Rapport de Projet Alla2 : Prédiction du Prix du Bitcoin avec le Machine Learning

---

Introduction

La Fascination pour la Cryptomonnaie : Un Croisement entre Technologie et Philosophie

Dans un monde où la technologie et la finance se rencontrent, les cryptomonnaies émergent comme une révolution fascinante. Ma passion pour ce domaine ne se limite pas à l'aspect technologique ; elle s'étend à la philosophie qui sous-tend ces monnaies numériques. Le Bitcoin, en particulier, représente bien plus qu'une simple devise : c'est un symbole de décentralisation et d'innovation, défiant les normes établies du système financier.

Cette fascination m'a conduit à explorer les profondeurs du Bitcoin, non seulement en tant que monnaie, mais aussi en tant que phénomène complexe et dynamique. L'intersection de la technologie blockchain, de la cryptographie, et des principes économiques crée un terrain fertile pour l'expérimentation et l'innovation. C'est dans cet esprit que j'ai entrepris un projet ambitieux : utiliser le machine learning pour prédire les fluctuations du prix du Bitcoin.

Présentation du Projet

Le projet que j'ai mené se concentre sur la prédiction de l'évolution du prix du Bitcoin pour le lendemain. Cette tâche, loin d'être triviale, représente un défi majeur dans le domaine de la finance et de la technologie. L'objectif était double : non seulement prédire la direction du mouvement du prix (augmentation ou baisse), mais aussi explorer les limites et les possibilités offertes par les techniques modernes de machine learning dans ce contexte.

Plan du Rapport

Collecte et Préparation des Données :

Modélisation et Approches :

Validation et Métriques d'Évaluation :

Résultats et Analyse :

Réflexions et Perspectives d'Amélioration :

Collecte et Préparation des Données

Introduction

La précision des prédictions dans le domaine des cryptomonnaies dépend fortement de la qualité et de la pertinence des données utilisées. Mon projet sur la prédiction des prix du Bitcoin a débuté par une étape fondamentale : la collecte et la préparation minutieuse des données. Cette phase a non seulement établi les bases pour les modèles de machine learning, mais a également été une aventure en soi, riche en apprentissages et en défis.

Collecte des Données : Un Choix Stratégique de Sources

J'ai sélectionné des sources de données fiables et complètes pour assurer la robustesse de mon analyse. Les API de BitQuery, CryptoCompare et CoinGecko ont été mes principales sources, fournissant des informations détaillées sur l'historique des prix, les volumes de transactions, et la difficulté de minage du Bitcoin. Ces plateformes ont été choisies pour leur précision et leur fiabilité, des aspects cruciaux dans le contexte fluctuant des cryptomonnaies.

Création de Features Techniques

La préparation des données a impliqué un travail approfondi sur les indicateurs techniques, essentiels pour analyser et prédire les mouvements du marché. Voici les features que j'ai calculées :

- EMA (Exponential Moving Average) 26 et 12 : Des moyennes mobiles pondérées qui mettent l'accent sur les données récentes.
- MACD (Moving Average Convergence Divergence) : Un indicateur de tendance suivi de près dans le trading de cryptomonnaies.
- RSI (Relative Strength Index) : Un outil pour mesurer la dynamique des mouvements de prix.
- Volume Relatif : Comparaison du volume actuel au volume moyen sur une période étendue.

- OBV (On-Balance Volume) : Un indicateur qui utilise le volume de flux pour anticiper les changements de prix.
- ATR (Average True Range) : Pour évaluer la volatilité du marché.
- Bandes de Bollinger : Utiles pour identifier les conditions de surachat ou de survente.
- Oscillateur Stochastique et Momentum : Pour mesurer le momentum et les conditions de marché extrêmes.

Défis et Solutions

La collecte et la préparation des données ont présenté plusieurs défis. Le premier a été de garantir la fiabilité et la pertinence des données dans un contexte où les sources gratuites sont souvent limitées. Le calcul des indicateurs techniques a nécessité une compréhension approfondie de chaque indicateur et de son impact potentiel sur les prédictions du modèle. J'ai dû également veiller à la cohérence et à la normalisation des données pour assurer leur efficacité dans les modèles de machine learning.

Conclusion

Cette première phase du projet a été essentielle, soulignant l'importance d'une base de données solide et bien préparée en machine learning. Les données collectées et transformées ont jeté les fondations pour les étapes suivantes de modélisation et d'analyse, ouvrant la voie à des insights potentiellement révélateurs sur le comportement futur du prix du Bitcoin.

Modélisation et Approches

Introduction

Après avoir préparé une base solide de données, je me suis concentré sur la phase de modélisation. Cette étape est essentielle pour déterminer la capacité du système à prédire avec précision les mouvements du prix du Bitcoin. J'ai exploré plusieurs modèles de machine learning, en mettant l'accent sur leur fonctionnement, leur pertinence, et les paramètres clés.

Les Modèles Exploités

1. Logistic Regression

Fonctionnement : La régression logistique est un modèle statistique qui estime la probabilité qu'une variable dépendante appartienne à une certaine catégorie. Dans ce cas, elle prédit la probabilité que le prix du Bitcoin augmente ou diminue.

Pertinence : Ce modèle est apprécié pour sa simplicité et son efficacité dans les problèmes de classification binaire. Il est particulièrement utile pour comprendre l'impact de chaque feature sur la prédiction.

Hyperparamètres Principaux : 'C' pour la régularisation et le 'solver' pour l'optimisation.

2. XGBoost

Fonctionnement : XGBoost (eXtreme Gradient Boosting) est une implémentation avancée de l'algorithme de boosting de gradient. Il construit séquentiellement des modèles plus simples, en se concentrant sur les erreurs des modèles précédents pour améliorer les prédictions.

Pertinence : XGBoost est réputé pour sa performance et sa vitesse, surtout avec des ensembles de données de grande dimension. Il est efficace pour capter des patterns complexes dans les données.

Hyperparamètres Principaux : 'colsample_bytree', 'learning_rate', 'max_depth', 'n_estimators', 'subsample'. Une recherche grid sera utilisée pour optimiser ces paramètres.

3. LSTM (Long Short-Term Memory)

Fonctionnement : Les LSTM sont une forme spéciale de réseaux de neurones récurrents, capables de capturer des dépendances à long terme dans les séquences de données. Ils sont particulièrement adaptés pour traiter des séries temporelles.

Pertinence : Dans le contexte des données temporelles du Bitcoin, les LSTM sont idéaux pour comprendre et prédire les tendances sur la base de l'historique des prix et d'autres indicateurs

Hyperparamètres Principaux : Nombre de couches, nombre de neurones par couche, longueur des séquences d'entrée.

Stratégie de Sélection des Features

Le feature engineering est un pilier essentiel dans ce projet de prédiction des prix du Bitcoin. Cette étape va bien au-delà de la simple sélection de données ; elle implique une analyse minutieuse et une compréhension profonde des dynamiques du marché des cryptomonnaies. Les features telles que l'EMA, le MACD, et le RSI ont été soigneusement choisies pour leur pertinence dans l'analyse des tendances du marché. En parallèle, une exploration créative a été menée pour développer de nouvelles features, combinant des indicateurs existants et expérimentant avec différentes transformations des données. L'objectif était de capturer au mieux les nuances et les fluctuations du marché, tout en maintenant l'équilibre entre la richesse des informations et la simplicité du modèle pour éviter le surajustement.

Conclusion

Cette phase de modélisation a été un équilibre entre l'analyse technique et la stratégie de sélection des features. En utilisant la régression logistique, XGBoost et les LSTM, j'ai exploré diverses facettes des données du Bitcoin. Chaque modèle, avec ses hyperparamètres, a révélé des insights uniques sur le marché. Maintenant, l'étape suivante consiste à appliquer la métrique d'évaluation "earn_metric" et la méthode de validation walk forward, essentielles pour tester l'efficacité des modèles dans un marché en constante évolution.

Validation et Métriques d'Évaluation

Introduction

L'évaluation des modèles de machine learning est une étape cruciale pour déterminer leur efficacité et leur applicabilité. Dans ce projet, j'ai utilisé deux métriques principales pour évaluer les modèles de

classification : une métrique personnalisée que j'ai développée, nommée "earn_metric", et la métrique classique d'accuracy.

Validation Walk Forward

La validation walk forward est une technique de validation croisée adaptée aux séries temporelles. Elle consiste à entraîner le modèle sur une fenêtre de données historiques, puis à tester le modèle sur la période suivante, en avançant progressivement la fenêtre d'entraînement et de test.

Importance : Cette méthode est particulièrement pertinente pour les données temporelles, comme celles du Bitcoin, car elle respecte l'ordre chronologique des données. Elle permet d'évaluer la capacité du modèle à s'adapter aux changements du marché et à faire des prédictions précises sur des données non vues auparavant.

Application : J'ai appliqué la validation walk forward pour tous les modèles testés, ce qui a permis une évaluation réaliste de leur performance dans un contexte de marché dynamique et en évolution.

La Métrique Personnalisée : Earn Metric

Concept et Fonctionnement

La "earn_metric" est une innovation conçue pour évaluer la performance des modèles en termes de gains financiers potentiels. L'idée derrière cette métrique est de mesurer non seulement la précision des prédictions, mais aussi leur impact économique réel lorsqu'elles sont appliquées dans un contexte de trading de Bitcoin.

Processus : La métrique calcule le rapport entre les gains réalisés en suivant les probabilités prédites par le modèle et les gains obtenus en restant investi à 100% en Bitcoin sur une période de n_days.

Application : Pour chaque prédiction, le modèle détermine une probabilité de hausse ou de baisse du prix du Bitcoin. Cette probabilité est utilisée pour ajuster la répartition du capital entre Bitcoin et USD. La métrique évalue ensuite les gains réalisés en suivant cette stratégie sur une période donnée, comparés aux gains d'un investissement constant en Bitcoin.

Importance

Cette métrique est particulièrement pertinente dans le contexte du trading de cryptomonnaies, car elle va au-delà de la simple précision des prédictions. Elle offre une perspective pratique sur la manière dont les prédictions peuvent être utilisées pour maximiser les gains financiers.

La Métrique d'Accuracy

En parallèle à la "earn_metric", j'ai également utilisé la métrique d'accuracy pour évaluer les modèles.

Définition : L'accuracy mesure la proportion de prédictions correctes par rapport au total des prédictions. C'est une métrique standard dans la classification binaire, indiquant simplement combien de fois le modèle a correctement prédit si le prix du Bitcoin allait augmenter ou baisser.

Rôle : Bien que l'accuracy ne fournisse pas d'insights sur les gains financiers, elle reste un indicateur important de la performance générale du modèle. Elle permet de comparer les modèles de manière standardisée et de s'assurer que les prédictions sont globalement fiables.

Conclusion

En combinant la validation walk forward avec les métriques "earn_metric" et accuracy, j'ai pu évaluer de manière complète et réaliste les modèles de classification. La validation walk forward a joué un rôle clé en respectant la séquentialité des données et en testant l'adaptabilité des modèles aux conditions changeantes du marché. La "earn_metric", avec son approche innovante axée sur les gains financiers, et l'accuracy, en tant que mesure standard de précision, ont ensemble fourni une évaluation équilibrée, soulignant à la fois la précision et la rentabilité potentielle des modèles dans le trading de Bitcoin.

Résultats et Analyse des Modèles

Introduction

Après une série d'expérimentations et d'ajustements sur différents modèles et leurs paramètres, j'ai identifié le modèle le plus performant pour la prédiction des mouvements de prix du Bitcoin. Cette section présente les résultats clés et les enseignements tirés de cette phase de modélisation.

Le Meilleur Modèle : XGBoost

Configuration Optimale

Le modèle qui a montré les meilleures performances est un modèle XGBoost, configuré avec les hyperparamètres suivants :

colsample_bytree: 1.0
learning_rate: 0.01
max_depth: 5
n_estimators: 200
subsample: 0.8

Cette configuration a été le résultat d'un processus itératif de tuning des hyperparamètres, où j'ai cherché à équilibrer la capacité du modèle à capturer des patterns complexes sans tomber dans le surajustement.

Performance

Le modèle XGBoost, avec cette configuration, a démontré une capacité supérieure à prédire correctement les mouvements de prix du Bitcoin, en se basant sur les métriques d'accuracy et la "earn_metric". Il a surpassé les autres modèles, y compris la régression logistique, en termes de précision et de potentiel de gains financiers.

Le Défi avec LSTM

Malgré les avantages théoriques des LSTM pour les séries temporelles, je n'ai pas réussi à développer un modèle LSTM qui surpasse les autres approches.

Limitations : Le principal obstacle a été le manque d'expérience et de ressources pour optimiser pleinement les LSTM. Ces modèles nécessitent une compréhension approfondie et une capacité de calcul significative pour l'ajustement fin des hyperparamètres et l'architecture du réseau.

Potentiel : Je suis convaincu qu'avec plus de ressources et d'expertise, un modèle LSTM pourrait être développé pour surpasser les approches actuelles. Les LSTM ont un potentiel énorme dans la

modélisation des séries temporelles, en particulier avec des données aussi dynamiques que celles du Bitcoin.

Améliorations Possibles du XGBoost

Bien que le modèle XGBoost ait été le plus performant, il existe toujours un potentiel d'amélioration.

Optimisation Continue : Des ajustements supplémentaires des hyperparamètres, une sélection plus fine des features, et l'exploration de techniques avancées comme le stacking ou l'ensembling pourraient améliorer davantage sa performance.

Données Supplémentaires : L'intégration de données supplémentaires, en particulier celles accessibles via des API payantes, pourrait enrichir le modèle et affiner ses prédictions.

Conclusion

Les résultats obtenus avec le modèle XGBoost sont prometteurs, mais ils soulignent également les défis et les opportunités d'amélioration. Le potentiel non réalisé du LSTM et les possibilités d'optimisation du XGBoost ouvrent la voie à des réflexions plus approfondies sur ce projet. Dans le prochain chapitre, je me pencherai sur les leçons tirées, les défis rencontrés, et les perspectives d'amélioration pour des modèles encore plus performants.

Réflexions et Perspectives d'Amélioration

Ce projet de prédiction des prix du Bitcoin a été une aventure enrichissante, offrant de nombreuses leçons et révélant des pistes d'amélioration. Dans cette conclusion, je réfléchis aux aspects clés du projet et envisage les directions futures pour le rendre encore plus performant et accessible.

L'Importance Cruciale des Données

La qualité des données est un pilier fondamental dans tout projet de machine learning, et ce projet ne fait pas exception. L'accès à des données plus diversifiées et détaillées pourrait considérablement améliorer la performance des modèles.

API Glassnode : L'une des avenues prometteuses est l'utilisation de l'API Glassnode, qui offre des données approfondies sur les cryptomonnaies. Bien que son coût soit actuellement prohibitif pour moi, l'accès à ces données pourrait ouvrir de nouvelles perspectives pour affiner les prédictions.
Richesse des Features : Des features plus sophistiquées, telles que celles disponibles via Glassnode, pourraient apporter des insights plus nuancés sur le marché du Bitcoin, améliorant ainsi la précision des modèles.

Exploration Continue des LSTM

Le potentiel des LSTM dans ce projet reste largement inexploré. Avec des ressources et une expertise accrues, je suis convaincu qu'un modèle LSTM optimisé pourrait surpasser les performances actuelles.

Optimisation et Ressources : Une exploration plus poussée des architectures LSTM, combinée à une puissance de calcul accrue, pourrait mener à un modèle extrêmement efficace pour les prédictions de séries temporelles.

Expertise Spécialisée : L'acquisition de compétences plus approfondies dans le domaine des réseaux de neurones récurrents serait un atout majeur pour exploiter pleinement le potentiel des LSTM.

Vers une Plateforme en Ligne

L'idée de rendre ce projet accessible au grand public est séduisante. Une plateforme en ligne permettrait à d'autres de bénéficier des insights générés par le modèle.

Accessibilité : En développant une interface utilisateur intuitive, le modèle pourrait être utilisé par des traders de tous niveaux pour informer leurs décisions d'investissement.

Perfectionnement Préalable : Avant de lancer une telle plateforme, je souhaite peaufiner davantage le modèle pour garantir sa fiabilité et sa pertinence pour les utilisateurs.

Exploration d'Autres Algorithmes

Bien que le modèle XGBoost ait montré d'excellentes performances, l'exploration d'autres algorithmes reste une piste intéressante.

Random Forest et Autres : Des algorithmes comme Random Forest ou d'autres techniques avancées pourraient offrir des perspectives différentes et potentiellement plus efficaces pour la prédiction des prix du Bitcoin.

Comparaison et Complémentarité : Tester divers algorithmes permettrait non seulement de comparer leurs performances, mais aussi d'explorer des combinaisons de modèles pour une prédiction plus robuste.

Conclusion

Ce projet a été un parcours d'apprentissage intense, soulignant l'importance des données de qualité, l'exploration de nouvelles techniques de modélisation, et l'ouverture vers des applications pratiques. Les perspectives d'amélioration sont vastes, allant de l'accès à des données plus riches à l'exploration de nouvelles méthodologies de modélisation. L'objectif final est de créer un outil de prédiction non seulement performant mais aussi accessible, contribuant ainsi à la communauté du trading de cryptomonnaies.