이수강

1999년 | 남 | 서울 마포구

010-3772-9916
swangle2100@gmail.com

자기소개서

안녕하세요. 귀사의 데이터 분석가 직무에 지원하게 된 이수강입니다. 홍익대학교 경영학전공 졸업유예생이고, 2025년 2월 졸업 예정입니다. 4학년 1학기 "비즈니스프로그래밍2" 과목과 "빅데이터기반경영 / 사업타당성분석" 과목을 수강하면서 데이터 분석 분야를 접하게 되었고, 학기 중에 팀 프로젝트로 "한국 프로야구 관중수 예측 모델"을 만들어 보면서 데이터 분석이 적성에 맞는다는 것을 알게 되었습니다. 학기가 끝난 이후에도 후속 연구 제의를 받아 프로야구 관중수 예측 모델을 발전시켜 논문을 작성해 학회지에 게재하는 등 데이터 분석 능력을 키우고 있습니다.

머신러닝을 활용한 팀 프로젝트인 "한국 프로야구 관중수 예측 모델" 연구를 소개해 드리겠습니다. 해당 분석은 팀원이 모기업의 지원이 없으면 운영이 안되는 한국 프로야구 구단의 수익 구조 문제를 제기해 주었고, 제가 그를 해결하기 위한 방안으로 미국 MLB에는 이미 대부분의 구단이 활용하고 있는 티켓 가격 변동 전략을 제안하면서 시작되었습니다. 티켓 가격 변동 전략을 활용하기 위해서는 관중수 예측을 잘할 수 있어야 한다는 게 팀의 생각이었습니다. 2022년부터 2023년의 정규리그 1440경기 데이터를 활용했습니다. 요인은 기존 연구들에서 참고한 날씨, 요일, 순위, 기대득점, 승률, 선발 투수, 그리고 독자적으로 생각해낸 디시인사이드 구단별 게시글 조회수까지 수집하여 선형 회귀 모델을 제작했습니다. 그 결과 R-Squared 값 0.8 이상을 가지고, RMSE 값 2400 대의 꽤나 성능이 좋은 모델을 만들 수 있었습니다. 하지만 기존 연구들에서 등장한 거의 모든 독립변수를 사용하는 바람에 무거운(다차원의) 모델이 만들어졌습니다. 특히 홈, 원정 팀 선발 투수를 인코딩 하는 과정에서 차원이 급격하게 증가하여 차원의 저주 문제가 발생하여 선발 투수 요인을 제거할 수밖에 없었습니다. 학기가 끝나고, 수업을 담당하셨던 교수님들의 제의를 받아 기존 연구 주제 그대로 후속 연구를 하게 되었습니다. 후속 연구의 목적은 학술지 게재가 목적이었기 때문에 조금 더 요인 선택에 신중할 필요가 있었습니다. 기존의 관중 수 예측 연구들에서 공통적으로 사용된 요인들을 임의로 선택하였고, 임의성을 완화하기 위해 Feature Selection 기법 중 미리 선택할 요인의 수나 퍼센트를 정해두지 않고 최저 선택 요인 수만 선택하면 되는 Recursive Feature Elimination 기법을 선택하여 요인을 선택했습니다. 미세먼지 농도나 연승/연패 여부가 RFE 과정에서 탈락하여 기존 numerical 데이터에서 미세먼지의 경우 환경부 기준 미세먼지 상황 분류인 "좋음, 보통, 나쁨, 매우 나쁨"의 카테고리 데이터로 바꾸었고, 연승/연패의 경우에도 이상치로 판단되는 3시그마 이상의 값중 최저치인 6연승/연패와 연승/연패의 시작인 2연승/연패의 중간인 4연승/연패 이상 여부 데이터로 바꾸어서 RFE를 다시 실행했습니다. 그 결과 미세먼지 매우 나쁨 여부, 4연승/연패 이상 여부 카테고리가 채택되어 기존에 의도한 대로 기본 모델을 설계할 수 있었습니다. 그리고 기존 연구들과의 차별점이었던 "디시인사이드 구단별 조회수"가 논문에 게재되기에 부적절한 데이터 출처라는 지도 교수님의 지적이 있었습니다. 이를 개선하고자 2022년 ~ 2024년의 기간 동안 네이버 데이터 랩에서 각 구단별로 구단 이름을 검색한 값을 가져와서 각 경기에 홈 키워드 스코어, 원정 키워드 스코어라는 요인을 추가했습니다. 그 결과 R-Squared 값이 기존 모델 0.6에서 키워드 모델 0.65로 상승하였고, 선형 방정식의 요인 계수 중 홈, 원정 키워드가 다른 요인을 모두 제치고 가장 중요한 요인으로 채택되는 연구 결과에 도달했습니다.

내공이 부족한 상황에서 기술적, 기획적인 측면에서 주도적인 역할을 담당했기 때문에 프로젝트를 진행하면서 공부가 많이 됐습니다. 기술적인 부분을 도맡아서 연구를 진행한다는 점, 특히 처음 알게 된 기법을 바로 프로젝트에 사용한다는 것 부담도 있었지만 제가 생각한 가설이, 그리고 여기저기서 퀘스트 재료 모으듯이 긁어온 데이터들이 현실의 문제를 해결하는 데 도움을 준다는 사실을 원동력 삼아 즐겁게 프로젝트를 진행할 수 있었습니다.

앞으로도 데이터 분석이라면 흥미를 잃지 않고 스스로 동기부여를 받으며 일할 수 있다는 자신감이 있습니다. 좋은 팀원 분들과 좋아하는 분야에 대한 열정을 함께 태울 날을 기대하겠습니다.