Année 2022/2023

Université Paris-Nanterre

UFR SEGMI

**Exploitation des données massives en finance**

**<Application de la régression Ridge et Lasso pour prédire le prix de l'action Hermès>**

M2 GDA

YANG Lulu

TIAN Minghao

**1 Introduction**

Afin d'appliquer la régression Ridge et la régression Lasso pour prédire les tendances du prix des actions de Hermès, nous avons trouvé des données de négociation d'actions pour quatre sociétés (Hermès, LV, Dior et Kering) du 1er janvier 2000 au 19 mai 2022 à partir de données accessibles au public. Cet ensemble de données contient le prix d'ouverture, le prix de clôture, le prix haut, le prix bas, le prix ajusté et le volume de transactions des actions des quatre sociétés.

Après le traitement des données, nous avons utilisé uniquement les données boursières de LV, Dior et Kering comme données d'entrée, entraînées par des modèles de régression linéaire, de régression Ridge et de régression Lasso, et avons directement prédit les informations complètes sur le prix des actions de Hermès. Les résultats montrent que toutes les informations sur le prix des actions de Hermès, à l'exception du volume des transactions, sont prédites de manière crédible sans aucun paramètre d'entrée pour Hermès. La régression Lasso obtient un avantage significatif par rapport à la régression linéaire et à la régression Ridge. Enfin, la précision de la prédiction a été encore améliorée en appliquant la régression Lasso aux autres informations sur les actions de Hermès en ajoutant le cours d'ouverture de l'action de Hermès aux données d'entrée. Ce projet démontre que les données boursières d'une société peuvent être prédites par le même type de données boursières de la société, et qu'avec la connaissance supplémentaire du prix d'ouverture des actions de la société, des prédictions précises peuvent être faites pour le reste du prix des actions de la société.

**1.1 Introduction à la méthode**

La régression est une technique utilisée dans le processus de modélisation pour analyser la relation entre les variables, et la façon dont les variables affectent le résultat. Dans ce document, la régression linéaire, la régression Ridge et le Lasso sont appliqués pour prédire les prix des actions respectivement.

**La régression linéaire** est une analyse de régression qui utilise une fonction des moindres carrés pour modéliser la relation entre une ou plusieurs variables indépendantes et la variable dépendante. Elle est simple et couramment utilisée, mais n'est généralement pas efficace dans les situations complexes.

**La régression Ridge** est une méthode de régression biaisée spécifiquement conçue pour l'analyse de données covariables, et est essentiellement une forme modifiée de l'estimation par les moindres carrés. La méthode permet d'obtenir des coefficients de régression plus réalistes en renonçant à la nature non biaisée des moindres carrés.

**La régression Lasso** est très similaire à la régression Ridge dans la mesure où il ajoute également un terme de biais à la fonction d'optimisation de la régression afin de réduire l'effet de la covariance. La différence est que la régression Lasso utilise un biais de valeur absolue comme terme de régularisation.

**1.2 Description de l'ensemble des données**

L'ensemble de données utilisé dans cet article contient des données quotidiennes sur les transactions boursières de quatre sociétés (Hermès, LV, Dior et Kering) du 1er janvier 2000 au 19 mai 2022.

表格

描述已自动生成

Figure 1.1 Aperçu de l'ensemble de données

PA, MC.PA, CDI.PA et KER.PA sont les noms des actions de Hermès, Louis Vuitton, Christian Dior et Kering respectivement. Adj Close, Close, High, Low, Open et Volume désignent respectivement le prix ajusté, la clôture, le haut du jour, le bas du jour, l'ouverture et le volume.

Le prix ajusté est la valeur après prise en compte complète des différentes actions de l'entreprise et est souvent utilisé pour vérifier les rendements historiques ou pour une analyse détaillée des performances passées. Le cours d'ouverture, également appelé cours d'ouverture du marché, est le prix auquel s'effectue la première vente ou le premier achat d'un titre par action après l'ouverture de chaque journée de négociation sur une bourse. Le cours de clôture est le prix auquel la dernière transaction d'un titre a été effectuée avant la fin de l'activité boursière de la journée. S'il n'y a pas de transaction ce jour-là, le dernier prix négocié est utilisé comme prix de clôture. Les prix les plus élevés et les plus bas sont les prix les plus élevés et les plus bas, respectivement, auxquels une action individuelle a été négociée ce jour-là. Le volume est le nombre de transactions négociées pour une opération donnée dans une unité de temps.

Nous avons essayé de prédire le cours complet de l'action Hermès en utilisant le cours de l'action Louis Vuitton, Christian Dior et Kering.

**2 Développement**

**2.1 Traitement des données**

Les données d'origine sont un tableau csv de 22941 lignes et 8 colonnes, et les informations de plusieurs entreprises sont empilées ensemble, ce qui rend difficile la division ultérieure des données.Il y a un décalage dans la quantité de données pour les quatre sociétés, ce qui indique que les informations sur le prix des actions à certaines dates sont incomplètes et qu'il peut y avoir une situation où il n'y a pas d'enregistrement du prix des actions d'une société. Lors de la vérification de la table inverse réelle, il a également été constaté qu'il n'y avait pas d'informations sur le prix de l'action pour une société donnée à une date donnée. Pour résoudre ce problème, nous avons d'abord regroupé le tableau total par symbole, puis nous avons effectué une fusion horizontale de deux tableaux par date, en ne fusionnant que les lignes qui avaient la même date.

表格

描述已自动生成

Figure 2.1 Aperçu des données traitées

Après avoir fusionné puis modifié les intitulés des colonnes, les résultats sont présentés à la figure 2.1. Après les ajustements ci-dessus, le tableau de données original est passé de 22941 lignes et 8 colonnes à 5724 lignes et 29 colonnes.

**2.2 Analyse des données**

L'histogramme de la distribution de toutes les variables a d'abord été visualisé et les résultats sont présentés dans la figure 2.2.

一些文字和图案

中度可信度描述已自动生成

Figure 2.2 Histogramme de la distribution de chaque variable

La figure ci-dessus montre que la distribution de l’information dans les variables est approximative, à l’exception de l’information VOLUME pour chaque entreprise. Cela montre qu’il est difficile de prévoir le trafic futur.

Afin d'examiner plus en détail la relation entre le VOLUME et le reste des informations sur le prix des actions, Hermès a été utilisé comme exemple pour visualiser l'ensemble des informations sur le prix des actions de Hermès, comme le montre la figure 2.3.

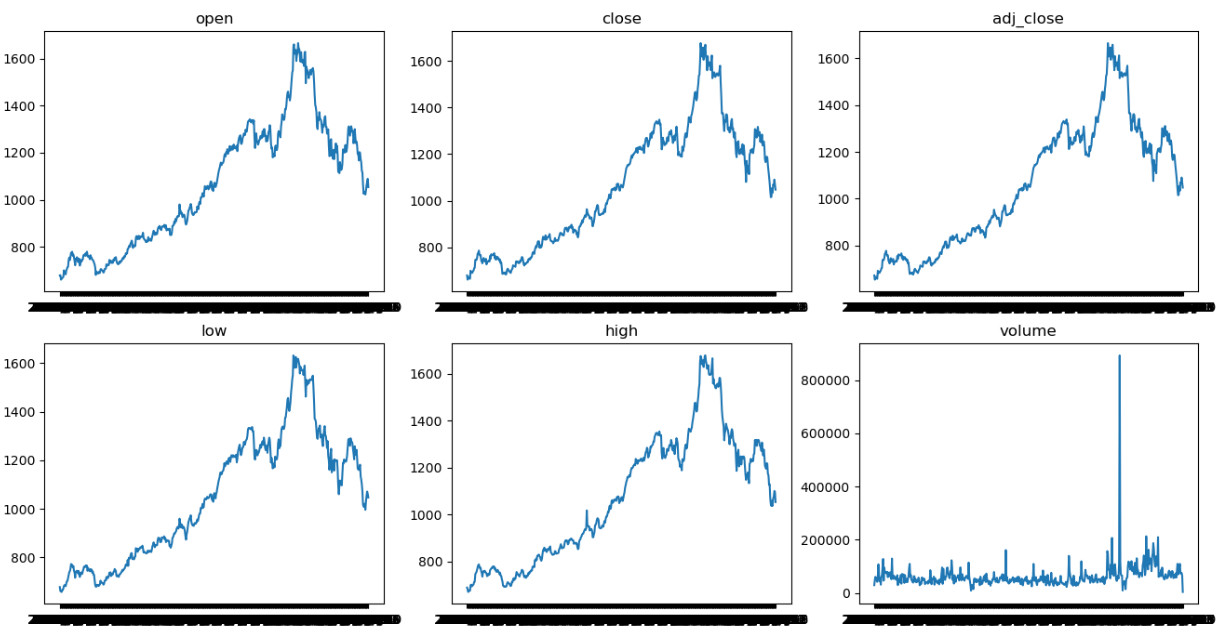


Figure 2.3 Informations sur l'évolution du cours de l'action Hermès 2020 depuis le début de l'année

Conformément à l'analyse ci-dessus, la tendance des flux et des autres informations sur le prix des actions est plus différente. La raison spécifique de ce phénomène peut être expliquée par le fait que toutes les autres informations sont corrélées avec le prix de l'action, tandis que le prix de l'action et le nombre de transactions du jour sont moins corrélés.

Pour vérifier la relation entre les entreprises ayant les mêmes informations boursières, les prix de clôture ajustés des quatre entreprises ont été analysés. Comme le montre la figure 2.4(a), il existe une corrélation plus significative entre les mêmes informations sur les actions de différentes sociétés.

图表, 散点图

描述已自动生成 图表

描述已自动生成

Figure 2.4 (a) Relations entre les prix de clôture ajustés pour les quatre entreprises (b) Tableau de corrélation pour les données complètes

Pour vérifier davantage les résultats ci-dessus, des corrélations ont été calculées pour toutes les variables de données et reportées dans le tableau de corrélation, comme le montre la figure 2.4(b). Il ressort clairement du graphique que les informations relatives au cours de l'action de chaque société sont fortement corrélées, ce qui peut s'expliquer par le fait que les quatre sociétés considérées dans ce document sont toutes des entreprises françaises de produits de luxe dont le contenu commercial est similaire, ce qui rend évidente la corrélation entre leurs cours. Deuxièmement, la présence d'une zone sombre régulière dans le tableau de corrélation, qui regroupe toutes les données de volume pour chaque entreprise, suggère que le nombre de transactions est moins corrélé avec le prix de l'action lui-même. La forte corrélation des informations sur le prix des actions entre les différentes sociétés suggère qu'il est possible d'utiliser les informations sur le prix des actions de seulement trois sociétés pour prédire les informations sur le prix des actions de Hermès lui-même.

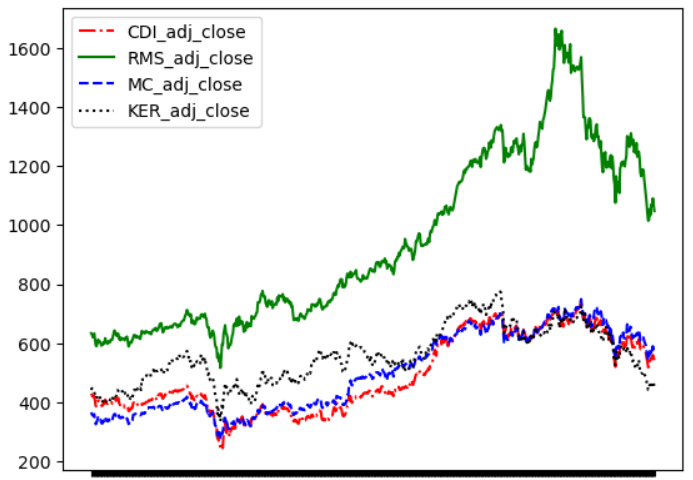
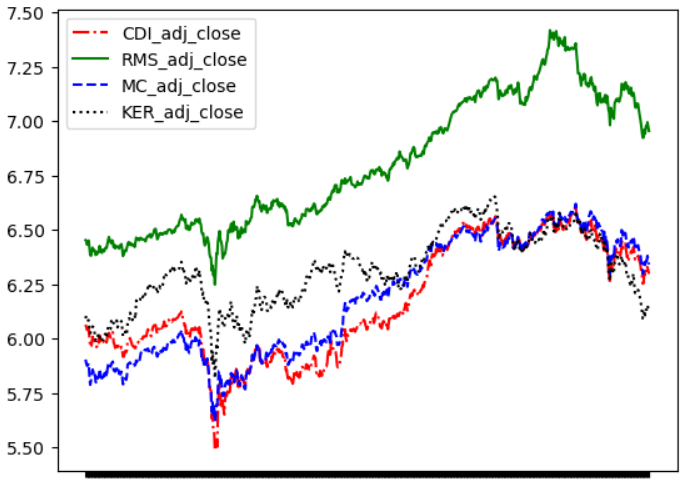
 

Figure 2.5 (a) La tendance des cours de clôture ajustés pour quatre marques sur une période donnée. (b) Logarithmique, la tendance du cours de clôture ajusté de quatre marques sur une période de temps.

Afin d’afficher les informations sur le cours des actions des quatre marques de manière plus intuitive, nous prenons les cours de clôture ajustés des quatre marques à un moment donné et les affichons dans la Figure 2.5 (a). Nous pouvons voir que les cours boursiers des quatre marques ont une tendance similaire. Cela montre qu’il existe une corrélation entre les marques de luxe dans les informations boursières dont nous avons besoin. Dans le même temps, la ligne verte des informations sur le cours de l’action Hermès représentée dans la figure est nettement éloignée des autres données, ce qui augmente la difficulté des prévisions ultérieures du modèle. La Figure 2.5(b) est le résultat logarithmique de toutes les données de la figure de gauche. Il est clair que les données se sont considérablement réduites tout en maintenant les tendances et les relations relatives. Cela facilite les prédictions ultérieures du modèle. Par conséquent, nous utiliserons des données logarithmiques pour l’entraînement et la prédiction de modèles, qui est également un moyen auxiliaire couramment utilisé dans l’apprentissage automatique.

**2.3 Application du modèle de régression**

Pour appliquer le modèle de régression, il faut d'abord diviser les ensembles de formation et de test, ainsi que les données d'entrée et de sortie. Nous avons divisé les 4000 premiers ensembles de données par nombre dans l'ensemble d'entraînement et les 1724 ensembles de données restants dans l'ensemble de test, pour un rapport de division global proche de 7:3. Comme les données sont triées dans le temps, cela revient à utiliser un modèle entraîné de 2000 à 2015 pour faire des prédictions sur les prix des actions de 2016 à 2022. Ensuite, nous avons utilisé les données de Hermès, Christian Dior et Kering comme entrées pour prédire respectivement l'ouverture, la fermeture, le haut, le bas, la fermeture ajustée et le volume total des transactions de Louis Vuitton.

Après avoir divisé les données, nous avons appliqué les modèles LineraRegression, Ridge et Lasso dans sklearn.linear\_model pour ajuster les régressions respectivement.

Le modèle linéaire et les moindres carrés :

Soit un vector d’input X′ = (X1, X2, . . . , Xp), on prédit l’ouput Y via le modèle

黑色的钟表

低可信度描述已自动生成

yi = β0 +β1x1i +···+βpxpi +εi, εi ∼ NIID(0,σ2)

On oublie le principe du sans-biais, et la précision de la prévision : les OLS donnent une estimation sans biais mais généralement avec une large variance. Avec beaucoup de variables explicatives, on veut généralement sélectionner un sous-ensemble qui a les effets les plus important.

La régression Ridge (et Lasso) utilise la méthode de contraction des coeﬀicients (shrinkage coeﬀicients) qui introduit un petit biais mais diminue drastiquement la variance.

文本, 信件

描述已自动生成

Équation de régression Ridge

文本, 信件

描述已自动生成

Équation de régresssion Lasso

Nous pouvons utiliser l'erreur quadratique moyenne (RMSE) pour mesurer la qualité du modèle, mais comme la RMSE est sensible à la magnitude de la variable dépendante (cible) et de la variable indépendante (caractéristiques), elle présente certaines lacunes dans la mesure de l'adéquation du modèle aux données. Nous pouvons donc utiliser une autre mesure de l'adéquation, le coefficient de détermination (R2).

Le coefficient de détermination, R2, est le rapport entre la somme des carrés de la régression et la somme totale des carrés, et sa valeur est comprise entre 0 et 1. Sa magnitude reflète la contribution relative de la régression, c'est-à-dire le pourcentage de la variation totale de la variable dépendante Y qui est expliquée par la relation de régression. R2 est la mesure la plus couramment utilisée pour évaluer le mérite d'un modèle de régression.

En général, un R2 plus élevé indique un degré plus élevé d'explication de la variable dépendante par la variable indépendante et un pourcentage plus élevé de la variation totale causé par la variable indépendante. Plus les points d'observation sont denses autour de la ligne de régression.

**3 Résultats**

**3.1 Comparaison des résultats**

Afin d'appliquer la régression Ridge et la régression Lasso, il faut rechercher les coefficients alpha des deux et sélectionner les coefficients appropriés pour l'ajustement. Comme le montre le tableau 3.1, les différents coefficients alpha ont un impact important sur les résultats (où plus le score r2 est proche de 1, meilleur est le résultat et plus petite est la RMSE).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| alpha | 0.0001 | 0.001 | 0.01 | 0.1 | 1 |
| r2 score | 0.93 | **0.95** | 0.77 | 0.12 | -17.94 |
| RMSE | 0.11 | **0.10** | 0.21 | 0.42 | 1.99 |

Tableau 3.1 Impact des différents coefficients alpha sur les résultats de prédiction des prix de clôture ajustés dans le cadre de la régression Lasso

Nous avons utilisé RidgeCV et LassoCV de sklearn.linear\_model pour rechercher les coefficients et sélectionner les coefficients les plus appropriés à appliquer aux deux modèles. Les résultats de la prédiction finale de la régression linéaire, de la régression Ridge et de la régression Lasso pour le prix d'ouverture, le prix de clôture, le prix haut, le prix bas, le prix de clôture ajusté et le volume total des transactions de Hermès sont présentés dans le tableau 3.2.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | adj\_close | open | close | low | high | volume |
| Linear | R2 score | 0.70 | 0.68 | 0.67 | 0.65 | 0.69 | -24.94 |
| RMSE | 0.25 | 0.25 | 0.25 | 0.26 | 0.25 | 2.29 |
| Ridge | R2 score | 0.70 | 0.68 | 0.68 | 0.66 | 0.69 | -24.67 |
| RMSE | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 | 2.28 |
| Lasso | R2 score | **0.93** | **0.92** | **0.92** | **0.92** | **0.93** | **-10.45** |
| RMSE | **0.12** | **0.12** | **0.12** | **0.13** | **0.12** | **1.52** |

Tableau 3.2 Résultats statistiques de la régression linéaire, de la régression Ridge et de la régression Lasso pour toutes les informations sur les actions des sociétés de Hermès

Toutes les données sont optimales sous la régression Lasso, ce qui montre pleinement la performance supérieure de la régression Lasso. Dans le même temps, les flux de stocks sont imprévisibles dans les trois modes.

La figure ci-dessous montre les avantages de la régression Lasso par rapport à la régression linéaire et de la régression de Ridge de manière plus intuitive.

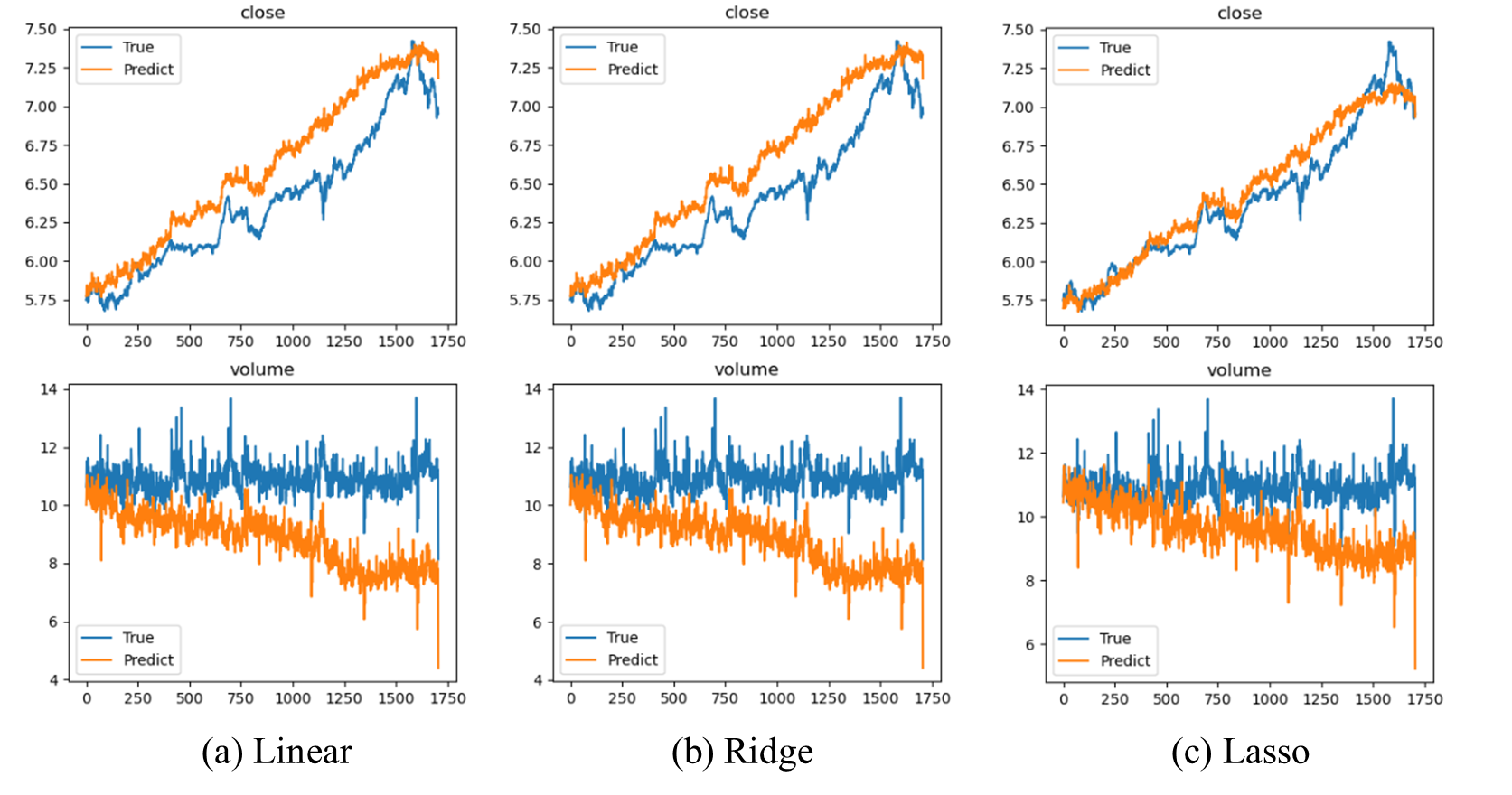


Figure 3.1 Prévisions de prix de clôture et de trafic Hermès sous régression linéaire, régression de Ridge et régression Lasso

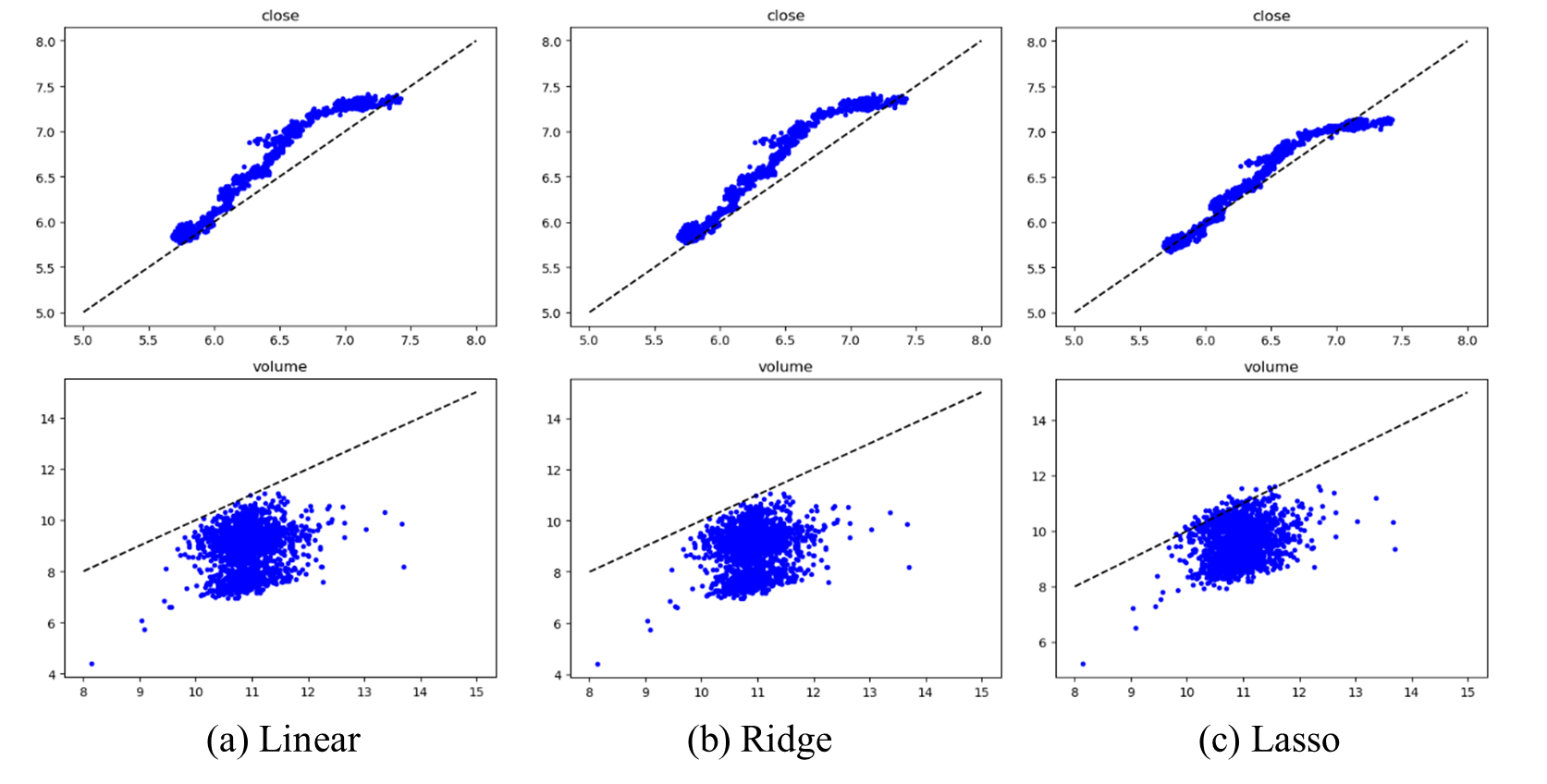


Figure 3.2 Diagramme de dispersion des régressions linéaires, Ridge et Lasso prédisant le prix de clôture et le nombre de transactions pour les sociétés de Hermès

Les diagrammes de dispersion fournissent une vue plus visuelle des résultats de la régression. Comme le montre la figure 3.2, nous comparons les résultats prédits des trois modèles pour le prix de clôture et le nombre de transactions. Où la position d'un point est déterminée par l'ensemble de ses coordonnées horizontales et verticales, et où ses coordonnées horizontales et verticales sont respectivement les résultats réels et prédits. La ligne pointillée noire du graphique indique la position où les coordonnées horizontales et verticales sont exactement équivalentes. Par conséquent, plus le point de dispersion est proche de la ligne pointillée, meilleur est le résultat de la prédiction par régression. L'amélioration de la régression Lasso par rapport à la régression linéaire et à la régression Ridge est également apparente dans le graphique.

**3.2 Ajouter la régression Lasso après le prix d’ouverture**

Après les calculs et l'analyse ci-dessus, nous avons constaté que la régression Lasso était celle qui traitait le mieux le problème. Pour améliorer encore les résultats, nous avons également ajouté le prix d'ouverture de Hermès pour la journée aux données d'entrée, puis nous avons fait des prédictions pour le prix de clôture de Hermès, le prix élevé et d'autres informations sur les actions.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | adj\_close | close | low | high | volume |
| Before | R2 score | 0.93 | 0.92 | 0.92 | 0.93 | -10.45 |
| RMSE | 0.12 | 0.12 | 0.13 | 0.12 | 1.52 |
| After | R2 score | **0.99** | **0.99** | **0.99** | **0.99** | **-8.14** |
| RMSE | **0.01** | **0.01** | **0.01** | **0.01** | **1.36** |

Tableau 3.3 Comparaison des résultats avant et après l'ajout du prix d'ouverture

Comme le montre le tableau 3.3, avec l'ajout du prix d'ouverture comme "point d'ancrage" pour l'action, les informations sur le prix de chaque action sont mieux prédites. Les scores R2 pour les prix de clôture, minimum et maximum sont tous très proches de 1. Cela indique que les prix de clôture, de clôture ajustée et maximum et minimum des actions de la société cible peuvent être prédits avec une grande précision en utilisant les informations sur la variation du prix de l'action de la société concernée, combinées au prix d'ouverture du jour.

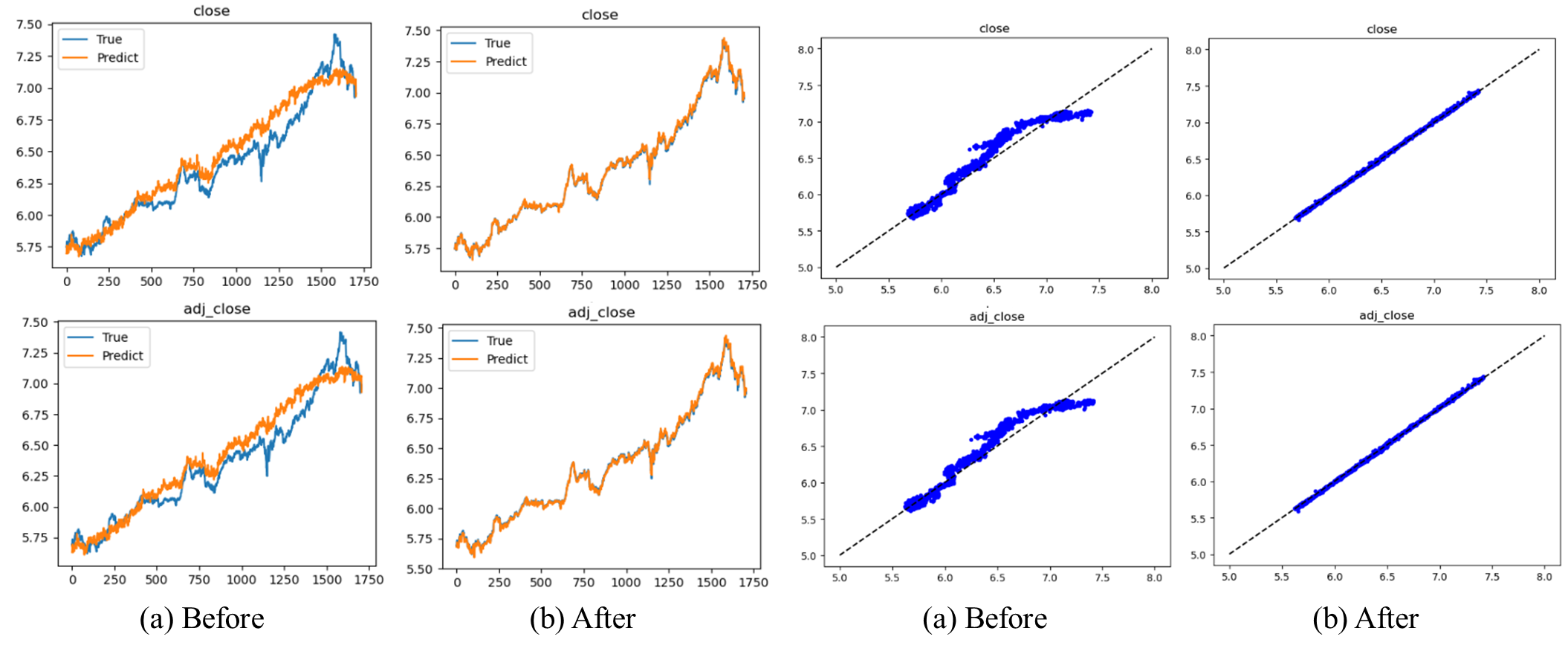
Dans le même temps, les indicateurs de volume, bien qu’améliorés, restent imprévisibles. Par conséquent, cette amélioration valide davantage le fait qu’il est impossible de prédire le volume des transactions de la société cible le jour même en se basant uniquement sur les cours des actions et le volume des transactions des autres sociétés.

Figure 3.3 Avant et après l’ajout du prix d’ouverture, Lasso revient à comparer les résultats du cours de clôture de l’action et à ajuster le cours de clôture

La figure 3.3 montre une comparaison des résultats pour les prix de clôture de la Hermès et les prix de clôture ajustés après ajout du prix d'ouverture de la Hermès. Nous n'avons pris que le prix de clôture et le prix de clôture ajusté (voir le code pour le graphique complet) et les avons représentés respectivement sur une ligne et un graphique en nuage de points. L'augmentation de la précision des prévisions est clairement visible dans les graphiques.

**4 Conclusion**

Le but de ce projet était d'appliquer la régression Ridge et la régression Lasso pour prédire les informations sur les stocks de LV. Nous avons trouvé des données sur les transactions boursières de quatre sociétés de luxe (Hermès, LV, Dior et Kering) du 1er janvier 2000 au 19 mai 2022 à partir de données accessibles au public. L'ensemble de données contient le prix d'ouverture, le prix de clôture, le prix le plus élevé, le prix le plus bas, le prix ajusté et le volume de transactions des actions des quatre sociétés.

Après avoir traité et analysé les données, nous avons constaté que les corrélations telles que les prix des actions et les changements entre plusieurs entreprises étaient significatives, et que le nombre de transactions était moins corrélé. Nous avons ensuite appliqué la régression linéaire, la régression Ridge et la régression Lasso pour entraîner les modèles et prédire les informations sur les transactions boursières de Hermès de 2015 à la plus récente, respectivement. Les résultats montrent que les trois régressions peuvent être entraînées en utilisant les données passées pour faire des prédictions efficaces sur les données de négociation futures. Les résultats montrent que la régression de Ridge n’améliore pas significativement les données de ce projet, mais la régression de Lasso obtient les meilleurs résultats pour tous les indicateurs.

Pour améliorer encore la prédiction, nous avons finalement ajouté le cours d'ouverture du jour de Hermès aux données d'entrée afin de tenter de prédire d'autres informations sur les actions de Hermès. Les résultats montrent qu'il est possible de faire des prédictions très précises sur les autres informations boursières de la société cible en se basant uniquement sur les informations boursières d'autres sociétés liées et sur le cours d'ouverture de la société cible.

Ce projet montre qu’en appliquant les informations sur les stocks passés à la formation, divers modèles de régression peuvent être utilisés efficacement pour les prévisions futures. Les informations boursières des sociétés liées dans le même secteur peuvent prédire efficacement le cours de l’action et les changements de la société cible, et une prédiction de haute précision des données boursières de la société cible peut être obtenue après avoir ajouté une petite quantité d’informations supplémentaires.

**5 Reference**

Sources de données：<https://www.datacastle.cn/dataset_list.html>

Berk (2008). Statistical learning from a regression perspective

James et al. (2013). An introduction to statistical learning