나만의 데이터 분석 End- to End 프로젝트 기획서

[2023 OUTTA 부트캠프 데이터분석반] 최혜린(010-8696-0840)

1. 주제 의식 및 개요

전국 만 25세 ~ 79세를 대상으로 진행된 교육통계서비스의 '평생학습 개인실태조사'에 따르면, 25 ~ 34세의 경우 평생학습 프로그램의 학습매체로 책이나 판서 수업 보다는 인터넷 강의 또는 컴퓨터(전자책, 태블릿PC, 스마트폰 등)를 선호한다고 답한 비율이 70.6%이었다. 학습방법의 경우 집단학습보다 개인학습을 선호한다고 답한 비율이 57.6%로 더 높게 나타났다. (5점 척도로 조사하여 '4점:다소 그렇다' 또는 '5점:매우 그렇다'로 응답한 비율) 한편 "성공적인 직장생활을 위해서는 지식과 기술을 끊임없이 향상시켜야 한다."는 태도로 학습에 임한다고 답한 점수가 모든 유형 중 77.8점으로 가장 높았다. "성인이 되어서도 지속적으로 학습을 하는 사람은 일자리를 잃을 가능성이 적다."의 학습 태도가 75.2점으로 뒤를 이었다. (5점 척도를 100점으로 환산) 이와 관련하여 「취업, 이직, 창업」, 「일하는데 필요한 기능 습득」, 「성과급, 연봉 등 소득증대」, 「고용안정」 등 직업과 관련된 목적을 가진 비형식 교육에 참여하기 위해서 연평균 27만원을 투자한다. 이를 종합하면, **사회 초년생에 해당하는 25~34세 성인 학습자들의 경우 자신의 직무 역량을 향상시킬 목적으로 온라인 학습 기반 평생 학습 커리큘럼에 참여하는 것을 선호하며, 평생학습기관 주관 프로그램에는 연평균 27만원을 자기부담하는 경향을 보인다는 결론에 도달한다.**

평생학습, 특히 성인들의 소프트웨어(SW) 역량 함양을 주제로 하는 평생학습 컨텐츠에 주목해야한다. 이미 정부는 2021년부터 폭발적으로 증가하고 있는 SW 인력 수요에 대한 대응으로 2025년까지 총 41만 4천 명의 SW인재 양성 목표의 종합대책을 내놓았다. SW 중심대학 확대를 통한 전공자 증대, 기업 주도 재직자의 직무 역량 향상을 위한 '프로젝트 기반 현장 훈련 지원'과정 등에 정부가 훈련비, 인건비 등을 지원한다. 인공지능(AI), 사물인터넷(IoT), 빅데이터 등 SW 신기술은 국가 성장의 핵심 동력이기에 관련 능력을 보유한 청년 인력 배출과 양성에 적극 투자하겠다는 것이 정부의 입장이었다. 네이버의 공익목적 교육사업을 주관하는 커넥트재단 역시 SW 교육 방식으로써 프로젝트형 학습과 더불어 높은 접근성, 고른 학습 기회, 유동성, 편의성을 장점으로 갖는 온라인 학습에 주목하고 있다. "SW가 산업의 중심이 되는 미래에는 생애 전반에 걸쳐 지속적인 배움이 필요하다. 누구나 경제적 부담 없이 필요한 최신 기술을 원하는 시기에 배울 수 있도록, 교육으로 다가올 미래를 준비한

다.”는 슬로건을 내걸고 2011년부터 성인 SW교육 플랫폼인 부스트캠프, 부스트코스를 운영하고 있다. 온라인 학습 기반인 코칭스터디의 경우 현재까지 9151명이 학습에 참여한 것으로 파악된다.

본인 역시 성인 학습자로써 플립 러닝(Flipped Learning : 온라인 사전 학습과 (오프라인) 프로젝트 학습이 혼합된 학습 방식) 기반 데이터분석 부트캠프에 참여하고 있는 입장에서 이러한 ‘성인 학습자를 대상으로 하는 SW 관련 온라인 학습’과 관련된 데이터를 분석해보고자 하였다. 데이터 수집 가능성, 데이터분석 능력을 고려했을 때 우선은 SW 교육을 위한 온라인 강좌의 데이터셋에 초점을 두었다.

1. 게시된 강좌 최초 개설일(timestamp) 및 콘텐츠 지속 기간(content duration)을 사용하여 강의 품질(강의의 질이 높다 = 콘텐츠 지속 기간이 길다고 정의)이 학생의 지불 의지(price)에 어떤 영향을 미치는지 파악해볼 수 있겠다.
2. 혹은 누적 수강생 수(subscribers), 리뷰의 수(reviews), 탑재된 강의의 수(lectures), 강의 난의도(level), 평점(rating) 및 콘텐츠 지속 기간을 기반으로 Udemy 강좌의 가격 예측을 해볼 수도 있다.
3. 분류 모델을 사용하여 인기 강좌와 어떤 요인이 강좌의 인기에 유의미한 영향을 주었는지 파악 해볼 수도 있을 것이다.

2. 데이터 수집 방법 (데이터 출처 포함)

https://www.kaggle.com/datasets/thedevastator/udemy-courses-revenue-generation-and-course-anal?resource=download

> 3.1-data-sheet-udemy-courses-web-development.csv

분석의 방향을 잡아 줄 데이터를 캐글에서 발견하였다. Udemy는 K-MOOC(한국형 오픈형 온라인 학습 과정)과도 연계되어 있는 미국 본사 세계 최대 온라인 교육 플랫폼이다. 학습 카테고리로는 프로그래밍 언어, 웹 개발, AI, 데이터 과학, 게임 개발, IT 자격증, 암호화폐&블록체인, 소프트웨어 개발&테스트, 비즈니스 분석&인텔리전스, 비즈니스&매니지먼트, 마케팅, 3D&애니메이션, 디자인, 커뮤니케이션, 자기 계발이 있다. Udemy의 경우 유료 컨텐츠가 주를 이룬다.

해당 데이터셋은 특히 ‘웹 개발(Web Development)’ 관련 강좌 정보에 초점을 두고 있다. 총 1206개의 행을 가지고 있다. 각 열은 course_title, url, price, num_subscribers, num_reviews, num_lectures, level, rating, content_duration, published_timestamp, subject로 구성되어 있다. 분석 과정 중에 데이터가 더 필요하다고 느껴 udemy 홈페이지에서 추가 데이터를 수집하더라도 칼럼명을 기준으로 필요 데이터를 정렬할 예정이다.

3. 예상되는 데이터 전처리 방법

해당 데이터는 csv 파일로 정형 데이터이다. 결측값, 파손값 등을 제거하는 데이터 정제부터 시작한다. ‘rating’ 등이 0인 데이터가 보이는데 강좌를 직접 검색하여 이것이 정말 평점 0점인지, 누락값인지 파악하여 그대로 분석에 사용하거나 값 채워넣기/행 삭제 등의 전처리를 할 생각이다. Beginner, Expert 등 범주형 데이터인 ‘level’은 One-Hot Encoding을 이용해서 숫자형 데이터로 변환할 것이다. 정형 데이터는 어떤 열을 중심으로 모델을 학습시킬지에 대한 고민이 중요하다. 전 행이 동일하게 Web Developement인 ‘subject’ 등의 열은 제거하는 것이 바람직하겠다. 시각화를 통해 변수 간 스케일 차이가 심할 경우 Min-Max Normalization이나 Z-score Normalization 등 데이터 정규화 역시 고려해 볼 것이다. 현재로서는 대부분의 요소들이 다 유의미해 보이기 때문에 주제에 따라 종속 변수를 잘 정해서 (예를 들어, 가격 예측 모델의 경우 가격) 회귀 모델을 위한 평균중심화나 로그 변환이 필요한지 여부를 파악해야겠다.

4. 데이터 분석 방법

시각화로써는 변수 간 상관관계 파악에 용이한 히트맵(Heatmap)은 반드시 사용할 것 같다.

목표하는 바가 데이터 내부에 있는 레이블된 데이터로 학습이 진행된다는 특징을 가진 ‘지도학습’을 주된 분석 방법으로 활용할 예정이다. 데이터셋 안의 요소 간 관계를 파악하고 그 요소를 예측하는 주제들을 떠올렸기 때문이다. 지도학습에는 분류(classification)와 회귀(regression)가 있는데 둘 모두 직접 사용해보고자 크게 잡은 세 가지 주제 중 두 가지에 대해 분석을 진행해보고 싶다. 데이터를 보니 모두 온라인 강의이나 무료인 경우와 유료인 경우로 나누어져 있었다. 이를 시작점으로 의사결정나무 방식을 적용시켜 가격과 관련된 주제를 풀어나갈 수도 있다. 해석도가 높은 선형 회귀 모델을 사용하여 예를 들어 ‘탑재된 강의의 수가 많을수록 강좌 가격이 더 나간다.’, ‘평점이 높고 리뷰 수가 많은 강좌 일수록 가격이 더 나간다.’ 등의 가설을 검정 해볼 수도 있다. 선형 회귀 모델을 사용한다면 다시 전처리 과정으로 돌아가서 로그 변환을 적용해 예측력까지 높여볼 생각이다.

5. 기대 효과 및 예상되는 결과

이 데이터분석 결과가 향후 온라인 학습 매체를 기반으로 하는 성인 SW교육 프로그램들의 질을 높이는 방안 마련에 일조하기를 기대한다. 학습 과정 기획 및 운영자들이 강사를 구하거나 강의의 수준, 길이, 가격 등을 정할 때 참고 자료로 쓰일 수 있다. 강좌의 운영 관련 데이터셋이나 수강생들의 학습 데이터셋을 구할 수만 있다면 이 데이터 분석 결과와 더해, 이 학습자가 성공적으로 프로그램을 완주할 것인지 아닌지 예측하는 모델을 구현할 수도 있겠다. 혹은 수강을 완료한 학습자가 관심을 가질 만한 다른 유사 컨텐츠를 추천해주는 시스템을 구현할 기반이 될 수도 있다. 온라인 강의 기반이더라도 OUTTA 처럼 집단 학습 방식이 학습자

의 학습에 효과적인지, 순수히 개인적으로 온라인 학습을 진행하는 것이 나은지 등을 비교하는 추가 분석/연구에도 도움이 되기를 기대한다.

6. 참고자료

- [통계조사] 교육통계서비스 | 평생학습 개인실태조사
  https://kess.kedi.re.kr/stats/intro?menuCd=0107&survSeq=2020&itemCode=01
- [신문 기사] ZDNET Korea, 임유경 기자(2021.6.9.) | 정부, 5년간 SW인재 41.3만명 키운다
  https://zdnet.co.kr/view/?no=20210609142515
- [홈페이지] 네이버 부스트코스 | 코칭스터디
  https://www.boostcourse.org/coachingstudy
- [홈페이지] K-MOOC | K-MOOC x udemy 수강안내서
  [통계조사] K-MOOC | 2023년 K-MOOC 수요 분석 및 활용 현황 조사
- 2023 OUTTA 부트캠프 데이터분석반 강의 자료