2023

MACHINE LEARNING : PREDICTION DU COURS BOURSIER D'APPLE

ELISA CORNUAULT
GODFROY AYISSI EKANI
SALEH CHENITTI

SOMMAIRE

I. DESCRIPTION ET METHODOLOGIE DE L'ETUDE

Dans notre travail nous nous intéressons à l'étude du cours de l'action Apple depuis sa création en 1985 jusqu'en 2022 en fonction de l'inflation américaine, le treasury yield, le federal fund rate, et l'indice S&P500, à travers une base de données, constituée et regroupée par ordre chronologique.

Nous verrons que dans cette étude, la question de la prévision des cours boursiers à travers l'usage des méthodes de Machine Learning met en exergue la difficulté de prévoir des cours avec exactitude. Puisque en effet, pour mener à bien ce projet, nous allons utiliser le Machine Learning. Deux types d'algorithme de machine Learning sont utilisés : l'apprentissage supervisé et l'apprentissage non supervisé. Notre étude sera sous la base de l'apprentissage supervisé. Pour ce faire, notre base de données a été divisée en deux : un échantillon d'entrainement et un échantillon test. Sur cet échantillon d'entrainement, les données sont déjà étiquetées et le système utilise cet échantillon d'entrainement, afin de pouvoir par la suite faire de la prévision sur l'échantillon test. Une autre distinction est importante à faire : le type de sortie que l'on attend de notre programme. Lorsque notre variable de sortie est continue, on parle de régression. Lorsque cette variable est discrète, on parle de classification. Donc notre cas, nous allons utiliser des modèles de régression et de classification.

II. PRESENTATION DE LA BASE DONNEES

L'objectif de ce projet est de pouvoir prédire l'évolution du cours boursier d'Apple en fonction de l'inflation américaine, le treasury yield américain (échéance 2 ans), le federal funds rate américain et le S&P500. Le federal funds rate représente l'intérêt que les banques se facturent pour des prêts de 1 jour (over night). Le treasury yield est le taux des obligations du Trésor américain (c'est le taux d'intérêt actuel dont bénéficierait un investisseur s'il achetait aujourd'hui des bons du Trésor américain à échéance 2 ans). Nous nous sommes concentrés sur la période Février 1985 – Décembre 2022, les données sont mensuelles et notre base de données contient au total 455 observations. Nous avons importé les données de l'indice boursier Apple et SP500 via le site API Yahoo finance, et nous avons utilisé Alpha Vantage pour l'extraction de l'inflation américaine, le treasury yield américain ainsi que le federal funds rate américain.

III. ANALYSE DESCRIPTIVE DES DONNEES

Avant toute prédiction, nous allons analyser notre base de données.

Les annexes 1 et 2 représentent l'évolution de chaque variable de 1985 à 2022. D'emblée, le federal funds rate et le treasury yield évoluent dans le même sens toute au long de la période d'observation et ces deux variables ont une tendance à la hausse de 1985 à 2022. Également, le S&P500 et le cours de Apple suivent une évolution similaire tout au long de la période étudiée et ces deux variables ont plutôt une tendance générale à la baisse de 1985 à 2022. Ce qui peut paraitre cohérent puisque Apple est une composante des 500 grandes sociétés cotées sur les bourses aux Etats-Unis, faisant partie du S&P500. Intuitivement, sachant qu'il existe une relation inverse entre les actions, les indices et les taux, on pourrait déduire que l'augmentation du federal funds rate et du treasury yield explique en partie la baisse de Apple et du S&P500. Enfin, la relation de l'inflation avec les autres variables n'est pas à priori mise en évidence graphiquement.

A la lumière de l'annexe 3, nous identifions les différentes relations entre les différentes variables utilisées dans l'étude. On note ici la relation linéaire et croissante entre le treasury yield et le federal funds rate, ainsi qu'une relation similaire entre le cours de Apple et le S&P500. Ce qui s'aligne avec les constations faites précédemment.

L'annexe 4 décrit la nature des corrélations entre des couples de variables. L'intuition évoquée précédemment est confirmée par cette matrice. En effet, une corrélation forte et positive entre le treasury yield et le federal funds rate existe. Il en est de même pour la corrélation entre le S&P500 et Apple. On peut donc dire que le S&P500 est une variable qui va fortement aider dans la prédiction de Apple.

On relève la corrélation négative entre le S&P500 et Apple, avec le treasury yield et le federal funds rate. Enfin, L'inflation ne semble être corrélée que faiblement avec les autres variables puisque son coefficient de corrélation est proche de 0. Par conséquent, l'ajout de cette variable pour expliquer le cours d'Apple ne devrait pas contribuer à priori à la prédiction. Il sera donc question d'explorer cela dans les développements subséquents.

IV. MISE EN APPLICATION DES METHODES DE PREDICTION

1. LA CLASSIFICATIOIN

D'après l'annexe 5, le cours d'Apple est une variable qui est continue car elle peut prendre un nombre infini de valeurs réelles possibles. C'est pourquoi, les modèles de prédiction de régression seraient le plus approprié. Cependant, pour pouvoir utiliser des modèles de classification, nous avons modifié la variable Y pour qu'elle devienne une variable Dummy : en prenant la valeur 0 si l'indice diminue d'un mois sur l'autre, et prenant la valeur 1 si l'indice augmente d'un mois sur l'autre. Cette variable transformée en variable Dummy nous permet d'avoir une variable discrète, permettant de faire de la classification. Nous arrivons donc à cette nouvelle base de données à l'annexe 6.

En outre, il est important de déterminer si notre data frame est équilibré ou déséquilibré (c'est-à-dire que certaines variables Y apparaissent en faible minorité). Un data frame déséquilibré peut poser problème dans la mesure où les algorithmes de machine learning ont du mal à apprendre sur cette minorité qui nous intéresse, ce qui aurait comme conséquence une mauvaise prédiction du label. Dans notre cas, d'après l'annexe 7, 56,04% des données représentent les rendements positifs et 44% les rendements négatifs. Par conséquent, nous n'avons pas besoin d'effectuer un rééchantillonnage car notre data frame est relativement équilibré.

Après avoir transformé notre variable en une variable discrète et après avoir vérifié l'équilibrage de notre data frame, nous pouvons passer à la séparation du data frame en échantillon train et échantillon test, avant de s'attaquer aux modèles de classification. En effet, avec cette séparation, l'algorithme pourra s'entrainer sur l'échantillon train, et ensuite faire la prédiction en utilisant l'échantillon test. Dans notre cas, 75% de notre data set seront des données d'entrainement, et 25% seront les données permettant de faire la prédiction.

Le tableau suivant résume l'ensemble des résultats des modèles utilisées pour prédire le cours d'Apple en fonction du treasury yield, federal funds rate, inflation et S&P500. Pour chacun des modèles, nous avons optimisé les hyperparamètres par la cross validation :

Modèle	Résultats	Matrice de confusion [Annexes 8, 9, 10, 11]
KNN	57,02%	Good classification for negativ return : 41,3% Good classification for positive return : 67,65%
Random Forest	39,47%	Good classification for negativ return : 95,65% Good classification for positive return : 1,47%
Naive Bayes	50,9%	Good classification for negativ return : 45,65% Good classification for positive return : 54,41%
SVM	59,65%	Good classification for negativ return : 0% Good classification for positive return : 100%

Les résultats ne sont pas très intéressants (surtout pour le modèle Random Forest). Le modèle le plus performant est le SVM, qui obtient un résultat égal à 59,65%, mais en ne sachant prévoir aucun rendement négatif d'après la matrice de confusion. A l'inverse, le Random Forest prédit correctement les rendements négatifs (95,65%), mais incorrectement les rendements positifs (1,47%), d'après la matrice de confusion. Ces résultats médiocres prouvent finalement que les variables explicatives ne permettent pas de bien prédire l'indice boursier Apple, en suivant les modèles de classification.

Grâce aux graphiques et à la matrice de corrélation précédents, nous avons supposé que la corrélation entre le cours d'Apple et l'inflation était relativement bas. Ce qui signifierait que la variable inflation ne serait pas utile pour expliquer le cours d'Apple (donc pour prédire cette dernière). Le tableau suivant est la preuve de cette supposition. En effet, nous pouvons remarquer que les résultats des modèles sont quasiment similaires entre l'ajout ou le retrait de la variable inflation, les résultats sont même meilleurs (excepté pour le modèle SVM) :

Modèle	Résultats avec Inflation	Résultats sans Inflation
KNN	57,02%	59,65%
Random Forest	39,47%	42,98%
Naive Bayes	50,9%	54,4%
SVM	59,65%	57,02%

Finalement le retrait de la variable, du fait de sa très faible corrélation avec l'indice boursier d'Apple, permet d'améliorer légèrement la prédiction de cette dernière.

N'ayant pas obtenu des résultats satisfaisants en utilisant la méthode classification, nous allons construire des modèles de régression afin d'essayer d'obtenir de meilleurs résultats avec la régression. Pour cela, nous allons reprendre la base de données de l'annexe 5 car dans ce data frame, la variable à prédire est une variable continue.

2. LA REGRESSION

Avant de construire ces modèles de régression, nous allons de nouveau séparer le data frame en échantillon train et test. Puis nous allons utiliser le modèle Lasso, ainsi que le Random Forest.

- **La régression Lasso**

La méthode du Lasso est une méthode de régularisation qui pénalise la valeur absolue des lambdas. La régularisation est une technique qui aidera un modèle à fonctionner sur des données qu'il n'a jamais vues auparavant. Donc cette méthode est utile dans une situation où nous n'avons pas autant

de données que souhaité (ce qui est notre cas), la régularisation du modèle aidera à généraliser le modèle, plutôt que de sur-ajuster les données dont on dispose. Cette méthode est également généralement utilisée lorsqu'on soupçonne de la multicolinéarité. Les résultats obtenus sont les suivants :

Eléments	Echantillon d'apprentissage	Echantillon de validation
Mean Square Error (MSE)	18.182	18.252
R^2	84.6%	87.5%

Le pouvoir explicatif du modèle dépasse les 80% mais cela demeure perfectible. Nous remarquons que l'échantillon d'entrainement permet d'améliorer la prédiction, en apprenant sur ces données, puisque l'échantillon de validation possède un R^2 de 87,5%, contre 84,6% sur l'échantillon d'apprentissage. Cependant, la moyenne des erreurs au carré de l'échantillon test est légèrement supérieure à celle de l'échantillon d'apprentissage (comme nous l'illustre également l'annexe 13). Mais cet écart est négligeable, dans une certaine mesure les résultats obtenus avec le Lasso sont satisfaisants et l'écart des MSE est relativement faible.

L'annexe 12 confirme les résultats du tableau précédent car ce graphique traduit le comportement du modèle par rapport à la tendance des données d'entrainement. De manière générale, le modèle prédit bien les valeurs, par contre à partir d'un certain seuil on observe que la performance du modèle se réduit. Cependant, confronter les résultats du Lasso à une autre méthode apparait comme judicieux afin de choisir le meilleur modèle.

- **Random Forest**

Pour prévoir les cours de Apple nous avons utilisé une méthode de Machine Learning supervisé : Random Forest ou foret aléatoire, Il est basé sur l'idée d'ensemble d'arbres de décision, ce qui signifie qu'il construit plusieurs arbres de décision indépendants et utilise leur moyenne pour effectuer une prédiction. Son application nous donne les résultats suivants :

Eléments	Echantillon d'apprentissage	Echantillon de validation
Mean Square Error (MSE)	0,242	0,29
R^2	99.1%	97,8%

A la lumière de ce dernier, on observe que d'une manière générale, le pouvoir explicatif du modèle (caractérisé ici par son $R^{2)}$ dépasse les 90%. Les prévisions faites par le modèle sont donc très honorables. Toutefois, la moyenne des erreurs au carré pour l'échantillon test est légèrement supérieur à celle de l'échantillon d'apprentissage (ces erreurs sont tout de même très bas) et le pouvoir explicatif du modèle en utilisant l'échantillon de validation est légèrement plus faible que pour l'échantillon d'apprentissage.

Pour rendre cela beaucoup plus clair, nous choisissons de représenter les résultats précédents par l'annexe 14. Ce graphique permet d'évaluer la performance du Random Forest et détermine si le modèle est sur ou sous-ajusté. La visualisation de la performance du modèle sur les données d'entrainement et de test a permis de voir comment le modèle se comporte par rapport à la tendance des données d'entrainement. Dans notre cas, le modèle est bien entrainé car les prévisions du modèle suivent de près les données de validation. Ce modèle est donc capable de généraliser les tendances de la formation à de nouvelles données.

Les variables choisies prédisent correctement le cours de Apple selon cette méthode, car le pouvoir explicatif des deux modèles de régression (R^2) reste élevé dans les deux cas.

V. CONCLUSION DE L'ETUDE

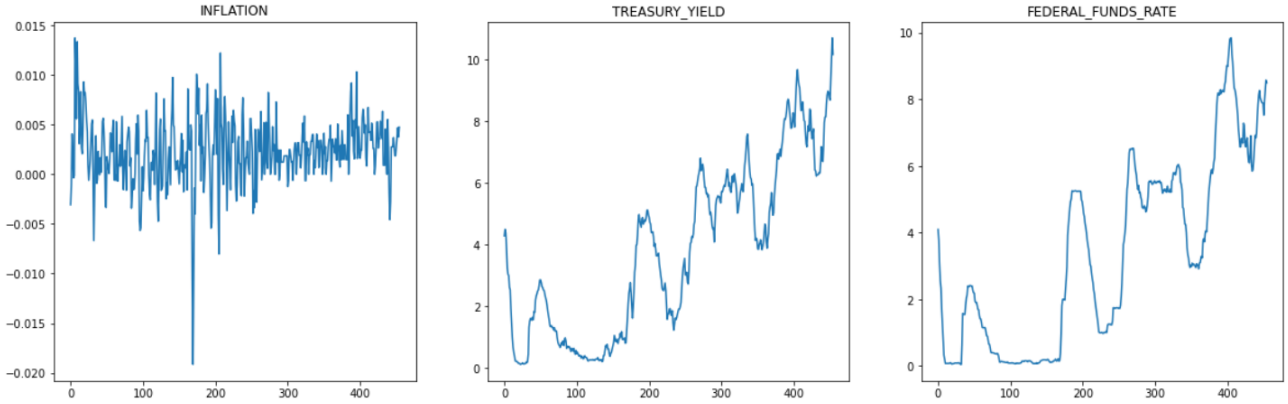
En dernière analyse, cette étude nous a permis de mettre en application les deux méthodes du Machine Learning : la classification et la régression. Nous avons dans un premier temps utilisé la méthode de classification pour prédire au mieux le cours boursier d'Apple en fonction de l'inflation américaine, le treasury yield, le federal funds rate et le S&P 500. Les résultats des modèles réalisés n'ont pas été très satisfaisants ; raison pour laquelle nous avons appliqué la régression. Les deux modèles de régression ont donné des résultats plus concluants grâce au pouvoir explicatif particulièrement élevé dans les deux cas (R^2) et malgré une marge d'erreur relativement plus importante sur l'échantillon test en comparaison avec l'échantillon train.

La prédiction des cours boursier est particulièrement intéressante à étudier car, elle lève le voile sur les problématiques liées au choix des données afin d'éviter les problèmes d'overfitting et underfitting, qui sont les deux problèmes les plus contraignants du Machine Learning. L'overfitting ou surajustement se produit lorsqu'un modèle ajuste trop bien les données d'entrainement, il est donc dans l'incapacité de prédire avec précision des données de test. A l'inverse, l'underfitting ou sous-ajustement se produit lorsque le modèle ne peut pas déterminer une relation significative entre les données d'entrée et de sortie (présence de biais élevé). Dans notre cas, les problèmes d'ajustement pourraient justement expliquer la faiblesse de nos prédictions. En effet, pour les modèles de classification, le SVM a prédit la totalité des rendements positifs et aucun rendement négatif, et le Random Forest a, à l'inverse, prédit quasiment à la perfection les rendements négatifs (96%) et quasiment aucun rendements positifs (1,5%). En ce qui concerne les modèles de régression, que ce soit pour le cas du Lasso ou du Random Forest, les erreurs sont plus grandes dans l'échantillon test que l'échantillon train mais avec un pouvoir explicatif très élevé.

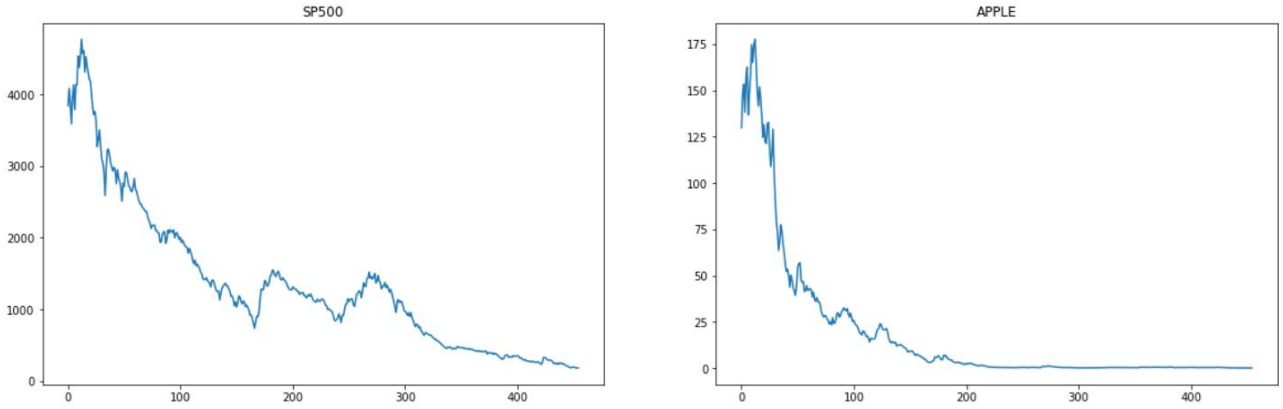
Pour remédier à ces problèmes d'ajustement, et pour améliorer la prévision du cours boursier, il aurait fallu avoir soit une plus grande base de données afin que l'algorithme puisse avoir une plus grande diversité de données et augmenter son temps d'entrainement, ou soit choisir d'autres variables explicatives.

VI. ANNEXES

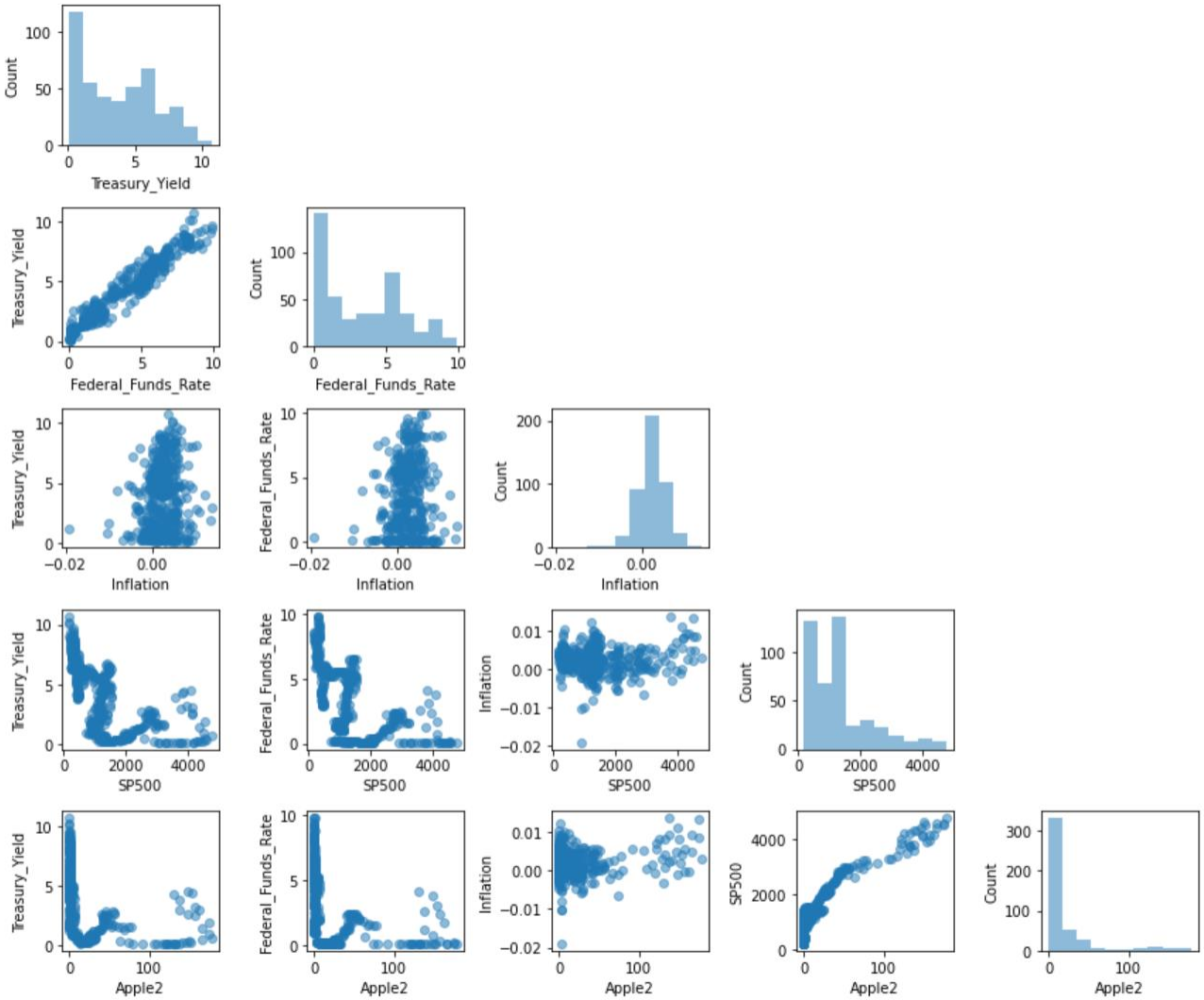
Annexe 1 : Graphiques Inflation, Treasury Yield, Federal Funds Rate



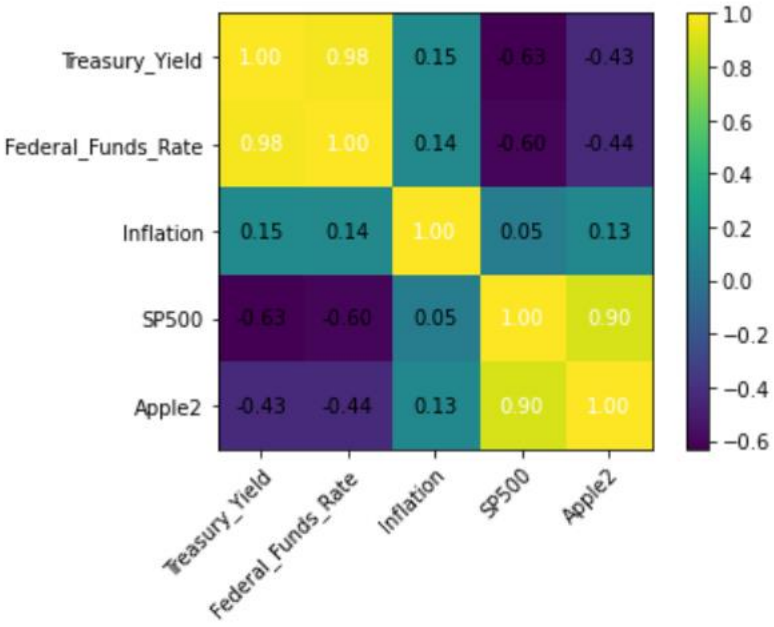
Annexe 2 : Graphiques S&P 500 et Apple



Annexe 3 : Relations entre les variables



Annexe 4 : Matrice de corrélation



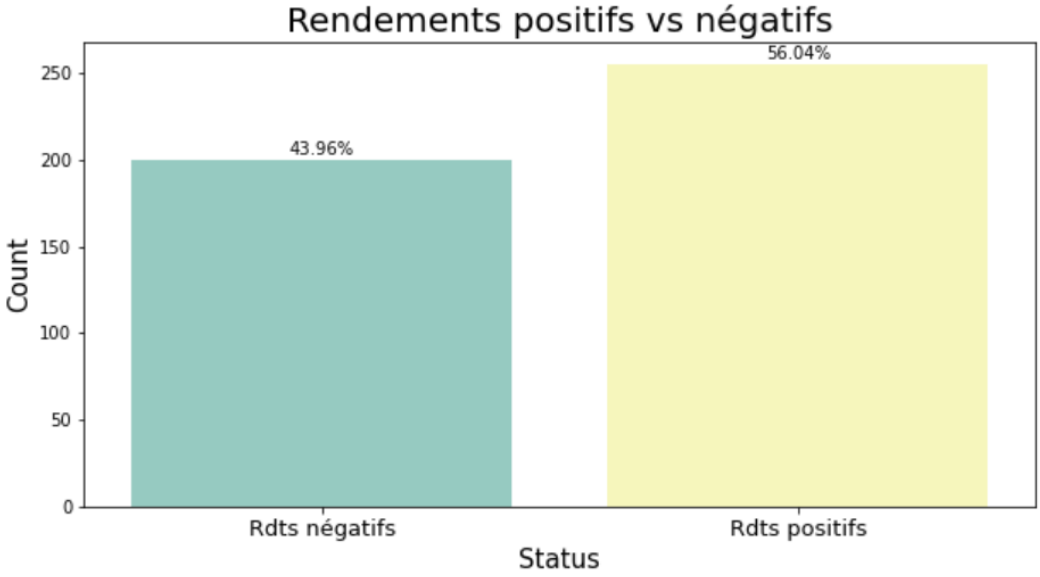
Annexe 5 : Dataframe – Régression

	Time	Treasury_Yield	Federal_Funds_Rate	Inflation	SP500	Apple2
0	2022-12-01	4.29	4.10	-0.003070	3839.500000	129.929993
1	2022-11-01	4.50	3.78	-0.001010	4080.110107	148.029999
2	2022-10-01	4.38	3.08	0.004056	3871.979980	153.339996
3	2022-09-01	3.86	2.56	0.002151	3585.620117	138.199997
4	2022-08-01	3.25	2.33	-0.000354	3955.000000	157.220001
...	...	...	...	...	...	...
450	1985-06-01	8.69	7.53	0.002796	191.850006	0.080357
451	1985-05-01	9.39	7.97	0.003742	189.550003	0.077567
452	1985-04-01	10.09	8.27	0.004699	179.830002	0.094866
453	1985-03-01	10.71	8.58	0.003774	180.660004	0.098772
454	1985-02-01	10.17	8.50	0.004739	181.179993	0.110491

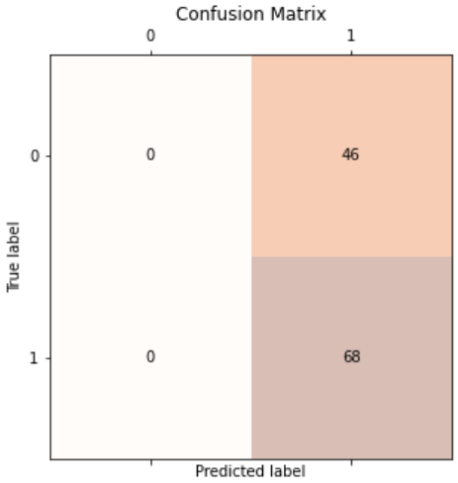
Annexe 6 : Dataframe – Classification

	Time	Treasury_Yield	Federal_Funds_Rate	Inflation	SP500	Apple
0	2022-12-01	4.29	4.10	-0.003070	3839.500000	0.0
1	2022-11-01	4.50	3.78	-0.001010	4080.110107	0.0
2	2022-10-01	4.38	3.08	0.004056	3871.979980	1.0
3	2022-09-01	3.86	2.56	0.002151	3585.620117	0.0
4	2022-08-01	3.25	2.33	-0.000354	3955.000000	0.0
...	...	...	...	...	...	...
450	1985-06-01	8.69	7.53	0.002796	191.850006	1.0
451	1985-05-01	9.39	7.97	0.003742	189.550003	0.0
452	1985-04-01	10.09	8.27	0.004699	179.830002	0.0
453	1985-03-01	10.71	8.58	0.003774	180.660004	0.0
454	1985-02-01	10.17	8.50	0.004739	181.179993	0.0

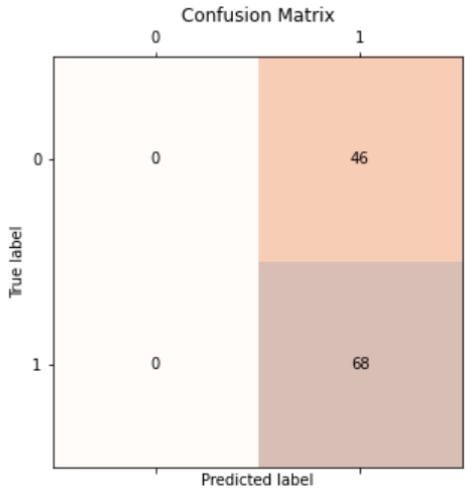
Annexe 7 : Vérification dataframe équilibré/déséquilibré



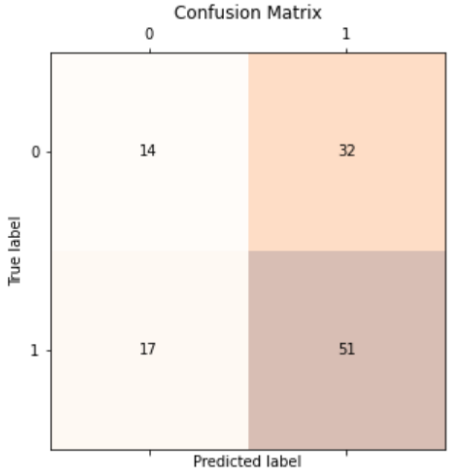
Annexe 8 : Matrice de confusion – KNN



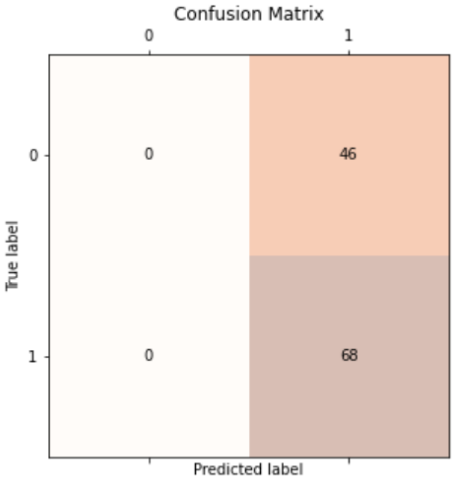
Annexe 9 : Matrice de confusion – Random Forest



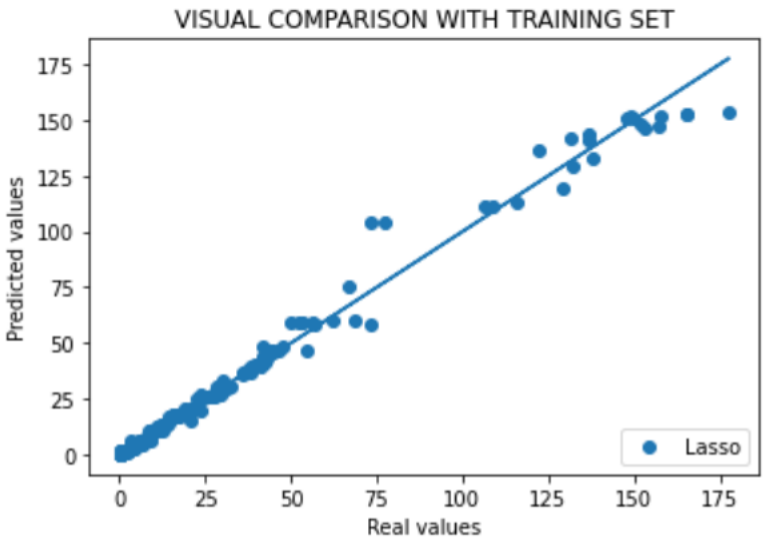
Annexe 10 : Matrice de confusion – Naive Bayes



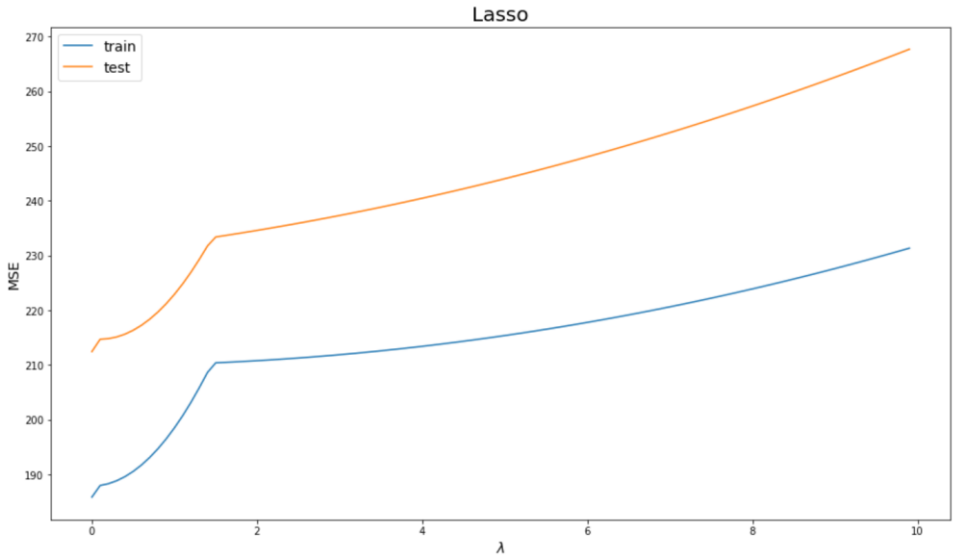
Annexe 11 : Matrice de confusion – SVM



Annexe 12 : Graphique – Lasso



Annexe 13 : Graphique – Lasso MSE



Annexe 14 : Graphique – Random Forest