Introduction

L'objectif de ce travail est de prédire le type de trader : HFT (High Frequency Trading, utilisent des stratégies de trading à haute fréquence) ou NON HFT (Les traders non-HFT incluent une gamme de participants au marché qui n'utilisent pas de stratégies de trading à haute fréquence). Cependant, il existe un troisième type de trader « trader Mix » que nous avons dû supprimer dans la data-frame car il utilise à la fois les méthodes du trading HFT et Non HFT. Ainsi, garder le trader Mix peut biaiser les résultats de nos analyses puisqu'il sera difficile d'identifier la méthode réellement utilisée à ce moment-là.

Par ailleurs, les bases données contiennent plusieurs variables explicatives permettant de caractériser le comportement de trading. Ces variables explicatives comprennent des mesures de temps entre différents évènements, des mesures de fréquence de trading et des mesures de durée des annulations de transactions. Nous faisons l'hypothèse que ces variables explicatives peuvent aider à détecter les traders HFT ou NON HFT, car ces derniers ont tendance à avoir des temps moyens plus courts et de volumes de trading plus élevés.

Exploration, Nettoyage et préparation des données

Avant de commencer à construire nos modèles, nous avons pris le soin d'analyser et de comprendre la structure des données. Nous les avons nettoyé, car nous avons constaté qu'il y avait des valeurs manquantes, ce qui pourrait fausser les prédictions. Ensuite, nous avons créé une variable binaire « type trader » qui prend les valeurs 1 si le trader est du type HFT et 0 sinon. Cette variable binaire nous permet de mieux construire les modèles.

Enfin, nous avons divisé les données en un ensemble de données d'entraînement (training set) pour la construire des modèles, et un ensemble de données de test (test set) pour réaliser les prédictions sur les modèles. Pour ce faire, nous avons transformé en log certaines variables afin de limiter les valeurs extrêmes des fréquences.

Par ailleurs, j'ai calculé les corrélations entre les variables explicatives (x) et la variable dépendante (type trader). L'analyse de la corrélation et visualisation de la distribution de chaque variable (voir les histogrammes dans R), nous ont permis de supprimer certaines variables non utiles et de ne garder que celles qui sont fortement positivement corrélées avec le « type trader » (voir tableau 1). Cela permet de renforcer la précision des analyses (voir tableau 2).

Tableau 1 : Résultats des corrélations

Variables	**otr**	**ocr**	**omr**	**nbtradevenuemic**	**nbsecondwithatleatonetrade**
Type trader	0.36129465	0.01815309	0.25410895	0.42268459	0.327240615

Tableau 2 : Données finales en sortie

Dataset AMF 2021	**Données d'entraînement et de test**
Nous sommes passés de 86 observations de 15 variables à 48 observations de 8 variables	Train_set : 39 obs. of 8 variables Test set : 9 obs. of 8 variables

Choix des modèles

Pour prédire le type de trader, nous avons défini 3 modèles de classification : le modèle de regression logistique (glm), le modèle de forêt aléatoire (rf) et le modèle "svmRadial" : Support Vector Machines (SVM) avec noyau radial. Nous comparerons ces modèles afin de choisir le meilleur.

Le modèle glm : nous l'utilisons pour identifier les variables explicatives les plus importantes pour prédire la variable binaire dépendante. Cette sélection de variables nous permet de réduire le nombre de variables à considérer et d'améliorer la précision et l'efficacité du modèle.

Le modèle rf : nous l'utilisons, car il est capable de combiner plusieurs arbres de décision pour améliorer la précision des prédictions, le modèle offre des prédictions précises, une robustesse aux données manquantes et des mesures d'importance de variables, tout en étant rapide et facile à utiliser (voir tableau 3).

Le modèle svmRadial : nous l'avons utilisé pour faire face à la complexité et la non-linéarité des données, mais aussi pour capturer la structure sous-jacente des données que les autres modèles ne sont pas capables. Cela peut conduire à de meilleures performances de prédiction.

Analyse des résultats

Modèle glm : la déviance (~0,0000000002263) résiduelle est très proche de zéro, cela indique que le modèle "glm" ajuste très bien les données. Le modèle donne également un AIC très faible (16), cela indique que le modèle est à la fois bien ajusté et relativement simple (si l'on compare avec les autres modèles). En conséquent, le modèle peut être considéré comme meilleur (voir tableau 3).

Modèle rf : le modèle montre qu'à chaque fois qu'un nœud est divisé en deux sous-nœuds, seules trois variables sont prises en compte pour cette division. Le modèle a réussi à ajuster les données avec une certaine précision (mean of squared residuals 0,01797138). Le modèle est capable d'expliquer 87,8 % de la variabilité de la variable de réponse "type trader" (voir tableau 3).

Modèle svmRadial : les résultats montrent que le modèle a une erreur de formation (training error) de 0,17459. Cela signifie que le modèle prédit la variable cible avec une erreur moyenne d'environ 17,46%. Le modèle compte 24 vecteurs de support (support vectors), qui sont les points de données les plus proches de la frontière de décision (voir tableau 3).

Tableau 3 : Résultats des modèles

Modèle glm	Modèle rf	Modèle svm
Degrees of Freedom: 38 Total (i.e. Null); 31 Residual Null Deviance : 36.71 Residual Deviance : 0,000000000226 AIC: 16	Mean of squared residuals: 0.01797138 % Var explained: 87.8	Gaussian Radial Basis kernel Hyperparameter : sigma = 0.1796 Number of Support Vectors : 24 Training error : 0.17459

Les valeurs dans le tableau 4 représentent les niveaux et/ou les catégories des classes de trader humain (Non HFT) et ordi (HFT). Pour le modèle rf par exemple, les valeurs 0,001, 0,008 et 0,77033 correspondent respectivement à la fréquence de prédiction "Non HFT" pour la classe "Non HFT" (vraie négative), la fréquence de prédiction "HFT" pour la classe "Non HFT" (faux positif) et la fréquence de prédiction "HFT" pour la classe "HFT" (vraie positive). Pour l emodèle glm, la raison pour laquelle les deux valeurs sont identiques est que le modèle fournit une seule probabilité de prédiction, qui est la probabilité d'appartenance à la classe "Non HFT". Pour le modèle svmRadial, il n'y a pas de faux positif car toutes les prédictions sont inférieures à la valeur de seuil de 0,5 (voir tableau 4).

Tableau 4 : Fréquences de prédiction pour les catégories des classes de trader humain (non HFT) et ordi (HFT)

Classes	Modèle glm	Modèle rf	Modèle svm
"Non HFT" (vraie négative)	-26.566 (Non HFT)	0.001 (Non HFT)	0.0072851 (Non HFT)
"Non HFT" (faux positif)	-26.566 (HFT)	0.008 (HFT)	
"HFT" (vraie positive)		0,77033 (HFT)	0.2798 (HFT)

Les résultats du tableau 5 montrent la précision de chaque modèle sur les données de test. Les modèles glm et rf ont une précision de 100% pour les classes de trader humain (Non HFT) et ordinateur (HFT). Le modèle svmRadial a une précision de 0.7777778, cela signifie que 77,78% de ses prédictions sont correctes. La précision est une métrique importante pour évaluer les performances d'un modèle. Le modèle glm et rf sont de meilleurs modèles pour prédire le type de trader (voir tableau 5).

Tableau 5 : Résultats d'évaluation de la performance des modèles avec précision

Modèle glm	Modèle rf	Modèle svm
1	1	0.7777778

Conclusion

En somme, pour évaluation les différents modèles sur la data challenge AMF 2021, nous avons visualisé les prédictions de chaque modèle avec les matrices de confusions pour les classes de trader humain et HFT. Les résultats indiquent que le modèle glm est le meilleur avec un pourcentage de prédiction de 99% contre 73% et 40% respectivement pour les modèles rf et svmRadial (voir tableau 6). Enfin, pour répondre à la question de savoir qui opère derrière le trade, nous pouvons utiliser le modèle qui a obtenu la meilleure précision sur le jeu de test. C'est ainsi que nous avons utilisé utilisé le modèle de forêt aléatoire (random forest) pour la meilleure prédiction. Les résultats montrent que le modèle de forêt aléatoire a une précision de 87,8% pour prédire qui opère derrière le trade. Cela suggère que le modèle n'est pas très précis dans la prédiction de cette variable (voir R). Le modèle glm est le meilleur modèle parmi les modèles testés.