Bank Marketing Dataset (Predicting Term Deposit Subscriptions)

4조 김채윤, 문희원,
이동규, 이주환,
이중기, 장준규

Bank Marketing

I. 도입

Telemarketing is a method of direct marketing to advertise or sell of goods or services by phone.

Find the best strategies to **improve** for the next marketing campaign. How can the financial institution have a greater **effectiveness** for future marketing campaigns?

Propose a **data mining** (DM) approach to **predict** the success of telemarketing calls for selling bank long-term deposits.

Acquired Data Site:
https://www.kaggle.com/janiobachmann/bank-marketing-dataset

Bank Marketing

II. 데이터 소개 및 처리

Overview

Dataset info

Number of variables	17
Number of observations	11162
Missing cells	0 (0.0%)
Duplicate rows	0 (0.0%)
Total size in memory	1.4 MiB
Average record size in memory	136.0 B

Variables types

Numeric	7
Categorical	6
Boolean	4
Date	0
URL	0
Text (Unique)	0
Rejected	0
Unsupported	0

Warnings

`balance` has 774 (6.9%) zeros — Zeros

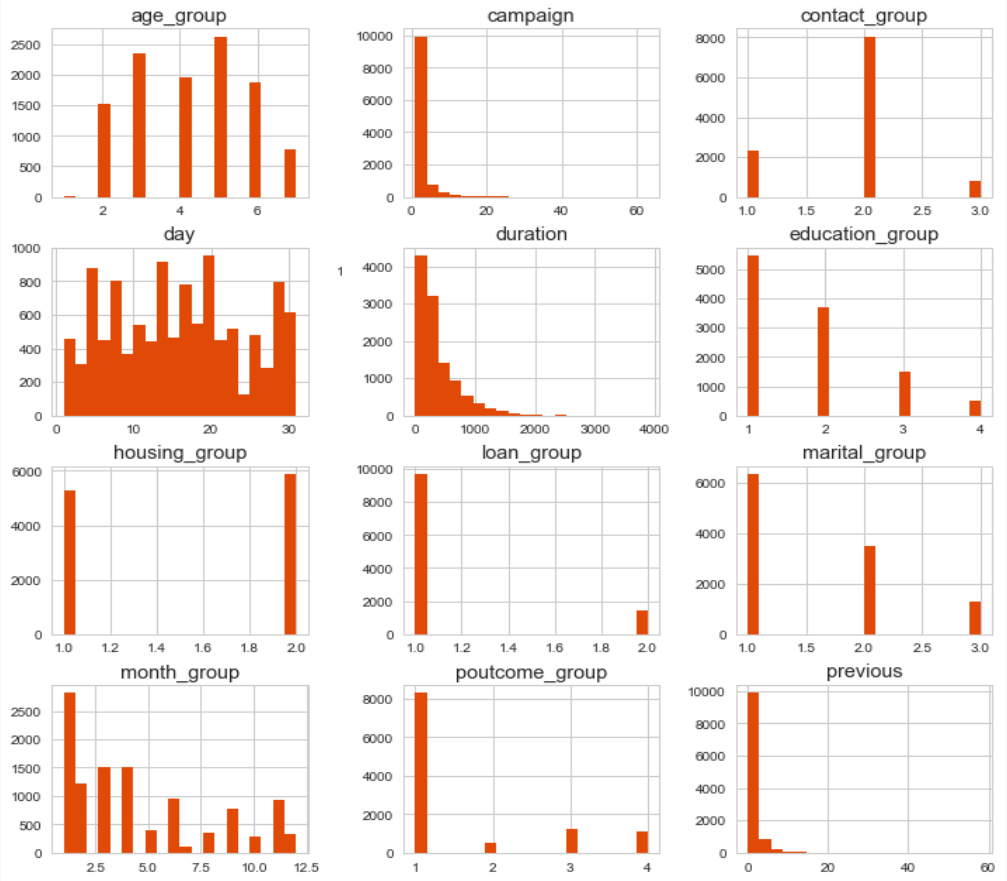
`previous` has 8324 (74.6%) zeros — Zeros

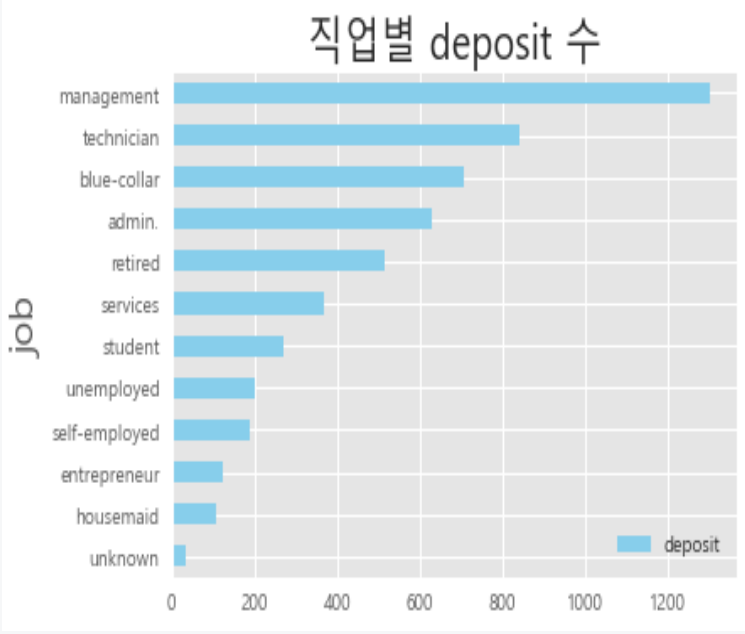
데이터 소개

변수	변수설명	변수	변수설명
Default	파산여부	**Month**	마지막 연락 월
Duration	통화 지속 시간	**Day**	마지막 연락 일
Campaign	캠페인 기간 연락횟수	**Previous**	이전 캠페인에서 이 고객에게 연락한 횟수
Pday	마지막 연락으로부터의 일수	**Poutcome**	이전 캠페인에서의 결과

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	deposit
0	59	admin.	married	secondary	no	2343	yes	no	unknown	5	may	1042	1	-1	0	unknown	yes
1	56	admin.	married	secondary	no	45	no	no	unknown	5	may	1467	1	-1	0	unknown	yes
2	41	technician	married	secondary	no	1270	yes	no	unknown	5	may	1389	1	-1	0	unknown	yes
3	55	services	married	secondary	no	2476	yes	no	unknown	5	may	579	1	-1	0	unknown	yes

II. 데이터 소개 및 처리
데이터 분포도





Data Cleaning

```
df['housing']=df['housing'].map({'no':0.,'yes':1}).astype('float32')

df['job']=df['job'].map({'management':0., 'blue-collar':1, 'technician':2., 'admin.':3, 'services':4,
                         'retired':5, 'self-employed':6, 'student':7, 'unemployed':8, 'entrepreneur':9,
                         'housemaid':10., 'unknown':11}).astype('float32')
```

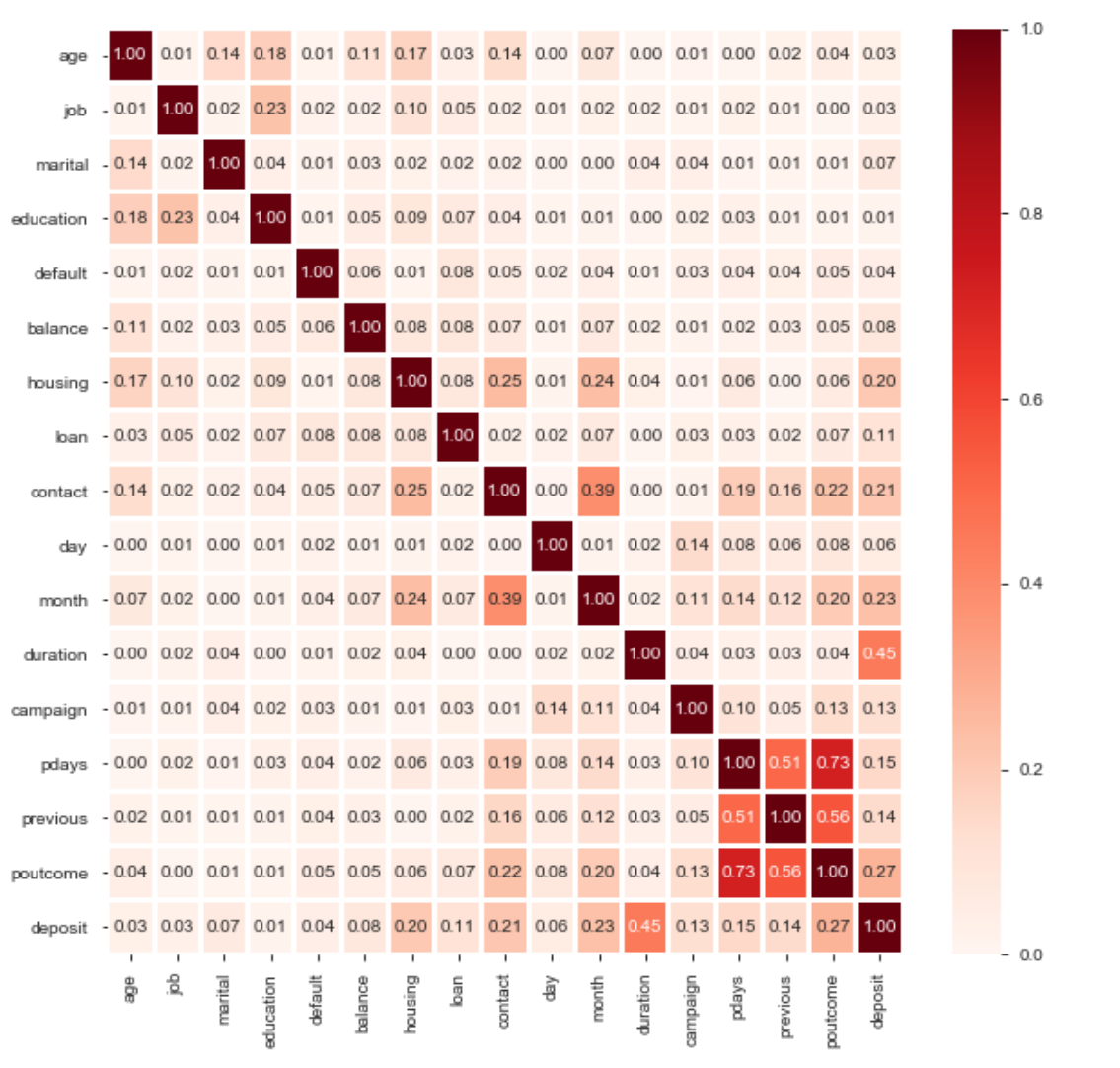
Standardization

```
sc = preprocessing.StandardScaler()
data=sc.fit_transform(data)
```

Applying OneHotEncoder

```
train_label = utils.to_categorical(train_label) # 0~9 -> one-hot vector
test_label = utils.to_categorical(test_label) # 0~9 -> one-hot vector
```

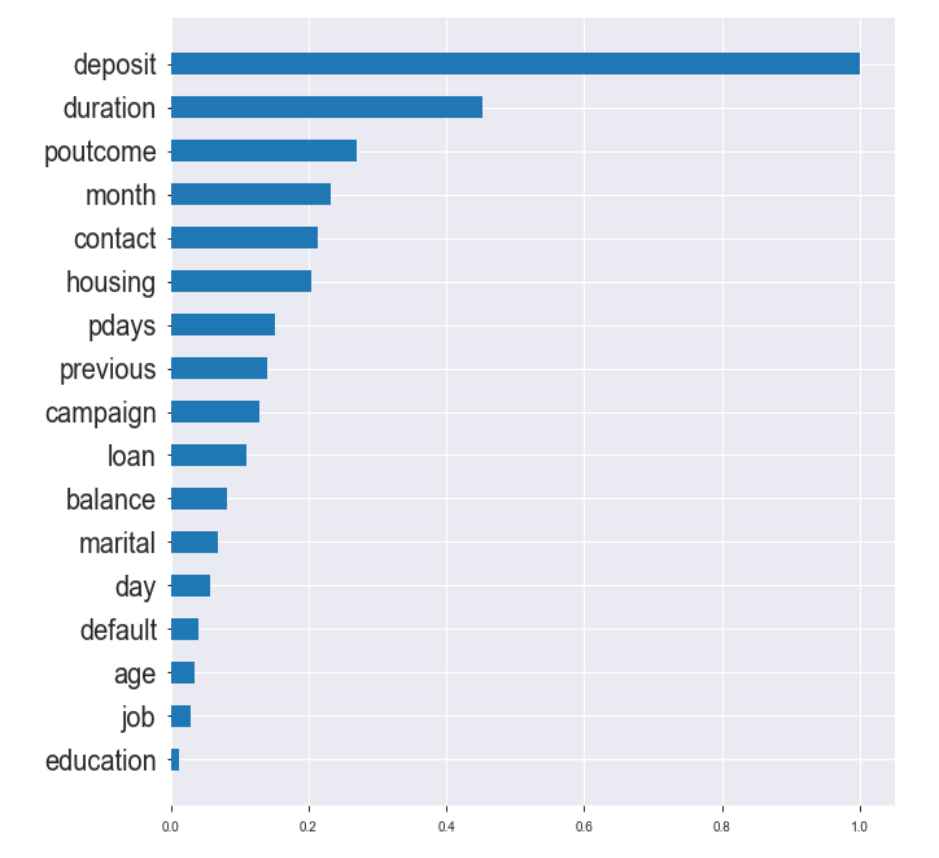
II. 데이터 소개 및 처리
피어슨 상관 관계



상관계수가 1에 가깝다는 것은 두 변수 간의 강한 양의 상관관계가 있다.

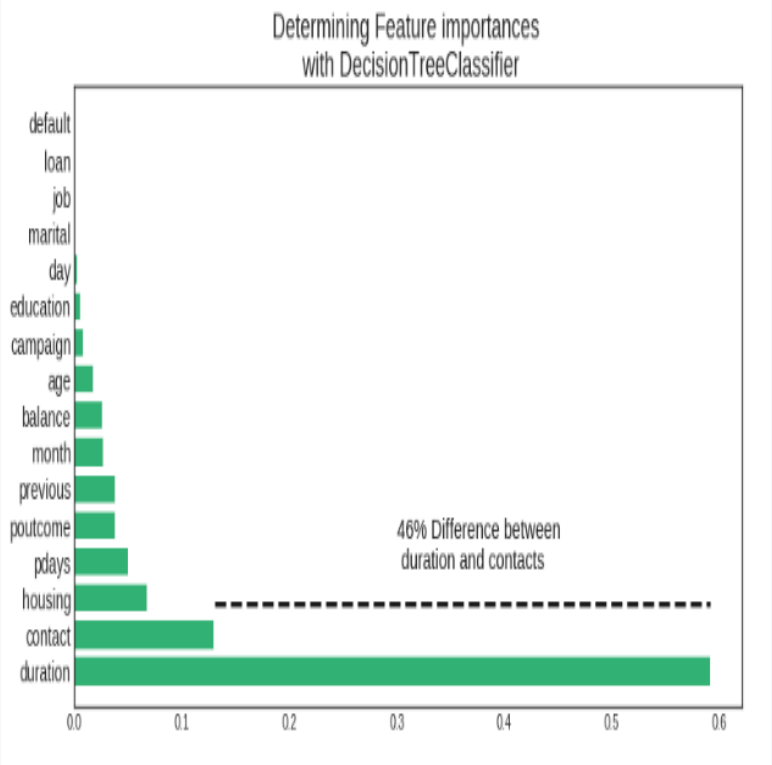
0에 가까울 수록 두 변수 사이의 관계가 없다.

II. 데이터 소개 및 처리
상관 관계로 부터 중요한 열 추출



머신러닝을 활용해 특징 중요도 추출

Decision Tree Classifier

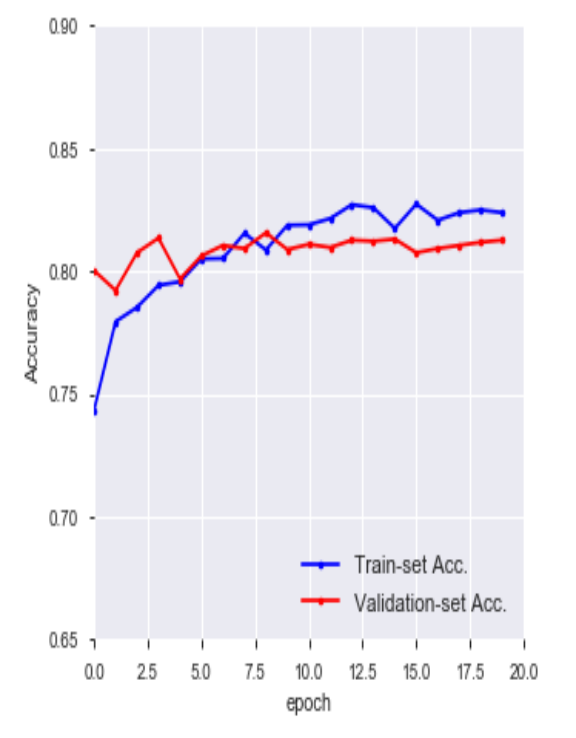


https://www.kaggle.com/janiobachmann/bank-marketing-campaign-opening-a-term-deposit

Gradient Boosting Classifier

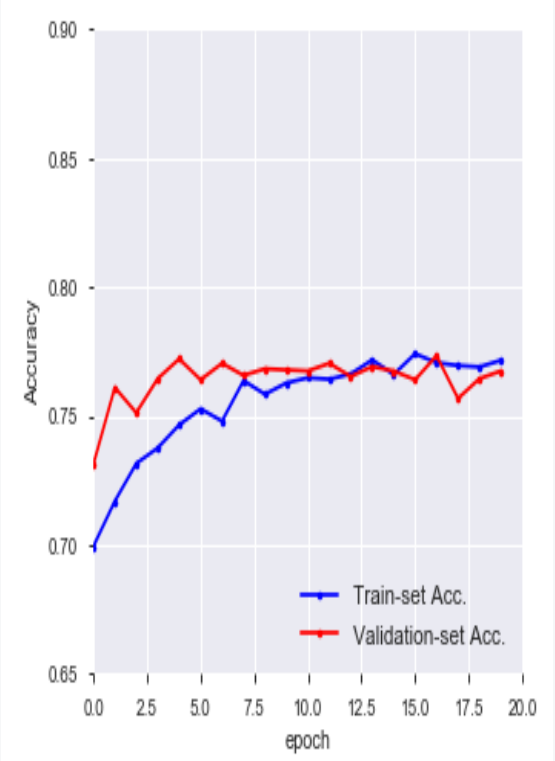
특정 열을 선택하여 딥러닝한 결과

상관도 0.2 이상 열



loss (cross-entropy) : 0.40047685369052755
test accuracy : 0.8187519

상관도가 큰 상위 2개 열



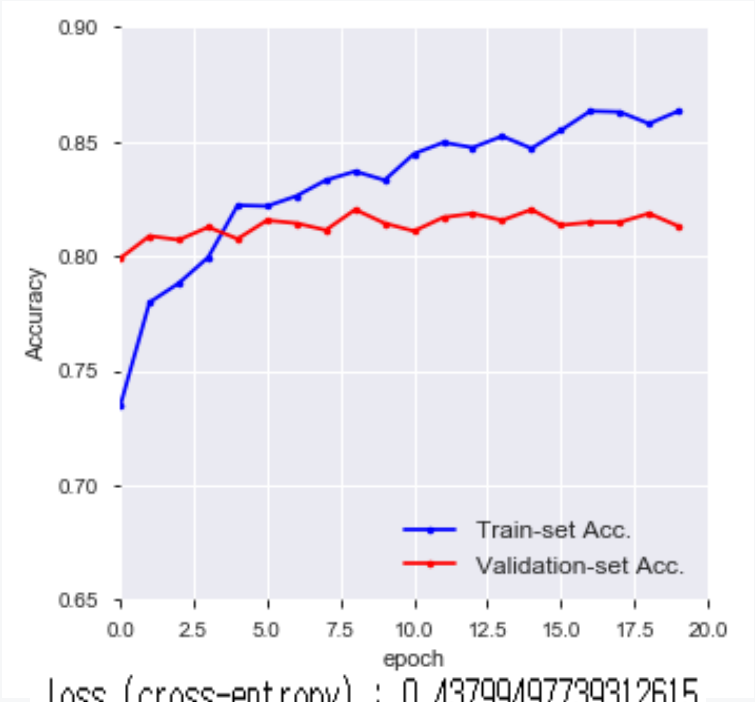
loss (cross-entropy) : 0.48313790471669843
test accuracy : 0.77724695

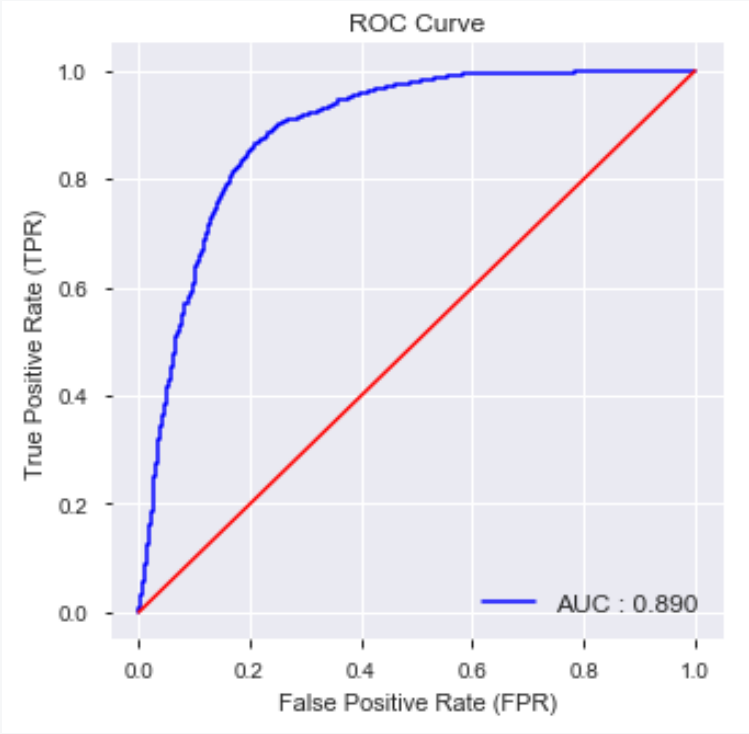
모델	
Trian : Test	7:3
Batch size	100
은닉층의 수	4
은닉층 1 노드 수	256
은닉층 2 노드 수	512
은닉층 3 노드 수	512
은닉층 4 노드 수	256
활성화함수	elu

Bank Marketing

III. 결과

III. 결과
특정 열을 선택하여 딥러닝한 결과





	Precision	recall	F1-score	Support
Not Deposit	0.87	0.79	0.83	1792
Deposit	0.78	0.86	0.82	1557
Micro avg	0.82	0.82	0.82	3349
Macro avg	0.82	0.82	0.82	3349
Weighted avg	0.83	0.82	0.82	3349

Bank Marketing

IV. 요약 및 결론

요약 및 결론

1) 상관도 분석

- 상관도가 높은 열 이외에 다른 열이 추가하여 계산을 했음에도 정확도에 치명적인 영향은 없었다.

2) 하이퍼 파라미터

- 딥러닝 모델에 사용되는 **하이퍼 파라미터**(hyper-parameter)를 적절히 선택하는 것은 중요하지만 **데이터 전처리**(data processing)가 더 중요한 요소라고 생각한다.

3) 머신러닝/딥러닝에 적합한 문제는 무엇일까?

- 우리가 알고자 하는 현상 혹은 문제를 **적절히 정의**하는 것이 중요하다.

B a n k M a r k e t i n g

감사합니다