Lecture – Ingénierie des données et pipelines d’apprentissage automatique

Temps estimé nécessaire : 8 minutes

Pipelines d'ingénierie des données

Dans le monde actuel axé sur les données, les organisations recherchent constamment des informations précieuses et utilisent des algorithmes avancés pour prendre des décisions éclairées. Les pipelines d'ingénierie des données sont la base de projets axés sur les données réussis. Ils gèrent la collecte, la transformation et le stockage de grandes quantités de données brutes. Les ingénieurs de données conçoivent et mettent en œuvre des systèmes robustes pour gérer les données à grande échelle. Ils utilisent des outils et des technologies pour nettoyer, transformer et intégrer différentes sources de données dans un format fiable.

La qualité des données est essentielle dans l'ingénierie des données. Les ingénieurs créent des pipelines de données efficaces pour traiter les données de manière fluide. Ces pipelines impliquent l'extraction de données à partir de diverses sources, la réalisation des modifications nécessaires et leur stockage dans des systèmes de stockage ou d'analyse. Les ingénieurs de données collaborent avec les data scientists, les analystes et les parties prenantes pour comprendre leurs besoins et leur fournir des données propres et accessibles.

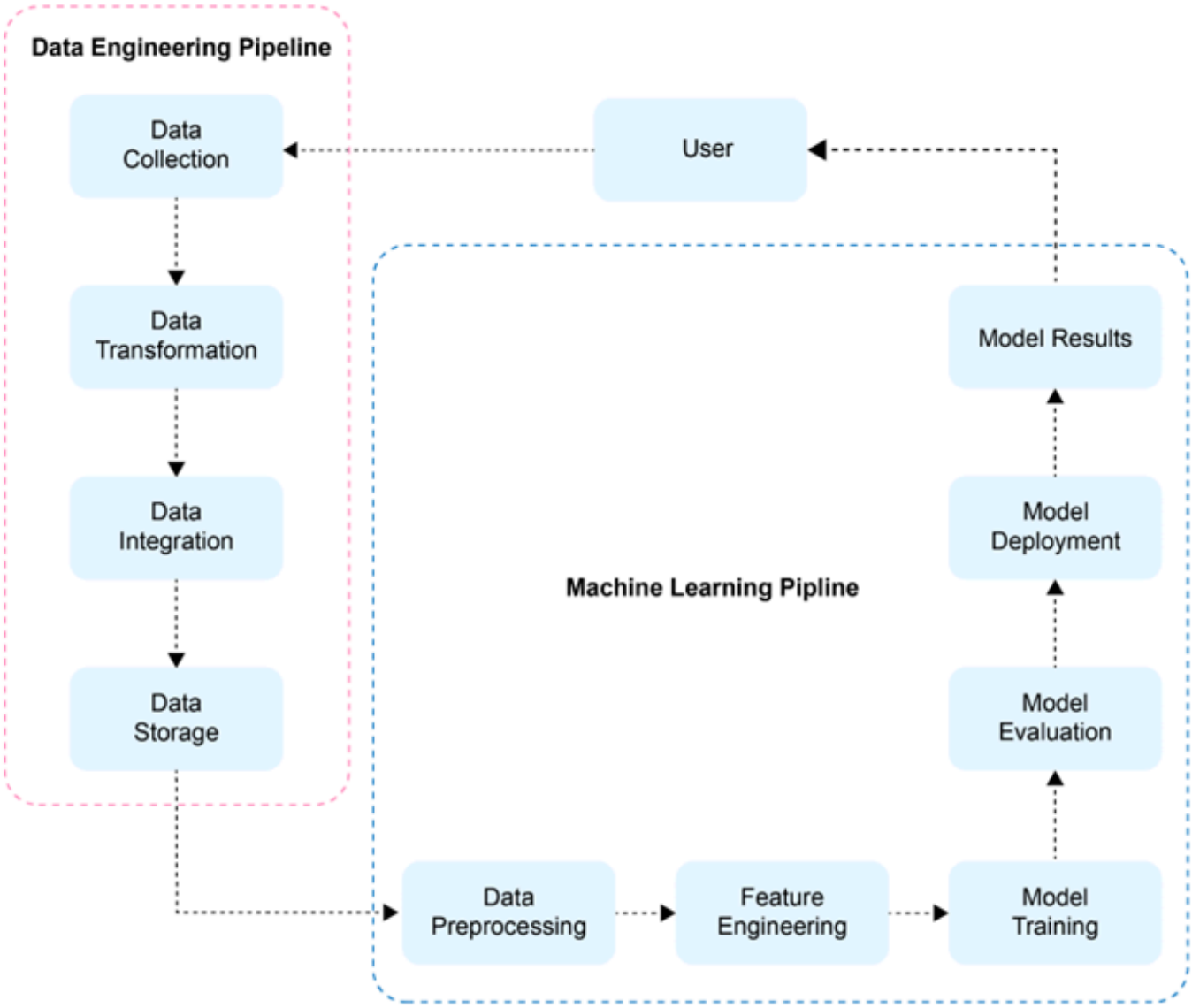
Examinons de plus près les quatre éléments clés qui composent les pipelines d’ingénierie des données :

1. **Collecte de données** : les ingénieurs de données sont chargés de collecter des données à partir de diverses sources. Cela comprend les bases de données, les API, le scraping Web, les plateformes de streaming, etc. En combinant des données provenant de plusieurs sources, les ingénieurs de données garantissent un ensemble de données complet et diversifié avec lequel travailler.

2. **Transformation des données** : une fois les données collectées, elles subissent une série de transformations pour garantir leur qualité et leur utilité. Cela implique des étapes de prétraitement telles que le nettoyage des données pour supprimer les erreurs et les incohérences, la gestion des valeurs manquantes et la résolution des valeurs aberrantes. Les ingénieurs de données appliquent également des techniques telles que la normalisation et la standardisation des données pour garantir l'uniformité et la comparabilité entre différents points de données.

3. **Intégration des données** : les organisations traitent souvent des données provenant de différentes sources et sous différents formats. Les ingénieurs de données jouent un rôle crucial dans l'intégration de ces données non liées pour créer un ensemble de données unifié et cohérent. Cela implique de fusionner des données provenant de diverses sources, d'effectuer des jointures de données pour combiner des informations connexes et d'agréger des données pour obtenir une vue consolidée.

4. **Stockage des données** : les données traitées et intégrées doivent être stockées dans un référentiel adapté pour une accessibilité et une évolutivité faciles. Les ingénieurs de données utilisent des systèmes d'entreposage de données ou des lacs de données pour stocker les données de manière sécurisée et efficace. Ces référentiels constituent la base d'une analyse plus approfondie et servent de centre centralisé pour les opérations basées sur les données.



Les pipelines d'ingénierie des données sont conçus pour gérer efficacement de gros volumes de données. Ils assurent la fluidité du flux de données, de la collecte au stockage, jetant ainsi les bases des analyses et des informations ultérieures.

Pipelines d'apprentissage automatique

Les pipelines d'apprentissage automatique jouent un rôle crucial dans l'extraction d'informations précieuses à partir de données à l'aide d'algorithmes. Ces pipelines couvrent l'ensemble du processus de création, de formation et de déploiement de modèles d'apprentissage automatique.

Les data scientists et les ingénieurs en machine learning collaborent étroitement pour optimiser les pipelines et améliorer les performances des modèles. Des techniques telles que l'échantillonnage des données, l'extraction de caractéristiques et la sélection de modèles garantissent des prévisions précises et des résultats significatifs. De plus, des concepts tels que la validation croisée, le réglage des hyperparamètres et des stratégies de déploiement de modèles efficaces sont intégrés aux pipelines de machine learning pour créer des solutions robustes et évolutives.

Explorons les cinq éléments essentiels qui composent ces pipelines :

1. **Prétraitement des données** : les données brutes nécessitent souvent un prétraitement avant de pouvoir être utilisées efficacement pour des tâches d'apprentissage automatique. Le prétraitement des données implique la gestion des valeurs manquantes, le traitement des variables catégorielles par codage ou codage à chaud, et la normalisation des caractéristiques numériques pour les amener dans une plage

cohérente. Cette étape garantit que les données sont dans un format adapté aux étapes suivantes du pipeline.

2. **Ingénierie des fonctionnalités** : l'ingénierie des fonctionnalités est le processus de sélection, de création ou de transformation des fonctionnalités dans l'ensemble de données pour améliorer les performances des modèles d'apprentissage automatique. Cela implique l'extraction de fonctionnalités pertinentes, la création de nouvelles fonctionnalités basées sur les connaissances du domaine ou l'application de techniques telles que la réduction de la dimensionnalité pour réduire la complexité des données. L'ingénierie des fonctionnalités joue un rôle crucial dans la capture des modèles et des relations sous-jacents dans les données.

3. **Entraînement du modèle** : à ce stade, des algorithmes d'apprentissage automatique sont appliqués à l'ensemble de données prétraitées et conçues pour apprendre des modèles et faire des prédictions. Les scientifiques des données et les ingénieurs en apprentissage automatique sélectionnent les algorithmes appropriés en fonction de la nature du problème et des données disponibles. Les modèles sont entraînés à l'aide de données étiquetées, ce qui leur permet d'apprendre à partir des modèles et de faire des prédictions ou des classifications précises.

4. **Évaluation du modèle** : les modèles formés doivent être évalués pour évaluer leurs performances et leur généralisabilité. Diverses mesures, telles que l'exactitude, l'erreur quadratique moyenne (MSE), la précision, le rappel et le score F1, sont utilisées pour évaluer les capacités prédictives des modèles. Des techniques de validation croisée sont utilisées pour valider les performances des modèles sur des données non vues, garantissant qu'ils peuvent être généralisés bien au-delà des données de formation.

5. **Déploiement du modèle** : le modèle le plus performant est déployé dans un environnement de production, où il peut faire des prédictions ou générer des informations sur des données nouvelles et invisibles. Cette étape implique l'intégration du modèle dans des systèmes existants ou la création d'API qui permettent une intégration facile avec d'autres applications. Les stratégies de déploiement de modèles garantissent que les modèles sont évolutifs, robustes et capables de gérer des flux de données en temps réel.

Les pipelines d’apprentissage automatique permettent aux organisations d’exploiter efficacement les algorithmes et de prendre des décisions basées sur les données, permettant ainsi l’automatisation, l’optimisation et les capacités prédictives.

Combler le fossé : pipelines d'ingénierie des données et pipelines d'apprentissage automatique

Les pipelines d'ingénierie des données et les pipelines d'apprentissage automatique peuvent avoir des objectifs différents, mais ils sont étroitement liés et se soutiennent mutuellement. Les pipelines d'ingénierie des données fournissent des données propres et bien structurées, ce qui est essentiel pour les pipelines d'apprentissage automatique réussis. Sans données de qualité, les modèles d'apprentissage automatique ne peuvent pas faire de prédictions précises ni trouver d'informations pertinentes.

Les pipelines d'ingénierie des données alimentent les pipelines d'apprentissage automatique avec les données nécessaires à la formation et à l'évaluation des modèles. En retour, les pipelines d'apprentissage automatique fournissent des informations précieuses pour améliorer le processus d'ingénierie des données. Les informations et les prévisions issues des modèles d'apprentissage automatique permettent d'identifier les domaines dans lesquels l'ingénierie des données peut être améliorée. Cette boucle de rétroaction continue garantit que les pipelines d'ingénierie des données s'adaptent aux besoins évolutifs des flux de travail d'apprentissage automatique.

Applications concrètes

La collaboration entre les pipelines d'ingénierie des données et les pipelines d'apprentissage automatique permet aux organisations d'atteindre divers objectifs. Voici quelques scénarios courants :

1. **Prise de décision en temps réel** : les pipelines d'ingénierie des données traitent et fournissent des données en temps réel, ce qui permet aux modèles d'apprentissage automatique de prendre des décisions opportunes et éclairées. Cela est essentiel pour des applications telles que la détection des fraudes, les systèmes de recommandation et la tarification dynamique. Les ingénieurs de données s'assurent que les données sont collectées et traitées rapidement, fournissant des données d'entrée pour les prévisions d'apprentissage automatique en temps réel.

2. **Traitement évolutif des données** : les pipelines d'ingénierie des données sont conçus pour gérer efficacement de grandes quantités de données, garantissant ainsi évolutivité et performances. Les pipelines d'apprentissage automatique peuvent tirer parti de cette capacité pour traiter des ensembles de données volumineux et former des modèles complexes. En optimisant le traitement et le stockage des données, les ingénieurs de données permettent aux pipelines d'apprentissage automatique de gérer des analyses de données à grande échelle et de générer des informations précieuses.

3. **Déploiement et surveillance des modèles** : les pipelines d'ingénierie des données facilitent le déploiement transparent des modèles d'apprentissage automatique dans les environnements de production. Ils garantissent la cohérence, la sécurité et la maintenabilité des données, tandis que les pipelines d'apprentissage automatique permettent la surveillance des modèles et l'optimisation des performances. Les ingénieurs de données travaillent aux côtés des ingénieurs d'apprentissage automatique pour garantir un déploiement et une intégration efficaces des modèles, ainsi que des pipelines de données robustes et évolutifs.

Conclusion

Les pipelines d'ingénierie des données et les pipelines d'apprentissage automatique jouent un rôle essentiel dans le paysage axé sur les données. L'ingénierie des données établit une infrastructure de données fiable, tandis que les pipelines d'apprentissage automatique exploitent des algorithmes pour extraire des informations précieuses. Grâce à la collaboration et à l'intégration, ces pipelines transforment les données brutes en informations exploitables, permettant aux organisations de prendre des décisions éclairées et de rester compétitives dans le monde actuel axé sur les données. La synergie entre l'ingénierie des données et les pipelines d'apprentissage automatique révèle la véritable valeur des données et stimule l'innovation dans divers domaines.

Auteur(s)

Ramesh Sannareddy, Pooja Bhardwaj

Autre(s) contributeur(s)

Andrew Pfeiffer