DATA MINING APPROACH FOR

KORICHI Abdel-Rahmen
DORLIAT Hector
2017-2018

Table des Matières

Introduction

Le bitcoin, contraction des mots anglais bit (unité de mesure binaire) et coin (pièce de monnaie) est une monnaie cryptographique adossée à un système de paiement de pair-à-pair qui n'existe que sous forme numérique.

Cette monnaie a été inventée en 2008 et son logiciel open source publié en 2009. Son créateur, qui se cache derrière le pseudonyme de Satoshi Nakamoto, n'a toujours pas été identifié.

Le bitcoin permet d'acheter des biens et des services, il peut aussi être échangé contre d'autres devises. De nombreux sites Internet mais aussi des magasins en dur acceptent cette monnaie dont l'un des principaux avantages réside dans le faible coût des frais de transaction.

Contrairement aux monnaies traditionnelles, le bitcoin n'est pas administré par une autorité bancaire unique, il fonctionne de manière décentralisée à travers un ensemble de noeuds. Ces derniers forment le réseau par lequel se font toutes les transactions. Un registre public sécurisé appelé blockchain ou " chaîne de blocs " tient l'historique de toutes ces opérations.

Des utilisateurs volontaires mettent à disposition leur temps et la puissance de calcul de leurs ordinateurs pour administrer la blockchain. Cette opération appelée le " minage " permet à ces personnes d'être rémunérées, en bitcoin bien sûr. La valeur du bitcoin est maintenue par les logiciels de minage qui adaptent l'intensité des calculs au nombre de mineurs actifs. Plus il y a de mineurs de bitcoin, plus les calculs sont complexes.

Plusieurs plateformes proposent la conversion de dollars, euros ou yuans en Bitcoins. Le taux de change de cette monnaie a connu de grandes variations qui sont suivies quotidiennement par des indices comme celui que propose les palateformes CoinDesk, CoinBase, ou encore Bitfinex. En 2011 par exemple, on obtenait un Bitcoin pour 4,15 euros. Deux ans plus tard, en décembre 2013, il fallait débourser 860 euros. En décembre 2017, la valeur atteignait le chiffre vertigineux de 16.000 euros.

La valeur du Bitcoin est considérée par certains comme trop incertaine pour être prédictible. Plusieurs personnes ont cherché à développer des mythologies et des modèles pour améliorer la probabilité de faire du profit dans leurs investissements. Le taux de réussite global de ces méthodologies et de ces modèles sont généralement trop bas pour qu'ils soient utilisés pour des investissements réels. L'une des raisons principales est l'énorme fluctuation du marché.

I Formalisation

1 État des lieux

Prédire et comprendre le marché des cryptomonnaies est un problème challengeant, et les travaux faits en fouille de données dans ce domaine sont encore assez rares. En revanche, on peut considérer que les marchés boursiers et les marchés des cryptomonnaies sont analogues sur de nombreux points. Et, dans ce cas, la documentation est plus vaste. Néanmoins, il semble que le marché des cryptomonnaies est beaucoup plus fluctuant. Et d'après nos recherches, il est important de prendre en compte la couverture médiatique pour bien comprendre la fluctuation du cours d'une cryptomonnaie.

2 Objectifs

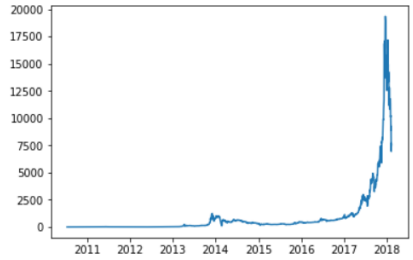
Ce papier a pour objectif de montrer comment des techniques de Data Mining peuvent être utilisées dans l'analyse du cours de la cryptomonnaie la plus célèbre à ce jour, le Bitcoin. Le but de notre démarche a été de rechercher un ou des schémas caractéristiques à l'évolution du prix du Bitcoin. Nous avons opté pour une posture prédictive. Ce qui nous intéressait n'était donc pas de trouver de l'information sur des éléments arrivant pendant l'évolution du prix, mais la précédant, ou, éventuellement, la suivant. Nous voulions alors identifier une configuration, ou une suite d'événements, propice à une brusque évolution du cours du Bitcoin.

3 Outils

Cette problématique définie, nous avons décidé de joindre notre problème à l'un des éléments introduits en cours (examen) : la fouille de séquences. En effet celle-ci se conforme bien à l'étude d'une série chronologique telle que l'évolution du cours d'une monnaie. L'extraction de séquences fréquentes intéressantes pouvait en outre nous permettre d'utiliser, et surtout d'implémenter directement afin de mieux maîtriser, l'algorithme de fouille de données que nous connaissons le mieux : APriori. Ce dernier nécessitera une certaine adaptation que nous décrirons plus loin. Le langage Python nous a semblé convenable et même pratique pour ce travail. Nous avons rédigé un Notebook (Jupyter) pour illustrer notre code par des graphes et tableaux.

```
plt.show()
```

```
Max length = 2763
Max time = 2761 days 23:00:00
```



In [3]:
```
df['timestamp'] = df['timestamp'].dt.strftime('%Y-%m-%d')
df2 = df.set_index('timestamp')
df2.head(7)
```

Out[3]:

timestamp	close	high	low	open	time	volumefrom	volumeto
2010-07-17	0.04951	0.04951	0.04951	0.04951	1279324800	20.00	0.9902
2010-07-18	0.08584	0.08585	0.05941	0.04951	1279411200	75.01	5.0900
2010-07-19	0.08080	0.09307	0.07723	0.08584	1279497600	574.00	49.6600
2010-07-20	0.07474	0.08181	0.07426	0.08080	1279584000	262.00	20.5900
2010-07-21	0.07921	0.07921	0.06634	0.07474	1279670400	575.00	42.2600
2010-07-22	0.05050	0.08181	0.05050	0.07921	1279756800	2160.00	129.7800

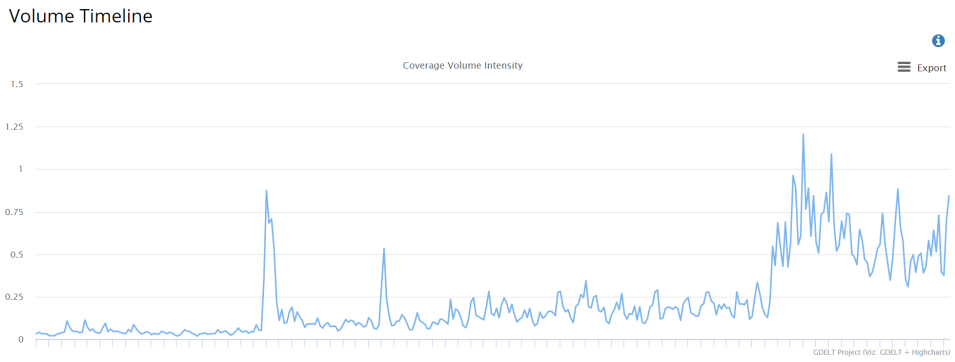
II Préparation des données

1 Recherche et import

Les premières et plus importantes données qu'il fallait nous procurer étaient évidemment celles concernant le prix du Bitcoin sur différentes périodes de temps. Pour le Bitcoin, comme pour la plupart des autres cryptomonnaies, plusieurs sites existent et proposent l'historique de son prix d'échange (en dollar). Nous avons utilisé l'API de *cryptocompare.com*, qui fournit tout l'historique par minute, par heure ou par jour de n'importe quelle cryptomonnaie très facilement sous format Json. Nous avons alors utilisé un script python simple permettant de récupérer depuis l'url de l'API ce fichier pour les valeurs jour par jour et depuis le début de collecte par le site, et de retourner un dataframe de la library *Pandas* correspondant.

La littérature concernant l'évolution des cours de bourse fait souvent état d'un fait : le cours d'une valeur s'explique souvent par sa propre valeur aux temps précédents. Mais en dehors de ses variations, nous avons tout de même voulu inclure dans nos séquences d'événements une variable décrivant la médiatisation du Bitcoin. La notoriété soudaine du Bitcoin étant par-

fois invoquée pour expliquer les fortes hausses de sa valeur, nous pensions qu'un indice de sa popularité à un temps précis pourrait peut-être expliquer son cours, et que si son impact était négligeable, il s'agirait d'une information presque aussi intéressante. Pour trouver des données sur la couverture médiatique, la tâche s'est avérée plus compliquée. Nous avons d'abord longuement recherché un moyen de scraper les données d'occurrence de termes ou de nombre de followers de certaines pages de *Twitter* puis du nombre de topics créés ou encore du nombre d'abonnés aux *sub* appropriés de *Reddit*, mais nous nous sommes vite confrontés à plusieurs problèmes : l'historique des données n'était pas forcément existant, et lorsqu'il l'était nous avons appris après échanges sur des forums que les données ne nous étaient pas accessibles (seuls certains graphes générés). Finalement, nous sommes tombés sur un site qui recense la couverture médiatique de plusieurs sujets, dont les Bitcoins : *https://blog.gdeltproject.org/* :



Nous avons donc importé ces données qui étaient sous format csv.

2 Nettoyage et description

Nous avons dans un premier temps créé à partir des données récupérées les variables **buzz**, **écart**, et **variation**:

La variable buzz est une variable qualitative qui nous indique l'importance de la couverture médiatique du Bitcoin. Cette variable peut prendre 4 valeurs: 'd', 'D', 'u', et 'U'. Nous les avons déterminées en regardant l'écart entre les valeurs de la couverture médiatique, et la moving average au temps précédent (ce afin que les variations "courantes" autour d'une même moyenne ne soient pas considérées comme exceptionnelles). 'd' indique que l'écart est faible et négatif (couverture médiatique faiblement décroissante), 'D' indique une forte décroissance. 'u' indique que la médiatisation est croissante, et 'U' indique un fort buzz.

La variable écart correspond à la différence des prix à l'ouverture et à la fermeture du marché sur une journée. Elle nous permet de déterminer les variations du prix.

Et donc, enfin, la variable variation est une variable qualitative qui nous indique l'importance de l'évolution du cours du Bitcoin et s'étend de '-3' (pic de décroissance) à '3' (pic de croissance). '0' indique une stagnation (évolution du prix inférieure à un certain seuil).

Les seuils de croissance et décroissance ont été testés avec plusieurs pourcentages. Les seuils qui nous ont finalement semblé pertinents sont $< 0.5\%$ pour définir une stagnation, $0.5\% < . < 5\%$ pour une variation simple (valeurs 1 et -1), $5\% < . < 10\%$ pour une variation intermédiaire (valeurs 2 et -2) et $> 10\%$ pour une forte variation (valeurs 3 et -3).

Enfin, nous avons concaténé ces deux nouvelles données dans la même table que les données précédentes, et donc indicées par les dates d'occurrence.

3 Création des séquences

Notre base de données de séquences a été construite à partir des colonnes 'buzz' et 'variation'. Le but est de se concentrer sur les séquences au moment des pics de prix, nous avons donc centré les séquences sur les valeurs les plus extrêmes de la colonne 'variation' : 3 et -3, c'est à dire lorsqu'il y a eu des variations d'un ordre supérieur à 10% sur une même journée. Une séquence est donc toujours de taille impaire avec une valeur extrême au milieu. Si un jour correspond à une faible/forte augmentation/diminution de la couverture médiatique du Bitcoin, la séquence de variation contiendra aussi cette information.

Par exemple: ['1'], ['0'], ['3', 'U'], ['0','d'], ['-2']

Cet exemple indique un prix en hausse, puis stagnant suivit d'une forte augmentation du cours accompagné d'une forte couverture médiatique le même jour, suivi d'une stagnation du prix couplée à une baisse de la couverture médiatique, puis d'une baisse relativement forte du cours.

La base de données contiendra alors autant de séquences à analyser qu'il y a de valeurs extrêmes de variation. Nous dotons notamment la fonction chargée de la création des séquences d'un paramètre permettant de modifier la taille des séquences générées.

III Extraction des informations

1 Implémentation A priori

Une fois que nos séquences ont été construites, nous avons codé les fonctions qui sont utiles à l'implémentation de l'algorithme A Priori, pour la recherche de sous séquences fréquentes. Comme nous nous intéressons particulièrement aux pics de valeurs, nous avons ajouté la conditions que les valeurs extrêmes devaient être inclues dans ces sous séquences. A partir de ça, on a crée une fonction 'recherche' qui lance l'algorithme A Priori pour différentes tailles de séquences (donc nombre de jours) en entrée.

Nous avons ensuite lancé la fonction 'recherche' plusieurs fois en augmentant graduellement le support minimal ainsi que la taille minimale des sous séquences que l'on voulait obtenir.

2 Résultats

On a observé que dans les sous-séquences que l'on obtenait à partir de séquences en entrée de longueur 25, donc sur 25 jours, une structure revenait très souvent: ['-2'], ['-3'], ['3'], ['0'], ['3'], ['-1']

qui correspond à une diminution moyenne, puis forte, puis une grosse augmentation, puis une stagnation, une autre grosse augmentation, et enfin une légère diminution. Cette structure est apparue 15 fois en un an!

D'autres structures apparaissent souvent, et sont liées à la couverture médiatique. Elle montrent que le pic de médiatisation succède toujours au pic de prix. Donc si vous investissez alors que tous les média en parlent de partout, ce n'est sans doute pas le meilleur moment pour investir!

Conclusion

On a donc pu trouvé des comportements très intéressants dans l'évolution du cours du Bitcoin. Il serait intéressant d'appliquer ces résultat à des décisions d'investissement réels, afin de tester leur validité. De plus, on pourrait extrêmement facilement appliquer ces algorithmes à n'importe quelle cryptomonnaies, et comparer les résulats.