

2025학년도 2학기 기계학습의 이해

프로젝트 전체 보고서

제출일: 2025.12.12.(금)

소속: AI융합대학 Language&AI융합전공

학번: 202400353

이름: 최은율

1. Introduction

1.1 프로젝트 배경 및 문제 정의

F1 경기는 평균 2시간이 소요되며, 시차로 인해 한국에서는 주로 새벽 시간에 중계됩니다. 바쁜 현대인이 모든 경기를 생중계로 챙겨보는 것은 현실적으로 어렵습니다. 또한 경기 결과만으로는 그 과정이 얼마나 치열했는지 알 수 없기 때문에, 놓친 경기에 대해 "풀타임으로 다시 볼 가치가 있는가?"를 판단하기 어렵습니다. 이에 따라 경기 내용 자체가 얼마나 역동적이었는지를 데이터로 확인하고, [전체 관람 / 하이라이트 / 요약본] 중 최적의 시청 방식을 제안하여 효율적인 시청을 돋는 AI 서비스를 기획했습니다.

1.2 프로젝트 목표

본 프로젝트의 목표는 F1 경기 데이터를 분석하여 '재미'를 정량화 및 예측하는 머신러닝 모델을 개발하고, 이를 실제 웹 서비스로 구현하여 사용자에게 실질적인 시청 가이드를 제공하는 것입니다.

2. Process

2.1 데이터 수집 및 Feature Engineering

FastF1 라이브러리를 활용하여 2021~2025년 시즌의 모든 레이스 데이터를 수집했습니다. 단순한 기록 데이터만으로는 경기의 역동성을 설명하기 어려워, 도메인 지식을 활용한 파생 변수를 설계했습니다.



한국외국어대학교
HANKUK UNIVERSITY OF FOREIGN STUDIES

- Chaos Score (혼란도): 경기의 흐름을 바꾸는 변수들에 가중치를 부여하여 수치화했습니다.
 - $Chaos\ Score = SC \times 3 + Red\ Flag \times 5 + VSC \times 1$
- Position Gains: 드라이버들의 출발 그리드 대비 최종 순위 상승 횟수를 집계하여 추월 활성도를 반영했습니다.

데이터 분석 결과, 'Chaos_Score'과 'Position_Gains'의 상관계수는 0.22로 낮게 나타났습니다. 이는 사고가 많다고 해서 반드시 치열한 배틀이 일어나는 것은 아님을 시사하며, 두 요소를 독립적인 변수로 활용했습니다.

2.2 모델링 및 학습 전략

F1 경기 데이터는 변수 간의 관계가 매우 복잡하고 비선형적입니다. 또한 전체 데이터셋의 크기가 약 120개 내외로 작아 딥러닝 모델을 적용하기에는 Overfitting 위험이 커집니다. 이에 따라 1) 비선형적 상호작용 포착에 유리하고, 2) 적은 데이터에서도 안정적인 성능을 보이며, 3) 변수 중요도(Feature Importance)를 통해 예측의 근거를 해석할 수 있는 Random Forest Regressor를 최종 모델로 선택했습니다.

- 학습 데이터 분할 (Expanding Window): 시계열 데이터의 특성을 고려하여, 미래의 정보를 미리 보지 않도록 데이터를 순차적으로 누적하며 검증하는 Expanding Window 기법을 적용했습니다. 데이터셋의 크기가 작기 때문에 과거의 모든 정보를 최대한 활용하여 모델의 일반화 성능을 높이고자 했습니다. (Train: 2021~2024년 누적 / Test: 2025년)

초기 학습 단계에서는 데이터 부족으로 인한 과소적합 경향이 관찰되었습니다. 이를 해결하기 위해 Warm Start 기법을 활용하여 n_estimators를 점진적으로 늘려가며 학습 곡선(Learning Curve)을 분석했습니다. 그 결과, 데이터가 누적될수록 검증 성능이 우상향하며 안정화되는 패턴을 확인했습니다. 또한 변수 중요도 분석 결과, 도메인 지식을 바탕으로 설계한 Chaos_Score와 Lead_Changes(선두 교체)가 재미 예측에 가장 결정적인 영향을 미치는 변수임이 입증되었습니다.

2025년 테스트 데이터셋에 대한 최종 성능은 다음과 같습니다.

- R2 Score: 0.249 – '재미'라는 주관적 영역의 높은 분산으로 인해 설명력 자체는 높지 않았습니다.

- MAE (평균 절대 오차): 1.04 (10점 만점 기준) – 그러나 실제 평점과 예측값의 오차가 평균 1점 내외라는 것은, 사용자에게 "볼 만한 경기인가?"를 제안하는 추천 시스템의 목적(Classification)으로서는 충분히 유효한 성능이라고 판단했습니다.

3. Service Architecture

학습된 모델을 Streamlit 프레임워크를 사용하여 실제 사용 가능한 웹 서비스로 구현했습니다. 사용자는 두 가지 모드로 서비스를 이용할 수 있습니다.

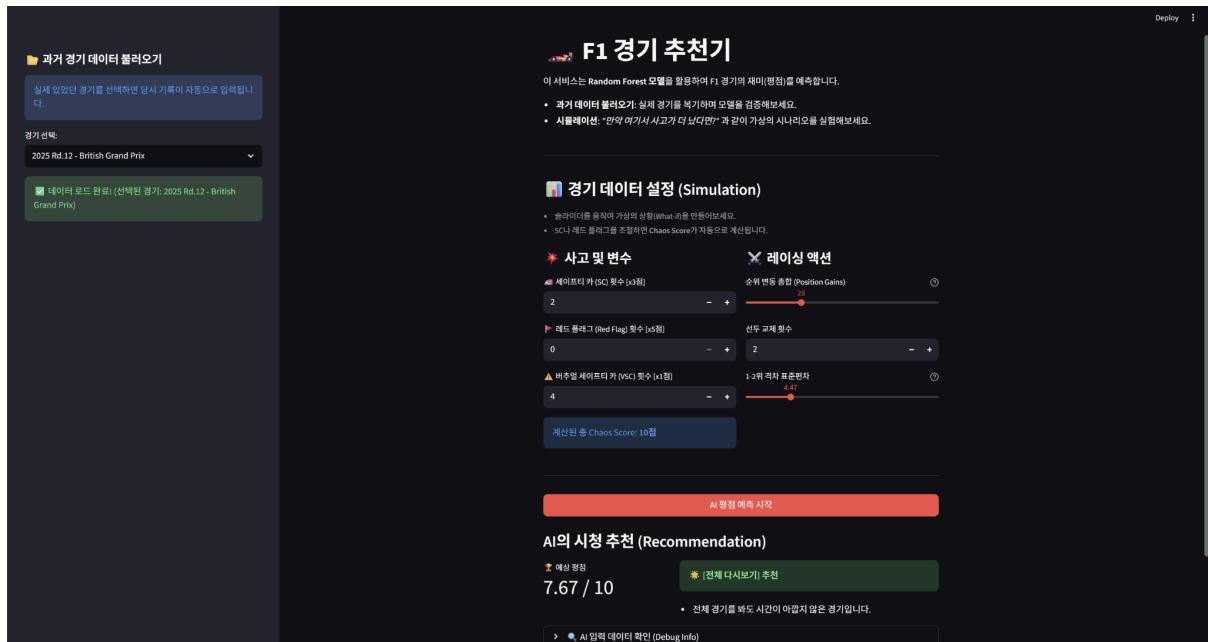
1. Preset Mode (과거 데이터 검증): 2021~2025년 실제 경기 데이터를 불러와 모델의 예측 결과를 검증합니다.
2. Simulation Mode (가상 시나리오): 슬라이더를 통해 사고 횟수, 추월 빈도 등을 조절하며 "만약 사고가 더 발생했다면?"과 같은 가상 시나리오(What-if)를 실험합니다.

시스템은 사용자 입력을 받아 Scaling 후 모델에 전달하며, 예측된 점수(0~10점)에 따라 경기 시청 가이드라인(풀경기, 하이라이트, 요약본)를 시각적으로 출력합니다.

4. Real Usage Results

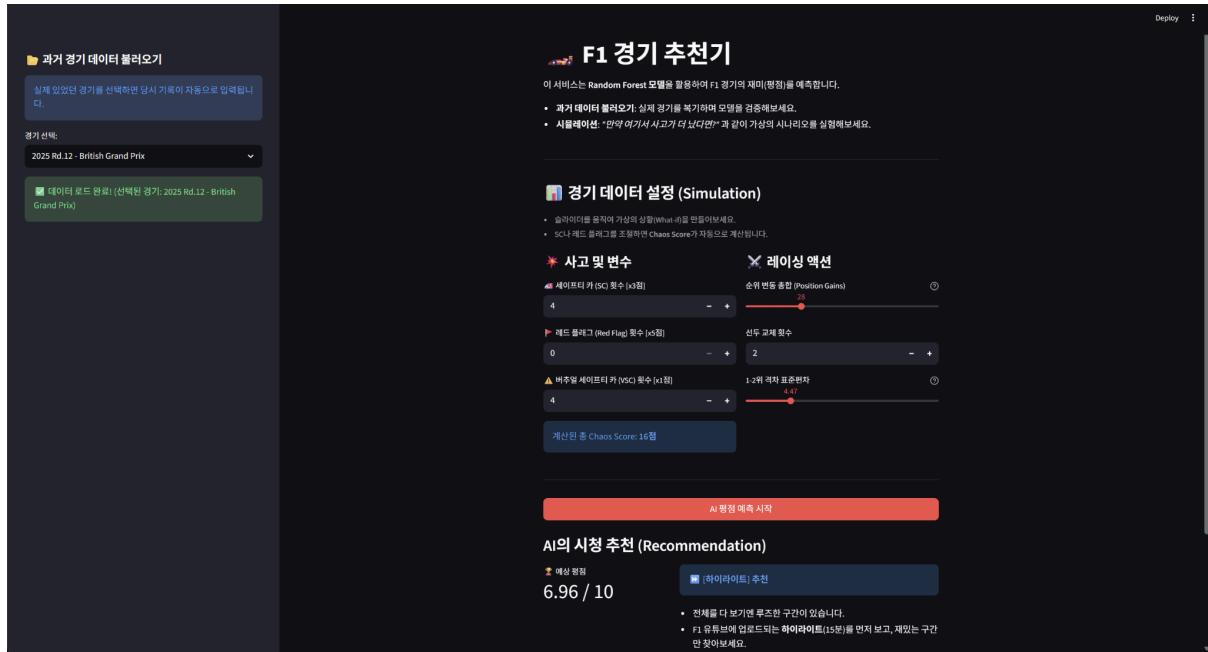
실제 서비스를 5회 이상 사용하며 프리셋 검증과 가상 시뮬레이션을 진행했습니다. 다양한 시나리오를 통해 모델이 경기의 맥락을 올바르게 이해하고 있는지 테스트했습니다.

4.1 2025 영국 GP (고득점)



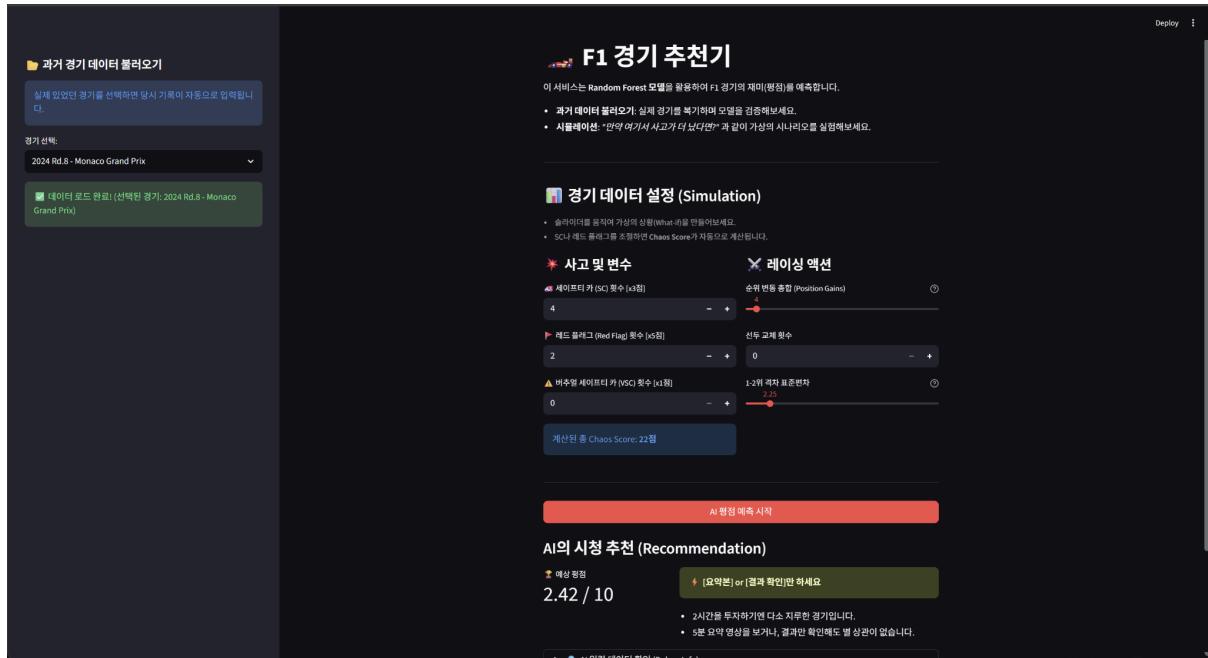
가장 먼저 모델의 고득점 예측 능력을 검증하기 위해 2025년 영국 그랑프리 데이터를 입력했습니다. 이 경기는 적절한 혼란(Chaos Score 10점)과 활발한 순위 변동(Position Gains 28회)이 어우러진 경기였습니다. 모델은 이를 고려하여 7.67점이라는 높은 평점을 부여했으며, 가이드라인에 따라 사용자에게 [전체 다시보기]를 추천했습니다. 이는 모델이 적당한 변수와 레이싱 액션의 조화를 '재미'로 잘 인식하고 있음을 보여줍니다.

4.2 2025 영국 GP 시뮬레이션



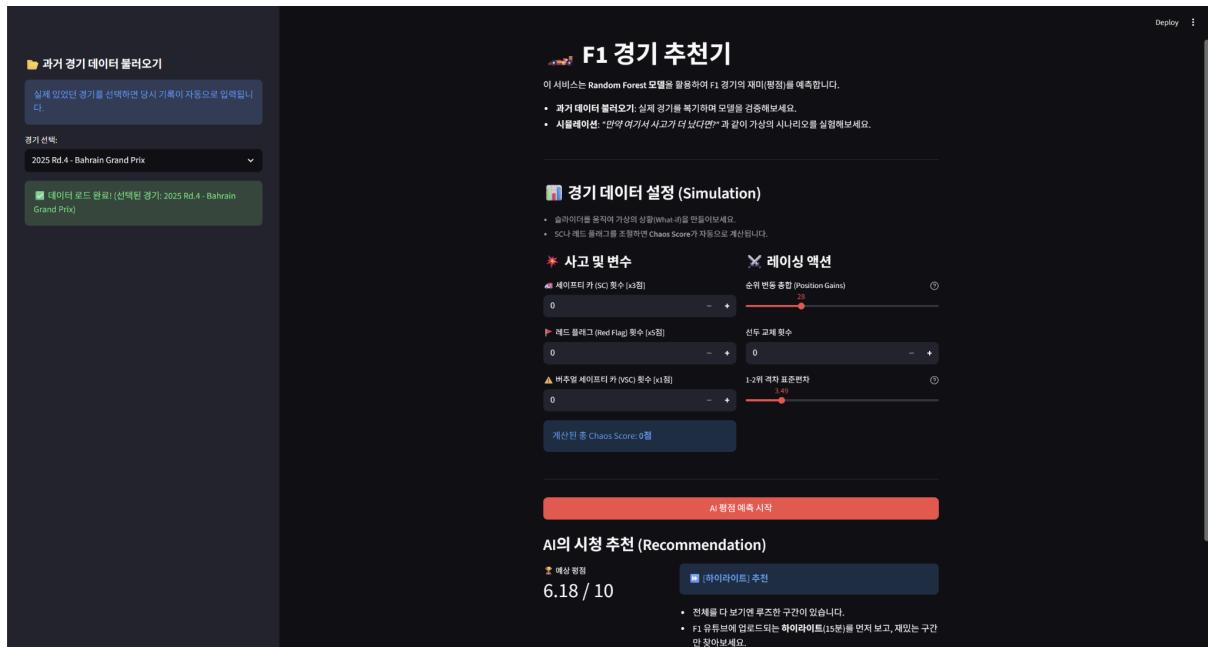
앞선 4.1의 영국 그랑프리(명경기) 설정에서 흥미로운 가설을 실험해 보았습니다. "사고가 더 많이 나면 더 재밌어질까?"라는 가정하에 세이프티 카(SC) 횟수를 2회에서 4회로 늘려보았습니다. 그 결과, 예상과 달리 평점은 7.67점에서 6.96점으로 오히려 하락했습니다. 이는 무조건적인 혼란이 재미를 보장하지 않으며, 너무 잦은 경기 중단은 오히려 흐름을 끊어 재미를 반감시킨다는 미묘한 패턴을 모델이 학습했음을 확인할 수 있습니다.

4.3 2024 모나코 GP (저득점)



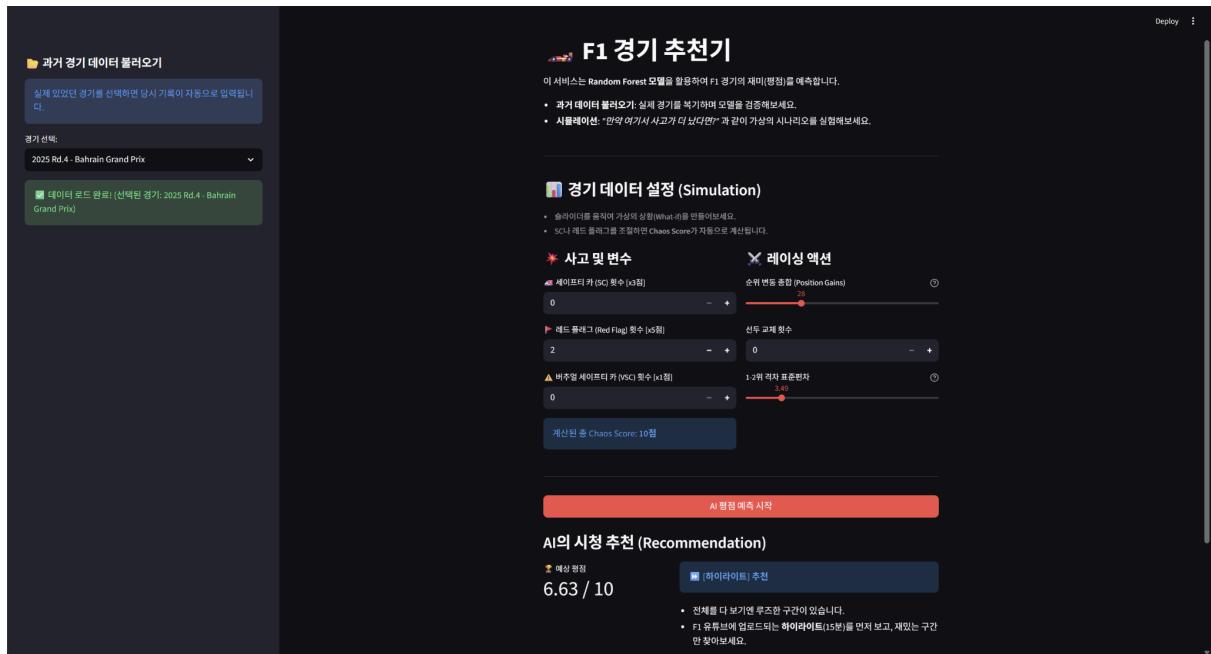
다음으로 모델의 변별력을 확인하기 위해, F1 팬들 사이에서 지루하기로 악명 높았던 2024년 모나코 그랑프리 데이터를 입력했습니다. 이 경기는 사고(Red Flag)로 인한 Chaos Score는 22점으로 다소 높았으나, 서킷의 특성상 추월이 적게 일어난 경기입니다. 모델은 높은 Chaos Score에도 불구하고, 추월이나 선두교체와 같은 레이싱 액션의 부재를 치명적인 감점 요인으로 판단하여 2.42점이라는 매우 낮은 점수를 예측했습니다. 모델이 단순히 Chaos Score가 높다고 높은 평점을 예측하지 않는다는 것을 확인할 수 있었습니다.

4.4 2025 바레인 GP



평범한 경기에 대한 예측을 테스트하기 위해 2025년 바레인 그랑프리 데이터를 불러왔습니다. 사고(Chaos)가 전혀 없는 깔끔한 경기였지만, 적당한 수준의 순위 변동(28회)이 있었던 케이스입니다. 모델은 이를 6.18점으로 예측하며 [하이라이트] 시청을 추천했습니다. 이는 전체를 다 보기에는 다소 루즈할 수 있지만, 주요 장면 위주로 즐기기에는 충분하다는 합리적인 판단을 내린 것으로 확인했습니다.

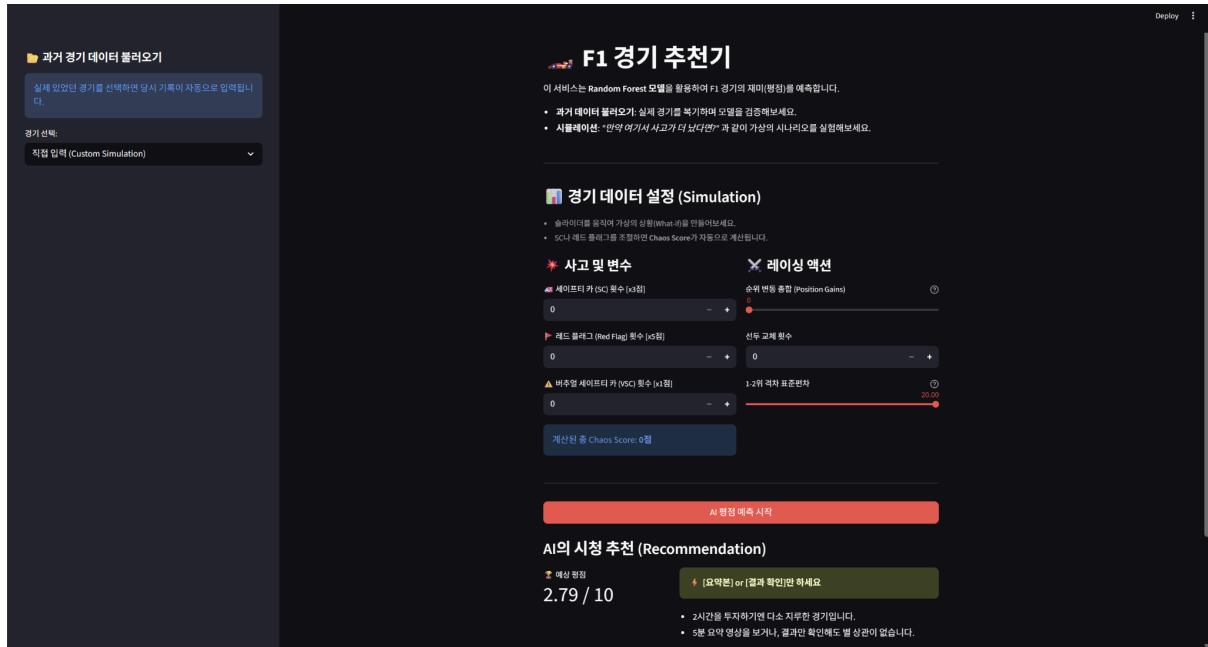
4.5 2025 바레인 GP 시뮬레이션



앞선 4.4의 바레인 GP(평점 6.18점)는 사고가 없는 깔끔한 경기였습니다. 여기에 "만약 사고가 발생해 변수가 생겼다면?"이라는 가정을 대입하여 Red Flag를 2회 추가해 보았습니다(Chaos Score 0점 → 10점). 그 결과, 예상 평점은 6.18점에서 6.63점으로 소폭 상승했습니다. 이는 4.3번 케이스와 대조되는 결과로, "적절한 수준의 혼란과 변수"는 경기의 긴장감을 높여 재미를 배가시키는 긍정적인 요소로 작용한다는 것을 모델이 학습했음을 보여줍니다.



4.6 최저점 테스트



마지막으로 모든 재미 요소를 극단적으로 설정한 최악의 상황을 시뮬레이션했습니다. 결과는 2.79점이었습니다. 흥미로운 점은 이 점수가 앞선 4.2의 모나코 GP(2.42점)보다 소폭 높다는 것입니다. 모나코 GP는 높은 Chaos Score에도 불구하고 추월이 거의 없었습니다. 즉, 모델은 경기가 자꾸 중단되어 흐름이 끊기는데, 추월은 없는 상황을 아무 사건 없이 매끄럽게 끝난 경기보다 더 시청 가치가 낮은 '최악의 경험'으로 판단했다고 볼 수 있습니다. 이는 모델이 단순히 수치를 더하는 것이 아니라, 변수 간의 부정적 상호작용(Negative Interaction)까지 학습했다고 볼 수 있습니다.

5. Conclusion

5.1 배운 점

첫째, 도메인 지식의 중요성입니다. 단순히 데이터를 모델에 넣는 것보다, F1의 특성을 반영해 직접 설계한 Chaos_Score가 평점 예측의 핵심 요인임을 확인했습니다. 둘째, End-to-End 개발 경험입니다. 데이터 수집부터 전처리, 모델 학습, 그리고 최종 사용자를 위한 웹 서비스 개발까지 머신러닝 프로젝트의 전 과정을 경험하며 실무적인 역량을 길렀습니다.

5.2 한계 및 개선 방향

현재 모델이 사용하는 Position_Gains 변수는 출발과 도착 순위의 차이(Net)만을 반영합니다. 이로 인해 경기 도중 치열하게 순위를 주고받은 횟수나, 피트 스톱 전략으로 인한 일시적 순위

변화(Gross)는 반영되지 않아 중위권의 치열한 배틀을 과소평가할 위험이 있습니다. 향후에는 랩(Lap)별 데이터를 초 단위로 분석하여 '실제 추월 시도 횟수'를 파생 변수로 추가한다면 이를 해결할 수 있을 것입니다.

또한, 데이터 누수를 막기 위해 Expanding Window split을 적용했으나, F1 데이터는 일반적인 시계열 데이터와 달리 과거의 패턴이 미래를 완벽하게 설명하지 못한다는 한계를 지닙니다. 매년 변경되는 차량 기술 규정과 드라이버 라인업 변경 등으로 인해 데이터의 연속성이 끊어지는 '구조적 변화'가 발생하기 때문입니다. 향후에는 단순한 과거 데이터의 누적보다는, 시즌별 규정 변화 가중치를 적용하거나 시즌 내 데이터에 집중하는 방식이 더 유효할 수 있다고 판단했습니다.

Appendix. F1 Terminology (용어 설명)

본 보고서와 프로젝트에서 사용된 F1 전문 용어에 대한 설명입니다.

Term	Description
Grand Prix (GP)	'그랑프리'라고 읽으며, 하나의 F1 대회(경기)를 의미합니다. (예: 영국 GP, 모나코 GP)
Safety Car (SC)	큰 사고 발생 시 트랙에 투입되는 실제 차량입니다. 모든 F1 차량이 SC 뒤로 줄을 서서 서행해야 합니다.
Virtual Safety Car (VSC)	작은 사고 시 발령되며, 실제 차량 투입 없이 전 구간 속도 제한이 걸립니다.
Red Flag	심각한 사고나 악천후로 경기가 일시 중단되는 상황입니다. 모든 차량이 피트(Pit)로 복귀해야 합니다.
Grid (Starting Grid)	예선(Qualifying) 성적에 따라 정해진 출발 순서입니다.
Position Gains	드라이버가 출발 순위(Grid)보다 얼마나 높은 순위로 경기를 마쳤는지를 나타내는 지표입니다.

