Livre blanc AWS

Principes de prévisions de séries temporelles avec Amazon Forecast

Principes de prévisions de séries temporelles avec Amazon Forecast: Livre blanc AWS

Table of Contents

Principes de prévisions de séries temporelles avec Amazon Forecast

Date de publication : 1er septembre 2021 (Historique du document)

Les entreprises utilisent aujourd'hui absolument tout, des simples feuilles de calcul aux logiciels de planification financière complexes, pour tenter de prévoir avec précision les futurs résultats commerciaux tels que la demande de produits, les besoins en ressources et les performances financières. Ce document présente la prévision, sa terminologie, ses défis et ses cas d'utilisation. Ce document utilise une étude de cas pour renforcer les concepts de prévision, les étapes de la prévision, et indique commentAmazon Forecast peut aider à résoudre les nombreux défis pratiques dans les problèmes de prévision du monde réel.

Présentation

La prévision est la science qui consiste à prédire l'avenir. En utilisant des données historiques, les entreprises peuvent comprendre les tendances, prévoir ce qui pourrait se produire et quand, et à leur tour, intégrer ces informations dans leurs plans futurs pour tout, de la demande de produits à la planification des stocks et à la dotation en personnel.

Compte tenu des conséquences des prévisions, la précision est importante. Si une prévision est trop élevée, les clients risquent de surinvestir dans les produits et le personnel, ce qui entraîne un gaspillage des investissements. Si les prévisions sont trop faibles, les clients risquent de sous-investir, ce qui entraîne un manque de matières premières et de stocks, créant ainsi une mauvaise expérience client.

Aujourd'hui, les entreprises essaient d'utiliser tous les moyens, des simples feuilles de calcul aux logiciels complexes de planification de la demande et de planification financière, pour générer des prévisions, mais il est difficile d'obtenir une précision élevée pour deux raisons :

- Premièrement, les prévisions traditionnelles peinent à intégrer de grands volumes de données historiques, manquant ainsi des signaux importants du passé qui se perdent dans le bruit.
- Deuxièmement, les prévisions traditionnelles intègrent rarement des données associées mais indépendantes, qui peuvent offrir un contexte important (comme le prix, les vacances/événements, les ruptures de stock, les promotions marketing, etc.) Sans l'historique complet et le contexte plus large, la plupart des prévisions ne parviennent pas à prédire l'avenir avec précision.

Amazon Forecast est un service entièrement géré qui permet de surmonter ces problèmes. Amazon Forecast fournit les meilleurs algorithmes pour le scénario de prévision en question. Il s'appuie sur les techniques modernes de machine learning (ML) et de deep learning lorsque cela est nécessaire pour fournir des prévisions très précises. Amazon Forecast est facile à utiliser et ne nécessite aucune expérience en matière de machine learning. Le service fournit automatiquement l'infrastructure nécessaire, traite les données et construit des modèles ML personnalisés/privés qui sont hébergés sur AWS et prêts à faire des prédictions. En outre, les techniques de machine learning évoluant rapidement, Amazon Forecast les intègre, de sorte que les clients continuent de bénéficier d'une plus grande précision, sans effort supplémentaire ou presque.

Votre infrastructure est-elle Well-Architected ?

Le cadre AWS Well-Architected vous permet de comprendre les avantages et les inconvénients des décisions que vous prenez lors de la création de systèmes dans le cloud. Les six piliers du cadre vous permettent d'apprendre les bonnes pratiques architecturales pour concevoir et exploiter des systèmes fiables, sécurisés, efficaces, rentables et durables. À l'aide du AWS Well-Architected Tool, disponible gratuitement dans la AWS Management Console, vous pouvez passer en revue vos charges de travail en fonction de ces bonnes pratiques en répondant à plusieurs questions pour chaque pilier.

Dans la Présentation détaillée du machine learning, nous nous concentrons sur la conception, le déploiement et l'architecture de vos charges de travail machine learning dans le AWS Cloud. Cette présentation complète les bonnes pratiques détaillées dans le cadre Well-Architected.

Pour obtenir davantage de conseils d'experts et de bonnes pratiques pour votre architecture de cloud — références aux déploiements d'architecture, diagrammes et livres blancs — consultez le Centre d'architecture AWS.

À propos des prévisions

Dans ce document, la prévision consiste à prédire les valeurs futures d'une série temporelle : l'entrée ou la sortie d'un problème est de nature temporelle.

Système de prévision

Un système de prévision comprend un ensemble diversifié d'utilisateurs :

- Les utilisateurs finaux, qui interrogent les prévisions pour un produit spécifique et décident du nombre d'unités à acheter ; il peut s'agir d'une personne ou d'un système automatisé.
- Les analystes métier/l'informatique décisionnelle, qui assistent les utilisateurs finaux, exécutent et organisent des rapports agrégés.
- Les spécialistes des données, qui analysent de manière itérative les modèles de demande, les effets causaux et ajoutent de nouvelles fonctionnalités pour apporter des améliorations incrémentales au modèle ou améliorer le modèle de prévision.
- Les ingénieurs, qui mettent en place l'infrastructure de la collecte des données et assurent la disponibilité des données d'entrée dans le système.

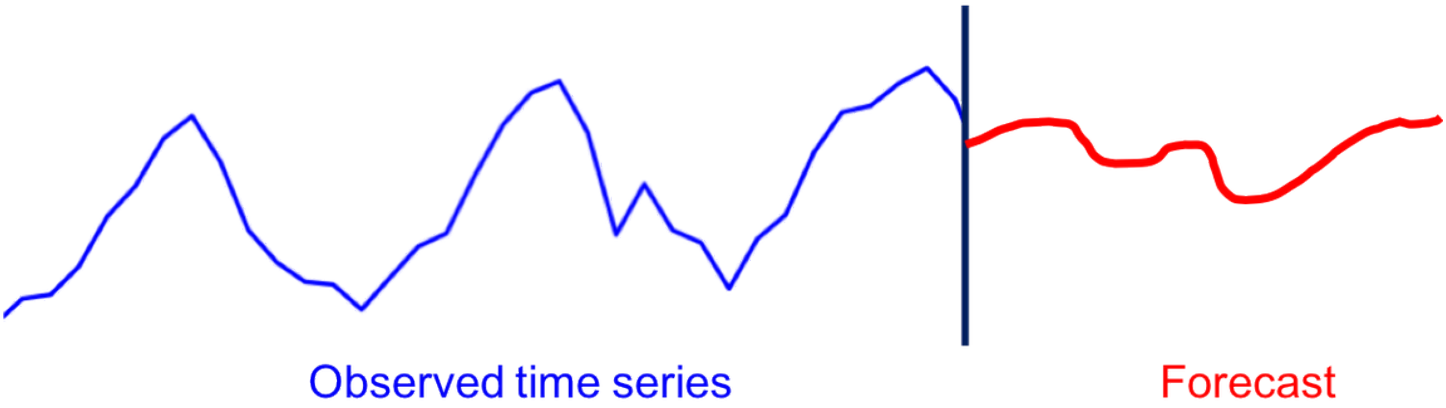
Amazon Forecast allège le travail des ingénieurs logiciels et permet aux entreprises disposant de capacités limitées en matière de science des données de tirer parti d'une technologie de prévision de pointe. Pour les entreprises disposant de capacités de science des données, un certain nombre de fonctionnalités de diagnostic sont incluses afin que les problèmes de prévision soient bien traités avec Amazon Forecast.

Où apparaissent les problèmes de prévision ?

Les problèmes de prévision se posent dans de nombreux domaines qui produisent naturellement des données de séries temporelles. Il s'agit notamment de la vente au détail, de l'analyse médicale, de la planification des capacités, de la surveillance des réseaux de capteurs, de l'analyse financière, de l'exploration des activités sociales et des systèmes de bases de données. Par exemple, les prévisions jouent un rôle clé dans l'automatisation et l'optimisation des processus opérationnels dans la plupart des entreprises qui permettent une prise de décision orientée données. Les prévisions de l'offre et de la demande de produits peuvent être utilisées pour la gestion optimale des stocks, l'ordonnancement du personnel et la planification de la topologie, et constituent plus

généralement une technologie essentielle pour la plupart des aspects de l'optimisation de la chaîne d'approvisionnement.

La figure suivante résume le problème de la prévision lorsque l'on se base sur une série temporelle observée qui présente un modèle (dans cet exemple, la saisonnalité), et que l'on crée une prévision sur une période donnée. L'axe horizontal représente le temps allant du passé (à gauche) au futur (à droite). L'axe vertical représente les unités mesurées. En tenant compte du passé (en bleu) jusqu'à la ligne verticale noire, l'identification du futur (en rouge) représente la tâche de prévision.



Vue d'ensemble des tâches de prévision

Points à prendre en considération avant de tenter de résoudre un problème de prévision

Les principaux enjeux à comprendre avant de résoudre des problèmes de prévision sont les suivants :

- Devez-vous résoudre un problème de prévision ?
- Pourquoi résolvez-vous le problème de prévision ?

En raison de l'omniprésence des données de séries temporelles, on peut facilement trouver des problèmes de prévision partout. Cependant, il convient de se demander s'il est vraiment nécessaire de résoudre un problème de prévision ou si vous pouvez le contourner complètement sans sacrifier l'efficacité de la prise de décision dans l'entreprise. Il est important de poser cette question car, scientifiquement parlant, la prévision fait partie des problèmes les plus difficiles en matière de machine learning.

Prenons l'exemple des recommandations de produits pour un détaillant en ligne. Ce problème de recommandation de produits peut être formulé comme un problème de prévision où, pour chaque

paire client-unité de gestion des stocks (référence), vous prévoyez le nombre d'unités d'un article spécifique que ce client particulier achètera. Cette formulation du problème présente un certain nombre d'avantages. L'un des avantages est que la composante temporelle est explicitement prise en compte, ce qui vous permet de recommander des produits en fonction des habitudes d'achat des clients.

Cependant, les problèmes de recommandation de produits sont rarement formulés comme un problème de prévision, car la résolution d'un tel problème de prévision est beaucoup plus difficile (par exemple, la rareté de l'information au niveau client-référence et l'échelle du problème) que la résolution directe du problème de recommandation. Par conséquent, lorsque vous envisagez une application de prévision, il est important de tenir compte de l'utilisation en aval de la prévision et de savoir s'il est possible de résoudre ce problème en utilisant une autre approche.

Amazon Personalize peut vous aider dans ces cas de figure. Amazon Personalize est un service de machine learning qui permet aux développeurs de créer des recommandations individuelles pour les clients qui utilisent leurs applications.

Après avoir déterminé que vous devez résoudre un problème de prévision, la prochaine question à se poser est la suivante : pourquoi résoudre ce problème de prévision ? Dans de nombreux contextes métier, les prévisions ne sont généralement qu'un moyen d'atteindre un objectif. Par exemple, pour la prévision de la demande dans un contexte de vente au détail, la prévision peut servir à prendre des décisions de gestion des stocks. Le problème de prévision est généralement une donnée d'entrée pour un problème de décision, qui peut à son tour être modélisé comme un problème d'optimisation.

Parmi les exemples de tels problèmes de décision, citons le nombre d'unités à acheter ou la meilleure approche pour traiter le stock existant. D'autres problèmes de prévision commerciale incluent la prévision de la capacité des serveurs ou la prévision de la demande de matières premières/pièces dans un contexte de fabrication. Ces prévisions peuvent être utilisées comme entrées pour d'autres processus, soit pour des problèmes de décision comme ci-dessus, soit pour des simulations de scénarios, qui sont ensuite utilisées pour la planification sans modèles explicites. La règle selon laquelle la prévision n'est pas une fin en soi souffre des exceptions. Dans le cas des prévisions financières, par exemple, la prévision est utilisée directement pour constituer des réserves financières ou est présentée aux investisseurs.

Pour comprendre l'objectif des prévisions, posez-vous les questions suivantes :

- Combien de temps dans le futur devez-vous prévoir ?
- À quelle fréquence devez-vous générer des prévisions ?
- Y a-t-il des aspects spécifiques des prévisions que vous devriez approfondir ?

Étude de cas : problème de prévision de la demande de détail pour une entreprise du e-commerce

Pour illustrer plus en détail les concepts de prévision, prenons le cas d'une entreprise du e-commerce qui vend des produits en ligne. L'optimisation des décisions dans la chaîne d'approvisionnement (par exemple, la gestion des stocks) est essentielle à la compétitivité de base de cette entreprise, car elle permet d'avoir le nombre exact de produits dans les lieux d'exécution appropriés. Il s'agit essentiellement de disposer d'une large sélection disponible avec des délais de livraison plus courts et des prix compétitifs, ce qui entraîne une plus grande satisfaction des clients. L'entrée principale dans le système logiciel de la chaîne d'approvisionnement est une prédiction de la demande ou la prévision des ventes potentielles de chaque produit du catalogue. Ces prévisions permettent de prendre d'importantes décisions en aval, dont les principales sont les suivantes :

- Planification au niveau macro (prévisions stratégiques) : pour une entreprise dans son ensemble, quelle est la croissance prévue en termes de ventes/recettes totales ? Où l'entreprise devrait-elle être (plus) active géographiquement ? Comment doit-on organiser la main-d'œuvre ?
- Prévision de la demande (ou des stocks) : combien d'unités de chaque produit sont censées être vendues par site ?
- Activité promotionnelle (prévision stratégique) : comment organiser les promotions ? Les produits doivent-ils être liquidés ?

La suite de l'étude de cas se concentre sur le deuxième problème, qui fait partie de la famille des problèmes de prévision opérationnelle (Januschowski & Kolassa, 2019). Ce document suit les principales préoccupations : données, modèles (prédicteurs), inférences (prévisions) et mise en production.

Pour cette étude de cas, il est important de garder à l'esprit que le problème de la prévision est un moyen de parvenir à une fin. Bien que les prévisions soient d'une importance cruciale pour l'entreprise, les décisions prises en aval de la chaîne d'approvisionnement sont encore plus importantes. Dans notre étude de cas, ces décisions sont prises par des systèmes d'achat automatisés qui s'appuient sur des modèles d'optimisation mathématique issus de la recherche opérationnelle. Ces systèmes visent à minimiser les coûts attendus pour l'entreprise.

Le mot clé est attendu, ce qui signifie que les prévisions doivent couvrir non seulement un avenir possible, mais tous les avenirs possibles, avec une pondération appropriée en fonction de la

probabilité d'un résultat particulier. À cette fin, le facteur clé pour la prise de décision en aval est une distribution complète des valeurs prévisionnelles plutôt qu'une simple prévision ponctuelle. La figure suivante illustre une prévision probabiliste (également appelée prévision par densité). Notez que vous pouvez facilement dériver une prévision ponctuelle (l'avenir le plus probable) de cette prévision probabiliste, mais passer d'une prévision ponctuelle à une prévision probabiliste est plus difficile.

À partir d'une prévision probabiliste, vous pouvez obtenir différentes statistiques et adapter les résultats pour vous aider dans la décision que vous souhaitez prendre. L'entreprise de e-commerce peut avoir un certain nombre de produits clés pour lesquels elle ne veut presque jamais être en rupture de stock. Dans ce cas, utilisez un quantile élevé (par exemple, le 90^{e} percentile), ce qui signifie que les produits seront en stock 90 % du temps. Pour d'autres produits, tels que les produits pour lesquels les remplacements sont plus faciles à trouver (comme les crayons), l'utilisation d'un percentile inférieur peut être plus appropriée.

Dans Amazon Forecast, vous pouvez obtenir facilement différents quantiles à partir de la prévision probabiliste.

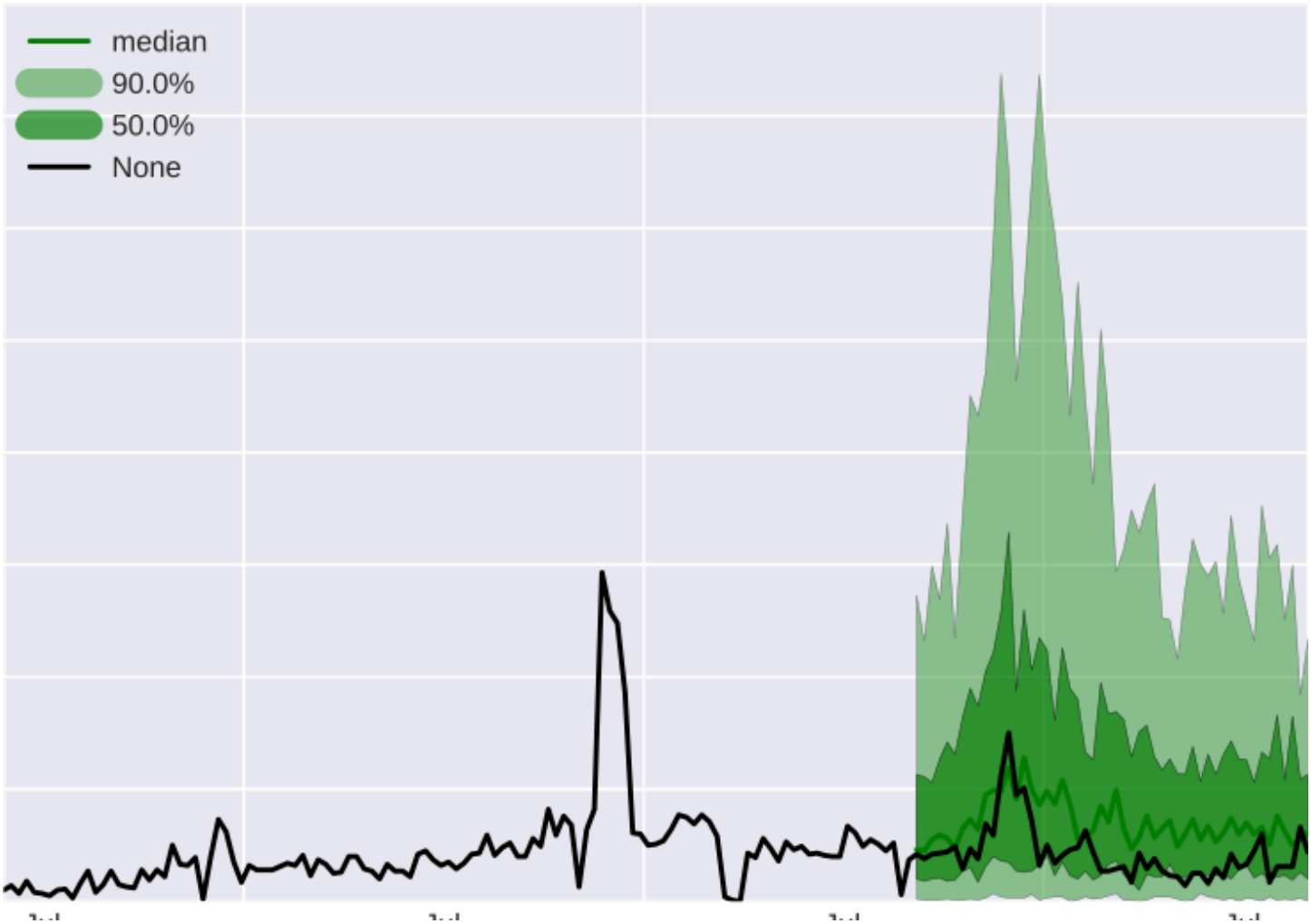


Illustration d'une prévision probabiliste

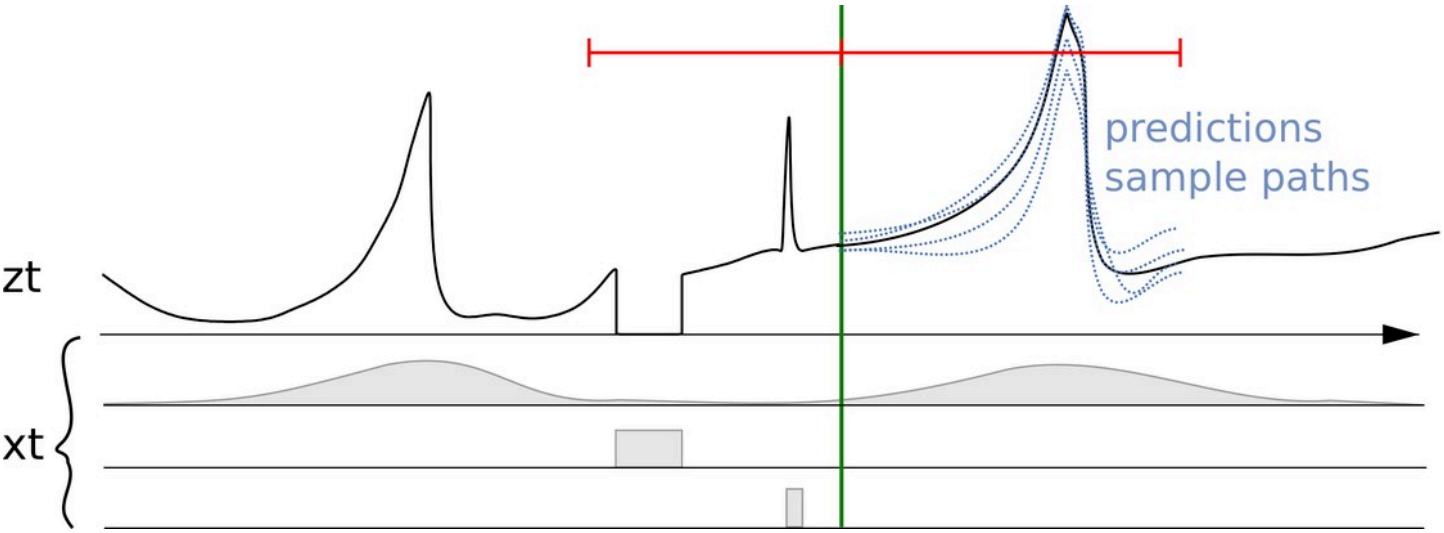
Dans la figure précédente, la ligne noire représente les valeurs réelles ; la ligne vert foncé est la médiane de la distribution des prévisions ; la zone ombrée vert foncé est l'intervalle de prévision dans lequel vous vous attendez à ce que 50 % des valeurs tombent ; et la zone vert clair est l'intervalle de prévision dans lequel vous vous attendez à ce que 90 % des valeurs réelles tombent.

Les sections suivantes couvrent les étapes nécessaires à la résolution du problème de prévision pour cette entreprise, notamment :

- Collecte et agrégation de données (étape 1)
- Préparation des données (étape 2)
- Création d'un prédicteur (étape 3)
- Évaluation des prédicteurs (étape 4)
- Automatiser la génération de prévisions (étape 5)

Étape 1 : Collecter et agréger des données

La figure suivante montre un modèle mental du problème de prévision. L'objectif est de prévoir la série temporelle z_t dans le futur, en utilisant autant d'informations pertinentes que possible pour rendre la prévision aussi précise que possible. Par conséquent, la première étape, et la plus importante, consiste à collecter autant de données correctes que possible.



Une série temporelle z_t avec des caractéristiques ou covariables associées (x_t) et plusieurs prévisions

Dans la figure précédente, plusieurs prévisions sont affichées à droite de la ligne verticale. Ces prévisions sont des échantillons de la distribution de la prévision probabiliste (ou, inversement, peuvent être utilisées pour représenter la prévision probabiliste).

Les informations principales à enregistrer pour un commerce de détail sont les suivantes :

- Les données de vente de la transaction : par exemple, l'unité de gestion des stocks (référence), le lieu, l'horodatage et les unités vendues.
- Données détaillées de la référence : métadonnées d'un article. Par exemple, la couleur, le service, la taille, etc.
- Données sur les prix : séries temporelles de prix de chaque article avec horodatage.
- Données d'information sur les promotions : différents types de promotions, soit sur une collection d'articles (catégorie), soit sur des articles individuels avec horodatage.
- Données d'information sur les stocks : pour chaque unité de temps, l'information indiquant si une référence était en stock ou achetable ou si elle était en rupture de stock.

- Données de localisation : l'emplacement d'un article ou d'une vente à un moment donné peut être représenté par un `location_id` ou un `store_id` de chaîne de caractères, ou par une géolocalisation réelle. Les géolocalisations peuvent être le code du pays plus un code postal à cinq chiffres, ou des coordonnées `latitude_longitude`. L'emplacement est considéré comme une « dimension » des ventes transactionnelles.

Dans Amazon Forecast, les données historiques de la quantité à prévoir sont appelées séries temporelles cibles (TTS). Pour le commerce de détail, la TTS correspond aux données de vente transactionnelles. D'autres données historiques, qui sont connues exactement au même moment que chaque transaction de vente, sont appelées les séries temporelles associées (RTS). Pour le commerce de détail, la RTS comprendrait des variables de prix, de promotion et de stock.

Notez que les informations sur les stocks sont importantes car ce problème est centré sur la demande prévue et non sur les ventes, mais l'entreprise enregistre uniquement les ventes. Lorsqu'une référence est en rupture de stock, le nombre de ventes est inférieur à la demande potentielle. Il est donc important de savoir et d'enregistrer quand de telles ruptures de stock surviennent.

Parmi les autres jeux de données à prendre en compte figurent le nombre de visites de pages Web, les détails sur les termes de recherche, les réseaux sociaux et les informations météorologiques. Il est souvent important de disposer de données pour le passé et pour l'avenir afin de pouvoir les utiliser dans des modèles. C'est une exigence de nombreux modèles de prévision et lors du processus de backtesting (décrit dans la section Étape 4 : Évaluation des prédicteurs).

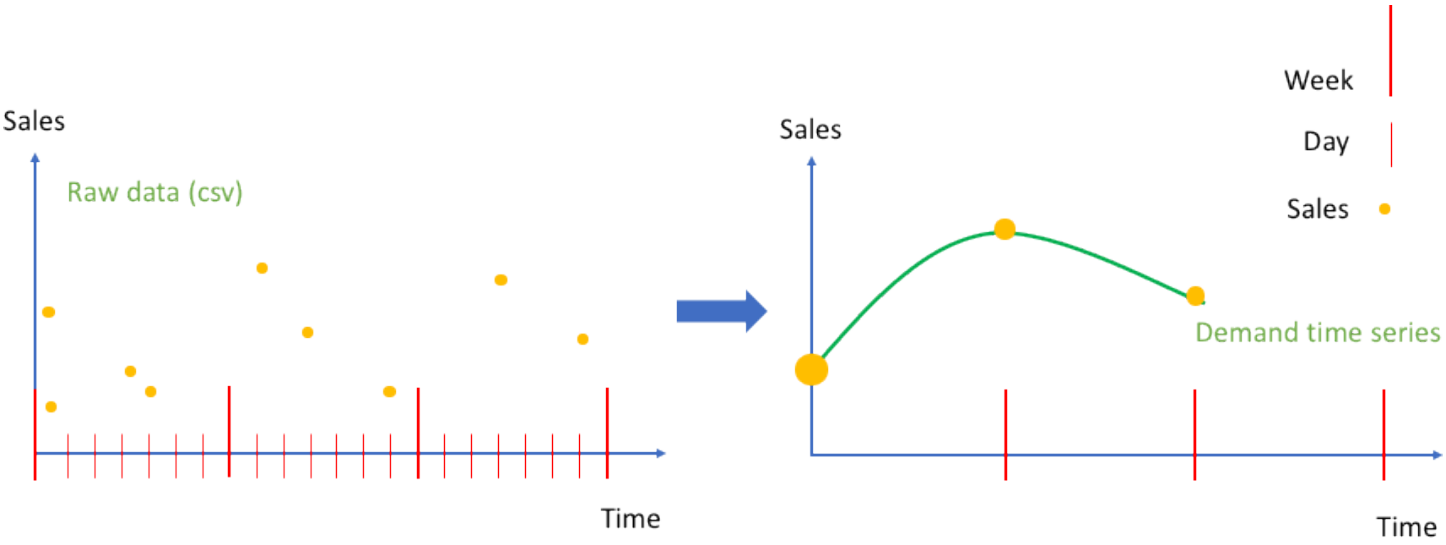
Pour certains problèmes de prévision, la fréquence des données brutes correspond naturellement à celle du problème de prévision. Les exemples incluent la demande du volume du serveur, qui est échantillonné par minute, lorsque vous souhaitez effectuer des prévisions à une fréquence minute par minute.

Les données sont souvent enregistrées à une fréquence plus précise, ou simplement à des horodatages arbitraires à l'intérieur d'une plage de temps, mais le problème de la prévision concerne un niveau de granularité plus grossier. Cela se produit fréquemment dans les études de cas sur le commerce de détail, où les données de vente sont normalement enregistrées sous forme de données transactionnelles ; par exemple, le format consiste en un horodatage avec une précision fine indiquant le moment où les ventes ont eu lieu. Dans le cas d'utilisation des prévisions, ce faible niveau de granularité n'est peut-être pas nécessaire ; il peut être plus approprié d'agréger ces données en ventes horaires ou quotidiennes. Ici, le niveau d'agrégation correspond au problème en aval ; par exemple, la gestion des stocks ou la planification des ressources.

Exemple

Dans la figure suivante, le graphique de gauche présente un exemple de données brutes sur les ventes clients qui peuvent être saisies dans Amazon Forecast sous la forme d'un fichier de valeurs séparées par des virgules (CSV). Dans cet exemple, les données de vente sont définies sur une grille temporelle quotidienne plus précise, et le problème consiste à prévoir la demande hebdomadaire sur la grille temporelle plus large dans le futur. Amazon Forecast procède à l'agrégation des valeurs quotidiennes d'une semaine donnée dans le cadre de l'appel d'API `create_predictor`.

Le résultat transforme les données brutes en un ensemble de séries temporelles bien formées avec une fréquence hebdomadaire fixe. Le graphique de droite illustre cette agrégation sur la série temporelle cible à l'aide de la méthode d'agrégation par somme par défaut. Les autres méthodes d'agrégation incluent la moyenne, le maximum, le minimum ou le choix d'un seul point (par exemple, le premier). Le niveau de granularité et la méthode d'agrégation doivent être choisis de manière à correspondre au mieux à l'utilisation commerciale des données. Dans cet exemple, la valeur agrégée est alignée sur l'agrégation hebdomadaire. L'utilisateur peut définir d'autres méthodes d'agrégation en utilisant la clé `FeaturizationMethodParameters` du paramètre `FeaturizationConfig` de l'API `create_predictor`.



Agrégation des données de ventes brutes sous forme d'événements (à gauche), dans une série temporelle à intervalles réguliers (à droite)

Étape 2 : Préparer les données

Une fois que vous disposez des données brutes, vous devez gérer les complications, telles que les données manquantes, et vous assurer que vous préparez les données pour les modèles de prévision qui rendent le mieux compte de l'interprétation voulue.

Comment gérer les données manquantes

La présence de valeurs manquantes dans les données brutes est un phénomène courant dans les problèmes de prévision réels. Une valeur manquante dans une série temporelle signifie que la vraie valeur correspondante à chaque moment avec la fréquence spécifiée n'est pas disponible pour un traitement ultérieur. Les valeurs peuvent être marquées comme manquantes pour plusieurs raisons.

Les valeurs manquantes peuvent être dues à l'absence de transaction ou à d'éventuelles erreurs de mesure (par exemple, parce qu'un service qui contrôlait certaines données ne fonctionnait pas correctement ou parce que la mesure ne pouvait pas s'effectuer correctement). Le principal exemple de ce dernier point dans l'étude de cas sur le commerce de détail est une rupture de stock dans la prévision de la demande, ce qui signifie que la demande ne correspond pas aux ventes de ce jour-là.

Des effets similaires peuvent survenir dans des scénarios de cloud computing lorsqu'un service a atteint une limite (par exemple, quand les instances Amazon EC2 dans une Région AWS donnée sont toutes occupées). Un autre exemple de valeurs manquantes est celui d'un produit ou d'un service qui n'a pas encore été lancé ou dont la production a cessé.

Les valeurs manquantes peuvent également être insérées par les composants de traitement des caractéristiques, afin d'assurer une longueur égale des séries temporelles avec un remplissage. Si elles sont suffisamment répandues, les valeurs manquantes peuvent avoir un impact significatif sur la précision d'un modèle.

Exemple 1

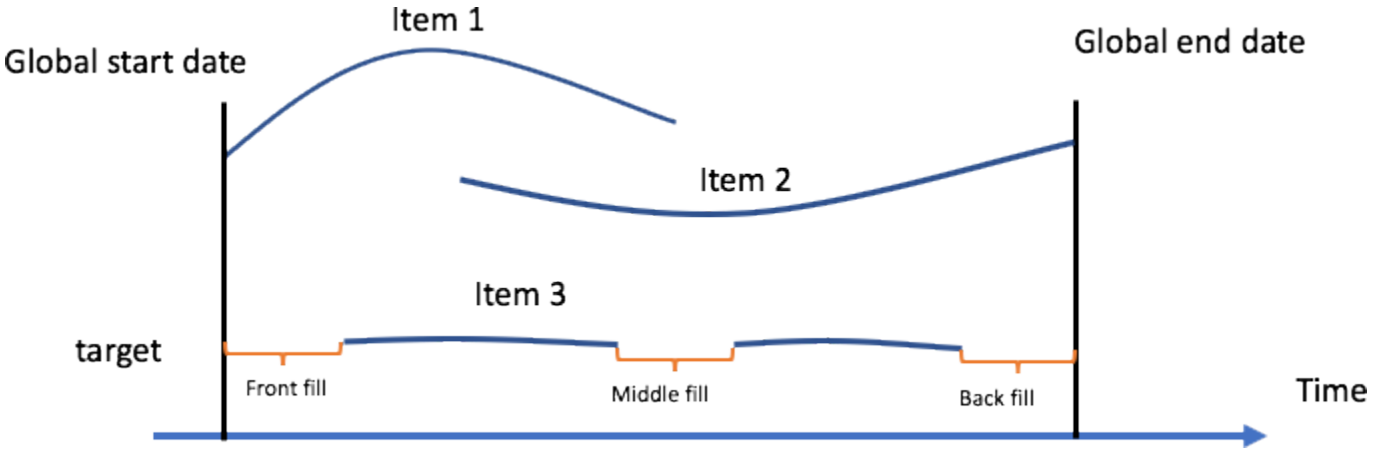
Le remplissage consiste à ajouter des valeurs normalisées aux entrées manquantes dans votre jeu de données. Dans la figure suivante, les différentes stratégies de traitement des valeurs manquantes dans Amazon Forecast — remplissage avant, intermédiaire, en amont et futur — sont illustrées pour l'élément 2 dans un jeu de données de trois éléments.

Amazon Forecast prend en charge le remplissage à la fois pour la série temporelle cible et la série temporelle associée. La date de début globale est définie comme la date de début la plus proche des

dates de début de tous les éléments de votre jeu de données. Dans l'exemple ci-dessous, la date de début globale correspond à l'article 1. De même, la date de fin globale est définie comme la dernière date de fin de la série temporelle pour tous les articles, qui se produit pour l'article 2.

Le remplissage avant complète toutes les valeurs depuis le début de la série temporelle donnée jusqu'à la date de début globale. Au moment de la publication de ce document, Amazon Forecast n'active aucun remplissage avant et permet à toutes les séries temporelles de commencer à des moments différents. Le remplissage intermédiaire indique les valeurs qui ont été renseignées au milieu de la série temporelle (par exemple, entre les dates de début et de fin des éléments), et le remplissage en amont à partir de la dernière date de cette série temporelle jusqu'à la date de fin globale.

Pour la série temporelle cible, les méthodes de remplissage intermédiaire et en amont ont une logique de remplissage par défaut de zéro. Le remplissage futur (qui s'applique uniquement à la série temporelle associée) complète toute valeur manquante entre la date de fin globale des articles et l'horizon de prévision spécifié par le client. Les valeurs futures sont requises pour utiliser le jeu de données de séries temporelles associé avec Prophet et DeepAr+, et facultatives pour CNN-QR.



Stratégies de gestion des valeurs manquantes dans Amazon Forecast

Dans la figure précédente, la date de début globale indique la date de début la plus ancienne parmi les dates de début de tous les articles, et la date de fin globale indique la date de fin la plus récente par rapport aux dates de fin de tous les articles. L'horizon de prévision est la période sur laquelle Forecast fournit des prévisions pour la valeur cible.

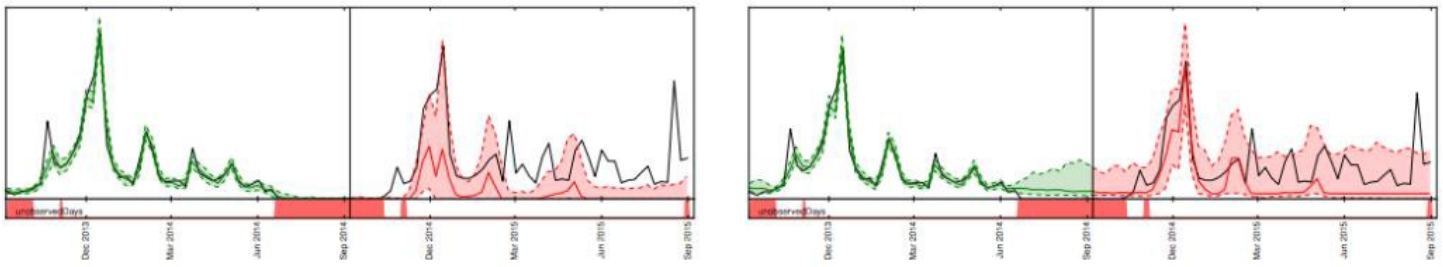
Il s'agit d'un scénario courant dans l'étude sur le commerce de détail qui représente des ventes nulles pour les données transactionnelles relatives aux articles disponibles. Ces valeurs sont traitées comme de vrais zéros et utilisées dans le composant d'évaluation des métriques. Amazon Forecast permet à l'utilisateur d'identifier les valeurs réellement manquantes et de les coder sous forme de

nombres (NaN) à traiter par les algorithmes. Ce document examine ensuite pourquoi ces deux cas diffèrent et quand chacun d'eux est utile.

Dans l'étude de cas sur le commerce de détail, l'information selon laquelle un détaillant a vendu zéro unité d'un article disponible diffère de l'information selon laquelle zéro unité d'un article indisponible est vendue, soit dans les périodes hors de son existence (par exemple, avant son lancement ou après l'arrêt de sa production), soit dans les périodes dans son existence (par exemple, partiellement en rupture de stock, ou lorsqu'il n'y avait pas de données de vente enregistrées pour cette plage de temps). Le remplissage à zéro par défaut est applicable dans ce premier cas. Dans ce dernier cas, même si la valeur cible correspondante est généralement nulle, des informations supplémentaires sont transmises par la valeur marquée comme manquante. La bonne pratique consiste à conserver les informations indiquant qu'il manquait des données et à ne pas les supprimer. Consultez l'exemple suivant pour comprendre pourquoi il est important de conserver les informations.

Amazon Forecast prend en charge des logiques de remplissage supplémentaires basées sur la valeur, la moyenne, la médiane, le minimum et le maximum. Pour les séries temporelles associées (par exemple, prix ou promotion), aucune valeur par défaut n'est spécifiée pour les méthodes de remplissage intermédiaires, en amont ou futures, car la logique des valeurs manquantes correcte varie en fonction du type d'attribut et du cas d'utilisation. La logique de remplissage prise en charge pour les séries temporelles associées inclut le zéro, la valeur, la moyenne, la médiane, le minimum et le maximum.

Pour effectuer le remplissage des valeurs manquantes, spécifiez les types de remplissage à mettre en œuvre lorsque vous appelez l'opération `CreatePredictor`. La logique de remplissage est spécifiée dans les objets `FeaturizationMethod`. Par exemple, pour coder une valeur qui ne représente pas des ventes nulles d'un produit indisponible dans la série temporelle cible, marquez une valeur comme réellement manquante en définissant le type de remplissage égal à NaN. Contrairement au remplissage par zéro, les valeurs codées avec la valeur NaN sont traitées comme réellement manquantes et ne sont pas utilisées dans le composant d'évaluation de la métrique.



L'effet du remplissage par 0 par rapport au remplissage par NaN sur les prévisions pour le même article

Dans la figure précédente, dans le graphique de gauche, les valeurs situées à gauche de la ligne noire verticale sont remplies de 0, ce qui donne lieu à une prévision sous-biaisée (à droite de la ligne noire verticale). Dans le graphique de droite, ces valeurs sont marquées comme NaN, ce qui permet d'obtenir des prévisions appropriées.

Exemple 2

La figure précédente illustre l'importance de gérer correctement les valeurs manquantes pour un modèle spatial à états linéaires, tel qu'ARIMA ou ETS. Celui-ci trace la prévision de la demande pour un article qui est partiellement en rupture de stock. La zone d'entraînement est affichée dans le graphique de gauche en vert, la plage de prédiction dans le panneau de droite en rouge et la véritable cible en noir. Les prévisions médianes, p10 et p90, sont indiquées respectivement sur la ligne rouge et dans la région ombrée. La partie inférieure montre les articles en rupture de stock (80 % des données) marqués en rouge. Dans le graphique de gauche, les zones en rupture de stock sont ignorées et remplies par 0.

Les modèles de prévision supposent donc qu'il y a beaucoup de zéros à prévoir et que, par conséquent, les prévisions sont trop basses. Dans le graphique de droite, les zones en rupture de stock sont traitées comme de véritables observations manquantes, et la demande devient incertaine dans la région en rupture de stock. Les valeurs manquantes pour les articles en rupture de stock étant correctement marquées comme NaN, vous ne constatez aucun sous-biais dans la plage de prévision de ce graphique. Amazon Forecast comble ces lacunes en matière de données, vous permettant ainsi de gérer correctement les données manquantes, sans avoir à modifier explicitement toutes leurs données d'entrée.

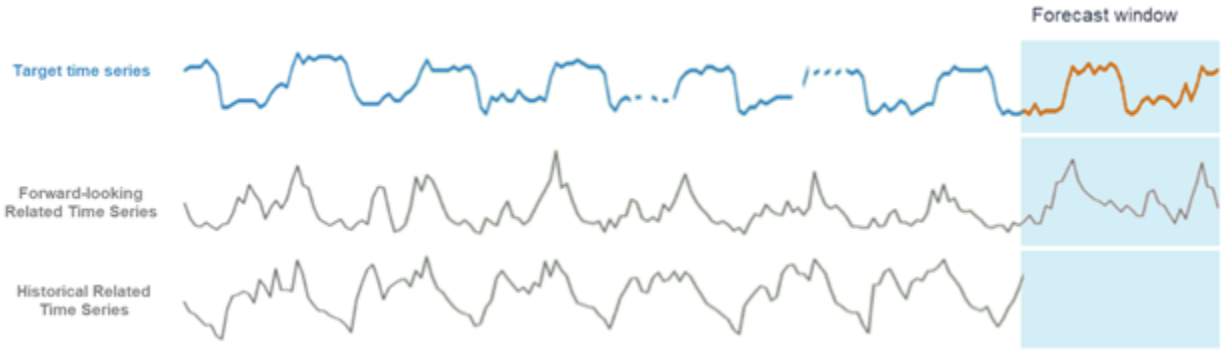
Concepts d'organisation de fonction de séries temporelles associées

Amazon Forecast permet aux utilisateurs de saisir des données associées afin d'améliorer la précision de certains modèles de prévision pris en charge. Ces données peuvent être de deux types : des séries temporelles associées ou des métadonnées d'éléments statiques.

> **Note**
>
> Les métadonnées et les données associées sont appelées fonctions en machine learning, et covariables en statistiques.

Les séries temporelles connexes sont des séries temporelles qui ont une certaine corrélation avec la valeur cible, et qui devraient apporter une certaine force statistique à la prévision sur la valeur cible. En effet, elles fournissent une explication en termes intuitifs. Consultez Amazon Forecast: predicting time-series at scale (Amazon Forecast : prédiction de séries temporelles à grande échelle) pour obtenir un exemple. Contrairement à la série temporelle cible, les séries temporelles associées sont des valeurs connues dans le passé qui peuvent avoir un impact sur la série temporelle cible, et peuvent avoir des valeurs connues dans le futur.

Dans Amazon Forecast, vous pouvez ajouter deux types de séries temporelles associées : des séries temporelles historiques et des séries temporelles prospectives. Les séries temporelles liées à l'historique contiennent des points de données jusqu'à l'horizon de prévision, et ne contiennent pas de points de données dans l'horizon de prévision futur. Les séries temporelles liées à l'avenir contiennent des points de données jusqu'à et dans l'horizon de la prévision.



Différentes approches concernant l'utilisation de séries temporelles associées avec Amazon Forecast

Exemple 3

La figure suivante montre un exemple d'utilisation de séries temporelles associées pour prédire la demande future d'un livre populaire. La ligne bleue représente la demande dans la série temporelle cible. Le prix est indiqué par la ligne verte. La ligne verticale représente la date de début des prévisions, et les prévisions relatives aux deux quantiles sont affichées à droite de la ligne verticale.

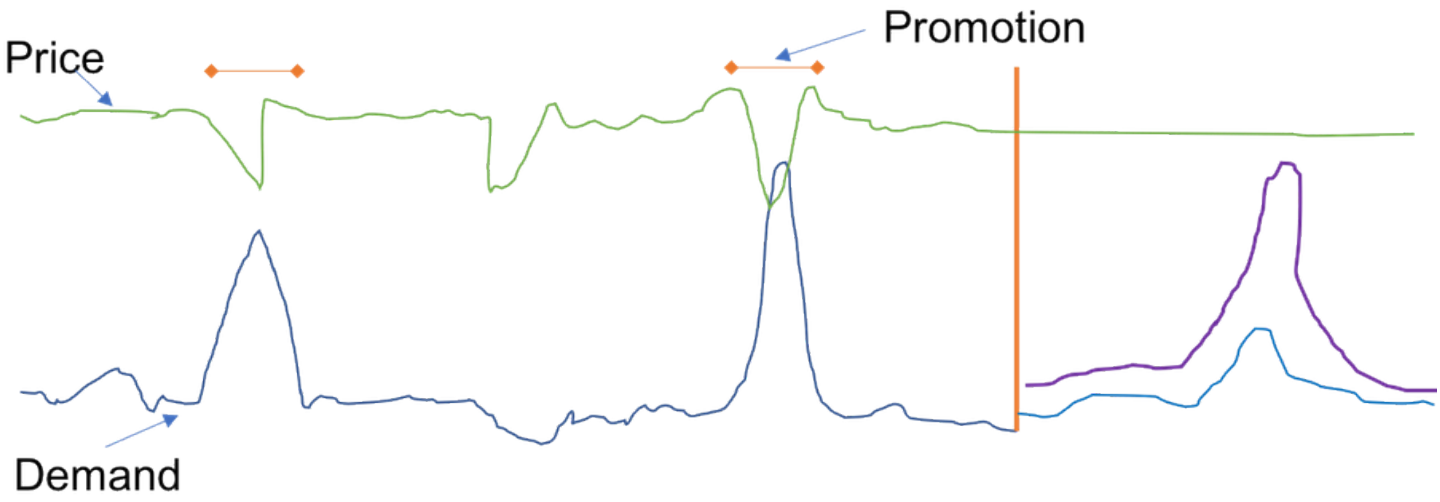
Cet exemple utilise une série temporelle associée prospective qui s'aligne sur la série temporelle cible au niveau de granularité de la prévision, et qui est connue à tout moment (ou la plupart du temps) dans le futur et dans la plage de la date de début de la prévision à la date de début de la prévision incrémentée par l'horizon de la prévision (date de fin de la prévision).

La figure suivante montre également que le prix est une caractéristique appropriée à utiliser, puisque vous pouvez voir des corrélations entre une baisse du prix et une augmentation des ventes du

produit. Les séries temporelles associées peuvent être fournies à Amazon Forecast par le biais d'un fichier CSV distinct, contenant la référence de l'article, l'horodatage et les valeurs des séries temporelles associées (dans ce cas, le prix).

Amazon Forecast prend en charge les méthodes d'agrégation, telles que la moyenne et la somme pour les séries temporelles cibles, mais pas pour les séries temporelles associées. Par exemple, il est peu judicieux d'additionner un prix quotidien et un prix hebdomadaire, et il en va de même pour les promotions quotidiennes.

Amazon Forecast peut intégrer automatiquement des informations sur la météo et les vacances dans un modèle en incluant des jeux de données de fonctionnalités intégrés (consultez `SupplementaryFeature`). Les informations météorologiques et les jours fériés peuvent avoir une incidence significative sur la demande du commerce de détail.



Les ventes d'un article donné (en bleu, à gauche de la ligne rouge verticale)

Les métadonnées des articles, également connues sous le nom de variables catégoriques, sont d'autres caractéristiques utiles qui peuvent être introduites dans Amazon Forecast. Consultez Amazon Forecast: predicting time-series at scale (Amazon Forecast : prédiction de séries temporelles à grande échelle) pour obtenir un exemple). La principale différence entre les variables catégoriques et les séries temporelles associées est que les variables catégoriques sont statiques : elles n'évoluent pas dans le temps. Parmi les exemples courants dans le secteur du commerce de détail, citons les couleurs des articles, les catégories de livres et les indicateurs binaires indiquant si un téléviseur est un téléviseur intelligent ou non. Ces informations peuvent être récupérées par les algorithmes de deep learning pour apprendre les similitudes entre les unités de gestion des stocks (références), en supposant que des références similaires affichent des ventes similaires. Comme ces métadonnées n'ont pas de dépendance temporelle, chaque ligne du fichier CSV de métadonnées

d'articles ne comprend que la référence de l'article et l'étiquette ou la description de la catégorie correspondante.

Étape 3 : Créer un prédicteur

Un prédicteur peut être créé de deux manières : en exécutant AutoML ou en sélectionnant manuellement l'un des six algorithmes intégrés d'Amazon Forecast. Lors de l'exécution d'AutoML, au moment de la rédaction de ce document, Amazon Forecast teste automatiquement les six algorithmes intégrés et choisit celui qui présente les plus faibles pertes moyennes sur les quantiles 10, 50 (médiane) et 90.

Amazon Forecast propose quatre modèles locaux :

- Moyenne mobile intégrée autorégressive (ARIMA)
- Lissage exponentiel (ETS)
- Séries temporelles non paramétriques (NPTS)
- Prophet

Les modèles locaux sont des méthodes de prévision qui adaptent un modèle unique à chaque série temporelle individuelle (ou à une combinaison article/dimension spécifique), puis utilisent ces modèles pour extrapoler les séries temporelles dans le futur.

ARIMA et ETS sont des versions évolutives de modèles locaux populaires issus du package de prévisions R. NPTS, une méthode locale développée par Amazon, présente une différence majeure par rapport aux autres modèles locaux. Contrairement aux simples prévisionnistes saisonniers, qui fournissent des prévisions ponctuelles en répétant la dernière valeur ou la valeur à une saisonnalité appropriée, NPTS produit des prévisions probabilistes. NPTS utilise un indice temporel fixe, où l'indice précédent (T - 1) ou la saison passée (T - tau) représente la prédiction pour l'intervalle de temps T. L'algorithme échantillonne aléatoirement un indice temporel (t) dans l'ensemble {0, ..., T - 1} pour générer un échantillon pour l'intervalle de temps actuel T. NPTS est particulièrement efficace pour les séries temporelles intermittentes (parfois aussi appelées « sparse ») comportant de nombreux zéros. Forecast comprend également la mise en œuvre en Python de Prophet, un modèle structurel bayésien de séries temporelles.

Amazon Forecast propose deux algorithmes globaux de deep learning :

- DeepAR+
- CNN-QR

Les modèles globaux entraînent un seul modèle sur l'ensemble des séries temporelles d'un jeu de données. Ceci est particulièrement utile lorsqu'il existe des séries temporelles similaires dans un ensemble d'unités transversales. Par exemple, des regroupements de séries temporelles de la demande de différents produits, des charges de serveurs et des requêtes de pages web.

En général, plus le nombre de séries temporelles augmente, plus l'efficacité de CNN-QR et de DeepAR+ augmente. Ce n'est pas toujours le cas pour les modèles locaux. Les modèles de deep learning peuvent également générer des prévisions pour de nouvelles références avec peu ou pas de données historiques sur les ventes. C'est ce que l'on appelle les prévisions de démarrage à froid.

	Neural Networks		Flexible Local Algorithms	Baseline Algorithms		
	CNN-QR	DeepAR+	Prophet	NPTS	ARIMA	ETS
Computationally intensive training process	High	High	Medium	Low	Low	Low
Accepts historical related time series*	✓	✗	✗	✗	✗	✗
Accepts forward-looking related time series*	✓	✓	✓	✗	✗	✗
Accepts item metadata (product color, brand, etc)	✓	✓	✗	✗	✗	✗
Suitable for sparse datasets	✓	✓	✗	✓	✗	✗
Performs Hyperparameter Optimization (HPO)	✓	✓	✗	✗	✗	✗
Allows overriding default hyperparameter values	✓	✓	✗	✓	✗	✗
Suitable for What-if analysis	✓	✓	✓	✗	✗	✗
Suitable for Cold Start scenarios (forecasting with little to no historical data)	✓	✓	✗	✗	✗	✗

Comparez les algorithmes disponibles dans Amazon Forecast

Pour plus d'informations sur les séries temporelles associées, consultez Related Time Series (Série temporelle associée).

Étape 4 : Évaluer les prédicteurs

Un flux de travail classique dans le domaine du machine learning consiste à entraîner un ensemble de modèles ou une combinaison de modèles sur un ensemble d'apprentissage et à évaluer leur précision sur un jeu de données d'attente. Cette section traite de la manière de fractionner les données historiques et des métriques à utiliser pour évaluer les modèles de prévision des séries temporelles. Pour les prévisions, la technique du backtesting est le principal outil pour évaluer la précision des prévisions.

Backtesting

Un cadre d'évaluation et de backtesting approprié est l'un des facteurs les plus importants pour faire d'une application de machine learning un succès. Vous pouvez vous appuyer sur des backtests réussis avec vos modèles pour gagner en confiance sur le pouvoir prédictif futur des modèles. En outre, vous pouvez ajuster les modèles via l'optimisation des hyperparamètres (HPO), apprendre des combinaisons de modèles et activer le méta-apprentissage et AutoML.

La série temporelle prédisant le temps caractéristique la différencie, en termes de méthodologie d'évaluation et de backtesting, des autres domaines du machine learning appliqué. Habituellement, dans les tâches de ML, pour évaluer l'erreur prédictive dans un backtest, vous divisez un jeu de données par éléments. Par exemple, pour la validation croisée dans les tâches liées aux images, vous vous entraînez sur un certain pourcentage des images, puis vous utilisez d'autres parties pour le test et la validation. Dans le domaine de la prévision, il est nécessaire de procéder à un fractionnement essentiellement temporel (et, dans une moindre mesure, par éléments) afin de s'assurer que les informations de l'ensemble d'entraînement ne se retrouvent pas dans l'ensemble de test ou de validation, et que vous simulez le cas de production aussi fidèlement que possible.

Le fractionnement dans le temps doit être effectué avec soin car vous ne voulez pas choisir un seul point dans le temps, mais plusieurs points. Sinon, la précision est trop dépendante de la date de début de la prévision, telle que définie par le point de séparation. Une évaluation des prévisions en continu, dans laquelle vous effectuez une série de fractionnements sur plusieurs points dans le temps et produisez le résultat moyen, donne des résultats de backtest plus solides et plus fiables. La figure suivante illustre quatre fractionnements de backtest différents.

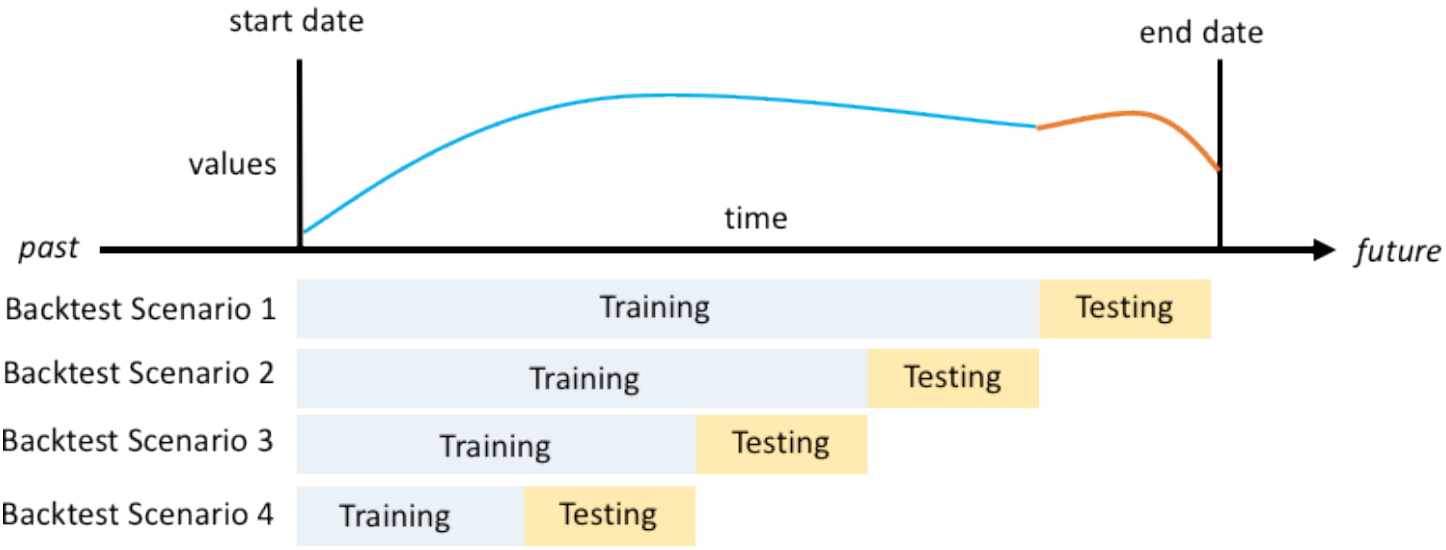


Illustration de quatre scénarios de backtesting différents avec une taille croissante de l'ensemble d'entraînement, mais une taille constante de l'ensemble de test

Dans la figure précédente, tous les scénarios de backtesting disposent de données disponibles dans leur intégralité pour pouvoir évaluer les valeurs prévues par rapport aux valeurs réelles.

La raison pour laquelle plusieurs fenêtres de backtest sont nécessaires est que la plupart des séries temporelles dans le monde réel sont normalement non stationnaires. L'entreprise de e-commerce de l'étude de cas est basée en Amérique du Nord et une grande partie de sa demande de produits est déterminée par le pic du quatrième trimestre, avec des pics particuliers autour de Thanksgiving et avant Noël. Pendant la saison des achats du quatrième trimestre, la variabilité de la série temporelle est plus élevée que pendant le reste de l'année. En disposant de plusieurs fenêtres de backtest, vous pouvez évaluer les modèles de prévision dans un cadre plus équilibré.

Pour chaque scénario de backtest, la figure suivante montre les éléments de base dans la terminologie d'Amazon Forecast. Amazon Forecast fractionne automatiquement les données en jeux de données d'entraînement et de test. Amazon Forecast décide comment fractionner les données d'entrée en utilisant le paramètre `BackTestWindowOffset` spécifié en tant que paramètre dans l'API `create_predictor` ou en utilisant sa valeur par défaut de `ForecastHorizon`.

Dans la figure suivante, vous voyez le premier cas, plus général, où les paramètres `BackTestWindowOffset` et `ForecastHorizon` ne sont pas égaux. Le paramètre `BackTestWindowOffset` définit une date de début de prévision virtuelle, représentée par la ligne verticale en pointillés dans la figure suivante. Il peut être utilisé pour répondre à la question hypothétique suivante : si le modèle était déployé ce jour-là, quelle serait la prévision ? Le paramètre

`ForecastHorizon` définit le nombre d'intervalles de temps à partir de la date de début de la prévision virtuelle pour effectuer la prédiction.

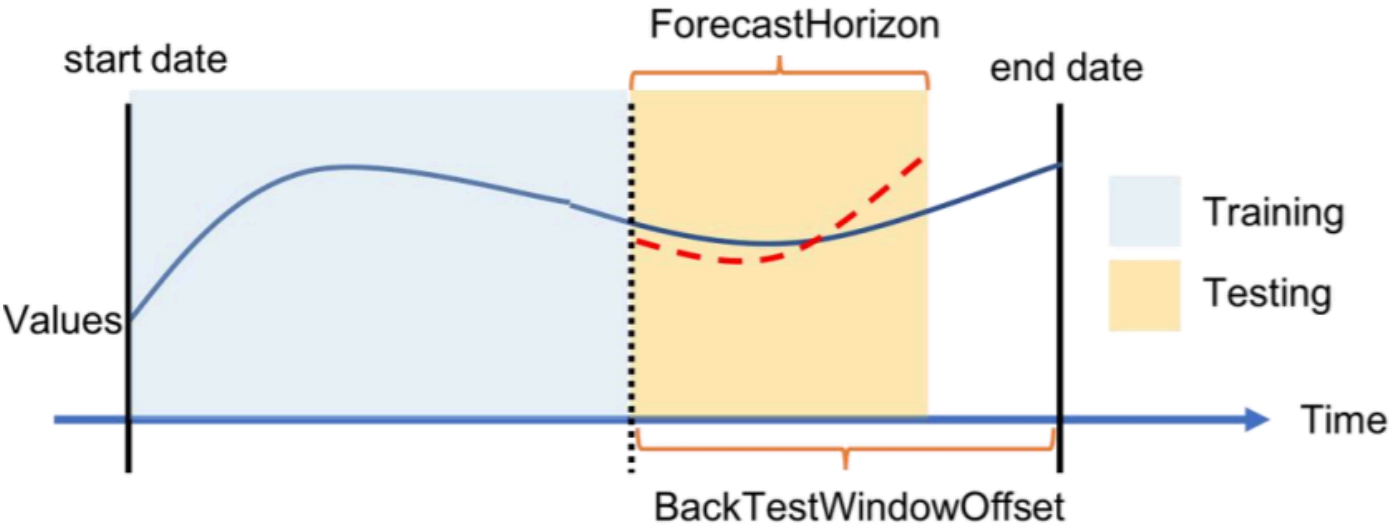


Illustration d'un scénario de backtest unique et de sa configuration dans Amazon Forecast

Amazon Forecast peut exporter les valeurs prévisionnelles et les mesures de précision générées pendant le backtesting. Les données exportées peuvent être utilisées pour évaluer des éléments spécifiques à des points de temps et des quantiles spécifiques.

Quantiles de prédiction et métriques de précision

Les quantiles de prévision peuvent fournir une limite supérieure et inférieure pour les prévisions. Par exemple, l'utilisation des types de prévision 0,1 (P10), 0,5 (P50) et 0,9 (P90) fournit une plage de valeurs connue sous le nom d'intervalle de confiance de 80 % autour de la prévision P50. En générant des prévisions à P10, P50 et P90, vous pouvez vous attendre à ce que la valeur réelle se situe entre ces limites 80 % du temps.

Ce document traite plus en détail des quantiles à l'étape 5.

Amazon Forecast utilise les mesures de précision de la perte de quantile pondérée (wQL), de la racine carrée de l'erreur quadratique moyenne (RMSE) et du pourcentage d'erreur absolu pondéré (WAPE) pour évaluer les prédicteurs pendant le backtesting.

Perte de quantile pondérée (wQL)

La métrique d'erreur de perte de quantile pondérée (wQL) mesure la précision de la prévision d'un modèle à un quantile spécifié. Elle est particulièrement utile lorsqu'il existe des coûts différents pour

la sous-estimation et la surestimation. La définition de la pondération (τ) de la fonction wQL intègre automatiquement différentes pénalités en cas de sous-estimation et de surestimation.

$$\mathrm{wQL}[\tau] = 2\frac{\sum_{i,t}[\tau \max(y_{i,t} - q_{i,t}^{(\tau)}, 0) + (1-\tau)\max(q_{i,t}^{(\tau)} - y_{i,t}, 0)]}{\sum_{i,t}|y_{i,t}|}$$

Fonction wQL

Où :

- τ — Un quantile dans l'ensemble {0,01, 0,02,..., 0,99}
- qi,t(τ) — Le quantile τ prédit par le modèle.
- yi,t — La valeur observée au point (i,t)

Pourcentage d'erreur absolu pondéré (WAPE)

Le pourcentage d'erreur absolu pondéré (WAPE) est une métrique couramment utilisée pour mesurer la précision des modèles. Il mesure l'écart global des valeurs prévues par rapport aux valeurs observées.

$$\mathrm{WAPE} = \frac{\sum_{i,t}|y_{i,t} - \hat{y}_{i,t}|}{\sum_{i,t}|y_{i,t}|}$$

WAPE

Où :

- yi,t : la valeur observée au point (i,t)
- ŷi,t : la valeur prédite au point (i,t)

Forecast utilise la prévision moyenne comme valeur prédite, ŷi,t.

Racine carrée de l'erreur quadratique moyenne (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{nT}\sum_{i,t}(\hat{y}_{i,t} - y_{i,t})^2}$$

La racine carrée de l'erreur quadratique moyenne (RMSE) est une métrique couramment utilisée pour mesurer la précision des modèles. Comme l'équation WAPE, elle mesure l'écart global des estimations par rapport aux valeurs observées.

Où :

- yi,t : la valeur observée au point (i,t)
- ŷi,t : la valeur prédite au point (i,t)
- nT : le nombre de points de données dans un ensemble de test

Forecast utilise la prévision moyenne comme valeur prédite, ŷi,t. Lors du calcul des métriques de prédiction, nT désigne le nombre de points de données dans une fenêtre de backtest.

Problèmes avec WAPE et RMSE

Dans la plupart des cas, les prévisions ponctuelles qui peuvent être générées en interne ou par d'autres outils de prévision devraient correspondre aux prévisions du quantile ou de la moyenne p50. Pour WAPE et RMSE, Amazon Forecast utilise la prévision moyenne pour représenter la valeur prédite (yhat).

Pour tau = 0,5 dans l'équation wQL[tau], les deux pondérations sont égales, et le wQL[0.5] se réduit au pourcentage d'erreur absolu pondéré (WAPE) couramment utilisé pour les prévisions ponctuelles :

$$\mathrm{wQL}[0.5] = 2\frac{\sum_{i,t} 0.5[\max(y_{i,t} - q_{i,t}^{(0.5)}, 0) + \max(q_{i,t}^{(0.5)} - y_{i,t}), 0]}{\sum_{i,t} |y_{i,t}|} = \frac{\sum_{i,t} |y_{i,t} - q_{i,t}^{(0.5)}|}{\sum_{i,t} |y_{i,t}|}$$

où yhat = q(0.5) est la prévision calculée. Un facteur d'échelle de 2 est utilisé dans la formule wQL pour annuler le facteur 0,5 afin d'obtenir l'expression exacte de WAPE [médiane].

Notez que la définition ci-dessus de l'équation WAPE diffère de l'interprétation courante du pourcentage d'erreur absolu pondéré (MAPE). La différence réside dans le dénominateur. La manière dont l'équation WAPE est définie ci-dessus évite le problème de la division par 0, un problème courant dans les scénarios du monde réel tels que le e-commerce dans l'étude de cas, qui vendra souvent 0 unité d'une référence donnée un jour donné.

Contrairement à la métrique de perte par quantile pondéré pour tau non égal à 0,5, le biais inhérent à chaque quantile ne peut pas être pris en compte par un calcul comme l'équation WAPE, où les

pondérations sont égales. Parmi les autres inconvénients de l'équation WAPE, on peut citer le fait qu'elle n'est pas symétrique, qu'elle présente une surinflation des erreurs en pourcentage pour les petits nombres et qu'elle n'est qu'une métrique ponctuelle.

La RMSE est le carré du terme d'erreur dans l'équation WAPE et une mesure d'erreur commune dans d'autres applications ML. La métrique RMSE favorise un modèle où les erreurs individuelles sont d'une magnitude constante, car de grandes variations dans l'erreur augmenteront la RMSE de façon disproportionnée. En raison de l'erreur quadratique, quelques valeurs mal prédites dans une prévision autrement bonne peuvent augmenter le RMSE. De plus, en raison des termes au carré, les termes d'erreur plus petits ont moins de poids dans la RMSE que dans la WAPE.

Les métriques de précision permettent une évaluation quantitative des prévisions. Elles sont cruciales, notamment pour les comparaisons à grande échelle (la méthode A est-elle globalement meilleure que la méthode B). Cependant, il est souvent important de compléter ces éléments par des visuels pour les différentes références.

Étape 5 : Générer et utiliser les prévisions pour la prise de décision

Une fois que vous disposez d'un modèle qui répond au seuil de précision requis pour votre cas d'utilisation spécifique (tel que déterminé par le backtesting), l'étape finale consiste à déployer le modèle et à générer des prévisions. Pour déployer un modèle dans Amazon Forecast, vous devez exécuter l'API Create_Forecast. Cette action héberge un modèle créé en s'entraînant sur l'ensemble du jeu de données historique (contrairement à `Create_Predictor`, qui divise les données en un ensemble d'entraînement et un ensemble de test). Les prédictions du modèle générées sur l'horizon de prévision peuvent alors être consommées de deux manières :

- Vous pouvez interroger les prévisions pour un élément particulier (en spécifiant l'élément ou la combinaison élément/dimension) en utilisant l'API `Query_Forecast` depuis la AWS CLI ou directement via la AWS Management Console.
- Vous pouvez générer les prévisions pour toutes les combinaisons d'éléments et de dimensions pour tous les quantiles à l'aide de l'API `Create_Forecast_Export_Job`. Cette API génère un fichier CSV qui est stocké en toute sécurité dans un emplacement Amazon Simple Storage Service (Amazon S3) de votre choix. Vous pouvez ensuite utiliser les données du fichier CSV et les insérer dans vos systèmes en aval utilisés pour la prise de décision. Par exemple, vos systèmes de chaîne d'approvisionnement existants peuvent intégrer directement les résultats d'Amazon Forecast afin de faciliter la prise de décision concernant la fabrication de certains articles.

Prévisions probabilistes

Amazon Forecast peut générer des prévisions à différents quantiles, ce qui est particulièrement utile lorsque les coûts de sous-estimation et de surestimation sont différents. Comme pour l'étape d'entraînement du prédicteur, des prévisions probabilistes peuvent être générées pour des quantiles compris entre p1 et p99.

Par défaut, Amazon Forecast génère des prévisions aux mêmes quantiles que ceux utilisés lors de l'entraînement du prédicteur. Si les quantiles ne sont pas spécifiés lors de l'entraînement du prédicteur, les prévisions seront générées à p10, p50 et p90 par défaut.

Pour la prévision p10, on s'attend à ce que la valeur réelle soit inférieure à la valeur prédite dans 10 % des cas, et la métrique wQL[0,1] peut être utilisée pour évaluer sa précision. Cela signifie que la

prévision P10 est une sous-prévision dans 90 % des cas, et que si elle était utilisée pour stocker les stocks, l'article serait épuisé dans 90 % des cas. La prévision P10 pourrait être utile lorsque l'espace de stockage est limité ou que le coût du capital investi est élevé.

> **Note**
>
> La définition formelle d'une prévision quantile est Pr(valeur réelle <= prévision au quantile q) = q. Techniquement, un quantile est un percentile/100. Les statisticiens ont tendance à dire « niveau du quantile P90 », car c'est plus facile à dire que « quantile 0,9 ». Par exemple, une prévision au niveau du quantile P90 signifie que l'on peut s'attendre à ce que la valeur réelle soit inférieure à la prévision dans 90 % des cas. Plus précisément, si au temps = t1 et au niveau de quantile = 0,9, la valeur prédite = 30, cela signifie que la valeur réelle au temps = t1, si vous avez 1 000 simulations, devrait être inférieure à 30 pour 900 simulations. Pour 100 simulations, la valeur réelle devrait être supérieure à 30.

D'autre part, la prévision P90 est une sur-prévision dans 90 % des cas, et elle est utile lorsque le coût d'opportunité de ne pas vendre un article est extrêmement élevé, ou lorsque le coût du capital investi est faible. Pour une épicerie, la prévision P90 pourrait être utilisée pour des produits comme le lait ou le papier toilette, pour lesquels le magasin ne veut jamais être à court et ne voit pas d'inconvénient à ce qu'il en reste toujours dans les rayons.

Pour la prévision p50 (souvent également appelée prévision médiane), on s'attend à ce que la valeur réelle soit inférieure à la valeur prévue dans 50 % des cas, et la métrique wQL[0.5] peut être utilisée pour évaluer sa précision. Lorsque le surstockage n'est pas trop préoccupant et que la demande pour un article donné est modérée, la prévision par quantile p50 peut être utile.

Visualisation

Amazon Forecast permet de tracer des prévisions de manière native dans la AWS Management Console. En outre, vous pouvez tirer parti de la pile complète de science des données Python. Consultez pour cela Amazon Forecast Examples (Exemples Amazon Forecast). Amazon Forecast permet d'exporter les prévisions sous forme de fichier CSV via l'API `ExportForecastJob`, ce qui permet aux utilisateurs de visualiser les prévisions dans l'outil analytique de leur choix.

Visualisation fournie dans la console Amazon Forecast pour différents quantiles

Résumé du flux de travail et des API de prévision

Le tableau suivant fait correspondre chaque étape du flux de travail de prévision avec l'API Amazon Forecast correspondante.

Tableau 1 : Étapes de la prévision et API d'Amazon Forecast

Étape	API	Fonctions de l'API
Étape 1 : Collecter et agréger des données Étape 2 : Préparer les données	`Create_Dataset_Group` , `Create_Dataset` , `Create_Dataset_Import_Job`	1. Définissez le domaine avancé (vente au détail, métriques, etc.) du problème. 2. Définissez le schéma des différents jeux de données (cible, associé, métadonnées de l'élément). 3. Importez des données d'Amazon S3 vers Amazon Forecast.
Étape 3 : Créer un prédicteur Étape 4 : Évaluer les prédicteurs	`Create_Predictor`	1. Exécute l'ETL. 2. Divise les données en ensembles d'entraînement et de test et entraîne le modèle. 3. Vous pouvez également utiliser `Create_predictor_backtest_Export_job` pour exporter les résultats du backtest en CSV pour calculer les métriques au niveau des éléments.

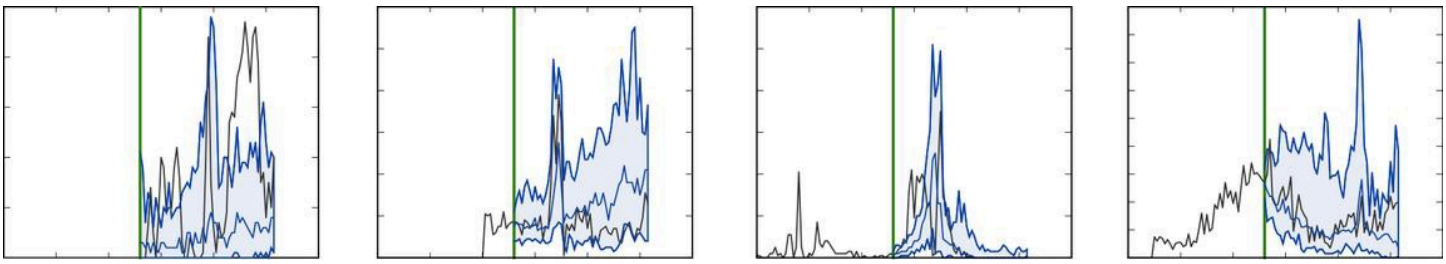
Étape	API	Fonctions de l'API
Étape 5 : Générer et utiliser les prévisions pour la prise de décision	`Create_Forecast`	1. Entraîne/héberge le modèle. 2. Génère des prédictions sur l'horizon de prévision pour un quantile spécifiqu e d'intérêt (par exemple, tout nombre entier compris entre 1 et 99, y compris la moyenne).
	`Query_Forecast` . `Create_Forecast_Ex port_Job`	Vous permet de consommer les prévisions créées par `Create_Forecast`

Utilisation d'Amazon Forecast pour des scénarios courants

Vous pouvez également effectuer des analyses de type « et si » en générant différentes prévisions en fonction de l'évolution de variables externes (par exemple, les prix ou les promotions). Par exemple, dans l'exemple d'étude de cas sur le e-commerce, vous pouvez créer différentes prévisions en fonction des promotions que vous prévoyez. Vous pouvez prévoir la demande d'un produit avec une remise de 10 %, puis de 20 %, afin de comprendre la quantité de produit que vous devrez stocker pour répondre à la demande. Pour ce faire, il suffit de créer des groupes de jeux de données uniques et de mettre à jour les séries temporelles correspondantes dans chacun d'eux, en fonction du scénario choisi.

En outre, vous pouvez également générer des prévisions pour des éléments sans historique (parfois appelé le problème du démarrage à froid). Cette approche nécessite la création d'un prédicteur utilisant DeepAR+ ou CNN-QR avec des métadonnées (telles qu'un jeu de données de métadonnées d'articles) pour générer des prévisions pour le nouvel article.

La figure suivante présente des exemples de quatre références différentes telles qu'elles apparaissent dans des problèmes réels de prévision opérationnelle.



Exemples de quatre références différentes telles qu'elles apparaissent dans des problèmes réels de prévision opérationnelle.

Dans l'image précédente, les valeurs réelles observées sont à gauche de la ligne verticale, et les prévisions en bleu sont à droite de la ligne verticale, comparées aux valeurs réelles en noir. Notez que l'historique de chaque référence individuelle, à gauche de la ligne verticale, n'est pas indicatif de son évolution à droite de la ligne verte.

Mise en œuvre des prévisions dans la production

Après avoir réalisé le flux de travail de bout en bout d'Amazon Forecast, il est essentiel d'identifier les principales différences entre les API `Create_Predictor` et `Create_Forecast` et de savoir quand chacune doit être utilisée.

La première est utilisée principalement lors de la validation du concept pour évaluer la précision/les métriques du modèle, tandis que la seconde est utilisée pour générer des prévisions dans un environnement de production.

Une fois en production, `Create_Predictor` ne doit pas être exécuté à chaque fois qu'une prévision doit être générée, mais seulement lorsque le modèle doit être ré-entraîné en raison de changements dans les données ou dans le cadre d'une fréquence préétablie (par exemple, toutes les deux semaines ou tous les mois). Comme les jeux de données sont mis à jour avec de nouvelles données, seul `Create_Forecast` doit être exécuté pour générer des prévisions pour le nouvel horizon de prévision.

En production, vous devez également automatiser vos importations de jeux de données et vos opérations de prévision afin de générer de nouvelles prévisions sur une base continue. Aujourd'hui, vous pouvez y parvenir en configurant des tâches cron à l'aide d'une combinaison de journaux Amazon CloudWatch Events, AWS Step Functions et de fonctions AWS Lambda. La configuration de la tâche cron automatise à son tour les appels à l'API Amazon Forecast pour l'importation/l'entraînement ou la génération de prévisions. Enfin, il est essentiel de gérer vos ressources et de les supprimer à intervalles réguliers afin de ne pas dépasser les limites du système prescrites par

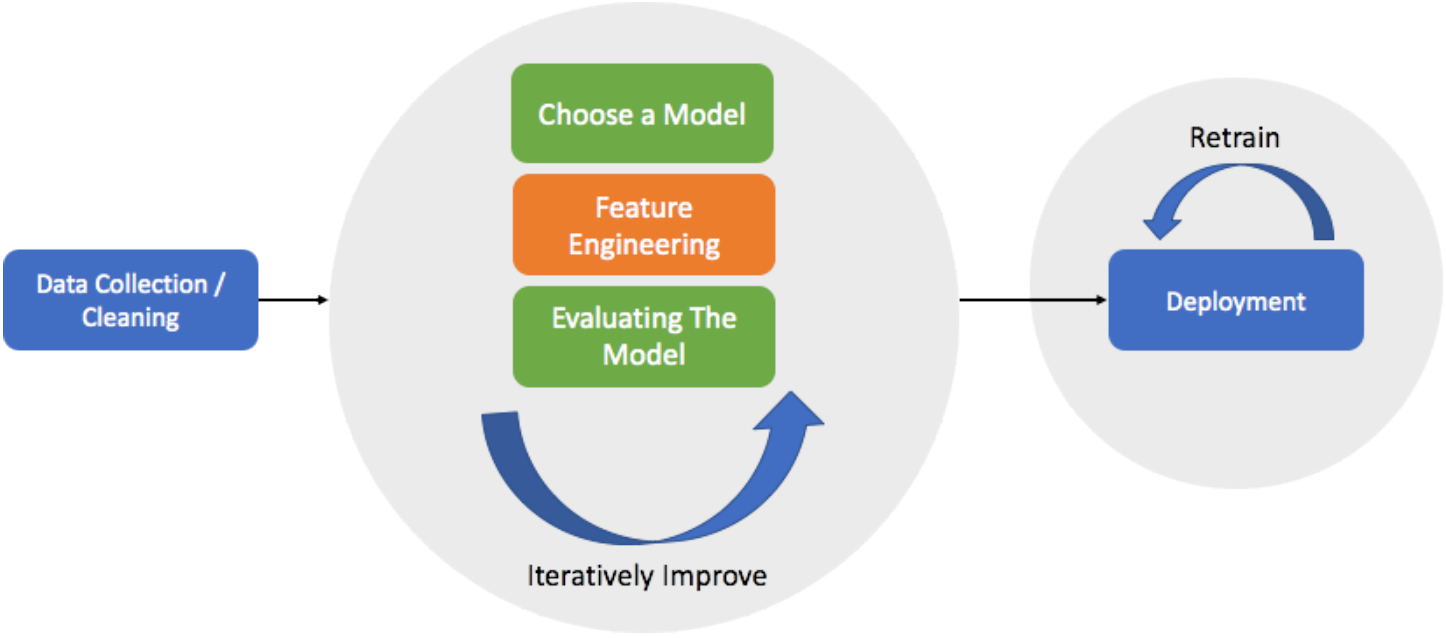
le service. Reportez-vous à cet article de blog incluant Amazon Redshift pour en savoir plus sur la configuration des tâches planifiées.

Conclusion

Januschowski et Kolassa (2019) proposent une classification des problèmes de prévision alignés sur les décisions que les entreprises doivent prendre, notamment les décisions stratégiques, tactiques, et opérationnelles. À chaque niveau de décision correspondent des tâches de prévision.

Les problèmes de prévision opérationnelle et tactique se caractérisent par le fait qu'ils contiennent de grandes quantités de données et nécessitent généralement un degré élevé d'automatisation. Différentes méthodes de prévision répondent à ces problèmes. Les méthodes de prévision locales fonctionnent généralement bien pour les problèmes de prévision stratégique, les méthodes basées sur le deep learning pour les problèmes de prévision opérationnelle, et pour les cas intermédiaires, il peut être nécessaire d'expérimenter un peu. Bien que cet article ait abordé les problèmes de prévision opérationnelle, Amazon Forecast n'a pas d'opinion arrêtée sur les modèles qu'il propose et inclut des modèles qui répondent aux problèmes de prévision stratégique, opérationnelle et tactique.

Le processus de résolution des problèmes de prévision opérationnelle peut être décomposé en étapes de base allant de la collecte et de la préparation des données à la construction et au déploiement du modèle. En général, il est plus utile de considérer cela comme un processus itératif plutôt que linéaire. Par exemple, lorsque les modèles et les cas d'utilisation sont mieux compris, il peut être utile de revenir à la phase de collecte des données. Le développement de modèles est également très itératif en soi.



Processus de développement simplifié pour la mise en production d'un modèle de prévision.

Participants

Ont contribué à la préparation du présent document :

- Yuyang Wang, chercheur principal en machine learning, Services verticaux d'IA
- Danielle Robinson, chercheuse appliquée, Services verticaux d'IA
- Tim Januschowski, directeur, Sciences appliquées ML
- Namita Das, responsable de produit senior, Services verticaux d'IA
- Christy Bergman, architecte de solutions spécialisée senior en IA/ML
- Kris Tonthat, rédacteur technique, documentation IA/ML

Autres lectures

Pour plus d'informations sur les prévisions de séries temporelles et les méthodes de deep learning, consultez :

- Documentation Amazon Forecast
- Blog sur la disponibilité générale d'Amazon Forecast
- Now available in Amazon SageMaker: DeepAR algorithm for more accurate
- Amazon SageMaker DeepAR now supports missing values, categorical and time series features, and generalized frequencies
- Amazon Forecast utilise désormais les réseaux de neurones conventionnels (CNN) pour entraîner les modèles de prévision jusqu'à 2 fois plus rapidement avec une précision jusqu'à 30 % supérieure
- Amazon Forecast prend désormais en charge les mesures de précision pour les éléments individuels
- Mesurez la précision de vos modèles de prévisions pour optimiser vos objectifs métier avec Amazon Forecast
- Indice météorologique Amazon Forecast - inclusion automatique de la météo locale pour améliorer la précision de votre modèle de prévision
- Articles scientifiques sur les modèles de prévision des séries temporelles
- Page GitHub d'exemples d'Amazon Forecast
- Centre d'architecture AWS

Annexe A : FAQ

Q : Comment démarrer avec Amazon Forecast ?

1. Tout d'abord, vous aurez besoin d'un Compte AWS.
2. Ensuite, ouvrez le service Forecast dans la AWS Management Console, créez un groupe de jeu de données, et importez un fichier `.csv` dans le jeu de données de la série temporelle cible (obligatoire). Les données minimales requises pour commencer sont des données historiques pour la quantité que vous voulez prédire, comme l'électricité par horodatage et par ménage.
3. Enfin, créez un modèle en exécutant CreatePredictor et générez des résultats en exécutant CreateForecast. Pour plus de détails, consultez la page de documentation Getting Started (Démarrage).

Consultez également le guide GitHub Introduction and Best Practices (Introduction et Bonnes pratiques).

Q : Amazon Forecast est-il adapté à mes besoins ?

Tous les problèmes de machine learning ne sont pas des problèmes de prévision. La première question à se poser est la suivante : « Mon problème métier inclut-il les séries temporelles dans son énoncé ? » Par exemple, avez-vous besoin d'une valeur particulière uniquement à une date et une heure précises dans le futur ? La prévision n'est pas adaptée aux problèmes généraux et statiques (où la date/heure particulière n'a pas d'importance), tels que la détection des fraudes ou la recommandation de titres de films aux utilisateurs. Il existe des solutions beaucoup plus rapides aux problèmes statiques.

En plus de disposer de données de séries temporelles, les données elles-mêmes doivent être « denses » et présenter un long historique. Le tableau suivant en fait un résumé :

Tableau 2 — Critères et classes d'algorithmes Amazon Forecast

Critères	Classe d'algorithme Amazon Forecast
Grand jeu de données comprenant jusqu'à cinq millions de séries temporelles présentant des tendances sous-jacentes similaires + effets	Deep learning propriétaire d'Amazon Forecast DeepAR+, CNN-QR

Critères	Classe d'algorithme Amazon Forecast
saisonniers + données associées. Chaque série temporelle doit présenter un long historiqu e, idéalement plus de deux ans si vous essayez de capturer des événements annuels, et chaque série temporelle doit avoir plus de 300, idéalement au moins 1 000 points de données.	
Petit jeu de données avec 1 à 100 séries temporelles, où la majorité des séries temporell es comportent plus de 300 points de données + effets saisonniers + données associées.	Prophet
Petit jeu de données contenant de 1 à 10 séries temporelles, la majorité des séries temporelles comportant plus de 300 points de données et des effets saisonniers.	ETS, ARIMA
Jeu intermittent (clairsemé avec de nombreux 0) contenant de 1 à 10 séries temporelles, où la majorité des séries temporelles comportent plus de 300 points de données.	NPTS propriétaire d'Amazon Forecast
Petit jeu de données (régulier ou clairsemé) contenant de 1 à 10 séries temporelles, où la majorité des séries temporelles contiennent moins de 300 points de données.	Les données sont trop petites pour Amazon Forecast. Essayez plutôt ETS dans Excel ou les modèles statistiques traditionnels ARIMA et Prophet.

La bonne pratique consiste à s'entraîner avec le mode AutoML dans votre Predictor, dès la première utilisation de vos données. AutoML exécutera automatiquement tous les algorithmes (les algorithmes DL sont exécutés avec l'optimisation HPO activée), afin d'apprendre quel algorithme fonctionne le mieux sur vos données.

Q : Comment dois-je aborder les données manquantes ? À partir de quel moment sont-elles trop nombreuses pour générer des prévisions raisonnables ?

Il se peut qu'il y ait des problèmes dans l'enregistrement des données ou que le niveau d'agrégation des données soit trop faible ou trop élevé. La règle générale est la suivante : la longueur des prévisions ne peut pas dépasser le tiers des données d'entraînement.

Outre la quantité de données manquantes, il convient également de tenir compte de l'imputation des données manquantes. Vous pouvez convertir tous les 0 en valeurs nulles et laisser Amazon Forecast faire le gros du travail pour imputer automatiquement les valeurs manquantes. Amazon Forecast détectera automatiquement si les valeurs manquantes sont dues à l'introduction de nouveaux produits (démarrage à froid) ou à des produits en fin de vie. Vous pouvez utiliser plusieurs logiques de valeurs manquantes, notamment la valeur, la médiane, le minimum, le maximum, le zéro, la moyenne et le NaN (séries temporelles cibles uniquement). Consultez la documentation pour la syntaxe de remplissage de valeurs nulles.

- « frontfill » (remplissage avant) : (TTS uniquement) désigne les éléments nouveaux ou démarrés à froid et la manière dont vous souhaitez traiter les valeurs nulles avant que l'article ne commence à avoir un historique
- « middlefill » (remplissage intermédiaire) : désigne les valeurs nulles situées au milieu des valeurs des séries temporelles
- « backfill » (remplissage en amont) : désigne les articles en fin de vie et la manière dont vous souhaitez traiter les valeurs nulles une fois qu'un article ne se vend plus
- « futurefill » (remplissage futur) : (RTS uniquement) désigne les valeurs nulles qui apparaissent après la fin des données d'entraînement

Q : Mes données historiques d'entrée n'ont pas de valeurs négatives, mais je vois des valeurs négatives dans les prévisions de la demande. Pourquoi cela se produit-il ? Que puis-je faire pour éviter cela ?

Pour tous les modèles autres que NPTS (entraînés sur les données non négatives) et DeepAR (avec fonction de vraisemblance binomiale négative), rien ne garantit la génération de nombres positifs. La solution consiste à passer à l'un des modèles susmentionnés, ou à tronquer les valeurs prévisionnelles à des valeurs non négatives.

Q. Pourquoi les métriques de précision diffèrent-elles selon les quantiles ? L'erreur ne devrait-elle pas être la même puisque le modèle est identique ?

Consultez Weighted Quantile Loss (wQL) [Perte de quantile pondérée (wQL)] pour obtenir plus d'explications sur la relation entre la pondération et le quantile.

Imaginez que vous ayez toutes les prévisions pour trois quantiles différents : p10, p50, p90. Les trois prédictions elles-mêmes sont des variables aléatoires. Les précisions sont calculées séparément entre les réalisations et les prévisions à chaque niveau de quantile. Vous pouvez voir un tableau de « wQL », pertes quantiles pondérées, comme ci-dessous. Les valeurs wQL n'ont aucune relation déterministe entre elles. (Rappelons que la perte signifie l'erreur, elle n'est donc pas ordonnée ; les prévisions quantiles, cependant, sont ordonnées). Ainsi, il n'y a aucune raison pour que p90 wQL soit plus grand que p50 wQL, par exemple.

Tableau 3 — Exemples de quantiles prévisionnels

	A	B	C
1	P10 wQL	P50 wQL	P90 wQL
2	0,18647	0,50879	0,30428

Q : Comment puis-je améliorer la précision des prévisions ?

La précision des prévisions dépend de la disponibilité des bonnes données, en quantité et en qualité suffisantes. Si la précision n'est pas satisfaisante, il peut être utile de comprendre dans quelle mesure le problème est prévisible (ou dans quelle mesure les données sont aléatoires/bruitées/stationnaires). Parmi les autres facteurs à prendre en compte, citons l'évaluation de différents modèles et paramètres d'hyperparamètres, ainsi que l'incorporation de fonctionnalités supplémentaires en utilisant les jeux de données de séries temporelles et de métadonnées d'éléments associés. Pour des suggestions spécifiques, consultez ce document de bonnes pratiques sur GitHub.

Q : J'ai un algorithme qui fonctionne très bien pour mon cas d'utilisation, et il n'est pas proposé dans Amazon Forecast. Que dois-je faire ?

L'équipe d'Amazon Forecast sera heureuse de vous aider dans ce cas d'utilisation. Contactez l'équipe de service d'Amazon Forecast par e-mail : <amazonforecast-poc@amazon.com>.

Annexe B : Références

Januschowski, Tim et Kolassa, Stephan. A classification of business forecasting problems. Foresight: The International Journal of Applied Forecasting. 2019

Salinas, David, Flunkert, Valentin, Gasthaus, Jan et Januschowski, Tim. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting. 2019

Gasthaus, Jan, Benidis, Konstantinos, Wang, Yuyang, Rangapuram, Syama Sundar, Salinas, David, Flunkert, Valentin et Januschowski, Tim. {Probabilistic Forecasting with Spline Quantile Function RNNs. La 22e conférence internationale sur l'intelligence artificielle et les statistiques. 2019

Januschowski, Tim, Gasthaus, Jan, Wang, Yuyang, Salinas, David, Flunkert, Valentin, Bohlke-Schneider, Michael et Callot, Laurent. Criteria for classifying forecasting methods. International Journal of Forecasting. 2019 (Connexion requise)

Januschowski, Tim, Gasthaus, Jan, Wang, Yuyang, Rangapuram, Syama et Callot, Laurent. Deep Learning for Forecasting. Foresight: The International Journal of Applied Forecasting. 2018

Januschowski, Tim, Gasthaus, Jan, Wang, Yuyang, Rangapuram, Syama Sundar et Callot, Laurent. Deep Learning for Forecasting: Current Trends and Challenges. Foresight: The International Journal of Applied Forecasting. 2018

Bose, Joos-Hendrik, Flunkert, Valentin, Gasthaus, Jan, Januschowski, Tim, Lange, Dustin, Salinas, David, Schelter, Sebastian, Seeger, Matthias et Wang, Yuyang. Probabilistic demand forecasting at scale. Proceedings of the VLDB Endowment. 2017

Historique du document

Pour être informé des mises à jour de ce livre blanc, abonnez-vous au flux RSS.

Modification	Description	Date
Livre blanc mis à jour	Mises à jour.	1er septembre 2021
Publication initiale	Première publication du livre blanc.	4 février 2020

> **Note**
>
> Pour vous abonner aux mises à jour RSS, un plug-in RSS doit être activé pour le navigateur que vous utilisez.

Mentions légales

Les clients sont responsables de leur propre évaluation indépendante des informations contenues dans ce document. Le présent document : (a) est fourni à titre informatif uniquement, (b) représente les offres et pratiques actuelles de produits AWS, qui sont susceptibles d'être modifiées sans préavis, et (c) ne crée aucun engagement ou assurance de la part d'AWSet de ses affiliés, fournisseurs ou concédants de licences. Les produits ou services AWS sont fournis « en l'état » sans garantie, représentation ou condition, de quelque nature que ce soit, explicite ou implicite. Les responsabilités et obligations d'AWSenvers ses clients sont déterminées par les contrats AWS. Le présent document ne fait pas partie d'un contrat entre AWS et ses clients, et ne le modifie pas.

Glossaire AWS

Pour connaître la terminologie AWS la plus récente, consultez le Glossaire AWS dans la Référence générale AWS.