대본

[모델 검증: Custom Loss Function]

먼저, 저희는 LG 해커톤의 리더보드의 기준 및 목표에 맞추어 Loss Function을 수정했습니다.

이를 통해, 저희의 모델은 상대적으로 높은 판매량을 보이는 제품들에 대한 예측 성능을 증가시킬 수 있었고, [1]을 할 수 있었습니다.

해당 Custom Loss Function은 여러 단계를 통해 계산됩니다. 첫 단계에서는 예측 값과 실제 값의 차이의 절대값을 계산합니다. 그 다음에는 예측 값과 실제 값 중 큰 값을 선택하여 해당 값으로 나누어차이 값을 정규화 합니다.

그리고 특정 일에 대한 모든 제품의 판매량 합계를 계산하여, 이를 통해 각 제품의 판매량 비중을 구합니다. 이 과정에서 아주 작은 값을 분모에 추가함으로써 분모가 0이 되는 것을 방지하였습니다.

마지막으로, 판매량 비중과 정규화된 오차값을 이용하여 최종 손실 값을 계산합니다. 이렇게 계산된 손실 값은 판매량이 큰 제품의 예측 오차가 작은 제품의 예측 오차보다 더 큰 영향을 미칩니다. 즉, 판매량이 많은 제품에 대한 예측의 정확도가 높아야 모델의 성능이 좋다고 평가될 수 있습니다.

예를 들면, 한 제품의 실제 판매량이 1000개이고 예측값이 1100개인 경우, 그 차이는 100개입니다. 반면 다른 제품의 실제 판매량이 10개이고 예측값이 20개라면 차이는 10개입니다. 이 경우, 절대적인 차이만을 고려하면 두 제품 모두 예측이 잘못되었다고 판단할 수 있지만, 저희의 Custom Loss Function을 통해 볼 때, 첫 번째 제품의 예측 오차가 두 번째 제품의 예측 오차보다 더 중요한 것으로 간주합니다.

[모델 알고리즘: LSTM]

저희는 DACON에서 제공한 기본 LSTM의 구조를 변경했습니다.

기존 모델에서는 제품 고유 특성을 먼저 학습한 후에, 시계열 데이터만으로 학습을 진행하기 때문에, 제품의 특성에 따른 판매량 변동성을 충분히 학습하지 못할 가능성이 있었습니다.

그래서 각 시점마다 제품의 고유 특성과 판매량 데이터를 함께 학습할 수 있도록 모델에 입력하는 학습 데이터의 구조를 변경했습니다. 그 결과, 제품의 특성에 따라 판매량의 변동성을 더 정확하게 학습하였고, 예측 성능이 향상 되었습니다.

결과적으로, 각 시점마다 제품의 고유한 특성과 판매량 데이터를 함께 학습하는 것이, 단순히 시계열 데이터만을 사용하여 학습하는 것보다 더 효과적임을 확인할 수 있었습니다.

[모델 알고리즘: 변수 및 파라미터 설명]

다음으로 모델에 학습시킬 변수와 파라미터를 설정했습니다.

LSTM 시계열 학습에서 "Window Size"는 모델이 입력으로 받는 연속적인 시계열 데이터 기간을 의미합니다. "Predict Size"는 모델이 예측해야 하는 미래의 시점의 수로, 여기서는 21일 입니다. "Step Size"는 Window가 얼마나 이동하는지 나타내며, 저희는 21로 설정했습니다.

Step Size를 21로 설정한 이유는, 과적합을 줄이는 데 도움이 될 수 있다고 판단했기 때문입니다. 각 학습 샘플이 겹치지 않고 독립적인 예측을 수행하게 되어, 모델이 훈련 데이터에 대해 과도하게 최적화되지 않도록 했습니다.

또한, 모델은 충분한 과거 정보를 바탕으로 예측을 수행해야 합니다. Window Size가 너무 길어지면 계산 복잡도가 증가하고, 긴 Sequence의 경우에는 기울기 소실(vanishing gradient)이 발생하여 모델 성능이 하락할 수 있습니다. 그래서 저희는 알맞은 "Window Size"를 선택하기 위해서, 동일한 모델에서 Window의 변화에 따라 성능의 변화를 확인하고 최적의 Window Size를 선택했습니다.

[모델의 실용성 및 활용 가능성]

본 모델은 1 epoch당 학습 속도가 매우 빠르게 설계되었습니다. 이로 인해, 대량의 데이터가 유입되었을 때에도 모델은 이를 신속하게 받아들이고 처리할 수 있습니다.

가벼운 모델 구조 덕분에, 빠른 학습 및 예측이 가능하며, 이는 실시간 환경에서의 활용성을 크게 향상시킵니다.

신속한 처리 능력은 사용자에게 즉각적인 피드백을 제공할 수 있게 하여, 사용자 경험을 풍부하게 만듭니다.

[실제 현업에서의 적용 가능]

본 모델은 평가 데이터의 30%에 대하여 0.57178의 성능을 보였으며, 평가 데이터 전체 100%에 대해서는 0.57898의 성능을 나타냈습니다. 이 결과는 모델이 다양한 데이터 범위와 상황에도 안정적인 성능을 유지할 수 있음을 보여줍니다.

이는 모델이 대규모 데이터셋에서도 성능의 안정성을 유지하며 활용할 수 있음을 의미합니다. 특히, 데이터의 양이 다양할 때도 성능의 일정한 수준을 보장하므로, 실제 업무에서의 다양한 상황과 요구사항에 유연하게 대응할 수 있습니다.