Financial Stock Market Forecast using Data Mining Techniques in Morocco

22 November 2022

Soukaina CHRIT
SOUKAINA.CHRIT@etu.uae.ac.ma
Fatima EL JAIMI
FATIMA.ELJAIMI@etu.uae.ac.ma

Description

Avec l'émergence des technologies de l'information telles que le World Wide Web et les réseaux sociaux, une grande quantité de données devient accessible à tous. Récemment, Data Mining a reçu une attention croissante en tant que moyen d'analyser et de traiter une grande quantité de données financières et l'une des tâches difficiles pour les chercheurs, les commerçants et les investisseurs est de prévoir les mouvements des cours des actions. La prédiction des prix précis des actions présente une tâche difficile pour tous les praticiens du marché boursier comme les commerçants et les investisseurs.

Nous allons construire un modèle pour prédire les mouvements des cours des actions en utilisant des techniques de Data Mining avec Artificial Intelligence Classifier.

Les résultats de ce projet aideront les professionnels de la finance et investisseurs à avoir une bonne connaissance de l'évolution du cours de l'action marocaine. Cela renforcera davantage leur confiance à faire plus d'affaires avec moins de risques.

Objectives

Les objectifs de ce projet sont les suivants :

- Étudier les données existantes de l'échange marocain pour analyser automatiquement les données financières pour prédire les tendances futures des prix.
- Étudier les méthodes de Data Mining qui pourraient être utiliser dans ce cas, évaluez les résultats et choisir la méthode la plus efficace.
- Tester la méthode et la déployer.

Description des données

Pour ce projet on a choisi les données historiques journalier des actions de la bourse de Casablanca, avec une somme de 75 sociétés cotées, de tous secteurs et toutes les industries, entre la période de 2010 et 2022.

Notre projet est un problème de time séries, ce qui explique le choix des données historiques. On a aussi opté pour des enregistrements journaliers pour plus de précision et une meilleure qualité de données cotée complétude.

Les données sont collectées d'une *manière manuelle*, en téléchargeant les données de chaque société depuis le site https://fr.investing.com/, qui est une plateforme financière et un site d'actualités, *l'un des trois principaux sites Web financiers mondiaux au monde*. Il propose des cotations de marché, *des informations sur les actions*, etc.

La taille de notre jeu de données est de 140 244 enregistrements avec 7 champs, contenant une majorité de données quantitatives, un champ Date et un champ de données qualitatives.

La date est présentée sous la forme DD/MM/YYYY. Les données numériques sont formatées en 2 chiffres décimale. Les symboles **K** et **M** sont utilisées pour représenter respectivement les milliers et les millions. La devise utilisée est le dirham marocain MAD. Et les sociétés sont présenter par l'abréviation de leurs noms.

Extrait des données :

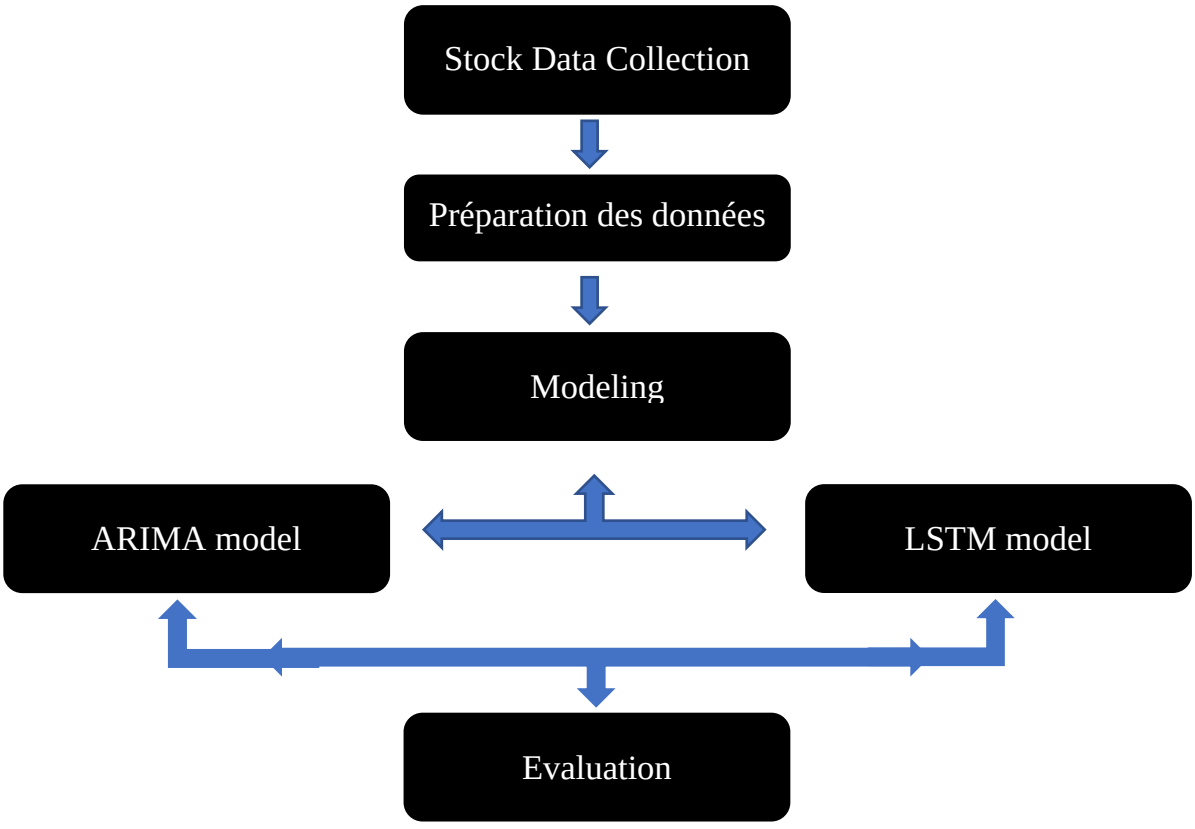
Date	Dernier	Ouv.	Plus Haut	Plus Bas	Vol.	Variation %	Societe
10/11/2022	6,59	6,55	6,63	6,55	5,57K	0,0154	ADH
09/11/2022	6,49	6,79	6,79	6,49	70,26K	-0,0241	ADH
08/11/2022	6,65	6,88	6,88	6,65	116,94K	-0,0148	ADH
07/11/2022	6,75	6,96	6,96	6,75	56,19K	-0,0146	ADH
04/11/2022	6,85	6,86	6,96	6,85	15,72K	-0,0044	ADH
03/11/2022	6,88	6,98	6,98	6,86	31,01K	-0,0029	ADH

Description des articles en relation

Erdinç Altay et M. Hakan Satman de l'Université d'Istanbul - Faculté d'économie, dans leur étude, ont comparé les performances de prévision des stratégies ANN et de régression linéaire à la Bourse d'Istanbul et ils ont obtenu des preuves de la performance statistique et financière des modèles ANN. Selon eux, bien que les statistiques de précision des prévisions hors échantillon basées sur des données quotidiennes et mensuelles des modèles ANN (RMSE, MAE et Theil's U) n'ait pas surpassé les modèles de régression alternatifs, ils ont obtenu des preuves significatives d'une meilleure prédiction de la direction du marché de ANN. Pour les données quotidiennes, hebdomadaires et mensuelles, respectivement, les modèles ANN peuvent prédire avec précision la direction des indices boursiers jusqu'à 57,8 %, 67,1 % et 78,3 %.

Dans leur article intitulé "Analyse systématique et examen des techniques de prédiction du marché boursier", Dattatray P. Gandhmal et K. Kumar présentent une analyse approfondie de 50 articles de recherche qui suggèrent des méthodologies telles que le Bayesian model, Fuzzy classifier, Artificial Neural Networks (ANN), Support Vector Machine (SVM) classifier et d'autres. L'ANN et le Fuzzy classifier sont les méthodes les plus souvent utilisées pour obtenir des prévisions boursières précises.

Diagramme



Algorithmes utilisés

ARIMA model

ARIMA (abréviation de "Autoregressive Integrated Moving Average") est un modèle statistique qui peut être utilisé pour prévoir des données de séries chronologiques (time series data), telles que les cours des actions. Il s'agit d'une généralisation du modèle ARMA (moyenne mobile autorégressive) plus simple, et est capable de tenir compte des tendances dans les données en utilisant un processus appelé différenciation.

Le modèle ARIMA est spécifié par trois paramètres : p, d et q.

Le paramètre d représente l'ordre de différenciation, qui est une mesure du nombre de fois que les données ont été différenciées afin de les rendre stationnaires (c'est-à-dire pour supprimer les tendances).

Le paramètre p représente l'ordre de la composante autorégressive (AR) du modèle, qui est une mesure du nombre de pas de temps passés utilisés pour prédire le pas de temps actuel.

$$y_t = a_0 + \sum_{n=1}^{p} a_n y_{t-n} + \epsilon_t$$

Le paramètre q représente l'ordre de la composante moyenne mobile (MA) du modèle, qui est une mesure du nombre d'erreurs de prévision passées utilisées pour prédire le pas de temps actuel.

$$r_t = b_0 + \sum_{n=1}^{q} b_n r_{t-n} + \epsilon_t$$

Pour ajuster un modèle ARIMA à une série chronologique, les étapes suivantes sont généralement suivies :

Vérifiez la stationnarité des données à l'aide d'un test statistique ou en inspectant visuellement les données pour les tendances. Si les données ne sont pas stationnaires, appliquez une différenciation pour les rendre stationnaires.

Déterminez les valeurs optimales des paramètres p, d et q.

Ajustez le modèle ARIMA final aux données en utilisant les valeurs p, d et q sélectionnées.

Utilisez le modèle ajusté pour faire des prévisions de la série chronologique dans le futur.

Résultats

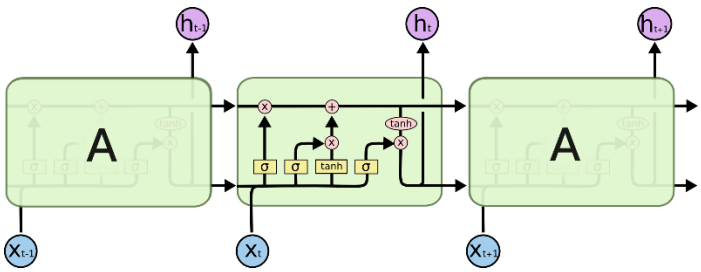
	AIC	BIC	MAE	MSE
ATW	13895.2	13912.5	2.9	46.2
BCP	11300.1	11317.4	1.6	12.7
BOA	11689.9	11730.2	1.9	35.5
CDM	8132.5	8151.8	15.6	909.9

LSTM :

Suite aux très mauvais résultats obtenus avec le modèle ARIMA, nous avons décidé d'utiliser un autre réseau de neurones appelé LSTM.

Les réseaux à mémoire longue et courte durée – généralement simplement appelés « LSTM » sont un type particulier de RNN, capable d'apprendre des dépendances à long terme. Ils ont été introduits par Hochreiter & Schmidhuber (1997), et ont été raffinés et popularisés par de nombreuses personnes dans les travaux suivants. Ils fonctionnent extrêmement bien sur une grande variété de problèmes, et sont maintenant largement utilisés.

Les LSTM sont explicitement conçus pour éviter le problème de dépendance à long terme. Se souvenir d'informations pendant de longues périodes est pratiquement leur comportement par défaut, pas quelque chose qu'ils ont du mal à apprendre !



- **Modeling**

Après avoir préparé et remodelé les données à utiliser pour le modèle LSTM, nous avons essayé plusieurs architectures et hyperparamètres pour obtenir le meilleur modèle adapté à notre situation.

Le modèle que nous avons construit à la fin contient trois couches.

La couche d'entrée est la couche LSTM (6 features et 20 timesteps), vous pouvez trouver l'architecture du modèle ci-dessous :

```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_2 (LSTM)               (None, 256)               269312

 dense_4 (Dense)             (None, 32)                8224

 dense_5 (Dense)             (None, 1)                 33

=================================================================
Total params: 277,569
Trainable params: 277,569
Non-trainable params: 0
```

Résultats

Attijari Wafa Bank - Results

Banque Centrale Populaire - Results

Bank of Africa - Results

Crédit du Maroc - Results

	MSE	RMSE
ATW	0.00067	0.03
BCP	**0.00021**	0.01
BOA	0.00113	0.03
CDM	0.00103	0.03

Les références :

- https://fr.investing.com/
- https://towardsdatascience.com/time-series-models-d9266f8ac7b0
- https://medium.com/@humble_bee/rnn-recurrent-neural-networks-lstm-842ba7205bbf
- https://link.springer.com/article/10.1007/s00521-020-04867-x
- Time Series Forcasting in Python by Marco Peixeiro
- Statsmodels.org official documentation