이란 통신사 데이터셋 분석

TEAM2 이런이란
2022113156 이예령
2022113170 장정의
2022113185 전지현

TABLE OF CONTENTS

INTRODUCTION & DATASET SELECTION

이란 통신사 데이터셋

이란 통신사 데이터베이스에서 수집된 데이터로 이탈을 제외한 모든 속성은 9개월 동안 집계하였고,
이탈 속성은 12개월 말 고객의 상태

Feature 소개

Call failure, Complains, Subscription length, Charge amount, Seconds of use, Frequency of use,
Frequency of SMS , Distinct called numbers, Age group, Traffic plan, Status, Age, Customer value, Churn

데이터 선택 이유

1) 너무 과학적이거나 난해한 feature가 섞이지 않은 데이터
2) features 및 instance의 개수가 너무 많지 않은 데이터
3) EDA 과정을 통해 유의미하게 탐색하고 활용할 수 있는 데이터.

INTRODUCTION & DATASET SELECTION

가설

1) Call failure, Complains, Seconds of use가 큰 영향
2) Age, Age group간의 공선성이 있을 것

목표

기존 feature 중 다중공선성을 줄이고 적은 feature을 사용하여 정확도가 0.8 이상인 모델을 개발

방향성

범주를 크게 넘어서는 심각한 이상치는 없으나, 이상치가 조금 있는 관계로 시각화하여 판단.
VIF, PCA 등 여러 기법을 사용하기.

DATA PREPROCESSING

- 데이터 확인

	Call Failure	Complains	Subscription Length	Charge Amount	Seconds of Use	Frequency of use	Frequency of SMS	Distinct Called Numbers	Age Group	Tariff Plan	Status	Age	Customer Value	Churn
0	8	0	38	0	4370	71	5	17	3	1	1	30	197.640	0
1	0	0	39	0	318	5	7	4	2	1	2	25	46.035	0
2	10	0	37	0	2453	60	359	24	3	1	1	30	1536.520	0
3	10	0	38	0	4198	66	1	35	1	1	1	15	240.020	0
4	3	0	38	0	2393	58	2	33	1	1	1	15	145.805	0

- 데이터 모양 확인

(3150, 14)

DATA PREPROCESSING

- 데이터 정보 확인

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3150 entries, 0 to 3149
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Call  Failure           3150 non-null   int64
 1   Complains               3150 non-null   int64
 2   Subscription  Length    3150 non-null   int64
 3   Charge  Amount          3150 non-null   int64
 4   Seconds of Use          3150 non-null   int64
 5   Frequency of use        3150 non-null   int64
 6   Frequency of SMS        3150 non-null   int64
 7   Distinct Called Numbers 3150 non-null   int64
 8   Age Group               3150 non-null   int64
 9   Tariff Plan             3150 non-null   int64
 10  Status                  3150 non-null   int64
 11  Age                     3150 non-null   int64
 12  Customer Value          3150 non-null   float64
 13  Churn                   3150 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 344.7 KB
```

- 띄어쓰기 2번 된 columns 1번으로 수정
- 결측치 없는지 최종 확인

```
Call Failure              0
Complains                 0
Subscription Length       0
Charge Amount             0
Seconds of Use            0
Frequency of use          0
Frequency of SMS          0
Distinct Called Numbers   0
Age Group                 0
Tariff Plan               0
Status                    0
Age                       0
Customer Value            0
Churn                     0
dtype: int64
```

DATA PREPROCESSING

- 데이터 통계량 확인 - Charge Amount 확인 필수 (max 10 불가)

	Call Failure	Complains	Subscription Length	Charge Amount	Seconds of Use	Frequency of use	Frequency of SMS	Distinct Called Numbers	Age Group	Tariff Plan	Status	Age	Customer Value	Churn
count	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000
mean	7.627937	0.076508	32.541905	0.942857	4472.459683	69.460635	73.174921	23.509841	2.826032	1.077778	1.248254	30.998413	470.972916	0.157143
std	7.263886	0.265851	8.573482	1.521072	4197.908687	57.413308	112.237560	17.217337	0.892555	0.267864	0.432069	8.831095	517.015433	0.363993
min	0.000000	0.000000	3.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	15.000000	0.000000	0.000000
25%	1.000000	0.000000	30.000000	0.000000	1391.250000	27.000000	6.000000	10.000000	2.000000	1.000000	1.000000	25.000000	113.801250	0.000000
50%	6.000000	0.000000	35.000000	0.000000	2990.000000	54.000000	21.000000	21.000000	3.000000	1.000000	1.000000	30.000000	228.480000	0.000000
75%	12.000000	0.000000	38.000000	1.000000	6478.250000	95.000000	87.000000	34.000000	3.000000	1.000000	1.000000	30.000000	788.388750	0.000000
max	36.000000	1.000000	47.000000	10.000000	17090.000000	255.000000	522.000000	97.000000	5.000000	2.000000	2.000000	55.000000	2165.280000	1.000000

- Charge Amount 이상치 제거

```
Index : 473, Value: 10.0
Index : 2173, Value: 10.0
Index : 2273, Value: 10.0
Index : 2373, Value: 10.0
Index : 2673, Value: 10.0
Index : 2923, Value: 10.0
Index : 2973, Value: 10.0
```

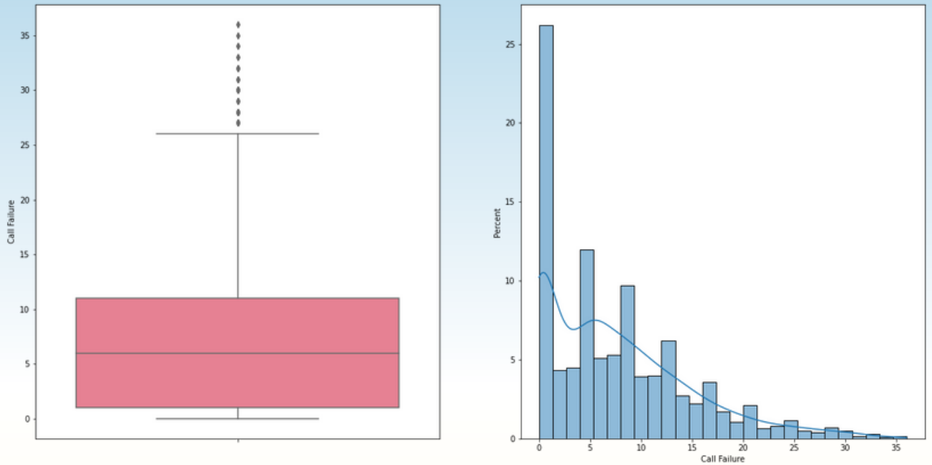
- 제거 확인

(3150, 14)

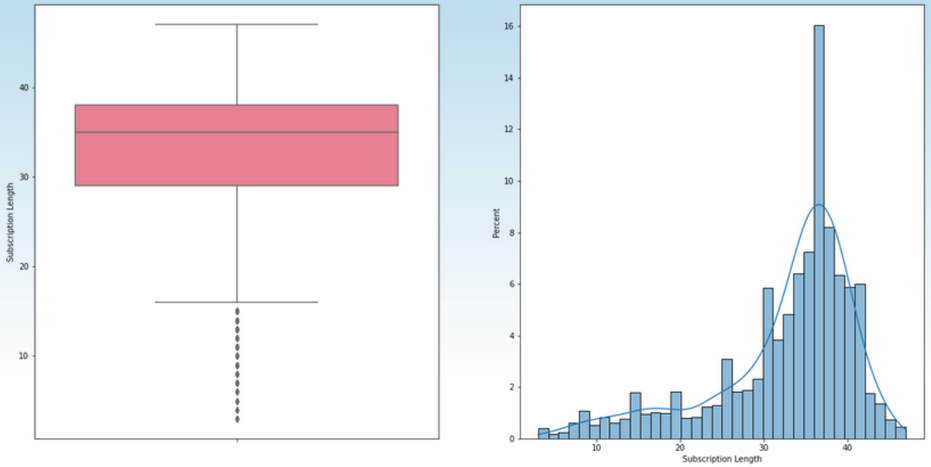
- 다시 연속적인 인덱스 생성

EDA

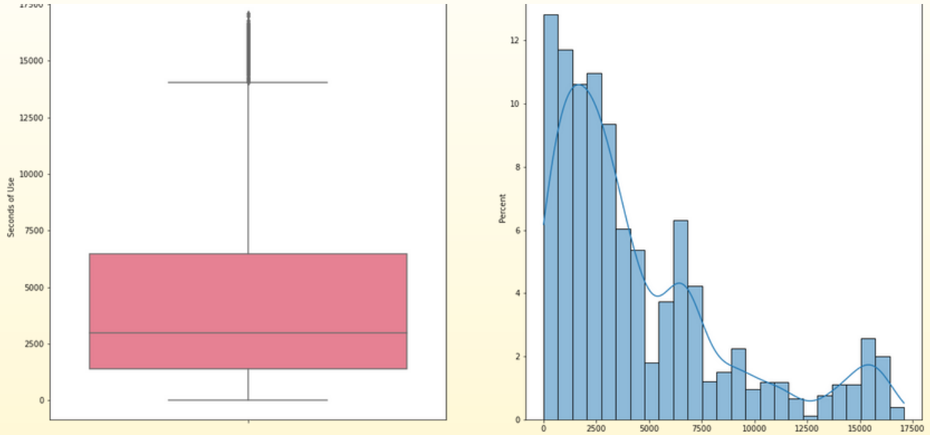
Call Failure



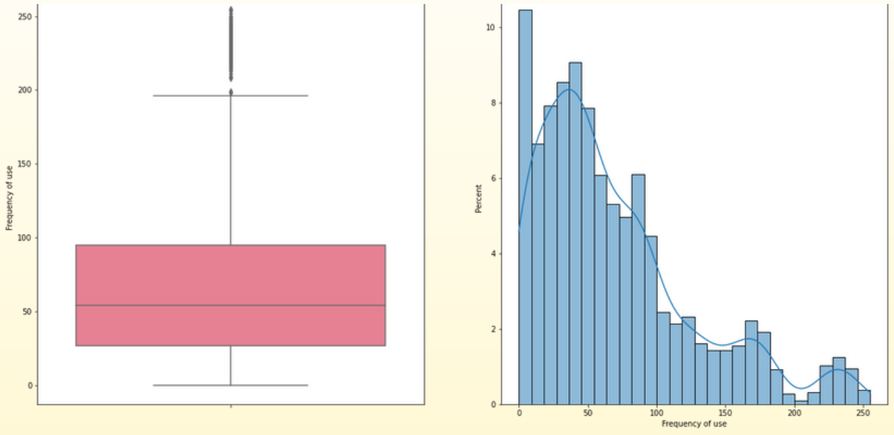
Subscription length



Seconds of Use

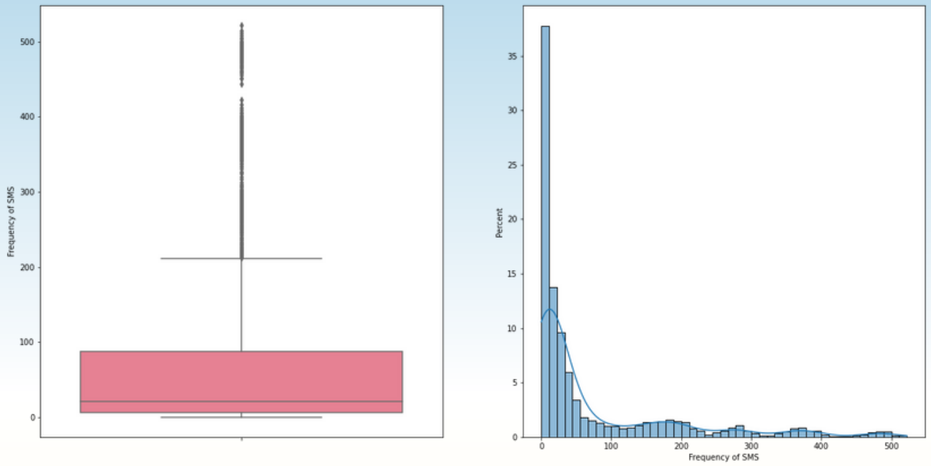


Frequency of use

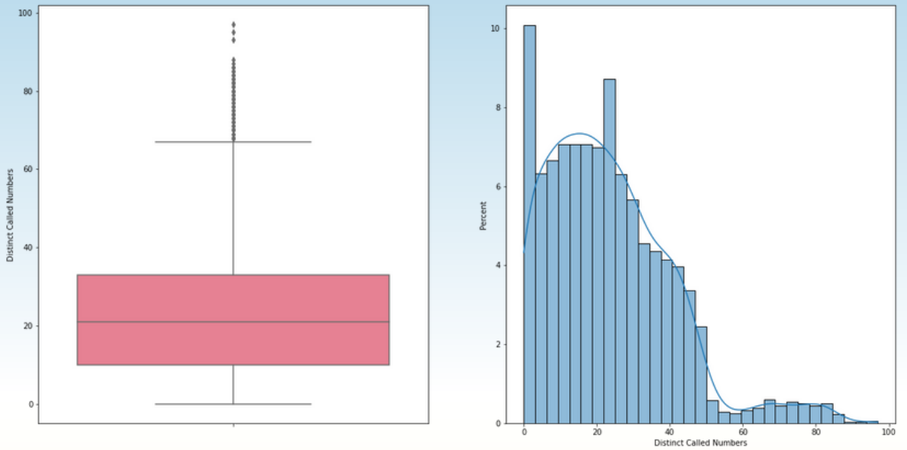


EDA

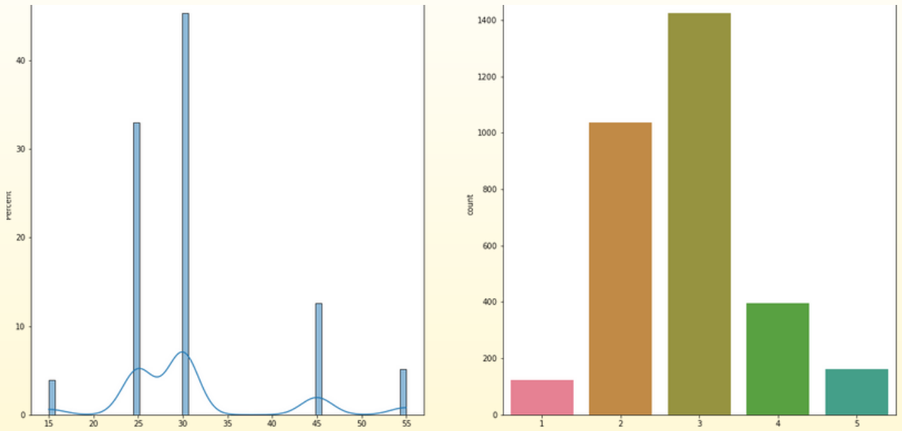
Frequency of SMS



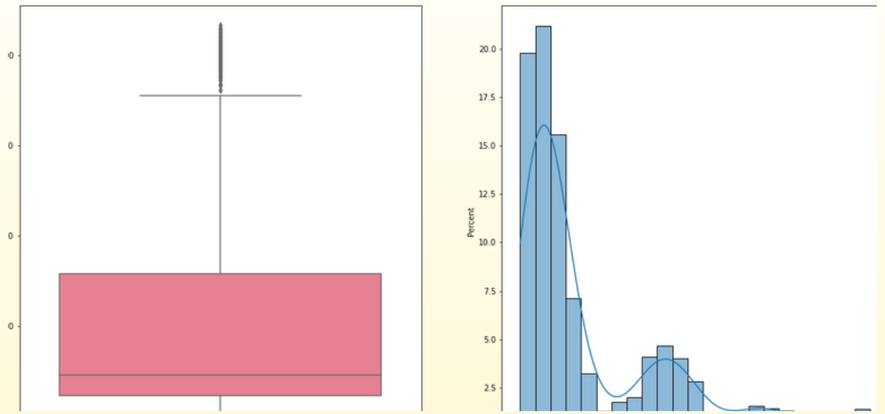
Distinct Called Numbers



Age / Age Group



Customer Value



- 이상치는 존재하나, 연속된 데이터 분포를 보이므로 이상치 제거X

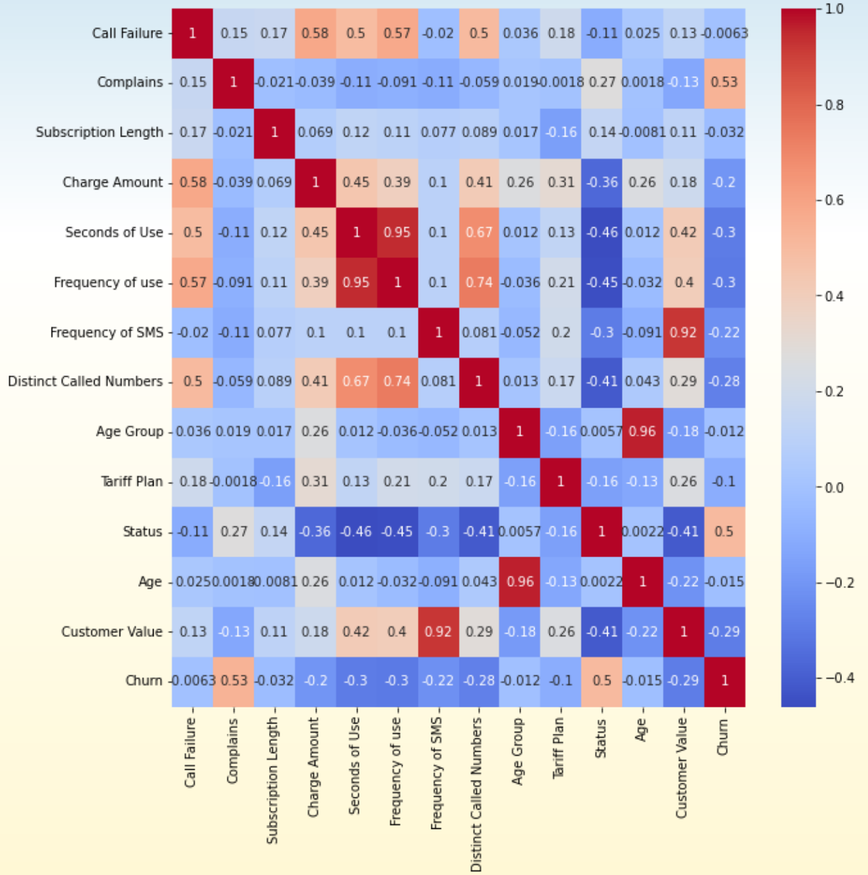
EDA

- VIF 지표를 기준으로 다중공선성 제거

VIF

	feature	VIF
0	Call Failure	6.050645
1	Complains	1.249507
2	Subscription Length	15.176198
3	Charge Amount	4.072552
4	Seconds of Use	48.678331
5	Frequency of use	46.627292
6	Frequency of SMS	58.654443
7	Distinct Called Numbers	6.971451
8	Age Group	155.605362
9	Tariff Plan	19.744892
10	Status	15.421378
11	Age	187.567685
12	Customer Value	95.967924

correlation



- 변수 선택 기준

1. VIF > 10 삭제
2. cluster 간의 공선성 확인
>> 공선성이 있는 변수들을 모두 삭제하지 않고 선별

- seconds of use , frequency of use, distinct of numbers

- frequency of SMS, customer value

- age group, age

MODEL SELECTION& BUILDING

MODEL1

X : Call Failure, Complains, Seconds of Use
(가설 1)

Log Likelihood: -564.8132114788167

MODEL2

X: 모든 feature

Log Likelihood: -482.7194320714108

MODEL3

X: 모든 feature들 PCA하여 사용

Log Likelihood: -528.955159362062

MODEL4

X: VIF < 10 이하인 features

Log Likelihood: -498.1581343953569

- **MODEL간의 성능을 LR Test Statistic과 p-value로 비교한 결과:**

model2 > model1 / model 2 > model3 / model2 > model4

MODEL SELECTION& BUILDING

MODEL 4 선택 이유

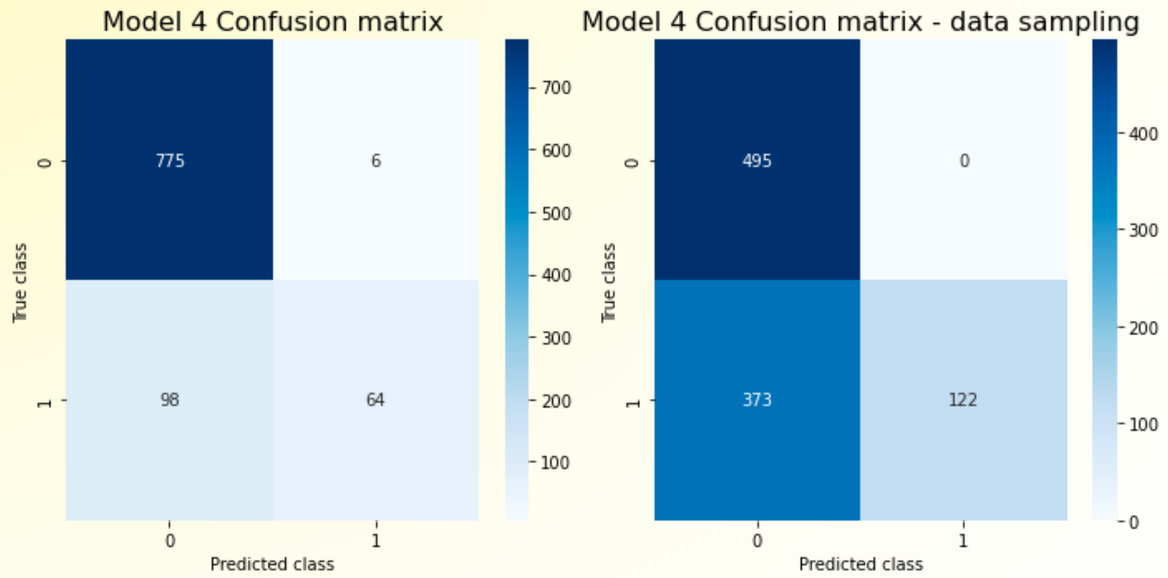
- 데이터셋의 **class label 개수 불균형** → logistic regression모델 간의 **성능 비교 지표**로 accuracy보다 l**og likelihood, LR Test Statistic**과 **p-value** 종합적으로 고려

- model 2의 성능이 model 4보다 좋아보이지만, 변수 전체 사용 보다는 **일부만 사용하여 비슷한 성능을 구현하는 모델**이 더욱 좋은 모델이라 판단

- PCA 모델을 사용하지 않는 것이 **변수 해석 가능성**에 있어 용이함

MODEL EVALUATION

MODEL4 - SAMPLING DATA 적용



Log Likelihood: -853.6103244769533
Logistic Regression sampling data score : 0.62323

MODEL 4 (default data) Precision, recall, f1 score : 0.91429, 0.39506, 0.55172
MODEL 4 (data sampling) Precision, recall, f1 score : 1.00000, 0.24646, 0.39546

기존 데이터는 0(유지)을 0(유지)으로 예측한 경우가 압도적으로 많았음
라벨의 비율을 맞춘 데이터는 정확도 감소, 1(이탈)을 0(유지)로 예측한 경우 증가

CONCLUSION

가설

1) Call failure, Complains, Seconds of use가 큰 영향 - **EDA 탐색 결과 그러함.**
2) Age, Age group간의 공선성이 있을 것 - **EDA 탐색 결과 그러함.**

결과

1) 정확도는 전체 feature을 사용한 Model2가 가장 높았음
2) 그러나 변수의 개수, 다중공선성, 해석가능성 등을 종합적으로 고려했을 때 model4를 최적의 모델으로 선정함
3) 선정한 모델을 더욱 개선시키기 위하여, class lable의 불균형 문제를 해결하고자 데이터 샘플링을 진행함
   이후 다시 성능을 확인해보니 정확도가 감소하고, 통신사 이탈을 통신사 유지로 예측하는 경우가 많았음

의미

1) 대표적으로 통화 실패 횟수, 서비스 불만 여부가 고객 이탈에 가장 큰 영향을 끼쳤음을 알 수 있음
2) 모델을 활용하여 같은 특징을 가진 고객의 이탈 여부를 예측 가능
2) 통화 서비스 개선, 이탈할 것으로 예측되는 고객 사전 관리 등 통신사의 고객 유치에 프로젝트 결과 활용 가능

향후 연구

1) 라벨의 비율이 1:1인 데이터로 EDA, VIF 확인, PCA 등의 모델링 과정을 다시 진행해 볼 것
2) 다른 통신사의 데이터에도 모델링 과정을 적용해보고 결과를 비교해 볼 수 있음

Thank's For Watching

TEAM2 이런이란
2022113156 이예령
2022113170 장정의
2022113185 전지현

Reference

https://seaborn.pydata.org/generated/seaborn.color_palette.html#seaborn.color_palette
https://stackoverflow.com/questions/48185090/how-to-get-the-log-likelihood-for-a-logistic-regression-model-in-sklearn
https://velog.io/@livelikesloth/P-Value
https://steadiness-193.tistory.com/198 https://www.kaggle.com/code/aymenkhouja/customer-churn-prediction/notebook