AWS 백서

Amazon Forecast에서 시계열 예측 원칙

Copyright © 2023 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon Forecast에서 시계열 예측 원칙: AWS 백서

Copyright © 2023 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon의 상표 및 브랜드 디자인은 Amazon 외 제품 또는 서비스와 함께, Amazon 브랜드 이미지를 떨어뜨리거나 고객에게 혼동을 일으킬 수 있는 방식으로 사용할 수 없습니다. Amazon이 소유하지 않은 기타 모든 상표는 Amazon과 제휴 관계이거나 관련이 있거나 후원 관계 여부에 관계없이 해당 소유자의 자산입니다.

Table of Contents

요약 및 개요 i
개요 1
Well-Architected를 적용하고 있으신가요? 2
예측에 대해 3
예측 시스템 3
예측 문제는 어디에서 발생하나요? 3
예측 문제를 해결하기 전에 고려할 사항 4
사례 연구: 전자 상거래 비즈니스의 소매 수요 예측 문제 6
1단계: 데이터 수집 및 집계 8
예 9
2단계: 데이터 준비 11
누락된 데이터 처리 방법 11
예 1 11
예 2 13
기능화 및 관련 시계열 개념 13
예 3 14
2단계: 예측기 생성 16
4단계: 예측 변수 평가 18
백테스트 18
예측 분위 및 정확도 지표 20
가중 분위 손실(wQL) 20
가중 절대 백분율 오차(WAPE) 20
평균 제곱근 오차(RMSE) 21
WAPE 및 RMSE의 문제점 21
5단계: 의사 결정을 위한 예측 생성 및 사용 23
확률적 예측 23
시각화 24
예측 워크플로 및 API 요약 25
일반적인 시나리오에 Amazon Forecast 사용 26
프로덕션 환경에 예측 구현 26
결론 28
기여자 29
추가 자료 30
부록 A: 자주 묻는 질문 31

부록 B: 참고 문헌 34
문서 이력 35
고지 사항 36
AWS 용어집 37

Amazon Forecast에서 시계열 예측 원칙

게시 날짜: 2021년 9월 1일(문서 이력)

오늘날 기업은 제품 수요, 리소스 요구, 재무 성과 등의 향후 비즈니스 성과를 정확하게 예측하기 위해 단순 스프레드시트, 복잡한 재무 계획 소프트웨어 등의 다양한 도구를 사용하고 있습니다. 이 백서에서는 예측, 용어, 문제 및 사용 사례를 소개합니다. 이 문서에서는 사례 연구를 통해 예측 개념, 예측 단계 및 Amazon Forecast가 실제 예측 문제에서 발생하는 여러 가지 실제 문제를 해결하는 데 어떻게 도움이 되는지 참조하여 설명합니다.

개요

'예측'은 미래를 예견하는 과학입니다. 이전 데이터를 사용하여 기업은 추세를 파악하고, 어떤 일이 언제 일어날 지 예측하고, 그 정보를 제품 수요부터 재고 계획 및 인력 배치에 이르는 모든 것에 대한 향후 계획에 반영할 수 있습니다.

예측의 결과를 고려할 때 정확도가 중요합니다. 예측이 너무 높으면 고객이 제품 및 직원에 과도하게 투자하게 되어 투자가 낭비될 수 있습니다. 예측이 너무 낮으면 고객의 적게 투자하게 되어 원자재 및 재고가 부족해져 고객 경험이 저하될 수 있습니다.

오늘날 기업들은 간단한 스프레드시트부터 복잡한 수요/재무 계획 소프트웨어에 이르기까지 모든 것을 사용하여 예측을 생성하려고 하지만 다음과 같은 두 가지 이유로 높은 정확도를 유지하기는 여전히 어렵습니다.

- 첫째, 전통적인 예측은 대량의 과거 데이터를 통합하는 데 어려움을 겪고 있으며 노이즈로 손실되는 과거의 중요한 신호를 놓치고 있습니다.
- 둘째, 전통적인 예측은 관련성이 있지만 독립적인 데이터를 거의 포함하지 않고 있습니다. 이러한 데이터는 가격, 휴일/이벤트, 품절, 마케팅 프로모션 등과 같은 중요한 컨텍스트를 제공할 수 있습니다. 전체 기록과 더 넓은 맥락 없이는 대부분의 예측이 미래를 정확하게 예측하지 못합니다.

Amazon Forecast는 이러한 문제를 해결하는 완전 관리형 서비스입니다. Amazon Forecast는 현재 예측 시나리오에 가장 적합한 알고리즘을 제공합니다. 적절한 경우 최신 기계 학습(ML)과 딥 러닝을 활용하여 매우 정확한 예측을 제공합니다. Amazon Forecast는 사용하기 쉬우며 기계 학습 경험이 없어도 됩니다. 이 서비스는 필요한 인프라를 자동으로 제공하고, 데이터를 처리하고, AWS에서 호스팅되고 예측이 가능한 맞춤형/사설 ML 모델을 구축합니다. 또한 기계 학습 기술이 빠른 속도로 계속 발전함에 따라 Amazon Forecast는 이러한 기술을 통합하여 고객이 최소한의 추가 노력으로 정확도를 지속적으로 개선할 수 있도록 합니다.

Well-Architected를 적용하고 있으신가요?

AWS Well-Architected Framework는 클라우드에서 시스템을 구축하는 동안 사용자가 내리는 의사 결정의 장단점을 이해하는 데 도움이 됩니다. 이 프레임워크의 6가지 원칙을 통해 안정적이고 안전하며 효과적이고 비용 효율적이며 지속 가능한 시스템을 설계 및 운영하기 위한 아키텍처 모범 사례를 배울 수 있습니다. AWS Management Console에서 추가 요금 없이 사용할 수 있는 AWS Well-Architected Tool을 사용하면 각 원칙에 대한 여러 질문에 답하여 이러한 모범 사례와 비교하며 워크로드를 검토할 수 있습니다.

기계 학습 렌즈에서는 AWS 클라우드에서 기계 학습 워크로드의 설계, 배포 및 아키텍처 구성 방법을 집중적으로 살펴봅니다. 이 렌즈는 Well-Architected 프레임워크에 포함된 모범 사례에 추가됩니다.

클라우드 아키텍처에 대한 보다 전문적인 지침과 모범 사례(참조 아키텍처 배포, 다이어그램 및 백서)는 AWS 아키텍처 센터를 참조하세요.

예측에 대해

이 문서에서 예측이란 시계열의 미래 값 예측을 의미합니다. 문제에 대한 입력 또는 출력은 시계열 특성을 갖습니다.

예측 시스템

예측 시스템에는 다음과 같은 다양한 사용자가 포함됩니다.

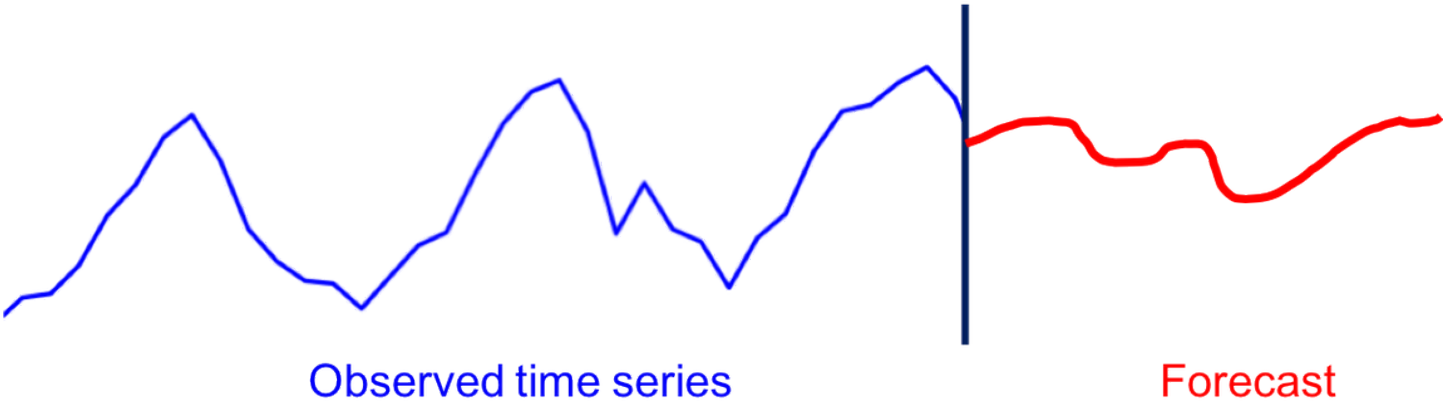
- 최종 사용자: 특정 제품에 대한 예측을 쿼리하고 구매할 제품 수를 결정합니다. 이는 사람일 수도 있고 자동화된 시스템일 수도 있습니다.
- 비즈니스 분석가/비즈니스 인텔리전스: 최종 사용자를 지원하며 집계 보고를 실행하고 구성합니다.
- 데이터 과학자: 수요 패턴과 인과 관계를 반복적으로 분석하고 새로운 기능을 추가하여 모델을 점진적으로 개선하거나 예측 모델을 개선합니다.
- 엔지니어: 데이터 수집의 인프라를 설정하고 시스템에 입력 데이터의 가용성을 보장합니다.

Amazon Forecast는 소프트웨어 엔지니어의 작업을 줄여 주며 데이터 과학 역량이 제한된 기업에서도 최신 예측 기술을 활용할 수 있도록 합니다. 데이터 과학 역량을 갖춘 기업의 경우 Amazon Forecast에 포함된 다양한 진단 기능을 통해 예측 문제를 잘 해결할 수 있습니다.

예측 문제는 어디에서 발생하나요?

예측 문제는 자연적으로 시계열 데이터를 생성하는 여러 영역에서 발생합니다. 여기에는 소매 판매, 의료 분석, 용량 계획, 센서 네트워크 모니터링, 재무 분석, 소셜 활동 마이닝 및 데이터베이스 시스템이 포함됩니다. 예를 들어 예측은 데이터 기반 의사 결정을 지원하는 대부분의 비즈니스에서 운영 프로세스를 자동화하고 최적화하는 데 중요한 역할을 합니다. 제품 공급 및 수요에 대한 예측은 최적의 재고 관리, 직원 일정 관리 및 토폴로지 계획에 사용될 수 있으며, 일반적으로 공급망 최적화의 대부분 측면에서 중요한 기술입니다.

다음 그림은 패턴(이 예에서는 계절성)을 나타내는 관측된 시계열을 기반으로 특정 기간 동안 예측이 생성될 때 발생하는 예측 문제에 대한 요약을 보여 줍니다. 가로축은 과거(왼쪽)에서 미래(오른쪽)로 이동하는 시간을 나타냅니다. 세로축은 측정된 단위를 나타냅니다. 검은색 수직선까지의 과거(파란색)를 고려하여 미래(빨간색)를 식별하는 것이 예측 과제입니다.



예측 작업 개요

예측 문제를 해결하기 전에 고려할 사항

예측 문제를 해결하기 전에 이해해야 할 가장 중요한 질문은 다음과 같습니다.

- 예측 문제를 해결해야 하나요?
- 예측 문제를 해결하는 이유는 무엇인가요?

시계열 데이터가 어디에나 있기 때문에 어디서나 예측 문제를 쉽게 찾을 수 있습니다. 그러나 핵심 질문은 예측 문제를 해결할 필요가 있는지 또는 비즈니스의 효율적인 의사 결정에 영향을 미치지 않으면서 예측 문제를 완전히 피할 수 있는지 여부입니다. 과학적으로 말하자면 예측은 기계 학습에서 가장 어려운 문제 중 하나이기 때문에 이 질문을 던지는 것이 중요합니다.

예를 들어 온라인 소매업체를 위한 제품 추천을 고려해 보겠습니다. 이 제품 추천 문제는 각 고객 재고 관리 단위(SKU) 쌍에 대해 특정 고객이 구매할 특정 항목의 개수를 예측하는 예측 문제로 간주할 수 있습니다. 이 문제 공식화에는 여러 가지 이점이 있습니다. 한 가지 이점은 시간 구성 요소가 명시적으로 고려되므로 고객의 구매 패턴에 따라 제품을 추천할 수 있다는 것입니다.

그러나 제품 추천 문제를 예측 문제로 공식화하는 경우는 거의 없습니다. 이러한 예측 문제를 해결하는 것은 추천 문제를 직접 해결하는 것보다 훨씬 어렵기 때문입니다(예: 고객-SKU 수준에서의 정보 희소성 및 문제의 규모). 따라서 예측 애플리케이션을 생각할 때는 예측의 다운스트림 사용량을 고려하고 대체 접근 방식을 사용하여 이 문제를 해결할 수 있는지 여부를 고려하는 것이 중요합니다.

이 경우 Amazon Personalize가 도움이 될 수 있습니다. Amazon Personalize는 개발자가 애플리케이션을 사용하는 개별 고객에게 맞는 추천을 쉽게 생성할 수 있는 기계 학습 서비스입니다.

예측 문제를 해결해야 한다고 판단한 후 다음으로 던져야 할 질문은 예측 문제를 해결하는 이유에 대한 것입니다. 많은 비즈니스 환경에서 예측은 일반적으로 목적을 달성하기 위한 수단에 불과합니다. 예를

들어, 소매 상황에서 수요 예측의 경우 예측을 사용하여 재고 관리 결정을 내릴 수 있습니다. 예측 문제는 일반적으로 의사 결정 문제에 대한 입력이며, 이는 다시 최적화 문제로 모델링될 수 있습니다.

이러한 의사 결정 문제의 예로는 구매할 상품 수량이나 기존 재고를 처리하기 위한 최선의 방법이 포함됩니다. 기타 비즈니스 예측 문제에는 서버 용량 예측 또는 제조 상황에서의 원자재/부품 수요 예측이 포함됩니다. 이러한 예측은 위와 같은 의사 결정 문제 또는 시나리오 시뮬레이션에 대한 다른 프로세스의 입력으로 사용될 수 있으며, 이러한 정보는 명시적 모델 없이 계획 수립에 사용될 수 있습니다. 예측이 그 자체로 끝이 아니라는 규칙에는 예외가 있습니다. 예를 들어 재무 예측에서 예측은 재무 준비금을 모으는 데 직접 사용되거나 투자자에게 제공됩니다.

예측의 목적을 이해하려면 다음 질문을 고려하세요.

- 얼마나 오랫동안 미래를 예측해야 하나요?
- 얼마나 자주 예측을 생성해야 하나요?
- 예측에서 자세히 살펴봐야 할 특정 측면이 있나요?

사례 연구: 전자 상거래 비즈니스의 소매 수요 예측 문제

예측 개념을 더 자세히 설명하기 위해 온라인으로 제품을 판매하는 전자 상거래 비즈니스를 예로 들어 보겠습니다. 공급망에서의 의사 결정(예: 재고 관리)을 최적화하는 것은 적절한 주문 처리 위치에 정확한 수의 제품을 보유하는 데 도움이 되므로 이 비즈니스의 핵심 경쟁력에 매우 중요합니다. 이는 기본적으로 더 짧은 배송 시간과 경쟁력 있는 가격으로 다양한 제품을 갖출 수 있다는 것을 의미하며, 이는 고객 만족도를 높입니다. 공급망 소프트웨어 시스템에 대한 주요 입력은 카탈로그에 있는 모든 제품의 수요 예측 또는 잠재적 판매 예측입니다. 이 예측을 통해 중요한 다운스트림 결정을 내릴 수 있으며, 그 중 핵심은 다음과 같습니다.

- 거시적 계획(전략적 예측): 비즈니스 전체의 경우 총 매출/수익 측면에서 예상되는 성장률은 얼마인가요? 지리적으로 비즈니스가 (더) 활발히 활동해야 하는 곳은 어디인가요? 인력을 어떻게 배치해야 하나요?
- 수요(또는 재고) 예측: 매장 당 판매될 것으로 예상되는 각 제품은 몇 개인가요?
- 프로모션 활동(전술적 예측): 프로모션은 어떻게 진행해야 하나요? 제품을 처분해야 하나요?

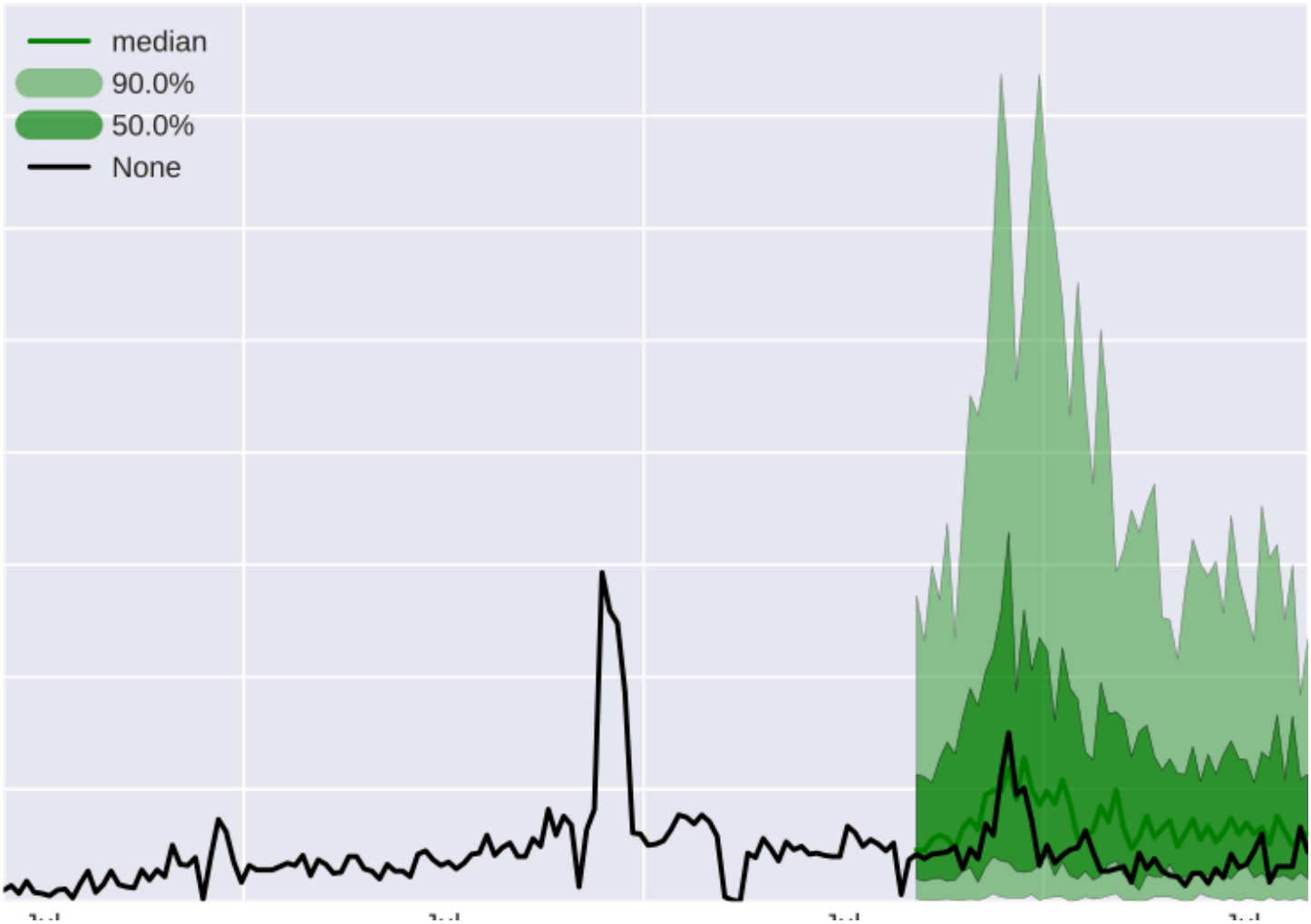
나머지 사례 연구는 운영 예측 문제 계열의 일부인 두 번째 문제에 초점을 맞추고 있습니다 (Januschowski & Kolassa, 2019년). 이 문서는 데이터, 모델(예측 변수), 추론(예측) 및 생산화와 같은 주요 관심사를 다룹니다.

이 사례 연구에서는 예측 문제가 목적을 위한 수단이라는 점을 염두에 두는 것이 중요합니다. 예측은 비즈니스에 매우 중요하지만, 다운스트림 공급망 결정은 훨씬 더 중요합니다. 사례 연구에서는 운영 연구의 수학적 최적화 모델을 사용하는 자동 구매 시스템이 이러한 결정을 내립니다. 이러한 시스템은 비즈니스에 필요한 예상 비용을 최소화하려고 합니다.

여기서 핵심 단어는 예상으로, 이는 특정 결과의 확률에 따라 적절한 가중치를 부여하여 예측이 가능한 한 가지 미래뿐만 아니라 가능한 모든 미래를 포함해야 한다는 뜻입니다. 이를 위한 다운스트림 의사 결정의 핵심 요소는 한 지점의 예측이 아니라 예측 값을 완전히 분배하는 것입니다. 다음 그림은 확률적 예측(밀도 예측이라고도 함) 을 보여줍니다. 이 확률적 예측에서 단일 지점 예측(가장 가능성이 높은 미래)을 쉽게 도출할 수 있지만, 지점 예측에서 확률적 예측으로 전환하는 것은 더 어렵습니다.

확률적 예측을 통해 다양한 통계를 얻고 원하는 결정을 내리는 데 도움이 되도록 결과를 조정할 수 있습니다. 전자 상거래 비즈니스에는 품절이 거의 발생하지 않는 여러 주요 제품이 있을 수 있습니다. 이 경우 높은 분위수(예: 90번째 백분위수)를 사용하면 제품 재고가 확보되는 시간의 90%로 변환됩니다. 대체품을 쉽게 찾을 수 있는 상품(예: 연필)과 같은 다른 상품의 경우 더 낮은 백분위수를 사용하는 것이 더 적절할 수 있습니다.

Amazon Forecast에서는 확률적 예측에서 다양한 분위수를 쉽게 얻을 수 있습니다.



확률적 예측의 그림

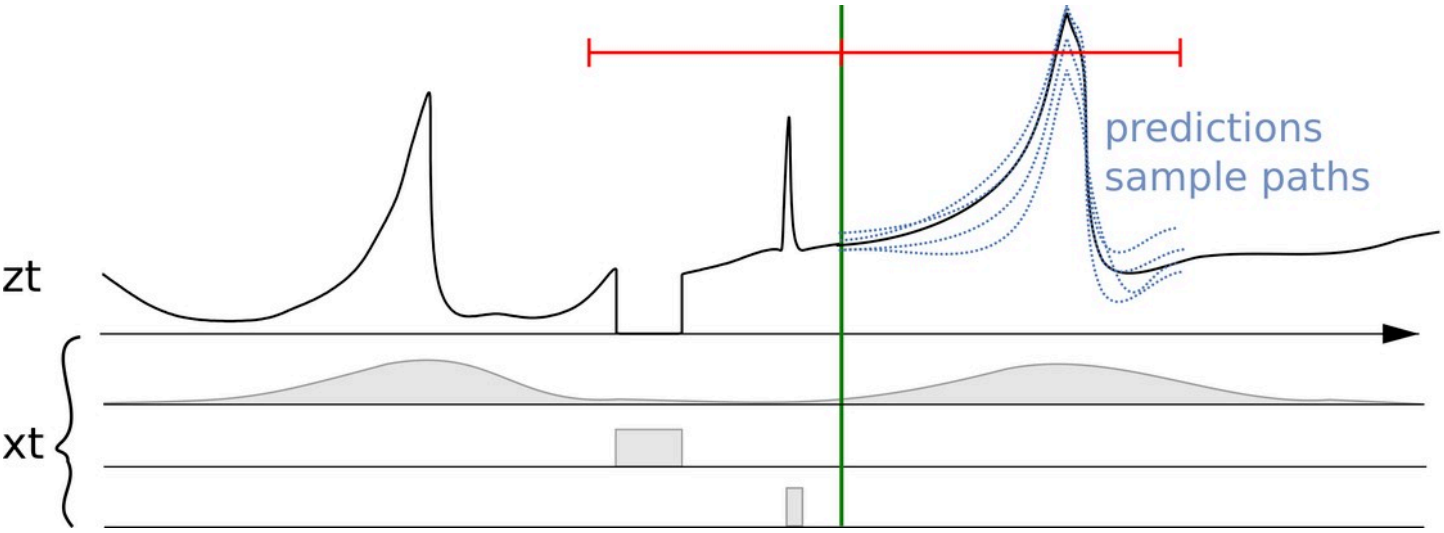
위 그림에서 검은색 선은 실제 값이고, 짙은 녹색 선은 예측 분포의 중앙값이며, 짙은 녹색 음영 영역은 값의 50% 가 하락할 것으로 예상되는 예측 구간이며, 연한 녹색 영역은 실제 값의 90% 가 하락할 것으로 예상되는 예측 구간입니다.

다음 섹션에서는 다음을 포함하여 이 비즈니스의 예측 문제를 해결하는 데 필요한 단계를 다룹니다.

- 데이터 수집 및 집계(1단계)
- 데이터 준비(2단계)
- 예측 변수 만들기(3단계)
- 예측 변수 평가(4단계)
- 예측 생성 자동화(5단계)

1단계: 데이터 수집 및 집계

다음 그림은 예측 문제에 대한 멘탈 모델을 보여줍니다. 목표는 가능한 한 많은 관련 정보를 사용하여 미래의 시계열 z_t를 예측하여 가능한 한 정확하게 예측하는 것입니다. 따라서 가장 중요한 첫 번째 단계는 올바른 데이터를 최대한 많이 수집하는 것입니다.



시계열 z_t와 관련 특징 또는 공변량 (x_t) 및 다중 예측값

위 그림에서는 수직선 오른쪽에 여러 예측이 표시됩니다. 이러한 예측은 확률적 예측 분포의 샘플입니다(또는 반대로 확률적 예측을 나타내는 데 사용할 수 있음).

소매 비즈니스에서 기록해야 할 주요 정보는 다음과 같습니다.

- 거래 판매 데이터 - 예: 재고 관리 단위(SKU), 위치, 타임스탬프 및 판매 수량
- SKU 상품 세부 정보 데이터 - 상품의 메타데이터. 예: 색상, 부서, 크기 등.
- 가격 데이터 - 타임스탬프가 있는 각 상품의 가격 시계열.
- 프로모션 정보 데이터 - 다양한 유형의 프로모션(상품 모음(카테고리) 또는 타임스탬프가 있는 개별 상품)
- 재고 정보 데이터 - 시간 단위마다 해당 SKU의 재고 또는 구매 가능 여부를 비교한 SKU의 재고 상태 정보
- 위치 데이터 - 특정 시점의 상품 또는 판매 위치를 `location_id` 또는 `store_id` 문자열이나 실제 지리적 위치로 표시. 지리적 위치는 국가 코드와 5자리 우편번호 또는`latitude_longitude` 좌표일 수 있습니다. 위치는 거래 판매의 '차원'으로 간주됩니다.

Amazon Forecast에서는 예측할 수량의 과거 데이터를 대상 시계열(TTS)이라고 합니다. 소매 비즈니스의 경우 TTS는 거래 판매 데이터입니다. 모든 판매 거래와 정확히 동시에 알려진 기타 과거 데이터를 관련 시계열(RTS)이라고 합니다. 소매 비즈니스의 경우 RTS에는 가격, 프로모션 및 재고 변수가 포함됩니다.

이 문제는 판매가 아닌 예측 수요에 중점을 두고 있지만 비즈니스에서는 판매량만 기록하므로 재고 정보가 중요하다는 점에 유의하세요. SKU가 품절되면 판매 수량이 잠재 수요보다 적으므로 이러한 품절 이벤트가 발생하는 시기를 파악하고 기록하는 것이 중요합니다.

고려해야 할 기타 데이터 세트로는 웹페이지 방문 수, 검색어 세부 정보, 소셜 미디어, 날씨 정보 등이 있습니다. 이 데이터를 모델에서 사용할 수 있으려면 과거 그리고 미래의 데이터를 사용할 수 있어야 하는 경우가 많습니다. 이는 많은 예측 모델 및 백테스트(4단계: 예측 평가 섹션 참조)에서 요구되는 사항입니다.

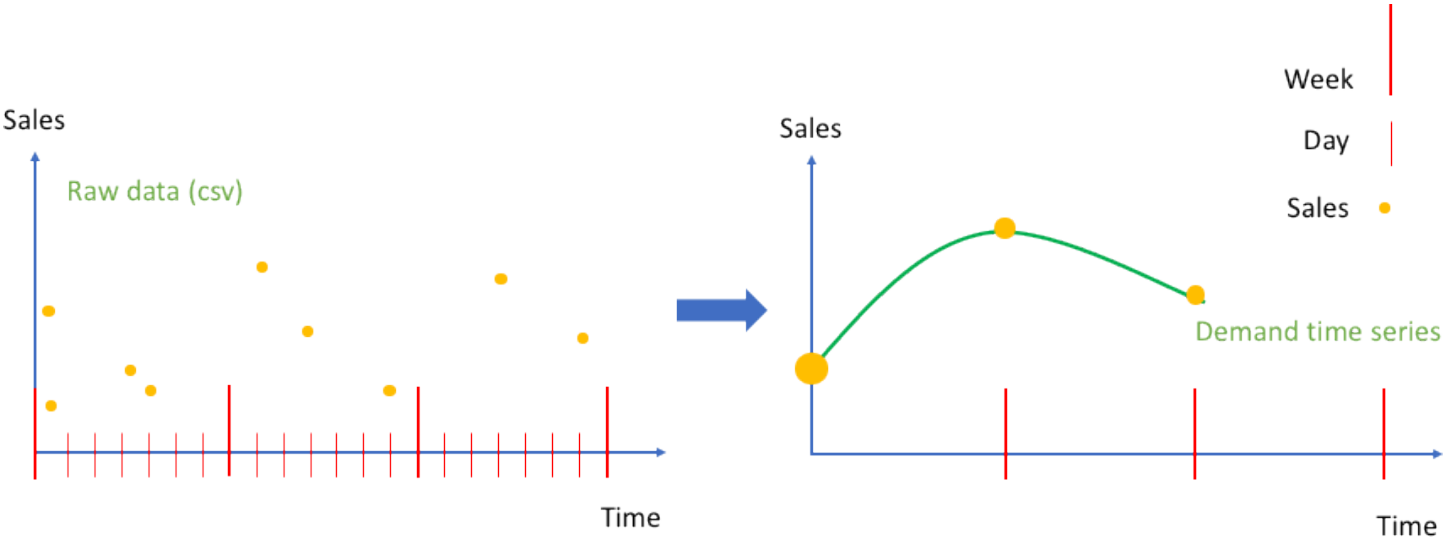
일부 예측 문제의 경우 원시 데이터의 빈도가 예측 문제의 빈도와 자연스럽게 일치합니다. 분 단위로 예측하려는 경우 분 단위로 샘플링되는 서버 볼륨 요청을 예로 들 수 있습니다.

데이터는 종종 더 정밀한 빈도로 기록되거나 특정 시간 범위 내에서 임의의 타임스탬프로 기록되지만, 예측 문제는 더 세부적으로 기록됩니다. 이는 일반적으로 판매 데이터가 거래 데이터로 기록되는 소매 사례 연구에서 흔히 볼 수 있습니다. 예를 들어 형식은 판매가 발생한 시점을 세밀하게 나타내는 타임스탬프로 구성됩니다. 예측 사용 사례에서는 이러한 낮은 세분 수준이 필요하지 않을 수 있으며 이 데이터를 시간별 또는 일별 매출로 집계하는 것이 더 적절할 수 있습니다. 여기서 집계 수준은 재고 관리 또는 리소스 계획과 같은 다운스트림 문제에 해당합니다.

예

다음 그림의 왼쪽 그래프는 Amazon Forecast에 CSV(쉼표로 구분된 값) 파일로 입력할 수 있는 원시 고객 판매 데이터의 예를 보여줍니다. 이 예시에서는 판매 데이터가 더 세밀한 일일 시간 그리드로 정의되며, 문제는 더 조밀한 시간 그리드에서 미래의 주간 수요를 예측하는 것입니다. Amazon Forecast는 `create_predictor` API 호출에서 특정 주의 일별 값을 집계합니다.

결과는 원시 데이터를 고정된 주간 빈도의 올바른 형식의 시계열 컬렉션으로 변환합니다. 오른쪽 그래프는 기본 합계 집계 방법을 사용하여 대상 시계열에 대한 이 집계를 보여줍니다. 다른 집계 방법에는 평균, 최대값, 최소값 또는 단일 지점 선택(예: 첫 번째 지점)이 포함되어 있습니다. 데이터의 비즈니스 사용 사례에 가장 잘 맞는 집계 수준 및 방법을 선택해야 합니다. 이 예에서는 집계된 값이 주간 집계에 맞춰 정렬됩니다. `create_predictor` API의 `FeaturizationConfig` 파라미터`FeaturizationMethodParameters` 키를 사용하여 사용자가 다른 집계 방법을 설정할 수 있습니다.



원시 판매 데이터를 이벤트로 집계하고(왼쪽) 동일한 간격의 시계열로 나타냄(오른쪽)

2단계: 데이터 준비

원시 데이터를 사용할 수 있게 되면 데이터 누락과 같은 복잡한 문제를 해결하고 의도한 해석을 가장 잘 포착할 수 있는 예측 모델에 사용할 데이터를 준비해야 합니다.

누락된 데이터 처리 방법

실제 예측 문제에서 흔히 발생하는 문제는 원시 데이터에 누락된 값이 있는 경우입니다. 시계열에서 누락된 값은 지정된 빈도의 모든 시점에서 해당하는 실제 값을 추가 처리에 사용할 수 없음을 의미합니다. 값이 누락된 것으로 표시되는 이유로 여러 가지가 있을 수 있습니다.

트랜잭션이 없거나 측정 오류가 발생할 수 있기 때문에 누락된 값이 발생할 수 있습니다(예: 특정 데이터를 모니터링한 서비스가 제대로 작동하지 않았거나 측정이 올바르게 수행되지 않은 경우). 소매 사례 연구에서 후자의 주요 예로 수요 예측의 품절 상황을 들 수 있습니다. 즉, 수요가 해당 날짜의 판매량과 같지 않다는 뜻입니다.

서비스가 한도에 도달하면 클라우드 컴퓨팅 시나리오에서도 비슷한 결과가 발생할 수 있습니다(예: 특정 AWS 리전의 Amazon EC2 인스턴스가 모두 사용량이 많음). 누락된 값 발생의 또 다른 예는 제품 또는 서비스가 아직 출시되지 않았거나 생산이 중단된 경우가 있습니다.

누락된 값을 특성 처리 구성 요소로 삽입하여 패딩이 있는 시계열 길이가 같도록 할 수도 있습니다. 누락된 값이 상당히 많을 경우 모델의 정확도에 큰 영향을 미칠 수 있습니다.

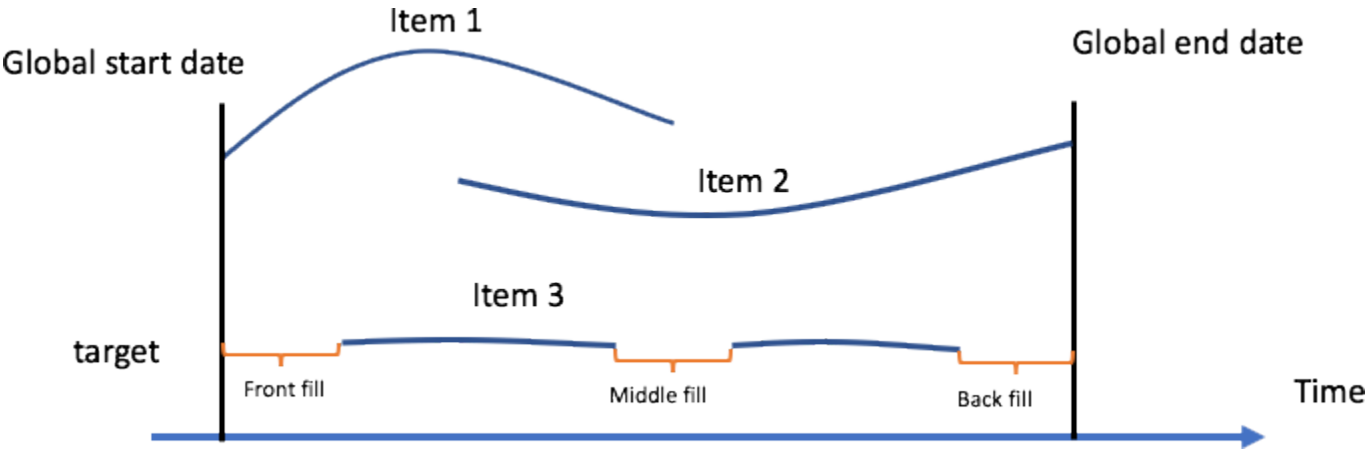
예 1

채우기는 데이터 세트에서 누락된 항목에 표준화된 값을 추가하는 프로세스입니다. 다음 그림에서는 세 항목으로 구성된 데이터 세트의 항목 2에 대해 Amazon Forecast의 누락된 값을 처리하는 다양한 전략(전면, 중간, 뒤로 및 앞으로 채우기) 을 보여 줍니다.

Amazon Forecast는 대상 시계열 및 관련 시계열 모두에 대한 채우기를 지원합니다. 글로벌 시작 날짜는 데이터 세트에 있는 모든 항목의 시작 날짜 중 가장 빠른 시작 날짜로 정의됩니다. 아래 예에서는 항목 1에 대한 글로벌 시작 날짜가 표시됩니다. 마찬가지로 글로벌 종료 날짜는 모든 항목에 대한 시계열의 가장 늦은 종료 날짜로 정의되며 항목 2에 해당합니다.

전면 채우기는 특정 시계열의 시작부터 글로벌 시작일까지의 모든 값을 채웁니다. 이 문서를 게시할 때에는 Amazon Forecast에서 전면 채우기를 설정하지 않으므로 모든 시계열이 서로 다른 시점에서 시작될 수 있습니다. 중간 채우기는 시계열의 중간(예: 항목의 시작 날짜와 종료 날짜 사이)에 채워진 값을 나타내며, 뒤로 채우기는 해당 시계열의 마지막 날짜부터 전체 종료 날짜까지 채워진 값을 나타냅니다.

대상 시계열의 경우 중간 및 뒤로 채우기 방법의 기본 채우기 로직은 0입니다. 앞으로 채우기(관련 시계열에만 적용)는 항목의 글로벌 종료 일자와 고객이 지정한 예측 기간 사이에 누락된 값을 모두 채웁니다. Prophet 및 DeepAR+와 함께 관련 시계열 데이터 세트를 사용하려면 미래 값이 필요하며 CNN-QR의 경우 선택 사항입니다.



Amazon Forecast에 누락된 가치 처리 전략

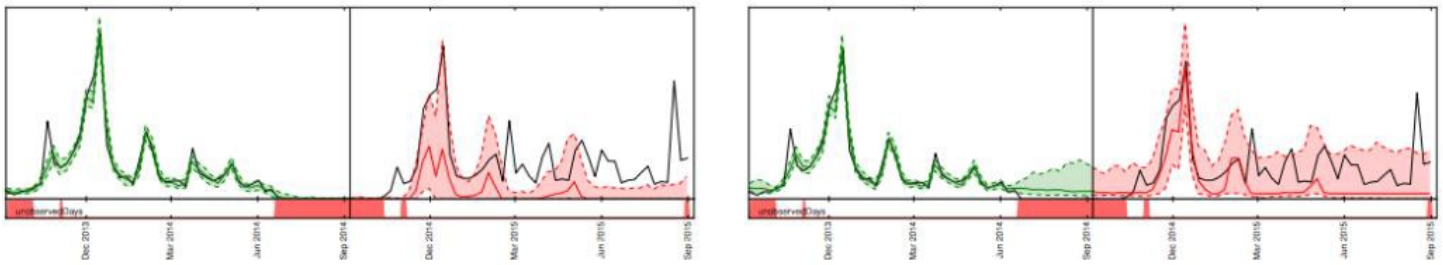
위 그림에서 글로벌 시작 날짜는 모든 항목의 시작 날짜 중 가장 빠른 시작 날짜를 나타내며, 글로벌 종료 날짜는 모든 항목의 종료 날짜 중 가장 늦은 종료 날짜를 나타냅니다. 예상 구간(Forecast Horizon)은 예측이 목표 값에 대한 예측을 제공하는 기간입니다.

이는 소매 연구에서 흔히 볼 수 있는 시나리오로, 가용 품목에 대한 거래 데이터에서 판매량이 0인 경우를 나타냅니다. 이러한 값은 실제 0으로 처리되며 지표 평가 구성 요소에 사용됩니다. Amazon Forecast를 사용하면 사용자가 실제로 누락된 값을 식별하고 알고리즘이 처리할 수 있도록 NaN(숫자가 아닌 값)으로 인코딩할 수 있습니다. 이 백서에서는 다음으로 이 두 가지 경우가 왜 다른지, 그리고 각각이 언제 유용한지 살펴봅니다.

소매 사례 연구에서 소매업체가 판매 가능한 품목을 0개 판매했다는 정보는 해당 품목이 존재하지 않는 기간(예: 출시 전 또는 지원 중단 후) 또는 해당 품목이 있는 기간(예: 일부 품절 또는 해당 기간 동안 기록된 판매 데이터가 없는 경우)에 판매가 불가능한 품목을 0개 판매했다는 정보와 다릅니다. 전자의 경우에는 기본 0 채우기를 적용할 수 있습니다. 후자의 경우 해당하는 목표 값이 일반적으로 0이지만 누락된 것으로 표시된 값에는 추가 정보가 전달됩니다. 가장 좋은 방법은 누락된 데이터가 있다는 정보를 보존하고 이 정보를 버리지 않는 것입니다. 정보 보관이 중요한 이유에 대한 설명은 다음 예를 참조하세요.

Amazon Forecast는 값, 평균, 중앙값, 최소값 및 최대값의 추가 채우기 로직을 지원합니다. 관련 시계열(예: 가격 또는 프로모션)의 경우 올바른 누락값 로직이 속성 유형 및 사용 사례에 따라 달라지므로 중간, 뒤로 또는 앞으로 채우기 방법에 대한 기본값이 지정되지 않습니다. 관련 시계열에 지원되는 채우기 로직에는 0, 값, 평균, 중앙값, 최소값 및 최대값이 포함됩니다.

누락 값 채우기를 수행하려면 `CreatePredictor` 작업을 호출할 때 구현할 채우기 유형을 지정합니다. 채우기 로직은 `FeaturizationMethod` 객체에 지정됩니다. 예를 들어, 대상 시계열에서 판매가 불가능한 제품의 판매량이 0이 아닌 값을 인코딩하려면 채우기 유형을 NaN으로 설정하여 값을 실제로 누락된 값으로 표시하세요. 0으로 채워진 값과 달리 NaN으로 인코딩된 값은 실제로 누락된 것으로 간주되며 지표 평가 구성 요소에 사용되지 않습니다.



0 채우기와 NaN을 이용한 채우기 결과가 동일 품목에 대한 예측에 미치는 영향

위 그림의 왼쪽 그래프에서는 검은색 세로 선의 왼쪽에 있는 값이 0으로 채워져 있어 예측이 과소 편향되어 있습니다(수직 검은색 선의 오른쪽). 오른쪽 그래프에서 이러한 값은 NaN으로 표시되어 적절한 예측으로 이어집니다.

예 2

위 그림은 ARIMA 또는 ETS와 같은 선형 상태 공간 모델에서 누락된 값을 올바르게 처리하는 것의 중요성을 보여줍니다. 부분적으로 품절된 품목에 대한 수요 예측을 표시합니다. 학습 영역은 왼쪽 그래프에서 녹색으로, 오른쪽 패널의 예측 범위는 빨간색으로, 실제 목표는 검은색으로 표시됩니다. 중앙값, p10 및 p90 예측치는 각각 빨간색 선과 음영 영역에 표시됩니다. 아래쪽에는 빨간색으로 표시된 품절 상품(데이터의 80%)이 표시됩니다. 왼쪽 그림에서는 품절 영역이 무시되고 0으로 채워집니다.

그 결과 예측해야 할 0이 많다고 가정하고 예측 모델이 생성되어 너무 낮은 예측값이 나오게 됩니다. 오른쪽 그림에서는 품절 영역이 실제 누락된 관측치로 간주되고 품절 영역에서는 수요가 불확실해집니다. 품절 품목에 대한 누락값이 NaN으로 적절하게 표시되어 있으면 이 그림의 예측 범위에 과소 편향이 나타나지 않습니다. Amazon Forecast는 이러한 데이터 격차를 좁히므로 모든 입력 데이터를 명시적으로 수정하지 않고도 누락된 데이터를 쉽고 올바르게 처리할 수 있습니다.

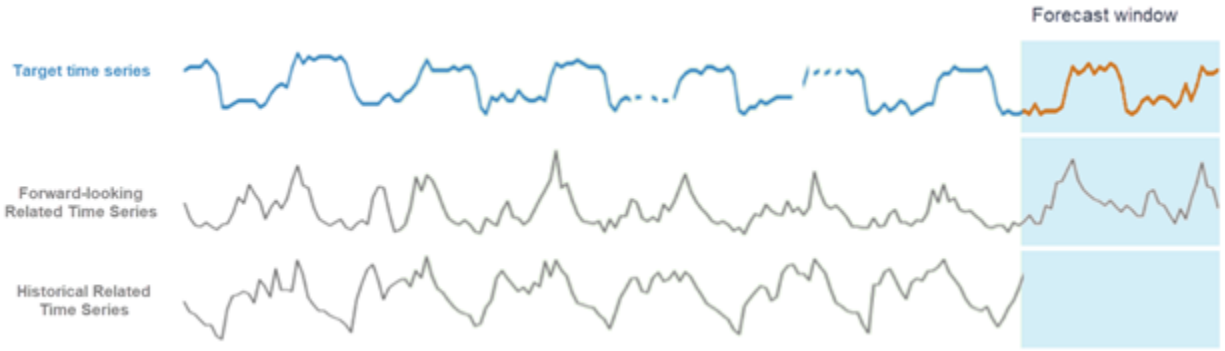
기능화 및 관련 시계열 개념

Amazon Forecast에서는 사용자가 관련 데이터를 입력하여 지원되는 특정 예측 모델의 정확도를 높일 수 있습니다. 이 데이터에는 관련 시계열 또는 정적 항목 메타데이터의 두 가지 유형이 있습니다.

> **Note**
>
> 메타데이터 및 관련 데이터를 기계 학습에서는 기능이라고 하고 통계에서는공변량이라고 합니다.

관련 시계열은 목표값과 어느 정도 상관관계가 있는 시계열을 말하며, 직관적인 용어로 설명을 제공하므로 목표값을 예측하는 데 통계적 강점을 어느 정도 제공할 수 있습니다(예시는 Amazon Forecast: 대규모 시계열 예측 참조). 대상 시계열과는 달리 관련 시계열은 과거에 알려진 값으로, 대상 시계열에 영향을 미칠 수 있으며 미래에도 알려진 값을 가질 수 있습니다.

Amazon Forecast에서는 두 가지 유형의 관련 시계열, 즉 과거 시계열 및 미래 예측 시계열을 추가할 수 있습니다. 과거 관련 시계열에는 예측 기간까지의 데이터 포인트가 포함되며 향후 예측 기간 내에는 데이터 포인트가 포함되지 않습니다. 미래 예측 관련 시계열에는 예측 기간까지의 데이터 포인트와 예측 기간 내의 데이터 포인트가 포함됩니다.



Amazon Forecast에서 관련 시계열을 사용하는 것과 관련된 다양한 접근 방식

예 3

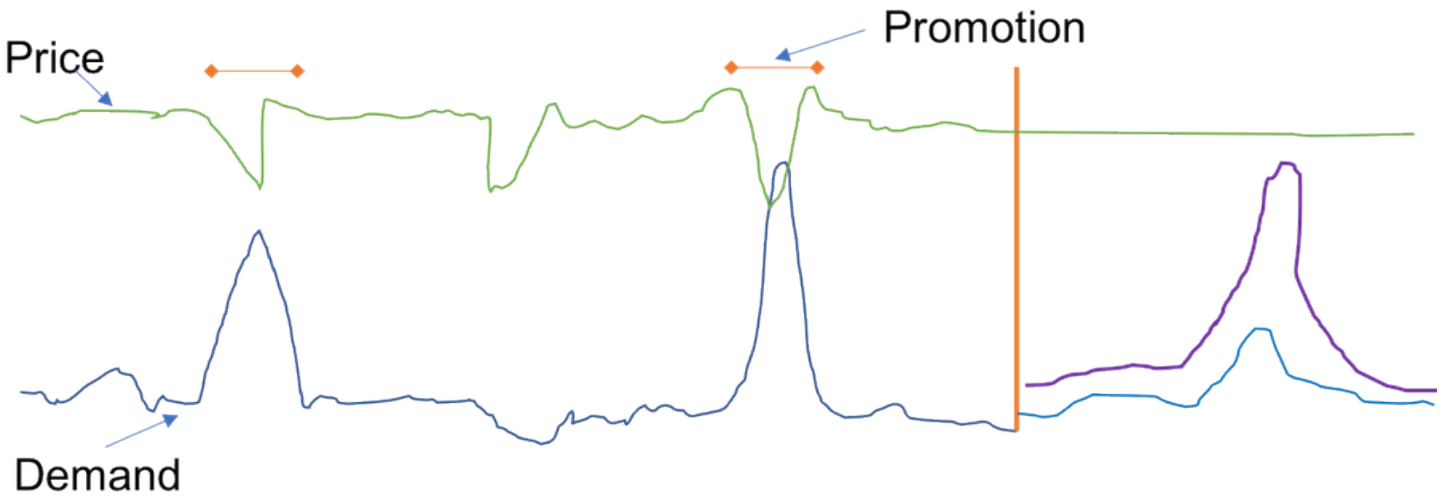
다음 그림은 관련 시계열을 사용하여 인기 도서의 향후 수요를 예측하는 방법의 예를 보여줍니다. 파란색 선은 목표 시계열의 수요를 나타냅니다. 가격은 초록색 선으로 표시됩니다. 세로선은 예측 시작 날짜를 나타내며 두 분위의 예측은 수직선 오른쪽에 표시됩니다.

이 예에서는 예측 세부 수준에서 목표 시계열에 맞춰 정렬되며 예측 시작 일자부터 예측 대상 기간(예측 종료 일자)만큼 증가된 범위 내에서 향후의 모든(또는 대부분) 시간을 알 수 있는 미래 예측 관련 시계열을 사용합니다.

다음 그림은 가격 하락과 제품 판매 증가 사이의 상관 관계를 확인할 수 있기 때문에 가격이 사용하기에 적합한 기능임을 보여줍니다. 항목 SKU, 타임스탬프 및 관련 시계열 값(이 경우 가격)이 포함된 별도의 CSV 파일을 통해 관련 시계열을 Amazon Forecast에 제공할 수 있습니다.

Amazon Forecast는 대상 시계열에 대한 평균 및 합계와 같은 집계 방법을 지원하지만 관련 시계열에 대해서는 지원하지 않습니다. 예를 들어 일일 가격을 주간 가격으로 합산하는 것은 거의 의미가 없으며 일일 프로모션의 경우에도 마찬가지입니다.

Amazon Forecast는 내장된 기능 데이터 세트를 포함하여 날씨 및휴일 정보를 모델에 자동으로 통합할 수 있습니다(`SupplementaryFeature` 참조). 날씨 정보 및 휴일은 소매 수요에 큰 영향을 미칠 수 있습니다.



특정 품목의 판매량(파란색으로 표시, 빨간색 세로선 왼쪽)

범주별 변수라고도 하는 항목 메타데이터는 Amazon Forecast에 입력할 수 있는 기타 유용한 기능입니다(예시는 Amazon Forecast: 대규모 시계열 예측 참조). 범주별 변수와 관련 시계열 간의 주요 차이점은 범주별 변수는 정적이며 시간이 지나도 변하지 않는다는 것입니다. 일반적인 소매업의 예로는 상품 색상, 책 카테고리, TV의 스마트 TV 여부를 표시하는 이진 표시기 등이 있습니다. 이 정보는 유사한 SKU의 판매량이 비슷하다고 가정할 때 재고 관리 단위(SKU) 간의 유사성을 학습하는 딥 러닝 알고리즘을 통해 수집할 수 있습니다. 이 메타데이터에는 시간 종속성이 없으므로 항목 메타데이터 CSV 파일의 각 행은 항목 SKU와 해당 카테고리 라벨 또는 설명으로만 구성됩니다.

2단계: 예측기 생성

예측기는 AutoML을 실행하거나 내장된 6가지 Amazon Forecast 알고리즘 중 하나를 수동으로 선택하는 두 가지 방법을 통해 생성할 수 있습니다. 이 문서 작성 시점에서, AutoML을 실행할 때 Amazon Forecast는 6개의 내장 알고리즘을 자동으로 테스트하고 10번째, 50번째(중앙값) 및 90번째 분위 중에서 평균 분위 손실이 가장 낮은 알고리즘을 선택합니다.

Amazon Forecast는 네 가지 로컬 모델을 제공합니다.

- Autoregressive Integrated Moving Average(ARIMA)
- Exponential Smoothing(ETS)
- Non-Parametric Time Series(NPTS)
- Prophet

로컬 모델은 단일 모델을 각 개별 시계열(또는 특정 항목/차원 조합)에 맞춘 다음 해당 모델을 사용하여 시계열을 미래로 추정하는 예측 방법입니다.

ARIMA 및 ETS는 R 예측 패키지의 인기 로컬 모델을 확장 가능한 버전으로 제공합니다. Amazon에서 개발한 로컬 방법인 NPTS는 다른 로컬 모델과 비교했을 때 핵심적인 차이가 있습니다. 마지막 값이나 적절한 계절성 값을 반복하여 포인트 예측을 제공하는 단순한 계절 예측기와 달리 NPTS에서는 확률적 예측을 생성합니다. NPTS는 고정 시간 인덱스를 사용하며, 여기서 이전 인덱스(T - 1) 또는 지난 시즌(T - tau)이 시간 단계 T에 대한 예측값입니다. 알고리즘은 집합 {0,..., T - 1} 의 시간 인덱스 (t)를 무작위로 샘플링하여 현재 시간 단계 T에 대한 표본을 생성합니다. NPTS는 0이 많은 간헐적(희소라고도 함) 시계열에 특히 효과적입니다. Forecast에는 베이지안 구조적 시계열 모델인 Prophet의 Python 구현도 포함됩니다.

Amazon Forecast는 두 가지 글로벌 딥 러닝 알고리즘을 제공합니다.

- DeepAR+
- CNN-QR

글로벌 모델은 데이터 세트의 전체 시계열 컬렉션에 대해 단일 모델을 학습시킵니다. 이는 횡단면 단위 집합에 유사한 시계열이 있을 때 특히 유용합니다. 이러한 시계열 그룹화에는 다양한 제품, 서버 로드 및 웹 페이지에 대한 요청이 필요합니다.

일반적으로 시계열 수가 증가할수록 CNN-QR과 DeepAR+의 효능이 증가합니다. 로컬 모델의 경우 항상 그런 것은 아닙니다. 딥 러닝 모델을 사용하여 과거 판매 데이터가 거의 없거나 전혀 없는 상태에서 신규 SKU에 대한 예측을 생성할 수도 있습니다. 이를 콜드 스타트 예측이라고 합니다.

	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)	✓	✓	✗	✗	✗	✗

Amazon Forecast에서 사용할 수 있는 알고리즘 비교

관련 시계열을 자세히 알아보려면 관련 시계열을 참조하세요.

4단계: 예측 변수 평가

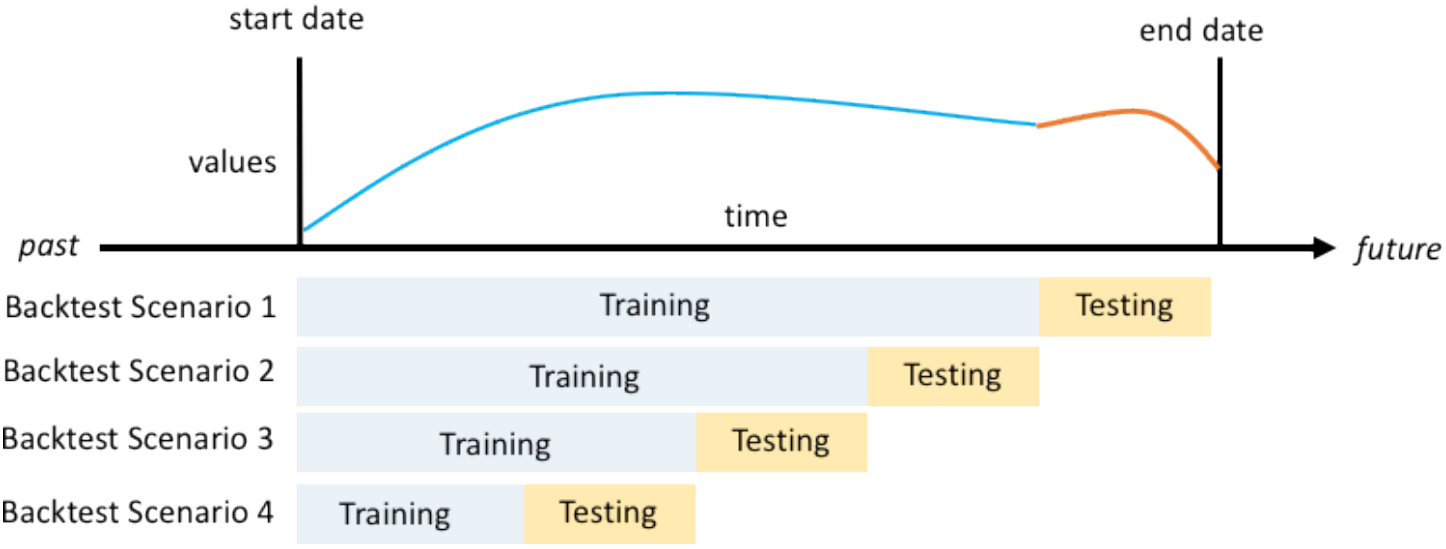
기계 학습의 일반적인 워크플로는 훈련 세트에서 모델 세트 또는 모델 조합을 트레이닝하고 홀드아웃 데이터 세트에서 정확도를 평가하는 것으로 구성됩니다. 이 섹션에서는 과거 데이터를 분할하는 방법과 시계열 예측에서 모델을 평가하는 데 사용할 지표에 대해 설명합니다. 예측의 경우 백테스트 기법은 예측 정확도를 평가하는 주요 도구입니다.

백테스트

적절한 평가 및 백테스트 프레임워크는 기계 학습 애플리케이션을 성공으로 이끄는 가장 중요한 요소 중 하나입니다. 모델의 성공적인 백테스트를 통해 모델의 미래 예측력에 대한 확신을 얻을 수 있습니다. 또한 하이퍼파라미터 최적화(HPO)를 통해 모델을 조정하고, 모델 조합을 학습하고, 메타 학습 및 AutoML을 활성화할 수 있습니다.

시계열 예측 특성 시간은 평가 및 백테스트 방법론 측면에서 응용 기계 학습의 다른 분야와 차별화됩니다. 일반적으로 ML 작업에서는 백테스트의 예측 오류를 평가하기 위해 데이터 세트를 항목별로 분할합니다. 예를 들어 이미지 관련 작업의 교차 검증을 위해 사진의 일정 비율을 학습시킨 다음 다른 부분을 테스트 및 검증에 사용합니다. 예측을 할 때는 주로 시간별로(항목별로 낮게) 구분해서 훈련 세트의 정보가 테스트 또는 검증 세트로 유출되지 않도록 하고 생산 사례를 최대한 가깝게 시뮬레이션해야 합니다.

한 시점을 선택하는 것이 아니라 여러 지점을 선택하려면 시간별 분할을 신중하게 수행해야 합니다. 그렇지 않으면 분할 지점으로 정의된 예측 시작 날짜에 따라 정확도가 너무 크게 좌우됩니다. 여러 시점에 걸쳐 일련의 분할을 수행하고 평균 결과를 출력하는 롤링 예측 평가를 통해 더욱 강력하고 신뢰할 수 있는 백테스트 결과를 얻을 수 있습니다. 다음 그림은 네 가지 다른 백테스트 분할을 보여줍니다.



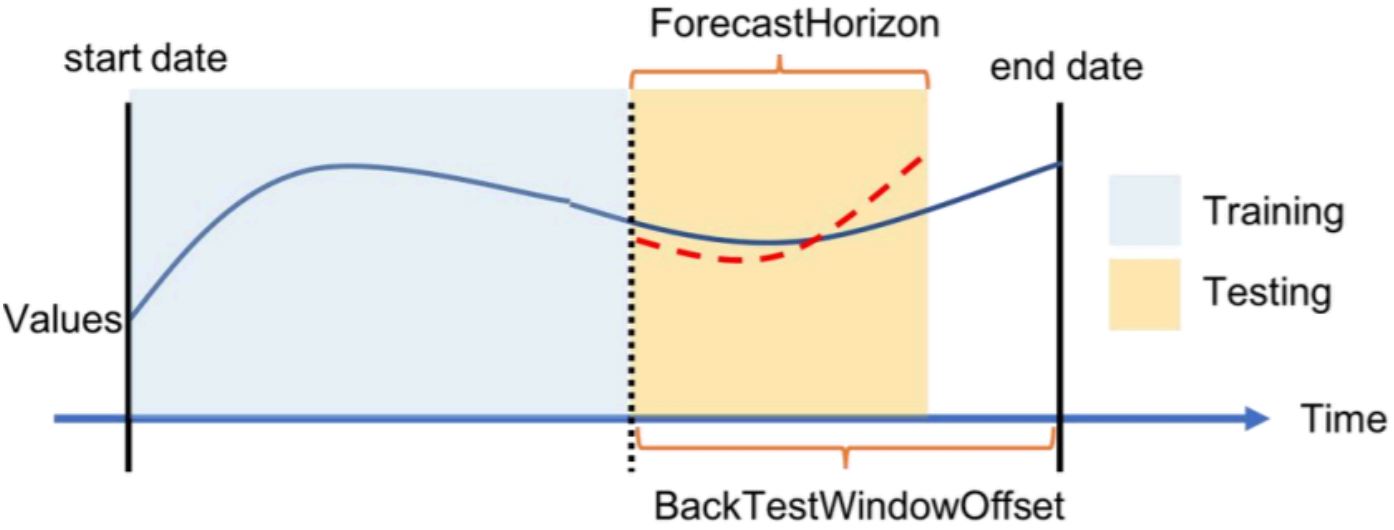
학습 세트 크기는 증가하지만 테스트 크기는 일정하게 유지되는 네 가지 백테스트 시나리오의 그림

위 그림에서 모든 백테스트 시나리오에는 예측된 값을 실제 값과 비교하여 평가할 수 있는 전체 데이터가 포함되어 있습니다.

여러 개의 백테스트 창이 필요한 이유는 실제 세계의 대부분의 시계열이 일반적으로 고정되지 않기 때문입니다. 북미에 기반을 둔 사례 연구의 전자 상거래 비즈니스는 제품 수요의 대부분이 4분기 성수기에 주도되며 특히 추수감사절과 크리스마스 이전에 최고조에 달합니다. 4분기 쇼핑 시즌에는 시계열 변동성이 해당 연도의 나머지 시기보다 높습니다. 여러 개의 백테스트 창을 통해 보다 균형 잡힌 환경에서 예측 모델을 평가할 수 있습니다.

다음 그림은 각 백테스트 시나리오에 대한 Amazon Forecast 용어의 기본 요소를 보여줍니다. Amazon Forecast는 자동으로 데이터를 학습 및 테스트 데이터 세트로 분할합니다. Amazon Forecast는 `create_predictor` API에서 파라미터로 지정되거나 기본값인 `ForecastHorizon`을 사용하는 `BackTestWindowOffset` 파라미터를 사용하여 입력 데이터를 분할하는 방법을 결정합니다.

다음 그림에서는`BackTestWindowOffset` 및`ForecastHorizon` 매개변수가 같지 않은 보다 일반적인 전자의 경우를 볼 수 있습니다. 이 `BackTestWindowOffset` 매개변수는 가상 예측 시작 날짜를 정의하며, 다음 그림에서 수직 점선으로 표시됩니다. '이 날 모델을 배포할 경우 예측치는 어떻게 될까요?'와 같은 가상의 질문에 답하는 데 사용할 수 있습니다. `ForecastHorizon`은 가상 예측 시작 날짜부터 예측까지 걸리는 시간 단계 수를 정의합니다.



Amazon Forecast에서의 단일 백테스트 시나리오 및 해당 구성 그림

Amazon Forecast는 백테스트 중에 생성된 예측값과 정확도 지표를 내보낼 수 있습니다. 내보낸 데이터는 특정 시점 및 분위에서 특정 항목을 평가하는 데 사용할 수 있습니다.

예측 분위 및 정확도 지표

예측 분위는 예측의 상한과 하한을 제공할 수 있습니다. 예를 들어, 예측 유형 0.1(P10), 0.5(P50) 및 0.9(P90)를 사용하면 P50 예측에 대한 80% 신뢰 구간이라고 하는 값 범위가 제공됩니다. P10, P50, P90에서 예측을 생성하면 실제 값이 80% 의 확률로 해당 경계 사이에 있을 것으로 예상할 수 있습니다.

이 백서에서는 5단계의 분위에 대해 더 자세히 설명합니다.

Amazon Forecast는 가중 분위 손실(wQL), 평균 제곱근 오차(RMSE) 및 가중 절대 백분율 오차(WAPE) 정확도 지표를 사용하여 백테스트 중에 예측기를 평가합니다.

가중 분위 손실(wQL)

가중 분위 손실(wQL) 오류 지표는 지정된 분위에서의 모델 예측 정확도를 측정합니다. 과소 예측과 과대 예측에 따른 비용이 서로 다를 때 특히 유용합니다. wQL 함수의 가중치 (τ)를 설정하면 과소 예측과 과대 예측에 대한 서로 다른 페널티가 자동으로 적용됩니다.

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

wQL 함수

여기서 각 항목은 다음과 같습니다.

- τ - {0.01, 0.02,..., 0.99} 집합 내의 분위
- qi,t(τ) - 모형이 예측하는 τ-분위
- yi, t - (i, t) 지점에서의 관측값

가중 절대 백분율 오차(WAPE)

가중 절대 백분율 오차(WAPE)는 모델 정확도를 측정하는 데 일반적으로 사용되는 지표입니다. 예측값과 관측값의 전체 편차를 측정합니다.

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

WAPE

여기서 각 항목은 다음과 같습니다.

- yi, t - (i, t) 지점에서의 관측값
- ŷi, t - (i, t) 지점에서의 예측값

예측은 평균 예측값을 예측값인 ŷi,t로 사용합니다.

평균 제곱근 오차(RMSE)

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

평균 제곱근 오차(RMSE)는 모델 정확도를 측정하는 데 일반적으로 사용되는 지표입니다. WAPE와 마찬가지로 관측값에서 추정치의 전체 편차를 측정합니다.

여기서 각 항목은 다음과 같습니다.

- yi, t - (i, t) 지점에서의 관측값
- ŷi, t - (i, t) 지점에서의 예측값
- nT - 테스트 세트의 데이터 포인트 수

예측은 평균 예측값을 예측값인 ŷi,t로 사용합니다. 예측 변수 지표를 계산할 때 nT는 백테스트 창의 데이터 포인트 수입니다.

WAPE 및 RMSE의 문제점

대부분의 경우 내부적으로 또는 다른 예측 도구를 통해 생성할 수 있는 포인트 예측은 p50 분위 또는 평균 예측과 일치해야 합니다. WAPE와 RMSE 모두에 대해 Amazon Forecast는 평균 예측을 사용하여 예측값 (yhat) 을 나타냅니다.

wQL[tau] 방정식의 tau = 0.5인 경우 두 가중치가 동일하며 wQL[0.5]은 포인트 예측에 일반적으로 사용되는 가중 절대 백분율 오차(WAPE)로 감소합니다.

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

여기서 yhat = q(0.5) 는 계산된 예측값입니다. wQL 공식에서 배율 인수 2는 0.5 인수를 상쇄하여 정확한 WAPE[중앙값] 식을 구하는 데 사용됩니다.

참고로 위의 WAPE 정의는 평균 절대 백분율 오차(MAPE)에 대한 일반적인 해석과 다릅니다. 차이점은 분모에 있습니다. 위에서 정의한 WAPE를 사용하면 사례 연구의 전자 상거래 비즈니스와 같은 실제 시나리오에서 흔히 발생하는 문제인 0으로 나누는 문제를 피할 수 있습니다. 이러한 문제는 특정 날짜에 특정 SKU를 0개 판매하는 경우가 많습니다.

0.5가 아닌 tau에 대한 가중 분위 손실 지표와 달리 가중치가 동일한 WAPE와 같은 계산으로는 각 분위의 고유한 편향을 포착할 수 없습니다. WAPE의 다른 단점은 대칭적이지 않고, 작은 숫자에 대한 백분율 오차가 과도하게 늘어나며, 점 단위 지표에 불과하다는 것입니다.

RMSE는 WAPE의 오차 항의 제곱이며 다른 ML 애플리케이션의 일반적인 오차 지표입니다. RMSE 지표는 개별적인 오차의 크기가 일정한 모델을 선호합니다. 오차의 변동이 커지면 RMSE가 과도하게 커지기 때문입니다. 오차의 제곱근이기 때문에 그렇지 않았다면 양호한 예상에서 몇 개의 잘못 예측된 값이 RMSE를 증가시킬 수 있습니다. 또한 제곱 항으로 인해 오차 항이 작을수록 WAPE보다 RMSE에서 가중치가 작아집니다.

정확도 지표를 통해 예측을 정량적으로 평가할 수 있습니다. 특히 대규모 비교(방법 A가 방법 B보다 전체적으로 우수한지 등)의 경우 이는 매우 중요합니다. 그러나 개별 SKU에 대한 시각 자료로 이를 보완하는 것이 중요한 경우가 많습니다.

5단계: 의사 결정을 위한 예측 생성 및 사용

특정 사용 사례에 필요한 정확성 임계값을 충족하는 모델을 확보했다면(백테스트를 통해 결정) 마지막 단계는 모델을 배포하고 예측을 생성하는 것입니다. Amazon Forecast에 모델을 배포하려면 Create_Forecast API를 실행해야 합니다. 이 액션은 데이터를 교육 세트와 테스트 세트로 분할하는 `Create_Predictor`과는 달리 전체 과거 데이터 세트에 대해 교육하여 만든 모델을 호스팅합니다. 그러면 예측 기간 동안 생성된 모델 예측을 다음 두 가지 방식으로 사용할 수 있습니다.

- AWS CLI에서 `Query_Forecast` API를 사용하거나 AWS Management Console을/를 통해 직접 특정 항목에 대한 예측을 쿼리할 수 있습니다(항목 또는 항목/차원의 조합을 지정하여).
- `Create_Forecast_Export_Job` API를 사용하여 모든 분위의 모든 항목 및 차원 조합에 대한 예측을 생성할 수 있습니다. 이 API는 선택한 Amazon Simple Storage Service(S3) 위치에 안전하게 저장되는 CSV 파일을 생성합니다. 그런 다음 CSV 파일의 데이터를 사용하여 의사 결정에 사용되는 다운스트림 시스템에 연결할 수 있습니다. 예를 들어, 기존 공급망 시스템은 Amazon Forecast의 결과를 직접 수집하여 특정 SKU의 제조와 관련된 의사 결정을 내리는 데 도움이 될 수 있습니다.

확률적 예측

Amazon Forecast는 다양한 분위로 예측을 생성할 수 있으며, 이는 과소 예측과 과다 예측 비용이 서로 다를 때 특히 유용합니다. 예측기 교육 단계와 마찬가지로 p1과 p99 사이의 분위에 대한 확률적 예측을 생성할 수 있습니다.

기본적으로 Amazon Forecast는 예측기 교육 중에 사용된 것과 동일한 분위로 예측을 생성합니다. 예측기 교육 중에 분위를 지정하지 않으면 기본적으로 p10, p50 및 p90에서 예측이 생성됩니다.

p10 예측의 경우 실제 값이 10%의 확률로 예측 값보다 낮을 것으로 예상되며 wQL[0.1] 지표를 사용하여 정확도를 평가할 수 있습니다. 즉 P10 예측은 90%의 확률로 예측보다 낮게 표시되며, 재고 비축에 사용할 경우 90%의 확률로 품목이 품절됩니다. P10 예측은 저장 공간이 많지 않거나 투자 자본 비용이 높을 때 유용할 수 있습니다.

> **Note**
>
> 분위 예측의 공식 정의는 Pr(실제 값 <= 분위에서의 예측) = q입니다. 엄밀히 말하면 분위는 백분위수/100입니다. 통계학자들은 'P90 분위 수준'이라고 말하는 경향이 있는데, 이는 '분위 0.9'보다 말하기가 더 쉽기 때문입니다. 예를 들어 P90 분위 수준 예측은 실제 값이 90% 의 확률로 예측값보다 작을 것으로 예상될 수 있음을 의미합니다. 구체적으로 시간=t1이고 분위 수준=0.9일 때 예측값=30이면, 1,000개의 시뮬레이션이 있는 경우 시간=t1에서의 실제 값은 900

개의 시뮬레이션에서 30 미만일 것으로 예상되고 100개의 시뮬레이션에서는 실제 값이 30을 초과할 것으로 예상됩니다.

반면 P90 예측은 90%의 확률로 예측보다 높게 표시되며, 품목 미판매로 인한 기회 비용이 매우 높거나 투자 자본 비용이 낮을 때 유용합니다. 식료품점의 경우 우유나 화장지 같이 매장에서 절대 소진되는 것을 원하지 않으며 항상 진열대에 개의치 않고 둘 수 있는 제품에 P90 예측기를 사용할 수 있습니다.

p50 예측(종종 중간값 예측이라고도 함)의 경우 실제 값은 50%의 확률로 예측값보다 낮을 것으로 예상되며 wQL[0.5] 지표를 사용하여 정확도를 평가할 수 있습니다. 재고 과잉이 크게 우려되지 않고 특정 품목에 대한 수요가 적당한 경우 p50 분위 예측이 유용할 수 있습니다.

시각화

Amazon Forecast를 사용하면 기본적으로 AWS Management Console에서 예측을 플로팅할 수 있습니다. 또한 전체 Python 데이터 과학 스택을 활용할 수 있습니다(Amazon Forecast 예시 참조). Amazon Forecast를 사용하면`ExportForecastJob` API를 통해 예측을 CSV 파일로 내보낼 수 있으므로 사용자는 선택한 분석 도구에서 예측을 시각화할 수 있습니다.

Amazon Forecast 콘솔에서 다양한 분위로 제공되는 시각화

예측 워크플로 및 API 요약

다음 표는 예측 워크플로의 각 단계를 해당 Amazon Forecast API와 일치시킵니다.

표 1: 예측 단계 및 Amazon Forecast API

단계	API	API 함수
1단계: 데이터 수집 및 집계 2단계: 데이터 준비	`Create_Dataset_Group` , `Create_Dataset` , `Create_Dataset_Import_Job`	1. 문제에 대한 상위 수준 도메인(소매업, 지표 등)을 설정합니다. 2. 다양한 데이터 세트(대상, 관련, 항목 메타데이터)에 대한 스키마를 정의합니다. 3. Amazon S3에서 Amazon Forecast로 데이터를 가져옵니다.
3단계: 예측기 생성 4단계: 예측기 평가	`Create_Predictor`	1. ETL을 수행합니다. 2. 데이터를 교육/테스트 세트로 분할하고 모델을 교육시킵니다. 3. 선택적으로 `Create_predictor_backtest_Export_job` 을/를 사용하여 항목 수준 지표를 계산하기 위해 CSV로 백테스트 결과를 내보냅니다.
5단계: 의사 결정을 위한 예측 생성 및 사용	`Create_Forecast`	1. 모델을 교육/호스팅합니다. 2. 특정 관심 분위(예: 평균을 포함하여 1에서 99 사이의 모든 정수)에 대한 예측 기간을 생성합니다.

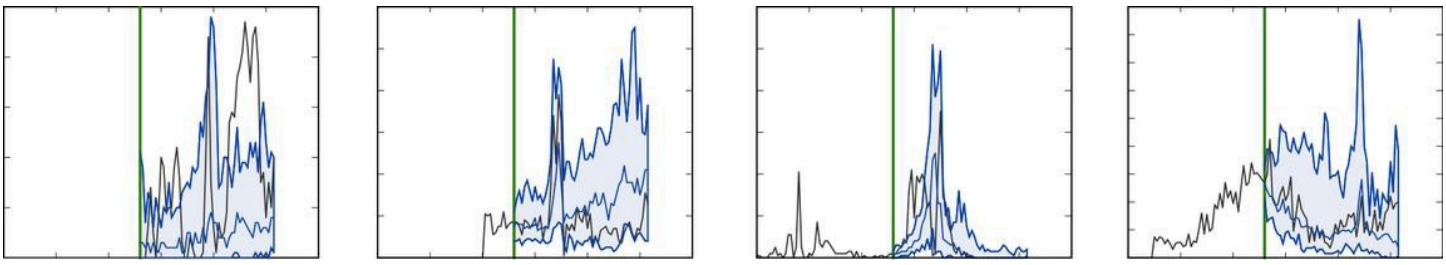
단계	API	API 함수
	`Query_Forecast` . `Create_Forecast_Export_Job`	`Create_Forecast` 에서 생성한 예측을 사용할 수 있습니다.

일반적인 시나리오에 Amazon Forecast 사용

외부 변수(예: 가격 또는 프로모션)의 변화를 기반으로 다양한 예측을 생성하여 가정 분석을 수행할 수도 있습니다. 예를 들어, 전자 상거래 사례 연구 예시에서는 계획 중인 프로모션을 기반으로 다양한 예측을 만들 수 있습니다. 10% 할인된 가격으로 제품에 대한 수요를 예측한 다음 20% 할인된 가격으로 제품 수요를 예측하여 수요를 충족하기 위해 재고를 보유해 놓아야 하는 제품의 수량을 파악할 수 있습니다. 이는 관심 시나리오를 기반으로 고유한 데이터 세트 그룹을 설정하고 각각의 관련 시계열을 업데이트하여 달성할 수 있습니다.

또한 이전 기록이 없는 품목(콜드 스타트 문제라고도 함)에 대한 예측을 생성할 수도 있습니다. 이 방법을 사용하려면 DeepAR+ 또는 CNN-QR을 메타데이터(예: 항목 메타데이터 데이터 세트)와 함께 사용하여 새 항목에 대한 예측을 생성해야 합니다.

다음 그림은 실제 운영 예측 문제에 나타나는 네 가지 SKU의 예를 보여줍니다.



실제 운영 예측 문제에 나타나는 네 가지 SKU의 예

위 이미지에서 검정색 실제 값과 비교하였을 때 관측된 실제값은 수직선의 왼쪽에 있고 파란색으로 표시된 예측값은 수직선의 오른쪽에 있습니다. 수직선 왼쪽에 표시된 각 개별 SKU 내역은 녹색 선 오른쪽에 표시된 SKU 내역의 진전을 의미하지 않습니다.

프로덕션 환경에 예측 구현

종합적인 Amazon Forecast 워크플로를 완료했으면 `Create_Predictor` 및 `Create_Forecast` API 간의 주요 차이점과 각 API를 사용해야 하는 시기를 식별하는 것이 중요합니다.

전자는 주로 개념 증명 과정에서 모델 정확도/지표를 평가하는 데 사용되는 반면 후자는 프로덕션 환경에서 예측을 생성하는 데 사용됩니다.

프로덕션 단계에서는 `Create_Predictor`을/를 예측을 생성해야 할 때마다 실행할 필요가 없으며, 데이터 변경으로 인해 또는 사전 설정된 주기(예: 격주 또는 월별)의 일부로 모델을 재교육해야 하는 경우에만 실행하면 됩니다. 데이터 세트가 새 데이터로 업데이트되므로 새 예측 기간에 대한 예측을 생성하는 데만 `Create_Forecast`을/를 실행하면 됩니다.

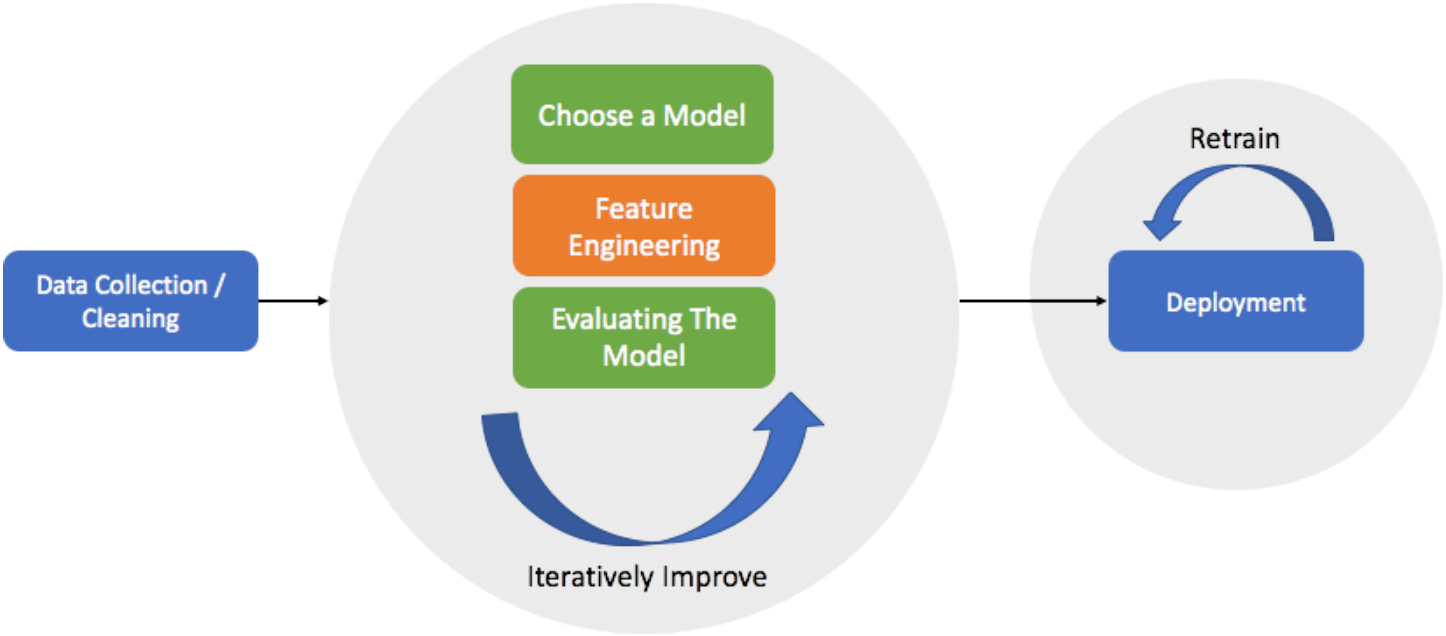
프로덕션 환경에서는 순차적으로 새로운 예측을 생성하기 위해 데이터 세트 가져오기 및 예측 작업도 자동화해야 합니다. 현재는 Amazon CloudWatch Events 로그, AWS Step Functions 및 AWS Lambda 함수를 조합해서 cron 작업을 설정하여 이를 달성할 수 있습니다. cron 작업 설정은 차례로 가져오기/재교육 또는 예측 생성을 위한 Amazon Forecast API 호출을 자동화합니다. 마지막으로, 서비스에 규정된 시스템 제한을 초과하지 않도록 리소스를 관리하고 정기적으로 삭제하는 것이 중요합니다. 예약된 작업 설정에 대한 자세한 내용은 Amazon Redshift를 비롯한 이 블로그 게시물을 참조하세요.

결론

Januschowski와 Kolassa(2019년)는 전략적이고전술적인 운영 결정을 포함하여 비즈니스가 취해야 하는 결정에 맞춰 예측 문제를 분류합니다. 각 의사 결정 수준에는 해당하는 예측 작업이 있습니다.

운영 및 전술적 예측 문제는 많은 양의 데이터를 포함하는 것이 특징이며 일반적으로 높은 수준의 자동화가 필요합니다. 다양한 예측 방법이 이러한 문제를 해결합니다. 로컬 예측 방법은 일반적으로 전략적 예측 문제에 적합하고 딥 러닝 기반 방법은 운영 예측 문제에 적합하며, 이 사이 문제인 경우에는 약간의 실험이 필요할 수 있습니다. 이 백서에서는 운영 예측 문제에 대해 논의했지만 Amazon Forecast는 제공하는 모델에 대해 편향적이지 않으며 전략적, 운영 및 전술적 예측 문제를 해결하는 모델을 포함합니다.

운영 예측 문제 해결 프로세스는 데이터 수집 및 준비부터 모델 구축 및 배포에 이르는 기본 단계로 나눌 수 있습니다. 일반적으로 이를 선형 프로세스보다는 반복 프로세스로 간주하는 것이 가장 유용합니다. 예를 들어 모델과 사용 사례가 더 잘 이해되면 데이터 수집 단계로 돌아가는 것이 합리적일 수 있습니다. 모델 개발 자체도 매우 반복적입니다.



예측 모델을 프로덕션 환경으로 가져오는 간소화된 개발 프로세스

기여자

이 문서를 작성하는 데 도움을 주신 분들입니다.

- Yuyang Wang, AI 버티컬 서비스 선임 기계 학습 과학자
- Danielle Robinson, AI 버티컬 서비스 응용 과학자
- Tim Januschowski, ML 응용 과학 매니저
- Namita Das, AI 버티컬 서비스 선임 제품 관리자
- Christy Bergman, 선임 AI/ML 전문 솔루션 아키텍트
- Kris Tonthat, AI/ML 문서 기술 작성자

추가 자료

시계열 예측 및 딥 러닝 방법에 대한 자세한 내용은 다음을 참조하세요.

- Amazon Forecast 설명서
- Amazon Forecast 일반 가용성 블로그
- 이제 Amazon SageMaker에서 사용 가능: DeepAR 알고리즘을 통한 정확도 향상
- Amazon SageMaker DeepAR, 이제 누락값, 범주형 및 시계열 기능, 일반화된 빈도 지원
- Amazon Forecast, 이제 CNN(합성곱 신경망)을 사용하여 최대 2배 빠른 속도와 최대 30% 더 높은 정확도로 예측 모델 훈련 가능
- Amazon Forecast, 이제 개별 항목에 대한 정확도 측정 지원
- Amazon Forecast를 통한 예측 모델 정확도의 간편한 측정을 통해 비즈니스 목표 최적화
- Amazon Forecast Weather Index – 현지 날씨를 자동으로 포함하여 예측 모델 정확도 향상
- 시계열 예측 모델에 관한 과학 논문
- Amazon Forecast 샘플 GitHub 페이지
- AWS 아키텍처 센터

부록 A: 자주 묻는 질문

Q: Amazon Forecast를 시작하려면 어떻게 해야 합니까?

1. 먼저 필요한 것은 AWS 계정입니다.
2. 그런 다음 AWS Management Console에서 Forecast 서비스를 열고 데이터 세트 그룹을 만든 다음 `.csv` 파일을 대상 시계열 데이터 세트로 가져옵니다(필수). 시작하는 데 필요한 최소 데이터는 예측하려는 수량에 대한 과거 데이터입니다(예: 각 가구의 타임스탬프당 전기).
3. 마지막으로 CreatePredictor를 실행하고 CreateForecast를 실행해 결과를 생성하여 모델을 생성합니다. 자세한 내용은 시작하기 설명서 페이지를 참조하세요.

GitHub 소개 및 모범 사례 가이드도 참조하실 수 있습니다.

Q: Amazon Forecast가 저에게 적합한가요?

모든 기계 학습 문제가 예측 문제인 것은 아닙니다. 첫 번째 질문은 '비즈니스 문제의 설명에 시계열이 포함되어 있나요?'입니다. 예를 들어, 미래의 특정 시간 및 날짜에만 특정 값이 필요할까요? 예측은 사기 탐지 또는 사용자에게 영화 제목 추천과 같은 일반적이고 정적인 문제(특정 날짜/시간이 중요하지 않은 경우)에는 적합하지 않습니다. 정적 문제에는 훨씬 더 빠른 해결책이 있습니다.

시계열 데이터를 보유하는 것 외에도 데이터 자체도 '밀도가 높고' 기록이 길어야 합니다. 이 내용은 다음 표에 요약되어 있습니다.

표 2 - 기준 및 Amazon Forecast 알고리즘 클래스

기준	Amazon Forecast 알고리즘 클래스
유사한 기본 패턴 + 계절 영향 + 관련 데이터를 가진 최대 5백만 개의 시계열이 있는 대규모 데이터 세트 각 시계열은 기록 길이가 길어야 하고 연간 이벤트를 캡처하려는 경우 기록이 2년 이상이 이상이어야 하며, 각 시계열에는 300개 이상, 이상적으로는 1,000개 이상의 데이터 포인트가 있어야 합니다.	Amazon Forecast 전용 딥 러닝 DeepAR+, CNN-QR

기준	Amazon Forecast 알고리즘 클래스
대부분의 시계열에 300개 이상의 데이터 포인트 + 계절 영향 + 관련 데이터가 있는 1~100s 시계열의 작은 데이터 세트	Prophet
대부분의 시계열에 300개 이상의 데이터 포인트 + 계절 영향이 있는 1~10s 시계열의 작은 데이터 세트	ETS, ARIMA
대부분의 시계열에 300개 이상의 데이터 포인트가 있는 1~10s 시계열이 있는 간헐적(0이 많은 희소)	Amazon Forecast 전용 NPTS
대부분의 시계열에 300개 미만의 데이터 포인트가 있는 1~10s 시계열의 작은 데이터 세트(일반 또는 희소)	Amazon Forecast에 비해 데이터가 너무 작습니다. 대신 Excel에서 ETS를 사용하거나 기존 통계 모델인 ARIMA 및 Prophet을 사용해 보세요.

가장 좋은 방법은 데이터를 처음으로 분석할 때 Predictor에서 AutoML 모드를 사용하여 학습하는 것입니다. AutoML은 모든 알고리즘(HPO가 켜진 상태에서 실행되는 DL 알고리즘)을 자동으로 실행하여 데이터에 가장 적합한 알고리즘을 학습합니다.

Q: 누락된 데이터에 대해 어떻게 대응해야 하나요? 합리적인 예측을 하기에 너무 많은 시간이 필요한 때는 언제인가요?

데이터 기록에 문제가 있거나 데이터의 집계 수준이 너무 낮거나 너무 높은 경우일 수 있습니다. 일반적인 규칙은 예측 길이가 학습 데이터의 1/3보다 길 수 없다는 것입니다.

누락된 데이터의 양 외에 또 다른 고려 사항은 누락된 데이터의 대체입니다. 0을 모두 null로 변환하고 Amazon Forecast에서 누락된 값을 자동으로 입력하도록 할 수 있습니다. Amazon Forecast는 신제품 출시(콜드 스타트) 또는 수명이 다한 제품으로 인해 누락된 값이 발생하는지 여부를 자동으로 감지합니다. 값, 중앙값, 최소값, 최대값, 0, 평균 및 nan(대상 시계열만 해당)을 비롯한 여러 누락값 로직을 사용할 수 있습니다. null 채우기 구문에 대한 설명서를 참조하세요.

- 'frontfill' - (TTS만 해당) 신제품 또는 콜드 스타트 제품 및 해당 제품에 기록이 생기기 전에 null을 처리하는 방법을 나타냅니다.
- 'middlefill' - 시계열 값 중간에 있는 null을 나타냅니다.
- 'backfill' - 수명이 다한 제품과 제품 판매가 중단된 후 null을 처리하는 방법을 나타냅니다.

- 'futurefill' - (RTS만 해당) 학습 데이터가 끝난 후 발생하는 null을 나타냅니다.

Q: 입력된 과거 데이터에는 음수 값이 없는데 수요 예측에 음수 값이 표시됩니다. 이런 현상이 발생하는 이유는 무엇인가요? 이 문제를 방지하려면 어떻게 해야 하나요?

NPTS(음수가 아닌 데이터에서 학습됨) 및 DeepAR(음-이항 가능성 함수 사용)을 제외한 모든 모델의 경우 양수 생성을 보장할 수 없습니다. 해결책은 앞서 언급한 모델 중 하나로 변경하거나 예측 값을 음수가 아닌 값으로 자르는 것입니다.

Q: 정확도 지표가 분위에 따라 다른 이유는 무엇인가요? 모델이 같으니 오차도 같아야 하지 않나요?

가중치가 분위에 따라 어떻게 달라지는지에 대한 자세한 설명은 가중 분위 손실(wQL)을 참조하세요.

p10, p50, p90이라는 세 가지 다른 분위에 모든 예측이 있다고 가정해 봅시다. 세 예측 자체는 랜덤 변수입니다. 정확도는 각 분위 수준에서 실제와 예측치 간에 별도로 계산됩니다. 아래에서 'WQL', 즉 가중 분위 손실 표를 볼 수 있습니다. wQL 값은 서로 결정적인 관계가 없습니다. (리콜 손실은 오류를 의미하므로 순서가 지정되지 않습니다. 하지만 분위 예측은 순서대로 정렬됩니다.) 따라서 예를 들어 p90 wQL이 p50 wQL보다 커야 할 이유가 없습니다.

표 3 - 예측 분위 예제

	A	B	C
1	P10 wQL	P50 wQL	P90 wQL
2	0.18647	0.50879	0.30428

Q: 예측 정확도를 높이려면 어떻게 해야 하나요?

예측 정확도는 적절한 데이터를 적절한 양과 품질로 사용할 수 있는지 여부에 따라 달라집니다. 정확도가 만족스럽지 않다면 문제가 얼마나 예측 가능한지(또는 데이터가 얼마나 랜덤인지/노이즈가 있는지/고정적인지) 이해하는 것이 합리적일 수 있습니다. 고려해야 할 다른 요소로는 다양한 모델 평가, 하이퍼파라미터 설정, 관련 시계열 및 항목 메타데이터 데이터 세트를 사용한 추가 기능 통합 등이 있습니다. 구체적인 제안 사항은 GitHub의 이 모범 사례 문서를 참조하세요.

Q: 제 사용 사례에 잘 맞는 알고리즘이 있는데 Amazon Forecast에서는 제공되지 않습니다. 어떻게 해야 하나요?

Amazon Forecast 팀이 이 사용 사례에 대해 기꺼이 도와드리겠습니다. 이메일을 통해 Amazon Forecast의 서비스 팀에 문의하세요<amazonforecast-poc@amazon.com>.

부록 B: 참고 문헌

Januschowski, Tim and Kolassa, Stephan. A classification of business forecasting problems(비즈니스 예측 문제의 분류). Foresight: The International Journal of Applied Forecasting(Foresight: 국제 응용 예측 저널). 2019년

Salinas, David, Flunkert, Valentin, Gasthaus, Jan and Januschowski, Tim. DeepAR: Probabilistic forecasting with autoregressive recurrent networks(DeepAR: 자기 회귀 순환 네트워크를 사용한 확률적 예측). International Journal of Forecasting(국제 예측 저널). 2019년

Gasthaus, Jan, Benidis, Konstantinos, Wang, Yuyang, Rangapuram, Syama Sundar, Salinas, David, Flunkert, Valentin and Januschowski, Tim. {Probabilistic Forecasting with Spline Quantile Function RNNs(스플라인 분위수 함수 RNN을 사용한 확률적 예측). The 22nd International Conference on Artificial Intelligence and Statistics(인공 지능 및 통계에 관한 국제 컨퍼런스 제22차). 2019년

Januschowski, Tim, Gasthaus, Jan, Wang, Yuyang, Salinas, David, Flunkert, Valentin, Bohlke-Schneider, Michael and Callot, Laurent. Criteria for classifying forecasting methods.(예측 방법을 분류하는 기준). International Journal of Forecasting(국제 예측 저널). 2019년 (로그인 필요)

Januschowski, Tim, Gasthaus, Jan, Wang, Yuyang, Rangapuram, Syama and Callot, Laurent. Deep Learning for Forecasting(예측을 위한 딥 러닝). Foresight: The International Journal of Applied Forecasting(Foresight: 국제 응용 예측 저널). 2018년

Januschowski, Tim, Gasthaus, Jan, Wang, Yuyang, Rangapuram, Syama Sundar and Callot, Laurent. Deep Learning for Forecasting: Current Trends and Challenges(예측을 위한 딥 러닝: 현재 동향 및 과제). Foresight: The International Journal of Applied Forecasting(Foresight: 국제 응용 예측 저널). 2018년

Bose, Joos-Hendrik, Flunkert, Valentin, Gasthaus, Jan, Januschowski, Tim, Lange, Dustin, Salinas, David, Schelter, Sebastian, Seeger, Matthias and Wang, Yuyang. Probabilistic demand forecasting at scale(규모에 따른 확률적 수요 예측). Proceedings of the VLDB Endowment(VLDB Endowment 회의록). 2017년

문서 기록

이 백서의 업데이트에 대한 알림을 받으려면 RSS 피드를 구독하세요.

변경 사항	설명	날짜
백서 업데이트됨	업데이트	2021년 9월 1일
최초 게시	백서를 처음 게시했습니다.	2020년 2월 4일

> **Note**
>
> RSS 업데이트를 구독하려면 사용 중인 브라우저에서 RSS 플러그인이 활성화되어 있어야 합니다.

고지 사항

고객은 본 문서에 포함된 정보를 독자적으로 평가할 책임이 있습니다. 본 문서는 (a) 정보 제공만을 위한 것이며, (b) 사전 고지 없이 변경될 수 있는 현재의 AWS 제품 제공 서비스 및 사례를 보여 주며, (c) AWS 및 자회사, 공급업체 또는 라이선스 제공자로부터 어떠한 약정 또는 보증도 하지 않습니다. AWS 제품 또는 서비스는 명시적이든 묵시적이든 어떠한 종류의 보증, 진술 또는 조건 없이 '있는 그대로' 제공됩니다. 고객에 대한 AWS의 책임과 법적 책임은 AWS 계약서에 준하며 본 문서는 AWS와 고객 간의 계약에 포함되지 않고 계약을 변경하지도 않습니다.

© 2021 Amazon Web Services, Inc. 또는 자회사. All rights reserved.

AWS 용어집

최신 AWS 용어는 AWS 일반 참조서의 AWS 용어집을 참조하세요.