다변량 및 빅데이터 분석

20180835 정보통계학과 정지윤

신용카드 데이터를 이용한 고객 세분화

20180835 정지윤

1. 서론

충성도 높은 고객 식별, 고객 세분화, 타깃 마케팅 및 기타 마케팅 산업의 사용 사례에 사용할 수 있는 고객 신용카드 정보 데이터 세트이다. 고객의 신용카드 정보 데이터를 이용하여 마케팅 업계에 충성도가 높은 고객을 알아보고자 한다.

2. 고객의 신용카드 정보 데이터

2.1 데이터 출처 및 변수 설명

본 논문에 사용된 데이터는 Kaggle(https://www.kaggle.com/)에서 제공한 데이터로 고객의 신용카드 데이터를 나타낸 것이다. 분석에 사용될 변수는 다음의 표 2.1과 같다.

표2.1 변수명

변수	변수이름	변수설명
id	Sl_No	고객 일련 식별 번호
X1	Age_Credit_Limit	고객의 평균 신용카드 한도
X2	Total_Credit_Customer	고객이 소유한 총 신용카드
X3	Total_visits_bank	고객의 총 은행 방문 수
X4	Total_visits_online	고객의 총 온라인 은행 이용 수
X5	Total_calls_made	고객의 총 전화 은행 이용 수

2.2 기초통계량

본 분석인 다변량 분석에 들어가기 전, 변수의 기초통계량을 정리하여 변수의 특성에 대해 알아보았다. 자료의 관측개수는 각 변수별로 결측값이 없이 모두 660개로 동일하다.
변수 X1인 최솟값과 최댓값의 차이가 매우 크므로 고객의 평균 신용카드 한도의 차이가 심한 것을 알 수 있다. 변수 X3,X4,X5를 통해 고객 중 은행을 전혀 이용하지 않는 고객과 온라인 은행을 전혀 이용하지 않는 고객과 전화 은행을 전혀 이용하지 않는 고객이 있음을 확인할 수 있다. 또한 고객은 평균적으로 5개의 신용카드를 가지고 있음을 알 수 있다.
나무상자 그림에서 X1 변수의 값이 나머지 변수들의 값과 차이가 커서 제대로 된 확인이 불가능했기에 단위가 동일한 변수들은 따로 묶어서 나무상자그림을 그렸다. X1(신용카드 한도)와 X4(온라인 은행 이용 수)는 이상값을 가지고 있음을 볼 수 있다. 또 3가지 은행 이용 방법 중에 X5(전화 이용)가 평균이 가장 높고 데이터의 퍼짐 정도도 가장 큰 것을 확인할 수 있다.

표 2.2 연속형 변수들의 기초통계량

변수	관측개수	최솟값	중앙값	평균	최댓값	표준편차
X1	660	3000	18000	34574	200000	37625.49
X2	660	1.00	5.00	4.71	10.00	2.17
X3	660	0.00	2.00	2.40	5.00	1.63
X4	660	0.00	2.00	2.61	15.00	2.94
X5	660	0.00	3.00	3.58	10.00	2.87

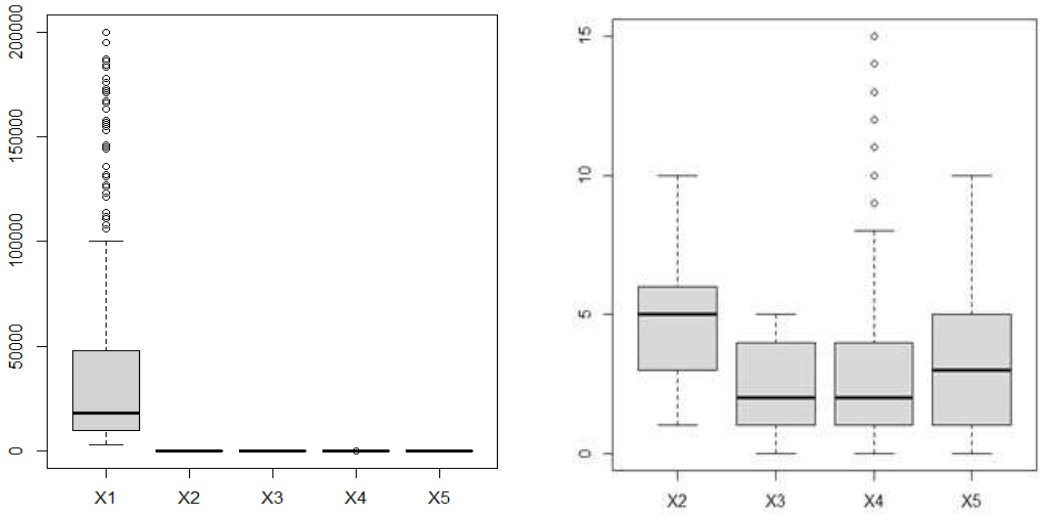

그림2.1 연속형 변수들에 대한 boxplot

3. 다변량 통계분석

3.1 상관분석

상관분석(correlation analysis)이란 두 변수 간에 선형관계를 파악하기 위한 통계적 기법으로 상관계수는 변수 간 관계의 정도 혹은 방향에 대한 척도이다. 이때 두 변수 간의 관계의 강도를 상관관계(Correlation, Correlation coefficient)라한다. 상관계수는 값이 -1에 가까울수록 두 변수가 강한 음의 상관관계를 가지며, 1에 가까울수록 두 변수가 강한 양의 상관관계를 갖는다. 또한 상관계수가 0에 가까우면 두 변수 사이에는 선형의 상관관계가 없다는 것을 의미한다. X1과 X2의 상관계수는 0.61로 가장 높은 양의 상관관계를 보이며 X1과 X4의 상관계수는 0.55로 두번째로 높은 양의 상관관계를 보인다. X1과 X3의 상관관계가 -0.10으로 가장 약한 것을 알 수 있다. 그림3.1은 변수들 간의 상관관계를 보다 보기 편하게 하기 위하여 시각화한 것이다.

표 3.1 변수간 상관행렬

변수	X1	X2	X3	X4	X5
X1	1.00	0.61	-0.10	0.55	-0.41
X2	0.61	1.00	0.32	0.17	-0.65
X3	-0.10	0.32	1.00	-0.55	-0.51
X4	0.55	0.17	-0.55	1.00	0.13
X5	-0.41	-0.65	-0.51	0.13	1.00

그림3.1 상관행렬의 시각화

3.2 주성분 분석

주성분 분석은 여러 개의 반응변수로 얻어진 다변량 데이터에 대해, 분산-공분산 구조를 변수들의 선형결합식으로 설명하고자 하는 접근 방법이다. 주성분 분석의 목적은 크게 차원 축소, 변동이 큰 축 탐색, 주성분을 통한 데이터의 해석이라고 할 수 있다.

표 3.2 주성분 계수

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.51	1.37	0.566	0.527	0.494
Proportion of Variance	0.46	0.37	0.064	0.056	0.049
Cumulative Proportion	0.46	0.83	0.896	0.951	1.000

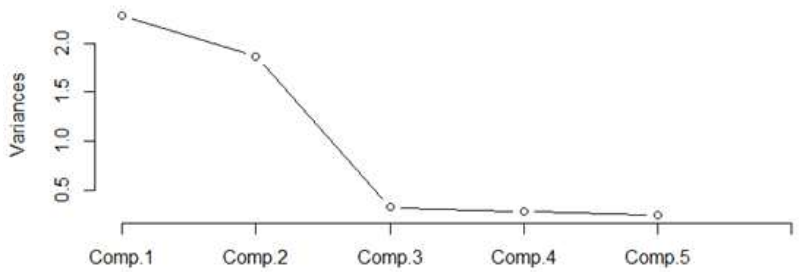

그림 3.2 스크리 그래프

주성분 분석 시 데이터에 대한 정보의 손실을 최소화하기 위해 적절한 개수의 주성분을 선택하는 방법으로 전체 변이에의 공헌도와 스크리 그래프를 활용하였다. 전체 변이의 최소 80%가 되도록 주성분의 수를 결정한다고 했을 때 2번째 주성분이 설명하는 누적값이 83%로 2개의 주성분으로 결정하였다. 스크리 그래프를 보았을 때 고유값의 변화가 작아져 경사가 완만해지는 부분은 3번째 주성분부터이므로 3개의 주성분으로 결정된다. 두 방법의 결과 다르므로 평균 고유값 방법을 사용하여 결정하도록 한다. 상관행렬을 사용하였기 때문에 평균 고유값은 1이 된다. 1을 넘는 고유값은 2개이므로 주성분 개수는 2개로 결정하였다.

표 3.3 주성분 계수

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
X1	-0.489	0.403		-0.309	0.709
X2	-0.598		0.285	0.741	-0.105
X3	-0.280	-0.587	0.614	-0.445	
X4	-0.112	0.665	0.305	-0.318	-0.592
X5	0.559	0.224	0.670	0.236	0.364

표 3.3은 Credit 데이터에 대한 주성분 계수를 나타낸 것으로, 두 개의 주성분에 대해 선형결합식으로 나타내면 다음과 같다.

$$Y_1 = -0.489X_1 - 0.598X_2 - 0.280X_3 - 0.112X_4 + 0.559X_5$$
$$Y_2 = 0.403X_1 - 0.587X_3 + 0.665X_4 + 0.224X_5$$

첫 번째 주성분은 전체 분산의 46%를 설명한다. 가장 많은 분산을 설명하고 있으며 고객이 소유한 총 신용카드(X2)와 고객의 총 전화 은행 이용 수(X5)의 계수들의 절댓값이 비교적 큰 것을 확인할 수 있다. 두 번째 주성분은 전체 분산의 37% 정도를 설명한다. 고객의 총 은행 방문 수(X3)와 고객의 총 온라인 은행 이용 수(X4)의 계수들의 절댓값이 비교적 큰 것을 확인할 수 있다.

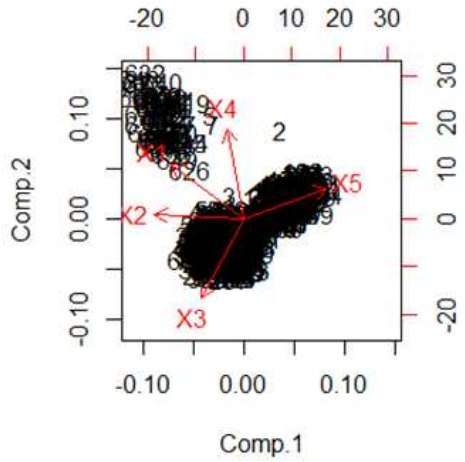

그림 3.3 상관행렬을 이용한 주성분 그래프

표 3.3과 그림 3.3을 이용하여 해석해 보았을 때, 첫 번째 주성분에서 X5의 값이 커질수록 주성분의 값이 커지고, X2와 X1의 값이 커질수록 주성분의 값이 작아지고 그래프의 화살표 방향이 반대이므로 X5와 (X2,X1)의 대비성분이다. 고객의 평균 신용카드 한도가 낮고 소유한 총 신용카드가 개수가 적을수록 고객의 총 전화 은행 이용 수는 늘어남을 알 수 있다. 두 번째 주성분에서 X4와 X1의 값이 커질수록 주성분의 값이 커지고, X3의 값이 커질수록 주성분의 값이 작아지고 그래프의 화살표 방향이 반대이므로 (X4,X1)와 X3의 대비성분이다. 이는 고객의 총 은행 방문 수가 감소할수록 고객의 온라인 은행 이용 수가 늘어나고 와 평균 신용카드 한도가 커짐을 알 수 있다.

3.3 인자분석

인자분석은 공분산 구조를 몇 개의 관측 불가능한 '인자'로 설명하려는 것으로 변수들 간에 내재하고 있는 공통의 구조를 파악하고, 데이터의 특성을 몇 개의 인자로 축약하여 설명하고자 하는 것이 인자분석의 목적이다. 즉, 인자분석은 여러 개 변수들의 상관성 구조를 나타내는 몇 개의 인자로 분석하는 것으로 인자분석을 통하여 데이터 분석자는 인자가 형성하는 차원과 그 차원에서의 변수의 위치 및 의미를 파악할 수 있다.

표3.4 인자의 회전 전과 Varimax인자 회전 후 인자 적재값

	회전 전		회전 후	
변수	Factor1	Factor2	Factor1	Factor2
X1	0.850	0.203	0.710	0.510
X2	0.791	-0.296	0.844	
X3		-0.813	0.348	-0.721
X4	0.476	0.730	0.165	0.856
X5	-0.615	0.567	-0.784	0.292

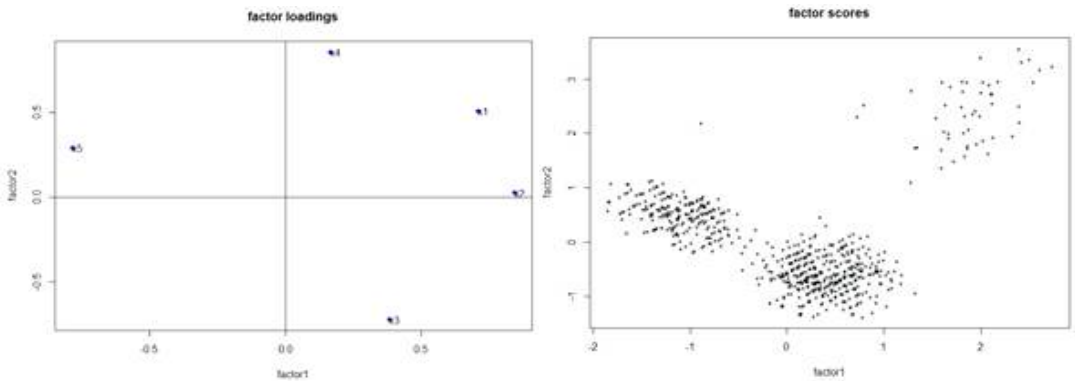

그림 3.4 인자 패턴과 인자 산점도

인자분석을 시행한 결과를 해석해 보면, 인자 적재값이 0.5정도 넘어가면 인자에서 높은 비중을 차지하는 것으로 판단할 수 있다. Factor1에서 X1(고객의 평균 신용카드 한도)과 X2(고객이 소유한 총 신용카드) 변수의 인자 적재값이 0.5보다 큰 값을 가지므로 '신용카드 관련 인자'라고 해석할 수 있다. Factor2에서 X1(고객의 평균 신용카드 한도)과 X4(고객의 총 온라

인 은행 이용 수) 변수의 인자 적재값이 0.5보다 큰 값을 가지므로 '온라인에서의 평균 신용카드 한도 관련 인자'라고 해석할 수 있다. 이를 기반으로 Varimax 회전 후 인자 모형은 아래와 같다. 그림을 통해 확인하였을 때도 인자1에서 인자 적재값이 높은 변수는 X1과 X2가 높은 것을 확인할 수 있다. 인자2에서도 동일하게 X1과 X4 변수의 인자 적재값이 가장 큰 것을 확인할 수 있다. 그리고 인자 모형은 다음과 같다.

$$X_1 = 0.710 \times F_1 + 0.510 \times F_2 + \epsilon_1$$
$$X_2 = 0.844 \times F_1 + \epsilon_2$$
$$X_3 = 0.348 \times F_1 - 0.721 \times X_2 + \epsilon_3$$
$$X_4 = 0.165 \times F_1 + 0.856 \times F_2 + \epsilon_4$$
$$X_5 = -0.784 \times F_1 + 0.292 \times F_2 + \epsilon_5$$

3.4 군집분석

군집분석이란, 관찰대상인 개체들을 유사성에 근거하여 보다 유사한 집단으로 분류하는 다변량 분석 기법이다. 그러한 군집분석의 첫 번째 목적은 적절한 군집으로 나누는 것이고 두 번째 목적은 각 군집의 특성, 군집간의 차이 등에 관한 탐색적 연구를 하는 것이다. 계층적 군집 방법은 최단연결법, 최장연결법, 평균연결법, ward의 계층적 군집 방법이 있고 비계층적 방법에는 k-means방법이 있다. 모든 계층적 방법을 시행해본 결과 군집 간 정보손실을 최소화하는 방법으로 ward 방법이 적합하다고 판단한다. 따라서 본 분석에서는 Ward의 계층적 군집 분석법과 비계층적 군집분석법인 k-means를 사용한다.

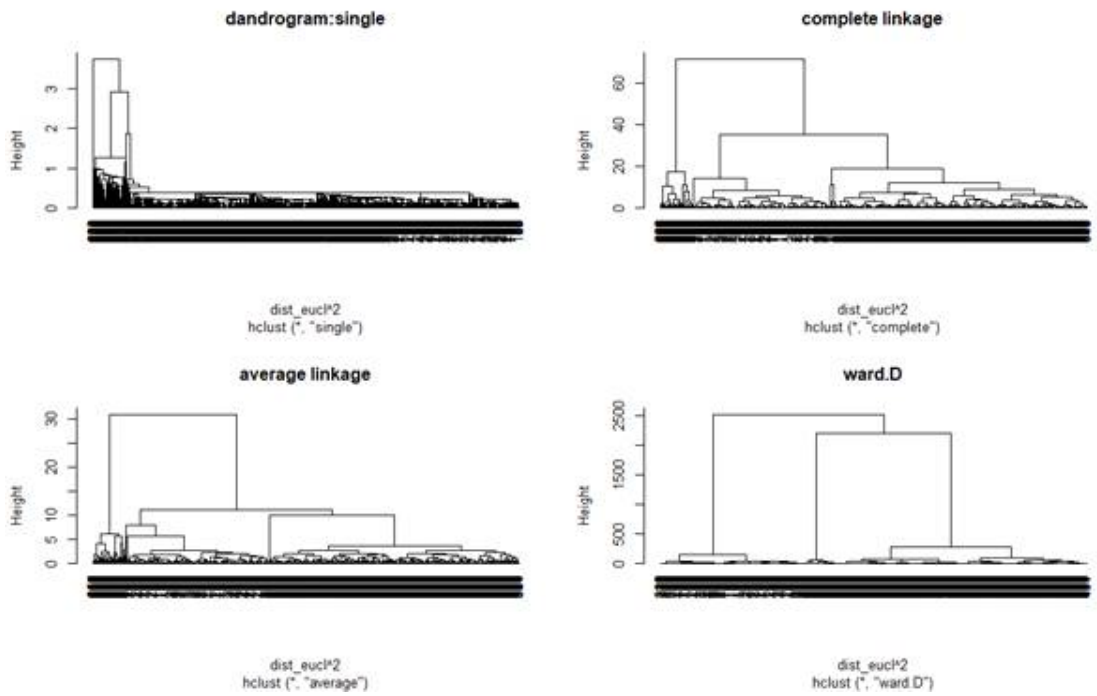

그림3.5 계층적 군집방법을 이용한 덴드로그램

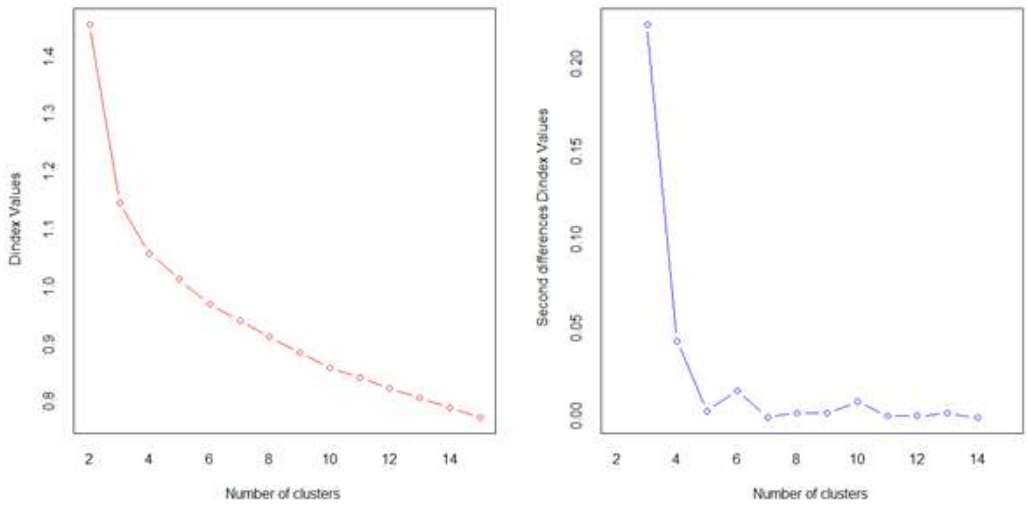

그림3.6 nbclust를 이용한 군집 개수 추정 결과

nblust를 이용하여 군집한 결과는 아래 그림3.6과 같다. 그래프를 통해 3개의 군집으로 정했다. Ward 방법을 통해 군집을 나누어 본 결과, 군집1에 385개, 군집2에 225개, 군집3에 50개로 분류된다.

표3.5 Ward 방법을 이용한 군집 결과

변수	군집1	군집2	군집3
고객번호	1 3 4 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 … 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 (총 385개)	2 6 8 9 10 11 12 13 14 15 … 224 225 226 227 228 289 314 (총 225개)	5 7 613 614 615 616 … 651 652 653 654 655 656 657 658 659 660 (총 50개)

표3.6 군집 별 평균값

	X1	X2	X3	X4	X5
Cluster1	2.83	1.86	-1.10	2.83	-0.87
Cluster2	-0.021	0.373	0.666	-0.553	-0.553
Cluster3	-0.60	-1.06	-0.90	0.32	1.15

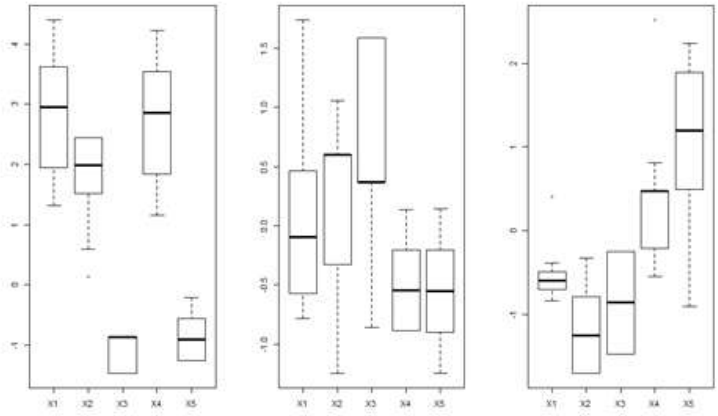

그림3.7 군집 별 boxplot

표3.6은 k-means 군집분석을 통한 각 군집 별 평균값을 나타낸 것이다. 군집1은 X1,X4의 평균이 가장 높게 나왔으며 군집2는 X3의 평균이 가장 높게 나왔고 군집3은 X5의 평균이 가장 높게 나왔다. 그림3.7은 각 그룹별로 어떤 변수가 가장 높은 평균값을 가지는지 시각적으로 쉽게 파악할 수 있다.

4. 결론

고객의 신용카드 정보 데이터를 이용하여 충성도에 따른 고객 세분화를 알아보았다. 먼저 기초통계량을 통해 데이터의 특징을 살펴보았다. 고객의 평균 신용카드 한도 범위가 크고 은행을 이용하는 3가지 방법에 대해 평균적으로 가장 많이 이용하는 방법은 전화를 사용한 은행 이용이었고 그 다음으로 온라인 은행, 오프라인 은행 순임을 알았다. 하지만 전화를 사용한 은행 이용은 3가지 방법 중에서도 편차가 있는 편임을 보았다. 상관분석을 통해 변수 간에 선형관계를 확인해 보았다. 그 중에서 고객의 평균 신용카드 한도와 고객이 소유한 총 신용카드의 상관성이 가장 높았고, 고객의 평균 신용카드 한도와 오프라인 은행의 방문 수에 대한 상관성이 가장 낮았다. 다음으로 주성분 분석을 통해 2개의 주성분을 선택하였다. 전체 변이를 최소 80%가 넘도록 하였고 추가 적으로 평균 고유값인 1을 넘는 2개의 고유값을 통해 주성분 개수를 2개로 결정하였다. 첫 번째 주성분이 46%로 가장 많은 분산을 설명한다. 상관행렬을 이용한 주성분 그래프를 통해 첫 번째 주성분은 고객의 평균 신용카드 한도가 높고 고객이 소유한 총 신용카드 수가 많을수록 고객이 전화를 사용하여 은행을 이용한 총 수가 적어짐을 알 수 있다. 두 번째 주성분은 고객의 평균 신용카드 한도가 높고 고객이 온라인을 통해 은행을 이용하는 총 수가 클수록 고객이 오프라인으로 은행을 방문하는 총 수는 적음을 알 수 있다. 인자분석의 경우 주성분과 동일한 기준을 통해 2개의 인자를 선택하였다. 첫 번째 인자는 X1(고객의 평균 신용카드 한도) 과 X2(고객이 소유한 총 신용카드 수)의 인자 적재값이 0.5 이상이므로 '신용카드 관련 인자'라고 정했다. 두 번째 인자는 X1(고객의 평균 신용카드 한도) 과 X4(고객의 총 온라인 은행 이용 수)의 인자 적재값이 0.5 이상이므로 '온라인에서의 평균 신용카드 한도 관련 인자'라고 정했다. 마지막으로 군집분석에서 계층적 군집 방법에서는

Ward 방법을 사용하였고 비계층적 군집방법으로는 k-means 방법을 이용하여 분석하였다. 두 분석에 차이가 거의 없었고 각 군집은 충성도에 따라 3개의 군집으로 나누었다. 전체 관측 개수는 약 군집1은 385개, 군집2는 225개 군집3은 50개로 분류되었다. 그 중에서도 군집1이 평균적으로 높은 수치를 기록하는데 이는 세 개의 군집 중에 충성도가 가장 높다고 할 수 있다. 따라서 충성도가 높은 고객은 평균적으로 신용카드의 한도가 높고 신용카드의 개수가 많으며 은행 이용 수가 많다는 특징을 가지고 있음을 알 수 있었다.

요약

고객의 신용카드 정보를 이용하여 충성도가 높은 고객에 대한 특징을 알아보았다. 대체로 신용카드 한도가 높고 신용카드를 보유한 수가 많고 온라인 은행을 이용하는 횟수가 높을수록 충성도가 높은 고객으로 분류되는 것을 알 수 있었다.

참고 문헌 및 사이트

김재희 (2020) R 다변량 통계분석(개정판), 교우사.
덕성여자대학교 정보통계학과 (2019) 다변량 통계분석 모음집, 덕성여자대학교
케글(Kaggle) Credit Card Customer Data | Kaggle