L사 CVR Prediction

Machine Learning 수행 과제

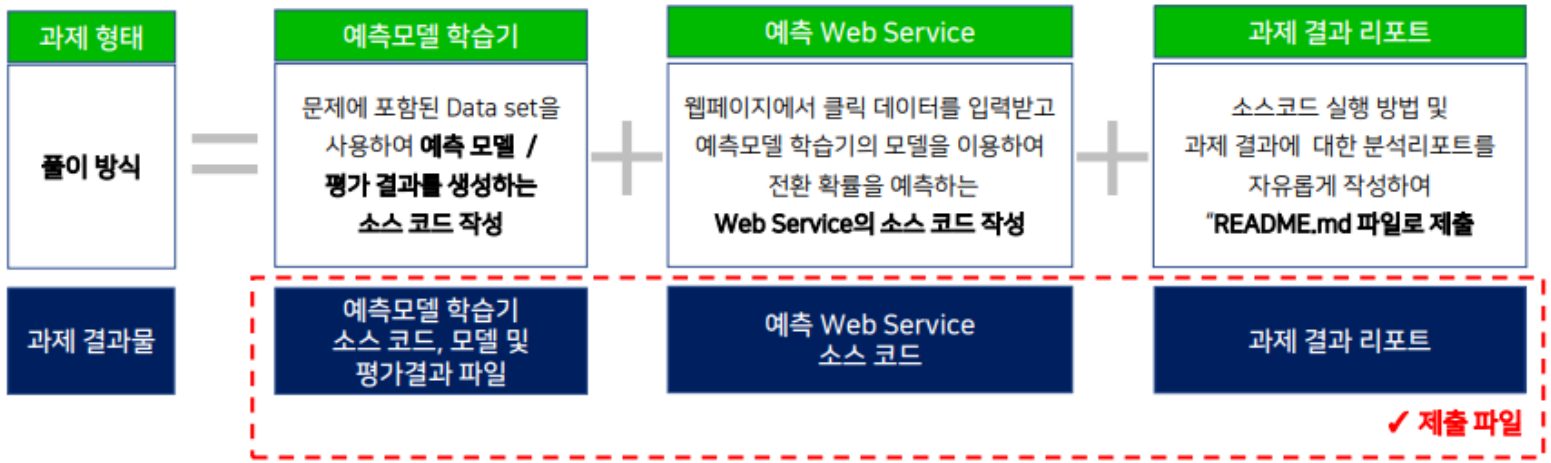
과제명 : CVR Prediction

- 유저 클릭 히스토리를 바탕으로 광고 전환(conversion) 확률을 예측하는 모델을 구현합니다.
- 주어진 Dataset내의 학습데이터를 통해 예측 모델을 생성하고 평가데이터로 모델의 성능을 평가하고 분석합니다.
- Dataset 형식의 데이터를 임의로 입력 받아 예측 CVR을 구하는 웹페이지를 구현합니다.
- 위 작업에 대한 1) 예측모델 학습기의 소스코드, 모델 및 평가결과 파일, 2) 예측 Web Service의 소스코드,
  3) 과제결과 리포트를 제출해 주세요.

개요

- 과제는 다음 페이지에 제시되는 문제를 확인하여 진행합니다.
- 개발 언어 : Python 권장
- 실행 결과를 확인할 수 있도록 실행 조건을 README.md 파일에 작성해 주세요.
- 과제 구현 시 공통으로 사용할 수 있는 코드는 모듈화 해주세요.
- 모델의 성능은 평가 범위가 아닙니다.

과제 풀이 방식



CVR Prediction

[LINE PLUS] AD Platform 채용 연계형 인턴쉽(Machine Learning) 과제

소스코드 실행방법

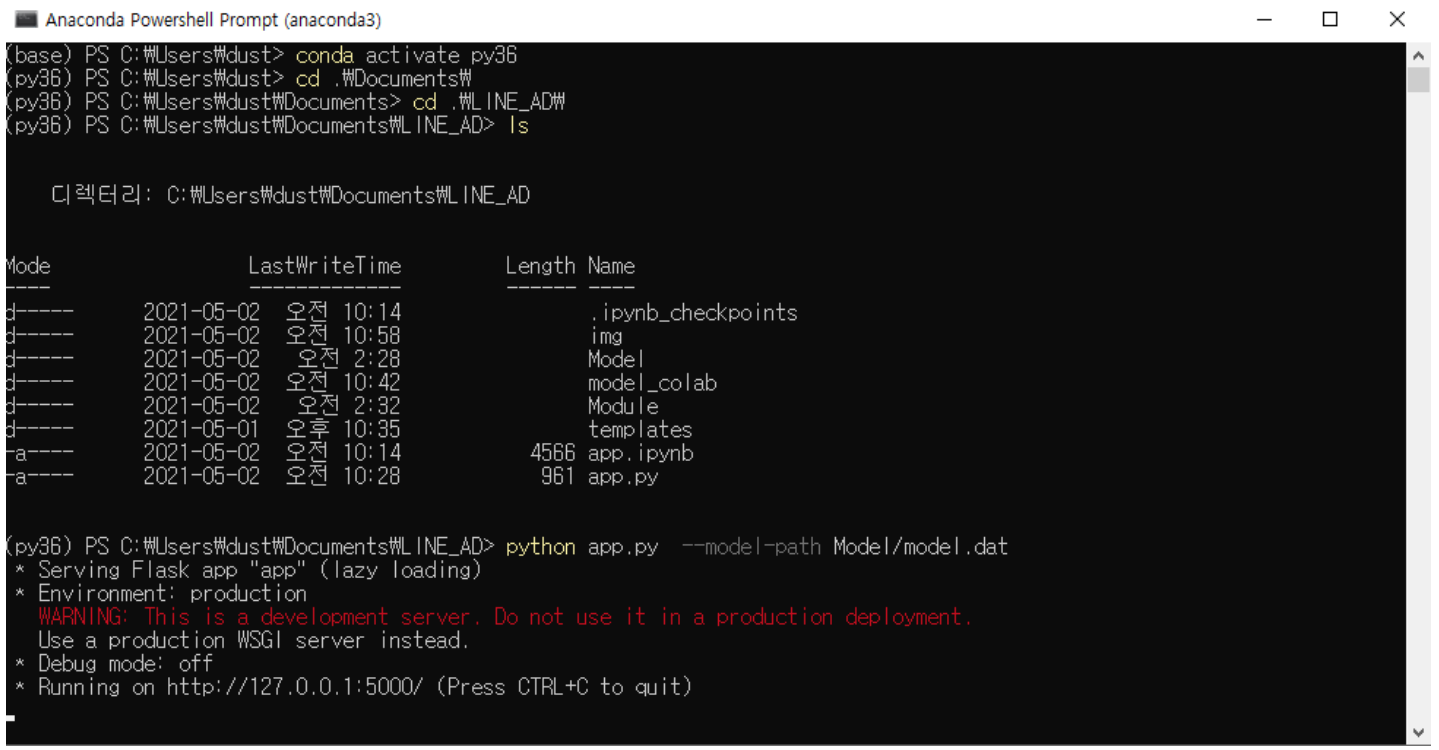
예측모델 학습기 소스코드 실행방법

- 실행환경 : Google Colab (python 3.7.10)
- 외부 라이브러리 : numpy
- 사용 모델 : FFM (직접구현)
- optimizer & regularizer : Adagrad + CosineAnnealingLR (직접구현)

1. 압축파일내의 예측모델 학습기내의 파일들을 코랩으로 이동(권장)
2. dataset 폴더에 과제 데이터 저장
3. train.ipynb 파일 실행

train.ipynb 핵심 코드 설명

실행 화면



Flask를 통해 구현

예측모델 Web Service 소스코드 실행방법

- 실행환경 : Conda env python 3.7.10

```
# Conda PowerShell Prompt 실행
conda activate ... # 가상 환경 실행 권장
cd ... # app.py가 저장된 위치로 폴더경로이동
python app.py --model-path Model/model.dat #run app.py
```

출력된 경로(http://127.0.0.1:5000/)로 이동하여 예측 수행
1.line 파라미터 예측

```
#line 파라미터 예시
0\t-1\t-1\t1598891820\t-1\t0.0\t-1\t7E56C27BFF0305E788DA55A029EC4988\t-1\t-1\t-1\t-1\t-1\t-1\t-1\t-1\
```

입력 후 submit 버튼 클릭

```
#결과 예시
0 0.07376863680866319
```

2.모델 다시불러오기
reload 버튼 클릭

```
Model Reloaded
```

과제결과 분석 리포트

```
              precision    recall  f1-score   support

           1       0.89      1.00      0.94    889930
           0       0.00      0.00      0.00    110070

accuracy                             0.89   1000000
macro avg          0.44      0.50      0.47   1000000
weighted avg       0.79      0.89      0.84   1000000
```

1에 대해서는 예측을 정밀도, 재현율이 다 높지만 0에 대해서는 예측을 하지 못하고 있다.
이는 원래 데이터가 불균형이기 때문에 일어난 것으로 추측된다. 이러한 문제점을 해소하기 위해서는

데이터 측면

1. undersampleing : 클래스를 같은 비율이 되게 데이터를 줄여 학습, 하지만 정보의 손실
2. oversampleing : 클래스를 같은 비율이 되게 데이터를 증가하여 학습. 대표적으로 SMOTE나 GAN을 이용한 데이터를 생성할 수 있다. 하지만 실제 데이터와의 괴리가 있을 수 있다.
3. 데이터 전처리 - click_timestamp : 시계열 컬럼이기에 이에 대한 전처리를 하여 추가 피처로 사용.
4. 데이터 전처리 - 고유값개수를 기준으로 판별하여 컬럼별로 grouping ex) A,A,A,B,B,C,D,E -> A,A,A,B,B,C,C,C
5. 데이터 처리 - 결측치 : 일정한 값으로 치환(빈도 등 고려), ML/DL 을 통한 치환

모델 측면

1. 다른 모델과의 앙상블 : FFM은 상호작용만을 학습하기 때문에 컬럼별 정보가 손실될 수 있다. 각 모델별로 전처리가 다를 수 있다.
2. 다른 Loss사용 : mse등 다른 모델을 사용하거나, 클래스별로 가중치를 부여