Seoul Software ACademy kaggle

2024.08.20

Mini Project

데이터분석 결과 보고

The Movies Dataset

: 데이터 AI 개발 과정 조유경

목차

Table of Contents

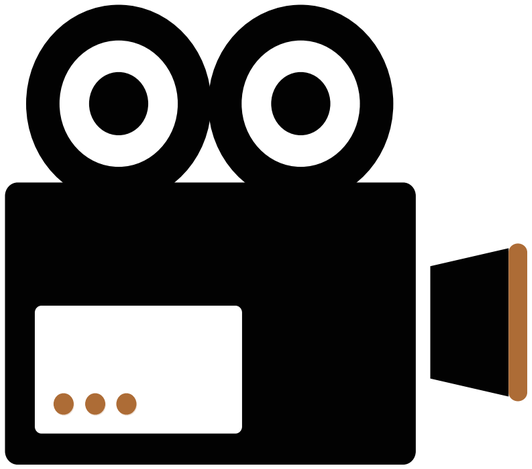
분석 목적 및 배경

Analysis purpose and background

" 영화 키워드와 영화 평점간의 상관관계 분석 "

분석 목적 및 배경

Analysis purpose and background



- ## 첫번째 배경 요인

  영화의 평점을 예측하는 것은 영화의 흥행 여부를 사전에 미리 점검할 수 있습니다. 영화의 흥행 여부를 예측함에 따라 영화제작에 대한 투자 의사결정에 도움을 줄 수 있습니다.

- ## 두번째 배경 요인

  영화 키워드는 영화의 주요 주제나 장르를 나타내며, 이를 기반으로 평점을 예측할 수 있을지 상관관계를 알아보려는 목적이 있습니다. 영화의 흥행에 영향을 주는 요인을 확인할 수 있다면 영화제작과정에서 해당 요인들을 고려하여 흥행가능성을 높일 수 있을 것입니다.

- ## 세번째 배경 요인

  키워드와 평점 간의 관계를 이해하면, 영화의 성공 요인을 분석하고 더 나은 추천 시스템을 구축할 수 있습니다. 고객의 니즈를 저격할 수 있는 추천시스템 구축에 도움이 될 것입니다.

데이터 소개 및 분석 과정

Introduction to data and analytics courses

The Movies Dataset

영화의 메타데이터와 키워드를 통해 영화 평점과 키워드 간의 상관관계를 분석하고,
이를 기반으로 영화 평점 예측 모델을 구축하는 데 사용

1 Keyword.csv

영화별 주요 키워드 정보를 포함한 데이터셋

Variable	Definition	DataRange	DataType
keywords	키워드	-	범주형

2 Movies_metadata.csv

영화의 메타데이터를 포함한 데이터셋

Variable	Definition	DataRange	DataType
vote average	영화 평균 평점	0.0 ~ 10.0	수치형 (연속형)

3 Ratings.csv

사용자 영화 평점 정보를 포함한 데이터셋

Variable	Definition	DataRange	DataType
rating	사용자가 부여한 평점	0-5	범주형 (순서형)

데이터 전처리

Data pre-processing

결측치 처리

" Keywords Dataset"

```
Keywords Dataset - Missing Values:
 id          0
keywords    0
dtype: int64
```

" Movies Meta Dataset"

```
Meta Dataset - Missing Values:
vote_average                6
vote_count                  6
dtype: int64
```

- vote_average column의 결측치는 평균으로 대체

json형식 data 파싱

" json형식의 data 리스트로 변환 "

```
df_merged['keywords'] = df_merged['keywords'].apply(lambda
x: [d['name'] for d in ast.literal_eval(x)])
```

- json형식의 데이터는 가공 처리가 어려우니 리스트로 변환
- 리스트의 경우 모든 keyword data를 확인하여 빈도수 파악시 코드실행 시간이 오래걸림
- 모델링 진행시 PCA(주성분분석)을 하여 차원 축소 진행

데이터 전처리

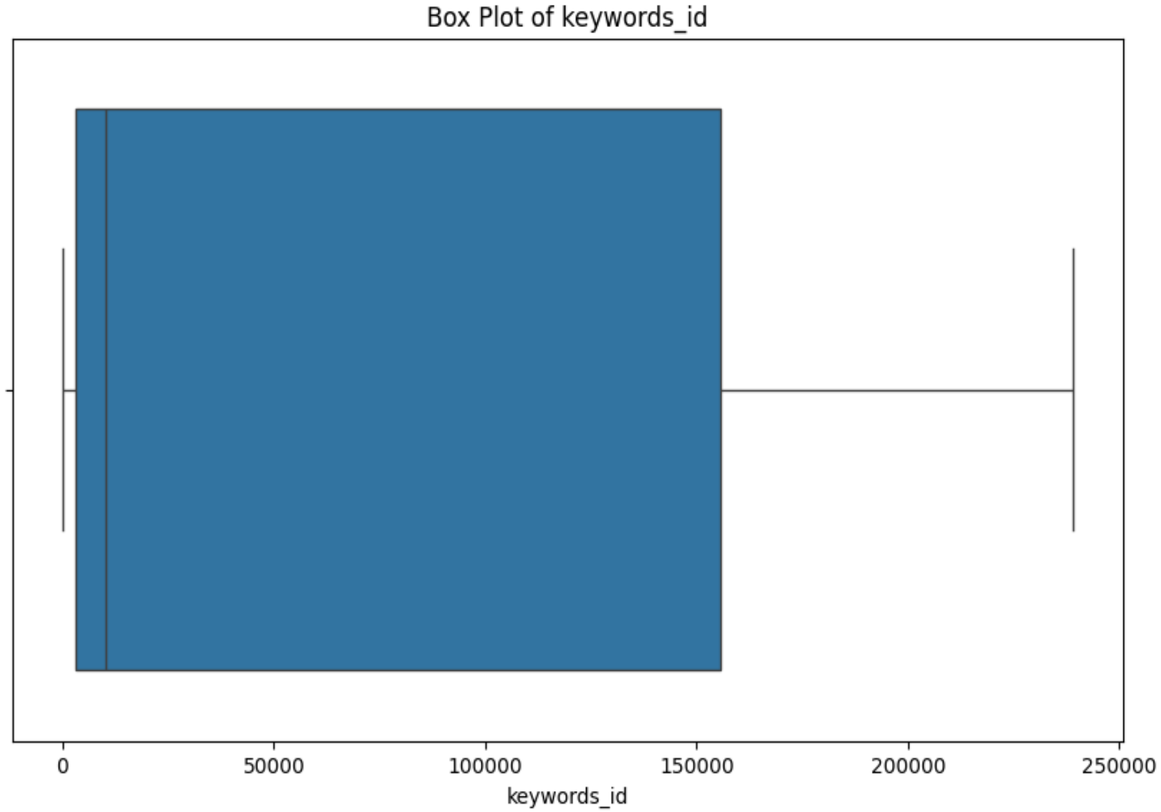
Data pre-processing

이상치 확인

" Keywords Dataset"

```
1 # 이상치
2 plt.figure(figsize=(10, 6))
3 sns.boxplot(x=df_keywords_expanded['keywords_id'])
4 plt.title('Box Plot of keywords_id')
5 plt.show()
```

- 중앙값이 왼쪽에 치우쳐있는 것으로 보아 데이터 분포가 크게 왜곡되어 있음을 확인
- 모든 Data는 수염 범위 안에 위치
- 이상치 시각화 결과 특별히 처리해야할 이상치는 없음



데이터 소개 및 분석 과정

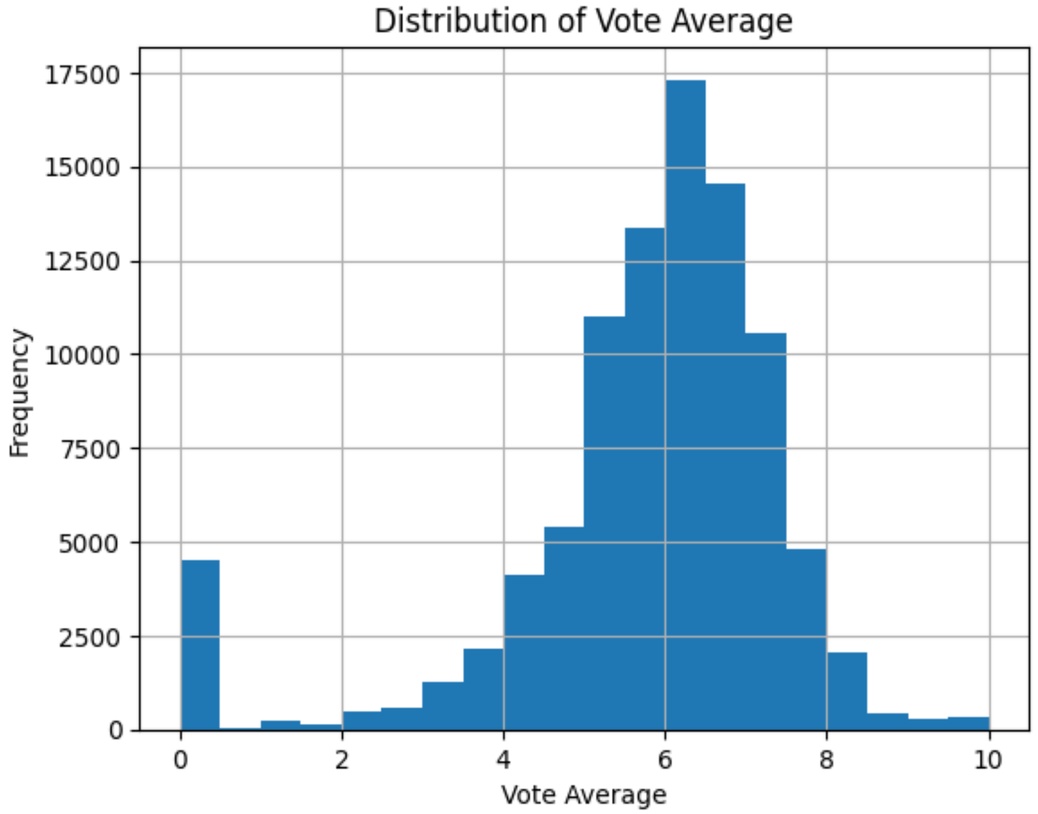
Exploratory Data Analysis

1 Movies_metadata.csv

" 데이터 분포 확인 "

히스토그램 시각화

- 4-8점 사이에서 정규분포 모양을 보임.
- 대체로 영화들이 평균적인 평가를 받았다는 것을 의미.
- 다만 0과 같은 최하점수에서 특이점을 보임
- 극도로 낮은 평가를 받은 영화가 존재함을 알 수 있음.
- 특이점을 보이는 영화의 특성을 파악해볼 필요가 있음.



데이터 소개 및 분석 과정

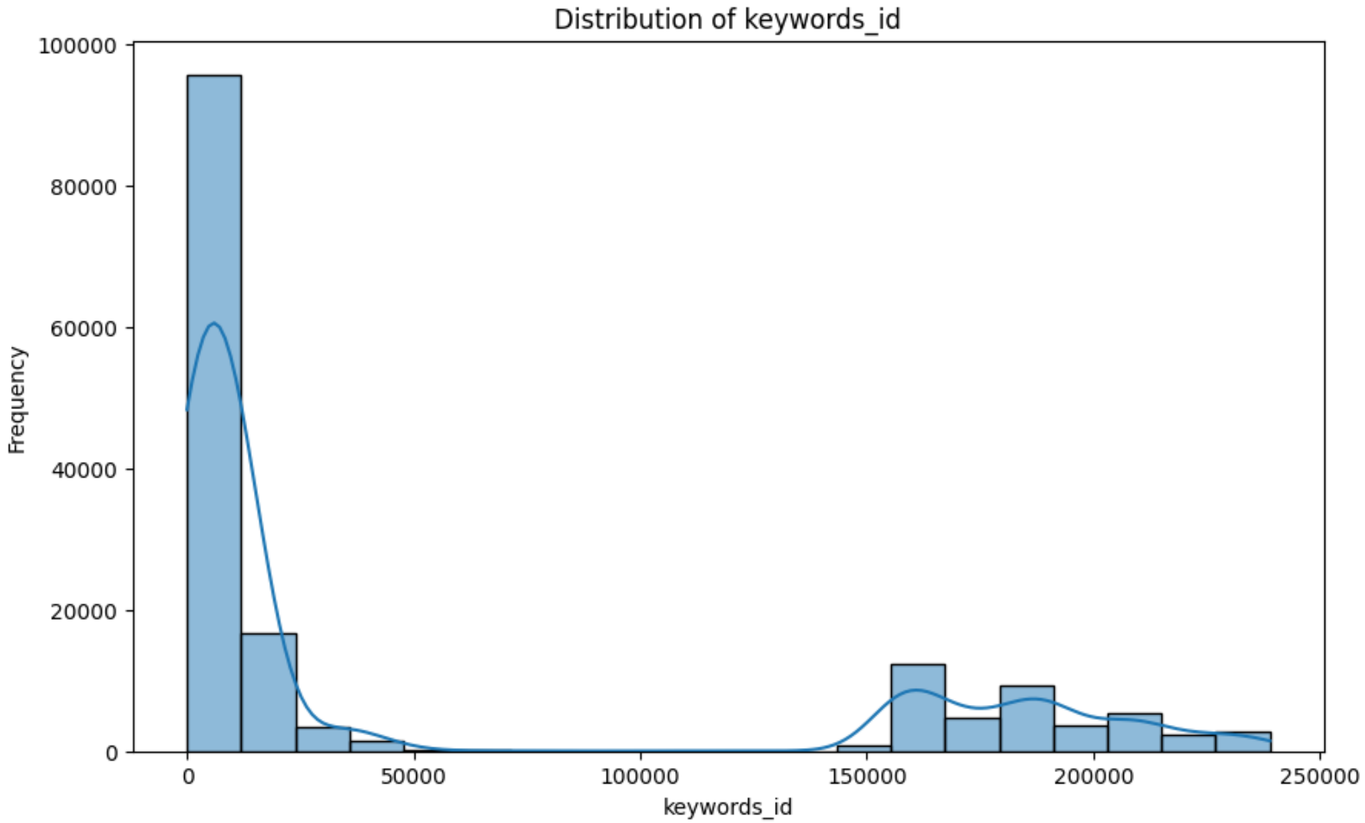
Exploratory Data Analysis

2 Keyword.csv

" 데이터 분포 확인 "

KDE(커널 밀도 추정) 그래프 시각화

- 데이터 편향성이 높음(positive skewness)
- 언급횟수가 적은 keyword가 극단적으로 많이 존재
- 150,000에서 200,000 사이에 몇 개의 군집 (cluster)이 존재 -> 특징을 확인해볼 필요 있음
- 해당 부분 전처리 작업 후 모델링 적용 필요



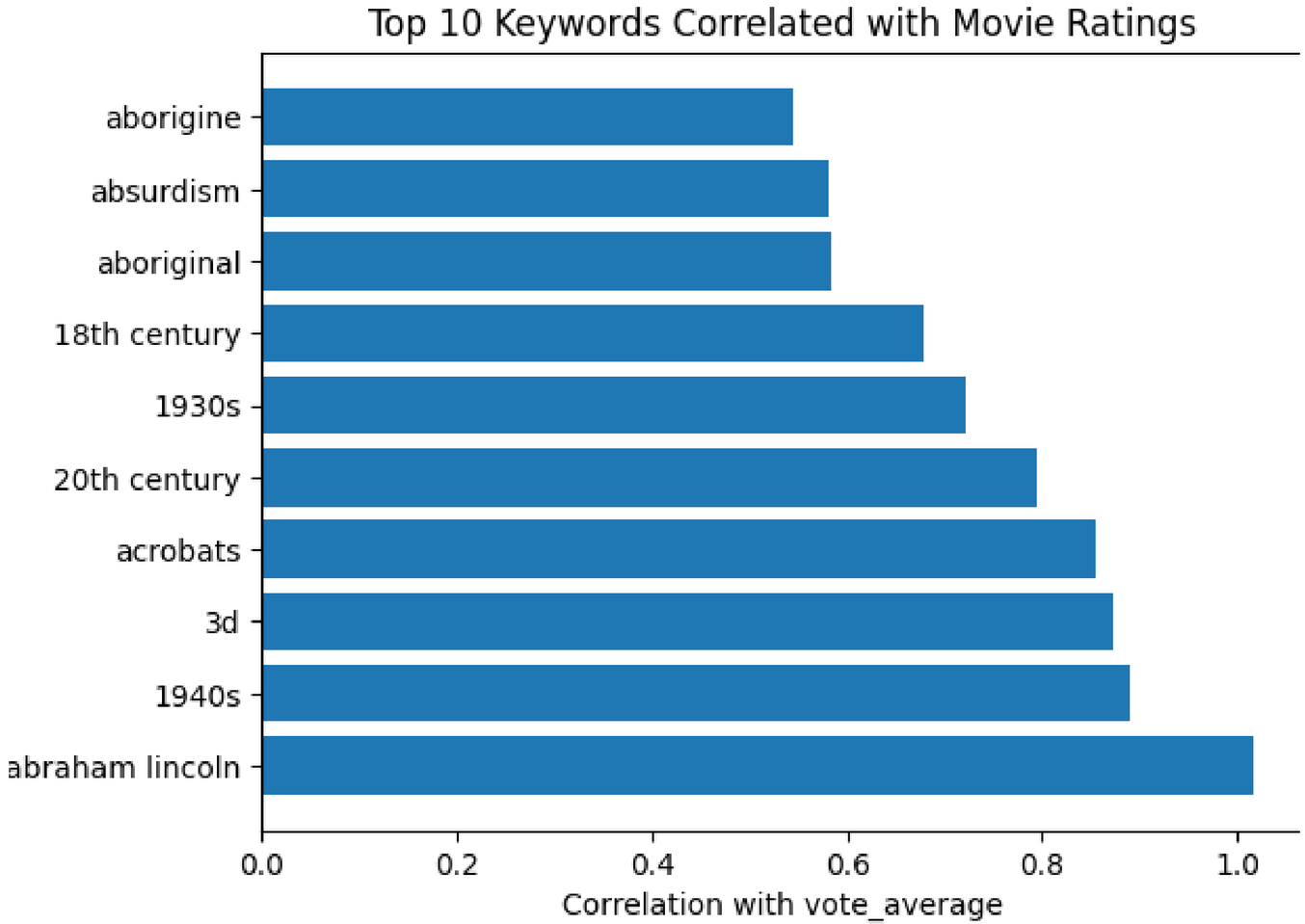
모델 선택 및 평가

LinearRegression

" Linear Regression 모델을 적용한
평점 예측을 통해 평점과 관련하여
중요도가 높은 keyword 추출 "

키워드 빈도 계산 및 필터링 → 키워드 벡터화 (CountVectorizer) → PCA 차원 축소

- keyword Top 10을 살펴보면, 특정 시기, 연도와 관련된 keyword가 눈에 띈다는 것을 알 수 있음.



LinearRegression 모델을 사용하여 추출한
평점과 관련하여 중요도가 높은 keyword Top 10

모델 선택 및 평가

Random Forest / XGBoost

랜덤포레스트(Random Forest)

```python
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score

# 1. 랜덤 포레스트 모델 생성
rf_model = RandomForestRegressor(n_estimators=100, random_state=42

# 2. 모델 학습
rf_model.fit(X_train, y_train)

# 3. 예측
y_pred_rf = rf_model.predict(X_test)

# 4. 성능 평가 (R²  스코어와 MSE)
r2_rf = r2_score(y_test, y_pred_rf)
mse_rf = mean_squared_error(y_test, y_pred_rf)

print(f"Random Forest R²  Score: {r2_rf}")
print(f"Random Forest Mean Squared Error: {mse_rf}")
```

```
Random Forest R²  Score: 0.03761432462367942
Random Forest Mean Squared Error: 3.5870270340622485
```

Random Forest 모델을 사용한 모델링

XGBoost

```python
import xgboost as xgb

# 1. XGBoost 모델 생성
xgb_model = xgb.XGBRegressor(n_estimators=100, random_state=42)

# 2. 모델 학습
xgb_model.fit(X_train, y_train)

# 3. 예측
y_pred_xgb = xgb_model.predict(X_test)

# 4. 성능 평가 (R²  스코어와 MSE)
r2_xgb = r2_score(y_test, y_pred_xgb)
mse_xgb = mean_squared_error(y_test, y_pred_xgb)

print(f"XGBoost R²  Score: {r2_xgb}")
print(f"XGBoost Mean Squared Error: {mse_xgb}")
```

```
XGBoost R²  Score: 0.024309943372273435
XGBoost Mean Squared Error: 3.6366154438248963
```

XGBoost 모델을 사용한 모델링

모델 선택 및 평가

Model selection and evaluation

항목	모델 설명	모델 성능 평가
LinearRegression	단순 선형 회귀 모델 립 변수와 종속 변수 간의 선형 관계를 모델링	R^2 Score: 0.0314693308136811 96 Mean Squared Error: 3.609930802774298
Random Forest	여러 결정 트리를 결합 예측 성능을 높이는 앙상블 학습 모델	R^2 Score: 0.037614324623679 42 Mean Squared Error: 3.5870270340622485
XGBoost	Gradient Boosting 알고리즘을 기반 고성능 앙상블 모델	R^2 Score: 0.024309943372273435 Mean Squared Error: 3.6366154438248963

결과 및 해석

Results and interpretation

결과 Results	• Linear Regression : (R^2 Score: 0.0315) / (MSE: 3.6099) • Random Forest: (R^2 Score: 0.0376) / (MSE: 3.5870) • XGBoost: (R^2 Score: 0.0243) / (MSE: 3.6366)
해석 interpretation	• Linear Regression: 키워드와 평점의 선형 관계를 확인 불가능. 설명력 낮음. • Random Forest: 비선형 관계를 일부 확인하였으나, 여전히 설명력 낮음 • XGBoost: 과적합으로 인해 성능이 떨어졌을 가능성 있음.

결론 및 향후작업

Conclusions and next steps

*" 키워드 Data의 경우, 평점과의 **뚜렷한 상관관계를 파악하기 어려움**. "*

*" **과적합**될 가능성이 높으며, 빈도수만으로 가중치를 부여하기엔 **한계**가 있음. "*

- 클러스터링 적용 : 비슷한 특성을 지닌 키워드들간의 클러스터링 진행 후 재분석
- 추가 데이터 활용 : 배우, 감독, 장르, 제작비, 제작사 등 추가적인 데이터를 포함하여 각 요인과 평점간의 통계검정분석을 실행 후 재모델링
- 하이퍼파라미터 튜닝 : 3가지 모델의 하이퍼파라미터를 튜닝하여 성능을 향상
- 데이터 전처리 개선 : 키워드 외의 다른 요인을 포함한 더 나은 전처리 방법 모색
- 다양한 모델 적용 시도 : 다른 머신러닝 모델이나 딥러닝 모델을 시도