Description des données

Réalisé par : Rachida OUCHENE, Lylia TOUAZI, Koussaila HAMMOUCHE.
GROUPE 1

I.Introduction

Le jeu de données a été récolter des enchères eBay sur le produit iphone7. Nous étudierons un problème de classification binaires pour Le Shill bidding, pour pouvoir distinguer les enchères normales de celles frauduleuses. Shill bidding est la fraude aux enchères la plus courante, mais la plus difficile à détecter en raison de sa similitude avec le comportement d'enchères normal.

II.Présentation des données

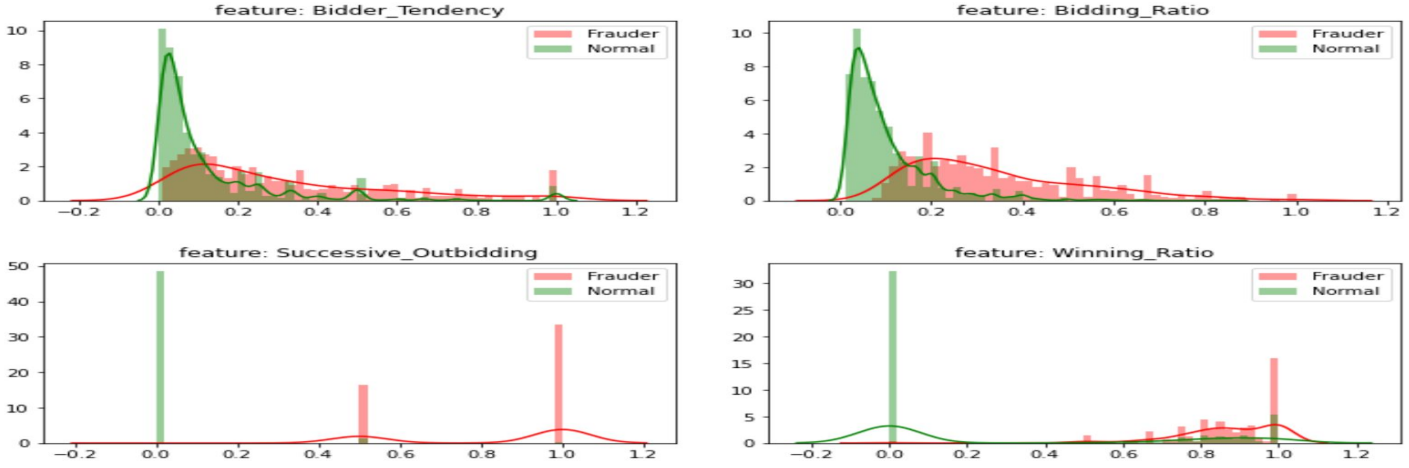
- Nombre d'individus : 6321.
- Nombre de variables : quantitative 8, qualitative 5.
- Analyses des valeurs manquantes : pas de valeurs manquantes.
- variable target (expliqué): Class (class 0 : normal, class 1 : fraudée).
- variables explicatives : Bidder_tendency, Bidding_Ratio, Successive_Outbidding, Winning_Ratio, Last_bidding, Auction_Bids, Starting_Price_Average, Early_Bidding ces dernieres sont de **types float**, Auction_Duration de **type int**.

III.Analyse univariée

- Analyse de la variable target (10.7 % de soumissionnaires fraudeurs "1", 89.3% normaux "0").
- Le jeu de données est normalisé : toutes les valeurs dans l'intervalle [0,1],sauf Auction_Duration qui signifie les jours de l'enchère (jour 1, jour 3, jour 5, jour 7 et jour 10).

IV.Analyse de comportement des variables avec la variable target

★ Après avoir représenté la distribution pour chacune des variables quantitatives, nous avons remarqué un comportement différent entre les deux classes pour les variables suivantes :



- **Bidder tendency** : la tendance normale varie de 0 à 0,2 tandis que celle d'une fraude varie de 0 à 1.
- **Bidding Ratio** : les enchères normales varient de 0 à 0,2 tandis qu'une fraude varie de 0,1 à 0,8.
- **Successive Outbidding :** dans le fraude 0,5 ou 1,0 tandis que véritable enchérisseur aura toujours 0.
- **Winning Ratio :** ratio de gain majoritaire d'un enchère normal est 0 alors que la fraude varie de 0,7 à 1.
- **Last_bidding, Auction_Bids,Starting_Price_Average ,Early Bidding, Auction_Duration :** leur distribution est pareille dans les deux classes.

  Donc nous pouvons supposer que les taux de Bidder_tendency, Bidding_Ratio, Successive_utbidding ainsi que Winning_Ratio semblent liés à l'action de fraude.

★ Le tableau ci-dessous représente la moyenne des variables explicatives quantitatives pour chaque classe.

Class	Bidder_Tendency	Bidding_Ratio	Successive_Outbidding	Last_Bidding	Auction_Bids	Starting_Price_Average	Early_Bidding	Winning_Ratio
0	0.122403	0.101775	0.016649	0.450286	0.227638	0.465605	0.423630	0.308242
1	0.310979	0.344268	0.832593	0.570463	0.264797	0.533181	0.489674	0.865322

Nous avons remarqué que les moyennes sont élevées dans les classes frauduleuses contrairement aux classes normales, nous pouvons dire à chaque fois qu'une variable est élevée alors elle fait référence à un comportement suspect .

V. Analyse bivariée

- Corrélation deux par deux pour des variables pertinentes

Dans cette partie nous présentons les ensembles de deux à deux de variables de notre DataSets qui ont une corrélation très importante,qui est plus proche de 1 .

Ensemble de variables	Corrélation
(Bidding_Ratio,Winning_Ratio)	0.642905
(Successive_Outbidding,Bidding_Ratio)	0.604828
(Successive_Outbidding,Class)	0.901035
(Last_Bidding,Early_Bidding)	0.950096
(Starting_Price_Average,Auction_Bids)	0.629086

- Par contre si nous séparons notre DataSets en deux ensembles par rapport à la variables Class nous trouvons les ensembles de variables les plus corrélées comme suits:

Ensemble de variables	Corrélation Class 0	Corrélation Class 1
(Last_Bidding,Early_Bidding)	0.959019	0.876006
(Starting_Price_Average,Auction_Bids)	0.626736	0.642655
(Bidding_Ratio,Winning_Ratio)	0.695769	pas trop élevé(-0.131250)

VII. Conclusion :

Après notre première étude des données SB on a remarqué différentes variables qui semblent être liées au fraude tel que Bidder tendency ,Bidding Ratio ,Successive Outbidding et Winning_Ratio. Il y en a aussi d'autres qui sont corrélés entre eux. Nous avons aussi remarqué que les valeurs des variables explicatives élevées font référence à un comportement suspect.