SCDC 2020

Track1

박민성 – 건국대학교 산업공학과
김동언 – 명지대학교 수학과/경제학과
정원렬 – 서울과학기술대학교 ITM 전공

1. 데이터 모델링

✓ Model 구축 및 현황

Logistic Regression

로지스틱 분석으로 범주형 결과의
자료에 대해서 Logit 값을 비교해서 분석 진행

RandomForest

분류한 데이터들의 일반화성능을 향상시키기
위하여 학습, 훈련

PCA

eigen vector 를 통하여 주성분 분석으로
차원을 축소시키는 방법

K-mean clustering

데이터 간의 거리를 계산하여 군집화 하는
방법

2. 현재 진행 상황

진행상황

- Samp_cst_feat.csv 과 samp_train.csv의 자료를 토대로 변수 중요도 분석
- 변수 상관 관계 분석 , 모델링 완료

분석상황

- Logistic Regression을 통한 MRC_ID_DI 가 0 or else 모델 구축
- PCA 와 K-means 를 통한 군집화
- RandomForest 를 통한 MRC_ID_DI 예측 모델링

2. 현재 진행 상황

✓ Model 선택 과정 및 데이터 분석 과정

Logistic Regression

➢ 로지스틱 분석으로 범주형 결과의 자료에 대해서 Logit 값을 비교해서 분석 진행

```
lr_clf = LogisticRegression(max_iter = 10000)
lr_clf.fit(x_train, y_train)

pred_lr = lr_clf.predict(x_test)
print("accuracy = {}".format(accuracy_score(y_test, pred_lr)))
print("MSE = {}".format(mean_squared_error(y_test, pred_lr)))
```

```
accuracy = 0.8610928242264648
MSE = 0.13890717577353523
```

Logit Regression Results

Dep. Variable:	MRC_ID_DI	**No. Observations:**	7086
Model:	Logit	**Df Residuals:**	6861
Method:	MLE	**Df Model:**	224
Date:	Mon, 21 Sep 2020	**Pseudo R-squ.:**	0.4009
Time:	20:52:57	**Log-Likelihood:**	-2066.6
converged:	False	**LL-Null:**	-3449.2
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
VAR002	0.9805	0.863	1.136	0.256	-0.711	2.672
VAR003	-1.3099	0.255	-5.133	0.000	-1.810	-0.810
VAR004	-6.8278	4.186	-1.631	0.103	-15.032	1.377
VAR005	-0.5489	0.793	-0.692	0.489	-2.103	1.005
VAR006	-6.0795	4.581	-1.327	0.184	-15.059	2.899

➢ MRC_ID_DI == 0 (온라인쇼핑몰 미사용) 을 0 나머지를 1로 설정 하고 모델링 한 결과

Accuracy 가 약 86% 로 상당히 높은 수치로 확인할 수 있었습니다. 이를 통해 쇼핑몰 사용 고객을 충분히 예측 할 수 있었습니다.

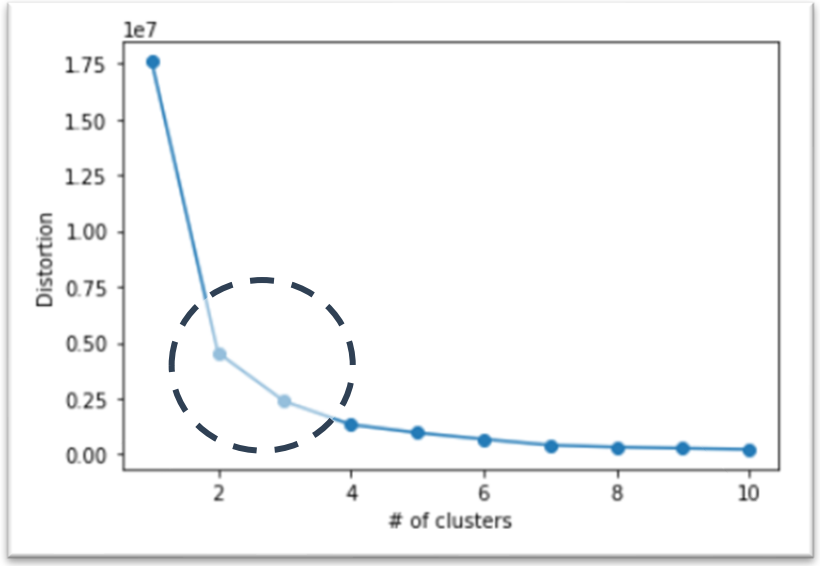
이를 통해 차후 데이터를 받는다면 Clustering에 대한 Insight를 추출 할 수 있을 것으로 예상

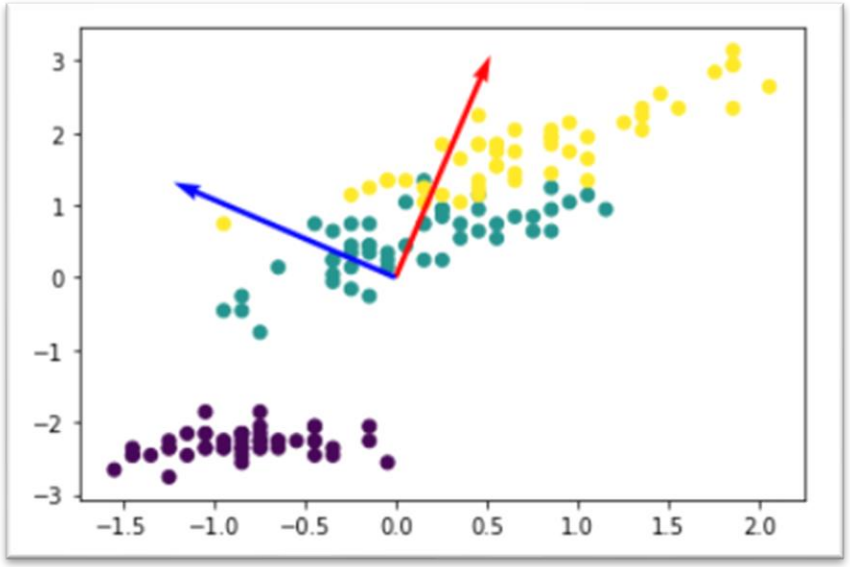
2. 현재 진행 상황

PCA

- 차원 축소를 통해 변수 225 개에서 N 개로 추출 하는 방식입니다. 우측에 있는 그래프를 보시면 더욱 자세히 해석 할 수 있습니다. 우측의 시각화 된 데이터는 PCA 를 통해 2차원 데이터로 차원 축소를 진행 하였습니다. 화살표는 데이터의 기저를 의미합니다.

  저희는 데이터 자체로 의미있는 군집단이 될 수 있을것이라고 가정하였으며 Kmeans를 통해 기울기가 완만한 구간을 최적의 군집수로 지정하여 군집화를 진행하였습니다.





- 저희는 데이터를 clustering 할 때 몇 개의 군집단을 사용 하는 것이 올바른 지에 대해 분석했습니다. 그 결과 Distortion이 완만히 낮아지는 단계인 2~4개의 군집을 설정한다면 데이터를 의미있게 구분 할 수 있을 것이라고 생각했습니다.

  위의 시각화 자료에서 볼 수 있듯 차이가 군집 별 차이가 의미있게 구분 되는 것으로 보아, 데이터 수령 후 더욱 깊은 분석을 할 수 있을 것입니다.

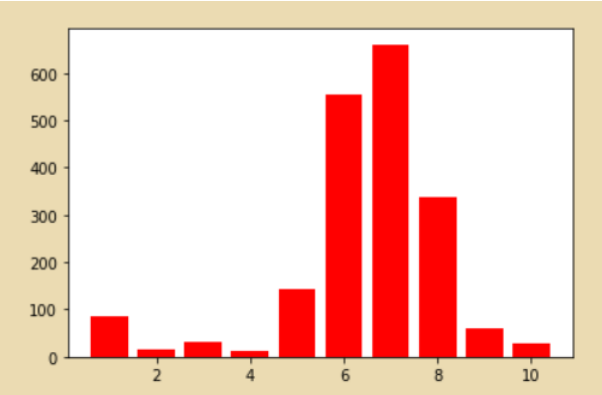
2. 현재 진행 상황

Random Forest

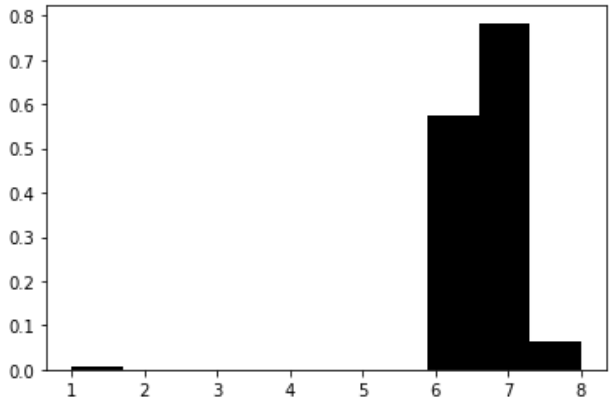
- 해당 Raw 데이터를 모두 input variable 로 ,MRC_ID_DI 를 target variable 로 Random Forest 를 돌린 결과 정확도가 0.87로 상당히 높은 수치가 나왔습니다. 하지만 이는 MRC_ID_DI 가 0 인 값이 상당히 많은 것으로 보아 의미 있는 결과가 아니였습니다.

  따라서 저희는 MRC_ID_DI 가 0 인 값을 제외한 클래스 분류를 용이하게 할 수 있을 것이라는 가정을 세우고 진행하였습니다.

  MRC_ID_DI 가 0인 값을 제외하여 RandomForest Model을 적용 해 본 결과 0.413이라는 결과를 추출 가능 하였습니다.



<기존 데이터의 MRC_ID_ID 빈도>



정확도 : 0.4134948096885813

- 하지만 histogram 으로 결과의 분포도를 분석 해 본 결과, MRC_ID_DI 가 6, 7인 값이 많은 것을 확인 할 수 있었습니다.

  이는 데이터 량이 균일하지 못하기 때문에 나오는 결과라고 예측 가능 하였습니다

  만약 데이터를 조금 더 받거나, 변수의 중요도, 변수의 상관관계를 예측 할 수 있다면 더 좋은 성능의 모델을 만들 수 있을 것으로 예상 합니다.

3. 계획 및 데이터 분석

- 공모전 주최 회사 (삼성카드) 에서 주는 데이터를 수령합니다.

- 수령한 데이터를 바탕으로 최 중요 변수 , 삭제 변수를 구합니다. 또한 변수들 간의 선형성을 분석하여 의미 있는 관계를 파악합니다.

- 데이터 분석을 완료 한 후 데이터 전처리 및 모델링을 완료합니다. 뿐만 아니라 모델링 결과와 데이터 상관관계를 활용하여 새로운 마케팅 방안을 생성합니다.

3. 계획 및 데이터 분석

✓ 중요변수 선택

방법 1.

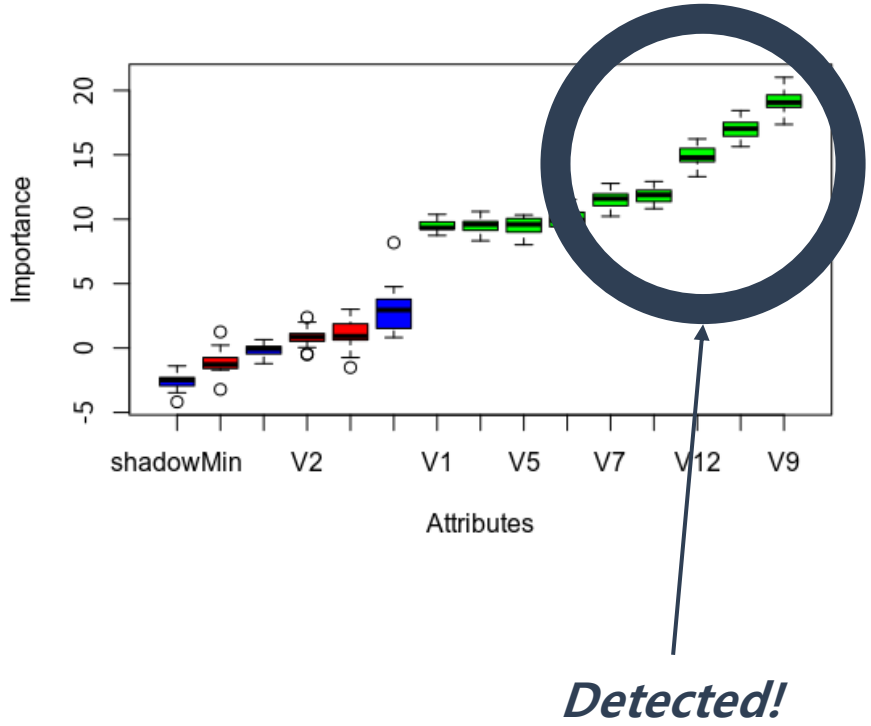
➢ BORUTA 알고리즘을 통해 변수 중요도 파악 및 무의미 한 변수 제거

```
Iteration:      85 / 100
Confirmed:      15
Tentative:      5
Rejected:       206
Iteration:      86 / 100
Confirmed:      15
Tentative:      5
Rejected:       206
Iteration:      87 / 100
Confirmed:      15
Tentative:      5
Rejected:       206
```

✓ 중요 변수 확인

✓ 잠정적 삭제 변수

✓ 삭제변수

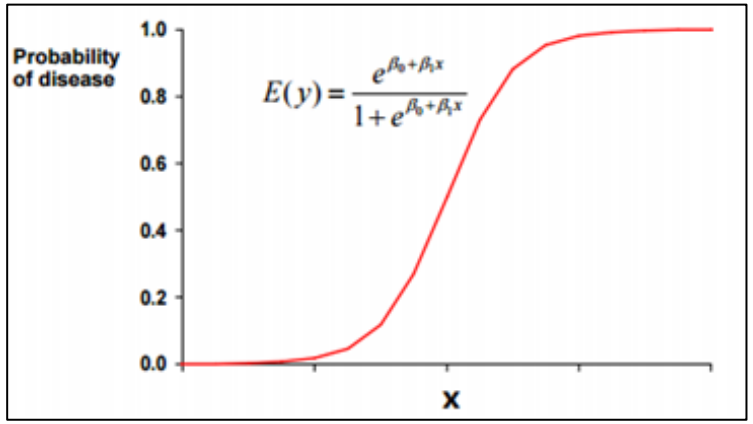


➢ BORUTA 알고리즘의 결과와 변수 상관관계를 복합적으로 판단하여 변수를 선택 및 삭제 할 예정입니다.

3. 계획 및 데이터 분석

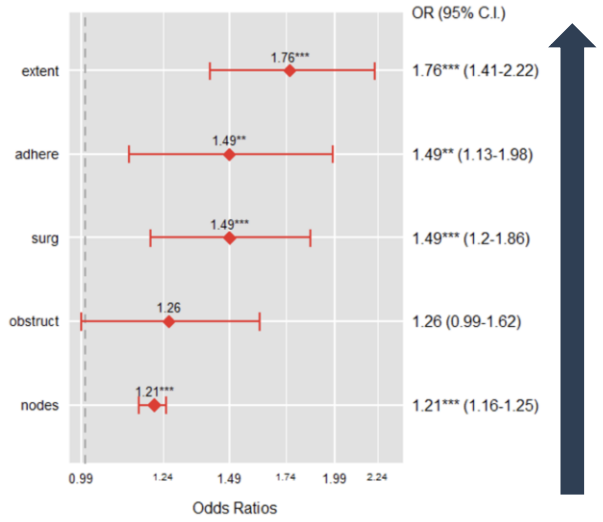
방법 2.

- Logistic 을 통한 Coefficient 추출 - 전진선택법



- 전제 1. – Logistic 모델 또한 선형회귀 방정식 변형이다.
- 전제 2. – 선형회귀 변수 분석 또한 Logistic 모델을 구하는 것에 있어 큰 영향을 줄 것이다.

- 모든 변수를 바탕으로 Logistic 모델을 만든다.
- Odds Ratio 를 계산하여 해당 변수가 결과에 영향을 미칠 확률을 계산한다.
- 영향력이 높은 Odds Ratio 를 구하여 영향력이 높다고 판단되는 변수들을 통해 새로운 Logistic 모델을 생성한다.
- 2~3번 과정을 반복하여 모델을 정립한다.



중요도 상승

4. 솔루션을 통한 마케팅 방안

방안 1.

- 모델링을 통해 얻을 수 있었던 Coefficient 가 가장 높은 변수를 파악, 이를 통해서 카드 이용에 영향이 큰 변수들을 분석해서 이를 통한 고객 마케팅 방안을 제시하려고 합니다.

Ex) Regression 파악 후 Odds Ratio 분석

방안 2.

- 이탈자와 비이탈자를 clustering 한 후, 해당 클래스 데이터를 분석하여 비이탈자 고객을 예측해서 고객을 집중적으로 관리할 수 있는 솔루션을 제시하려고 합니다.