1

안녕하세요. 저는 AI 부트캠프 18기 장건희입니다.

이번 프로젝트의 주제는 ‘이번 여름, 어떤 신메뉴을 개발해야 할까?’입니다.

2

카페 업계는 경쟁이 치열한 시장으로, 새로운 메뉴를 개발하여 시장에서 차별화를 추구하는 것이 중요합니다.

소비 트렌드에 대응하여 성공적으로 신메뉴를 개발할 경우, 새로운 고객들을 유치하고 기존 고객들의 재방문을 유도할 수 있습니다. 일시적인 매출 증대 뿐 아니라, 신메뉴를 바탕으로 마케팅 및 홍보 또한 적절하게 수행한다면 시장 경쟁력 강화로도 이어질 수 있습니다.

3

그러나 개발한 신메뉴과 성과를 보이지 않는 경우, 시간과 노력이 낭비됩니다. 실패한 신메뉴를 개발하는 동안 다른 중요한 프로젝트나 기회를 놓칠 수 있으며, 이로 인해 매출이 감소하고 시장경쟁력이 약화됩니다.

4

저희는 신메뉴 개발부서팀에 도움이 될 수 있도록 고객들의 구매 현황과 구매된 제품들의 특징을 바탕으로 판매량을 예측해주는 서비스를 개발하고자 합니다.

5

서비스를 개발하는 전체적인 흐름은 다음과 같습니다.

6

제품 데이터는 스타벅스 홈페이지에서 Selenium을 통해 동적 스크레이핑 방식으로 수집했습니다. 홈페이지에서 제공하지 않는 정보의 경우 직접 방문 및 검색을 통해 추가했습니다.

고객이나 주문 데이터의 경우, 외부에서 수집할 수 없었기 때문에, 프로젝트 진행을 위해 가상의 데이터를 생성했습니다. 이러한 작업은 모두 Python에서 수행했습니다.

7

이후 수집한 데이터를 SQL 데이터베이스로 관리하였습니다. 관리도구로는 TablePlus와 MySQL 를 사용했습니다.

8

오른쪽 그림은 DB에서 관리되고 있는 모습입니다.

9

이후 데이터의 특징을 동적으로 파악하기 위해서 테블로 퍼블릭을 이용하여 대시보드로 표현했습니다.

10

예측모델은 다중선형회귀 모델을 사용했습니다. 예측 모델 생성에 필요한 전 과정은 Python에서 수행했습니다.

이렇게 생성된 모델을 Flask에 넘겨서 Web 상에서 서비스를 제공하고 있습니다..

11 (tablue desktop으로 이동)

예측에 사용할 데이터를 한 번 살펴보겠습니다.

먼저 고객들의 특성입니다.

성비에 경우 여성 고객이 더 많이 주문했지만, 남성과 여성에 따른 제품선호도는 큰 차이를 보이지 않습니다.

(고객 연령대로 이동)

연령대도 마찬가지로 특정 연령대가 많이 나타나고는 있지만, 연령대에 따른 제품선호도는 차이가 없습니다.

따라서 신제품 개발 시 특정 고객층의 주문 기록에 초점을 맞추기 보다는, 고객 모두의 주문 기록을 살펴보았습니다.

(음료 판매 추이)

이 그래프는 지난 한해 저희 카페 음료수 판매량 기록입니다. 보시면 아시겠지만, 저희 카페는 여름 전후로 하여 기존보다 매출이 증가하는 추세를 보입니다.

여름에 방문이 늘어나는 특징을 보여서 그런지 대체적으로 차가운 제품을 많이 구매하고 있습니다.

(5~9 하이라이트)

특히 매출이 증가하는 5월에서 9월 사이로 집계를 해보면, 차가운 제품의 판매 비율이 더욱 증가하고 있음을 확인할 수 있습니다. 따라서 신제품은 차가운 음료를 개발한다고 가정했습니다.

뿐만 아니라 5월에서 9월 사이로 집계를 해보면, 전체 기간 대에 비해 일부 카테고리가 판매되지 않고 있다는 것을 확인할 수 있습니다. 따라서 여름철에 판매되지 않는 카테고리 제품 또한 예측에 고려하지 않겠습니다.

각 카테고리의 세부 내용을 확인해보면, (한 번씩 확인) 카테고리 내 개별 제품은 대부분 비슷한 판매량을 기록하고 있습니다. 단 에스프레소 카테고리의 경우, 아메리카노가 너무 많은 비율을 차지하고 있습니다.

(음료 세부 추이로 이동)

아메리카노는 전체 판매 기록을 확인했을 때에도 너무 많은 비율을 차지하고 있어, 아메리카노를 포함하고 예측 모델을 생성할 경우, 예측 모델이 아메리카노만 묘사하는 경우가 발생했습니다. 따라서 아메리카노 판매 기록은 훈련 데이터에서 제거했습니다.

(ppt 이동)

입력 데이터는 다음 목록과 같습니다. 각각의 연속형 변수는 다음의 범위에서 값을 가집니다.

범주형 변수의 경우 카테고리는 티, 에스프레소, 블렌디드 등을 나타냅니다.

카페인 함량은 시약처에서 제시하는 기준으로 고카페인과 저카페인으로 나누었습니다.

종류의 경우, 커피, 혼합, 과일음료 및 차 등으로 새롭게 구분하였는데, 간단한 예시는 아래 내용을 확인해주시면 감사하겠습니다.

12

현재 생성된 예측모델에서 주요 변수들의 변수 중요도는 다음과 같습니다.

범주형 변수에서는 카테고리가 판매량에 미치는 영향이 크게 나타났습니다.

연속형 변수에서는 칼로리가 높을수록 판매량에 미치는 악영향이 크게 나타났습니다.

api이동

그럼 짧게 서비스를 시연해보겠습니다.

다음과 같이 새로 개발한 신제품의 특징을 입력하면 예상 판매량이 나옵니다.

앞서 살펴본 변수중요도에서 판매량에 긍정적인 영향을 미치는 범주들을 선택하면 다음과 같이 판매량이 나옵니다.

반면 판매량에 부정적인 영향을 미치는 범주들을 선택하면 다음과 같이 나옵니다.

(ppt 이동)

13

프로젝트를 마무리하면서 들은 생각은, 전반적으로 이번 프로젝트는 아쉬움이 많이 남습니다.

먼저 데이터를 수집하는 과정에서도, 고객이나 주문 기록을 직접 수집하지 못해서 가상의 데이터를 활용했다는 부분이 서비스를 개발하고도 굉장한 찝찝함으로 남았습니다.

또 다음은 바로 서비스 시연입니다. 앞서 보셔서 아시겠지만 결과가 너무 초라합니다. 사실 데이터베이스에 변화가 생길때마다 예측모델이 바뀌기 때문에 이에 따라 변수중요도도 달라집니다. 따라서 결과 부분에 변수중요도 그래프도 같이 제시하고 싶었는데, 한글 시각화랑 겹치면서 자꾸 에러가 발생해서 넣지 못했습니다.

또 단순히 예측 판매량만 제시하는 게 아니라, 이러한 예상 판매량이 어떠한 근거를 바탕으로 신제품 개발에 있어서 긍정적인지, 부정적인지 표현하고 싶었습니다. 그런데 시간이 촉박하여 이 부분을 수정하지 못한 게 굉장히 아쉽습니다.

이상으로 발표를 마치겠습니다. 감사합니다.