碩士學位論文

RFM 모형의 가중치 선택에 관한 연구

Data driven selection methods of weights in RFM Model

濟州大學校 大學院

電算統計學科

金 東 錫

2021年 6月

RFM 모형의
가중치 선택에 관한 연구

Data driven selection methods of weights in RFM Model

指導敎授 金 鐵 洙

金 東 錫

이 論文을 理學 碩士學位 論文으로 提出함

2021年 6月

金東錫의 理學 碩士學位 論文을 認准함

審査委員長 _____㊞

委　　員 _____㊞

委　　員 _____㊞

濟州大學校 大學院

2021年 6月

Data driven selection methods of weights in RFM Model

Kim, Dong-Seok
(Supervised by professor Kim Chul-Soo)

A thesis submitted in partial fulfillment of the requirement for the degree of Master of Science

June. 2021.

This thesis has been examined and approved.

.....
Thesis Committee Chair, Lee Yun-Jung, Prof. of Computer science and Statistics.

.....
Thesis Committee Member, Kim Chul-Soo, Prof. of Computer science and Statistics.

.....
Thesis Committee Member, Lee Bong-Kyu, Prof. of Computer science and Statistics.

June. 2021.

Department of Computer science and Statistics
GRADUATE SCHOOL
JEJU NATIONAL UNIVERSITY

Abstract

Data driven selection methods of weights in RFM Model

Kim Dong-Seok
Department of Computer science and Statistics
The Graduate School of Jeju National University

As we enter the era of the 4th industrial revolution in earnest, various strategies using data are being proposed in the overall field. This made it possible to devise a new marketing methodology that can reduce material and time costs and maximize corporate profits through strategic decision-making.

In a rapidly changing lifestyle, customers express various desires, and in order to meet these customer expectations, companies can refine and analyze customer purchasing data to predict customer purchasing patterns and retain existing customers while recruiting new customers suggesting a way out. Among these methods, customer relationship management (CRM), which is widely used representatively, can be analyzed in various ways in connection with data mining techniques as a large amount of data is accumulated today. The RFM (Recency, Frequency, Monetary) model is a simple and convenient modeling method as one of the traditional customer relationship management techniques. However, the most important thing in RFM model design is to assign weights to each variable, and there is no clear standard for weighting so far.

In this paper, we propose a weight selection method that can design the RFM model more efficiently. One of the big data analysis techniques, the K-Means algorithm represented by clustering, is clustered and selected as the weight of the RFM model by utilizing the Coefficient of Variation (CV) values of the three components of each cluster: Recency, Frequency, and Monetary. The purpose is to calculate the RFM

score from the finally designed RFM model, assign it to each customer, and subdivide the entire customer into five graded groups.

A total of three datasets were used in this study, and each dataset consists of 6,116, 549,019, and 6,919 customer purchase date(Recency), purchase frequency(Frequency), and purchase amount (Monetary) are included.

Using Python's machine learning library, the optimal K value obtained through the Elbow method is applied to each dataset, and the final RFM model is designed by selecting weights through the statistics of R, F, and M variables of each group divided into K groups. Scores were given to all customers and classified into five groups.

This allowed us to overcome the limitations of purchase data characteristics without target variables, and K-Means Clustering allowed us to design RFM models with weights selected in an objective way that reflected the characteristics of clustered group variables, which enabled better customer segmentation.

Keywords : Customer Relationship Management, Cluster Analysis, K-Means Clsutering, RFM Analysis, RFM weight, Customer segmentation.

목 차

표 목차

그림 목차

제 1 장 서 론

제 1절 연구 배경

방대한 데이터가 출현하면서 각 기업의 이윤을 극대화하기 위한 마케팅 활동의 방향성은 빠르게 변화하고 있다. 급변하는 현대 사회에서 소비자들의 욕구는 더욱 다양하고 복잡해지고 있으며 이러한 고객을 확보하기 위한 각 기업 간의 경쟁은 치열해지고 있다. 이러한 환경 속에서 기업들이 고객의 특성을 파악하고 고객의 가치를 측정하여 적절한 마케팅 전략을 수립하는 것은 고객 가치를 상승시키고 지속적 경쟁 우위를 유지하는데 매우 중요하게 작용한다. (전희주, 2011)

CRM(Customer Relationship Management, 고객관계관리)은 기존의 고객은 유지하면서 신규 고객을 유치하려는 방안으로 널리 활용되는 기법으로서 오늘날 많은 기업에서 이를 기반으로 한 마케팅 활동을 펼치고 있다. 최근 정보통신 기술의 급격한 발달과 동시에 빅데이터 시대가 도래하면서 CRM에 필요한 대량의 데이터가 축적되고 있으며 고객의 구매 데이터의 특성을 대량의 데이터를 토대로 분석할 수 있게 되었다. 각 기업 입장에서는 이러한 데이터를 통해 불특정 다수 고객의 특성을 파악하고 특성별로 고객을 그룹화하여 개인화된 맞춤형 마케팅을 제공하여 마케팅 효율성을 극대화할 수 있는 기회의 시대가 열린 것이다. 이렇게 얻어진 방대한 데이터 속에서 유의미한 정보를 찾아내서 전략적으로 활용하는 데이터마이닝 기법 또한 적극적으로 사용되고 있다. 데이터마이닝이란 대량의 데이터 안에서 기존에는 알려지지 않았던 특정한 패턴이나 규칙을 찾아내어 모델링함으로써 유용한 지식을 추출하는 반복적 프로세스이다.(조혜정, 2001)

데이터마이닝 기법에는 여러 가지 방법이 제시되고 있는데 목적에 따라 유연하게 활용할 수 있으며 Pang-Ning Tan(2005)은 이러한 데이터 마이닝 기법을 크게 Clustering(군집화), Association Rules(연관규칙), Anomaly Detection(이상감지), Predictive Modeling(예측모델) 등으로 구분하였고 Bharati M.

Ramageri(2010)은 Classification(분류), Clustering(군집화), Predication(예측), Association rule(연관규칙), Neural networks(신경망)으로 구분했으며 이승영(2003)은 분류(classfication), 연속패턴(sequential pattern), 연관(association), 군집(Classification) 등으로 분류했다. 결국 데이터를 요약하고 군집하고 분류하며 데이터 간의 관계 및 성향, 패턴 인식 등을 파악하여 얻고자 하는 정보를 얼마나 잘 해석하고 분석할 수 있느냐가 데이터 마이닝의 핵심이라고 할 수 있다.(최영희, 2001)

고객을 세분화하기 위한 마케팅 전략으로는 상품에 의한 세분화, 고객 프로파일에 의한 세분화, RFM에 의한 세분화, LTV(Life Time Value, 고객 생애 주기)에 의한 세분화 등 여러 가지 있으며(이소영, 2004) 각 방법은 방법의 차이만 있을 뿐 고객을 그룹화하여 그룹별로 차별화 마케팅을 진행하기 위한 목적으로 사용된다. 환대산업 연구에서도 데이터마이닝을 활용한 구매 패턴 분석 기법이 중요하게 떠오르고 있는데 그중에서도 앞서 언급한 고객을 세분화하는 분석 방법인 RFM 분석기법이 자주 사용되고 있다.(이지민, 2020)

본 연구에서는 고객 거래 데이터의 최근성, 거래 빈도, 거래 규모 등의 데이터를 활용한 RFM 모형을 다룬다. Recency(최근성), Frequency(거래 빈도), Monetary(거래 규모)의 약자인 RFM 모형은 고객을 그룹화하는 편리한 모델링 방법으로 간단하게 설계할 수 있는 모형임에도 불구하고 예측력이 높아서 상당히 광범위하게 사용되고 있다.(김동훈, 2007)

이러한 RFM 분석에서 가장 중요한 것은 가중치를 선택하는 것이며 기업마다 가중치를 선택하는 기준이 다르므로 기업을 대표할 수 있는 가중치를 선택하는 것이 중요하다.(문영수, 2005) 그러나 RFM 모형의 점수를 산출하는 과정에서의 핵심 문제인 가중치 선택에 관한 기준은 여전히 불분명한 상태이다. 가중치를 선택하는 기준은 기업마다 혹은 관련 연구마다 상이한데 전통적으로는 기업 마케팅 담당자가 각 요소별 중요도를 직접 판단하고 분류하여 직관적으로 모형을 구축해왔다.(최희경, 2002)

이 밖에도 이소영(2004)은 최적의 가중치 선정을 위해 RFM 점수에서 상위 20% 해당하는 고객의 평균을 산출하여 전체 자료의 RFM 요소에 해당하는 상위 20%의 평균값을 비교하여 유사한 수준을 최적의 가중치로 선택하였으며 이윤성

(2010)은 빈도 및 점수를 활용하는 산출식을 제안하였고 정윤필(2009)은 연구하고자 하는 데이터의 해당 쇼핑몰의 마케팅 팀과의 토의를 거쳐 임의로 가중치를 부여하였다.

이렇게 그간의 선행연구들을 고찰해본 결과 RFM 모형 분석에 관한 연구는 오랫동안 지속하였지만, 획일화된 가중치 선택 방안은 거의 없는 상황이며 각 담당자의 직관으로 결정이 되는 경우가 많았다.

따라서 본 연구에서는 RFM 모형 구축 시 적절한 가중치를 선택하는 방법을 제시하고자 한다.

제 2절 연구 목적

본 연구의 목적은 고객 구매 데이터를 기반으로 한 RFM 모형 설계 시 가중치 선택에 관한 구체적인 방법을 제안하는 것에 있으며 과거에 여러 가지 연구에서 제시되었던 가중치 선택 방법 대신 데이터마이닝 기법과 기술 통계량을 근거로 하여 새로운 가중치 선택 방법을 제시하고자 한다. 이는 기존 RFM 모형의 가중치는 일반적으로 마케팅 담당자나 연구자의 직관으로 선택되거나 로지스틱 회귀분석 등의 복잡한 통계 기법을 활용하여 산출되었는데 본 연구에서는 기계학습 기법과 기술 통계량을 활용하여 고객이 갖는 변수의 특성을 고려하면서도 더욱 객관화된 방법으로 가중치를 산출하고자 한다. 연구의 세부적인 단계별 목적은 다음과 같다.

첫째, 데이터마이닝의 군집분석 기법의 하나인 K-means Clustering 방식을 활용하여 데이터를 군집화하고 많은 양의 데이터를 변수의 특성에 따라 나누고자 한다.

둘째, 군집화한 후 각 그룹의 특성을 나타내는 기술 통계량을 토대로 계산된 가중치 값을 제시하고자 한다.

셋째, 앞서 계산된 가중치를 최종 RFM 모형에 적용함으로써 실질적인 모형을 설계하도록 제시하고자 한다.

마지막으로, 위 방법을 통해 설계된 모형을 활용하여 전체 고객에게 RFM 점수를 부여하고 최종 5등급으로 세분화하여 각 집단의 차이를 비교해보고 적절히 분류되었는지 통계적 방법을 통해 검정한다. 본 연구에서는 총 3개의 데이터셋을 활용하였으며 여러 가지 가중치 선택 방법을 비교하는 데 의미가 있다.

제 3절 연구 방법 및 범위

본 연구는 K-means Clustering 기법을 활용하여 전체 고객 구매 데이터를 최적의 K개 그룹으로 분류하고 각 그룹의 R, F, M 세 요소의 평균, 표준편차 등의 기술 통계량을 기반으로 RFM 모형의 가중치를 제시해보고자 한다.
상세 연구 방법은 다음과 같다.

첫째, 연구에 활용할 고객 구매 데이터를 수집하고 Microsoft Excel, R, Python 등을 통해 데이터를 정제한다. 이때 활용된 데이터는 여러 연구에서 두루 쓰이는 3개의 데이터셋을 활용한다.

둘째, K-means Clustering 기법을 통해 정제된 데이터를 군집화한다. 이때 Python의 기계학습 라이브러리인 Scikit-Learn의 K-means Clustering 알고리즘을 활용하여 Elbow-method를 통해 최적의 K값을 찾아 데이터를 군집화한다.

셋째, 군집화된 각 그룹의 Recency, Frequency, Monetary 세 요소의 평균 및 표준편차를 이용해 CV(Coefficient of variation, 변동계수)를 구하고 그 비율을 통해 각 요소별 가중치 값을 제시한다.

마지막으로, 산출된 가중치 값을 토대로 최종 RFM 모형을 완성하고 전체 고객에게 부여하여 최종 5개의 등급으로 세분화하고 각 집단별로 어떠한 차이점이 있는지 살펴본다. 이후 기존의 여러 가지 연구에서 일반적으로 쓰이는 가중치 선택 방법을 참고하여 RFM 모형을 완성한 뒤 본 연구에서 제시하고자 하는 CV를 통해 산출된 가중치를 선택한 RFM 모형을 비교함으로써 각 모형으로 세분화한 고객의 5개 등급의 집단 사이에는 어떠한 차이점이 있는지 살펴본다.

제 4절 연구 구성

본 논문은 서론, 이론적 배경, 연구 방법, 연구 결과, 결론 등 크게 5장으로 구성되어 있으며 제 1장 서론에서는 연구 배경 및 목적, 방법 및 범위에 대해 서술한다.

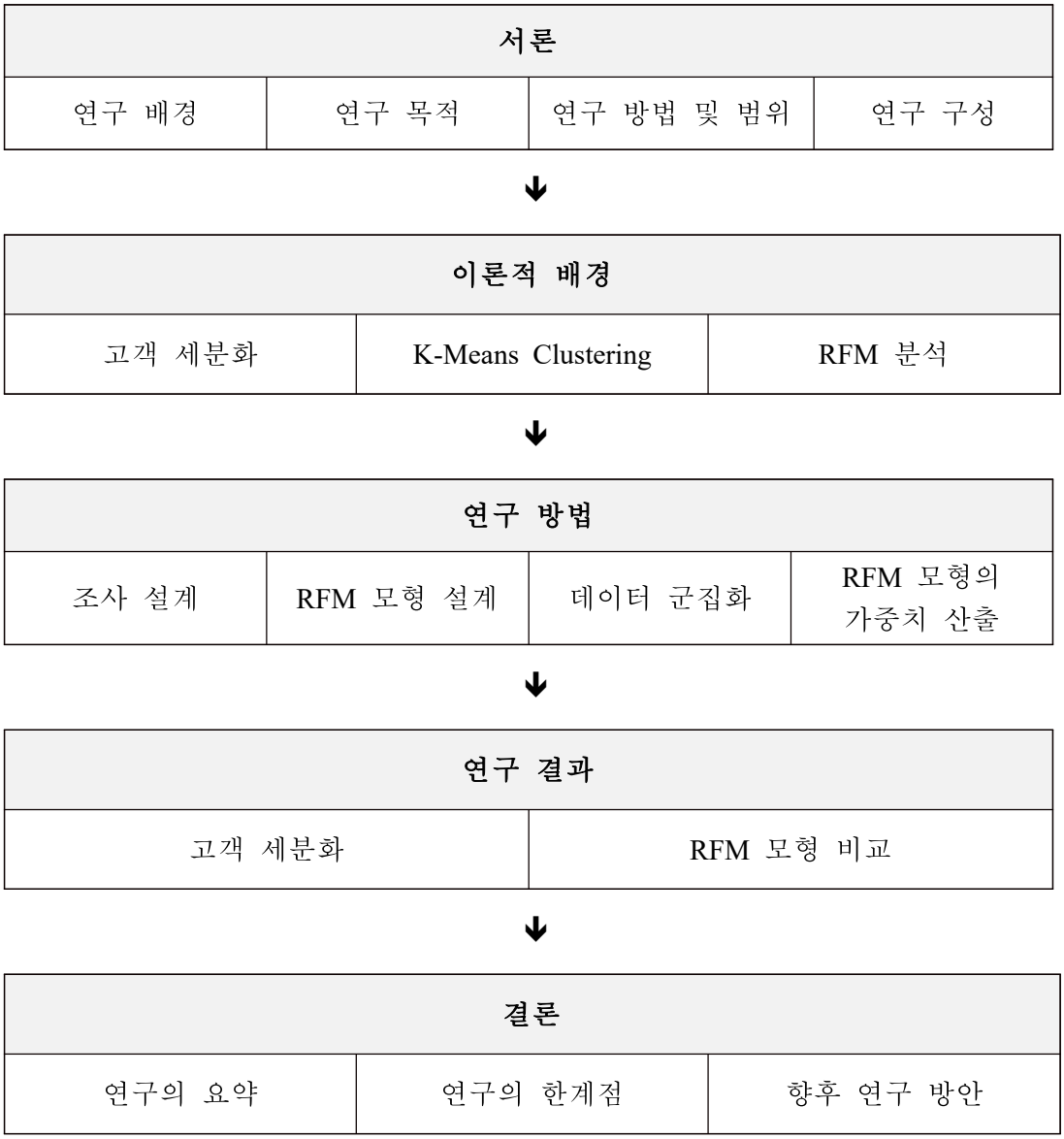
제 2장 이론적 배경에서는 고객세분화 및 K-Means clustering의 개념에 대한 내용과 절차에 대해 알아본다. 또한 RFM 모형의 개념 및 점수 산출 과정에 대한 내용과 가중치 선택에 관한 선행연구에 대해 기술한다.

제 3장 연구 방법에서는 각기 다른 특성이 있는 3가지 데이터셋을 선정하고 해당 데이터를 정제하여 RFM 분석을 실시한다. 또한 각종 통계적 기법 적용을 위한 Microsoft Excel, R, Python 등의 라이브러리 및 알고리즘을 사용한다.

제 4장 연구 결과에서는 제 3장의 연구 방법의 결과를 제시한다. 실제 데이터에 K-means Clustering을 적용하고 도출해 낸 각종 기술통계량을 기반으로한 가중치를 산출하고 최종 RFM 모델에 적용하여 비교한다.

제 5장에서는 연구의 결과를 요약하고 시사점 및 한계점을 기술한다.

본 연구 흐름도는 다음 <그림 1-1>과 같다.



<그림 1-1> 연구 흐름도

제 2 장 이론적 배경

제 1절 고객세분화

고객 세분화(Customer Segmentation)는 전체 고객을 여러 가지 변수를 토대로 유사한 고객끼리 세분화된 집단으로 분류하는 것을 의미한다. 즉, 기업이 제품 또는 서비스를 판매하기 위한 목표 대상을 선정하기 위해 고객들이 가진 중요한 정보의 차이를 기준으로 하여 전체 시장을 나누어 마케팅 효과를 극대화하고자 하는 것이다.(안범준, 2011)

이렇게 분류된 고객은 각 집단별 성향이나 특성이 있으므로 서로 다른 마케팅 전략 구현을 가능하게 하며 불특정 다수 고객이 아닌 특정 집단에 집중적인 판매나 홍보 활동을 진행하면서 시간적, 금전적 감소 효과를 얻을 수 있으며 더 높은 성과를 기대할 수 있다.(서현지, 2017)

고객을 세분화하는 방법은 나누는 기준에 따라 여러 가지 측면으로 분류할 수 있다. 이영진(2010)은 하나의 변수를 기준으로 구분하는 전통적 세분화 방법과 여러 가지 변수를 통해 얻은 다기준 스코어에 의한 방법 등 크게 두 가지로 구분하였다. 하나의 변수로 구분하는 방법은 사용성이 편리하다는 장점이 있지만 비교적 정확성이 떨어진다는 단점이 있다. 이러한 단점을 극복하고자 등장한 것이 다기준 스코어에 의해 구분되는 방법인데 대표적인 방식으로 RFM 기법, LTV 기법, 통계적 기법, 데이터 마이닝 기법 등이 있다.(정윤필, 2009)

전용준(2007)은 그리드 방법, 군집화 방법, 의사결정나무 방법 등 세 가지로 구분하였으며 그리드 및 군집화 방법은 기술적 분석으로, 의사결정나무 방법은 예측적 분석으로 분류하였다. 이처럼 고객을 세분화하는 방법은 여러 가지가 있지만 고객 세분화 연구의 목적은 각 기업의 실정에 맞는 세분화 기준 변수를 적절히 찾아내는 것에 있으며 실제 사용되는 변수는 각 기업의 상황에 따라 천차만별이어서 어떤 방법이 절대적으로 좋다는 것은 있을 수 없고 특정한 상황에 특

정한 기준이 적절했다고 검증하는 것이 중요하다.(황준경, 2006)

고객을 세분화하는 것에는 여러 가지 장점이 있는데 첫째, 기업 입장에서 수익성에 가장 밀접한 고객군과 그렇지 않은 고객군을 분류하여 수익성이 높은 고객군에게 마케팅 활동을 집중할 수 있게 하고 둘째, 충성 고객의 니즈를 파악하여 그에 맞는 제품이나 서비스를 제공하여 지속적인 관계 유지를 이어 나갈 수 있으며 셋째, 고객 서비스를 향상시키고 고객의 니즈를 충족시키기 위한 제품의 향상 또는 조정을 할 수 있으며 마지막으로 마케팅 비용의 효율을 높여 원가를 절감하고 수익은 증대시키는 기대효과를 가지고 올 수 있다.(Express Analytics)

본 논문은 RFM 분석을 통해 전체 고객을 특정한 개수의 집단으로 세분화하는데 목적이 있으며 기존의 마케팅 담당자에 의해 직관적으로 선택되거나 복잡한 통계 기법을 통해 얻어지는 RFM 모형의 가중치를 보다 간편하고 각 집단의 변수 특성이 더욱 잘 반영되도록 하는 방법을 연구하고자 한다.

제 2절 K-Means Clustering

K-Means Clustering은 데이터마이닝의 큰 분류인 비지도학습(Unsupervised Learning)의 대표적인 기법으로 군집분석 방식 중 가장 대표적으로 활용되는 분석 알고리즘이다.(이창해, 2015) 비교적 쉽고 직관적이며 각 군집별 유형 특징 파악이 용이하다는 장점 때문에 여러 문제에 쉽게 적용되었다.

이 알고리즘은 거리를 기반으로 하는 분류이며 K개의 중심 좌표와 각 데이터를 (식 2-1)과 같이 유클리디안 거리로 계산하여 그 평균 거리를 최소화하는 것으로 데이터를 가까운 K 군집에 배정한다.

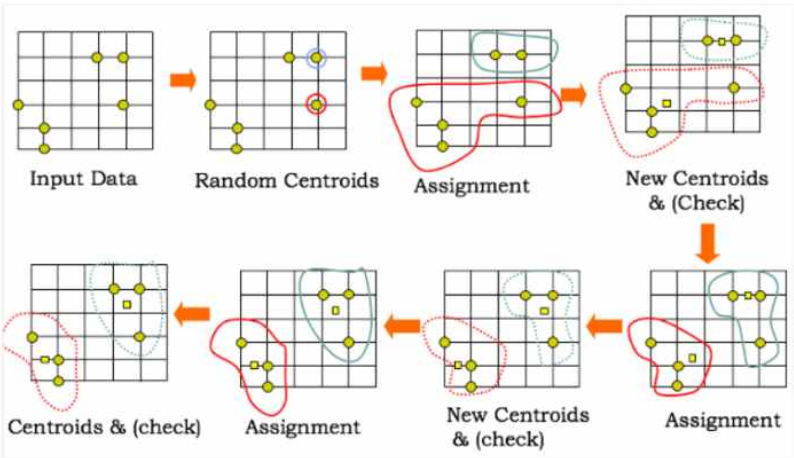
$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots(p_n-q_n)^2}=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \quad : (식\ 2\text{-}1)$$

K-Means Clustering 수행 절차는 <그림 2-1>과 같이 크게 4개로 구분할 수

있으며 이에 대한 자세한 절차는 <그림 2-2>와 같다.

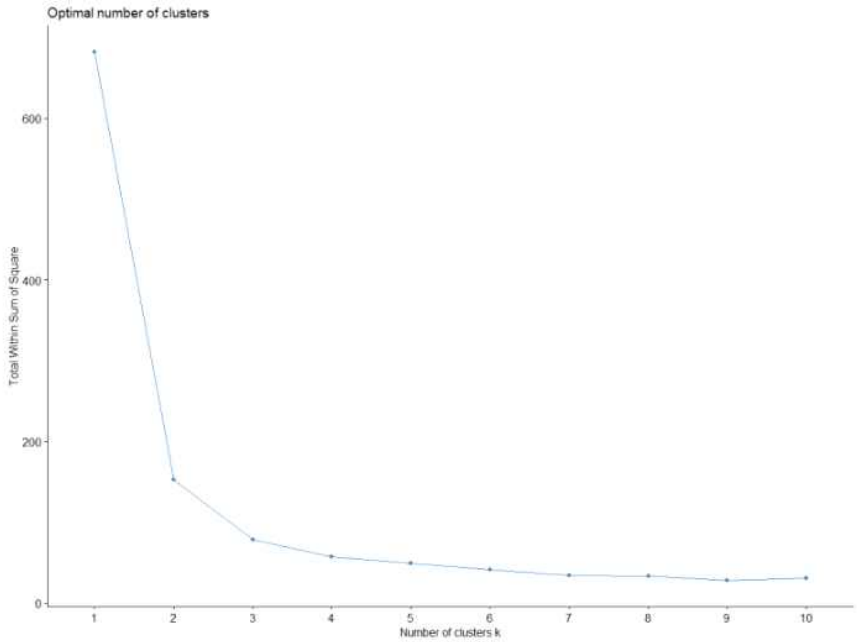
초기 K값 선택	초기 군집의 수 K개의 중심 좌표(Centroid) 계산
↓	
클러스터 배정	각 데이터를 중심 좌표와의 거리를 산출하여 가장 가까운 군집에 배정
↓	
신규 K값 선택	각 군집별 새로운 중심 좌표 계산
↓	
범위 확인	모든 데이터가 더 이상 군집에 배정되지 않을 때까지 반복 시행 후 K개의 최종 군집 형성

<그림 2-1> K-Means Clustering 4단계



<그림 2-2> K-Means Clustering 수행 절차

본 연구에서는 RFM 모형의 가중치를 선택하기 위해 위와 같은 K-Means Clustering 기법을 사용하였고 이때 최적 군집의 개수, 즉 K값을 찾는 방법으로 Elbow Method를 이용하였다.



<그림 2-3> Elbow Method 그래프 예시

Elbow Method는 클러스터 수를 순차적으로 늘려가면서 SSE(Sum of Squared Error, 오차제곱합)값을 계산하고 그 변화의 크기가 최소화되는 시점을 최적의 클러스터 개수로 판단하는 방법론이다.(김태진, 2020)
<그림 2-3>의 예시에서 x축 값이 3일 때 y축의 기울기가 급격히 완만해지는 것을 볼 수 있는데 이는 SSE의 감소율이 크게 작아지는 것이고 이 경우 최적의 클러스터 개수로 3을 선택할 수 있다.

제 3절 RFM 분석

1. RFM 분석의 개념

RFM은 CRM(Customer Relation Management, 고객관계관리)의 한 기법으로서 고객 세분화 작업 시 가장 흔하게 활용되는 점수부여방식(Scoring system)의 분석기법이다.(염창선, 2013)

고객 세분화를 통한 인터넷 쇼핑몰의 고객관계관리 RFM모델은 Bult & Wansbeek(1995)에 의해 처음 정립되었는데 이는 '얼마나 최근에 구입했는가(Recency)', '얼마나 자주 구입했는가(Frequency)', '얼마나 구매했는가(Monetary)' 등 세 가지 기준으로 고객의 수익성을 평가하는 모델이다.

R(Recency, 최근성), F(Frequency, 빈도), M(Monetary, 구매 금액)으로 이루어진 RFM 기법은 세 가지 요소를 가지고 고객 기여도에 근거하여 고객의 수익성을 평가하는 모델링 기법으로 매우 간편하게 접근할 수 있어 많은 기업에서 마케팅 방법으로 널리 사용하고 있는 분석 방법 중 하나이다.(김동훈, 2008)

일반적으로 R(최근성)은 가장 최근에 구매한 고객일수록 재구매율이 높아지고 구매 시점이 오래된 고객일수록 재구매율이 낮아지게 되는데 Gönül, Kim, & Shi(2000)에 따르면 최근에 온라인으로 물건을 주문한 고객은 최초 주문에 대한 보완이나 추가 주문, 물품을 받은 후 신뢰감 상승에 대한 연이은 주문 등 재구매로 이어질만한 여지가 많아지기 때문이다. 그러나 최근성과 구매행위의 상관관계는 물품에 따라 전혀 다르게 나타나기 때문에 이는 구매 시점과 품목의 정확한 데이터를 파악하기 전에 최근성과 구매의 명확한 상관관계는 알기 어렵다.(이지민, 2020)

F(빈도)는 일반적으로 특정 기간 동안의 구매 횟수를 비교하는 변수지만 각 기업 마케팅 담당자나 구매 데이터의 특성에 따라 다양하게 사용된다. 1년을 기준으로 한 고객별 구매 횟수를 사용하기도 하고 총 판매 품목 수량을 보기도 하며 통화 횟수, 입출금 횟수, 외래방문 횟수(이소영, 2005), 자연재해 빈도(김태진, 2020), 총 구매 수량(류귀열, 2006; 이영진, 2010) 등을 사용한 것이 그 예이다.

일부 마케팅 담당자는 구매 횟수를 고객으로 있었던 기간(duration of being the customer)으로 나눈 값을 사용하기도 하는데, 이렇듯 빈도는 조사하고자 하는 목적에 맞게 최선의 변수를 찾아서 대입시켜 측정하면 되는 것이다.(Hughes, 2000)

M(구매 금액)은 특정 기간 동안의 총 구매 금액의 합계를 볼 수도 있고 총금액을 고객으로 나누어 주문 당 평균 구매금액으로도 볼 수 있다.(Bult& Wansbeek, 1995; Hughes, 2000). F와 M은 구매 행위와 밀접한 상관관계를 갖는 경향이 있다.

2. RFM 모형의 설계

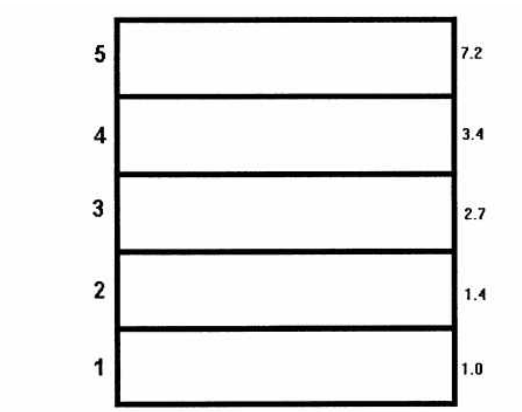
일반적으로 R, F, M 세 요소를 기반으로 고객들에게 점수를 각각 부여하여 상위 20% 단위, 혹은 10% 단위 등으로 고객을 구분하는 것이 RFM 분석의 기본이며 이러한 RFM 모형 설계에는 여러 가지 방법이 있는데 각 요소에 점수를 부여하고 전체 고객을 3, 5, 10개의 그룹으로 나누어 단일 지표로 나타내는 전통적인 방법과 R, F, M을 각각 세분화한 조합을 활용한 집단 세분화 방법이 대표적이다.(최희경, 2002)

전통적인 RFM 모형에서 가장 일반적으로 사용되는 것은 (식 2-2)와 같이 각 변수를 20%씩 5개의 집단으로 나누어 상위 집단에 속한 고객에게는 5점, 다음 20% 고객에게는 4점, 마지막 하위 20% 고객에게는 1점을 부여하는 방식으로 나열하는 것이다.

$$\text{점수} = R_i + F_i + M_i,\ i = 1, 2, 3, 4, 5 \quad : (\text{식 } 2\text{-}2)$$

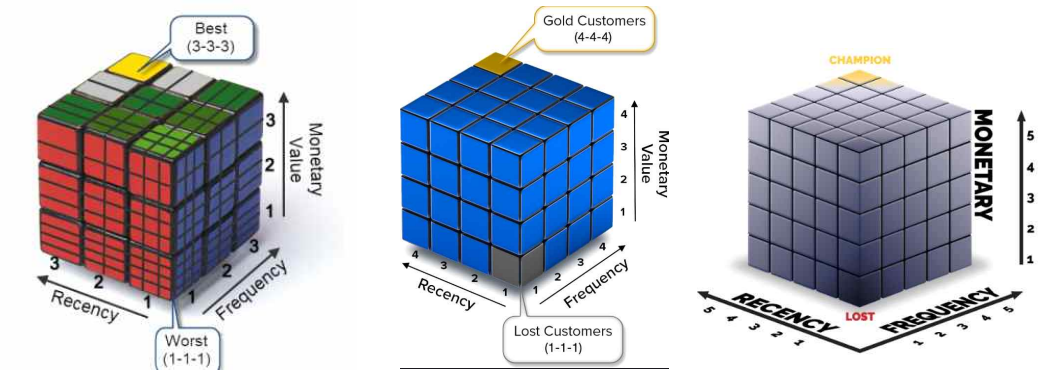
John R. Miglautsch(2000)에 따르면 이러한 RFM 모형은 반드시 균등하게 나눌 필요는 없으며 데이터가 특정 집단에 지나치게 많거나 적은 경우 이를 적절하게 분배할 수 있다. 이는 많은 기업의 구매 데이터에서 단 한 번의 주문만을

하는 고객의 비율이 30~60%를 차지하는데, 이 경우 F 변수에서 2점을 부여받는 고객 중 30%가 1점과 동일한 행동으로 한 것으로 여겨지게 되는 오류가 억제하기 위함이다. <그림 2-4>는 John R. Miglautsch가 제시하였는데, 전체 고객의 28%가 단 한 번의 구매만 일어났을 경우 2점을 받아야 하는 고객 중 1/3은 1점을 받은 사람과 동일하게 여겨질 수 있다는 것을 지적하였다.



<그림 2-4> 균등하게 나눈 F값의 분포 (출처: John R. Miglautsch)

각 기업은 이러한 방법에 따라 균등하거나 주관적인 범위로 각 집단에게 점수를 부여할 수 있으며 이렇게 되면 최종적으로 모든 고객은 세 요소에 대한 점수가 각각 매겨진다. Arthur Hughes(2000)의 제안에 따라 R, F, M 세 요소에 대해 각각 1점에서 5점까지의 점수를 매기면 5(R) * 5(F) * 5(M) = 125개(111, 112, 113, …, 553, 554, 555)의 세분화된 집단이 형성된다. 이 역시 반드시 5개씩 구분지을 필요는 없고 기업이 갖고 있는 고객 구매 데이터의 양에 따라 <그림 2-5>와 같이 3*3*3 = 27개, 4*4*4 = 64개, 5*5*5 = 125개 등 자유롭게 결정할 수도 있으며 중요시되는 특정변수나 중간값, 평균값을 활용하여 2*2*5 = 20개, 2*2*2 = 8개의 집단으로 구분할 수도 있다.(이지민, 2020)



<그림 2-5> RFM Cube

3. RFM 모형의 가중치 선택에 관한 선행연구

RFM 모형을 통해 세분화된 집단은 27개, 64개, 125개 등으로 그 수가 너무 많아서 실질적으로 마케팅 활동을 하기가 어려우므로 RFM 점수를 기준으로 등급을 나누게 되는데 일반적으로 20%씩 나누는 5등급 또는 10%씩 나누는 10등급 등을 많이 사용하고 있다.(류귀열, 2013) 하지만 이렇게 고객들을 5개 또는 10개의 등급으로 구분하더라도 이는 R, F, M 각 요소의 중요도를 판별할 수 있는 근거가 부족하였으며 이를 보완하고자 가중치 부여에 대한 연구들이 끊임없이 진행되어 왔다.(John R. Miglautsch(2000); 최희경(2002); 이소영(2004); 박희창(2010) 등)

이는 전통적인 RFM 모형에서 각 요소에 가중치를 부여하는 것으로 식으로 표현하면 (식 2-3)과 같다.

$$RFM\text{점수} = W1 \times Recency + W2 \times Frequency + W3 \times Monetary \quad : (\text{식 } 2\text{-}3)$$
$$(W1 + W2 + W3 = 1)$$

예를 들어 가중치가 고려되지 않은 RFM 방법으로 등급을 구분할 경우 R, F, M 각각의 점수가 5점, 4점, 1점인 고객과 1점 5점, 4점인 고객은 같은 등급으로

분류된다. 이때 가장 최근에 방문했지만 구매 금액이 적은 고객과 최근에는 방문하지 않았지만 과거에 자주 방문하면서 구매 금액도 높았던 고객을 동일한 집단으로 보는 것이 맞는지에 대한 의문점이 생기게 된다. 이 과정에서 해당 기업이나 마케팅 담당자가 중요시 생각하는 변수에 더 높은 가중치를 부여한다면 더욱 신뢰도 높은 고객 세분화를 할 수 있게 된다.

이러한 가중치 선택은 RFM 모형 설계에서 가장 중요한 과정이지만 현재는 명확한 기준이 없는 실정이다. 따라서 마케팅 담당자나 기존 연구에서 제안하는 방식에 의존하여 직관적으로 선택되고 있으며 이러한 가중치 선택 방법에는 여러 가지가 있지만 크게 3개로 구분할 수 있다.

1) 파레토 법칙에 의한 결정

파레토 법칙이란 사회 현상의 80%는 20%로 인해 발생한다는 경험법칙으로서 '상위 20%에 속하는 고객이 전체 매출의 80%를 차지한다.'는 이론으로 2대 8 법칙 또는 2080법칙이라고도 한다. 따라서 가중치를 선정할 때 상위 20%의 고객이 갖는 평균값을 활용하면 전체 고객 구매 데이터에 영향을 미칠 수 있다는 데에서 착안한 방법이다.

이소영(2004)은 CRM의 마케팅 전략 중 하나인 고객 세분화의 중요성을 강조하면서 부산 소재 D의료원의 외래 데이터를 활용해 최적의 RFM 모형을 구축하고자 최적의 가중치 선정 방법으로 파레토 법칙을 활용하였으며 다음과 같은 4단계 절차를 제시했다.

1단계	모든 경우의 수에 대한 가중치 목록 제시
↓	
2단계	RFM 점수의 상위 20%에 해당하는 고객 평균 산출
↓	
3단계	전체 자료 중에서 R, F, M 각 요소에 대한 상위 20% 평균값 산출
↓	
4단계	모든 가능한 임의의 가중치에 대해 2단계와 3단계에 해당하는 평균값과 비교하여 유사한 수준에서 최적의 가중치로 결정

<그림 2-6> 파레토 법칙을 활용한 가중치 선택 절차 (이소영, 2004)

즉, 가중치(W1, W2, W3)들의 합이 1이 되는 소수점 첫째 자리의 모든 경우의 수를 구하고 RFM 점수에서 상위 20% 고객들의 평균값과 전체 자료의 R, F, M의 상위 20%의 고객들의 평균값을 비교하여 가장 유사한 값이 나오는 구간의 경우의 수를 최적의 가중치로 결정하는 것인데 결과적으로 W1=0.2, W2=0.4, W3=0.4 로 선택하여 최종 RFM 모형을 제시하였고 분산분석을 통해 해당 모형에 의해 세분화된 그룹들이 적절하게 분리되었음을 증명하였다.

안요찬(2010)은 한방병원의 진료기록 정보를 기반으로 하는 체계적인 CRM 전략과 방법론 개발의 필요성을 지적하며 통합의료정보시스템의 외래 환자 데이터를 활용하여 고객 세분화를 실시하고 고객관계관리 시스템 구축 모형을 제안하였는데 이때 사용한 RFM 모형의 최적 가중치로 파레토 법칙을 활용하여 W1=0.2, W2=0.2, W3=0.6을 결정하였다.

이강태(2002) 또한 같은 방법으로 RFM 모형의 가중치를 W1=0.2, W2=0.1, W3=0.7로 결정하여 쇼핑몰 고객 자료에 대한 RFM 분석 결과를 통해 고객 가치 등급화를 실시하고 우수고객의 예측과 분류, 개별 고객의 수익률에 대해 살펴보고 최적의 수익모델을 제시하였다.

2) 로지스틱 회귀분석에 의한 결정

로지스틱 회귀분석이란 목표변수가 입력 변수들에 의해 어떻게 예측되는지 알아보기 위해 함수식으로 자료를 표현하여 분석하는 통계 방법이다.(최희경, 2002) 입력변수는 범주형 또는 연속형의 값을 갖지만 목표변수는 이분형의 값을 가지기 때문에 성공/실패, 합격/불합격, 구매함/구매 안 함과 같은 이분 변수가 포함된 데이터가 있는 경우 이 방법이 사용된다. 예를 들어 과거 1년간의 고객 구매 데이터가 있을 때 이를 앞의 8개월과 뒤의 4개월의 데이터로 분할한 후 앞의 8개월 데이터에서는 R, F, M값을 관측하고 뒤의 4개월 데이터에서는 해당 기간 동안의 구매 여부를 1(구매함)과 0(구매 안 함)으로 코딩하여 이분형 자료를 만든다. 이렇게 두 개로 나뉜 데이터를 비교하여 예측되는 값을 토대로 가중치로 선택할 수 있다.(Biokorea 이용구, 2007)

김동훈(2008)은 기업의 고객 관리의 중요성을 강조하며 고객이 이탈하지 않고 지속적으로 유지될 수 있는 방안으로 RFM 모형을 제시하면서 RFM 모형 설계의 핵심 문제는 가중치 부여에 있다고 하였으며, 기존 업계에서 사용되고 있는 직관적이고 관습적인 가중치를 통계적 기법으로 대신하고 각 집단 간의 유의성을 검증하여 그 특성을 비교하였다. 목표변수를 최근 3개월 이내의 구매 유무로 설정하고 해당 기간 내에 구매한 고객이 R, F, M 중 어떤 요인에 의해 구매했는지에 대한 연관성을 알아보고 로지스틱 회귀모형식에 의해 산출된 각 변수의 계수를 토대로 W1=0.440, W2=1.031, W3=0.379로 결정하였다.

RFM모형의 변형 모델인 TRFM 기법을 사용한 동상옥(2008)과 CRFM을 사용한 이영진(2010) 또한 가중치를 로지스틱 회귀분석을 통해 각각 W1=0.13, W2=0.01, W3=0.61, W4=0.25과 W1=0.10, W2=0.53, W3=0.14, W4=0.23의 값을 얻어냈으며 해당 가중치를 더한 가중합을 5로 나누고 100을 곱하여 최종 점수를 산출하였다.

3) 직관에 의한 결정

담당자의 직관으로 RFM 모형의 가중치를 결정하는 것은 복잡한 통계 기법이 들어가지 않아도 간편하게 적용할 수 있는 장점이 있다.

류귀열(2006)은 RFM 분석을 통해 전체 고객을 5개로 나누고 연관규칙과 의사

결정 나무를 통하여 각 집단별 유의한 변수들의 패턴을 찾아냄으로써 연관규칙과 의사결정나무의 비교분석과 동시에 이론적으로는 설명하기 힘든 복잡한 집단의 특성들에 대해 효과적으로 파악하는 방법을 제시하였다. 이때 활용되는 데이터의 보유 기관 특수성을 고려하여 가중치를 각각 W1=0.1, W2=0.2, W3=0.7로 선택하여 RFM 모형을 완성했다. 이는 직관적으로 해당 기관이 구매금액을 가장 중요하게 생각한다는 것을 알 수 있다.

정윤필(2009)은 시간의 흐름에 따른 구매 행동 및 고객등급의 변화를 파악할 수 있는 두 기간 기반의 고객 세분화 방법을 제시하고자 RFM 모형을 설계하였는데 이때 마케팅팀과의 토의를 거쳐 W1=0.3, W2=0.2, W3=0.5로 가중치를 부여했다. 이는 가장 중요하다고 생각하는 하나의 변수에 0.7을 부여한 연구와는 다르게 0.5를 부여함으로써 비교적 고르게 부여했다는 것을 직관적으로 확인할 수 있었다.

이 밖에도 이윤성(2010)은 일반적으로 널리 사용하는 로지스틱 회귀분석을 사용하려고 하였으나 목표 변수가 없는 데이터 특성을 고려하여 빈도에 점수를 고려하는 방법을 통해 가중치를 산출하였으며 김태진(2020)은 각 집단 별로 M값에 미치는 R과 F의 영향 값을 도출하여 피해 금액과의 상관관계를 분석한 계수를 바탕으로 가중치를 산출하였다. 류귀열(2013)은 한국과학기술정보연구원에서 제안한 기준을 바탕으로 부합되는 가중치 목록을 정하고 여러 개의 RFM 모형을 설계한 뒤 가중치 변화에 따른 RFM 분포 형태를 비교하여 각 기업의 목적에 따라 가중치를 결정할 수 있다는 결론을 제시하였다.

이렇듯 RFM 모형의 가중치 선택은 담당자나 데이터의 특성에 따라 매우 다양하게 결정될 수 있으며 RFM 모형의 고도화 연구 또한 지속적으로 이어지고 있다. 이렇게 다양한 분야의 선행연구에 널리 쓰이면서 변형된 RFM 모형이 제시되기도 하였는데 이는 기존 R, F, M 변수 외에 중요하다고 생각하는 변수를 추가하거나 대치하여 만드는 모형으로서 캠페인 반응 비율(Campaign)을 추가한 CRFM 모형, 월평균구매시기(Time)가 추가된 TRFM 모형, 광고 상품 총 구매기간(Period)이 추가된 RFMP모형 (이영진, 2010; 동상옥, 2008; 조광현, 2012)등이

있다.

또한 최근에는 R, F, M 세 요소의 가중합으로 환산된 점수에만 의존하는 등급뿐만 아니라 추가적인 세부 속성을 부여하는 연구가 등장했는데 이는 M 점수는 고려하지 않거나 참고용으로만 활용하고 R과 F 점수를 기반으로 고객을 분류하는 것이다.

B. Siva Jyothi(2020)은 R와 F의 점수를 기준으로 Champions, Loyal customers, Potential Loyalist, Recent Customers, Promising, Customers needing Attention, About to sleep, At Risk, Can't lose them, Hibernating, Lost로 총 11등급으로 구분하였고 Norma Ningsih(2020)은 여기에서 Lost를 뺀 10등급으로 세분화하였으며 Dr. Shailendra Kumar Srivastava(2019)는 Champions, Loyal Customer, Promising, Need Attention, About to sleep, At risk of competitors, Hibernating 등 총 7가지 속성으로 추가적인 세분화를 하였다.

<그림 2-7> R과 F점수 만을 활용하여 추가적인 세부 속성 구분

(출처: Guillaume Martin' github, 2018)

<그림 2-7>의 예시를 보면 R과 F의 점수가 모두 1~2점일 경우 Hibernating으

로, 모두 5점인 경우 Champions, F 점수는 높지만 R 점수가 낮은 경우는 Can't loose them, F 점수는 낮지만 R 점수는 높은 경우에는 New Customers로 구분하였으며 이는 RFM 변수의 가중합을 100%로 환산하여 20%씩 5등급으로 나누는 것과 함께 시너지 효과를 기대할 수 있고 더욱 효율적인 마케팅 전략을 제시하기 위한 수단으로 활용할 수 있다. 이는 1995년에 정의된 RFM 기법이 오늘날까지 많은 연구에 의해 끊임없이 발전하고 있다는 것을 보여준다.

선행연구에서 살펴보았듯이 RFM 분석은 RFM 모형을 설계하고 적절한 가중치를 선택하여 전체 고객을 대상으로 점수를 매기고 등급을 부여하여 각 집단별로 특성을 구분하여 타깃 마케팅에 활용하고자 하는 것을 알 수 있었다.

본 논문에서는 RFM 모형에 적절한 가중치를 선택하는 연구를 하고자 한다.

제 3 장 연구 방법

제 1절 조사 설계

1. 분석 대상

본 연구에서는 RFM 모형의 비교를 위해 고객 정보가 포함된 총 3가지의 구매 데이터셋을 활용하였다. 각 데이터셋은 기본적으로 고객을 고유하게 식별할 수 있는 고객 아이디와 더불어 RFM 모형 설정에 필요한 세 가지 요소인 구매일자(Recency), 구매빈도(Frequency), 구매금액(Monetary)을 포함한다.

(1) 데이터셋 A

데이터셋 A는 2000년 3월 10일부터 2001년 3월 19일까지 통신사에서 발생한 6,116건에 대한 고객 구매 자료이며 각 고객에 대한 다양한 인구통계학적 변수를 포함한다.

(2) 데이터셋 B

데이터셋 B는 UCI Repository에서 제공한 Online Retail Data Set으로 Kaggle(예측모델 및 분석 대회 플랫폼)에서 고객 세분화 연구에 가장 많이 활용되는 데이터셋이다. 이 데이터는 2010년 12월 1일부터 2011년 12월 9일까지 약 1년간 거래된 549,019건의 판매 데이터를 포함한다.

(3) 데이터셋 C

데이터셋 C는 미국의 온라인 CD 판매 회사인 CDNOW에서 제공한 데이터셋으로 Fader and Hardie(2001)이 연구에 활용했던 1997년 1월 1일부터 1998년 6월 30일까지의 전체 데이터셋을 1/10로 구성한 체계적인 샘플 자료로서 총 6,919

개의 자료를 포함한다.

<표 3-1> ~ <표 3-3>은 데이터셋 A, B, C의 각 분석변수의 설명이다.

<표 3-1> 데이터셋 A - 분석변수 설명

변수	Label	설명
ID	회원번호	고객 고유 식별 코드
GENDER	성별	1: 남자, 2: 여자
AGE	연령대	10대, 20대 초반, 20대 후반, 30대 초반, 30대 후반, 40대, 50대 이상
WEDDING	결혼여부	미혼, 기혼
AREA	거주지역	강원, 경기, 경남, 경북, 광주, 대구, 대전, 부산, 서울, 울산, 인천, 전남, 전북, 제주, 충남, 충북
COSTMETHOD	결제유형	현금, 신용카드
JOB	직업	회사원, 자영업, 대학(원)생, 학생(초/중/고), 주부, 공무원, 교직자, 전문직, 의료인, 법조인, 군인, 종교인, 방송인, 농/축산업, 일용직, 기타
SELLING	구매금액	해당 구매에 대한 총 금액
AMOUNT	구매수량	해당 구매에 대한 총 구매 개수
RECECY	구매시기	상품 구매 시기를 월로 표시 (1: 1개월, 2: 2개월, 3: 3개월 등)
CNT	구매횟수	해당 고객에 대한 상품 구매 횟수

<표 3-2> 데이터셋 B - 분석변수 설명

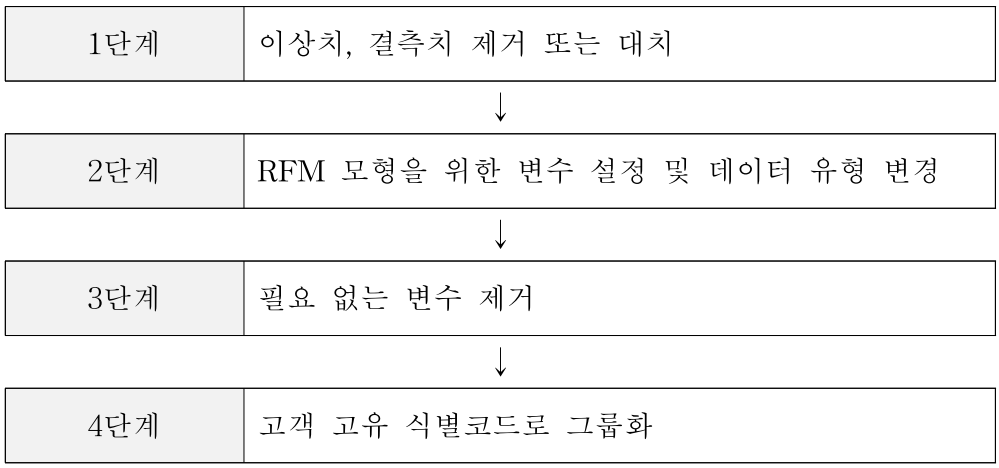
변수	Label	설명
InvoiceNo	영수증번호	6자리 정수로 되어있으며 취소 건은 'C'를 포함
StockCode	상품코드	5자리 정수로 구성
Description	상품명	상품명 상세 정보
Quantity	상품개수	해당 거래에 포함되는 총 상품 개수
InvoiceDate	구매일자	거래 발생 일시 (MM/DD/YYYY HH:MM)
UnitPrice	상품가격	단위: 파운드 (GBP)
CustomerID	회원아이디	회원 고유 식별 코드
Country	지역	구매 국가 (총 38개국)

<표 3-3> 데이터셋 C - 분석변수 설명

변수	Label	설명
V1	마스터번호	마스터 데이터셋과 연결된 번호
V2	고객번호	고객 고유 식별 코드
V3	구매일자	거래 발생 일시 (YYYYMMDD)
V4	구매개수	해당 거래에 포함되는 총 상품 개수
V5	구매금액	단위: 달러 (USD)

2. 데이터 전처리

실제 데이터는 결측치나 이상치 등의 잡음이 존재하며 불완전하므로 곧바로 연구에 활용하기에 다소 어려움이 있다. 따라서 본 연구에서는 <그림 3-1>과 같은 데이터 전처리 과정을 거쳤다.



<그림 3-1> 데이터 전처리 과정

본 연구에 사용된 각 데이터셋은 가장 먼저 결측치 및 이상치를 확인하여 완전히 제거하거나 다른 값으로 대치하였는데 예를 들면 데이터셋A에서 거래 취소인 데이터의 경우에는 상품개수와 상품가격이 결측값으로 확인되었기 때문에 해당 건은 모두 제거하였다. 이런 보편적인 전처리 과정 외에도 각 데이터 별로 구매 정보를 포함한 변수들은 RFM 모형 설계를 위해 R, F, M의 값으로 처리해주었는데 데이터셋B와 데이터셋C의 구매시기 변수는 각각 MM/DD/YYYY HH:MM, YYYY/MM/DD의 'DATE' 형태였기 때문에 기준이 되는 임의의 날짜를 정하고 그 차를 이용하여 정수형으로 변경하였다. 또한 본 연구에 무의미하거나 RFM 모형 설계에 있어서 제한이 되는 변수인 데이터셋A의 결제유형, 데이터셋B의 지역, 데이터셋C의 마스터번호 변수는 제거하였다. 마지막으로 거래 발생 건을 기준으로 되어있는 각 데이터셋을 고객고유식별코드를 기준으로 그룹화하여 최종적으로 데이터셋A 4,470명, 데이터셋B 4,338명, 데이터셋C 2,347명의 고객 구매 데이터로 각 데이터셋을 정리하였다.

3. 분석 방법

준비된 3개의 데이터셋은 통계 및 시각화 언어 R의 IDE인 RStudio를 통해 데이터 전처리 과정을 거쳤다. 이후 Python의 기계학습 라이브러리인 Scikit-Learn 패키지의 Elbow Method 및 K-Means Clustering 알고리즘을 통해 각 데이터셋을 적절한 K개의 군집으로 구분하였다. 이후 각 군집 별 R, F, M값의 평균 및 표준편차를 이용해 CV (Coefficient of Variation, 변동계수)를 구하고 그 비율을 활용하여 가중치를 산출하였다. 이후 최종 RFM 모형을 설계하고 각 고객에게 점수를 부여한 뒤 전체 고객을 5개 등급으로 세분화하였고 여러 가지 방법으로 가중치를 부여한 뒤 각 RFM 모형별 세분화된 고객 집단을 비교하였다.

제 2절 RFM 모형 설계

본 연구에서는 RFM 분석 시 일반적으로 널리 사용되는 Hughes(2000)가 소개하는 보편적인 방법의 RFM 모형을 활용하고자 각 데이터셋의 R, F, M 변수들을 5개의 범주로 구분하여 총 125(5*5*5)개의 집단으로 세분화한다. 단, 모든 범주에 대해서 20%씩 균등하게 구분하는 것은 아니고 각 데이터셋의 특성에 따라 유동적으로 설계한다. 이렇게 생성된 집단은 마케팅에 그대로 활용될 수도 있지만 일반적으로 분석의 용이함을 위해 각 집단에 RFM 가중치를 부여한 점수를 매겨 등급을 부여함으로써 더 적은 그룹으로 분할한다.(이소영, 2004) 각 항목의 자세한 설정은 다음과 같다.

1. R값 설정

R은 최근성을 나타내는 변수로써 '고객이 얼마나 가장 최근에 구매했는가?' 에 대한 지표를 나타낸다. 데이터셋A의 경우 고객 최근 방문일이 월 단위로 나누어져 있어 실질적으로 20%씩 구분하는 것은 어려웠다. 따라서 최대한 균등하게 분

배하되 데이터의 수치가 중앙 즉, 3점으로 몰리도록 범위를 설정하였다. 데이터셋B와 데이터셋C는 전체 데이터 기간을 20%씩 균등하게 나누어 구분하였다. 데이터셋 A, B, C의 R값에 대한 범위 및 분포는 <표 3-4> ~ <표 3-6>과 같다.

<표 3-4> 데이터셋 A - 구매시기에 의한 R값의 범위 및 분포

범위 (단위: 개월)	R 값	빈도	비율
1개월	5	395	8.84%
2개월 ~ 3개월	4	1243	27.74%
4개월 ~ 5개월	3	1166	26.85%
6개월 ~ 7개월	2	1115	13.09%
8개월 이상	1	551	9.78%

<표 3-5> 데이터셋 B - 구매시기에 의한 R값의 범위 및 분포

범위 (단위: 일)	R 값	빈도	비율
0일 ~ 12일	5	868	20.01%
14일 ~ 32일	4	880	20.29%
33일 ~ 71일	3	863	19.89%
72일 ~ 179일	2	866	19.96%
180일 이상	1	861	19.85%

<표 3-6> 데이터셋 C - 구매시기에 의한 R값의 범위 및 분포

범위 (단위: 일)	R 값	빈도	비율
1일 ~ 152일	5	472	20.11%
153일 ~ 441일	4	467	19.90%
442일 ~ 486일	3	478	20.37%
487일 ~ 512일	2	475	20.24%
513일 이상	1	455	19.39%

2. F값 설정

F는 구매횟수를 나타내는 변수로서 '고객이 얼마나 자주 구매했는가?'에 대한 지표를 나타낸다. 3개의 데이터셋 모두 1회만 구매한 경우가 전체 데이터의 53.15%, 34.42%, 51.12%씩 크게 차지하기 때문에 20%씩 균등한 배분은 할 수 없고 정보의 손실을 최소화하고자 1회, 2회, 3회, 4회, 5회 이상으로 설정하였다. 이는 데이터의 특성이 다름에도 비슷하게 나타나는 현상으로 고객 구매 데이터의 특징으로 볼 수 있다. 데이터셋 A, B, C의 F값에 대한 범위 및 분포는 <표 3-7> ~ <표 3-9>와 같다.

<표 3-7> 데이터셋 A - 구매횟수에 의한 F값의 범위 및 분포

범위 (단위: 회)	F 값	빈도	비율
5회 이상	5	299	6.69%
4회	4	216	4.83%
3회	3	336	7.52%
2회	2	1243	27.81%
1회	1	2376	53.15%

<표 3-8> 데이터셋 B - 구매횟수에 의한 F값의 범위 및 분포

범위 (단위: 회)	F 값	빈도	비율
5회 이상	5	1114	25.68%
4회	4	388	8.94%
3회	3	508	11.71%
2회	2	835	19.25%
1회	1	1493	34.42%

<표 3-9> 데이터셋 C - 구매횟수에 의한 F값의 범위 및 분포

범위 (단위: 회)	F 값	빈도	비율
5회 이상	5	388	16.53%
4회	4	150	6.39%
3회	3	207	8.82%
2회	2	405	17.26%
1회	1	1197	51.00%

3. M값 설정

M은 총 구매액을 나타내는 변수로써 '고객이 얼마나 많은 금액을 사용했는가?'에 대한 지표를 나타낸다. 데이터셋 A, B, C 모두 20%씩 균등한 비율로 분할하였다. 데이터셋 A, B, C의 M값에 대한 범위 및 분포는 <표 3-10> ~ <표 3-12>와 같다.

<표 3-10> 데이터셋 A - 구매금액에 의한 M값의 범위 및 분포

범위 (단위: 원(KRW))	M 값	빈도	비율
129,400 이상	5	893	19.98%
57,400 ~ 129,000	4	873	19.53%
39,200 ~ 57,000	3	877	19.62%
25,200 ~ 39,000	2	861	19.26%
2,000 ~ 25,000	1	966	21.61%

<표 3-11> 데이터셋 B - 구매금액에 의한 M값의 범위 및 분포

범위 (단위: 파운드(GBP))	M 값	빈도	비율
2059.4 이상	5	868	20.01%
942.3 ~ 2056.9	4	867	19.99%
490.2 ~ 942.3	3	868	20.01%
250.3 ~ 489.6	2	867	19.99%
3.8 ~ 250.2	1	868	20.01%

<표 3-12> 데이터셋 C - 구매금액에 의한 M값의 범위 및 분포

범위 (단위: 달러(USD))	M 값	빈도	비율
130.5 이상	5	470	20.03%
59.5 ~ 130.5	4	469	19.98%
32.3 ~ 59.3	3	469	19.98%
15.5 ~ 32.3	2	466	19.86%
4 ~ 15.4	1	473	20.15%

이렇게 각 데이터셋 별로 5개로 구분된 범위를 기준으로 R, F, M 변수들을 조합하여 R (5개) * F (5개) * M(5개) = 125개의 집단으로 세분화하였다. 예를 들어 어떠한 데이터의 RFM 조합이 (235) 인 경우 R값 = 2, F값 = 3, M값 = 5를

의미하는 것이며 (555) 조합이 가장 수익성이 큰 집단, (111) 조합이 가장 수익성이 낮은 집단으로 볼 수 있다. 단 여기에서 125개 모든 집단에 반드시 고객 데이터가 있는 것은 아니며 이는 각 데이터셋 별로 다르게 나타난다.

제 3절 데이터 군집화

본 연구에서는 RFM 모형의 가중치를 선택하기 위한 방법으로 K-Means Clustering 알고리즘을 활용하여 전체 고객 데이터를 K개의 그룹으로 군집화하며 이 과정은 크게 <그림 3-2>와 같다.

1단계	정제된 데이터 준비
↓	
2단계	R, F, M 변수에 대한 데이터 표준화
↓	
3단계	Elbow Method 방법을 활용한 최적의 K값 탐색
↓	
4단계	K개로 데이터 군집화

<그림 3-2> K-Means Clustering의 과정

앞의 RFM 모형을 설계하면서 전처리 과정을 마친 데이터셋을 준비하고 각 데이터셋 별로 가지고 있는 R, F, M 변수에 대해 데이터 표준화를 진행한다. 표준화는 데이터의 평균을 0, 분산 및 표준편차를 1로 만들어주는 데이터 전처리 과정의 하나로 데이터의 값이 너무 크거나 작은 경우 알고리즘 학습 과정에서 0에 수렴하거나 무한으로 발산하는 것을 방지하고 최적화 과정에서 안정성 및 수렴 속도를 향상시킨다. 또한 단위가 다른 여러 가지 변수를 함께 비교하는 경우 표

준화하여 더 큰 값을 가진 결과에 더 큰 영향을 미치는 것을 방지하기 위해 사용하게 된다.(Dev log, 2010)

<그림 3-3>은 데이터셋A의 표준화 이전 데이터 예시이다. 일반적으로 여러 가지 변수가 있는 데이터셋의 경우 표준화를 진행하는데, R, F, M 세 요소만 있는 데이터라도 그 단위가 모두 다르기 때문에 표준화 과정이 필요하며 표준화를 하지 않을 경우에는 잘못된 결과를 도출할 수 있다. 따라서 Python의 MinMax Scaler 함수를 활용하여 각 데이터셋에 대해 표준화 과정을 거치고 R, F, M 변수에 대한 데이터를 0부터 1사이의 값으로 변환 시킨다. <그림 3-4>는 <그림 3-3>의 표준화 결과를 보여준다.

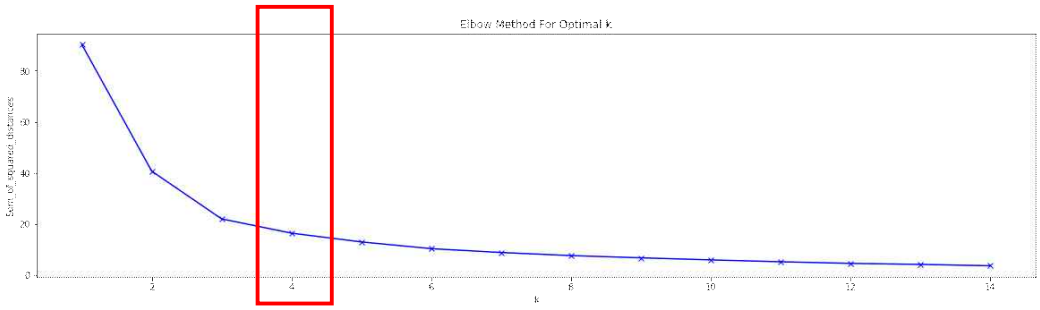
	recency	frequency	monetary
0	4	1	68000.0
1	1	1	408000.0
2	1	2	44000.0
3	2	1	7800.0
4	2	2	35000.0

<그림 3-3> 데이터셋A의 표준화 이전 데이터

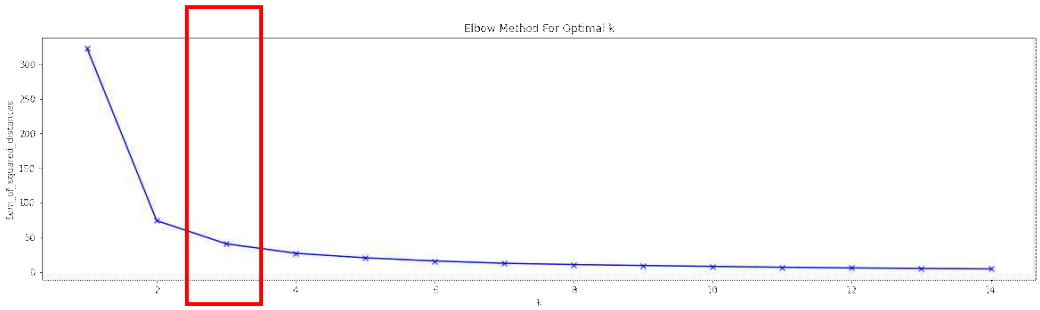
	recency	frequency	monetary
0	0.115385	0.000000	0.000987
1	0.000000	0.000000	0.006074
2	0.000000	0.014493	0.000628
3	0.038462	0.000000	0.000087
4	0.038462	0.014493	0.000494

<그림 3-4> 데이터셋A의 표준화 이후 데이터

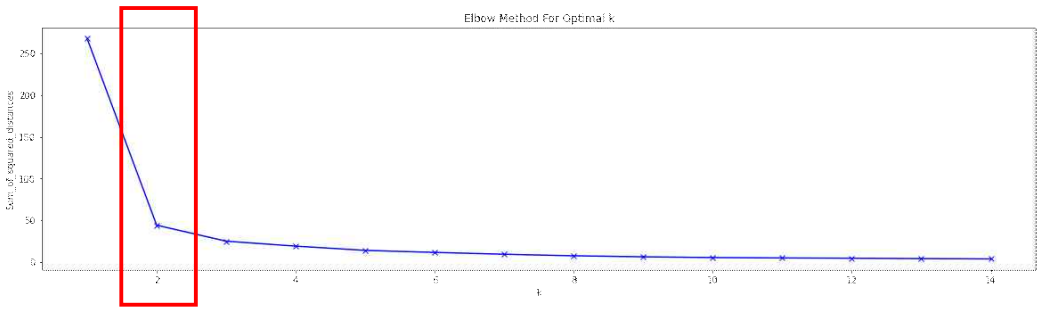
각 데이터셋의 표준화가 끝나면 Python의 Scikit-Learn 패키지의 Elbow Method 알고리즘을 활용하여 적절한 K값을 얻는다. 2장 2절에서 설명한 바와 같이 Elbow Method는 SSE값의 크기가 최소화되는 시점, 즉 기울기가 급격히 완만해지는 시점으로 K값을 찾을 수 있는데 <그림 3-5> ~ <그림 3-7>은 각 데이터셋에 Elbow Method 알고리즘을 적용하여 그래프로 도출한 것이다.



<그림 3-5> 데이터셋 A의 Elbow Method 그래프

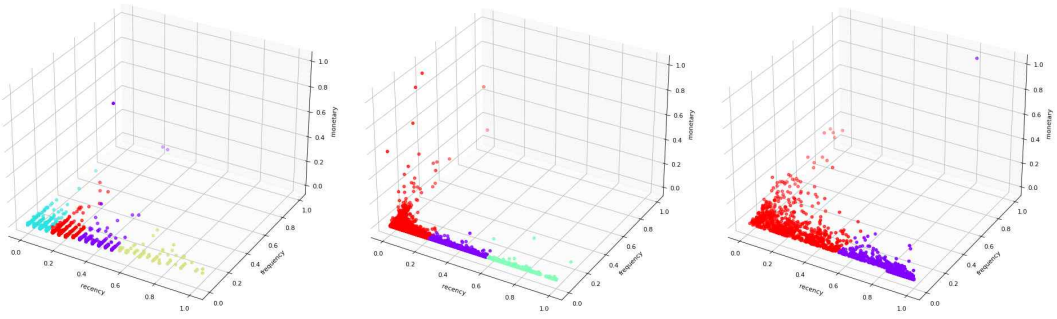


<그림 3-6> 데이터셋 B의 Elbow Method 그래프



<그림 3-7> 데이터셋 C의 Elbow Method 그래프

데이터셋A는 <그림 3-5>과 같이 K값이 4인 경우에 기울기가 완만해지는 것으로 확인되었고, 데이터셋B는 <그림 3-6> K가 2까지 급격히 떨어지다가 3부터 서서히 완만해지는 것으로 확인되었으며 데이터셋C는 <그림 3-7> 최적의 K값이 2로 확인하였다. 최종적으로 데이터셋A, B, C는 <그림 3-8>과 같이 각각 4개, 3개, 2개의 그룹으로 분류되었다.



<그림 3-8> 데이터셋 A, B, C의 K개 그룹 분류

K-Means Clustering을 통해 분류된 집단은 다음과 같이 정리할 수 있다. 데이터셋A는 K값을 4로 얻어 총 4개의 그룹으로 분류되었는데 각 그룹의 R값의 평균은 그룹2가 2.65로 가장 낮았으며 그룹3이 18.50으로 가장 높았다. F값의 최고 평균은 그룹3이 4.87로 가장 높았으며 그룹2와 그룹4는 각각 1.81, 1.96으로 근소한 차이로 그룹 4가 더 높은 것으로 나타났다. M값의 평균 또한 그룹2와 그룹 4는 비슷하게 나타났으며 그룹1과 그룹 3은 또렷한 차이를 보였다.

데이터셋B는 총 3개의 집단으로 분류하였는데 R, F, M 값 모두 확연한 차이를 두고 있다. 직관적으로 판단했을 때 그룹3이 가장 수익성이 낮은 고객의 집단이며 그룹 2가 수익성이 가장 높은 그룹이다.

데이터셋C는 2개의 집단으로만 분류하였는데 그룹 1과 2의 차이가 평균과 비교하여 매우 큰 차이를 보였다.

<표 3-13> ~ <표 3-15>는 각 데이터셋 A, B, C의 K개 그룹별 R, F, M의 기술통계량 및 빈도를 나타낸다.

<표 3-13> 데이터셋 A - K개 그룹별 R, F, M의 기술통계량 및 빈도

구분	평균 (X)			표준편차 (S)			빈도
	R	F	M	R	F	M	
전체	5.06	2.09	144,501	3.53	2.68	1,088,116	4,470
그룹1	11.25	3.73	475,902	1.63	6.01	3,945,569	315
그룹2	2.65	1.81	110,191	1.06	1.04	257,495	2,248
그룹3	18.50	4.87	247,456	3.10	2.84	432,141	125
그룹4	6.08	1.96	121,983	0.91	2.29	327,886	1,782

<표 3-14> 데이터셋 B - K개 그룹별 R, F, M의 기술통계량 및 빈도

구분	평균 (X)			표준편차 (S)			빈도
	R	F	M	R	F	M	
전체	92.06	4.27	2,054	100.01	7.70	8,989	4,338
그룹1	153.53	2.22	800	38.10	1.65	1,677	798
그룹2	31.86	5.46	2,709	25.27	9.09	10,746	2,914
그룹3	293.94	1.35	607	45.11	1.53	3,585	626

<표 3-15> 데이터셋 C - K개 그룹별 R, F, M의 기술통계량 및 빈도

구분	평균 (X)			표준편차 (S)			빈도
	R	F	M	R	F	M	
전체	371.39	2.94	104	178.10	4.18	217	2,347
그룹1	482.33	1.48	51	50.74	1.68	174	1,624
그룹2	122.20	6.23	223	86.12	5.90	254	723

제 4절 RFM 모형의 가중치 산출

본 연구의 핵심 목적인 RFM 모형의 가중치를 선택하기 위한 방법으로 앞서 K-Means Clustering으로 분류된 각 그룹에 대한 기술통계량 값을 활용한다.
고객 i에 대한 RFM 모형 (식 3-1)에서 W1은 R 항목의 가중치를, W2는 F 항목의 가중치, W3은 M 항목의 가중치를 나타내며 이때 W1, W2, W3의 합은 1이 된다.

$$RFM_i = W1 \times Recency + W2 \times Frequency + W3 \times Monetary \quad : (식\ 3\text{-}1)$$
$$(W1 + W2 + W3 = 1)$$

본 연구에서는 위 모형의 가중치인 W1, W2, W3을 찾기 위해 K개로 분류된 그룹1, 그룹2, …, 그룹K에 대하여 각각 R, F, M 값의 CV(Coefficient of variation, 변동계수)를 산출하고 전체 집단과의 비율을 활용하고자 하였으며 식으로 표현하면 다음(식 3-2)과 같다.

$$w1 = \frac{\min(CV_{rn})}{CV_{r1} + CV_{r2} + ... + CV_{rk}} \quad (CV_{rk} = \frac{s_{rk}}{x_{rk}})$$

$$w2 = \frac{\min(CV_{fn})}{CV_{f1} + CV_{f2} + ... + CV_{fk}} \quad (CV_{fk} = \frac{s_{fk}}{x_{fk}})$$

$$w3 = \frac{\min(CV_{mn})}{CV_{m1} + CV_{m2} + ... + CV_{mk}} \quad (CV_{mk} = \frac{s_{mk}}{x_{mk}})$$

$$W1 = \frac{w1}{w1+w2+w3},\ W2 = \frac{w2}{w1+w2+w3},\ W3 = \frac{w3}{w1+w2+w3}$$

$$(W1 + W2 + W3 = 1)$$

$x_{r,f,m}$: R, F, M에 대한 각각의 평균

$s_{r,f,m}$: R, F, M에 대한 각각의 표준편차

$_k$: $K-Means\ Clustering$으로 나눠진 그룹 수

$CV_{rfm \cdot k}$: k개 그룹에 대한 각 R, F, M의 CV

$\min(CV_{rfm \cdot n})$: R, F, M의 CV 최솟값

(식 3-2)

K개로 분류된 각 그룹별로 R, F, M의 표준편차를 평균으로 나누어 CV를 산출하고 CV를 최소로하는 CVrn, CVfn, CVmn을 찾아 전체 CV의 합으로 다시 나누어 w1, w2, w3를 구하고 모든 가중치의 합을 1로 만들기 위해 w1, w2, w3 각각을 3개를 모두 더한 값으로 나누어 W1, W2, W3를 구하여 최종 가중치로 결정하였다.

이때 CV의 최솟값을 활용한 이유는 CV가 나타내는 특성 때문인데 CV는 표

준편차의 크기를 평균으로 나눈 것으로 크기나 단위가 다른 분포 간의 변동을 비교할 때 사용한다. 이때 CV값이 작다는 것은 해당 변수가 평균으로부터 변동성이 작다는 것을 의미하며 본 연구의 목적은 각 K개의 집단별 변수의 특성을 가중치에 반영하고자 하는 것이므로 각 변수의 평균으로부터 변동성이 가장 작은 수치를 찾아내기 위해 CV의 최솟값을 활용하였다. <표 3-16>은 각 데이터셋에 최종 산출된 가중치를 나타낸다.

<표 3-16> 데이터셋 A, B, C에 대한 가중치

	W1	W2	W3
데이터셋 A	0.4039	0.3168	0.2793
데이터셋 B	0.2498	0.4095	0.3407
데이터셋 C	0.1558	0.5447	0.2995

<표 3-16>과 같이 각 데이터셋의 K개 집단으로부터 얻어낸 가중치를 적용하여 (식 3-3)과 같이 최종 RFM 모형을 설계하였다.

$$\text{데이터셋}A)\ RFM = 0.4039 \times R + 0.3168 \times F + 0.2793 \times M \quad : (\text{식 } 3\text{-}3)$$

$$\text{데이터셋}B)\ RFM = 0.2498 \times R + 0.4095 \times F + 0.3407 \times M$$

$$\text{데이터셋}C)\ RFM = 0.1558 \times R + 0.5447 \times F + 0.2995 \times M$$

제 4 장 연구 결과

제 1절 고객 세분화

1. RFM 점수 부여

각 고객에게 부여되는 RFM 점수는 R, F, M 세 가지 요소의 조합으로 계산된다. 앞서 결정한 최종 RFM 모형을 (식 4-1)과 같이 백분위로 산출하여 RFM Score를 구한 뒤 고객에게 부여한다. RFM Score는 최저 20점부터 최고 100점까지의 값을 가진다.

$$RFM Score = (W1 \times Recency + W2 \times Frequency + W3 \times Monetary)/5 \times 100 \text{ : (식 4-1)}$$

데이터에 따라 최저점과 최고점의 한곗값은 다를 수 있으며 본 연구에서는 <표 4-1> ~ <표 4-3>과 같이 각 데이터셋의 점수 분포가 이루어졌다. 데이터셋 A <표 4-1>는 각 고객들의 RFM Score 점수가 최저 26.336점부터 최고 100점까지 부여되었으며 총 118개의 점수 범위가 나타났다. 26.336점을 부여받은 17명 고객의 R, F, M 평균은 10.3, 2.0, 18684.1로 나타났고 100점을 부여받은 19명 고객은 1.0, 10.1, 452490.5로 나타났다.

<표 4-1> 데이터셋 A - RFM Score에 따른 RFM 평균 및 빈도

RFM Score	R 평균	F 평균	M 평균	빈도
26.336	10.3	2.0	18684.1	17
28.078	6.5	1.0	15430.9	191
31.922	10.3	2.0	33043.2	44
32.672	13.0	3.0	22500.0	1
33.664	6.4	1.0	32275.8	140
...	...	...	...	...
88.828	1.0	8.0	42300.0	1
91.922	2.6	9.5	640796.8	28
93.664	1.0	4.0	255114.3	7
94.414	1.0	7.8	95602.0	10
100.000	1.0	10.1	452490.5	19

데이터셋B <표 4-2>는 20점부터 시작하여 100점까지의 점수가 부여되었으며 총 118개의 점수 범위가 나타났다.

<표 4-2> 데이터셋 B - RFM Score에 따른 RFM 평균 및 빈도

RFM Score	R 평균	F 평균	M 평균	빈도
20.000	282.5	1.0	146.4	330
24.996	125.6	1.0	142.4	146
26.814	277.5	1.0	349.8	208
28.190	236.6	2.0	165.4	26
29.992	51.8	1.0	161.0	129
...	...	...	...	...
90.008	47.7	8.6	5003.7	101
91.810	6.4	4.0	2923.0	17
93.186	5.4	6.9	1536.2	120
95.004	22.0	10.2	5797.5	198
100.000	5.0	17.6	10813.8	367

데이터셋C <표 4-3>는 20점부터 100점까지 나타났으며 총 73개의 점수 범위를 보였다. 3개의 데이터셋 중에 표본의 수가 가장 적은 만큼 점수의 범위 또한 가장 적게 나타났다. 데이터셋C의 점수 분포에서 특이한 것은 RFM Score가 94.010인 고객 3명의 M값의 평균을 보면 2335.2로 지나치게 높은 것을 볼 수 있는데 이는 customerID값이 1901인 고객의 R, F, M 값이 각각 445, 56, 6552.7로 이상치로 판단할 수 있는 값을 갖기 때문이다. 전체 고객 2,347명에 대해 M값을 기준으로 내림차순 정렬을 했을 때 6552.7의 값을 가진 1901 고객이 1위로 나타났으며 2위의 M값은 1943.58로 무려 3배가 넘는 차이를 보였다. 실제로 1901 고객을 제외하고 RFM Score를 다시 산출했을 때 RFM Score가 94.010인 구간의 R, F, M 평균은 69.8, 5.8, 105.2로 안정화된 값을 보였다. 일반적으로 이러한 이상치(outlier)는 정상값 분석에 안 좋은 영향을 미치기 때문에 정규화 등을 통해 제거하지만 소수의 이상치가 의미 있게 분석되는 경우도 있으며 특히 고객 세분화의 경우에는 VVIP, VVVIP 등을 구분할 때도 활용될 수 있기 때문에 본 연구에서는 이상치 제거 단계를 거치지 않았다.

<표 4-3> 데이터셋 C - RFM Score에 따른 RFM 평균 및 빈도

RFM Score	R 평균	F 평균	M 평균	빈도
20.000	528.2	1.0	12.6	163
23.116	499.6	1.0	12.5	159
25.990	527.8	1.0	23.7	139
26.232	474.4	1.0	12.5	149
29.106	500.2	1.0	23.6	126
...	...	...	...	...
90.894	248.0	5.4	95.8	30
93.768	460.3	24.7	2335.2	3
94.010	68.4	5.8	106.0	34
96.884	250.3	7.4	308.7	80
100.000	57.4	11.2	424.3	234

이렇게 데이터셋 A, B, C 별로 전체 고객을 대상으로 RFM Score를 통해 점수를 부여하였고 다음 단계에서는 고객이 갖는 점수를 기준으로 총 5개의 등급으로 세분화하였다.

2. 5개 등급 구분

앞서 부여된 RFM Score를 기준으로 전체 데이터를 20%씩 동일한 비율의 5개의 등급으로 최상위 그룹부터 Diamond, Platinum, Gold, Silver, Bronze로 세분화한다. 등급을 구분할 때는 데이터 분포에 상관없이 마케팅 담당자나 연구자가 직접 범위를 정할 수도 있는데 이 경우에는 각 등급 분류별 데이터의 분포가 지나치게 큰 차이가 날 수 있으며 그 점수에 대한 기준을 정하는데도 명확하지 않기 때문에 본 연구에서는 각 고객을 전체 고객 수를 기준으로 20%씩 동일한 비율로 세분화하는 방법을 적용하였다.

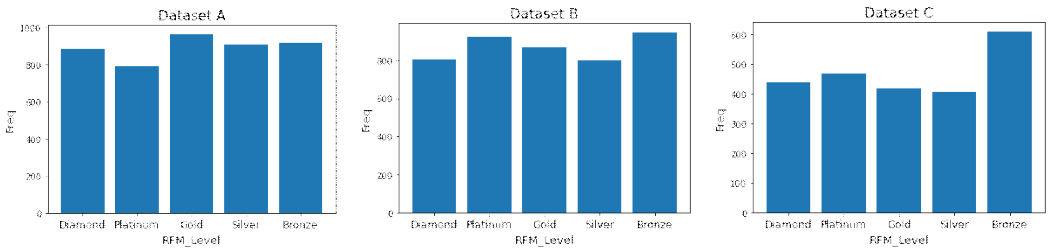
<표 4-4>는 데이터셋A, B, C를 5개 등급의 집단으로 세분화하고 각 집단의 점수 범위를 표로 나타낸 것이다.

<표 4-4> 데이터셋 A, B, C의 등급별 점수 범위

	데이터셋A	데이터셋B	데이터셋C
Diamond	63.2 이상	90 이상	84.9 이상
Platinum	55.8 ~ 63.1	63.6 ~ 88.2	53.8 ~ 83.1
Gold	48.7 ~ 55.4	46.4 ~ 63.6	36.9 ~ 52.2
Silver	40.8 ~ 48.1	33.2 ~ 45.4	29.1 ~ 35.1
Bronze	26.3 ~ 40.0	20.0 ~ 31.8	20.0 ~ 26.2

<표 4-4>와 같이 5개로 세분화된 각 집단은 20점을 시작으로 16점씩 동일한 차이로 나누어지지 않은 것을 확인할 수 있는데 이는 각 데이터셋이 갖는 변수의 특성이나 R, F, M값을 구분하는 범위가 다르기 때문이다. 데이터셋 A는 63.2 이상일 경우 Diamond 집단에 포함되는가 하면 데이터셋 B 90점 이상, 데이터셋

C는 84.9 이상이여야 Diamond 집단에 포함된다.



<그림 4-1> 데이터셋 A, B, C의 등급별 히스토그램

각 집단은 점수의 범위뿐만 아니라 집단을 구성하고 있는 고객의 분포도 다르게 나타났다. <그림 4-1>과 같이 데이터셋 A, B, C의 각 등급별 집단의 분포는 모두 다르게 나타났는데, 데이터셋A는 Gold 집단이 가장 많은 고객으로 구성되었고 Platinum 집단이 가장 적게 구성되었으며 데이터셋B와 C는 Bronze 집단이 가장 많고 Silver 집단이 가장 적은 것으로 나타났다. 이를 통해 각 데이터셋이 가지는 변수의 특성이나 데이터의 크기, R, F, M 세 요소에 대한 점수 범위, 가중치 등에 따라 세분화되는 고객의 집단은 얼마든지 달라질 수 있다는 것을 확인하였다.

이렇게 5개 등급으로 세분화된 집단이 적절히 분류되었는지에 대한 검토를 위해 분산분석(ANOVA, Analysis Of Variance)을 통해 각 집단별 R, F, M값에 대한 검정을 시행하였다. 분산분석은 여러 개의 집단 간 평균의 차이를 통계적으로 유의미한지를 판단하는 검정 방법으로 집단의 수가 2개 이상일 때 사용하며(김태윤, github) 본 연구에서는 Python 패키지를 활용하여 분산분석을 시행하였다.

<표 4-5> 데이터셋 A - 분산분석 결과

	df	sum_sq	mean_sq	F-value	P-value
R	4.0	4626.129552	1156.532388	101.294361	1.132161e-82
F	4.0	4860.902203	1215.225551	198.647142	5.380858e-157
M	4.0	6.072996e+13	1.518249e+13	12.96035	1.709595e-10

<표 4-6> 데이터셋 B - 분산분석 결과

	df	sum_sq	mean_sq	F-value	P-value
R	4.0	1.638832e+07	4.097080e+06	657.692562	0.000
F	4.0	76215.487375	19053.871844	456.660908	0.000
M	4.0	2.814006e+10	7.035016e+09	94.573823	3.069857e-77

<표 4-7> 데이터셋 C - 분산분석 결과

	df	sum_sq	mean_sq	F-value	P-value
R	4.0	5.101481e+07	1.275370e+07	1290.494982	0.000
F	4.0	17556.497711	4389.124428	620.248662	0.000
M	4.0	2.826506e+07	7.066264e+06	501.149155	2.415299e-312

<표 4-5> ~ <표 4-7>과 같이 데이터셋 A, B, C 모두 5개 등급의 집단에 대한 R, F, M 분산분석 결과 P-value값이 0.000 미만으로 이는 99% 신뢰 수준에서 집단 간의 평균 차이가 없다는 귀무가설을 기각함으로써 각 집단 간의 차이는 유의미하다고 할 수 있다.

3. 5개 등급 비교

앞 단계에서 세분화된 5개의 고객 집단별 RFM 점수의 범위를 확인하고 각 집단이 유의미하게 분류되었음을 살펴보았다. 다음 단계로 각 집단별 R, F, M 평균값과 빈도를 확인하고 그 차이점을 비교하였다.

<표 4-8>에서 데이터셋A의 5개 집단 R, F, M 세 요소의 평균을 살펴보면 R, F, M 세 요소의 값은 모두 상위그룹에서 하위그룹으로 갈수록 각 점수가 타당한 것으로 나타났다. Platinum 집단의 빈도가 790, Gold 집단의 빈도가 966으로 다소 차이가 있으나 대체로 적절히 5개로 잘 분류되었다.

<표 4-8> 데이터셋 A - 집단별 R, F, M 평균 및 빈도

	R 평균	F 평균	M 평균	빈도
Diamond	3.6	4.1	347125.5	887
Platinum	4.7	2.0	187129.6	790
Gold	5.1	1.6	134388.8	966
Silver	5.2	1.4	39788.3	908
Bronze	6.7	1.3	26380.4	919

데이터셋 B <표 4-9>의 5개 집단 R, F, M 평균을 살펴보면 R, F, M 모두 잘 구분된 것으로 나타났다. 3개의 데이터셋 중 집단별로 고객이 가장 고르게 분포되었고 각 집단의 R, F, M 평균값 또한 Diamond가 가장 높게 나타났으며 그다음으로 Platinum, Gold, Silver 순으로 순차적으로 낮아지다가 Bronze 집단에서 가장 낮은 것으로 나타났다.

<표 4-9> 데이터셋 B - 집단별 R, F, M 평균 및 빈도

	R 평균	F 평균	M 평균	빈도
Diamond	14.7	12.7	7292.6	803
Platinum	45.8	4.5	1864.0	924
Gold	73.7	2.3	893.9	866
Silver	108.9	1.4	444.3	798
Bronze	205.4	1.0	215.8	947

마지막으로 데이터셋C <표 4-10>를 살펴보면 데이터셋B와 마찬가지로 상위 집단에서 하위집단으로 갈수록 R, F, M 평균치는 점점 낮아지면서 고르게 나타나는 것을 확인하였다. 다만 Silver와 Bronze의 평균치가 1.0으로 동일하게 나왔는데 이는 데이터셋C의 Frequency 변수 즉, 구매빈도가 단 한 번의 구매만 했던 고객 수가 1,197명으로 전체 고객 수 2,347명 중 51%를 차지하는 데이터 특성 때문에 나타난 것으로 확인하였다. 이는 본 논문 2장-3절-2항에서 언급했던 '전체 고객 구매 데이터에서 단 한 번의 주문을 하는 고객 비율은 30~60%를 차지한다'는 John R. Miglausch(2000)의 주장과도 일치하는 내용이다.

<표 4-10> 데이터셋 C - 집단별 R, F, M 평균 및 빈도

	R 평균	F 평균	M 평균	빈도
Diamond	128.9	8.8	343.1	440
Platinum	251.8	2.9	97.9	470
Gold	437.6	1.6	58.7	419
Silver	498.9	1.0	31.6	409
Bronze	507.5	1.0	15.1	610

제 2절 RFM 모형 비교

1. 비교 모형 설계

여러 가지 가중치 선택 방법을 통한 RFM 모형 설계를 하고 데이터셋A, B, C 각각에 적용하여 본 연구와 동일하게 5개 등급 집단으로 세분화하고 그 차이를 확인하였다. 비교에 사용된 모형의 가중치 선택 방법은 파레토 법칙에 의한 결정, 직관에 의한 결정을 활용하였으며 목표변수가 없는 데이터셋A, B, C의 특성에 따라 로지스틱 회귀분석에 의한 결정 방법은 제외하였다.

이렇게 설계된 2개의 모형과 본 연구에서 제안하는 모형 등 총 3가지 모형을 비교하였으며 이때 각 모형별로 세분화된 집단은 R, F, M 평균치를 중심으로 직관적인 차이점을 비교하였다.

(1) 파레토 법칙에 의한 결정

<표 4-11>은 각 데이터셋별로 상위20% 매출과 전체 매출을 비교하여 백분위수로 나타낸 것이다.

<표 4-11> 데이터셋 A, B, C의 전체 매출 대비 상위 20% 매출 비교

	전체 매출	상위 20% 매출	상위 20% 비율
데이터셋A	645,923,625 원	488,118,090 원	75.6%
데이터셋B	8,911,407 파운드	6,649,437.461 파운드	74.6%
데이터셋C	243,675.26 달러	164,001.79 달러	67.3%

데이터셋 A, B, C의 매출 상위 20% 집단이 총 매출에 차지하는 비율은 각각 75.6%, 74.6%, 67.3%로 나타났다. 파레토 법칙이 제시하는 80%에 정확하게 미치진 않았으나 이는 상위 20% 집단의 중요성을 잘 보여주고 있다. 따라서 본 연구에서는 매출 기준의 상위 20% 집단과 가중치를 적용하지 않은 RFM 모형을 통

해 RFM Score를 부여한 상위 20% 집단의 R, F, M 평균값을 비교하여 가중치를 산출하였는데 이는 (식 4-2)와 같다.

$$\text{데이터셋}A)\ RFM = 0.12 \times R + 0.39 \times F + 0.49 \times M \quad : (\text{식 } 4\text{-}2)$$
$$\text{데이터셋}B)\ RFM = 0.12 \times R + 0.44 \times F + 0.45 \times M$$
$$\text{데이터셋}C)\ RFM = 0.20 \times R + 0.39 \times F + 0.41 \times M$$

데이터셋 3개 모두 M의 가중치가 가장 높고, R의 가중치가 가장 낮게 나타났다. 이는 각 데이터셋 A, B, C의 매출 상위 20%의 고객과 가중치를 부여하지 않은 RFM 모형을 통해 점수를 부여한 상위 20% 고객 사이에 R, F, M 평균값을 비교해본 결과 F와 M값은 근소한 차이로 나타났지만 R 값은 비교적 큰 차이를 보였기 때문인데 R의 가중치를 가장 적게 주면서 그 차이에 대한 편차를 줄이고자 함에 있다.

(2) 직관에 의한 결정

직관에 의한 결정은 W1= 0.1, W2=0.2, W3=0.7로 선택하여 최종적으로 (식 4-3)과 같은 모형을 설계하였다.

$$\text{데이터셋}A, B, C)\ RFM = 0.1 \times R + 0.2 \times F + 0.7 \times M \quad : (\text{식 } 4\text{-}3)$$

이는 한국과학기술정보연구원에서 중요도를 M, F, R로 선정하였고(류귀열, 2013) 기업 입장에서 매출에 대한 가중치를 중요하게 생각한다는 가정으로 설계한 모형으로 실제로 선행 연구의 사례에서도 직관에 의한 결정 시 중요도를 높게 측정한 경우가 가장 많았다. (류귀열, 2006; 정윤필, 2009 등)

2. 모형 비교

앞서 설계된 파레토 법칙을 활용한 모형, 직관에 의해 설계된 모형, 본 연구에서 제시하는 모형 등 총 세 가지로 결정된 RFM 모형을 각 데이터셋 A, B, C에 적용하여 고객을 세분화하고 그 등급별 집단의 특성을 살펴보았다. 이때 <그림 4-2>와 같이 각 모형은 모형A, 모형B, 모형C로 정했다.

모형A	파레토 법칙에 의한 RFM 모형
모형B	직관적으로 설계된 RFM 모형
모형C	본 연구에서 제시하는 RFM 모형

<그림 4-2> 세 가지 가중치 선택 방법을 적용한 각 모형의 정의

(1) 데이터셋 A

데이터셋A에 각 모형을 활용하여 고객을 5등급으로 세분화했을 때 R, F, M 세 요소에 대한 평균값은 <표 4-12>와 같다.

파레토 법칙을 활용한 모형 A는 R의 평균이 Diamond 집단을 제외하고 3개 집단에서 동일한 수치가 나왔으며 상위그룹에서 하위그룹으로 갈수록 수치가 낮아지는 반대의 양상을 보였다. 또한 F 평균은 Gold가 Platinum 집단보다 근소한 차이로 높게 나왔으며 금액은 상위그룹에서 하위그룹으로 갈수록 적절히 잘 낮아지는 것을 확인하였다. 직관적으로 설계된 모형 B 또한 R값의 수치가 최상위 그룹에서 하위그룹으로 갈수록 낮아지는 것을 확인하였으며 F값과 M값은 적절히 잘 분배된 것으로 나타났다.

모형A와 모형B 모두 가중치를 높게 준 M값은 각 집단별 타당한 수치가 나타났지만 R값은 하위그룹이 높은 점수를, 상위그룹이 낮은 점수를 보였다.

<표 4-12> 데이터셋 A - RFM 모형 A, B, C에 대한 5등급 집단 비교

	항목	Diamond	Platinum	Gold	Silver	Bronze
모형 A	R	6.9	4.8	4.8	4.3	4.8
	F	5.0	1.6	1.7	1.4	1.2
	M	420416.7	232594.0	52480.8	33237.2	18055.9
	빈도	800	927	900	887	956
모형 B	R	5.5	6.0	5.0	4.5	4.3
	F	3.9	2.2	1.7	1.5	1.3
	M	522518.0	129743.3	49652.1	31895.8	15670.0
	빈도	855	833	991	852	939
모형 C	R	3.6	4.7	5.1	5.2	6.7
	F	4.1	2.0	1.6	1.4	1.3
	M	347125.5	187129.6	134388.8	39788.3	26380.4
	빈도	887	790	966	908	919

(2) 데이터셋 B

데이터셋B <표 4-13>의 경우에는 모형 A, B, C 모두 적절히 잘 분배된 것으로 보인다. 단, 모형 A의 R값은 나머지 두 모형에 비해 편차가 고르지 못한 것으로 확인되었고 직관적으로 설계한 모형 B의 경우에는 Silver 집단의 고객이 673명이고 Bronze 집단의 고객이 1,066명으로 큰 차이가 나타났다.

<표 4-13> 데이터셋 B - RFM 모형 A, B, C에 대한 5등급 집단 비교

	항목	Diamond	Platinum	Gold	Silver	Bronze
모형 A	R	20.6	50.7	89.4	93.2	198.4
	F	12.3	4.4	2.3	1.4	1.0
	M	7034.8	1796.1	904.7	385.5	194.0
	빈도	865	857	879	809	928
모형 B	R	25.9	48.2	86.5	100.7	179.4
	F	12.4	4.4	2.4	1.6	1.1
	M	7750.9	1642.9	728.7	381.6	191.8
	빈도	814	919	866	673	1066
모형 C	R	14.7	45.8	73.7	108.9	205.4
	F	12.7	4.5	2.3	1.4	1.0
	M	7292.6	1864.0	893.9	444.3	215.8
	빈도	803	924	866	798	947

(3) 데이터셋 C

데이터셋C <표 4-14>에 세 개의 모형을 적용한 결과 또한 대체로 상위그룹에는 높은 점수, 하위그룹에는 낮은 점수가 잘 나타났지만 파레토 법칙을 활용한 모형 A의 R 값이 Silver가 Bronze보다 더 낮은 점수로 나타났다.

<표 4-14> 데이터셋 C - RFM 모형 A, B, C에 대한 5등급 집단 비교

	항목	Diamond	Platinum	Gold	Silver	Bronze
모형 A	R	127.2	270.5	444.3	507.0	501.5
	F	8.3	2.8	1.5	1.0	1.0
	M	315.7	97.3	51.6	26.8	12.5
	빈도	455	480	465	476	471
모형 B	R	136.3	289.4	421.7	484.2	501.5
	F	8.3	2.9	1.6	1.2	1.0
	M	340.3	96.2	47.1	26.7	12.5
	빈도	423	463	444	546	471
모형 C	R	129.7	250.9	438.9	498.8	507.5
	F	8.5	2.9	1.6	1.0	1.0
	M	318.7	97.6	58.8	31.5	15.1
	빈도	440	470	418	409	610

데이터셋 A, B, C에 세 가지 RFM 모형을 적용하고 비교해본 결과 전반적으로 최상위 그룹에서 높은 점수의 평균치가 나왔으며 하위그룹으로 갈수록 점수가 낮아지는 경향을 보였다. 다만 모형A와 B는 각 집단별 평균이 상위그룹보다 하위그룹에서 더 높은 수치가 나오거나 동일한 수치가 나오기도 했으며 특정 그룹에 인원이 몰리는 현상이 발생하기도 하였다. 이에 비해 K-Means Clustering 기법을 통해 데이터셋 변수의 동질성을 반영하고 CV의 최솟값으로 안정성까지 활용한 모형C는 세 모형 중 가장 안정적이고 이상적인 분포를 보였다.

제 5 장 결론

제 1절 연구의 요약

본 논문에서는 CRM(Customer Relationship Management, 고객관계관리)에서 일반적으로 널리 쓰이는 RFM 분석 기법에 사용되는 RFM 모형의 가중치를 선택하는 방안에 대한 연구를 진행하였다.

가중치 선택은 RFM 모형 설계 시 가장 중요한 단계로서 이때 결정된 가중치에 따라 각 고객에게 부여되는 RFM 점수가 달라지며 이 점수를 기반으로 고객을 세분화하는 과정에서도 각 집단의 특성이 달라진다.

적절한 가중치를 선택하기 위해 고객 아이디, 구매일자(Recency). 구매빈도(Frequency), 구매금액(Monetary) 등을 산출할 수 있는 정보가 포함된 총 세 가지의 고객 구매 데이터를 활용하였으며 R, Microsoft Excel 등을 이용해 1차 데이터 전처리 과정을 거쳤다. 이후 각 데이터셋을 K-Means Clustering 알고리즘을 활용하여 군집화하고 각 군집이 갖는 기술 통계량을 가지고 CV(Coefficient of Variation, 변동계수)를 산출하였다. K-Means Clsutering은 Elbow Method를 통해 최적의 K를 찾고 각 요소별 거리를 기반으로 최종 군집을 이루는 방법이므로 각 데이터셋이 가지고 있는 변수의 특성을 반영한다는 점을 이용하였고 이를 토대로 각 군집별 R, F, M값의 표준편차를 평균으로 나누어 CV값을 산출하고 CV를 최소로하는 값을 찾아 전체 CV의 합으로 다시 나누어 가중치를 산출하였다. 즉, K-Means Clustering 기법의 동질성과 CV 최솟값의 안정성을 반영하고자 하였다. 이때 Python의 기계학습 라이브러리인 scikt-learn을 주로 활용하였으며 pandas, matplotlib, numpy, seaborn, datetime 등 다양한 패키지를 사용하였다.

이렇게 선택된 가중치 W1, W2, W3을 통해 최종적으로 RFM 모형을 설계하고 데이터셋 A, B, C에 각각 적용하여 전체 고객을 5개의 등급으로 세분화하여

각 집단별 특성을 비교하였으며 분산분석을 통해 5개 집단 분류의 유의미함을 판단하였다.

이후 선행연구에서 제시된 파레토 법칙을 활용한 모형과 직관에 의한 모형을 설계하여 총 세 가지 모형을 비교하였는데 모두 고객을 적절하게 분류한 것으로 나타났다. 다만, 데이터 특성에 따라 상위그룹보다 하위그룹의 평균이 더 높게 나타나거나 동일하게 나타나는 경우가 있었는데 특히 데이터셋A의 경우에 파레토법칙을 활용한 모형이나 직관에 의한 모형은 R(구매시기)값에 대한 분류가 완전히 뒤바뀐 수치를 보이면서 M(구매금액)값에만 치우치는 경향을 나타낸 것과 반면에 본 연구에서 제시하는 모형은 모든 변수에 대해서 적절히 분류된 것을 확인할 수 있었다. 또한 데이터셋 B에 직관적으로 설계한 모형을 적용했을 때 Bronze 집단 고객이 1,066명으로 나타났는데 이는 673명으로 구분된 Silver 집단에 비해 훨씬 많은 수치로서 이렇게 특정한 집단에 몰리는 현상도 확인할 수 있었다.

본 연구에서는 오늘날까지도 널리 사용되는 RFM 모형의 가중치 선택에 있어서 보다 객관적인 방법을 제시하고자 하였다. 기존에는 기업 담당자나 연구자가 중요도가 높다고 생각하는 항목에 대해 높은 가중치를 부여하거나 복잡하고 어려운 통계 기법을 활용하여 가중치를 찾아 RFM 모형을 설계하였다. 본 연구에서는 이러한 과정 없이 기계학습 알고리즘과 각 변수가 갖는 기술 통계량을 가지고 동질성과 안정성을 갖는 간단하고 객관적으로 가중치를 선택하는 방안을 제시하였다. 이러한 가중치 선택 과정을 거친 RFM 모형을 기업에서 활용한다면 기존에 담당자가 사용하던 모형과 비교하여 더 나은 모형을 선택하거나 두 가지 모형을 적절히 병행한 방법의 세분화를 하면서 보다 나은 홍보 활동을 진행할 수 있을 것이다.

제 2절 연구의 한계점

RFM분석 기법은 고객의 R, F, M 변수를 통해 적절한 가중치를 부여하여 최

종 점수를 매기는 데 있어 매우 주관적인 성격을 갖는다. 이는 특정 고객을 어떤 등급의 집단으로 세분화하는가에 대한 문제인데 여기에는 명확하게 제시된 가이드가 없으며 데이터 특성뿐만 아니라 해당 기업의 특성이나 고객의 특성, 마케팅 담당자의 주관 등 달라질 수 있는 변수가 매우 많기 때문이다. 따라서 과거의 모형과 비교 검증을 하는 데 있어 R, F, M 평균값을 통한 비교 정도만 할 수 있었으며 통계적으로 제시할 수 있는 부분은 굉장히 제한적이었다. 이는 통계적으로는 유의미하다고 판단되더라도 담당자나 연구자의 주관에 의해 달라질 수 있기 때문이다.

또한 고객 구매데이터는 각 기업의 민감한 자료로서 대부분 폐쇄적이기 때문에 본 연구에서는 인터넷에 공개된 표준데이터나 해당 기관에서 제공한 데이터를 활용하였다. 따라서 제한된 표본과 변수만을 활용하여 연구를 진행하였으며 더 많은 표본과 다양한 변수를 적용하지 못한다는 한계가 있었다.

제 3절 향후 연구 방안

본 논문에서 제시된 RFM 모형의 가중치 선택 방법이 각 기업의 RFM 분석 시 활용되어 고객을 세분화하는 전략에 있어서 조금이나마 도움이 될 수 있기를 기대한다. 향후에는 더 많은 데이터셋의 다양한 변수를 활용하여 기업의 특성이나 고객에 특성까지 고려한 가중치를 선택하는 방안에 대한 연구가 이뤄질 필요가 있으며 고객을 단순히 5등급으로 세분화하는 것이 아닌 몇 개의 등급으로 세분화할지에 대한 연구도 필요하다. 또한 선행연구에서 제시한 가중치 선택방법과 본 연구에서 제시한 가중치 선택 방법에 대한 차이를 더 명확히 구분할 수 있는 방안을 제시할 필요가 있다.

더 많은 변수를 다양하게 적용하여 가중치를 산출해보는 연구와 새롭게 설계된 모형과 기존 모형들의 비교 분석을 하는 연구는 지금보다 더욱 효과적이고 신뢰도 높은 RFM 모형을 설계하는데 큰 도움이 될 수 있을 것이다.

참고문헌

<국내>

Dev log. (2010). [케라스] 무작정 튜토리얼12 - Scikit-learn의 Scaler. https://ebbnflow.tistory.com/137

김동훈. (2008). RFM분석을 통한 가전소비자의 전략적 구매 행동 연구. 성균관대학교 대학원, 석사학위논문.

김연형, 김재훈, 이석원. (2010). 고객관계관리와 데이터마이닝 - 사례 연구 중심으로 - (개정판). 교우사, ISBN-13: 978-89-8172-624-9.

김태진, 김성수, 전다희, 박상현. (2020). RFM모형을 활용한 지역별 재해 위험도 분석 방법론 제안. Journal of the Society of Disaster Information Vol. 16, No. 3, pp. 493-504, September 2020.

동상옥, 김규곤, 조성기. (2008). TRFM 모형을 이용한 패션기업의 고객세분화. Journal of the Korean Data Analysis Society, Vol. 10, No. 6(B), December 2008, pp 3267-3277

류귀열, 문영수. (2006). 연관분석을 이용한 데이터마이닝 기법에 관한 사례연구. Journal of the Korean Data Analysis Society, Vol. 8, No. 3, June 2006, pp. 1021-1033

류귀열, 문영수. (2013). RFM에서 등급부여 방법에 관한 연구. Journal of the Korean Data Analysis Society, 2013, 24(2), 245-255

문영수. (2005). 데이터 마이닝 기법에 관한 연구. 서경대학교 대학원, 석사학위논문.

박희창. (2010). 항목 알에프엠점수를 고려한 가중 연관성규칙. Journal of the Korean Data Analysis Society, 2010, 21(6), 1147-1154.

서현지. (2017). 빅 데이터를 이용한 고객 행태 분석에 대한 연구 -유통업 사례를 중심으로-. 가천대학교 대학원, 석사학위논문.

안범준, 송기정, 서광규. (2011). Latent Class Analysis 기반의 만족 고객 세분화를 이용한 고객만족경영 향상 방안. The Journal of the Korea Contents Association (한국콘텐츠학회논문지), Volume 11 Issue 12 / Pages.386-394 / 2011 / 1598-4877(pISSN) / 2508-6723(eISSN)

안요찬. (2010). 고객세분화를 통한 한방병원 고객관계관리시스템 구축 모형. 한국산업정보학회논문지 제15권 제5호 (2010, 12).

염창선, 정윤필. (2013). 고객세분화를 통한 인터넷 쇼핑몰의 고객관계관리. Journal of KIIT. Vol. 11, No. 12, pp. 159-167, Dec. 31, 2013.

이강태. (2002). eCRM 환경에서 LTV극대화를 위한 고객세분화 기법에 관한 연구. 전주대학교 대학원, 석사학위논문.

이소영, 최승배, 김규곤, 강창완. (2004). 고객 세분화를 위한 최적 RFM 모형 구축에 관한 연구. Journal of the Korean Data Analysis Sosietym Vol. 6, No. 6, December 2004, pp. 1829-1840

이소영, 최승배, 김규곤, 김형도, 강창완. (2005). 의료분야에서의 고객 세분화를

통한 수익성 모형 개발에 관한 연구 - 데이터마이닝 기법을 이용하여 -. Journal of the Korean Data Analysis Sosietym Vol. 7, No. 2, April 2005, pp. 523-532

이영진, 강창완, 김규곤, 최승배. (2010). CRFM 모형을 이용한 고객세분화와 모형평가. Journal of the Korean Data Analysis Society, Vol. 12, No. 6 (B), December 2010, pp. 3283-3293

이용구. (2007). BIKorea 칼럼.
http://www.bikorea.net/news/articleView.html?idxno=18

이윤성, 김규곤, 강창완. (2010). 고객세분화에서 TRFM을 반영한 고객생애가치모형의 개발. Journal of the Korean Data Analysis Society, Vol. 12, No. 6 (B), December 2010, pp. 3271-3282

이지민. (2020). 데이터마이닝 기법을 활용한 외식소비자들의 구매 연관성 분석. 경희대학교 대학원, 박사학위논문.

이창해. (2015). K-means 클러스터링 마이닝기법을 활용한 개인별 음원 추천 모형에 관한 연구. 연세대학교 대학원, 석사학위논문.

전용준. (2007). 고객세분화에 대한 이해와 활용방안 (BI 코리아)

전희주. (2011). 고객세분화에 기반한 생존분석을 활용한 고객 수명예측모델. 한국통계학회논문집, 2011, 18권, 6호, 687-696.

정윤필. (2009). 두 기간 RFM을 이용한 인터넷 쇼핑몰의 고객세분화. 부경대학교 대학원, 석사학위논문.

조광현, 박희창. (2012). 고객세분화를 위한 RFMP 기반 CFD 기법의 적용 방안. Journal of the Korean Data Analysis Society (June 2012) Vol. 14, No. 3 (B), pp. 1291-1300.

조혜정. (2001). 고객세분화를 위한 데이터마이닝 기법 비교. 동아대학교 대학원, 석사학위논문

최영희. (2001). 데이터마이닝을 이용한 데이터 활용에 관한 연구. 조선대학교 대학원, 석사학위논문.

최희경. (2002). 통계적 기법을 이용한 RFM 모형 비교 분석. 중앙대학교 대학원, 석사학위논문.

황준경. (2006). 고객 세분화를 통한 고객 분석 효과에 관한 연구. 서울시립대학교 대학원, 석사학위논문

<국외>

Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. Marketing Science, 14(4), 378-394.

Dr. Shailendra Kumar Srivastava. (2019). Potential Customer Segmentation and Customer Relationship Management strategies in Online Retail in India. Synerge, I.T.S journal of IT & Management, Vol. 17, No. 2, July-December, 2019.

Express Analytics. (2021). The many benefits of customer segmentation. https://expressanalytics.com/blog/the-many-benefits-of-customer-segmentation-20151119/

Fader and Hardie. (2001). Forecasting Repeat Sales at CDNOW: A Case Study. Interfaces Volume 31, Number 3, Part 2 of 2 May-June 2001.

Gönül, Kim, & Shi. (2000). Mailing_smarter_to_catalog_customers. JOURNAL OF INTERACTIVE MARKETING VOLUME 14 / NUMBER 2 / SPRING 2000

Guillaume Martin. (2018). Guillaume Martin' github. https://guillaume-martin.github.io/rfm-segmentation-with-python.html

Hughes, A. M. (2000). Strategic database marketing: the masterplan for starting and managing a profitable, customer-based marketing program (Vol. 12). New York: McGraw-Hill.

John R. Miglautsch. (2000). Thoughts on RFM scoring. Journal of Database Marketing Vol. 8, 1, 67 - 72.

Mrs. B. Siva Jyothi. (2020). Marketing Model for efficient CRM (Customer Relationship Management). Anil Neerukonda Institute of Technology & Sciences.

Mrs. Bharati M. Ramageri, Lecturer. (2010). DATA MINING TECHNIQUES AND APPLICATIONS. Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305.

Norma Ningsih. (2020). AplikasiAnalisis Segmentasi Pelanggan untuk Menentukan Strategi Pemasaran Menggunakan Kombinasi Metode k-Meansdan Model RFM. SISTEMASI: Jurnal Sistem Informasi ISSN:2302-8149 Volume 10, Nomor 1, Januari 2021: 139-151.

Pang-Ning Tan. (2005). Introduction to data mining.

Randall S. Collica. (2007). CRM Segmentation and Clustering, Using SAS Enter prise Miner. ISBN-13: 978-1590475089

Thomas-haslwanter. (2020). thomas-haslwanter' github, statsintro_python (ANOVA). https://github.com/thomas-haslwanter/statsintro_python/blob/master/ipynb/8_anovaOneway.ipynb

국문 초록

본격적인 4차 산업혁명 시대를 맞이하면서 전반적인 분야에서 데이터를 활용한 다양한 전략들이 제시되고 있다. 이는 기업의 전략적인 의사결정을 통해 물질적, 시간적 비용을 절감할 수 있고 기업 이윤을 극대화할 수 있는 새로운 마케팅 방법론을 고안할 수 있게 되었다.

급변하는 생활 패턴 속에서 고객은 다양한 욕구를 표현하고 있으며 기업은 이러한 고객의 기대에 부응하기 위해 고객의 구매 데이터를 정제하고 분석하여 고객 구매 패턴을 예측하고 기존 고객을 유지하면서 신규 고객을 모집 할 수 있는 방안을 제시하고 있다. 이러한 방법 중 대표적으로 널리 사용되고 있는 고객관계관리(CRM, Customer Relationship Management)는 오늘날 대량의 데이터가 축적되면서 데이터마이닝 기법과 연계하여 다양한 방법으로 분석할 수 있다. RFM(Recency, Frequency, Monetary)모형은 전통적인 고객관계관리 기법 중 하나로 간단하고 편리한 모델링 방법이며 그 예측력이 뛰어나 현재까지 많이 사용되고 있다. 그러나 RFM 모형 설계에서 가장 중요한 것은 각 변수에 가중치를 부여하는 것인데 현재까지 명확한 가중치의 기준은 제시된 것이 없는 상태다.

본 논문에서는 RFM모형을 보다 효율적으로 설계할 수 있는 가중치 선택 방안을 제시하고자 한다. 빅데이터 분석기술 중 하나인 군집분석에서 대표되는 K-Means 알고리즘을 활용하여 데이터를 군집화하고 각 군집의 Recency, Frequency, Monetary 세 가지 요소의 CV(Coefficient of Variation, 변동계수)값을 활용하여 RFM 모형의 가중치로 선택한다. 최종적으로 설계된 RFM 모형에서 RFM 점수를 산출하여 각 고객들에게 부여하고 전체 고객을 5개의 등급화된 집단으로 세분화하는데 목적이 있다.

본 연구에는 총 3개의 데이터셋이 활용되었으며 각 데이터셋은 6,116건, 549,019건, 6,919건의 고객 구매데이터로 이루어져 있고 인구통계학적 변수를 포함한 고객 정보와 더불어 RFM 분석에 필요한 세 가지 요소인 구매시기(Recency), 구매빈도(Frequency), 구매금액(Monetary)를 포함한다.

Python의 기계학습 라이브러리를 활용하여 Elbow method를 통해 얻어낸 최적의 K값을 각 데이터셋에 적용하고 K개로 나눠진 각 집단의 R, F, M 변수의 통계량을 통해 가중치를 선택하여 최종 RFM 모형을 설계하고 전체 고객들에게 점수를 부여하여 최종 5개의 집단으로 분류하였다.

이를 통해 목표변수가 없는 구매데이터 특성의 한계를 극복할 수 있었으며 K-Means Clustering을 통해 군집화된 집단별 변수의 특성이 반영된 객관적인 방법으로 가중치를 선택한 RFM 모형을 설계할 수 있었고 이를 통해 보다 나은 고객 세분화가 가능하였다.

핵심어: 고객관계관리, 기계학습, 군집분석, K-Means Clsutering, RFM분석, RFM 가중치, 고객세분화