안녕하세요 의뢰인님! 이번에 인기있는 음악은 어떤 특성을 가져야하는지에 대해 의뢰를 맡겨주신 박건후라고 합니다.

설명해드릴 목차는 다음과 같습니다. 데이터 설명, 전처리 과정, 모델링 및 해석, 결론 순으로 말씀드리겠습니다.

자, 그럼

첫 번째, 데이터 설명 및 문제 정의 입니다.

의뢰인님은 한국 음악 시장이 글로벌해짐에 따라 어떤 음악이 세계적으로 성공할 수 있는지 분석을 의뢰해주셨는데요. 과연, 대중적으로 성공적인 음악을 출시하기 위해서는 어떤 특성을 가져야할까요?

음악의 특성들입니다. 이런 것들이 있고, 대부분 0~1의 수치를 가집니다. 그리고 이 중에서 장르, popularity, danceability 정도만 중요하게 봐주시면 될 듯 합니다.

(넘기기)

다음은 문제 정의 및 가설 설정입니다. 문제 정의는 다음과 같구요.

그에 따른 가설은 다음과 같습니다.

다음으로 전처리 과정입니다.

먼저, 데이터 셋 확인입니다. (읽기)

다음은 타겟 전처리입니다. (읽기)

(읽기)

다음은 문자형 칼럼 전처리입니다. id와 name이 들어간 칼럼이 카디널리티가 높은 것을 확인할 수 있습니다. 이 칼럼들은 모델이 학습하기에는 Private한 데이터이므로 추후 삭제하도록 합니다. (그 다음 읽기)

다음은 수치형 칼럼 전처리입니다. (읽기)

그 후, 그래프를 그려서 수치형 데이터의 이상치를 확인했는데요. 'duration\_m'이 이상하다는 것을 확인하였습니다. 그리고 다른 칼럼에서도 이런 이상한 분포도를 확인할 수 있는데 이는 원래 존재하는 수치들이므로 그냥 사용하였습니다. (그 다음 읽기)

지금까지의 과정에서 도출한 것으로 중간 점검을 하겠습니다. (읽기)

네, 아직도 전처리가 남았죠…?

남은 전처리를 해줍니다. (읽기)

지금까지의 과정으로 해결할 수 있는 문제들이 있는데요. 먼저, 가장 인기있는 장르는 어떤 것일까 입니다.

출시된 장르의 수를 확인했을 때, chlidren’s music이 제일 많은 것을 알 수 있고,

인기도가 높은 탑10 장르를 확인하였을 때, pop이 제일 높은 것을 확인할 수 있습니다.

따라서, 가설 ‘가장 인기 있는 장르는 Pop일 것이다.’ 는 True입니다. 또한, 출시되는 갯수가 많다고 인기를 차지하는 비중이 높지 않다는 것을 알 수 있습니다.

다음은 재생시간은 어느 정도가 적당한지, 재생시간이 인기도에 영향을 주는지에 대해서인데요.

재생시간 분포를 확인한 결과, 데이터가 몰려있음을 볼 수 있습니다.

따라서, 재생시간이 7분을 초과하는 데이터를 삭제 후 확인해보면 3~4분 사이의 음원에서 인기도가 높음을 알 수 있으므로 이 시간대가 가장 적당합니다.

그리고 더 확실히 보기 위해 선 그래프로 확인해보면 특정 구간의 재생시간이 넘어가면 재생시간이 길어질수록 인기도가 하락함을 알 수 있습니다.

즉, 가설 ‘재생시간은 인기도에 영향을 주지 않을 것이다.’ 는 False입니다.

다음은 세 번째 순서인 모델링 및 모델 해석 입니다.

먼저, 중간 점검 때 체크해 둔 사항을 토대로 train set을 전처리 해줍니다.

(읽기)

그 후, (읽기) 02. 이 평가 지표들 중에서 auc 스코어를 사용할 것이고, 나머지 다른 평가 지표들은 전체적으로 비교해보고자 함수를 작성하였습니다. 그리고 기준 모델을 설정 후 점수를 계산해보았습니다.

다음은 모델링인데요. 저는 여기 있는 (읽기) 이 세 가지 모델로 데이터를 학습 후에 성능을 비교하고 최종 모델을 선택하였습니다.

(읽기)

(읽기)

(읽기)

이렇게 해서 나온 각 모델별 검증 점수와 그래프는 이와 같습니다.

모든 모델의 점수와 그래프를 비교한 결과, XGB 모델이 가장 뛰어나므로 최종 모델로 XGB 모델을 골랐습니다.

그 후, 최종 모델인 XGB 모델로 Test Set 예측하였구요

Test set의 AUC 점수가 Validation set의 AUC 점수랑 비슷하게 나온걸로 보아 나름 일반화가 잘 된 모델이라고 할 수 있습니다.

그렇다면 인기도에 가장 영향을 많이 주는 특성은 무엇일까요? 이는 Permutation Importance와 그 수치를 이용한 특성 중요도 파악할 수 있었습니다.

그로 인해, 영향을 가장 많이 주는 특성은 Genre임을 알 수 있었고, 가설 ‘인기도에 가장 영향을 많이 주는 특성은 Danceability일 것이다.’는 False라는 것이 밝혀졌습니다.

다음으로 Permutation Importance를 통한 특성 중요도 상위 2개의 특성의 PDP를 확인하였는데요. Pop, Rap, Rock인 장르일수록, 데시벨이 높을 수록, 인기도 상승에 영향을 준다는 것을 알 수 있었습니다. 남은 문제들을 이 PDP를 통해 해결해 볼 건데요.

먼저, 댄스 음악이 인기가 많은지, 발라드 음악이 인기가 많은지에 대해서 입니다. PDP를 통해 Danceability에 의한 인기도 파악해 본 결과

수치가 낮을 때보다 높을 때 인기도가 상승하는 것을 볼 수 있습니다. 따라서, 가설 ‘댄스 음악이 인기가 더 많을 것이다.’는 True 입니다.

마지막 문제인 Genre에 따라 Danceability가 예측에 영향을 주는 정도가 다른지는 2특성 PDP를 통해 파악할 수 있었는데요.

이 그래프를 보아 가설 ‘Genre에 따라 Danceability가 예측에 영향을 주는 정도가 다를 것이다.’는 True임을 알 수 있습니다.

자, 지금까지 모델 해석까지 완료해보았는데요. 마지막으로 결론입니다.

과연, 대중적으로 성공적인 음악(인기도가 높은 음악)을 출시하기 위해서는 어떤 특성을 가져야할까요? 이를 위해 각 특성의 PDP를 그려보았고, 그를 통해 분석하였습니다. 자세한 그래프는 PPT 맨 뒤 부록 참고하시길 바랍니다.

그리고, 그 특성은 다음과 같습니다. 중요특성은 (읽기) 참고할만한 특성은 따로 읽어보시는 것을 추천드립니다.

네, 지금까지

의뢰인님의 의뢰에 대한 분석이었습니다.

감사합니다.