BigData HW3_report
20170735
컴퓨터공학과
이승준

1. Binary_Classification

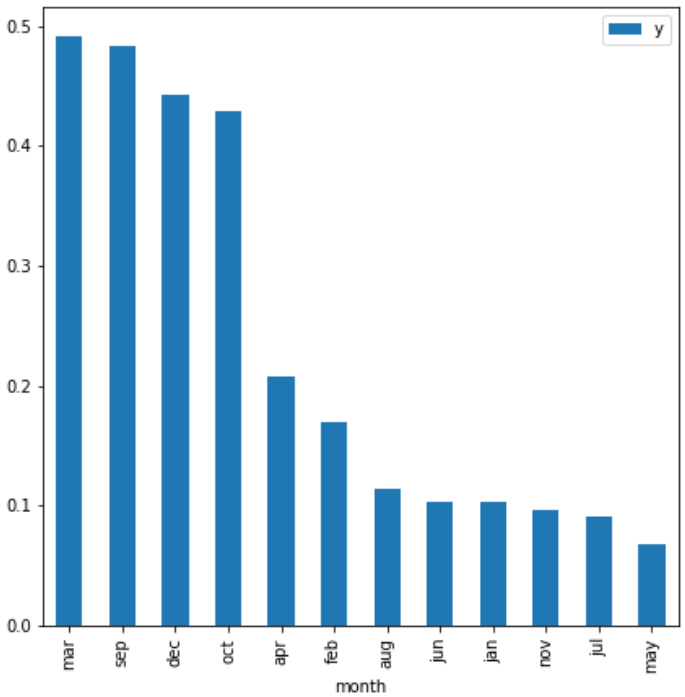
먼저, csv 파일에서 각각 train data 와 test 데이터들을 받고, string data들을 숫자들로 바꿨다.

```
dot_data = sktree.export_graphviz(clf, out_file=None)
graph = graphviz.Source(dot_data)
graph.render('/Users/leeseungjoon/Desktop/2018-2/BigData/HW3/Binary/discrete')
```

이 코드들을 이용해 각각의 노드들의 기준이, 어떤 컬럼에 해당하는 값의 크기가, 어떤 기준보다 작거나 큼으로 하고 있음을 알 수 있었다. 즉, 부등호를 이용해 데이터들을 나누는 것이다. 이에,

```
#f, ax = plt.subplots(1, 1, figsize=(7, 7))
#train_x[['poutcome','y']].groupby(['poutcome'],as_index=True).mean().sort_values(by='y',
ascending=False).plot.bar(ax=ax)
#train_x[['balance','y']].groupby(['balane'],as_index=True).mean().sort_values(by='y',
ascending=False).plot.bar(ax=ax)
```

이 코드들을 이용해 다음과 같은 그림을 얻을 수 있는데, month에 대한 column에서, mar, sep, 등등과 y가 1인 비율을 나타낸 것이다.



Decision tree의 기준이 부등호를 이용한다는 점과, 위의 그래프의 정보를 이용해, string 들을 정수형으로 바꿀 때 y에 영향을 많이 끼치는 순서대로 숫자로 바꿨다. 즉, 데이터들을 숫자로 바꿀 때 아무런 숫자나 넣은 것이 아니고, mar은 0, sep은 1 dec은 2 등등으로 바꾸며 부등호를 살릴 수 있도록 숫자를 바꿔줬다.

이후 train_data에 decisiontree를 fit하고, 이를 통해 test_data에 대한 예상값을 만들었다.

kaggle score는 0.51954 이다.

2. Crime_data

먼저, 1번과 똑같이 데이터를 읽어온다. 그 후, Dates 컬럼을 날짜, 시간 두 컬럼으로 나누었다. 년, 월, 일에 구분을 주기 위해 2018-10-11 이란 데이터는 각각 10000, 100, 1 을 곱해 20181011이란 데이터로 만들었다. time도 마찬가지로 그렇게 만들었다. 그런데 날짜에 관한 컬럼은 빼는게 정확도가 더 높게 나와, 날짜 컬럼은 뺀 것으로 채택했다. 시간 컬럼 또한 마찬가지였다.

나머지 string 데이터들은 숫자로 변환해 줬다. train에는 resolution이 있는데, test에는 없어서, train_data에서 'Resolution' 컬럼을 뺀 나머지들로 DeicisionTreeClassfier를 fit하고, 이에 따라 test_data의 resolution을 만들고, 전체 test_data로 crime_category를 예측했는데, resolution이 없는 것이 더 점수가 높게 나왔다.

Description column은 삭제했고, address 데이터들을 이용해보고자 했지만, 하지 않은 것과 점수가 똑같았다.

그리고 마지막에 decisiontreeclassifier로 test 데이터에 대한 crime_category를 예측했다.

Kaggle score는 2.57505 이다.