INTRODUCTION MACHINE LEARNING

Rapport de Machine Learning

Yassir BELARCHE
Aimen KHAOUID
Ayoub FARAJ
Mohammed Amine BOURICHI
Oussama MOUNAJJIM

ECOLE CENTRALE CASABLANCA, JANUARY 2024

Contents

1 Introduction

1.1 Contexte du Projet

Dans le monde dynamique des enchères en ligne, une préoccupation croissante émerge : la présence de "robots" - des offres automatisées générées par des logiciels - qui faussent le jeu équitable des enchères. Cette situation a engendré une frustration croissante parmi les enchérisseurs humains, qui se retrouvent souvent incapables de remporter des enchères face à ces concurrents logiciels. En conséquence, la fidélité et l'engagement des utilisateurs de base du site, qui sont essentiels à sa prospérité, diminuent considérablement.

Face à ce défi, le but principal de ce projet est de déceler les offres placées par des "robots" sur un site d'enchères en ligne. Cette identification permettra aux propriétaires du site d'étiqueter et d'éliminer ces utilisateurs de leur plateforme, rétablissant ainsi une activité d'enchères juste et équilibrée.

Ces données se composent de deux ensembles distincts : un ensemble de données sur les enchérisseurs, incluant des informations telles que l'identifiant, le compte de paiement, et l'adresse de l'enchérisseur, et un ensemble de données sur les offres, qui comprend 7,6 millions d'offres réalisées via des appareils mobiles.

En conclusion, ce projet vise à restaurer l'intégrité des enchères en ligne en identifiant et en éliminant la présence perturbatrice des robots, améliorant ainsi l'expérience des utilisateurs humains et la santé globale du site d'enchères.

1.2 Objectifs

L'objectif principal de ce projet est la détection et l'identification précise des offres effectuées par des robots dans le cadre des enchères en ligne. Cette démarche a plusieurs buts spécifiques :

1. **Améliorer la Justesse des Enchères :** En identifiant les offres automatisées, le projet vise à rétablir une concurrence loyale et équitable pour les enchérisseurs humains. Cela contribuera à maintenir l'intégrité et la confiance dans le système d'enchères.

2. **Augmenter la Satisfaction des Utilisateurs :** En éliminant les activités automatisées, le site peut offrir une meilleure expérience aux utilisateurs humains, ce qui est crucial pour retenir les clients fidèles et attirer de nouveaux utilisateurs.

3. **Analyse de l'Importance des Caractéristiques :** Comprendre quels facteurs contribuent le plus à la classification des enchères comme étant effectuées par des robots ou des humains. Cela inclut l'analyse des comportements tels que les motifs d'accès aux enchères et l'utilisation de multiples adresses IP.

En résumé, ce projet vise à allier la technologie de pointe en matière d'apprentissage automatique et l'analyse de données pour résoudre un problème réel et pressant dans le domaine des enchères en ligne, avec un impact positif à la fois sur les utilisateurs et sur l'écosystème global des enchères.

2 Analyse Exploratoire des Données

2.1 Présentation des Données

Descriptions des fichiers :

- **train.csv** - Ensemble de données d'entraînement issu de l'ensemble des enchérisseurs
- **test.csv** - Ensemble de données de test issu de l'ensemble des enchérisseurs
- **bids.csv** - Ensemble de données des enchères

Nous présentons désormais les colonnes des bases de données "bids" et "train" respectivement :

2.1.1 Base de données "bids"

- **bid_id** - Identifiant unique pour une offre soumise
- **bidder_id** - Identifiant unique d'un enchérisseur
- **auction** - Identifiant unique d'une enchère
- **merchandise** - La catégorie dans laquelle s'inscrit le produit mis en enchère et pour lequel une offre a été soumise.
- **device** - Modèle de téléphone d'un visiteur
- **time** Heure à laquelle l'offre est faite (transformée pour protéger la privacité de l'enchérisseur).
- **country** - Le pays auquel appartient l'IP
- **ip** - Adresse IP d'un enchérisseur (obfusquée pour protéger la vie privée).
- **url** - url d'où l'enchérisseur a été référé (obfusquée pour protéger la vie privée).

2.1.2 Base de données "train"

- **bidder_id** - Identifiant unique d'un enchérisseur
- **payment_account** - Compte de paiement associé à un enchérisseur (transformée pour protéger la vie privée).
- **address** - Adresse postale d'un enchérisseur (transformée 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.

2.1.3 Nombre d'enchérisseurs différents dans les deux bases de données

- Nombre d'enchérisseurs différents 'bidders' dans la base de données bids.csv : **6614**
- Nombre d'enchérisseurs différents 'bidders' dans la base de données train.csv : **2013**

3 Statistiques Descriptives

La section suivante détaille les caractéristiques de notre base de données, mettant en lumière les aspects quantitatifs des enchères, des enchérisseurs, des dispositifs utilisés, des adresses IP, des pays, des comptes de paiement et des adresses postales.

3.1 Vue d'Ensemble des Données

Notre base de données contient un total de 3,071,224 offres uniques provenant de 1,984 enchérisseurs différents. L'analyse révèle également la présence de 10 catégories de marchandises distinctes, 5,729 dispositifs différents et 1,030,950 adresses IP uniques. Les enchères proviennent de 198 pays différents, suggérant une large portée géographique et une diversité considérable des participants.

- Nombre total d'offres (bids) : 3,071,224
- Nombre total d'enchérisseurs (bidders) : 1,984
- Catégories de marchandises (merchandise) : 10
- Dispositifs (devices) utilisés : 5,729
- Adresses IP (ip) uniques : 1,030,950
- Pays (country) impliqués : 198
- Enchères (auctions) distinctes : 12,740
- Comptes de paiement (payment_account) : 1,984
- Adresses (address) : 1,984

3.2 Détails par Catégorie d'Outcome

L'analyse comparative entre les enchérisseurs humains et les robots révèle des différences notables en termes de comportement et d'activité.

3.2.1 Humains (Outcome = 0)

Les enchérisseurs humains affichent les statistiques suivantes :

- Offres uniques moyennes par utilisateur : 1,413.51
- Auctions uniques moyennes par utilisateur : 58.07
- Pays uniques moyens par utilisateur : 12.59
- IPs uniques moyennes par utilisateur : 581.26
- Devices uniques moyens par utilisateur : 73.95
- URLs uniques moyennes par utilisateur : 335.19

3.2.2 Robots (Outcome = 1)

Les enchérisseurs robots présentent les statistiques suivantes, suggérant un niveau d'activité plus élevé :

- Offres uniques moyennes par utilisateur : 4,004.04
- Auctions uniques moyennes par utilisateur : 145.04
- Pays uniques moyens par utilisateur : 26.32
- IPs uniques moyennes par utilisateur : 2,387.80
- Devices uniques moyens par utilisateur : 163.61
- URLs uniques moyennes par utilisateur : 544.58

Ces distinctions marquées entre les moyennes des humains et des robots suggèrent que les robots présentent une activité d'enchère plus intensive et diversifiée, manifestée par un nombre significativement plus élevé d'offres, une plus grande variété d'articles enchéris, et l'utilisation de multiples adresses IP et appareils.

3.3 Visualisations

Dans cette section, nous présentons des graphiques de ditribution des données dans la nouvelle base :

3.3.1 Distribution des enchérisseurs :

Le graphe ci-dessous illustre les ratios d'enchérisseurs humains et robots :

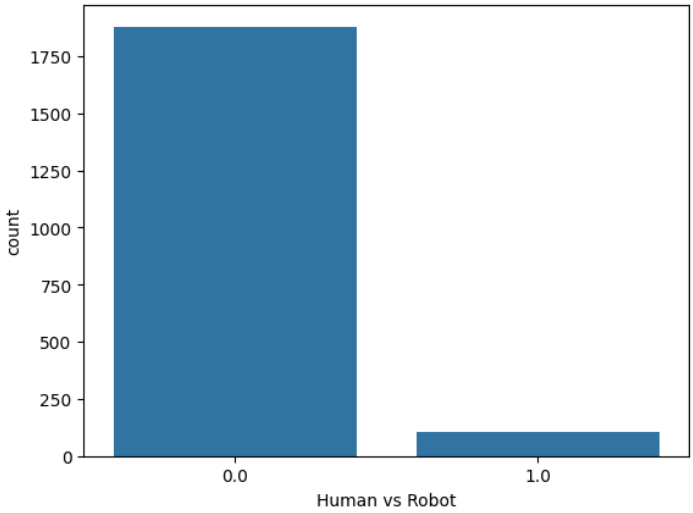


Figure 1: Proportion des humain et des robots dans la base de données

3.3.2 Distribution des catégories d'objets mis en enchère :

Le graphe ci-dessous montre le classement des catégories d'objets selon le nombre d'enchères générées :

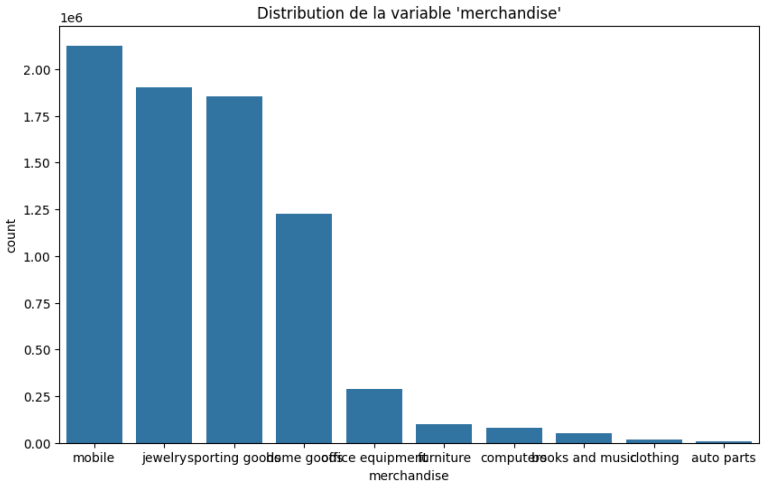


Figure 2: Distribution des enchères par catégories de marchandise

3.3.3 Distribution des objets les plus sollicités :

Le graphe ci-dessous montre les 10 enchères ayant reçu le plus d'offres par rapport à la nature de l'enchérisseur (humain ou robot) :

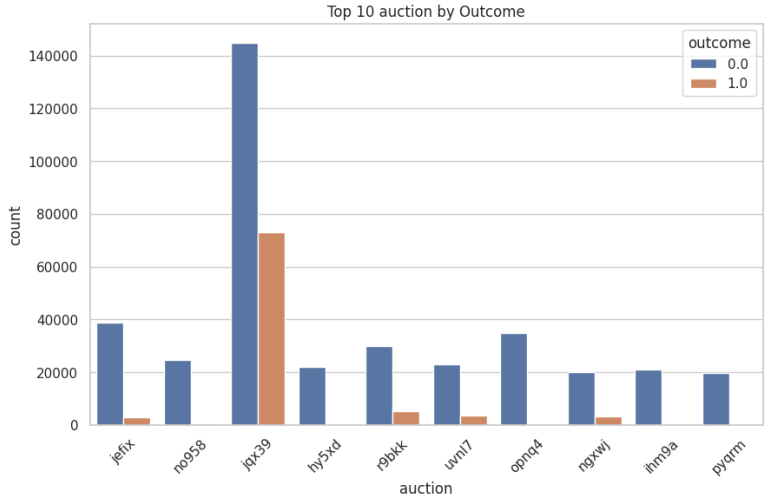


Figure 3: Références avec le plus d'enchères par nature d'enchérisseur

3.3.4 Distribution des enchères par marchandise :

Le graphe ci-dessous compare le nombre d'enchères effectuées par les enchérisseurs humains et robots pour chaque marchandise :

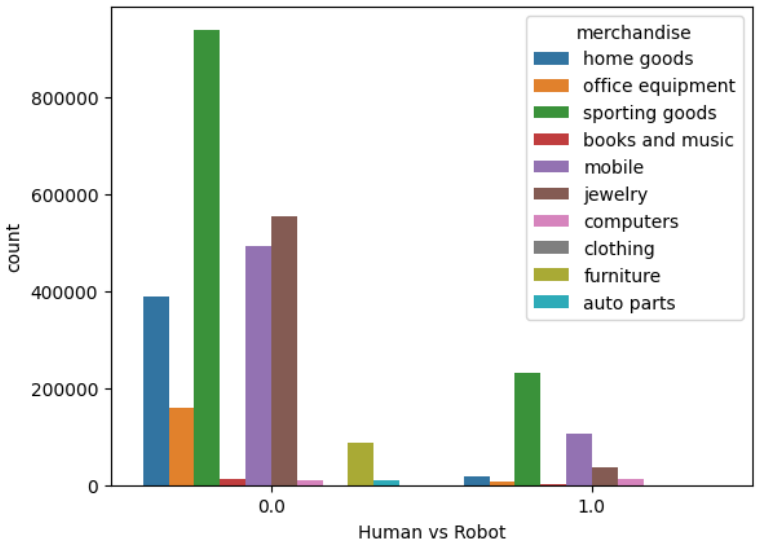
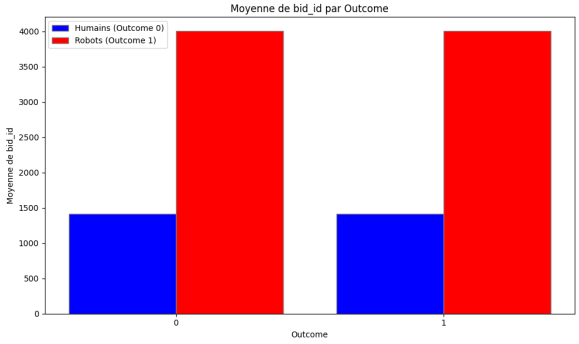


Figure 4: Enchères sur les marchandises par nature d'enchérisseur

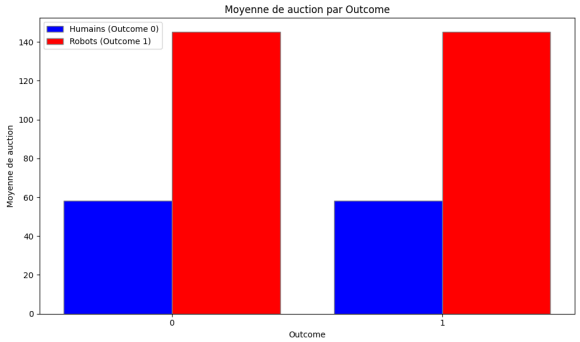
3.3.5 Comparaison enchérisseurs humains et robots :

Nous comparons ci-dessous la nature des enchérisseurs en fonction des moyennes des identifiants d'enchères, des références d'objets, du nombre d'appareils et du nombre de pays associés à leur participation à la vente aux

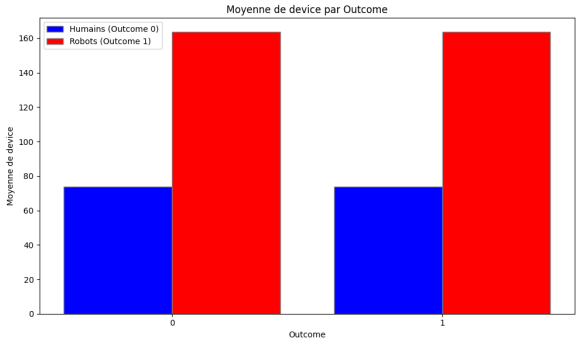
enchères. Les graphiques obtenus sont présentés ci-après :



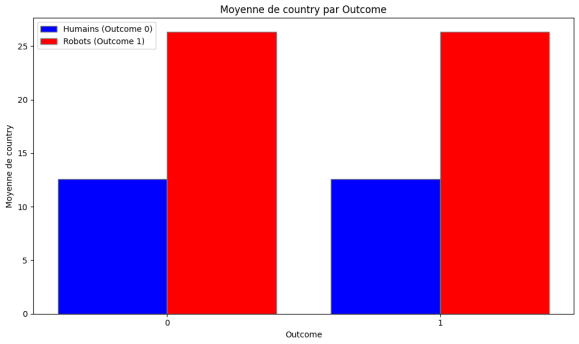
(a) Humain vs Robot : Moyenne identifiant d'enchère



(b) Humain vs Robot : Moyenne de références enchéries



(c) Humain vs Robot : Moyenne nombre d'appareils



(d) Humain vs Robot : Moyenne nombre de pays

Une observation remarquable se dégage du graphique, mettant en évidence une nette infériorité de la moyenne humaine par rapport à celle des robots pour chacune des variables prises en considération.

Cette disparité significative entre les moyennes humaines et robotiques suggère que les robots présentent une activité d'enchère plus soutenue. Plus précisément, ils semblent soumettre un nombre substantiellement plus élevé d'offres, ciblant un nombre plus important de références et provenant de diverses localisations.

4 Feature Engineering

Dans cette section, nous explorons l'importance cruciale de l'ingénierie des fonctionnalités (Feature Engineering), un processus essentiel pour optimiser les performances des modèles de machine learning en créant de nouvelles caractéristiques significatives à partir des données existantes. Le feature engineering englobe la sélection des caractéristiques pertinentes, la création de nouvelles caractéristiques informatives, ainsi que l'analyse de leur impact sur la qualité du modèle.

4.1 Sélection des Caractéristiques

La sélection des caractéristiques constitue une étape fondamentale dans la construction d'un modèle performant. Nous avons adopté une approche consistant à conserver les caractéristiques présentant une corrélation significative avec la variable cible, tout en éliminant celles qui sont redondantes ou peu informatives.

4.2 Création de Nouvelles Caractéristiques

Nous avons introduit plusieurs nouvelles caractéristiques pour enrichir notre jeu de données et capturer des informations pertinentes sur les comportements des enchérisseurs. Ces caractéristiques comprennent notamment :

- **Proportion d'IPs uniques par enchérisseur :** Cette caractéristique indique le nombre d'adresses IP uniques utilisées par chaque enchérisseur, pouvant révéler des comportements de connexion diversifiés ou des stratégies d'enchères automatisées.
- **Total d'URL uniques par utilisateur** : Représente la variété des ressources web que chaque utilisateur a visitées, possiblement reflétant l'étendue de ses intérêts ou de son activité.
- **Total d'enchères uniques par utilisateur** : Compte le nombre d'enchères distinctes auxquelles un utilisateur a participé, donnant une mesure de son engagement dans les enchères.
- **Proportion d'URLs uniques par enchérisseur** : Indique le nombre d'URLs uniques par enchérisseur, divisé par le nombre total d'enchères, offrant des insights sur le comportement de navigation des utilisateurs.
- **Nombre total d'enchères par utilisateur** : Indique le nombre total d'enchères auxquelles chaque utilisateur a participé, reflétant son activité sur la plateforme.
- **Nombre total d'adresses IP uniques par utilisateur** : Compte le nombre total d'adresses IP uniques utilisées par chaque utilisateur, offrant des informations sur la diversité de leurs connexions.

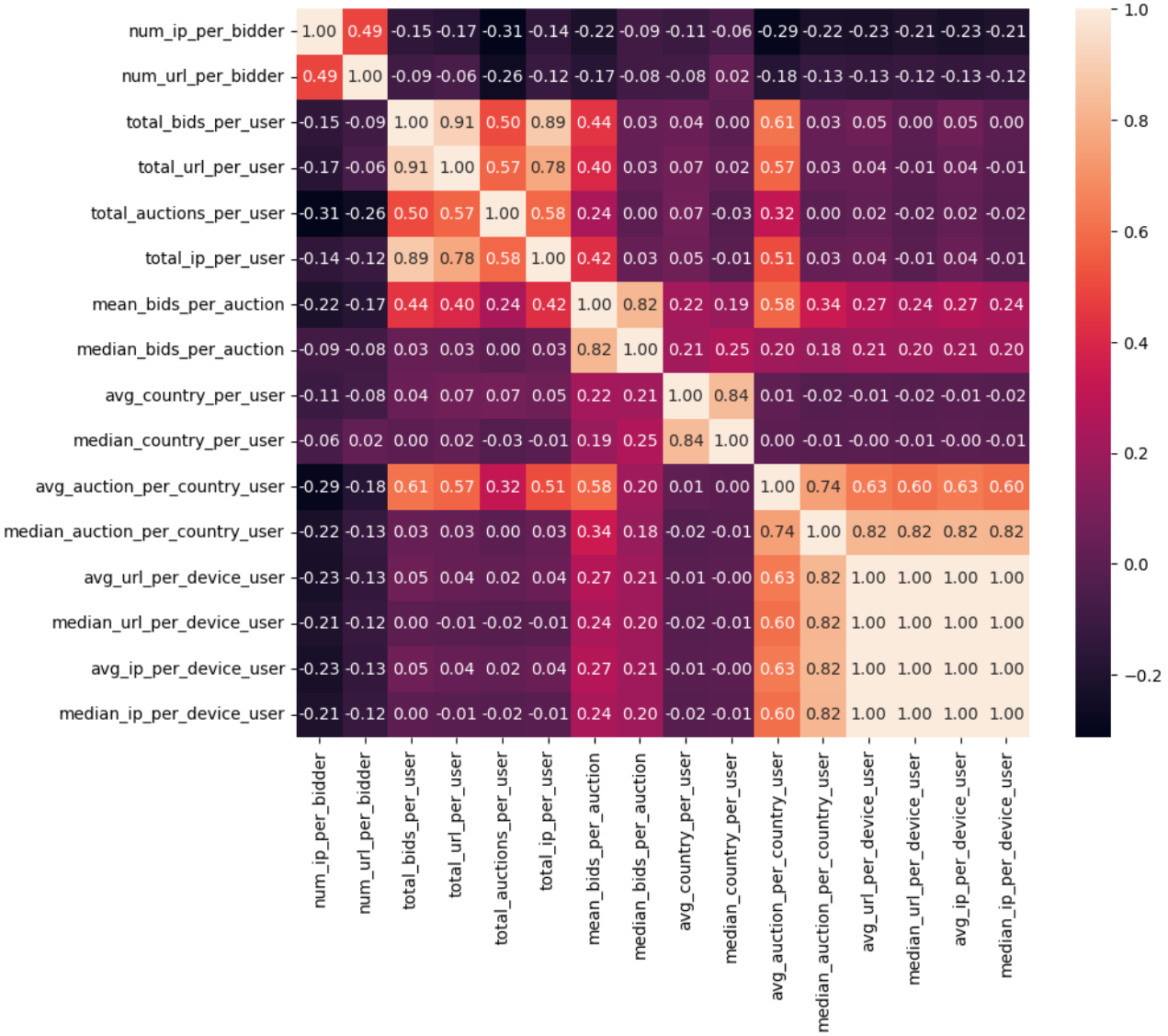
Ces nouvelles caractéristiques visent à capturer divers aspects du comportement des enchérisseurs, ce qui pourrait être essentiel pour la prédiction des enchérisseurs robots.

4.3 Justification et Impact sur le Modèle

L'introduction de ces nouvelles caractéristiques enrichit la représentation des données et renforce la capacité du modèle à distinguer les enchérisseurs humains des robots. Les caractéristiques sélectionnées sont choisies pour leur pertinence potentielle dans la prédiction de l'issue des enchères. Nous avons réalisé une analyse des corrélations entre les caractéristiques et la variable cible, ainsi qu'une cartographie des corrélations entre les caractéristiques elles-mêmes, pour évaluer leur impact et leur importance relative dans le modèle.

Nous avons inclus dans notre analyse deux graphiques essentiels :

Carte de chaleur des corrélations entre les caractéristiques numériques:



4.4 Equilibrage des données robots et humains :

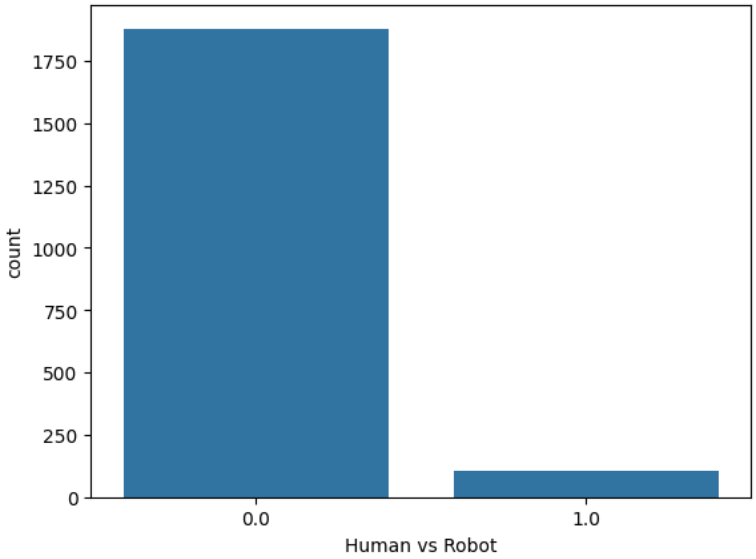


Figure 6: Proportion des humain et des robots dans la base de données non équilibrés

Pour pallier le déséquilibre entre les classes de robots et d'humains dans notre ensemble de données, nous avons adopté une approche innovante en utilisant un réseau de neurones génératif antagoniste (GAN). Notre processus a commencé par la normalisation des caractéristiques, après quoi nous avons isolé les données appartenant à la classe minoritaire, dans ce cas, les robots. Nous avons ensuite construit un GAN composé d'un générateur et d'un discriminateur, chacun constitué de couches denses et utilisant des fonctions d'activation relu et tanh.

Le générateur a été entraîné pour produire de nouvelles instances de données qui imitent la distribution de la classe minoritaire, tandis que le discriminateur a été entraîné pour différencier entre les instances générées et les véritables instances de la classe minoritaire. Après un entraînement intensif, le générateur a réussi à créer des données synthétiques réalistes qui ont été mélangées avec nos données existantes, ce qui a entraîné un ensemble de données plus équilibré. Cette nouvelle collection de données a permis de former des modèles de classification sur un terrain plus égalitaire, avec une représentation améliorée des motifs complexes inhérents à la classe minoritaire.

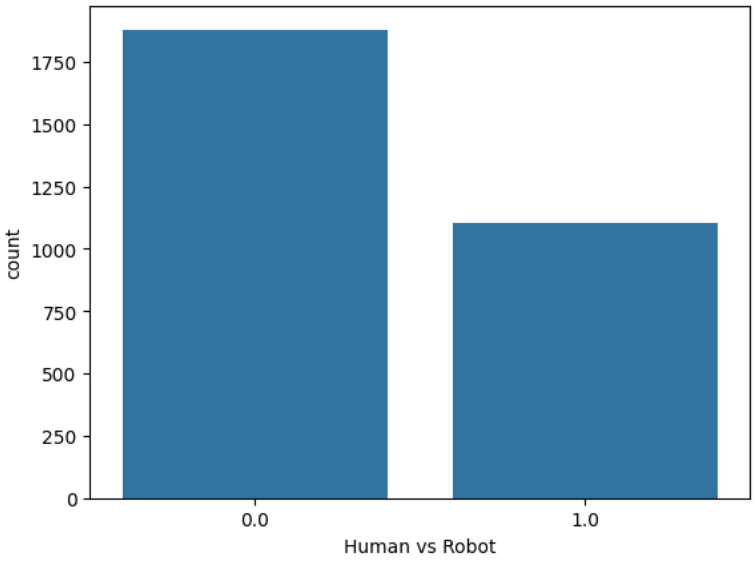


Figure 7: Proportion des humain et des robots dans la base de données équilibrés

5 Modélisation Prédictive

La modélisation prédictive repose sur le choix judicieux des algorithmes pour la classification des données. Les modèles sélectionnés sont basés sur une première analyse des performances, mettant en évidence trois candidats prometteurs : la Régression Logistique, le Random Forest et le Support Vector Machine.

5.1 Choix des modéles

5.1.1 Régression Logistique

La régression logistique est souvent le premier choix pour un problème de classification binaire en raison de sa simplicité et de son efficacité. Ce modèle est particulièrement utile lorsqu'il y a une relation linéaire entre les caractéristiques indépendantes et la probabilité logarithmique de la variable de réponse. En outre, il fournit non seulement des prédictions de classification, mais aussi des probabilités qui peuvent être utiles pour évaluer l'incertitude de la classification. Sa simplicité le rend facile à implémenter, à interpréter et à entraîner, ce qui en fait un choix robuste pour de nombreux scénarios.

5.1.2 Machine à Vecteurs de Support (SVM)

Le SVM est un modèle puissant et polyvalent, idéal pour les ensembles de données de taille moyenne et les espaces à grande dimensionnalité. Il est efficace dans les cas où la séparation entre les classes est claire et où il est important de trouver la meilleure marge de séparation. Le SVM utilise des fonctions de noyau pour transformer l'espace de caractéristiques et trouver un hyperplan optimal dans cet espace transformé, offrant ainsi une grande

flexibilité pour modéliser des relations non linéaires. C'est un choix privilégié pour des problèmes complexes où la précision est cruciale.

5.1.3 Forêt Aléatoire (Random Forest)

Le Random Forest est un modèle d'apprentissage ensembliste qui combine les prédictions de plusieurs arbres de décision pour produire des résultats plus stables et précis. Ce modèle est particulièrement utile pour les ensembles de données avec un grand nombre de caractéristiques et de données, car il peut capturer des interactions complexes entre les variables. Il est également robuste aux valeurs aberrantes et peut être utilisé pour estimer l'importance des caractéristiques, ce qui aide à comprendre le modèle. En raison de son approche d'ensemble, il est souvent plus performant que les modèles uniques, en particulier sur des ensembles de données complexes.

5.2 Implémentation des modéles

Dans le cadre de notre étude, nous avons implémenté trois modèles de classification différents : la régression logistique, le SVM (Support Vector Machine) et le Random Forest. L'objectif était d'évaluer et de comparer leurs performances sur notre base de données spécifique.

5.2.1 Préparation des Données

Nous avons commencé par charger notre ensemble de données dans un DataFrame Pandas. Ensuite, nous avons séparé les caractéristiques (features) de la variable cible (outcome). Pour cela, nous avons utilisé la fonction `train_test_split` de Scikit-Learn pour diviser les données en ensembles d'entraînement et de test, en veillant à maintenir une distribution équilibrée des classes.

5.2.2 Entraînement des Modèles

Trois modèles ont été initialisés : la Régression Logistique, le SVM et le Random Forest. Chaque modèle a été entraîné sur l'ensemble d'entraînement. Nous avons pris soin d'ajuster les paramètres de chaque modèle pour optimiser leurs performances.

5.2.3 Évaluation des Modèles

Après l'entraînement, nous avons évalué les modèles sur l'ensemble de test. Nous avons calculé la précision, le score F1 et l'aire sous la courbe ROC (ROC AUC) pour chaque modèle. Ces métriques nous ont permis de comprendre la performance de chaque modèle en termes de précision, de sensibilité et de spécificité.

5.2.4 Visualisation des Performances

Enfin, nous avons visualisé les performances des modèles à l'aide de courbes ROC. Ces courbes nous ont aidé à comparer visuellement l'efficacité de chaque modèle dans la classification des données. Un graphique a été créé pour illustrer les taux de vrais positifs par rapport aux faux positifs de chaque modèle.

6 Évaluation des Modèles

6.1 Métriques d'Évaluation

Lors de l'évaluation de nos modèles de classification - la régression logistique, le SVM et le Random Forest - nous avons sélectionné des métriques d'évaluation spécifiques pour capturer efficacement la performance de chaque modèle. Nous avons utilisé la précision (Accuracy), le score F1 et l'aire sous la courbe ROC (ROC AUC) comme nos principales métriques. La précision offre une mesure globale de la performance, indiquant la proportion de prédictions correctes par rapport au total. Cependant, dans le contexte de classes déséquilibrées, elle peut être trompeuse. C'est pourquoi nous avons inclus le score F1, qui est la moyenne harmonique de la précision et du rappel, fournissant un équilibre entre la sensibilité et la spécificité, et est particulièrement utile lorsque les coûts des faux positifs et des faux négatifs diffèrent. Enfin, l'aire sous la courbe ROC (ROC AUC) nous permet d'évaluer la performance des modèles indépendamment du seuil de classification, offrant une mesure complète de la probabilité de discrimination entre les classes.

6.2 Comparaison des Performances sur données non équilibrés

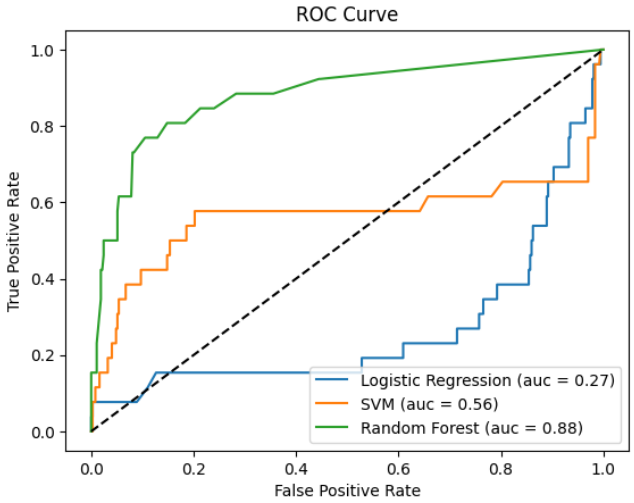


Figure 8: ROC Curve - Données non équilibrés

Les performances des modèles de classification, y compris la régression logistique, le SVM (Support Vector Machine) et le Random Forest, ont été évaluées en utilisant des métriques d'exactitude (Accuracy), de score F1 et de l'aire sous la courbe ROC (ROC AUC). Bien que les trois modèles présentent une précision élevée de 0.93, cette métrique ne reflète pas nécessairement leur efficacité à discriminer entre les classes dans un contexte où les classes sont déséquilibrées. Le score F1, qui prend en compte à la fois la précision et le rappel, révèle que le Random Forest excelle avec un score de 0.24 comparé au SVM qui n'a pas réussi à identifier correctement les vrais positifs, ce qui est indiqué par un score F1 de 0.00. La courbe ROC fournit une perspective plus complète,

montrant que le Random Forest a une AUC impressionnante de 0.88, suggérant une excellente performance dans la classification des résultats positifs par rapport aux négatifs. À l'opposé, la régression logistique et le SVM ont des AUC de 0.27 et 0.56 respectivement, ce qui indique que malgré une précision similaire, leur capacité à discriminer entre les classes est nettement moins efficace que celle du Random Forest. En conséquence, sur la base de notre analyse, le Random Forest émerge comme le modèle le plus performant pour cet ensemble de données spécifique.

Modèle	Précision	Score F1	AUC ROC
Régression Logistique	0.93	0.13	0.27
SVM	0.93	0.00	0.56
Random Forest	0.93	0.24	0.88

Table 1: Comparaison des métriques de performance des modèles de classification.

6.3 Comparaison des Performances sur données équilibrés

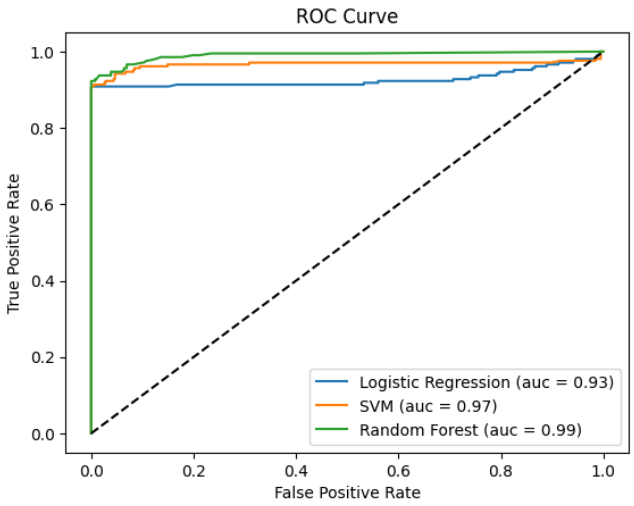


Figure 9: ROC Curve - Données équilibrés

Sur la base des métriques de performance affichées et de la courbe ROC après l'équilibrage des données, nous observons une amélioration significative de la performance pour les trois modèles de classification. La régression logistique montre une précision de 0.96 avec un score F1 de 0.95 et une AUC ROC de 0.93, indiquant une haute performance globale, tant en termes de précision que de capacité à discriminer entre les classes positives et négatives. Le SVM réalise des performances légèrement supérieures, avec une précision et un score F1 identiques, mais une AUC ROC de 0.97, suggérant une meilleure spécificité par rapport à la régression logistique. Néanmoins, c'est le Random Forest qui se distingue le plus nettement, atteignant une précision de 0.97, un score F1 de 0.95, et une AUC ROC exceptionnelle de 0.99, confirmant son statut de modèle le

plus performant post-équilibrage. Cette AUC ROC proche de la perfection illustre une excellente séparation des classes. Ces résultats indiquent que l'équilibrage des données a probablement permis une meilleure généralisation et une réduction des biais, menant à une performance accrue pour tous les modèles testés.

Modèle	Précision	Score F1	AUC ROC
Régression Logistique	0.96	0.95	0.93
SVM	0.97	0.95	0.97
Random Forest	0.97	0.95	0.99

Table 2: Comparaison des métriques de performance des modèles de classification.

6.4 Analyse des Résultats

Une analyse comparative des performances de nos modèles de classification avant et après l'équilibrage des données révèle des conclusions significatives. Initialement, les modèles affichaient une précision élevée; cependant, des AUC ROC relativement faibles pour la régression logistique et le SVM indiquaient des lacunes dans leur capacité discriminatoire. Plus particulièrement, le score F1 très bas du SVM signalait une faiblesse dans la détection des vrais positifs, probablement exacerbée par le déséquilibre des classes.

Après l'application de techniques d'équilibrage, nous observons une augmentation notable de l'AUC ROC pour la régression logistique et le SVM, et une amélioration remarquable pour le Random Forest, qui atteint une AUC de 0.99. Ces améliorations suggèrent que l'équilibrage des données a considérablement bénéficié à la capacité du Random Forest à gérer les fonctionnalités et à modéliser les interactions complexes.

De façon révélatrice, le score F1 uniformément élevé pour tous les modèles post-équilibrage indique une amélioration marquée dans la précision et le rappel, affirmant leur capacité accrue à identifier correctement les cas positifs tout en minimisant les erreurs. Bien que la précision reste élevée après l'équilibrage, c'est la progression des scores AUC ROC et F1 qui est particulièrement notable, offrant une évaluation plus nuancée et fiable de la performance dans le contexte d'un ensemble de données équilibré.

En conclusion, ces résultats mettent en lumière l'importance de l'équilibrage des classes pour améliorer la généralisation des modèles de classification, soulignant son rôle crucial dans l'amélioration des performances pour des applications réalistes.

7 Discussion

7.1 Interprétation des Résultats

L'amélioration des performances de classification après l'équilibrage des données suggère que les modèles étaient auparavant biaisés envers la classe majoritaire. Cela était particulièrement évident pour le SVM, dont le

score F1 a significativement augmenté après l'équilibrage. Le Random Forest a montré une capacité remarquable à s'adapter à l'équilibrage des données, comme en témoigne son AUC ROC exceptionnellement élevé, indiquant une capacité supérieure à distinguer entre les classes. Ces observations mettent en exergue la nécessité de traiter le déséquilibre des classes lors de la conception d'études de machine learning pour assurer une interprétation précise et équitable des résultats de modélisation.

7.2 Limitations

Bien que l'équilibrage des données ait entraîné une amélioration notable des performances des modèles, cette étude présente certaines limitations. Premièrement, les données générées par GAN pourraient ne pas parfaitement refléter la complexité et la diversité de la classe minoritaire. Deuxièmement, l'équilibrage des classes peut parfois conduire à un surajustement, en particulier si les données générées ne sont pas suffisamment variées. Enfin, l'interprétation des résultats pourrait être limitée par le manque de compréhension des processus sous-jacents qui régissent la génération de données, ce qui pourrait entraîner une confiance excessive dans les prédictions du modèle.

7.3 Suggestions pour des Recherches Futures

Pour les recherches futures, il serait bénéfique d'explorer d'autres techniques d'équilibrage des données, telles que l'under-sampling de la classe majoritaire ou le suréchantillonnage informé, pour comparer leur efficacité par rapport aux GANs. Il serait également judicieux d'incorporer des analyses de sensibilité pour évaluer la robustesse des modèles face à différentes distributions de classes. En outre, des recherches supplémentaires pourraient se concentrer sur l'amélioration de l'interprétabilité des modèles de forêt aléatoire et SVM, permettant une meilleure compréhension des caractéristiques qui influencent le plus les prédictions.

8 Conclusion

À travers notre étude, nous avons abordé le défi émergent dans les enchères en ligne posé par les robots enchérisseurs. En appliquant des techniques avancées d'apprentissage automatique et en équilibrant notre ensemble de données, nous avons pu améliorer considérablement la détection de ces agents automatisés. Les résultats obtenus indiquent que le Random Forest, en particulier, a bénéficié de manière significative de l'équilibrage des données, démontrant une capacité supérieure à discriminer entre les comportements humains et robotiques dans le contexte des enchères.

La comparaison des performances des modèles avant et après l'équilibrage des données révèle l'importance cruciale de préparer les données de manière à réduire le biais envers la classe majoritaire. Les améliorations observées dans les scores F1 et AUC ROC mettent en évidence l'efficacité des modèles post-équilibrage, en particulier dans la manière dont ils gèrent les données déséquilibrées et la qualité de leurs prédictions.

Cependant, notre recherche n'est pas sans limitations. La génération de données synthétiques via des GANs, bien que bénéfique, soulève des questions sur la fidélité absolue de ces données par rapport à la complexité réelle des comportements des enchérisseurs. Il existe également un risque de surajustement et une nécessité d'approfondir la compréhension des modèles utilisés.

À l'avenir, nous recommandons d'explorer une variété de techniques d'équilibrage des données pour comparer et potentiellement améliorer les performances des modèles. Des analyses de sensibilité plus poussées et une attention accrue à l'interprétabilité des modèles pourraient conduire à des insights plus nuancés et à une application pratique plus robuste.

En conclusion, ce projet illustre l'efficacité de l'apprentissage automatique dans l'identification des robots enchérisseurs et met en lumière les défis et opportunités associés à l'équilibrage des ensembles de données. Les méthodes et conclusions présentées ici pourraient servir de référence pour des recherches futures, contribuant ainsi à la lutte contre la fraude dans les enchères en ligne et à la promotion de pratiques commerciales équitables.