Regression Analysis 보고서

2011100038 박병진
2011100074 김혁준
2012100005 강나루

Contents

B. 데이터 전처리

D. 결론

A. Intro

C. 회귀분석

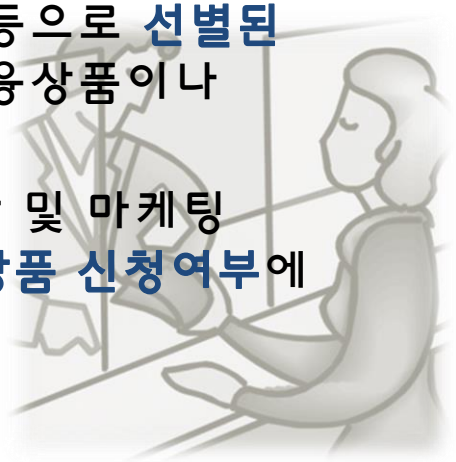
Bank term deposit

Bank Direct Marketing

√ **직접마케팅**

대면접촉이나 전화, 이메일, 편지 등으로 **선별된 고객에게 접촉**함으로써 새로운 금융상품이나 서비스에 대해 홍보하는 전략.

√ 해당 데이터는 고객들의 인적 사항 및 마케팅 전략 과정의 여러 정보, **정기예금상품 신청여부**에 대한 자료이다.



Objective :

어떤 고객들에게 홍보를 하여야 효과적으로 홍보할 수 있는가?

Description

Description

Variable	Type	Description
age	numeric	연령
job	categorical	직업의 종류 (관리직, 무직, 노동직, 기술직,서비스 등)
marital	categorical	혼인상태
education	categorical	최종학력(교육수준)
default	binary	부채 유무
balance	numeric	연평균 통장 잔고액수(단위 : 유로)
housing	binary	주택담보대출 유무
loan	binary	개인대출 유무

Description

Description

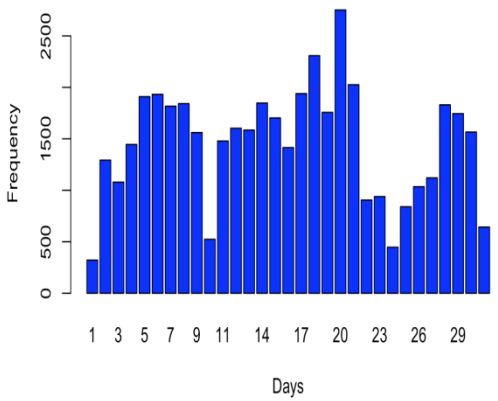
Variable	Type	Description
contact	categorical	연락방법(휴대전화, 집전화, unknown)
Day	numeric	가장 최근에 연락한 날짜의 날(1~31)
Month	categorical	가장 최근에 연락한 날짜의 달(1~12)
Duration	numeric	가장 최근에 연락한 날의 지속시간(단위 : 초)
Campaign	numeric	최근 홍보기간 동안에 해당 고객과 연락한 횟수
Pdays	numeric	최근거래 이후에 지난 기간(단위는 일)
Previous	numeric	최근 홍보 이전에 해당 고객과 연락한 횟수
Poutcome	categorical	이전 홍보의 결과(거래를 계속 하는지의 여부)
Y	binary	정기적금 가입여부

Removal

① 변수 제거

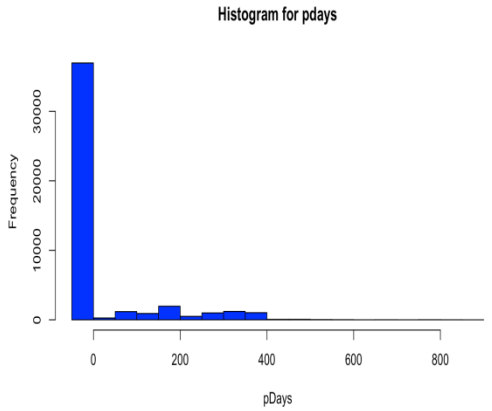
√ **day : 가장 최근 연락 시점의 일 (1일 ~ 31일)**

- Day의 경우 대체로 고르게 분포되어 있고, month 변수만으로도 정보가 충분하다고 판단되어 제거.



√ **pdays : 고객에게 연락 이후 경과일**

- 첫 고객의 경우 –1로 기록되어 있는데 전체 자료의 81%차지.
- 연락의 유무가 중요한 것으로 생각되어 previous와 유사한 정보라고 판단되어 제거.



Imputation

② 결측값 대체

Age

- 결측치(NA) : 4obs
- 연속형 변수이므로 대표적 imputation 방법인 median(39)로 대체.

```
      age
Min.   :18.00
1st Qu.:33.00
Median :39.00
Mean   :40.94
3rd Qu.:48.00
Max.   :95.00
NA's   :4
```

Marital

- 결측치(NA) : 5obs
- 연령에 따라 영향을 많이 받을 것이라 생각.
- 30세 이상 : married
  30세 미만 : single로 대체.

```
     marital
 divorced: 5207
 married :27209
 single  :12790
 NA's    :    5
```

기타 결측치

- Unknown 역시 결측치로 판단해야 하나 결측치의 수가 많아서 unknown을 하나의 범주로 취급.

variable	Miss
Job	288
Education	127
Contact	13020
Poutcome	36595

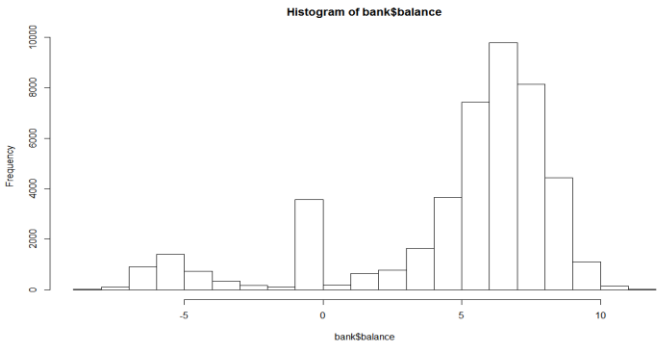
Transformation

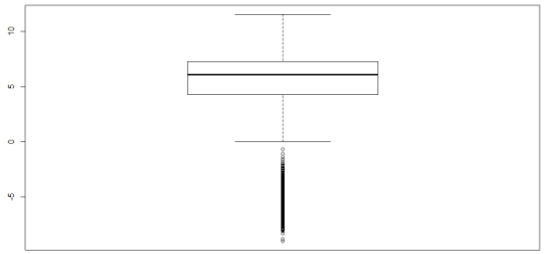
③ 변수 변형

√ Balance 변수 : 자료의 변수가 비대칭적(left-skewed)

- 정규성 모형을 만들기 위해서 log변환을 시도.
- 단순 로그 변환 시 음의 값을 가지는 obs 때문에 문제 발생
- log(0)의 값이 발생하는 것을 방지하기 위해서 로그의 진수에 1을 더해줌 => log(x+1)
- log의 진수가 음수인 경우 진수값에 -를 곱해주고 log앞에 -를 붙여줌.

```
bank$balance<-ifelse(bank$balance>=0,log(bank$balance+1),-log(-bank$balance+1))
```





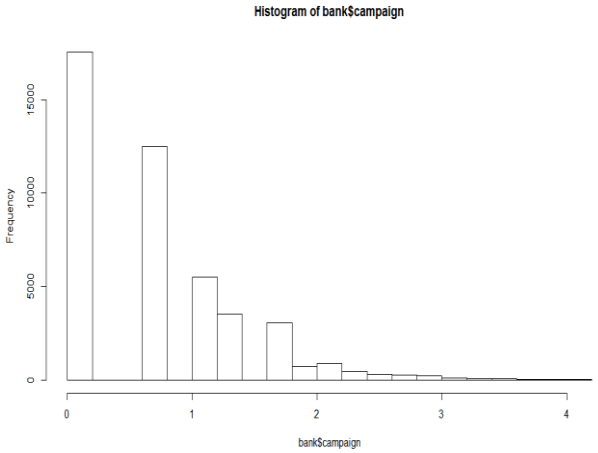
Transformation

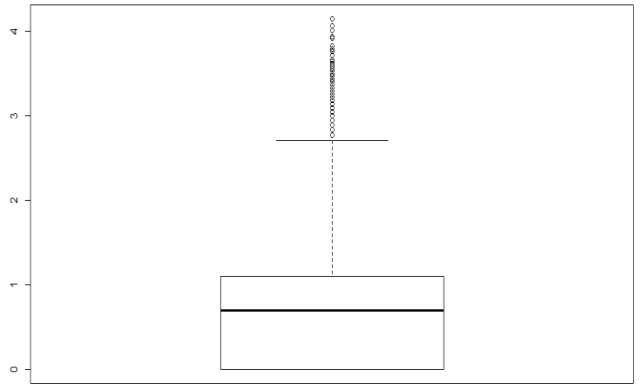
③ 변수 변형

√ Campaign 변수 : 자료의 변수 비대칭적(right-skewed)

- 정규성 모형을 만들기 위해서 log변환을 시도.
- 음의 값과 0이 존재하지 않으므로 단순 로그 변형

```
bank$campaign <- log(bank$campaign)
```





Response Variable

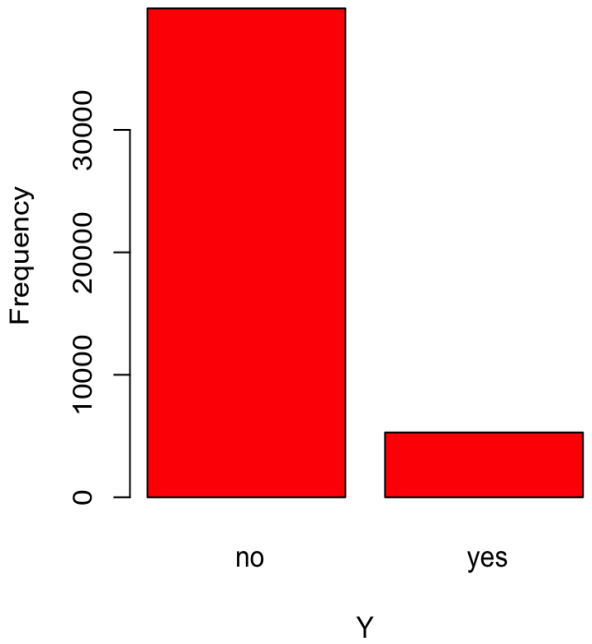
④ 반응 변수

√ **반응변수 y**

– 결측치(NA) 1개 존재

=>반응변수는 예측에 있어서 매우 중요한 정보라고 생각해서 **결측치를 제거**함.

– No가 전체 반응변수 중에서 88.3%를 차지하므로 Sensitivity가 Specificity에 비해서 상대적으로 **낮게** 나올 것으로 예상.



Model selection

예측력을 높이기 위해서 Validation set 접근법 사용.

> Validation set을 6:4으로 나누었을 때

```
> miss.err = 1-sum(diag(ctable))/sum(ctable) # Misclassification Rate
> miss.err
[1] 0.09885315
```

> Validation set을 7:3으로 나누었을 때

```
> miss.err = 1-sum(diag(ctable))/sum(ctable) # Misclassification Rate
> miss.err
[1] 0.09772353
```

Error rate을 통해서
Validation partition을 7:3으로 결정.

Model selection

적합한 변수 설정을 위한 Model selection

› Stepwise selection(AIC)

```
> fit2$anova
          Step Df    Deviance Resid. Df Resid. Dev      AIC
1             NA          NA      31814   15199.53 15281.53
2  - default   1 0.04871219      31815   15199.57 15279.57
3 - previous   1 0.75218979      31816   15200.33 15278.33
4      - age   1 0.90794475      31817   15201.23 15277.23
```

› Stepwise selection(BIC)

```
> fit2$anova
          Step Df    Deviance Resid. Df Resid. Dev      AIC
1             NA          NA      31814   15199.53 15639.01
2      - job  11 64.5300918      31825   15264.06 15585.63
3  - default   1  0.1554161      31826   15264.21 15575.06
4 - previous   1  0.9327082      31827   15265.14 15565.28
5      - age   1  4.4565240      31828   15269.60 15559.01
```

√ 오분류율을 살펴보았을 때

BIC(0.09764864) 〈 AIC(0.09772353)

그러나, job 변수가 중요할 것으로 생각되어

AIC방법 중 stepwise방법으로 적합

=〉 변수 default, previous, age를 제외

ROC Curve

ROC Curve를 통한 Cutoff설정

› Cutoff를 0.5로 설정했을 시

```
> ctable
       Predicted
Actual     0     1
   no  11499   295
   yes  1010   550
```

√ **Missclassification Rate :** 0.09772353

√ **Prediction accuracy :** 0.9022765

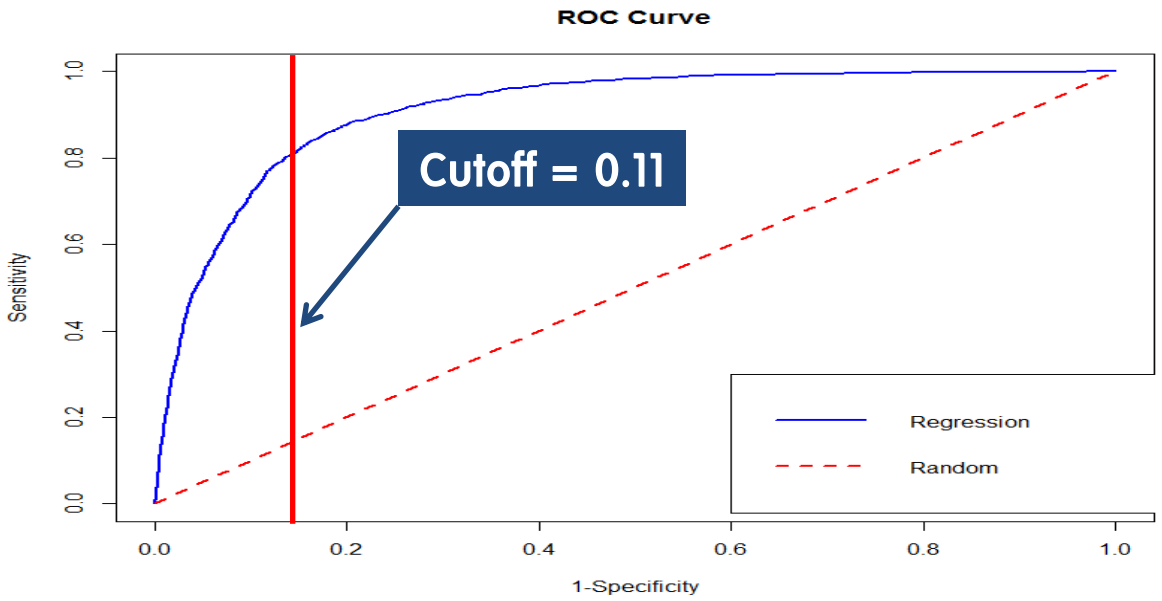
√ **Sensitivity :** 0.3525641

√ **Specificity :** 0.9749873

ROC Curve

ROC Curve를 통한 Cutoff설정

> ROC Curve



√ AUC면적이 0.9105532로 약 91%의 설명력을 가지고 있어 설명력이 좋은 것으로 판단.

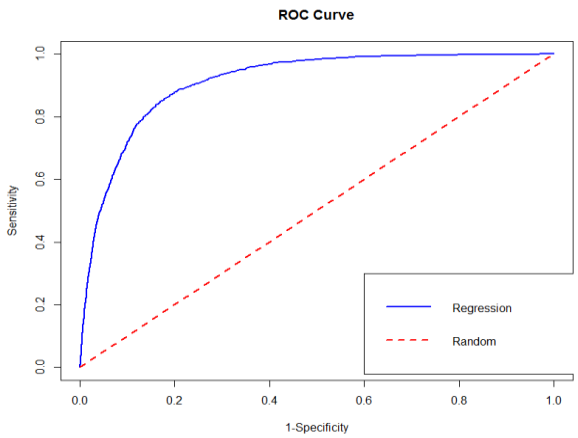
√ ROC Curve를 참고하여 Cutoff 설정. 원 자료에서 y값이 NO가 많은 점을 고려하여 민감도를 높이는 방향으로 Cutoff설정 => 0.11로 Cutoff 결정

(민감도 : 0.8532051, 특이도 : 0.8245718)

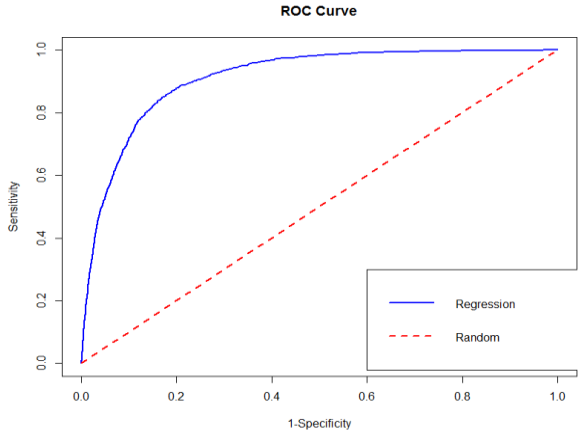
ROC Curve

ROC Curve를 통한 Cutoff설정

› Test Set vs Training Set



Training Set



Test Set

√ **AUC면적**이 각각 0.9084461과 0.9105532로 두 Set의 설명력은 비슷한 수준

√ **Test set**에서도 잘 적합 되고 있음을 확인할 수 있다.

Regression Model

Regression Coefficient 해석

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.495e+00	1.496e-01	-16.683	< 2e-16	***
job blue-collar	-3.409e-01	8.704e-02	-3.917	8.98e-05	***
job entrepreneur	-4.056e-01	1.511e-01	-2.683	0.007288	**
job housemaid	-5.168e-01	1.599e-01	-3.232	0.001230	**
job management	-1.763e-01	8.693e-02	-2.029	0.042496	*
job retired	1.993e-01	1.045e-01	1.907	0.056488	.
job self-employed	-2.232e-01	1.297e-01	-1.721	0.085168	.
job services	-1.602e-01	9.936e-02	-1.612	0.106900	
job student	3.743e-01	1.278e-01	2.929	0.003396	**
job technician	-1.955e-01	8.262e-02	-2.366	0.017992	*
job unemployed	-2.232e-01	1.324e-01	-1.685	0.091908	.
job unknown	-3.862e-01	2.877e-01	-1.342	0.179467	
marital married	-1.818e-01	7.064e-02	-2.573	0.010075	*
marital single	7.534e-02	7.576e-02	0.994	0.320027	
education secondary	1.492e-01	7.731e-02	1.929	0.053705	.
education tertiary	3.895e-01	8.894e-02	4.379	1.19e-05	***
education unknown	2.437e-01	1.234e-01	1.975	0.048308	*
balance	3.845e-02	6.884e-03	5.585	2.34e-08	***
housing yes	-6.979e-01	5.180e-02	-13.473	< 2e-16	***
loan yes	-3.642e-01	7.093e-02	-5.134	2.83e-07	***
contact telephone	-1.258e-01	8.821e-02	-1.426	0.153748	
contact unknown	-1.548e+00	8.720e-02	-17.758	< 2e-16	***
month aug	-6.708e-01	9.235e-02	-7.264	3.77e-13	***
month dec	7.511e-01	2.077e-01	3.616	0.000299	***
month feb	-2.628e-01	1.006e-01	-2.612	0.009014	**
month jan	-1.213e+00	1.418e-01	-8.554	< 2e-16	***
month jul	-7.856e-01	9.233e-02	-8.509	< 2e-16	***
month jun	3.271e-01	1.077e-01	3.036	0.002398	**
month mar	1.386e+00	1.400e-01	9.898	< 2e-16	***
month may	-4.701e-01	8.509e-02	-5.524	3.31e-08	***
month nov	-8.614e-01	9.850e-02	-8.744	< 2e-16	***
month oct	7.473e-01	1.306e-01	5.722	1.05e-08	***
month sep	8.128e-01	1.399e-01	5.811	6.21e-09	***
duration	4.214e-03	7.689e-05	54.809	< 2e-16	***
campaign	-2.816e-01	3.611e-02	-7.797	6.33e-15	***
poutcome other	2.223e-01	1.047e-01	2.123	0.033726	*
poutcome success	2.186e+00	9.485e-02	23.042	< 2e-16	***
poutcome unknown	-1.587e-01	6.783e-02	-2.340	0.019301	*

√ **job기준**으로 보았을 때 student는 다른 변수들의 조건이 동일할 때 **로짓**이 admin 직업을 가질 때에 비해서 **3.743e-01**만큼 **증가**한다고 해석할 수 있음.

Regression Coefficient 해석

√ 로짓을 증가시키는 조건

- job이 student인 경우
- 학력이 높을수록
- 결혼 여부의 경우 single의 p-value가 높게 나와서 판단불가
- balance가 많을수록
- 9,10,12월 달에 campaign을 시행했을 때
- duration이 길수록
- poutcome이 other이거나 success일 때

√ 로짓을 감소시키는 조건

- job이 housemaid, entrepreneur인 경우
- 기혼인 경우
- 주택담보대출이 있을 경우
- loan이 있을 경우
- 7,8,11월 달에 campaign을 시행했을 때
- 이번 캠페인에서 접촉횟수가 많은 경우

Regression Coefficient 해석

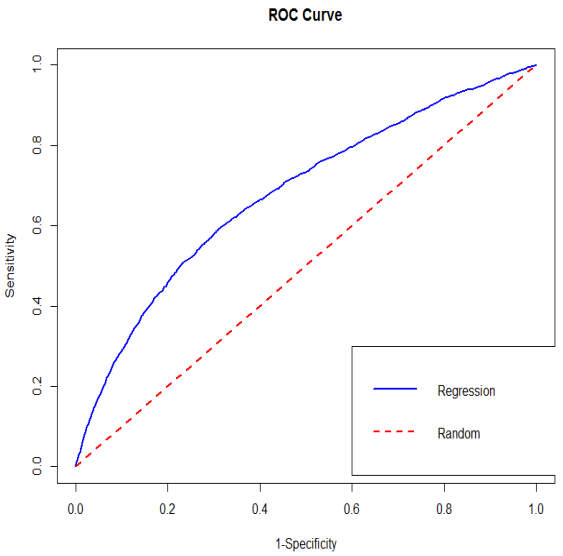
Bank term deposit data	Variable
Bank client data	Age, job, marital, education, default, balance, housing, loan
Campaign data	Contact, day, month, duration
Other attributes	Campaign, pdays, previous, poutcome

√ 고객과 직접 연관되는 변수들은 **bank client data**이므로 따로 떼어내어 볼 필요가 있다고 판단
(Campaign data나 Other attributes는 캠페인과 연관된 변수들로 홍보대상 고객분류에는 **필요가 없다**고 판단)

Regression Model

Regression Coefficient 해석

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.291183   0.143646 -15.950  < 2e-16 ***
age                 0.005275   0.002153   2.449 0.014311 *
job blue-collar    -0.373777   0.072650  -5.145 2.68e-07 ***
job entrepreneur   -0.523576   0.127179  -4.117 3.84e-05 ***
job housemaid      -0.559163   0.133853  -4.177 2.95e-05 ***
job management     -0.249817   0.072459  -3.448 0.000565 ***
job retired         0.408623   0.092417   4.421 9.80e-06 ***
job self-employed  -0.261633   0.107522  -2.433 0.014962 *
job services       -0.237074   0.082470  -2.875 0.004045 **
job student         0.480842   0.108115   4.448 8.69e-06 ***
job technician     -0.288327   0.068504  -4.209 2.57e-05 ***
job unemployed      0.008749   0.106751   0.082 0.934682
job unknown        -0.615180   0.236399  -2.602 0.009260 **
marital married    -0.169096   0.058278  -2.902 0.003713 **
marital single      0.214147   0.066111   3.239 0.001199 **
education secondary 0.238678   0.064529   3.699 0.000217 ***
education tertiary  0.537110   0.074184   7.240 4.48e-13 ***
education unknown   0.282475   0.101993   2.770 0.005613 **
default yes        -0.324950   0.190728  -1.704 0.088431 .
balance             0.068478   0.006055  11.309  < 2e-16 ***
housing yes        -0.720346   0.038297 -18.810  < 2e-16 ***
loan yes           -0.460992   0.059750  -7.715 1.21e-14 ***
```



√ ROC Curve에서 AUC를 계산하면 0.6788326으로 설명력이 다소 떨어진다.

따라서,

참고용으로 분석결과를 이용하는 것이 바람직해 보임

Conclusion

결론

√ 분석결과를 토대로 다음의 조건을 만족시키는 고객에게 홍보를 하는 것이 효과적이라 판단됨

- 직업이 student이거나 .admin. 직업이 housemaid, entrepreneur인 경우는 선호되지 않음.
- 고학력자
- 잔고가 많은 고객
- 주택담보대출과 loan이 있는 고객은 비선호
- 결혼여부의 경우 single일 경우(Bank client data 분석결과 참고)
- 나이의 여부는 크게 중요하지 않은 것으로 판단됨

Limitation

한계

√ 변수들간의 교호작용 고려X

– 교호작용이 없는 단순한 로지스틱 모형에 대해서 적합 시켰으나 변수들간의 교호작용이 있을 수 있음.

√ 범주형 자료 중 Unknown의 미처리

– Unknown의 수가 많아서 imputation시 위험성이 존재하여 하나의 범주로 두고 분석.

√ 변수변환의 문제점

– 변수변환을 시켰음에도 불구하고 정규성 모형을 완전히 충족시키지 못함.
=〉 더 좋은 transformation방법이 필요할 것으로 판단됨.