9회차 실습 소개

2021

I. 9회차 실습 소개

- **지도학습 및 비지도학습 총 7개 실습 문제**
  - 지도학습의 1, 2번, 3번은 필수, 4번은 선택
  - 비지도학습의 1, 2번은 필수, 3번은 선택

- **실습 시간**
  - 실습: 09:00~11:30, 13:00~15:40
  - 실습 리뷰: 16:00~17:30

- **실습 평가**
  - 분석 후 노트 파일 제출
    - 15:40에 지도학습 및 비지도학습 분석 결과가 포함된 노트파일 제출(1개 파일로 제출해주세요~)
    - won.sang.l@gmail.com으로 제출해주세요.

- **실습 리뷰 및 시상**
  - 16:00
  - 1등: 대상, 베라 3만원(1명), 2등: 최우수상, 베라 2만원(2명), 3등: 우수상, 베라 1만원(3명)
  - 모형 결과(지도학습은 성능 지표, 비지도학습은 정성적 평가), Optional 문제 수행 여부 등 고려
  - 교육 만족도 설문 후 스타벅스 커피 쿠폰 발송

II. 지도학습 실습

1. 보험 가격 예측
2. 신용카드 승인 분류
3. 은행 마케팅
4. (Optional) 인플루언서 선택

1. 보험 가격 예측

– **개요**

- 한 보험사의 1000명에 대한 데이터를 바탕으로 모델링을 하려고 합니다. 주어진 데이터를 활용해서 승인 여부를 모델링하고, 성능을 확인하세요.
- Y는 PremiumPrice이고, X는 그 외 변수(수치형 및 범주형) 입니다.
- 필요한 전처리를 적용: 결측치 Imputing, 수치형 X변수들에 대해 Scaling, Train, Test로 8:2로 파티셔닝
- 회귀모델링을 적용하고, Test셋에 대해서 RMSE를 계산하세요.

– **데이터**

- 데이터: Medicalpremium.csv
- 출처: Kaggle, https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction

2. 신용카드 승인 분류

- **개요**
  - 신용카드 신청자들에 대해서 해당 신청의 승인 여부를 모델링하고자 합니다. Target으로는 고객별로 Good 또는 Bad를 정의해야 하며, credit_record.csv의 status를 주로 참고할 수 있습니다.
  - 필요한 전처리를 적용: 결측치 Imputing, 수치형 X변수들에 대해 Scaling, Train, Test로 8:2로 파티셔닝
  - Test에 대해서 Accuracy, Recall, Precision 등을 구하세요.

- **데이터**
  - 두 데이터는 ID로 Merge할 수 있음
  - 신청 관련 데이터: application_record.csv.zip / 신용 관련 데이터: credit_record.csv.zip
  - 출처: Kaggle, https://www.Kaggle.com/datasets/rikdifos/credit-card-approval-prediction

3. 은행 마케팅

– **개요**

- 은행 고객들에 대한 데이터를 바탕으로, 신규 예금을 신청할 고객인지를 모델링하고자 합니다. Target으로는 y변수를 사용하며 yes, no의 값이 있습니다.
- 필요한 전처리를 적용: 결측치 Imputing, 수치형 X변수들에 대해 Scaling, Train, Test로 8:2로 파티셔닝
- Test에 대해서 Accuracy, Recall, Precision 등을 구하세요.

– **데이터**

- 데이터: bank_full.zip
- 출처: Kaggle, https://www.Kaggle.com/skverma875/bank-marketing-dataset

4. 인플루언서 선택

– **개요**

- 소셜 네트워크 상의 인플루언서들의 데이터 입니다. 각 행은 두 소셜 인플루언서에 대한 변수가 제시되었으며, Choice라는 컬럼은 두 인플루언서 중 사람들이 더 영향력 있다고 선택한 결과입니다.
- 필요한 전처리를 적용: 결측치 Imputing, 수치형 X변수들에 대해 Scaling, Train, Test로 8:2로 파티셔닝
- Test에 대해서 Accuracy, Recall, Precision 등을 구하세요.

– **데이터**

- 데이터: train.csv, test.csv
- 출처: Kaggle, https://www.kaggle.com/competitions/predict-who-is-more-influential-in-a-social-network/overview

III. 비지도학습 실습

1. Book Crossing 이용 도서 추천
2. Behance 패턴 발견
3. (Optional) 미국 국내 항공망 그래프 분석

1. Book Crossing 이용 도서 추천

- **개요**
  - Book Crossing 서비스에 사용된 데이터로 부터, 콘텐츠를 추천하는 모델링을 해보세요.
  - BX-Book-Ratings.csv를 중심으로 데이터를 분석하며, 필요 시 책에 대한 부가정보다 User에 대한 부가정보를 연결해서 사용할 수 있습니다.
  - 파티셔닝을 수행하며, 추천 결과에 대한 RMSE를 확인하세요.

- **데이터**
  - 데이터: BX-CSV-Dump.zip
  - 출처: http://www2.informatik.uni-freiburg.de/~cziegler/BX/

```
pd.read_csv("BX-Book-Ratings_utf8.csv", sep=";")
```

2. Behance 패턴 발견

- **개요**
  - Behance는 온라인에서 이미지를 열람할 수 있는 온라인서비스입니다. Appreciate 데이터를 이용해서 패턴을 발견해보세요.
  - 데이터에는 컬럼명이 없으며, 컬럼명을 부여할 경우, id, item, timestamp입니다.
  - Apriori와 Fpgrowth로 모두 패턴을 발견해보세요

- **데이터**
  - 데이터: Behance_appreciate_1M.gz
  - 출처: https://datarepo.eng.ucsd.edu/mcauley_group/gdrive/behance/

```
!wget https://datarepo.eng.ucsd.edu/mcauley_group/gdrive/behance/Behance_appreciate_1M.gz
!unzip Behance_appreciate_1M.zip
pd.read_csv("Behance_appreciate_1M", header=None, sep=" ")
```

3. 미국 국내 항공망 분석

- **개요**
  - 본 데이터는 2008년 미국 국내 항공 내역을 나타내고 있습니다. 이 데이터를 이용해서 항공네트워크를 분석해보세요.
    - Origin: 출발 공항 코드
    - Dest: 도착 공항 코드
  - 지도학습을 이용해서 각 공항 별 연착을 예측해보세요

- **데이터**
  - 데이터: 2008.zip
  - 출처: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7

Q&A