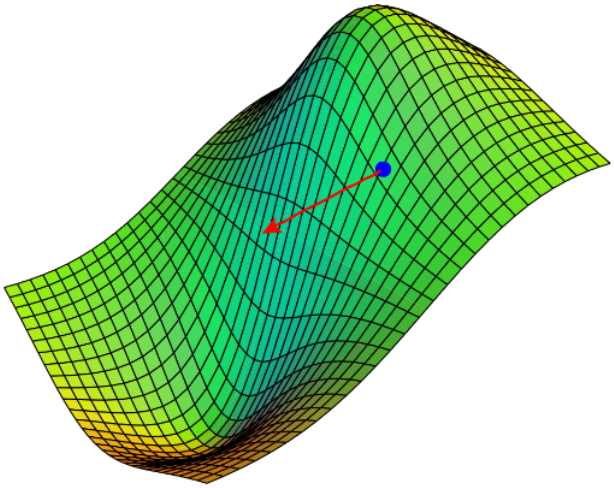
UNIVERSITÉ DE NANTES

IAE NANTES
ÉCONOMIE & MANAGEMENT

Quantitative Research on Market Predictability

- Synthèse -



Aurouet Lucas

Master EKAP
IAE Nantes
Under the supervision of Prof. O.DARNE
27/06/2021

1 Problématique de l'étude

Ce travail s'inscrit dans la continuité du Master II EKAP, économétrie appliquée. L'objectif est d'utiliser l'analyse de données pour confronter théorie et pratique, observer dans quelle mesure les deux coïncident ou s'opposent, faire sens des phénomènes observés pour mieux comprendre les tenants et aboutissants de l'environnement étudié. Ici, nous nous concentrons sur les marchés financiers, nous tentons de déterminer s'il est possible de prédire le cours d'une action à l'aide d'indicateurs dits "techniques". Selon la théorie des marchés efficients, stipulant que toute l'information disponible est déjà reflétée sur le marché, il est impossible d'anticiper les futurs mouvements de prix. Selon les praticiens, qui ont régulierement recours à ces indicateurs techniques, il existe de l'information résiduelle qui permet de déterminer à l'avance comment le prix d'une action est amené à évoluer. La problématique se résume donc à une opposition directe entre théorie et pratique: est-il possible d'utiliser les indicateurs techniques pour prédire le cours des actions ? Le cas échéant, les marchés seront considérés comme non efficients (tout du moins imparfaitement efficients), remettant en cause l'hypothèse de E.Fama[1].

2 Cadre de l'étude

Les indicateurs techniques sont des mesures calculées à partir de certaines caractéristiques des cours (prix, volume, volatilité). Ces mesures sont censées d'après les praticiens, identifier les tendances, points de retournements et autres mouvements de prix et permettent donc de connaître à l'avance le comportement du cours. C'est (entre autres) grâce à ce type de mesures que les investisseurs professionnels prennent la décision d'acheter ou vendre tel ou tel actif.

Cette pratique s'oppose fortement à la théorie de E.Fama selon laquelle les marchés financiers sont capables de réagir instantanément à n'importe quel choc. Le prix de chaque actif représente donc sa valeur prenant en compte l'ensemble des phénomènes qui ont pu influer par le passé. Il n'existe, selon l'hypothèse des marchés efficients, aucun actif sur ou sous-évalué, puisque chaque prix reflète parfaitement le sentiment des investisseurs vis-à-vis de l'état du marché. Il est donc vain de tenter de déterminer le futur cours de l'actif à l'aide de méthodes censées réaliser des prédictions à l'aide d'information résiduelle qui n'aurait pas été prise en compte par les autres acteurs du marché.

Ici nous avons choisis 6 indicateurs techniques pour nous aider à déterminer le cours des actions, ces indicateurs ont été choisis sur la base d'un raisonnement très simple; ils sont les indicateurs les plus utilisés en pratique. Parmi tous ceux disponibles, ils constituent donc ceux qui ont le plus de chance de contenir une information pertinente. Leur utilisation intensive par l'ensemble des praticiens est susceptible d'induire une corrélation entre les indicateurs et le cours de l'action, même si, intrinsèquement, ils ne contiennent aucune information. Le cas échéant, la corrélation est induite par un nombre suffisant d'acteurs qui basent leurs décisions sur des facteurs communs, entraînant de facto des mouvements de prix tels que prédisent les indicateurs, c'est une prophétie auto-réalisatrice.

[1] Efficient Capital Markets: A Review of Theory and Empirical Work - *E.Fama* - The Journal of Finance, Vol. 25, No. 2, *pp.383-417* - 1970

Les modèles sont entraînés sur 4 actions différentes, issues de marchés différents (Apple - technologie, Air France-KLM - transport aérien, Walmart - grande distribution, TESLA - automobile), nous espérons ainsi constater une différence de la performance de chaque modèle en fonction du marché et du comportement de chaque actif. Nous utilisons l'intégralité du jeu de données pour la régression logistique (08/02/2010 - 31/12/2019), les résultats des modèles de machine learning présentés ici sont obtenus depuis le jeu test (02/04/2017 - 31/12/2019).

3 Méthodologie

Les concepts méthodologiques utilisés pour ces travaux se déclinent en trois points: transformation des variables, modèles, métriques de performance.

- **Transformation des variables:** essentielles à l'utilisation de modèles d'analyse de données, les variables requièrent une attention toute particulière. La sélection des variables pertinentes et leur préparation constituent une étape sans laquelle les résultats finaux n'auraient pas d'utilité. Les indicateurs techniques sont d'abord utilisés dans les modèles de manière brute, c'est à dire en considérant la valeur de l'indicateur comme variable explicative. Puis nous avons ultérieurement transformé certaines de ces variables quantitatives en variables qualitatives de sorte à imiter les praticiens qui considèrent des valeurs limites au dessus/dessous desquelles l'indicateur indique un certain mouvement de prix à venir. La variable à expliquer doit également être transformée; les indicateurs sont utilisés en tant que signaux d'achat ou de vente, il faut donc que la variable d'intérêt reflète cette logique. On transforme donc le cours de l'action en une variable elle aussi qualitative, qui prend une valeur de 0 si le rendement est négatif, 1 s'il est positif[2].

- **Modèles:** trois modèles sont utilisés ici; la régression logistique, pour tester la significativité de la relation de manière formelle. La Machine à Supports Vectoriels (SVM pour Support Vector Machine) ainsi que le réseau de neurones pour la prédiction du rendement, positif ou négatif. L'intérêt de cette démarche est d'exploiter les avantages des différentes méthodes pour des problématiques précises. La régression logistique permet d'effectuer des tests statistiques, donc de vérifier la significativité des variables, de mesurer la quantité d'information obtenue. Mais la régression logistique est un modèle linéaire, qui ne serait pas en mesure de correctement modéliser les données si elles venaient à être non-linéaire, et qui conclurait à des prédictions peu convaincantes. A l'inverse les modèles de machine learning (SVM et réseau de neurones) sont des modèles "boites noires" qui ne permettent pas de rendre compte de la force de la relation entre nos différentes variables, ni même de la nature de cette relation. Ils permettent en revanche de déceler des relations fortement non-linéaires, et donnent des prédictions généralement meilleures que les modèles linéaires.

- **Métriques de performance:** résumer et chiffrer la performance de chaque modèle est essentiel, d'une part pour les comparer entre eux, et vis-à-vis d'un modèle "nul". Et d'autre part pour mesurer leur applicabilité dans la vie réelle. La comparaison

[2]Les jours où le rendement est strictement égal à 0 représentent une infime partie du jeu de données, recodée en 1.

des modèles entre eux permet notamment de choisir le meilleur modèle, celui qui offre les prédictions les plus justes, les plus précises, mais aussi de mettre en valeur des phénomènes qui permettent de mieux comprendre la structure des données. Si un modèle est meilleur qu'un autre, la différence nous indique quelle modélisation est la plus proche du processus de génération de données. Le meilleur modèle est donc celui qui explique au mieux le comportement du cours des actions. Nous utilisons ici le test de ratio de vraisemblance, qui permet de mesurer l'intérêt global d'un modèle, le test de Wald qui vérifie la significativité des variables ainsi que le pseudo R^2 qui mesure la quantité d'information obtenue. Ces tests seront utilisés pour la régression logistique, ils ne sont pas calculables avec des modèles de machine learning[3], d'où l'intérêt d'utiliser une régression de ce type. Nous utilisons également la precision, le recall, le F1 ainsi que la courbe ROC (Receiver Operator Characteristic), qui sont trois méthodes différentes basées sur le principe de ratio entre observations correctement identifiées par le modèle et observations incorrectement classifiées. Ces mesures seront utilisées pour les modèles de machine learning. Pour mesurer la performance économique des modèles nous avons choisis les ratios de Sharpe et de Sortino, deux mesures de rendements ajustées du risque, communément utilisées en pratique.

4 Résultats

Les résultats sont encourageants, après une première transformation des variables brutes en signaux, nous trouvons des relations significatives pour l'immense majorité des variables. Ces relations concordent d'autant plus avec la pratique, les signaux envoyés étant bien des signaux de vente et d'achat quand ils doivent l'être, et non pas l'inverse. Dans le cas contraire, la relation aurait été toute aussi significative mais la contradiction entre les signaux observés et les signaux tels qu'ils sont employés dans la pratique nous aurait poussé à croire que la relation n'est pas exploitable. La régression contient suffisamment d'information pour être meilleure (plus précise) qu'un modèle "nul" (absence de modèle) comme le montre le test de ratio de vraisemblance significatif au seuil de 99%. Le pseudo R^2 nous indique que la régression explique environ 10% de la variance des rendements[4], qui reste une valeur relativement haute comparée au reste de la littérature.

[3] Principalement à cause des méthodes d'estimation qui différent

[4] Transformés en variable binaire 1 si positif, 0 sinon.

Table 1: Régression Logistique

Actif	**AF.PA**	**TSLA**	**AAPL**	**WMT**
Variable	Coef.			
ADX	-	+*	+**	+
OBV	+**	+	+*	+
RSI_0	+***	+***	+***	+***
RSI_1	-***	-***	-***	-***
RSI_2	+***	+***	+***	+***
RSI_3	-***	-***	-***	-***
D_0	+***	+	+	+
D_1	-**	-.	-*	-***
D_2	+***	+***	+*	+.
D_3	-*	-*	-	-**
boll_0	-*	-***	-***	-***
boll_1	+**	+	+*	+
MACD_2	+	+	+**	+
Likelihood Ratio p-value	$3.1e^{-44}$	$3.8e^{-42}$	$1.4e^{-40}$	$2.3e^{-50}$
$Pseudo - R^2$	0.103	0.101	0.097	0.12

*Note: significativité: . = 0.1, * = 0.05, ** = 0.01, *** = 0.001*

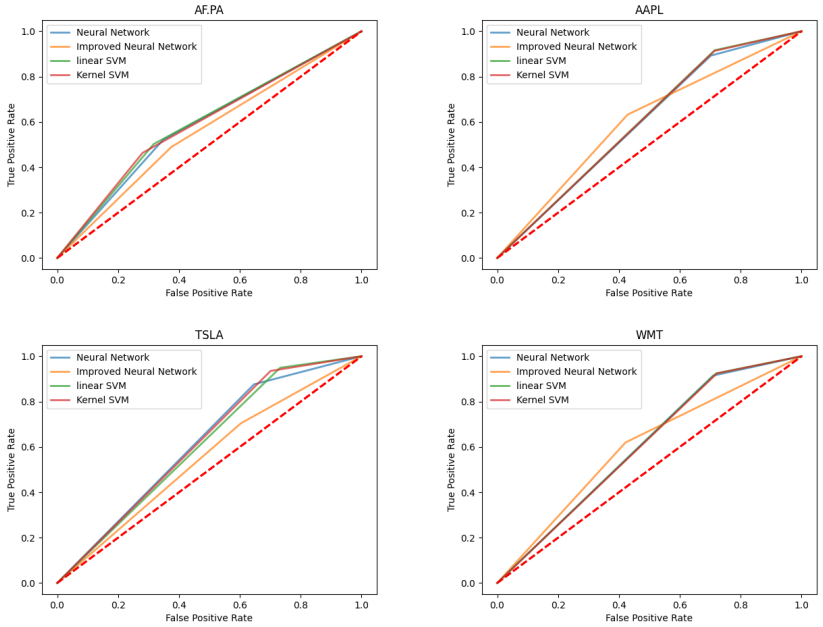
Table 2: Score F1 sur jeu test

	TSLA	AF.PA	AAPL	WMT
Réseau Standard	0.65	0.58	0.64	0.69
Réseau Non-linéaire	0.56	0.55	0.60	0.60
SVM Standard	0.66	0.59	0.67	0.69
SVM Kernel	0.66	0.59	0.66	0.69

Les méthodes de machine learning concluent également à des résultats satisfaisants, la précision globale des modèles sur tous les actifs est de l'ordre de 55% à 69% et les modèles semblent être stables puisqu'ils donnent d'aussi bons résultats sur des données sur lesquelles ils n'ont pas étés entraînés[5]. La ROC reflète cette précision avec une courbe qui tend vers le cadran supérieur gauche. Le réseau de neurones non-linéaire est le modèle qui subit la plus forte perte de précision entre les jeux de données d'entraînement et de test, ce qui nous porte à croire que le modèle est sur-ajusté, et qu'un modèle faiblement non-linéaire suffit à modéliser le cours des actions. Les trois autres modèles sont comparables en termes de performance comme le montre le courbe ROC. Les modèles sont mieux ajustés sur des séries moins volatiles, la volatilité influe sur le ratio precision/recall et il semble y avoir une corrélation négative entre la variance du cours et la performance finale du modèle. Ce phénomène est illustré par la sous-performance de tous les modèles sur l'action Air France-KLM comparativement aux autres actifs, ainsi qu'à la sur-performance des modèles sur l'action Walmart. D'une manière générale, nos modèles sont capables de déterminer avec une précision toute relative, mais significative, les futurs mouvements de prix.

[5]Une légère perte de précision est à déplorer mais cela ne constitue pas un problème bien au contraire cela montre que les modèles ne sont comportent pas de manière anormale.

Figure 1: ROCs sur le jeu test



Cette précision se traduit par des résultats économiques significatifs, le meilleur modèle (réseau de neurones standard) offre des ratios de Sharpe et de Sortino supérieurs à ceux d'une stratégie buy & hold sur l'indice de marché S&P 500. On obtient des ratios de Sharpe compris entre 1.87 et 2.34 et des ratios de Sortino compris entre 1.88 et 2.58 selon les actifs contre 0.54 et 0.57 respectivement, pour la stratégie benchmark. Les modèles permettent également d'obtenir, sur chaque actif, de meilleurs ratios qu'avec une stratégie sans modèle. Cette observation nous permet de savoir avec certitude que l'excès de rendement ajusté est dû à la précision des prédictions, non pas au fait que l'actif ai initialement de meilleurs ratios que le S&P 500, comme c'est le cas pour Tesla, Apple et Walmart.

Table 3: Résultats Economiques

Modèle Nul

	TSLA	AF.PA	AAPL	WMT	S&P 500
Sharpe	0.55	0.36	0.98	0.86	-
Sortino	0.62	0.40	1.07	0.94	-

Réseau de Neurones

	TSLA	AF.PA	AAPL	WMT	S&P 500
Sharpe	2.34	2.26	1.87	2.06	0.54
Sortino	2.53	2.58	1.88	2.46	0.57

5 Conclusion

Nous avons mis en évidence l'existence d'un pouvoir prédictif significatif des indicateurs techniques choisis. la régression logistique semble indiquer que les variables sont majoritairement significatives et explique une part non négligeable des rendements. Les modèles de machine learning parviennent à établir des prédictions suffisamment précises pour confirmer l'utilité les indicateurs techniques. Ces prédictions permettent, sur un nouveau jeu de données, de donner lieu à des mesures de performance ajustées du risque préférables à celle d'une stratégie sans modèle. Il semblerait qu'il soit possible de prédire, dans une certaine mesure, le cours des actions à l'aide d'indicateurs techniques. Les marchés ne sont pas parfaitement efficients. Néanmoins, pour réfuter de manière crédible l'hypothèse des marchés efficients, il faudrait étudier la stabilité des modèles dans le temps et l'espace, en allongeant la période de test, et en effectuant des mesures sur un plus grand nombre d'actifs. Il serait également intéressant d'approfondir la relation existante entre la volatilité et la précision du modèle.