BIG3003, Spring 2022 Homework #08 is due on June 14, 2022

2016044875 류재원

1. 분석 데이터와 머신러닝 개요

본 과제에서 분석하게 될 데이터는 FIFA20 이라는 게임의 데이터입니다.

(https://www.kaggle.com/datasets/stefanoleone992/fifa-20-complete-player-dataset)

널리 잘 알려진 FIFA 게임 시리즈는, 실제 축구 선수들의 데이터를 기반으로 만들어진 축구 게임입니다. 세계 축구 협회에 등록된 수만 명의 선수들은 여러 종류의 스탯(1~100)을 부여받고, 실제 포지션(공격수, 미드필더, 수비수, 골키퍼), 보조적인 능력, 신체 조건 등을 통해 실제와 아주 유사하게 구현됩니다. 다음은 그 예시입니다.



|  |  |
| --- | --- |
|  | 위 그림은 전 세계적으로 유명한 리오넬 메시 선수의 데이터입니다. 나이, 신장, 소속 국가 등이 잘 나와있습니다. 한편, 왼쪽 그림은 리오넬 메시 선수의 FIFA15 내 스탯입니다. 전 세계적으로 유명한 선수 답게, 거의 대부분의 스탯이 80~90 인 것을 알 수 있습니다. 이처럼 FIFA 게임 시리즈 내 스탯 데이터를 바탕으로, 그 선수의 포지션을 분류(Classfication) 하려 합니다. 이를 통해 선수의 스탯에 따른 적합 포지션을 산출할 수 있을 것입니다. 예를 들어, 리오넬 메시 선수와 손흥민 선수는 같은 공격수 클래스로 분류될 것입니다. |

미래에 축구 선수를 지망하는 사람들 중에는 어떤 능력을 갖추어야 그 포지션에 적합한 선수가 될 수 있을지를 판단하는데 어려움을 겪는 사람들이 있습니다. 전 세계에서 가장 영향력 있는 FIFA 시리즈의 스탯으로 실제 축구 선수의 포지션을 분류할 수 있다면, 이들의 능력 향상을 돕고 축구 선수로써의 기량 향상에 도움을 줄 수 있을 것이라 생각합니다.

2. 머신러닝 실행

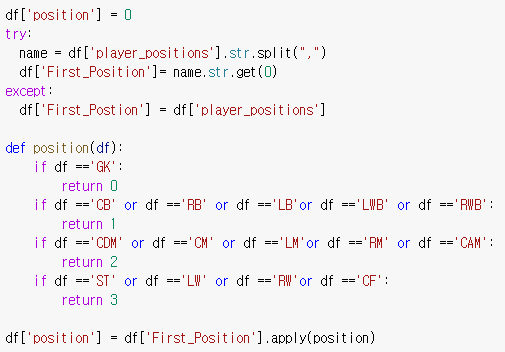
위의 Kaggle 웹사이트에서 데이터를 다운 받고, Python 언어 및 colab 플랫폼을 활용해 데이터 전처리 및 머신러닝을 진행하였습니다.

우선, 머신러닝을 위한 데이터를 불러왔습니다. Kaggle 웹사이트에서 다운받은 fifa15 ~ fifa20 의 데이터를 모두 활용하였으며, 총 데이터 개수는 11만개입니다.

테이블이(가) 표시된 사진

자동 생성된 설명

다음으로는, 종속변수를 구성하였습니다. 종속변수는 선수의 포지션이며, 각각 골키퍼 : 0, 수비수 : 1, 미드필더 : 2, 공격수 : 3 으로 분류하였습니다. 이 네 가지 클래스는 데이터 분석의 목표 변수입니다. 기존 position 컬럼을 대체하는 새로운 컬럼인 ‘position’을 생성하였습니다.



아래 그림은 위 과정을 통해 분류된 클래스의 현황입니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

미드필더가 약 4만명, 수비수가 약 3만명, 공격수가 2만명, 골키퍼가 1만명인 것을 알 수 있습니다. 이는 실제 축구에서 필요로 하는 포지션별 선수 수와 대략 일치합니다. 아래 그림은 실제 머신러닝에 활용될 피처, 반응변수를 추출한 것입니다. 이 피처들은 실제 fifa 게임에서 활용되는 데이터로 선수의 스탯에서 가장 기본적인 요소입니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

컬럼 중에는 포지션(공격수, 수비수 등)에 특화된 이름을 가진 것도, 그렇지 않은 것도 존재합니다. 대표적인 컬럼이 ‘power\_stamina’ 로, 어느 포지션에서 높은 스탯을 가질지 예측할 수 없습니다. 또한 종속 변수인 y를 정의하였습니다.

다음으로는 데이터 전처리를 진행하였습니다. 아래 그림은 리오넬 메시 선수의 데이터 중 일부입니다.



그림 내 특정 스탯에는 + 또는 – 가 존재하는 것을 알 수 있는데, 이는 fifa 시리즈 특유의 현실 반영 스탯 변화 시스템입니다. 이를 제거하기 위한 작업을 진행하였습니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

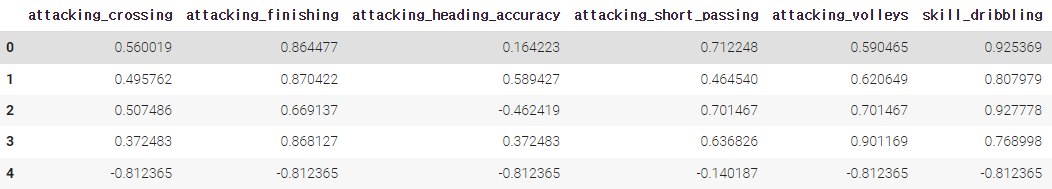
데이터 내 + 또는 – 가 포함되어 있을 경우, 앞쪽의 숫자만을 남기도록 전처리를 진행하였습니다. 아울러 기존의 문자열 형태를 수치 형태로 변경하였습니다.

탐색적 데이터 분석을 진행하던 중, 전처리가 필요한 요인을 발견하였습니다. 가령, 메시 선수는 전반적으로 우수한 선수이기 때문에 수비 스탯이 실제 수비수보다 높을 수도 있습니다. 이 경우 메시 선수가 수비수로 분류될 수도 있습니다. 따라서 이를 방지하고자 스탯의 ‘경향성’을 통한 상대적인 높낮이 비교를 도입하였습니다. 구체적으로는 각 row 별로 Standard-Scailing 을 통해 데이터를 전처리 하였습니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

아래는 최종적으로 확정된 독립변수의 예시입니다.



모든 데이터가 평균 0, 표준편차 1의 형태를 띄고 있는 것을 알 수 있습니다. 이를 통해 보다 정확한 분류가 가능할 것입니다. 다음으로 학습을 위해 데이터를 분류하였습니다.

분류 정확도 향상을 cross-validation 기법을 사용하고, 널리 잘 알려져 있는 RandomForest 방식을 사용하기로 하였습니다. RandomForest는 분류와 회귀 모두에 적합한 머신러닝 기법으로, 위 데이터처럼 연속형 데이터이나 결과값이 이산형 데이터인 상황에서 적합합니다. 다음과 같이 5번의 교차 검증을 진행해 평균값을 도출하였습니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

약 94%의 우수한 성능이 도출되었습니다. 이는 스탯을 기반으로 포지션을 약 94% 정확히 분류해 냈음을 의미합니다. 다음은 위와는 별개로 임의로 데이터를 분리하여 머신러닝을 진행한 후, f1-score를 산출한 결과입니다.

테이블이(가) 표시된 사진

자동 생성된 설명

테스트 데이터를 따로 분류하여 테스트를 진행한 결과, 마찬가지로 94%의 정확도를 기록하였습니다. 다음은 테스트 데이터를 활용한 예측 결과와 실제값 사이의 혼동행렬표입니다.

|  |  |
| --- | --- |
|  | 왼쪽의 혼동행렬표를 보면, 골키퍼 포지션을 오분류한 경우는 없었으나 수비수-미드필더, 미드필더-공격수를 오분류한 경우가 많았습니다. 이를 보완하면 더 좋은 모델을 구성할 수 있을 것입니다. |