**HD현대 대본**

**PPT 1**

안녕하세요 저는 경기대학교 컴퓨터과학과 석사과정 이상민입니다

문제 해결을 위한 솔루션 발표를 시작하겠습니다.

**PPT 2**

우선 저는 하나의 학습 데이터를 통해 시각화를 진행하였습니다.

시각화 결과 동일한 경향을 보이는 특성들을 확인할 수 있었고,

시계열 데이터이기 때문에 통계적인 세밀한 분석을 수행하였습니다

**PPT 3**

특성 별 기술 통계 분석은 통계치와 상관관계를 분석하였으며,

표 1에는 학습 데이터 16개 전체 값을 합쳐 계산하였습니다.

**PPT 4**

다음으로는 앞선 통계치를 기반으로 하는 분석 내용입니다.

시간은 다른 특성에 비해 큰 범위와 높은 표준 편차를를 가지고 있기에 데이터 포인트가 평균 값과 멀리떨어져있음을 확인하였으며, 왜도와 첨도를 통해 정규 분포보다 더 넓은 범위를 가지며, 높은 분포를 가진것을 확인하였습니다.

또한 시그널 A와 B는 작은 표준편차를 가지고 있으며, 평균 주위에 데이터 포인트가 밀집되어 있음을 확인하였습니다.

시그널 C가 시그널 중 가장 낮은 표준 편차를 가지고 있었으며, IQR값이 0으로 중앙값에 상당히 밀집되어 있음을 확인하였으나, 다른 시그널에 비해 첨도가 높았기에 분포가 상당히 높으며, 이상치가 존재할 가능성이 높습니다.

마지막으로 센서 값들중 A와 D는 표준편차가 다른 센서보다 높아 더 넓은 범위를 가고 있었으며, B와 C는 왜도가 1이상으로 분포가 오른쪽에 높은 값의 데이터 포인트가 존재하는것을 확인하였습니다.

**PPT 5**

다음으로는 특성간 상관관계를 분석하였습니다.

시간은 센서 B와 양의 상관관계를 그리고 시그널 A 센서 C와는 음의 상관관계를 보여 시간에 따라 3가지 특성의 증감 현상을 확인하였고,

시그널 A는 시그널 B, C, D와 양의 상관관계를 보임으로써 A의 변화가 다른 시그널 전체에 영향을 미치는 것을 확인하였습니다. 또한 시그널 B 또한 A와 유사한 관계를 보였습니다.

센서값들의 관계는 A와 B는 약한 상관관계를 보이나, 센서 C와 D는 중간 정도의 상관관계를 보임으로써 서로 유사한 관계를 가짐을 확인하였습니다.

**PPT 6**

이러한 전체적인 특성 별로 통계치를 반영하기 위해서 딕셔너리를 정의하여 전체 데이터에 반영하였으며,

**PPT 7**

앞에서 분석된 통계치 분석 내용과 상관관계를 기반으로 하여 통계치가 반영된 특성 값들 마다 다른 스케일링을 적용하였으며,

시그널은 A의 값의 변화가 다른 시그널에 영향을 미치기에 4개의 특성에 동일한 스케일링 방법을 적용하였으며, A, B, C가 작은 범위와 왜도를 가지고 있기에 민맥스 스케일링을 통해서 모든 특성이 동일한 스케일을 갖도록 하였습니다.

그리고 시간의 경우 다른 특성들의 증감 현상이 존재하기에 상관관계를 고려한 스케일링을 적용하기 어렵다고 판단하여, 시간 속성의 통계 분석 결과인 큰 범위와 높은 표준 편차를 갖기에 표준 정규화를 통해 다른 특성들과의 균형을 맞추었습니다.

마지막으로 센서 값들의 경우, 모두 양의 상관관계를 가지고 있었기에 하나의 스케일링 방법을 적용하였으며, 센서 C와 D의 경우 왜도가 높아 이상치의 영향을 많이 받을 수 있기에 중앙 값과 사분위 범위를 활용하여 로버스트 스케일링을 적용하여 이상치에 견고하게 하였습니다.

**PPT 8**

다음으로는 학습에 사용된 모델입니다.

저는 학습에 시계열 트랜스포머 모델을 활용하였으며, 해당 모델은 트랜스포머의 인코더 구조만을 사용하며, 포지셔널 인코딩을 활용하여 시계열 데이터의 시간적 특성을 활용합니다.

해당 모델은 다양한 테스크에 활용가능하다는 장점을 갖기에, 해당 대회의 문제처럼 시계열 데이터를 기반으로 하여 시계열 그래프를 예측하는 것이 아닌, 작업 중량을 예측하는 등의 문제에 효과적일 것이라고 판단하였습니다.

추가적으로 LSTM 모델과 D라이너 모델을 통해 실험을 진행하였으나, TST 모델보다 저조한 성능을 보였으며, 추가적으로 학습 과정에서의 loss를 확인하였으나, 제대로 학습이 이루어 지지 않는 것을 확인하였습니다.

**PPT 9**

다음으로는 학습 전략에 관해서 설명드리겠습니다.

먼저 첫번째로 커스텀 로스 함수를 정의하였으며, 해당 로스는 예측 결과가 목표값보다 클 경우 패널티를 부과하는 방향으로 학습을 진행하였으며, 패널티의 경우 20%를 부여하는 것이 학습 효과가 가장 뛰어남을 예측 성능을 통해 확인하였습니다.

**PPT 10**

두번째로는 딥러닝의 일반적으로 사용되며, 신뢰성있는 결과를 얻기 위한 K폴드 학습 방식을 통해 전체 학습을 진행하였습니다.

해당 방식을 통해 일반화 성능을 강화하였으며, 과적합 방지등의 효과를 꾀하였습니다.

**PPT 11**

마지막으로 예측 값 매핑 과정에 대해서 설명하겠습니다.

해당 과정을 통해 그림 11처럼 예측된 값들을 타겟으로 하는 값들로 후처리를 진행하였으며, 타켓 값 사이에 존재하는 값들의 경우 절대 값을 통하여, 반올림과 반내림을 수행하였습니다.

여기서 예측값이 0인 경우 102로 치환하였으며, 타겟 값의 범위에 0이 없는 것이 예측 성능을 증가시켰습니다.

**PPT 12**

마지막으로 결론에 대해서 말씀드리겠습니다.

저는 전체적인 데이터 분석을 수행하였으며, 통계적 분석과 특성간 상관관계를 분석하여 나온 결과들을 통해 전처리를 수행하였습니다.

그리고 학습 모델로 다양한 테스크에 적용 가능한 TST모델을 활용하여 학습을 진행하였으며, 학습 전략으로 오버래피드 MAE 로스 평션과 k-fold 방법과 예측 값 매핑을 통해 학습 과정을 진행하고 예측 값의 후처리를 진행하였습니다.

추가적으로 아쉬운점은 더 다양하고 최신 모델들을 통해 학습을 진행하지 못한 점과 하이퍼 파라미터 최적화를 수행하지 못한 점이 아쉬웠으며, 해당 방법들을 통해 더욱 성능을 높일 수도 있을것이라고 생각합니다.

이상으로 발표 마치겠습니다 감사합니다.