Song popularity prediction - (Base ML model)

목표 · Dataset 이해/ clean up (결측치제거)

· Song popularity 예측가능 regression 모델 생성

· 모델 평가 (R2, RMSE, ··)

순서

1. Data exploration → 2. EDA → 3. 전처리 → 4. Data manipulation. → 5. Feature selection / extraction.

6. predictive modeling → 7. outcome/conclusion.

1 Data exploration.

→ 필요한 library / dataset 로드.

· song_name 은 제거, target 변수: "song_popularity"

· df.info() 를 통해 null값 확인 → null 값 없음

· df.nunique().sort_values() 이용하여 변수확인

· numerical / categorical 변수 나누기.

2 EDA

· target 변수의 분포를 확인 (정규분포 따르는지 확인)

⇒ 전반적으로 정규분포적인 모양을 가진다.

· categorical 변수의 분포를 plot을 통해 확인.

- audio mode: 0/1 (이 2개 있다)
- time signature (0/1/2/3/4/5 - 4가 압도적이 많다)
- key (0~11) 골고루 분포한다.

· numerical 변수의 분포를 plot을 통해 확인

⇒ skewed 된 모양이 많이 보인다.

· pair plot 을 이용하여 모든 변수 간 상관관계를 파악.

3 데이터 전처리

· Duplicate row 삭제. (총 3개 삭제)

· categorical 변수를 numerical 로 전환.

- one-hot encoding
- dummy encoding

· outlier 제거 (IQR을 이용)

· 삭제된 data의 비율을 시각화

(삭제: 62.5%, 사용: 42.5%) Cleanup 까지 완료.

④ Data manipulation.

- train set / test set 으로 data 구분.
- Feature scaling (Standardization)

⑤ Feature selection / Extraction.

- 변수 간의 correlation (heatmap) 표현.

⇒ 다양한 correlation 이 보여진 것으로 판단.

- OLS를 통해 OLS Regression Result를 얻어낸다. (AIC, BIC 등..)

→ 여러 변수들 간의 상관성이 높은것을 판단.

⇒ 다중공선성 (multicollinearity)이 나타난다.

→ 1) minimal method - Variance Inflation Factor (VIF)

2) Automatic method - Recursive Feature elimination (RFE)

3) Featuring elimination using PCA decomposition.

1) VIF

- 다중공선성을 판단할 때 이용.
- 예측변수들이 상관성이 높아 추정 회귀 계수의 신뢰 크기를 측정한 것.
- 신뢰와 귀결수를 신뢰도 ↓ (VIF 값이 1근방이면 다중공선성 없이 모형을 신뢰, VIF기준 → 변수 선택법의)

2) RFE

- 재귀적인 feature selection 방식.
- 재귀 feature에 대해 훈련 후, 중요하지 않은 feature 제거.

3) PCA decomposition. (주성분분석)

⑥ predictive modelling

1) Multiple Linear Regression (MLR)

→ Actual model, prediction model의 분포도 plot으로 판단.

→ Training set / Testing set의 R2-score, MSE, RMSE 구함.

→ residual plot까지 그리기.

2) Ridge regression model.

→ 1)과 동일한 값 구한다.

3) Lasso regression model

4) Elastic-net

5) polynomial regression model

⑦ 위의 5개의 model을 비교.

→ 비교결과, model 별로 R2 값이 다양하게 나온다.

→ 5개의 model에 대한 비교결과 (plot)

→ 결과적으로, polynomial model의 R2 값이 가장 크다.

→ model 별 RMSE 값도 비교

↳ RMSE 값이 작을수록, training / test 의 값이 비슷할 수록 좋은 model.

→ RMSE 값으로는 MLR model이 가장 좋다. PLR은 overfitting 된 모습이다.

VIF / RFE / PCA

(Feature selection의 방식)

1. Filter method (Feature 간 상관성) - VIF (불안정성 O)
2. Wrapper method (Feature를 조정하여 모형 예측 → 성능파악, Feature 선택)
3. Embedded method (예측모형 최적화, 회귀계수 추정 과정에서 각 Feature를 선택)

- VIF: 10이 넘으면 다중공선성이 있는 것으로 판단. 5로 낮추라고 하기도.

→ VIF값이 무조건 10이 넘는다고 feature을 삭제해 줘야 아님, 상위 몇개만 제거하면서 확인.

- RFE: Backward 방식 중 하나. 모든 변수 다 포함 → 변수확률 진행. → 중요도가 낮은 것부터 삭제.

- PCA: 주성분에서만을 뽑아온다.