**프로젝트 요약**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **프로젝트 제목** | 프로젝트명 | | | | |
| **구성원** | **이름** | **학교** | **학과** | **전화번호** | **이메일** |
| 백승근 | 카이스트 | 컴퓨터공학 |  |  |
| 장유진 | 성균관대학교 | 소프트웨어 |  |  |
| 배지훈 | 중앙대학교 | 컴퓨터공학 |  |  |
| 김나원 | 중앙대학교 | 컴퓨터공학 |  |  |
| 강준후 | 경희대학교 | 컴퓨터공학 |  |  |
| **지도교수** | Prof. Eric T. Matson | | | | |
| **참여멘토** | Prof. John Springer | | | | |
| **프로젝트**  **필요성 및**  **개요** | 오늘 날 정보통신 분야의 화두 중 하나는 빅 데이터다. 다양한 분야에서 방대한 양의 데이터가 생겨나고, ‘정형화된 데이터’부터 ‘비정형 데이터’까지 여러 종류의 데이터들의 처리와 가공 역량의 필요성이 급증하는 시대다. 머신러닝을 통한 여러 응용분야에서 그 효용성은 입증되었으며, 빅 데이터의 활용을 통해 사람의 행동 패턴을 예측하기도 한다. 이는 특히 영리 목적의 기업 입장에서 경쟁력의 필수 요소이다.  이에 본 팀은 소비자의 소비 패턴 분석에 중점을 두고 프로젝트를 시작했으며, Kaggle Competition의 ‘KKBox’s Music Recommendation Challenge’를 주제로 진행했다. 소비자의 개인 정보(나이, 거주지역)와 과거 서비스 기록을 기본 데이터로 특정 노래를 청취한 후 한 달 이내에 재 청취할 가능성을 예측했다.  이 문제에 대해 머신러닝의 ‘Gradient Boosting’, ‘Neural Network’, ‘Random Forest’ 3가지 모델을 사용하였다. 각각의 모델마다 데이터 종류에 따라 가지는 장단점이 있기 때문에 여러 방향에서 문제에 접근하고자 노력했다. | | | | |
| **프로젝트 결과 요약** |  | | | | |
| **적용 기술** |  | | | | |

**프로젝트 수행 결과보고서**

* **프로젝트 소개 및 목표**

1. **프로젝트 소개**

본 프로젝트는 IITP에서 진행하는 IITP-Purdue 프로그램 내 캡스톤 디자인 프로젝트의 결과이다. 본 팀의 분야는 ‘Big data and Machine learning’ 이었으며 Kaggle사의 Competition 중 ‘KKBox’s Music Recommendation Challenge’를 주제로 프로젝트를 진행하였다. KKBOX는 대만의 음악 스트리밍 서비스를 제공하는 기업으로 이 대회에서 이용자와 노래에 관련된 데이터 셋을 제공한다. 이 데이터 셋을 기반으로 이용자가 어떠한 특정 노래를 청취 했을 때, 그 특정 노래를 한 달 내에 재 청취할 확률을 예측한다. 주어진 여러 종류의 데이터를 가공하고 머신러닝 알고리즘을 통해 예측에 적용한다. 개발 라이브러리로는 ‘TensorFlow’, ‘Scikit-Learn’을 사용하였고 시행 모델로는 ‘Light Gradient Boosting’, ‘Neural Network’, ‘Random Forest’ 가 있으며 개발언어로는 ‘Python’, ‘R’을 사용하였다.

1.1 빅 데이터(Big data)와 프로젝트 동기(Motivation)

현 세대에서 컴퓨터공학 분야의 진보는 빅 데이터(Big data)와 머신러닝(Machine Learning)을 제외하고 논할 수 없다. 그만큼 이미 우리의 현실 생활에서도 깊게 관여하고 있으며 특히 인공지능 분야에서는 그 역량을 여실히 드러내고 있다. 또한 방대한 양의 데이터를 여러가지 머신러닝 알고리즘을 통해 가공, 분석함으로써 기존에는 해결이 어려웠던 문제들에도 접근할 수 있게 되었다.

빅 데이터와 머신러닝의 급속한 발전은 비즈니스 분야에서도 주목받고 있다. 이는 빅 데이터가 세계적으로 통합된 환경에서 상거래에 큰 영향을 끼칠 수 있기 때문이다. 기업들은 빅 데이터 활용을 통해 전반적인 프로세스와 조직 구성도를 개선하며 직접적인 소비자의 행동 패턴을 분석, 예측하기도 한다. 기존에는 무의미했던 데이터를 통하여 새로운 가치를 창출해 내는 것이다.

이에 본 팀은 프로젝트의 주제를 빅 데이터와 머신러닝을 통한 소비자의 소비 패턴 분석에 중점을 두고 시작했다. 기업의 입장에서 소비자의 행태를 예측하고자 했기에 충분한 소비자의 데이터 셋이 필요했고, 기업 측에서 데이터 셋을 제공하는 ‘KKBox’s Music Recommendation Challenge’ 대회에 참가하게 되었다.

1. **프로젝트 목표**

본 프로젝트는 음악 스트리밍 서비스 사의 입장에서 최대 이익을 창출하는 목표의 부분으로서, 이용자(user)와 관련된 방대한 양의 데이터를 통해 소비자의 행동을 예측한다. 데이터 셋을 기반으로 이용자가 어떠한 특정 노래를 청취했을 때, 그 특정 노래를 한 달 내에 재 청취할 확률을 예측한다. 이 예측은 이용자의 새로운 노래와 아티스트에 대한 호감도와 연계될 수 있으며 새로운 이용자에게 효율적인 노래 추천 서비스까지 제공할 수 있게 된다. 이에 각 사용자의 특정 노래 별 재 청취율의 예측 정확도를 최대로 높이는 것을 목표로 한다.

2.1 활용 가능성

본 프로젝트는 이용자의 개인 정보와 청취 기록을 통해 재 청취율을 예측한다. 소비자 개개인의 경향을 예측한다는 점에서 다수를 위해 일반화된 서비스와 차별성을 둔다. 이러한 개인 맞춤 서비스(Customer Specific Solutions)는 고객과 장기적인 관계를 유지하는 데에 중추적인 역할을 하며 이는 판매 수익률과도 직접적인 연관이 있다.

이용자의 재 청취율을 예측한다는 것은 개개인마다의 특성을 추출할 수 있다는 것을 의미한다. 특정 노래의 재 청취율을 예측함으로써 장르, 작사가, 작곡가 등에 대한 선호도를 알 수 있고 연령별, 지역별 경향 또한 조사할 수 있다. 이는 이용자에게 노래 추천 서비스와 서비스 제공 플랜까지 맞춤으로 제공할 수 있음을 의미한다.

이렇게 고객의 개인 정보와 단기적으로 생성된 서비스 기록으로 소비 패턴을 예측할 수 있다면 이는 고객 관계 관리(Customer relationship management) 시스템에도 영향을 끼친다. 시장과 고객에 대한 이해도가 높아짐으로써 새로운 고객 가치 분석이 가능하게 된다. 또한 잠재 고객, 휴면 고객에 대한 전략적 영업 플랜을 수립할 수 있다.