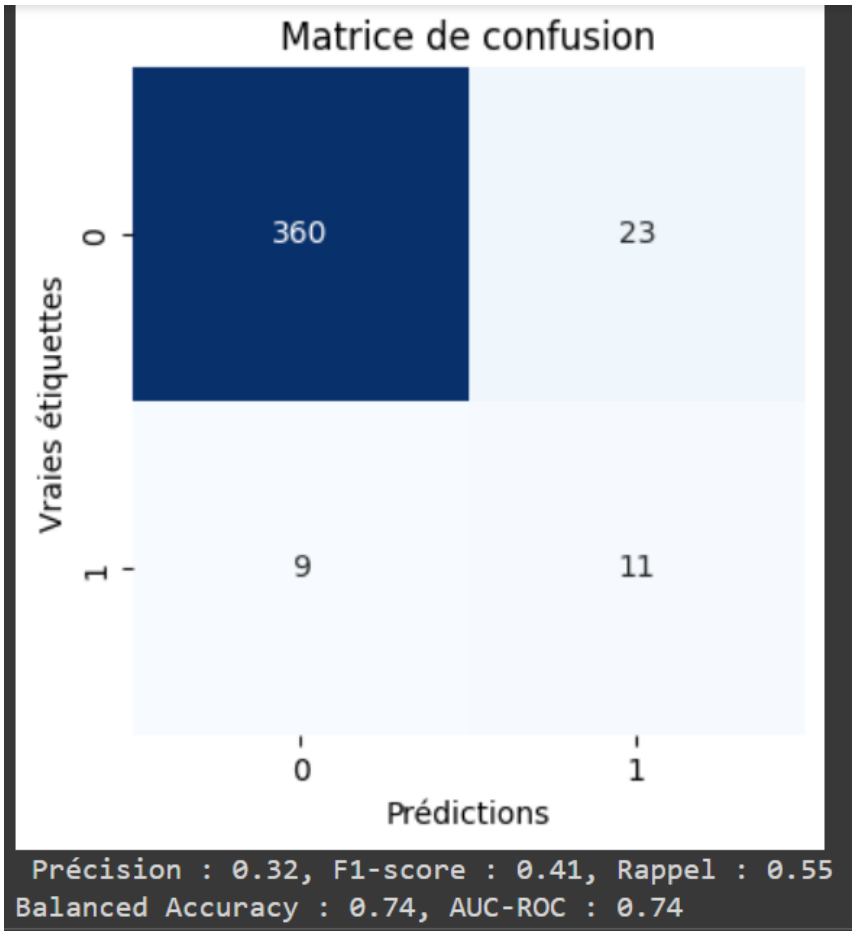


FIGURE 4 – Evaluation du modèle AdaBoostClassifier après SMOTE

4.3 KNN

Nous avons entraîné le modèle KNN sur les données déséquilibrés et nous avons obtenu ces résultats.

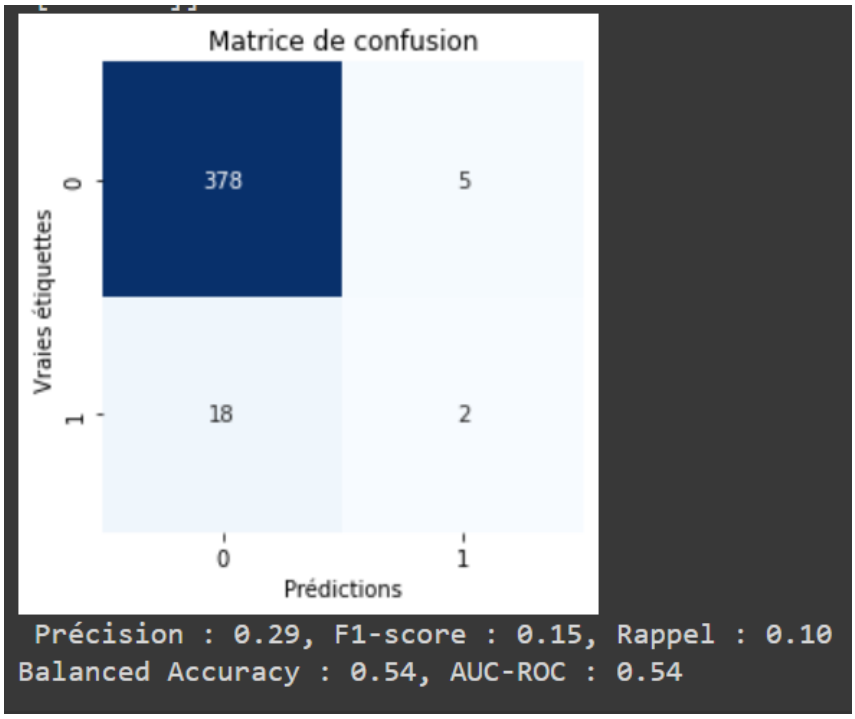


FIGURE 5 – Evaluation du modèle KNN

Ces résultats nous donne une idée que le modèle ne prédit pas bien les robots et qu'il n'est pas précis.

Pour résoudre ce problème, nous avons entraîné le modèle sur les données après SMOTE, nous avons obtenu ces résultats :

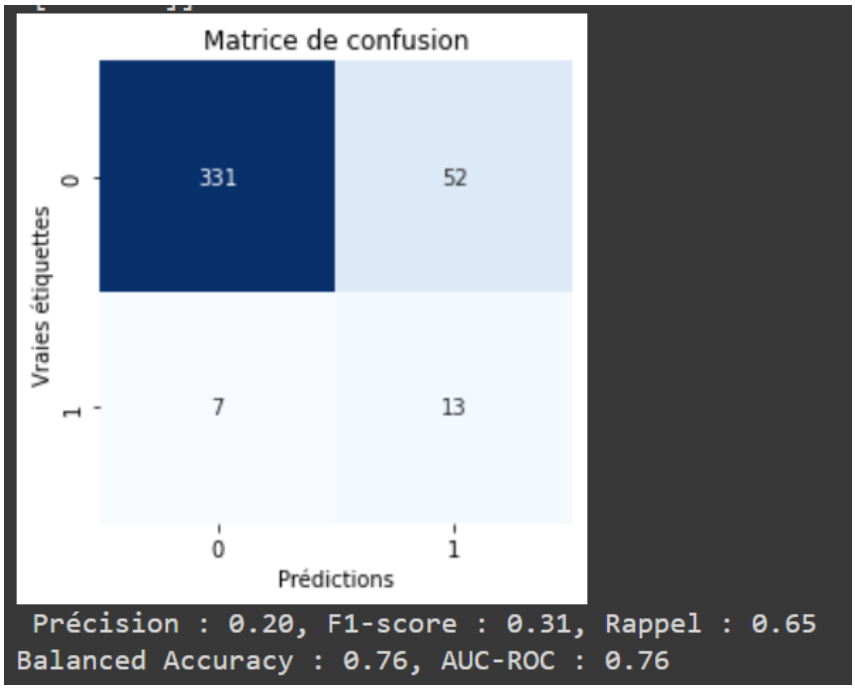


FIGURE 6 – Evaluation du modèle KNN après SMOTE

Alors, les métriques suivantes se sont améliorées : Balanced Accuracy,f1 score et le

Rappel. Par contre, nous pouvons voir que la précision a diminué, parce que les résultats se dont dégragés pour la classe des humains

4.4 Random Forest

Nous avons entraîné le modèle Random forest sur les données déséquilibrés et nous avons obtenu ces résultats.

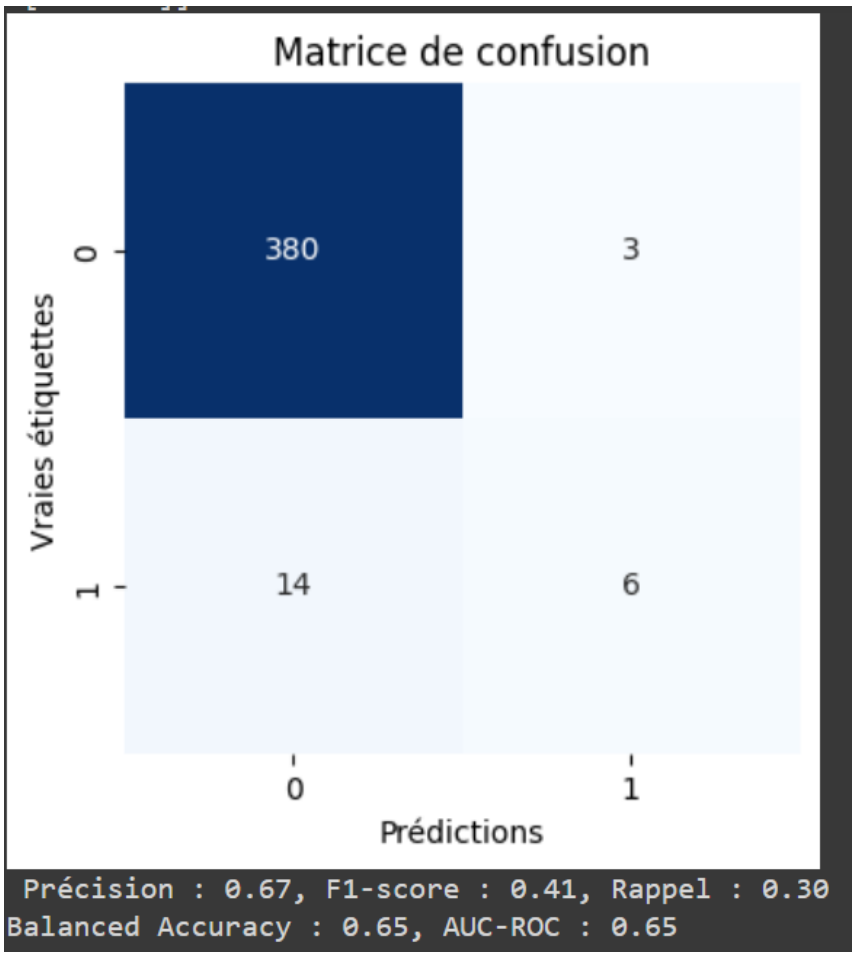


FIGURE 7 – Evaluation du modèle Random Forest

Cela signifie que le modèle prédit pas bien les robots et les humains, en le comparant avec les autres modèles.

Pour résoudre ce problème, nous avons entraîné le modèle sur les données après SMOTE, nous avons obtenu ces résultats :

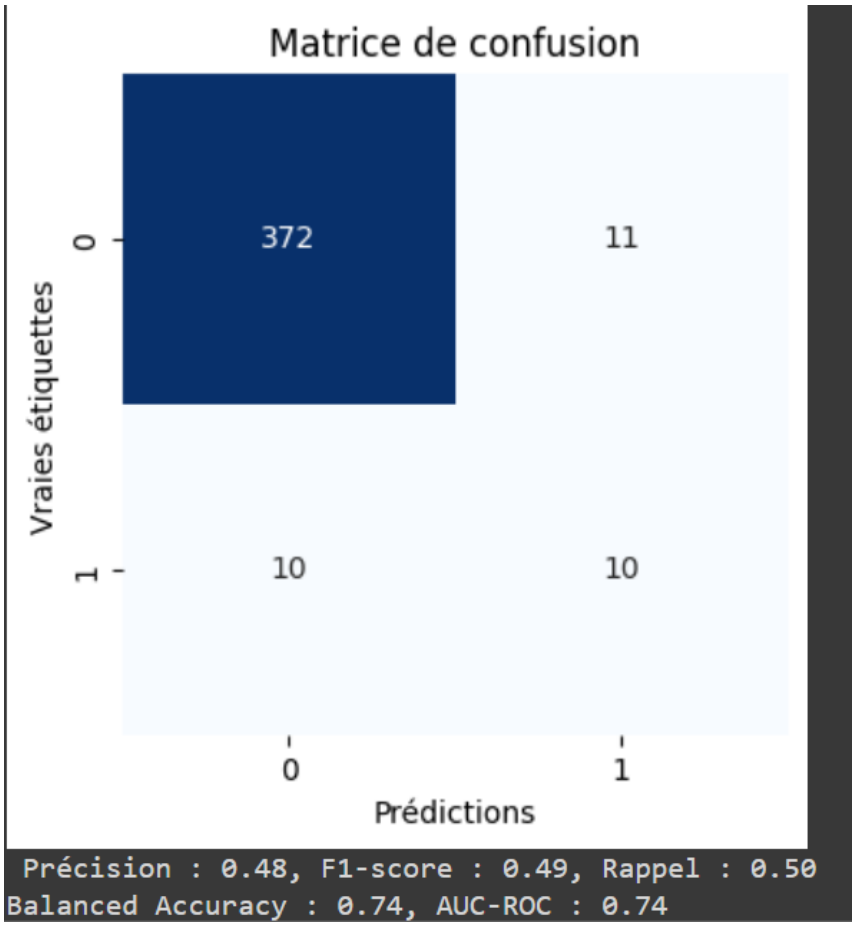


FIGURE 8 – Evaluation du modèle Random Forest après SMOTE

Alors, les métriques suivantes se sont améliorées : Balanced Accuracy,f1 score et le Rappel. Par contre, nous pouvons voir que la précision a diminué, parce que les résultats se dont dégragés pour la classe des humains.

4.5 Modèle SVM

Avant l'application du modèle SVM, nous avons opté à adopter l'approche PCA(n=2) pour ainsi obtenir les deux features les plus significatifs qui décrivent notre DataSet.

1ère partie : Nous avons entraîné le modèle SVM sur les données dés-équilibrées, ainsi nous avons obtenu les résultats suivants :

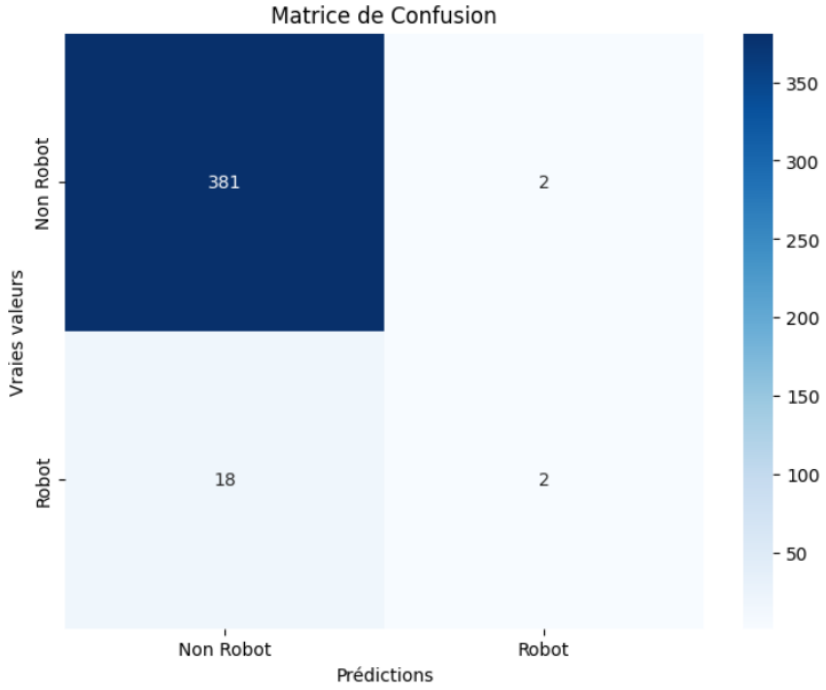


FIGURE 9 – Matrice de confusion du modèle SVM

AUC-ROC:0.55
balanced_accuracy:0.55

Classification Report:

	precision	recall	f1-score	support
0.0	0.95	0.99	0.97	383
1.0	0.50	0.10	0.17	20
accuracy			0.95	403
macro avg	0.73	0.55	0.57	403
weighted avg	0.93	0.95	0.93	403

FIGURE 10 – Evaluation du modèle SVM

2ème partie : L'application de la technique SMOTE avant l'utilisation du SVM a considérablement amélioré les performances du modèle. La balanced accuracy et l'AUC-ROC ont augmenté respectivement à 0.74, indiquant une meilleure capacité du modèle à traiter les classes déséquilibrées. En particulier, la sensibilité pour la classe minoritaire a augmenté de manière significative, passant de 0.10 à 0.65, soulignant ainsi l'efficacité de SMOTE dans l'amélioration de la détection des exemples de la classe moins représentée.

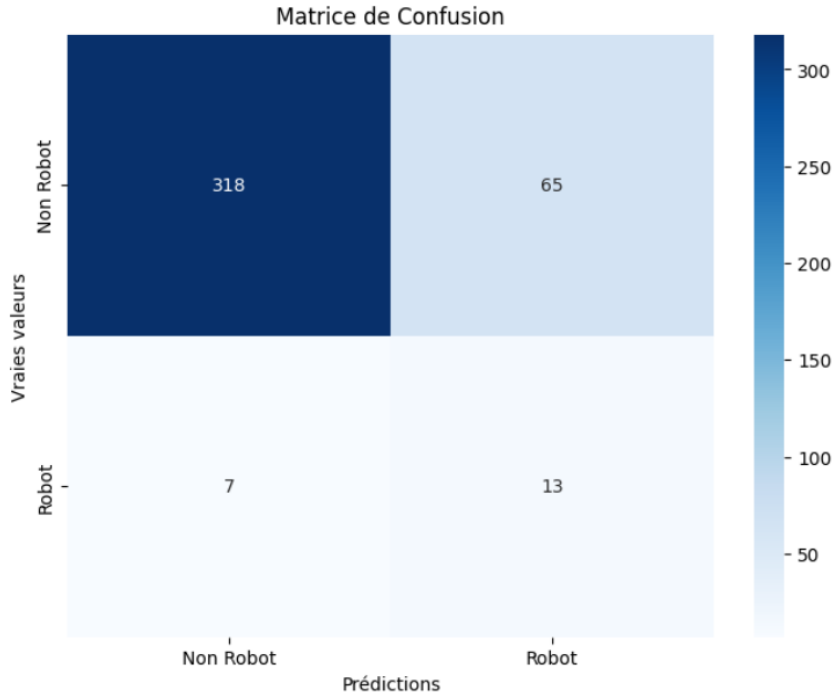


FIGURE 11 – Matrice du confusion SVM après SMOTE

```
AUC-ROC:0.74
balanced_accuracy:0.74

Classification Report:
              precision    recall  f1-score   support

         0.0       0.98      0.83      0.90       383
         1.0       0.17      0.65      0.27        20

    accuracy                           0.82       403
   macro avg       0.57      0.74      0.58       403
weighted avg       0.94      0.82      0.87       403
```

FIGURE 12 – Evaluation du modèle SVM après SMOTE

5 Conclusion

En conclusion, notre démarche pour améliorer la performance du modèle de classification des enchères humaines ou automatisées a suivi une approche méthodique en plusieurs étapes. Tout d'abord, nous avons effectué une analyse approfondie des caractéristiques en utilisant le Feature Engineering pour sélectionner les variables pertinentes. Ensuite, nous avons pris en compte le déséquilibre de classe en appliquant la technique SMOTE pour équilibrer les données.

Par la suite, nous avons évalué plusieurs modèles de classification, dont AdaBoost-Classifier, KNN, Random Forest, et SVM. Les résultats obtenus, exprimés en termes de

précision, F1-Score, rappel, Balanced-Accuracy, et AUC-ROC, ont fourni des insights précieux sur la performance de chaque modèle.

En analysant les performances des différents modèles, le modèle Random Forest se distingue par la meilleure précision et le F1-Score le plus élevé parmi les modèles évalués, avec des valeurs respectives de 0.48 et 0.49. Cela suggère que Random Forest est relativement plus capable de prédire avec précision les deux classes d'enchères, bien que son rappel soit modéré à 0.50. D'autre part, bien qu'AdaBoostClassifier affiche un rappel plus élevé à 0.55, sa précision et son F1-Score sont relativement plus faibles, indiquant potentiellement une plus grande sensibilité aux faux positifs. Les modèles KNN et SVM présentent des performances globalement inférieures, avec des scores de précision et de F1-Score considérablement plus bas, bien que le rappel pour KNN soit relativement plus élevé à 0.65.

En continuant cette démarche, nous recommandons d'explorer davantage l'optimisation des hyperparamètres pour les modèles sélectionnés, afin de maximiser leur efficacité. Cette étude constitue une base solide pour le développement d'un modèle robuste de classification des enchères, avec une attention particulière à la détection des enchères automatisées, qui est l'aspect principal de notre projet.

Modèle	**Précision**	**F1-Score**	**Rappel**	**Balanced-Accuracy**	**AUC-ROC**
AdaBoostClassifier	0.32	0.41	0.55	0.74	0.74
KNN	0.20	0.31	0.65	0.76	0.76
Random Forest	0.48	0.49	0.50	0.74	0.74
SVM	0.17	0.27	0.65	0.74	0.74

Table 1 – Performances des Modèles de Classification des Enchères