DACA

NGO 데이터 분석 : 3주차

2020.11.05 ~ 2020.11.11

빅데이터 응용학과 석사 1기 양선욱

목차

1. 요인분석 – 1.탐색적 요인분석

사용 데이터

PLED_NUM : 전체 플릿지 수

PLED_NUM_FULFILLED : 납부완료 플릿지 수

CHILD_NUM_COUNTRYCODE : 아동 국가 수

PAY_NUM : 납입 후원 횟수

PAY_SUM_PAYMENTAMOUNT : 납입총후원금액

PAY_NUM_NOPAY : 미납횟수

INTR_NUM_REQUEST : 인터랙션 요청

INTR_NUM_COMM : 인터랙션 횟수

분석 목적

- 후원자의 후원 상황을 설명할 수 있는 요인을 찾는다.
- 뜻 밖의 변수로 만들어지는 요인을 탐색한다.

분석 과정

1) 변수투입
2) 고유값 검사
3) 요인 수 지정
4) 요인적재량 검사
5) 불필요 변수 제거, 새로운 변수 투입
6) 2)로 가서 반복

핵심 코드

```
X = df[['PLED_NUM','PLED_NUM_FULFILLED','CHILD_NUM_COUNTRYCODE','PAY_NUM','PAY_SUM_PAYMENTAMOUNT',
'PAY_NUM_NOPAY','INTR_NUM_REQUEST','INTR_NUM_COMM']]

df = df.dropna()
#3. 탐색적요인분석
fa = FactorAnalyzer(method='principal',n_factors=3, rotation='varimax').fit(X)
print('fa : ',fa)
#4. 결과 출력
print('요인적재량 :\n', pd.DataFrame(fa.loadings_, index=X.columns))
print('\n공통성 :\n', pd.DataFrame(fa.get_communalities(), index=X.columns))
ev, v=fa.get_eigenvalues()
print('\n고유값 :\n', pd.DataFrame(ev))
print('\n요인점수 :\n', fa.transform(X.dropna()))
```

1. 요인분석 – 1.탐색적 요인분석

결과

	고유값
0	3.88
1	1.66
2	1.08
3	0.56

	0요인	1요인	2요인
전체플릿지수	0.938	0.019	0.185
납부완료 플릿지수	0.932	0.034	0.158
아동국가수	0.849	0.023	0.192
납입후원횟수	0.805	0.186	-0.368
납입총후원금액	0.765	0.267	-0.307
미납횟수	0.053	0.111	0.892
인터랙션 요청	0.078	0.921	-0.033
인터랙션 횟수	0.122	0.731	0.527

	공통성
전체플릿지수	0.914
납부완료 플릿지수	0.896
아동국가수	0.759
납입후원횟수	0.819
납입총후원금액	0.751
미납횟수	0.812
인터랙션 요청	0.855
인터랙션 횟수	0.827

결과 해석

- 고유값이 1이상인 변수가 3개이므로 요인의 수를 3개로 지정하였다.
- 공통값은 모두 0.5이상으로 충분하다.
- 3개의 요인에 각 변수들이 배정되었다.

실무적 인사이트

- 0요인은 모두 후원자의 후원 상황을 설명하는 변수들이다.
  -> 0요인을 '후원현황'라고 부르자.
- 1요인은 모두 인터랙션에 관한 변수들이다.
  -> 1요인을 '후원자인터랙션현황'이라고 부르자.
- 2요인은 변수가 하나뿐이지만 설명력이 충분하다. 미납에 관련된 변수를 더 찾아보자.

2. 분류예측분석 – 1.선형판별분석

사용 데이터

액티브비율= PLED_ACTIVE_NUM/PLED_NUM

PAY_RATE_NOPAY : 미납율

CHURN : 이탈여부

분석 목적

- 고객의 이탈여부를 예측하는 모델을 만든다.
- 이탈여부에 영향을 주는 변수들을 찾는다.

분석 과정

1) 변수투입
2) 독립변수 정규성 검사
3) 독립x독립 상관분석
4) 독립x종속 상관분석
5) 불필요 변수 제거, 새로운 변수 투입
6) 2)로 가서 반복
7) 선형판별분석
8) 사후 분석

분석 결과

*** 독립변수 정규성 검사**

독립변수	F값	p-value
액티브비율	0.61	0.000
미납율	0.47	0.000

*** 선형계수**

	선형계수
액티브비율	-31.869
미납율	4.033
절편	12.915

2. 분류예측분석 – 1.선형판별분석

결과

* 독립변수x종속변수 상관분석

	액티브비율	미납율	이탈여부
액티브비율	1.0		
미납율	-0.3	1.0	
이탈여부	-0.9	0.32	1.0

* 예측정확도 : 96.7%

* 분류 행렬표

	유지예측	이탈예측	총합
실제유지	755	32	787
실제이탈	0	190	190
총합	755	222	977

결과 해석

- 예측정확도는 96.7%로 준수하다.
- 미납율 변수 때문에 예측정확도가 내려갔다고 보여진다.
- 더미 데이터까지 써가면서 모든 데이터를 독립변수로 시도해 보았지만 최종으로 고른 두 가지 독립 변수 만큼의 결과가 나오지 않았다.
- 유지는 잘 예측했으나, 이탈 예측에서 오류가 나왔다.

실무적 인사이트

- 이탈예측의 정확도는 85.5%이다.
- 하지만 이탈 고객을 한 명도 놓치지 않았다.
- 고객의 이탈에 가장 큰 영향을 주는 것은 액티브비율이 줄어 드는 것이다.
- 하지만 이것은 이탈에 따른 결과일 수 있다.
- 이탈의 추세를 확인할 시계열 데이터가 필요하다.

2. 분류예측분석 – 2. 로지스틱

사용 데이터

- AGE : 나이
- 고객등급
- CHURN : 이탈여부

분석 목적

고객등급	우대	골드	실버	브론즈	새싹
이탈율	17%	14%	13%	20%	58%

- 새싹등급인가, 아닌가는 이탈여부에 영향을 줄 것이다.
- 이를 기반으로 고객이탈을 예측하는 모델을 만든다.

분석 과정

1) 변수투입
2) 로지스틱 분석
3) 불필요 변수 제거, 새로운 변수 투입
4) 2)로 가서 반복
5) 사후 분석

변수	coef	p-value	odds
절편	-1.018	0.000	0.361
AGE	-0.015	0.010	0.985
새싹등급	2.046	0.000	7.740

분석 결과

Pseudo R-squared: 0.088

정확도 : 81.7%

	유지예측	이탈예측	총합
실제유지	705	31	736
실제이탈	137	47	184
총합	842	78	920

결과해석 & 실무적 인사이트

- 후원자의 나이가 많을 수록 이탈은 소폭 감소한다.
- 새싹등급은 타 등급에 비해 이탈확률이 7.7배 높다.
- 해당 모델도 R-squared값이 낮아 미숙하다.
- 유지라고 예측했으나 실제로 이탈한 고객은 19.4%에 달한다.
- 해당 모델로는 고객을 나이를 늘리라는 결론밖에 못 내린다.

3. 군집분석 – 1.계층적 군집분석

사용 데이터

- 액티브플릿지비율
- 월후원율
- PLED_RATE_FULFILLED : 납부 완료 플릿지 비율

분석 목적

- 후원자들의 후원 성과에 따른 그룹화를 하여 맞춤형 대응을 하려고 한다.

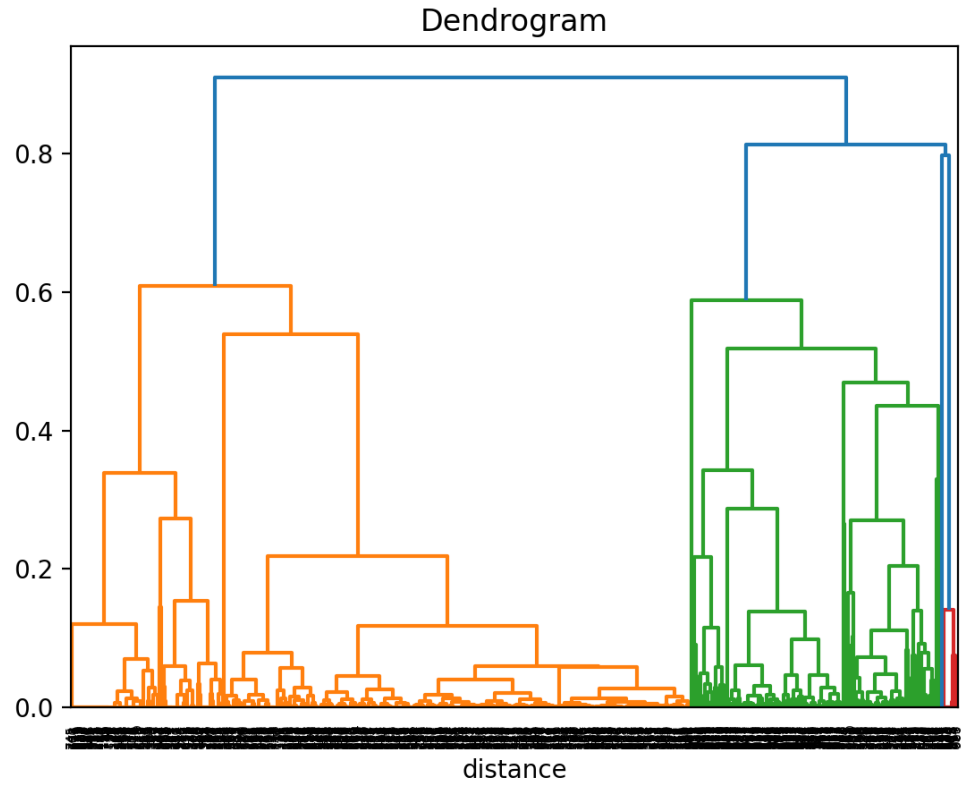
분석 과정

1) 변수투입
2) 계층적 군집분석
3) 불필요 변수 제거, 새로운 변수 투입
4) 2)로 가서 반복
5) 사후 분석

결과해석 & 실무적 인사이트

- 크게 두 가지의 그룹으로 나뉘어 진다.
- 무슨 요인으로 그룹이 나뉘어 졌는지 사후분석이 필요하다.
- 상대적으로 범위가 좁은 오른쪽 그룹에 선택과 집중 전략이 필요해 보인다.

분석 결과



3. 군집분석 – 2.비계층적 군집분석

사용 데이터

- 해외아동후원금
- 국내아동후원금
- 해외사업후원금
- 긴급사업후원금
- 국내사업후원금
- 북한사업후원금

분석 목적

- 6가지 후원 분류가 존재한다.
- 후원자들의 후원 방향이 어떻게 나뉘는지 알아본다.

분석 과정

1) 변수투입
2) k-means 군집분석
3) 군집 수 수정
4) 2)로 가서 반복
5) 사후 분석

핵심 코드

```
df2 = df1[['해외아동','국내아동','해외사업','긴급사업','국내사업','북한사업']]
#2. 비계층적 군집분석
df2 = df2.drop(415)
model = KMeans(n_clusters=6, max_iter=20, random_state=19).fit(df2)
```

결과

*지정된 군집수에 따른 군집의 분포

군집수	군집1	군집2	군집3	군집4	군집5	군집6
6	467	382	37	30	40	21
5	467	403	37	30	40	
4	470	440	37	30		
3	443	496	38			
2	502	475				

결과 해석

- 군집5, 군집6은 모두 군집2에 흡수된다.
- 군집수가 4개에서 3개로 줄어들면 큰 변화가 있다.
  - 군집수를 4개로 지정한다.

3. 군집분석 – 2.비계층적 군집분석

결과

	해외아동	국내아동	해외사업	긴급사업	국내사업	북한사업	인원수	비율
군집1	333,794	3,978	1,595	553	3,446	382	470명	48.4%
군집2	12,250	9,681	15,443	19,534	27,147	318	440명	45.3%
군집3	80,540	618,108	0	1,351	12,972	0	37명	3.8%
군집4	761,570	73,000	12,000	11,000	2,000	4,000	30명	3.0%

결과 해석

- 군집1은 극단적 해외아동 집중형이다. > 후원금 중 해외아동 비율 97%
- 군집2는 올라운드형이다. > 14%, 11.4%, 18.3% , 23%, 32%, 0.4%
- 군집3은 국내아동 집중형이다. > 국내아동 86%, 해외아동 11%
- 군집4는 우량 해외아동 집중형이다. > 해외아동 88%. 군집1보다 후원금이 크다.

실무적 인사이트

- 왜 해외아동에 대한 후원이 많은지에 대한 분석이 필요하다.
- 현재 본사의 TV광고는 해외 기아 아동 사업에 초점을 맞추고 있다.
- 해외아동 후원에 집중 중인 본사의 방향성에 따라 군집2에 속한 후원자들을 군집1, 군집4로 유도할 필요가 있다.
- 북한사업에 대한 재검토가 필요하다.

4. 비모수통계검정 – 1.적합도 검정

사용 데이터

-LONGEVITY_D : 가입 일수 - AGE : 나이 -이탈여부

분석 목적

-이전의 결과들로 부터 새싹등급의 이탈 비율이 가장 높다.

- 가입 일수에 따른 새싹등급의 이탈이 랜덤한지 알아본다.

- 또한 해당 고객들의 나이가 정규한지 알아본다.

분석 과정

1) 새싹등급, 가입기한 1년 이내 데이터 선정

2) 가입일수 순으로 이탈여부에 따른 RUN 검정 분석

3) 나이에 따른 Kolmogorov-Smirnov 검정 분석

4) 가입기한을 조정하여 재분석

가입일수		1년이내	1년이상 3년이내
RUN검정	통계값	-0.419	-2.470
	p-value	0.675	0.013
정규성검정	통계값	0.095	0.778
	p-value	0.230	0.002

결과 해석

- 가입기한이 1년 이내인 데이터는 이탈여부가 무작위이며 나이가 정규분포를 보인다.

- 가입기한이 1년 이상 3년 이내인 데이터는 이탈여부가 무작위가 아니며 나이 또한 정규분포 따르지 않았다.

실무적 인사이트

- 새싹등급이면서 가입기한이 1년 이상인 후원자의 이탈이 급속도로 많아진 이유에 대한 분석이 필요하다.

- 이들의 이탈은 외부적 요인이 적용되어 보인다.

- 후원자의 나이가 정규분포를 따르지 않으므로 나이로 분석을 할 때는 비모수 통계분석 기법을 활용해야 한다.

4. 비모수통계검정 – 2.동질성 검정

사용 데이터

-LONGEVITY_D : 가입 일수　　-이탈여부

분석 목적

- 골드등급과 실버등급의 비율은 35%, 30%로 비슷하다.
- 골드등급과 실버등급의 이탈율은 14%, 13%로 비슷하다.
- 두 집단의 연간 일탈자의 수가 같은지 알아본다.

분석 과정

1) 골드등급, 실버등급 가입기한에 따른 데이터 선정
2) Mann-Whitney U 검정

분석 결과

U통계량 : 24.5　　p-value : 0.22

분석 결과

가입기한	골드등급	실버등급
1년 이내	0	3
1년 이상 2년 이내	6	6
2년 이상 3년 이내	4	9
3년 이상 4년 이내	7	3
4년 이상 5년 이내	4	3
5년 이상 6년 이내	7	2
6년 이상 7년 이내	3	0
7년 이상	21	11

실무적 인사이트

- 골드, 실버등급의 연간 이탈자 수는 같다.
- 두 등급의 후원자에게 같은 이탈 방지 접근법이 사용될 수 있다.
- 추후 연간 이탈 비율로도 분석을 해본다.

4. 비모수통계검정 – 3.상관성 검정

사용 데이터

-고객등급 -이탈비율 -책임감(미납율 기반 점수)

분석 목적

- 이전의 분석으로 우대등급의 이탈비율이 골드, 실버등급 보다 높다는 것을 알았다.
- 우대등급의 이탈을 막기 위해 이탈비율과 상관성이 있는 데이터를 찾아야 한다.

분석 과정

1) 각 등급의 이탈비율 순위 지정
2) 비교 변수의 순위 지정
3) kendall 검정분석 수행
4) 변수를 바꾸어 재시도

분석 결과

상관계수: -0.99 p-value : 0.016

분석 결과

	이탈비율	책임감
우대	3	3
골드	4	2
실버	5	1
브론즈	2	4
새싹	1	5

결과분석 & 실무적 인사이트

- 고객등급의 이탈비율 서열과 책임감 서열은 연관성이 있다.
- 상관계수가 -0.9로 강력한 상관관계성을 보여준다.
- 우대등급의 책임감을 다른 등급보다 높도록 관리하면 이탈비율이 줄어들 것인가? -> 인과관계 모호
- 인과관계를 알기 위한 시계열 데이터가 필요하다.
- 다른 변수에 대한 추가적인 분석이 필요하다.