# SAMSUNG CARD DATA COMPETITION

## 잡아야 할 변수

팀: 황창현

Data Load

0~10개의 온라인 가맹점 카테고리

고객과 온라인 가맹점 이용(미이용) 데이터

고객 관련 변수 데이터

숫자형 범주형 구분 데이터

- xlsx 파일을 csv 파일로 수정.

Data Browse

진행 방향

## 온라인 가맹점 미이용 0 이용 1 치환

이유 : 가맹점 별이지만, 샘플 데이터라 양도 충분하지 않은 데이터.
카드사 입장에서 어떤 변수가 영향이 있을지 보고 싶을 지 생각 했을 때,
합쳐서 온라인 가맹점은 어떤 변수에 카드를 이용하나 라는 생각으로
단순 미이용 이용으로 치환. 후에 이진분류를 위함.

Data Cleaning ( = Data Preprocess)

데이터 불균형 : 미이용 80% 이용 20%

데이터 타입 : 숫자형, 범주형

진행방향

**모델에서 데이터 불균형 해소.**
**데이터 타입 고려.**

```
# 미이용 대다수 그 뒤로 오픈마켓 홈쇼핑 등.
train.MRC_ID_DI.value_counts(normalize=True)
```

```
# 숫자형 범주형 나누기
# dtype.dType.unique()
categorical_feature=list(dtype[dtype.dType=='categorical'].Variable_Name.values)
```

```
Train=pd.merge(train,cst,on=['cst_id_di'])
Train['MRC_ID_DI']=Train['MRC_ID_DI'].replace({1:1,2:1,3:1,4:1,5:1,6:1,7:1,8:1,9:1,10:1})
```

```
notFeature=['cst_id_di','MRC_ID_DI']
feature=[col for col in Train.columns if col not in notFeature]

X=Train[feature]
y=Train['MRC_ID_DI']
```

Data EDA

목적

## 온라인 가맹점 방문 고객 예측

변수 관련 전처리 부족 이유

**변수 추측과 마케팅 제안에 시간을 더 쏟았다.**

모델

**lightgbm (2.3.1)**

선정이유

**속도 빠르고, 꽤 정확하면서 아까 나눈 범주형 변수와 데이터 불균형까지 커버하는 제일 먼저 생각 난 모델.**

Baseline Model

**이진 분류 기법** (lightgbm binary classification)사용.

데이터는 80%는 train 데이터 20%는 test 데이터로 사용.

lightgbm 단점은 속도가 빠른대신 **과적합의 위험**이 있다.

단점 메꾸기 위해 k겹의 **교차검증** 한다.

학습 목적이 이진분류니까 오답에 패널티를 더 주는 binary logloss 손실함수 사용.

auc는 나중에 0과 1로 나뉘어야 하기 때문에 소수점있는 auc는 의미없다.

# Predict

교차검증 중 제일 좋았던 첫번째 모델을 최종 모델로 하고,

test 예측 및 전체 예측.

```
with open('final_model.sav','wb') as file:
    pickle.dump(modesl[0],file)
```

```
data=pd.concat([pd.DataFrame(model.predict(X_test,raw_score=True)),y_true.reset_index(drop=True)],axis=1)
data['yPred']=np.where(data[0]<-1.2,0,1)
data.columns=['판별함수','yTrue','yPred']
data.index=y_true.index
data=pd.concat([train.iloc[y_true.index]['cst_id_di'],data],axis=1)
```

```
fpr,tpr,threshold=sklearn.metrics.roc_curve(y_true,data['yPred'])
auc=sklearn.metrics.auc(fpr,tpr)
plt.plot(fpr,tpr)
plt.plot([0,1],[0,1],'k--')
plt.title("auc score : "+str(auc))
plt.show()
```

Marketing - Conclusion

# 중요한 변수는
# VAR003 VAR167 VAR 138 VAR103 VAR031 VAR017 VAR042 VAR036 VAR008 VAR010 VAR192 VAR062 .

만약. 새로운 카드 이벤트를 한다면

VAR003 (연관 변수)
사용량을 유도하는 이벤트를 하거나

VAR003 과 VAR167 의 (연관)소비를 일정량 이상 충족하는 고객을 위해 새로운 카드를 내는 것이다.

Marketing - analysis

변수 추측

지금 까지 내가 봐온 분석에 쓰인 카드데이터는
업종명, 이용건수 등 정수 및 문자형이었는데

숫자형(소수), 범주형(0,1)이었다.

일단 숫자는 특정 변수(VAR)의 증감율을 나타낸게 아닐까 생각했다.

예로, 전월(전년)대비 음식점 이용(결제) 증감 비율 같은.

그래서 숫자형의 양수는 1 음수는 0 으로
변수의 의미를 바꿨다.

그러면 변수의 의미는
예로, VAR003의 값이 1이면

**VAR003은 전월(전년)대비 이용(결제)이 증가한 변수**

가 되는것이다.

이렇게 접근했다. 그리고 apriori 알고리즘으로 분석해봤다.

현재 컴퓨터로 큰 데이터를 다루기가 불가능했다.
일부만 추출했다.

Marketing - interprete

itemsets 의 값들은 지지도(support) 0.9 이상의 값들을 가진다는 뜻이다.

VAR003 같은 경우는 0.99의 정말 온라인 가맹점을 이용하는 고객이든 아니든 1 의 값을 가졌다는 의미다.

내가 해석한 변수에 따르면 전월대비 증가한 변수라는 뜻이다.

그 외 VAR103, VAR167, VAR003, VAR138 이 한번에 묶여있는 것의 의미는 이 넷의 지지도 즉, 전체 데이터 중에서 저 넷이 같이 있을 비중이 크다는 것이다. VAR003이 증가 할 때 VAR167,VAR138,VAR103이 같이 증가했다는 뜻이다.

antecedents는 먼저 라는 뜻, consequents는 결과라는 뜻이다. 향상도 (lift) 를 보면 먼저 VAR103,VAR167,VAR138이 나오면 VAR104도 있을 경우가 높다는 뜻이다.
향상도는 지지도가 낮을 수록 높고, 신뢰도가 높을수록 높은데
마케팅 입장에서 **지지도가 낮으면 의미가 없기 때문에 지지도가 높으면서 신뢰도가 더 높아야 비즈니스가 된다.**

Marketing - Limit

한계점

방금 코드는 변수 중요도가 높은 상위 50개의 데이터. 그 중에서도 10개만 가지고 한것이다.

쉽게말해 원-핫 인코딩과 같기 때문에
**분석하는데 시간**이 오래걸린다.
**분석을 못한다.**

그래서 고객을 이용하는 고객 ,이용하지않는 고객으로 나눴지만, 변수가 많아 규칙을 찾기가 어려웠다.

**변수의 의미**를 몰라 가정하는 수 밖에 없었다.

0과 1로만 판단하는 apriori 알고리즘에서

0과 1 판단을 단지 양수란 이유로 했기 때문에
0.0001 도 1이므로 해석이 **과대해석**이 될 수 있다.

conviction 은 찾은 규칙이 얼마나 잘 못 됐을 까 하는 건데 conviction이 대부분 높다.