KBR 제24권 제3호 2020년 8월 http://dx.doi.org/10.17287/kbr.2020.24.3.125

머신러닝 기반의 보상형 크라우드펀딩 성공 예측 모델링

문 동 지*
윤 상 혁**
최 수 빈***
김 희 웅****

크라우드펀딩은 최근 자금 조달 경로로 이용되며 소셜미디어와의 접목을 통해 빠르게 성장하고 있다. 2018년 기준 세계 크라우드펀딩 규모는 93억 7천만 달러, 한국 크라우드펀딩 시장은 1.1억 달러로 추정된다. 그러나 국내 크라우드펀딩 실패 확률은 2019년 기준으로 38%에 달하며, 펀딩 프로젝트가 실패할 경우 참여자(창설자, 투자자, 플랫폼) 모두에게 큰 부담이 된다. 만약 프로젝트 초기에 펀딩 성공 여부를 예측할 수 있다면 시간적 금전적 손해를 예방할 수 있다. 이에 본 연구는 국내 크라우드펀딩을 대상으로 성공 여부를 예측하는 모델을 구축하고자 한다. 기존 연구들 대부분은 펀딩 프로젝트가 끝난 후의 데이터를 사용했지만, 본 연구에서는 크라우드펀딩 사이트 와디즈의 펀딩 초기인 7일 이내의 댓글 데이터와 펀딩 참여 건수을 수집하여 예측 변수로 사용하였다. 예측 모델링 기법은 Decision Tree, SVM, Naive Bayes, AdaBoost, Gradient Boosting, Random Forest, MLP와 같은 머신러닝 알고리즘을 활용하였다. 예측 결과 Gradient Boosting이 90% 넘는 정확도를 보였고, Support Vector Machine이 가장 높은 정밀도(Precision, 0.95)를 보였다. 본 연구는 머신러닝 기반의 예측 모델을 개발함으로써, 크라우드펀딩 초기 단계에서 펀딩 성공 여부를 예측할 수 있다는 실무적 의의가 있다.

주제어: 크라우드펀딩, 와디즈, 예측 모델링, 머신 러닝, 감성 분석

Ⅰ. 서 론

디지털 통신 기술 및 온라인 소셜 네트워크가 활성화되면서 “크라우드펀딩”이 주목받고 있다(Schwienbacher and Larralde, 2010; 김재일 외, 2019). 크라우드펀딩은 다수를 의미하는 크라우드(crowd)와 펀딩(funding)의 합성어로 웹이나 모바일을 통해 다수의 개인으로부터 자금을 모으는 행위를 의미한다(Haas et al., 2014). 크라우드펀딩은 일반적으로 세 명의 참여자를 통해 자금을 모으는 행위가 이루어진다: 첫 번째로, 자금 조달을 위한 제안이나 캠페인을 제안하는 프로젝트 창설자(Maker), 두 번째는, 크라우드펀딩을 후원하는 투자자(Supporter), 그리고 크라우드펀딩을 중계하는 플랫폼(Platform)이다. 창설자는 자금 조달을 위한 제안이나 펀딩 프로젝트를 제안하는 역할을 하게 된다. 투자자는 펀딩 프로젝트 투자 대가로 다양한 유형의 가치를 받고, 플랫폼은 창설자와 투자자에게 효율적인 투자과정이 이루어질 수 있는 환경을 조성하는 역할을 한다. 창설자와 투자자는 개인에서부터 조직에 이르기까지 다양하며 플랫폼은 다양한 창설

논문접수일: 2020. 05. 25. 1차 수정본 접수일: 2020. 07. 06. 게재확정일: 2020. 07. 29.
* CJ대한통운 사원(dongji.moon@cj.net), 제1저자
** 연세대학교 정보대학원 박사(scottyoon@yonsei.ac.kr), 교신저자
*** 연세대학교 정보대학원 석사(soobin924@yonsei.ac.kr)
**** 연세대학교 정보대학원 교수(kimhw@yonsei.ac.kr)

자의 크라우드펀딩 등록을 심사하고, 펀딩 프로젝트가 성공적으로 끝나면 중계 수수료를 받는다.

세계 크라우드펀딩 규모는 2018년을 기준으로 약 93억 7천만 달러로 추정되며, 향후 4년간 연평균 29%의 높은 성장률을 기대하고 있으며 2022년도에는 약 259억 2천만 달러 규모까지 확대될 것으로 전망되고 있다(Statista, 2018). 크라우드펀딩의 높은 성장과 함께 대표적인 글로벌 크라우드펀딩 사이트인 킥스타터(Kickstarter)는 수많은 펀딩 성공 사례가 있다. 국내에서도 크라우드펀딩 시장 관련 규제 완화가 시장 전체에 순기능을 주면서 기업들이 초기 홍보 효과와 린스타트(Lean Start)를 쉽게 경험할 수 있게 되면서(Ambani, 2017), 성공 사례를 만들고 있다.

크라우드펀딩의 대표적인 유형은 보상형(Reward-based), 대출형(Lending-based), 지분투자형(Equality-based), 기부형(Donation-based) 4가지로 나뉜다(Cumming, 2014). 본 연구에서는 4가지 유형 중에서 프로젝트 참여의 진입 장벽이 낮고, 누구나 참여가 쉬운 보상형 크라우드펀딩을 집중적으로 연구하고자 한다. 보상형 크라우드펀딩은 목표 펀딩 금액을 달성해야 성공적으로 보상을 받을 수 있으며 목표 금액에 달성하지 못하면 투자금 전체가 반환되는 "All or Nothing" 전략을 따른다. 따라서 프로젝트의 성공과 실패 여부를 미리 판단하는 것이 중요하다(Bi et al., 2017). 성공 여부를 예측할 수 있으면, 시간적 기회비용을 낮출 수가 있으며 자금 조달 기회를 상실하지 않고 펀딩을 완료할 수 있다. 하지만 2019년 기준 국내 크라우드펀딩 가운데 실패하는 비율이 2019년 기준 38%가 넘어서(크라우드넷, 2020) 크라우드펀딩 창설자에게 부담감을 주고 있다.

따라서 크라우드펀딩 플랫폼들은 시간적 비용을 감소시키고, 성공적인 자금 조달이 가능하게 하려고 펀딩 프로젝트 성공 여부를 예측하고자 노력하고 있다. 이에 비교해 국내 크라우드펀딩 관련 예측 연구가 많이 이루어지고 있지 않아 학술적 및 실무적 연구가 부족한 실정이다. 또한, 크라우드펀딩 성공 예측과 관련된 기존 연구들(Nam et al., 2018; Yu et al., 2018; Yuan et al., 2016) 의 한계점은 데이터 대부분을 펀딩 종료 후에 수집했다는 점이다. 그 이외 선행 연구들을 살펴본 결과, 예측 정확도가 다소 낮고(Greenberg et al., 2013) 데이터 수집 기준이 명확하지 않다(Li et al., 2016)는 한계를 가지고 있다.

따라서 본 연구의 목적은 보상형 크라우드펀딩 초기시점에 정확도 높은 성공 예측 모델을 제시하는 것이다. 이를 위해 먼저 보상형 크라우드펀딩의 댓글 정보, Facebook 지지 서명, 펀딩 참여 건수와 같은 데이터 정보를 추출해 펀딩 프로젝트의 성공 여부를 예측할 것이다. 또한, 프로젝트 초기이면서 성공 예측률이 높은 최적의 기간을 선정하고 예측모델링을 수행하여 결과를 검증해보고자 한다. 예측모델링은 Decision Tree, Support Vector Machine (SVM), Navie Bayes, AdaBoost, Gradient Boosting, Random Forest, Multi-Layer Perceptron (MLP) 을 활용해 최적의 머신러닝 알고리즘을 찾아보고자 한다. 마지막으로 기존 크라우드펀딩 예측모델링 연구(Greenberg et al., 2013; Kamath and Kamat, 2016) 에서 활용한 변수와 국내 환경에서 유의한 변수를 비교하기 위해서 로지스틱 회귀분석을 수행할 것이며, 기존 크라우드펀딩 결과 예측(Sawhney et al., 2016; Yu et al., 2018; 이강희 외, 2018)에서 사용하지 않은 예측 모델링 기법인 Gradient Boosting 기법을 추가로 사용해 예측 모델을 개발하고자 한다.

Ⅱ. 개념적 배경

2.1 크라우드펀딩

크라우드펀딩은 다수의 대중으로부터 자금을 조달하는 방식을 의미하며, 플랫폼을 통해 불특정 다수로부터 자금을 모으는 방식으로 정의할 수 있다(Haas et al., 2014). 국내 크라우드펀딩 산업의 주요 동향을 살펴보면 지난 3년간(2016~2018) 417개의 창업 및 벤처기업이 크라우드펀딩을 통해 755억 원의 자금이 조달되었다(건당 평균 1.6억 원). 특히 2018년에는 178개의 기업이 301억 원(185건)을 조달하는 등 이용 기업 수와 조달 금액이 꾸준히 증가하는 추세이다(금융위원회, 2019). 제도적으로도 2019년부터 연간모집 한도를 7억 원에서 15억 원으로 확대되면서 평균 조달 금액이 매우 증가하였다. 연간모집 한도가 늘어나면서 알고리즘 기반의 펀드 추천 서비스를 제공하는 '두물머리' 회사는 15억 원의 자금을 모집 성공하였다. 그 외에도 '지피페스트'는 음악페스티벌 개최 자금을 9.7억 원 조달하였고, '타임 기술'은 선진 군수 지원 사업을 위한 자금 9.3억 원을 조달 성공하였다(박규석, 2019).

다음으로, 크라우드펀딩 국내외 사례를 좀 더 구체적으로 논해보고자 한다. 세계 최대 크라우드펀딩 플랫폼은 킥스타터(Kickstarter)이다. 킥스타터 성공 사례로는 디자이너 Scott Wilson이 만든 시곗줄 제작 프로젝트가 있다. Scott Wilson은 시곗줄의 3D 도면을 동영상으로 설명하여 후원자를 모집해 몇 시간 만에 $100만의 투자금을 모집하였다. 그 외에도 미국 신생기업인 페블 테크놀로지스사는 스마트워치 '페블'의 개발 자금을 구하기 위해 2012년 킥스타터에서 펀딩 프로젝트를 시작해 몇 시간 만에 1,026만 달러를 모금에 성공하였다(Wikipedia, 2020). '페블' 사례를 계기로 킥스타터는 스타트업이나 예술 프로젝트 창작자가 자금을 모으기 위한 대표 수단으로 자리 잡을 수 있었다. 한국의 대표적인 크라우드펀딩 플랫폼은 와디즈(Wadiz)이다. 대표적인 성공 사례로는 영화 '노무현입니다'가 있다. 이 영화는 제작비가 부족하여 제작 중단 위기가 있었으나 와디즈의 크라우드펀딩으로 제작비를 모아 개봉까지 성공하였다. 이 외 성공 사례로 '영철버거' 프로젝트가 있다. 영철버거는, 고려대학교 앞에 있던 햄버거 가게로, 대형 햄버거 프랜차이즈 등장으로 폐업 위기를 맞았다. 이때 고려대학교 학생회에서 와디즈에서 '비긴 어게인 영철버거'라는 이름으로 영철버거 살리기 프로젝트를 시작해 5,000만 원을 모금하여 영업을 재개하였다(김병주, 2017).

크라우드펀딩의 투자 모델은 크게 4가지 유형으로, 보상형(Reward-based), 대출형(Lending-based), 지분투자형(Equality-based), 기부형(Donation-based)이 있다(Cumming, 2014). 보상형은 투자자가 투자에 대한 보상을 제조된 제품이나 서비스를 보상받는 형태이다. 대출형은 투자자의 자금이 대출형태로 제공되며 이자를 통해 일부 수익을 기대할 수 있다. 지분투자형은 크라우드펀딩으로 주식이나 수익 증권을 소유한다. 마지막으로 기부형은 투자자가 보상을 전제로 하지 않는 기부하는 형식의 펀딩이다. 본 연구에서는 다루게 될 주 모델은 보상형 크라우드펀딩이다. 보상형 크라우드펀딩은 2019년 기준 와디즈에서만 매달 700개 이상 오픈되고 있다(윤경희, 2019).

보상형 크라우드펀딩 시장이 이처럼 상승하는 이유 중 하나는 바로 선 주문 생산을 진행하는 유통 구조 때문이다. 이 유통 구조에서 사업자는 재고 부담을 덜 수 있으며 임대료, 인건비, 유통 수수료 등을 고려하지 않아도 된다는 장점이 있다. 소비자로서도 유통 구조가 최소화되면 최종 소비자 가격이 낮아지는 장점이 있다. 또한, 크라우드펀딩을 통한 성공은 대형 유통업체 입점까지 연결되는 루트가 될 수도

있다. 펀딩 성공이 후 과정이 공개적으로 진행되기 때문에 홈쇼핑, 대형마트, 백화점 등 문턱이 높은 대형 유통업체의 담당자 눈에 띄어, 좋은 조건으로 입점 제안을 받을 수 있기 때문이다. 하지만 보상형 크라우드펀딩 프로젝트가 실패할 경우 창설자뿐 아니라 투자자와 플랫폼 모두에게 시간적 금전적인 손해를 미치게 된다. 즉, 크라우드펀딩 성공 여부를 초기에 예측할 수 있다면 프로젝트 참여자 모두에게 도움이 될 것이다.

2.2 크라우드펀딩 예측 모델 관련 선행 연구

크라우드펀딩 성공 예측과 관련된 대부분의 선행 연구들은 킥스타터의 보상형 크라우드펀딩의 프로젝트의 데이터를 이용하여 성공 여부를 예측하였다. 선행 연구에서 사용된 예측 변수는 프로젝트에 대한 기본 데이터, 소셜미디어, 언어적, 비언어적 특징을 사용하였다. 크라우드펀딩 예측 관련 선행 연구들은 〈표 1〉과 같이 정리하였다.

Greenberg et al.(2013)는 킥스타터에서 펀딩 프로젝트가 시작할 때 알 수 있는 정보만을 이용하여 성공과 실패를 예측하고자 했다. 하지만, Random Forest를 사용했을 때 68%의 정확도로 다른 선행 연구보다 상대적으로 예측 정확도가 낮다는 연구 한계를 가지고 있다. Hui et al.(2016), Yu et al.(2018), Nam et al.(2018)의 연구들은 데이터를 펀딩이 종료 후에 수집해 예측 변수를 선정하고 예측모델링을 수행하였다. 실제 크라우드펀딩 시장 환경에서는 펀딩 성공 여부 예측이 초기에 진행되어야 한다는 점에서 이 연구들의 한계점이 있다. 이 외에도 Li et al.(2016)는 크라우드펀딩 성공 여부가 아닌 실제 펀딩 금액을 예측하는 연구도 있었다. 이 밖에도 크라우드펀딩 관련 연구는 국내외 크라우드펀딩 현황 분석(곽현・이호근, 2014; 권보람・김주성, 2013), 크라우드펀딩의 성공 요인(이정은・신형덕, 2014), 성공과 실패사례 분석(권혁인 외, 2014) 등이 있지만 크라우드펀딩 성공 예측 모델과 관련된 연구는 초기 단계다.

기존 연구들을 확인한 결과, 연구 환경이 글로벌 크라우드펀딩 플랫폼인 킥스타터에 집중되어 있으며, 예측 시기가 펀딩 종료 후라는 한계가 있다. 따라서 본 연구에서는 국내 크라우드펀딩 시장 환경에 맞춰, 크라우드펀딩 초기에 성공 여부를 예측할 수 있는 모델을 개발하고자 한다.

2.3 예측 모델링 기법

예측 모델링(Predictive Modeling)이란 미래를 예측하기 위해 사용되는 기법으로, 데이터를 수집하고 예측 모형을 설정하여 예측 수립 및 검증과 수정 등의 과정을 통해 모형을 만드는데 최근에는 머신러닝 알고리즘을 많이 활용하고 있다(Kuhn and Johnson, 2013). 이에, 본 연구에서는 크라우드펀딩 결과를 예측 모델링하기 위해 7가지 머신러닝 알고리즘: Decision Tree, Naive Bayes, SVM, Gradient Boosting, Random Forest, AdaBoost, MLP을 사용하고자 한다. 각 머신러닝 알고리즘별 설명은 다음과 같다.

Decision Tree는 지도 분류 학습에서 가장 유용하게 사용되고 있으며 예측된 결과로 입력데이터가 분류되는 클래스를 출력한다. 이를 위해 지니 지수(Gini index), 엔트로피 지수(entropy index) 등의 불순도 측도(impurity measures)가 이용된다. 분석 방법으로는 통계학에 기반한 CART(Classification And Regression Trees), CHAID 알고리즘, 기계학습 계열인 ID3, C4.5, C5.0등의 알고리즘이 있다. 가장 널리 쓰이는 의사결정나무 알고리즘인 CART는 범주형 반응변수의 경우 지니지수를, 연속형 반응변수의 경우 분산을 이용하여 최적의 설명변수를 찾아낸다(Lewis, 2000). 모형의 성능을 조절하기

〈표 1〉 크라우드펀딩 예측 관련 선행 연구

구분	저자	사용 변수	연구결과
Crowd-funding Success	Greenberg et al. (2013)	Funding Goal, Project Category, Reward_count, Duration, Twitter_url, Has_video, facebook_connected, facebook_friend, Content Sentiment, Number of sentences in project_description	Decision Tree: 67.68% Random Forest: 67.53% Boosted Decision Stump : 65.10% Logistic Regression: 65.09%
	Sawhney (2016)	Unigrams, Sentiment, Part-of-Speech Tagging, Flesch-Kincaid Readability, LDA, Conditional Probability	Naïve Bayes: 65% Support Vector: 92%
	Hui et al. (2016)	Goal Amount, Credit Score, Number of different amounts to be donated, (Project Info, News Article)DC-LDA Topic Modeling Result	Random Forest: 95% Neural Network: 94% Support Vector: 80%
	Yu et al. (2018)	Goal Amount, Category, Currency, Deadline, Launching Date, Backers, Country	MLP: 93% Random Forest: 92% AdaBoost: 92% SVM: 92% Decision Tree: 90% Logistic Regression: 89% Naïve Bayes: 71%
	남수현 외 (2018)	Product category, Reward, Number of replies, Frequency of introduction updates, Target amount, Number of SNS followers, Image, Video, Number of founders in a project	C5.0: 84% K-NN: 72% Logistic Regression: 57% Naïve Bayes: 58% Random Forest: 88% SVM: 87%
Funding amount	Li, et al. (2016)	Goal Amount, Duration,The number of Image, Video, Comment, Project Description, Risks&Challenges, FAQs, Counting the number of words, Geo-location, Category, Project Creators, Twitter bi-connected components	Cox: 77% Tobit: 78% BJ: 80% BoostCI: 81% Logistic: 83% Log_Logistic: 87%

위해서 파라미터(parameter)를 조정하는 작업이 필요한데, 연구자가 나무를 얼마나 성장(growing) 시킬 것인지, 성장한 나무를 어느 정도까지 어떤 식으로 가지치기(pruning)할 것인지 판단하여 수행한다.

Naive Bayes 분류기는 공통으로 모든 특성값은 서로 독립임을 가정하며 단일 알고리즘을 통한 훈련이 아닌 일반적인 원칙에 근거한 여러 알고리즘을 이용하여 훈련한다(Rish, 2001). 단순하면서도 정확한 추정능력을 발휘한다고 알려져 많은 분류 문제를 해결하고 있다(Cortes and Vapnik, 1995). 이 분류기에서, 모델 w는 각 클래스 c_i가 발생하는 다양한 확률, 즉 {p(C= c_i)}_(i=1)^k와 x의 요소 x_j 가 특정 클래스 c_i에서 발생할 확률로 구성된다. Naive Bayes 분류 기능은 가장 높은 사후 확률을 가진 클래스를 선택하며 작동한다.

Support Vector Machine(SVM)은 결정경계(decision boundary)와 훈련데이터 사이에 최대 여유(margin)를 가지는 초평면(hyperplane)을 설계하는 머신러닝 기법이다(Cortes and Vapnik, 1995; Scholkopf et al., 1998). 인공신경망 기법의 문제점으로 지적되는 과적합 문제를 페널티(Penalty) 항을 이용하여 피할 수 있으며, 또한 함수 근사에 있어서 이상치(outlier)에 둔감하므로 높은 일반화 성능을 자랑한다. 따라서 인공신경망 기법보다 상대적으로 예측력이 우수한 모델의 구현이 가능하다는 장점이 있다(Scholkopf et al., 1998).

Adaboost는 편향(Bias)을 감소시키는 목적으로 간단하면서 효과적인 알고리즘이다. 다른 학습 알고리즘의 결과물들에 가중치를 두어 더하는 학습 방법으로 가속화(Boosting) 분류기의 최종 결과물을 표현한다. 부스팅 방법에는 크게 AdaBoost, Gradient Boosting(GBM), Xgboost(Chen and Guestrin, 2016), Light GBM(Ke et al., 2017) 등이 존재하지만, 본 연구에서는 AdaBoost와 Gradient Boosting을 사용하고자 한다. Gradient Boosting은 Friedman(2001)에 의해 제안된 방법으로 1개의 의사결정 나무보다 여러 분류기의 예측을 종합함으로써 분류의 정확성을 보다 향상시킨 앙상블기법 중 하나이다. 모델을 만들고, 임의의 미분 가능한 손실 함수를 최적화함으로써 모델을 일반화하는 방법이다(Friedman, 2001). Gradient Boosting은 분석용 데이터 관측값의 가중치가 같은 상태에서 시작되어 형성된 분류기로 오 분류 된 관측값은 다음 관측값에 높은 가중치를 준다. 정확하게 분류된 관측값은 낮은 가중치를 주는 과정을 계속해서 반복하면서 최종 분류기를 형성한다.

Random Forest는 Breiman(2001)에 의해 개발된 분류기법으로서, 단일 의사결정 나무를 이용하는 것이 아닌 여러 개의 나무로 확장한 의사결정 나무의 메타학습(meta-learning) 형태를 보이는 머신러닝 기법이다(김성진・안현철, 2016). Random Forest를 학습시킬 때 약한 분류기인 의사결정 나무의 수를 결정하게 된다. 각각의 의사결정 나무는 무작위로 추출된 학습 데이터와 입력변수에 의해 학습된다. Random Forest가 갖는 장점으로는 무작위 복원 추출의 데이터를 사용하여 각각의 의사결정 나무를 학습시키기 때문에 잡음이나 이상치로부터 크게 영향을 받지 않으며 의사결정 나무를 종합하여 예측하는 경우에 전체의 정확도가 높아진다.

Multi-Layer Perception(MLP)은 입력층과 출력층 사이에 하나 이상의 중간층이 존재하는 신경망 모델로, 비선형(Sigmoid, ReLU 등) 활성함수(Activation Function)을 사용하여 역전파(Backpropagation) 방식으로 네트워크 가중치를 갱신하는 방식으로 학습한다(Rosenblatt, 1961). 신경망의 네트워크는 입력층(input layer), 은닉층(hidden layer), 출력층(output layer) 방향으로 연결되어 있으며, 각 마디 간 연결에는 가중치(weight)가 존재한다. 각 층 내의 연결과 출력층에서 입력층으로의 직접적인 연결이 존재하지 않은 전방향(Feedforward) 네트워크를 구성한다.

Ⅲ. 연구 방법론

3.1 데이터 수집: 와디즈(Wadiz)

와디즈(Wadiz)는 한국 1위 크라우드펀딩 플랫폼으로, 2018년 누적 펀딩금액 1,000억 원을 기록하였다.[1] 또한, 한국에서는 최초로 지분투자형 크라우

1) 와디즈, https://platum.kr/archives/115060

드펀딩 서비스를 제공하였고, 현재는 보상형 크라우드펀딩과 투자형 크라우드펀딩 서비스를 제공하고 있다. 본 연구에서 수집한 데이터는 보상형 크라우드펀딩 프로젝트로 2013년 6월부터 2019년 4월까지 진행된 프로젝트 7,316개이다. 프로젝트별 커뮤니티 댓글 260,868개와 펀딩 참여 건수 788,087개, Facebook 지지 서명 횟수 194,281개를 수집하였다. 댓글, Facebook 지지서명, 펀딩 참여 건수의 경우 '한달전', '이틀전' 등의 표시 처리와 댓글이 없는 프로젝트를 제외하고 최종적으로 총 5,912개의 프로젝트가 분석에 사용하였다.

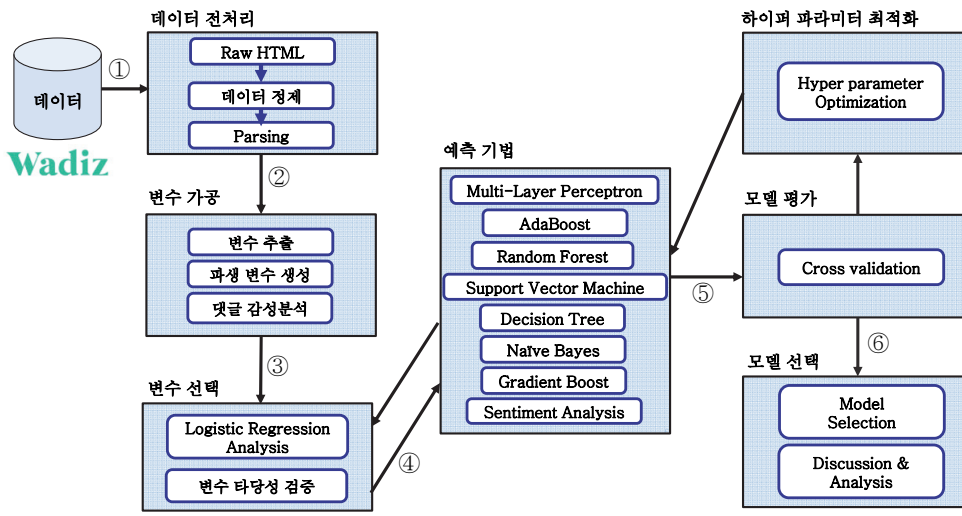
본 연구에서 활용한 예측 변수들은 카테고리(Category), 펀딩 기간(Term), 펀딩 시작 월(Month), 동영상 개수(Video), 사진 개수(Image), 목표 펀딩 금액(Amount), 펀딩 하트(Like) 개수는 Greenberg et al. (2013)의 연구를 바탕으로 예측 변수를 설정했다. 또한 크라우드펀딩의 언어적 특징을 연구한 Wang et al. (2018)을 바탕으로 댓글의 길이(C_Length), 댓글 개수(Comment), 댓글의 평균 감성점수(C_Senti) 등을 포함하였고, 와디즈만의 고유 변수인 Facebook 지지서명(Like)과 펀딩 참여 건수(Funding)를 추가로 예측 변수로 활용하였다. 여기서 프로젝트 창설자의 댓글 응답 시간 변수와 관련하여, 커뮤니티 게시판에 문의 댓글을 남겼을 때 메이커의 답변 속도를 분 단위로 측정하여 프로젝트별 평균 응답속도를 측정했다. 답변이 달리지 않은 프로젝트에 대해서는 최댓값보다 1분 더 높게 변수를 설정하여 응답속도가 0으로 측정되는 것을 방지했다. 또한, 댓글의 문의 길이(C_Length), 감성점수(C_Senti)를 프로젝트별 평균으로 계산해 각각 변수로 지정했다(〈표 2〉 참조).

3.2 연구 절차

본 연구의 연구 절차는 〈그림 1〉과 같다. 우선, Python을 이용하여 와디즈의 데이터를 크롤링을

〈표 2〉 변수 설명

구분	변수명	설명	출처
펀딩 정보	Category	펀딩 카테고리	Greenberg et al.(2013)
	Term	펀딩 기간	
	Month	펀딩 시작 월 (1월, 2월,…,12월)	
	Video	펀딩 소개 내 동영상 개수	
	Image	펀딩 소개 내 사진 개수	
	Amount	목표 펀딩 금액	
	Reward	리워드 개수	
	Like	하트(좋아요) 개수	
	New	새소식 개수	
댓글 정보	Comment	펀딩 시작 부터 7일 이내 댓글의 개수	Wang et al., (2018)
	C _Length	펀딩 시작 부터 7일 이내 댓글의 평균 길이	
	C_Senti	펀딩 시작 부터 7일 이내 댓글의 평균 감성점수	
	Reply_avg	창설자의 댓글 응답 시간(분)	
기타	Facebook	펀딩 시작 부터 7일 이내 페이스북 공유 건수	본 연구에서 추가
	Funding	펀딩 시작 부터 7일 이내 펀딩 건수	
목표 변수	Target	펀딩 성공 여부 (1: 성공, 0: 실패)	Greenberg et al.(2013)



〈그림 1〉 연구 절차

통해 수집한다 ①. 수집된 데이터 전처리 작업을 거친 후, 변수 가공 과정을 통해 변수 추출, 파생 변수 생성을 수행한다 ②. 그리고 추가로 댓글 감성 분석을 수행하고 그 결과를 예측 변수로 활용하여 펀딩 프로젝트의 결과를 예측하고자 한다. 여기서 감성분석은 소셜미디어 텍스트 마이닝을 위한 통합 애플리케이션인 Korean Natural Language Application (KoALA)[2]를 사용하여 명사의 감성분석을 하였다. KoALA는 한글 텍스트 마이닝에 특화된 애플리케이션으로, 학계뿐 아니라 산업계에서도 활발하게 활용되고 있다. KoALA는 자체 구축한 감성사전을 바탕으로 텍스트 데이터의 감성수준을 점수로 측정해 준다.

변수 선택 과정에서는 로지스틱 회귀분석을 이용하여 변수의 타당성을 검증하였으며 ③, 예측 모델링 과정을 반복하면서 모델의 설명력을 높이는 변수 집합을 찾도록 하였다 ④. 예측모델링 과정에서는 Boosting기법이 적용되지 않은 MLP, Support Vector Machine, Decision Tree, Naive Bayes 등에는 균일화 작업을 이룬 데이터를 이용하여 학습시켜 예측 모델링을 구축하였다. Boosting 기법이 이용된 AdaBoost, Random Forest, Gradient Boosting의 경우에는 Upsize sampling을 적용하지 않은 데이터를 이용하여 예측모델링을 구축하였다. 다음으로 모델 평가 과정과 하이퍼파라미터를 변화시켜가면서 최적의 모델을 찾기 위한 시뮬레이션을 수행하고 ⑤, 마지막으로 5가지 평가척도를 이용하여 모델을 선택하는 작업으로 진행하였다 ⑥.

3.3 데이터 처리 및 모델 성능 평가

본 연구에서 활용할 데이터 세트에서 목표 변수의 분포가 균일하지 않아서 oversize sampling 기법 중에서 많이 사용되는 Synthetic Minority Over-sampling Technique(SMOTE) 기법을 이용하여 균일화 작업을 진행하였다(Chawla et al., 2002). SMOTE는 비율이 낮은 분류의 데이터를 만들어 내는 방법으로 먼저 분류 개수가 적은 쪽의 데이터

2) KoALA, https://www.koala4text.com

의 샘플을 취한 뒤, 이 샘플의 k 최근접 이웃(k neighbor)을 찾고, 이 차이에 0~1 사이의 임의의 값을 곱하여 만든 새로운 샘플을 훈련데이터에 추가한다(Chawla, 2009). 결과적으로 기존의 샘플을 주변의 이웃을 고려해 약간씩 이동시킨 점들을 추가하는 방식으로 동작한다.

모델 성능 평가는 다음과 같이 진행했다. 일반적으로 머신러닝 모델은 손실 함수(Loss Function) 최적화 기준으로 성능 평가 지표를 선정한다. 클래스의 분포가 균일한 경우에는 정확도(Accuracy)와 ROC(Receiver-Operating Characteristic curve), AUC(Area Under Curve)를 선정한다. 하지만 모집단에 있는 객체가 비정상적이거나 관심 있는 계층에 속한 객체 수가 매우 적으면 계층 편중 현상이 발생하게 된다(Ezawa et al., 1996; Fawcett and Provost, 1996; Japkowicz and Stephen, 2002). 계층 편중 현상이 발생한 경우에는 정확도에 기반한 평가는 실효성이 없다(Provost and Fawcett, 2013). 이 경우 다수 계층에 대한 비율이 매우 높으므로, 정확도가 높다고 해서 머신러닝을 통한 실제 목적을 달성했다고 보기 어렵다.

Chawla(2009)에 따르면 목표 변수가 불균형을 이루고 머신러닝 사용 목표가 성공 예측 예측이면 정밀도(Precision)를 평가척도로 이용하는 것을 제안하였다. 이에 본 연구에서는 정밀도를 모델의 평가척도로 두고자 한다. 그리고 정밀도를 기준으로 K-중첩 교차분석(K-fold cross validation)을 이용하여 개발한 머신러닝 모델을 평가 및 검증한다(Moreno-Torres et al., 2012). K-중첩 교차분석 방법은 수집된 데이터의 크기가 비슷한 K개의 부분 집합(D_1,D_2, … ,D_k)으로 분할 후, 이들 부분 집합을 이용하여 학습 및 검정을 K번 반복 수행하는 검증 방법이다. 즉, 첫 번째 반복에서는 D_1을 제외한 나머지 부분 집합 D_2, … ,D_k를 이용하여 학습을 수행하고, 학습된 모델에 D_1을 적용하여 모델을 평가한다.

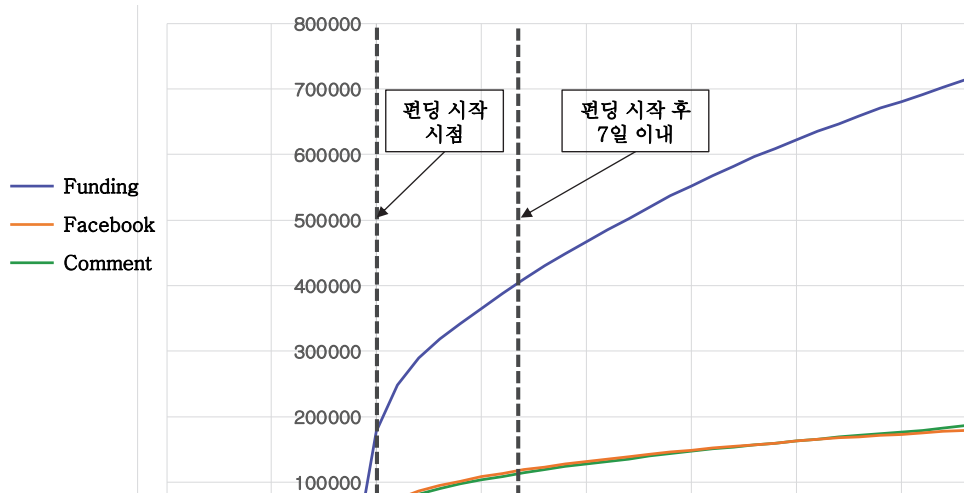
Ⅳ. 예측 모델링 및 평가

4.1 분석 결과

4.1.1 탐색적 데이터 분석

본 연구는 보상형 크라우드펀딩 성공 여부에 영향을 주는 요인을 파악하기 위해서 먼저 수집한 데이터를 바탕으로 탐색적 데이터 분석을 하였다. 수집된 데이터 바탕으로 분석 결과, 프로젝트 중 72.92%(4,305)가 성공하고 27.18%(1,607)가 실패한 것을 확인할 수 있었다. 월별 펀딩 건수를 본 결과, 1월은 641건, 2월은 380건, 3월은 300건, 4월은 434건, 5월은 374건, 6월은 428건, 7월은 455건, 8월은 540건, 9월은 430건, 10월은 567건, 11월은 710건, 12월은 653건으로 11월에 가장 많은 펀딩이 시작되었고 3월에 300건으로 가장 적은 펀딩이 시작되었음을 알 수 있었다. 또한, 크라우드펀딩의 카테고리별 분포를 분석해본 결과, 패션잡화가 1,194건으로 가장 많이 진행되었으며, 게임/취미는 69건, 스포츠는 151건, 교육/키즈는 163건, 공연/컬쳐는 248건, 반려동물은 283건, 뷰티는 399건, 리빙은 496건, 소셜/캠페인은 667건, 디자인소품은 804건으로 나타났다.

다음으로 펀딩 초기이면서 예측 정확도를 높일 수 있는 최적의 데이터 수집 기간을 탐색해보았다. 이를 위해 펀딩 참여 건수(Funding), 페이스북 지지서명(Facebook), 댓글 수(Comment)를 프로젝트 시작일부터 평균 종료 일자인 28.3일 동안의 일자별 비중으로 분석해 보았다. 그 결과, 펀딩 시작 시점에 펀딩 참여 건수는 16%, Facebook 지지서명



〈그림 2〉 프로젝트 누적 서포트, 지지서명, 댓글 분포

은 28%, 댓글은 26% 정도이다. 그리고 펀딩 시작 후 7일 이내에는 펀딩 참여 건수가 54%, Facebook 지지서명이 63%, 댓글이 58%의 분포를 보인 후 분포 증가가 완만해지는 것을 확인했다(〈그림 2〉 참조). 즉, 본 연구의 주요 예측 변수 데이터의 50% 이상이 펀딩 시작 후 7일 이내 수집되는 것을 확인하여 최적의 데이터 수집기간으로 판단되었다. 이에 우리는 전체 데이터 중에 펀딩 시작 후 7일 이내의 펀딩 참여 건수, Facebook 지지서명과 댓글 각각 387,602건, 113,614개, 109,199개를 분석에 이용했다.

〈표 3〉는 수치형 변수의 기술 통계량을 나타낸다. 댓글의 평균 길이(C_Length)는 77.5자이고, 댓글의 평균 감성 점수(C_Senti)는 0으로 중립을 보였다. 메이커의 댓글 응답속도(Reply_avg)는 454분이었다. 펀딩 시작 후 7일 이내의 댓글 개수(Comment)는 평균 84개를 포함하며 새소식(New)은 평균 5개, 페이스북 지지서명(Facebook)은 15개, 펀딩 참여횟수(Funding)는 83회인 것으로 나타났다. 펀딩 1개당 평균적으로 좋아요(Like). 수는 87개, 펀딩 지속기간(Term)은 28일, 리워드 개수(Reward)는 6개, 사진의 개수(Image)는 31개이고, 1개 이상의 동영상(Video)을 포함하고 있는 것을 알 수 있었다. 또한, 기술 통계량 분석 결과, 좋아요(Like), 댓글 개수(Comment), 댓글 응답속도(Reply_avg), 펀딩 참여횟수(Funding)의 표준 편차가 상대적으로 다른 변수에 비해 큰 것을 확인했다.

수집된 데이터를 기반으로 다음과 같이 데이터 전처리를 수행했다. 먼저 명목형 변수 데이터인 펀딩 카테고리(Category), 펀딩 시작 월(Month)을 원핫인코딩(One-hot-encoding)을 통해 더미 변수로 변환하였다. 또한, 수치형 데이터는 데이터 스케일을 맞춰주기 위하여 모든 값을 0에서 1사이로 Min-Max 스케일링하였다.

전처리 후, 예측 변수 간의 상관관계도 살펴보았다. 〈표 4〉을 통해 알 수 있듯, 댓글 길이(C_length)는 댓글 감성(C_Senti)과 음의 상관관계를 보이고, 새소식 개수(New)와 양의 상관관계를 보였다. 이는 커뮤니티 상에서 댓글 길이가 긴 수록 부정적인 댓글일 가능성이 있으며, 메이커는 부정적인 여론에 대응하기 위해 새소식을 올린다는 해석을 할 수 있다. 그리고 좋아요(Like)와 댓글의 개수(Comment)

〈표 3〉 기술 통계량

	Mean	St.d.	Min	Median	Max
Term	28.30	14.56	0.00	28.00	691.00
Video	1.12	1.48	0.00	1.00	15.00
Image	31.57	19.60	0.00	28.00	265.00
Amount	14.42	0.91	1.61	14.32	18.42
Reward	6.18	3.90	0.00	6.00	88.00
Like	87.92	193.07	0.00	39.00	5251.00
New	5.05	5.56	0.00	4.00	67.00
Comment	84.57	218.66	1.00	35.00	6,434.00
C_length	77.50	40.36	0.00	72.19	478.00
C_Senti	0.00	0.02	-0.20	0.00	00.25
Reply_avg	454.06	307.02	0.07	400.26	1,439.00
Facebook	15.54	35.34	0.00	6.00	1,055.00
Funding	83.03	196.07	0.00	23.00	5,061.00

〈표 4〉 상관관계 결과

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
Term (V1)	-												
Video (V2)	0.04	-											
Image (V3)	-0.11	0.12	-										
Amount (V4)	0.21	0.13	-0.06	-									
Reward (V5)	0.02	0.09	0.34	-0.05	-								
Like (V6)	0.00	0.17	0.22	0.05	0.08	-							
New (V7)	0.09	0.19	0.23	0.09	0.16	0.37	-						
Comment (V8)	0.10	0.18	0.15	0.15	0.08	0.73	0.42	-					
C_length (V9)	-0.11	0.06	0.19	-0.14	0.09	0.19	0.15	0.09	-				
C_Senti (V10)	0.06	-0.03	-0.09	0.03	-0.04	-0.10	-0.07	-0.05	0.09	-			
Reply_avg (V11)	0.06	-0.06	-0.17	0.01	-0.10	-0.09	-0.20	-0.09	-0.1	0.00	-		
Facebook (V12)	0.02	0.05	0.04	0.03	0.05	0.26	0.12	0.25	0.01	-0.01	-0.05	-	
Funding (V13)	-0.02	0.15	0.17	0.06	0.08	0.73	0.35	0.60	0.17	-0.09	-0.08	-0.05	-

가 0.73으로 높은 상관관계를 보이는 것으로 확인되었다. 댓글을 통해 펀딩 관련 자기 생각과 감정을 올린 서포터가 또 다른 표현 방법으로 '좋아요'를 눌렀을 가능성이 있다. 그리고 리워드의 개수(Reward)와 사진(Image)이 양의 상관관계를 보이는 것으로 보아 리워드가 많은 펀딩의 경우에는 각 리워드를 설명하기 위한 사진의 개수가 많은 것을 의미한다.

4.1.2 로지스틱 회귀 분석

본 연구에서는 크라우드펀딩 여부에 영향을 주는 변수들의 통계적 유의성을 파악하기 위해 로지스틱 회귀분석을 수행하였다(Desai et al., 2015; Greenberg et al., 2013). 로지스틱 회귀모형의 독립변수는 n개이며, 종속 변수는 1개로 0과 1 두 값만을 가진다. 독립변수 X들에 의해서 Z값이 변화하고 여기서 Z는 최종적으로 Event가 일어날 확률, 즉 Prob(Event)에 영향을 주는 지수로서 역할을 한다. 따라서 본 연구는 종속 변수를 펀딩 성공(=1), 실패(=0)로 나누고 로지스틱 회귀 분석을 적용하여 통계적으로 유의미한 예측 변수와 유의미하지 않은 변수를 파악했다.

분석 결과(〈표 5〉 참조), 펀딩의 성공에 통계적으로 유의미하게 긍정적인 영향을 주는 변수는 댓글의 평균 감성(C_Senti), 댓글의 개수(Comment), 새 소식 건수(New), Facebook 지지서명(Facebook), 펀딩 참여 건수(Funding), 펀딩 좋아요 개수(Like), 리워드 개수(Reward), 비디오 개수(Video)로 파악되었다. 펀딩 성공에 부정적인 영향을 주는 유의미한 변수는 목표 펀딩금액(Amount), 펀딩 기간(Term)인 것으로 확인되었다. 유의미하지 않은 예측 변수로는 댓글의 길이(C_length), 창설자의 댓글 응답속도(Reply_Avg), 사진의 총 개수(Image)가 확인되었다. 이에 본 연구에서는 펀딩의 성공에 영향을 미치는 변수로 예측 모델링을 개발하였다.

4.2 예측 모델링 개발

예측 모델링을 개발하기 위해 훈련데이터와 테스트데이터는 무작위로 7:3 비율로 나누되, 성공 프로

〈표 5〉 로지스틱 회귀 분석 결과

	coef	Std. err	z	P〉|z|
Intercept	14.75	0.81	18.15	0.00
Term	-0.02	0.00	-5.66	0.00
Video	0.08	0.03	2.26	0.02
Image	0.00	0.00	0.08	(0.94)
Amount	-1.13	0.06	-19.59	0.00
Reward	0.03	0.01	2.47	0.01
Like	0.02	0.00	11.39	0.00
New	0.39	0.02	18.76	0.00
Comment	0.01	0.00	5.34	0.00
C_length	0.00	0.00	0.52	(0.60)
C_Senti	4.14	2.03	2.04	0.04
Reply_avg	0.00	0.00	1.48	(0.14)
Facebook	0.01	0.00	2.30	0.02

젝트 비율(72.82%)과 실패 프로젝트 비율(27.18%)을 훈련데이터와 테스트데이터에 비슷하게 적용하였다. 훈련데이터는 K-중첩 교차분석(K-fold cross validation)을 통해 모델 평가 및 검증했다(Moreno-Torres et al., 2012). 데이터 사이즈를 고려해 본 연구에서는 5-중첩교차분석을 수행하였다. 5-중첩 교차분석을 통해 각 머신러닝의 최적의 하이퍼 파라미터(Hyperparameter)를 찾았다. 모델링은 Python의 머신러닝 모듈인 Scikit-learn(Pedregosa et al., 2011) 을 사용하였고, Grid Search를 이용하여 조정할 수 있는 각각의 하이퍼파라미터를 탐색했다.

테스트 결과, Decision Tree의 경우에는 Entropy를 결정지수로 사용하고 나무의 최대 깊이를 7로 설정했을 때 가장 좋은 성능을 보였다. Naive Bayes의 경우 Gaussian Naive Bayes 기법을 사용했을 때 높은 성능을 보이는 것을 확인했다. SVM의 경우에는 Penalty가 0.1, 커널 함수는 선형을 이용했을 때 가장 높은 성능을 보였다. Adaboost의 경우에는 estimator가 50개 learning rate이 0.5일 때 가장 높은 성능을 보였다. Gradient Boosting의 경우에는 learning rate이 0.5, estimator가 64개, 최대 깊이 18, minimum samples splits이 0.1, minimum samples leaf가 0.1, max feature가 5일 때 가장 높은 성능을 보였다. Random Forest의 경우에는 estimator가 50개, Entropy 결정지수, 최대 깊이 3으로 설정했을 때 좋은 성능을 보였다. 마지막으로 MLP의 경우 batch 크기는 0.5, epochs는 100으로 설정하고, 전체 네트워크는 입력층, 은닉층 2단, 출력층으로 구성되었다. 각 은닉층의 신경(neuron)은 34, 16일 때 최적인 것으로 확인했다. 각 층의 활성 함수(Activation Function)는 sigmoid를 사용하고, 옵티마이자(Optimizer)는 Adedelt를 이용하는 것이 가장 높은 성능을 보였다.

4.3 모델 검증 및 성능 평가

머신러닝 모델의 평가척도를 구하기 위해 혼동행렬을 사용하였다. 혼동행렬은 예측 값(Positive, Negative)과 실제 값(True, False)을 표로 만든 것이다. 본 연구에서 TP(True Positive)은 펀딩에 성공할 프로젝트에 성공할 것이라고 맞게 예측한 것이고, FP(False Positive)은 펀딩에 성공할 프로젝트를 실패할 것이라고 틀리게 예측하는 것이다. FN(False Negative)은 펀딩에 실패할 프로젝트를 성공할 것이라고 틀리게 예측한 것이고, TP(True Positive)는 펀딩에 실패할 프로젝트에 실패할 것이라고 맞게 예측한 것이다. 혼동행렬를 이용하여 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F-값, 민감도(Sensitivity), 특이도(Specificity)를 구할 수 있다(Swets, 1988).

정확도는 전체 결과 중에서 맞게 분류한 경우를

〈표 6〉 모델 성능 평가

	Accuracy (%)	Precision	Recall	F1-score	AUC
Decision Tree	85.17%	0.93	0.87	0.9	0.84
SVM	86.02%	**0.95**	0.85	0.9	**0.87**
Naïve Bayes	87.77%	0.91	0.93	0.92	0.84
Random Forest	88.34%	0.92	0.92	0.92	0.85
AdaBoost	89.57%	0.92	0.94	**0.93**	0.86
Gradient Boost	**90.30%**	0.92	**0.95**	**0.93**	**0.87**

나타낸다. 식으로 나타내면 TN + TP / TN + FP + FN + TP이다. 텍스트를 분류하고 정보를 검색하는 경우에는 정밀도와 재현율 척도가 자주 사용된다. 재현율은 TP / TP + FN, 정밀도는 TP / TP + FP로 나타낼 수 있다. 재현율과 정밀도의 조화 평균인 F-값을 이용해 모델을 평가하기도 한다. 통계분석, 패턴 인식 분야에서 민감도와 특이도가 자주 사용된다(Provost and Fawcett, 2013). 민감도는 재현율과 같으며, 특이도는 TN / TN + FP로 나타낼 수 있다. 마지막으로 민감도와 특이도가 어떤 관계가 있는지 표현한 그래프인 ROC Curve (Receiver-Operating Characteristic curve)와 ROC Curve 아래의 면적을 나타내는 AUC(Area Under Curve)도 있다.

〈표 6〉은 각 모델에 대한 평가척도: 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-값(F1-score), AUC 값을 나타낸다. 연구결과, 정확도 측면(Kuhn and Johnson, 2013) 에서는 Gradient Boosting이 가장 높은 성능(90.30%)을 보였다. 하지만 본 연구는 성공과 실패가 불균형 분포를 이루고 있어서 정확도 보다 정밀도로 평가하는 것이 더 적합하므로(Chawla, 2009), 성공을 더 잘 예측하는 모델을 찾는 척도인 정밀도를 기준으로 하면 SVM이 가장 높은 성능(0.95)을 보였다. 즉, SVM이 선형 함수를 커널로 갖고, 페널티는 0.1일 때 와디즈의 펀딩 프로젝트의 성공을 가장 잘 예측한다고 볼 수 있다. Decision Tree의 경우 정밀도가 0.93으로 SVM 다음으로 성공을 잘 예측했다. 또한, 간단한 규칙으로 분류 문제를 잘 해결하는 기법으로 알려진 Navie Bayes의 경우에도 Precision이 0.91로 비교적 높은 성능을 갖고 있다. Boosting 기법 중 하나인 Adaboost는 89.57%의 높은 정확도를 보였다. 반면 딥러닝 알고리즘 중 하나인 MLP는 은닉층을 2단으로 설계하고 다양한 파라미터를 변형시키면서 검증하였지만 다른 알고리즘에 비해 좋지 못한 결과를 얻었다.

Ⅴ. 토의 및 시사점

5.1 결과 토의 및 연구 한계

본 연구에서는 우선 로지스틱 회귀분석을 통해 크라우드펀딩 성공에 대한 예측 변수들의 통계적 유의성을 확인하였다. 분석 결과, 펀딩 성공에 긍정적으로 유의미한 영향을 주는 변수는 댓글의 평균 감성점수(C_Senti), 댓글의 개수(Comment), 새소식 건수(New), Facebook 지지서명(Facebook), 펀딩 참여 건수(Funding), 펀딩 좋아요 개수(Like), 리워드 개수(Reward), 비디오 개수(Video)로 파악되었다. 반면, 펀딩 성공에 부정적으로 유의미한 영향을 주는 변수는 목표 펀딩금액(Amount), 펀딩 기간(Term)인 것으로 확인되었다. 해당 변수들이 부정적인 영향을 주는 이유로는 목표 펀딩금액이 너무 높은 경우에는 프로젝트를 성공하기 어렵고, 프로젝트 지속기간이 긴 경우에는 프로젝트에 언제든 참여할 수 있다고 생각해 투자자들의 적극적인 참여도를 증가시키기 어려워 성공에 부정적인 영향을 미친 것으로 보인다.

유의미하지 않은 예측 변수로는 기존 연구에서 유의했던 댓글의 길이, 창설자의 댓글 응답속도(Wang et al., 2018), 사진의 개수(Greenberg et al., 2013; Yuan et al., 2016) 가 본 연구에서는 유의하지 않은 것으로 나타났다. 댓글의 길이의 경우, 와디즈에서 투자자 댓글은 자동으로 Facebook 지지서명 확인하는 링크를 포함되어 유의미한 투자자 의견이 부족한 것을 이유로 들 수 있다. 창설자의 댓글 응답속도(reply_avg)는 고객관리 차원에서 대부분 창설자가 빠른 속도로 응답하기 때문에 성공과

실패에 유의한 영향을 주지 않은 것이라고 해석할 수 있다. 마지막으로 사진의 총 개수(image)는 대부분의 펀딩이 글보다는 이미지로 프로젝트를 설명하고 있으므로 유의하지 않은 것으로 보인다.

또한, 본 연구는 펀딩 프로젝트 초기에 생성된 데이터만을 가지고 예측 모델링을 개발하였다. 본 연구에서는 7일 이내의 커뮤니티의 댓글의 감성점수, Facebook 지지서명 횟수, 펀딩 참여 횟수 등을 처리하여 예측 변수로 이용하며, 예측 모델링 기법으로는 Decision Tree, Naive Bayes, Support Vector Machine, Adaboost, Gradient Boosting, Random Forest, Multi-Layer Perception을 사용했다. 그중에서 Support Vector Machine이, 선형 커널 함수를 이용했을 때 정밀도가 0. 95로 가장 높은 성능을 보였다. 추가로 기존 연구(Sawhney et al., 2016; Yu et al., 2018; 이강희 et al., 2018)에서 사용하지 않은 기법인 Gradient Boosting을 사용하여 정확도 측면에서는 90.30%로 높은 성과를 얻었다. 기존 연구들과 예측 모델에 대한 절대적인 비교는 불가능하지만, 7일 이내의 예측 변수를 활용하여 90%가 넘는 예측력을 보여 좋은 성과를 얻었다고 할 수 있다.

하지만, 본 연구는 실제 7일 시점까지 알 수 있는 정보뿐 아니라 새 소식 개수, 좋아요 수 등 펀딩이 끝난 시점의 데이터를 포함하고 있어 완벽하게 7일 시점을 대변하지는 못한다는 한계점을 가지고 있다. 그러나 Greenberg et al.,(2013) 의 연구도 펀딩 시작 시점에 알 수 있는 정보를 이용하여 SVM으로 64%의 모델 성능을 보였고, 총 펀딩 횟수와 같은 펀딩이 끝난 시점의 데이터를 포함해야 정확도가 크게 상승하는 것을 확인하였다. 따라서 본 연구도 모든 예측 변수가 7일 시점을 대변하지 못한다는 한계를 갖고 있지만, 와디즈만의 새로운 예측 변수인 'Facebook 지지서명' 변수 등을 포함하고 0.9 이상의 정밀도를 가진 예측 모델을 개발했다는 데 의의가 있다. 또한, 본 연구에서는 변수 간의 상호작용 또는 두 개 이상의 입력 특성을 곱한 교차 곱에 대한 검토가 부족하였다. 특히, 본 연구에서 활용된 댓글의 평균 감성 점수(C_Senti), Facebook 지지서명(Facebook) 변수 간 상호작용에 관한 연구가 추가로 진행되면 확장된 인사이트를 얻을 수 있을 것이다.

향후 연구에서는 멀티미디어 정보를 적극적으로 활용하는 연구를 진행할 수 있다고 기대한다. 펀딩 프로젝트는 가독성을 위해 글보다는 사진에 글을 포함하는 것이 특징이다. Image Recognition API를 이용해 사진에서 글자를 추출하는 방법을 이용하여 언어적인 특징을 분석할 수 있을 것이다. 또한, 동영상의 정보와 섬네일(Thumbnail)의 정보를 이용하여 펀딩 프로젝트 관련 추가 정보를 추출할 수 있다. 또한, 향후 연구에서 국내외 다양한 크라우드펀딩 플랫폼들 특성을 파악하여 플랫폼별 예측 모델링을 개발하는 것을 제안한다.

5.2 연구의 학술적/실무적 의의

본 연구가 갖는 학술적 시사점은 다음과 같다. 첫 번째로 와디즈의 펀딩 프로젝트에 영향을 주는 요인들을 로지스틱 회귀분석을 사용하여 유의한 변수들을 파악하고, 이를 와디즈를 기준으로 해석하고 모델링에 적용하였다는 점에서 의의가 있다. 펀딩 성공에 유의미하지 않은 변수는 창설자의 댓글 응답속도, 사진의 개수이다. 유의한 변수로는 댓글의 평균 감성, 댓글의 개수, 새소식 건수, 펀딩 횟수, 프로젝트 좋아요 수, 리워드 개수, 비디오 개수이다. 펀딩 성공에 부정적인 영향을 주는 변수는 목표 펀딩금액, 프로젝트 지속기간인 것으로 확인되었다. 추가로 와디즈에서만 제공하는 기능인 'Facebook 지지서명'이 펀딩 성공 여부에 유의한 영향을 보이는 것을 증명했다. 즉, Facebook 지지서명이 플랫폼 내에서의 긍정적인 효과뿐 아니라, 소셜 네트워크상의

구전 효과와 네트워크 효과에 긍정적인 영향을 미쳤음을 의미한다.

두 번째, 프로젝트 초기 데이터를 가지고 예측 모델을 구축한 점이다. 선행 연구(Sawhney et al., 2016; Yu et al., 2018; Yuan et al., 2016)에서는 펀딩이 끝난 시점의 데이터를 활용하여 예측 모델을 구축했다. 반면에 본 연구에서는 프로젝트 초기이면서 예측 정확도를 높일 수 있는 최적의 날짜를 찾고자 노력하였다. 그 결과 펀딩 초기인 7일 이내에 커뮤니티의 댓글 수, Facebook 지지서명 횟수, 펀딩 참여횟수가 50%가 넘는 것을 확인했다. 즉, 펀딩 프로젝트가 시작하고 7일정도가 지나면 펀딩 성공 여부를 예측할 수 있다는 것을 실증적으로 증명하였다.

세 번째, 다양한 머신러닝 알고리즘을 적용해 높은 성능의 예측 모델을 만든 점이다. 본 연구에서는 Decision Tree, Naive Bayes, Support Vector Machine, Adaboost, Random Forest, Multi-Layer Perceptron을 이용해서 예측했다. 연구결과 Support Vector Machine이 선형 커널 함수를 이용했을 때 정밀도가 0.95로 가장 높은 성능을 보였다. 또한, 기존 연구에서 활용되지 않았던 Gradient Boosting 기법을 사용해 90%의 정확도로 높은 성능의 모델을 개발하였다.

최근, 스타트업 혹은 대기업에서도 신제품을 출시할 때 자금조달, 시장성 검증을 위해 크라우드펀딩을 이용하고 있다. 특히, 크라우드펀딩이 성공했을 때에는 홍보 효과를 누릴 수 있다는 점에서 프로젝트에 성공하는 것은 비즈니스적으로 중요하다. 이에 본 연구는 펀딩 프로젝트 결과를 초기에 예측하는 모델을 구축했다는 점에서 의의가 있다. 추가로 크라우드펀딩 참여자 별 실무적 시사점은 다음과 같다. 투자자 관점에서 펀딩 성공과 실패를 예측할 수 있어 시간적 기회비용을 낮출 수 있을 것이다. 창설자 측면에서는 본 연구의 결과를 활용하여 프로젝트의 성공과 실패에 대한 확률을 미리 파악한다면 실패할 프로젝트에 성공할 수 있도록 끌어낼 기회를 가질 수 있을 것이다. 그뿐만 아니라 성공하는 프로젝트에 대해서도 더 많은 참여를 촉구하기 위한 방향을 제안할 수도 있으리라 생각된다. 최근, 마지막으로 플랫폼 측면에서는 크라우드펀딩 프로젝트의 성공과 실패에 직접적인 관여를 하지는 않지만, 펀딩 초기에 예측할 수 있어 프로젝트의 성공과 실패 확률을 창설자에게 제공할 수 있을 것이다.

REFERENCES

Ambani, P.(2017), "Crowdsourcing new tools to start lean and succeed in entrepreneurship: Entrepreneurship in the crowd economy," *In Crowdfunding for Sustainable Entrepreneurship and Innovation* 37-53. IGI Global.

Bi, S., Liu, Z., and Usman, K.(2017), "The influence of online information on investing decisions of reward-based crowdfunding," *Journal of Business Research*, 71, 10-18.

Breiman, L.(2001), "Random forests." *Machine Learning*, 45(1), 5-32.

Chawla, N. V. (2009). "Data mining for imbalanced datasets: An overview," *In Data Mining and Knowledge Discovery Handbook*, 875-886. Springer.

Chen, T., and Guestrin, C.(2016), "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785-794.

Cortes, C., and Vapnik, V.(1995), "Support-vector networks," *Machine Learning*, 20(3), 273-297.

Cumming, G.(2014), "The new statistics: Why and how," *Psychological Science*, 25(1), 7-29.

Desai, N., Gupta, R., and Truong, K.(2015), "Plead or pitch? The role of language in kickstarter project success," *Technical Report*, Stanford University.

Ezawa, K. J., Singh, M., and Norton, S. W.(1996), "Learning goal oriented Bayesian networks for telecommunications risk management," *ICML*, 139-147.

Fawcett, T., and Provost, F. J.(1996), "Combining Data Mining and Machine Learning for Effective User Profiling," *KDD*, 8-13.

Friedman, J. H.(2001), "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, 1189-1232.

Greenberg, M. D., Pardo, B., Hariharan, K., and Gerber, E.(2013), "Crowdfunding support tools: Predicting success and failure," *In CHI'13 Extended Abstracts on Human Factors in Computing Systems*, 1815-1820.

Haas, P., Blohm, I., and Leimeister, J. M.(2014), *An Empirical Taxonomy of Crowdfunding intermediaries*.

Japkowicz, N., and Stephen, S.(2002), "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, 6(5), 429-449.

Kamath, R. S., and Kamat, R. K.(2016), "Supervised learning model for kickstarter campaigns with R mining," *International Journal of Information Technology, Modeling and Computing (IJITMC)*, 4(1), 19-30.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.(2017), "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, 3146-3154.

Kuhn, M., and Johnson, K.(2013), *Applied Predictive Modeling* (Vol. 26). Springer.

Li, Y., Rakesh, V., and Reddy, C. K. (2016), "Project success prediction in crowdfunding environments," *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 247-256.

Moreno-Torres, J. G., Saez, J. A., and Herrera, F.(2012), Study on the impact of partition-induced dataset shift on k-fold cross-validation. IEEE Transactions on Neural Networks and Learning Systems, 23(8), 1304-1312.

Nam, S., Jin, Y., and Kwon, O.(2018), "Online Document Mining Approach to Predicting Crowdfunding Success," *Journal of Intelligence and Information Systems*, 24(3), 45-66.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V.(2011), "Scikit-learn: Machine learning in Python," *The Journal of Machine Learning Research*, 12, 2825-2830.

Provost, F., and Fawcett, T.(2013), *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

Rosenblatt, F.(1961), *Principles of neurodynamics. Perceptrons and the theory of brain mechanisms*. Cornell Aeronautical Lab Inc Buffalo NY.

Sawhney, K., Tran, C., and Tuason, R.,(2016), *Using Language to Predict Kickstarter Success*. Stanford University: Stanford, CA, USA.

Scholkopf, B., Smola, A., and Muller, K.-R.(1998), "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, 10(5), 1299-1319.

Schwienbacher, A., and Larralde, B.(2010), "Crowdfunding of small entrepreneurial ventures," *Handbook of Entrepreneurial Finance*, Oxford University Press, Forthcoming.

Statista.(2018), Crowdfunding-Statistics and Facts | Statista. https://www.statista.com/topics/1283/crowdfunding/

Swets, J. A.(1988), "Measuring the accuracy of diagnostic systems," *Science*, 240(4857), 1285-1293.

Wang, N., Li, Q., Liang, H., Ye, T., and Ge, S.(2018), "Understanding the importance of interaction between creators and backers in crowdfunding success," *Electronic Commerce Research and Applications*, 27, 106-117.

Wikipedia.(2020), Pebble (watch). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Pebble_(watch)&oldid=954537348

Yu, P.-F., Huang, F.-M., Yang, C., Liu, Y.-H., Li, Z.-Y., and Tsai, C.-H.(2018), "Prediction of Crowdfunding Project Success with Deep Learning," *2018 IEEE 15th International Conference on E-Business Engineering (ICEBE)*, 1-8.

Yuan, H., Lau, R. Y., and Xu, W.(2016), "The determinants of crowdfunding success: A semantic text analytics approach," *Decision Support Systems*, 91, 67-76.

국내참고문헌

곽현, 이호근(2014), "크라우드펀딩 분야의 국내외 연구동향 분석," **정보화정책**, 21(4), 3-19.

권보람, 김주성(2013), "크라우드펀딩의 국내외 현황 및 입법 동향 분석," **전자통신동향분석**, 28(5).

권혁인, 이승하, 나윤빈(2014), "크라우드펀딩의 성공· 실패 사례분석: 커뮤니티 아트 프로젝트를 중심으로," **한국콘텐츠학회논문지**, 14(7), 125-136.

금융위원회(2019), "크라우드펀딩 주요 동향 및 향후 계획," **대한민국 정책브리핑**. 2019년 4월 11자 http://www.korea.kr/news/pressReleaseView.do?newsId=156326169

김병주(2017), 와디즈 플랫폼의 성공 사례를 통해 본 국내 크라우드펀딩 시장의 도전과 응전. 서울경제. 2017년 9월 11자. http://www.sedaily.com/NewsView/1OKZILAZ3P/GC1301

김성진, 안현철(2016), "기업신용등급 예측을 위한 랜덤 포레스트의 응용," **산업혁신연구**, 32(1), 187-211.

김재일, 박상철, 홍수지(2019), "크라우드소싱 플랫폼 창업 사례 연구: 브로스앤컴퍼니를 중심으로," **Korea Business Review**, 23(1), 29-56.

박규석(2019), "금융위, 크라우드펀딩 자금 조달 규모 755억 달성," CEOSCORE, 2019년 4월 11일자 http://www.ceoscoredaily.com/news/article.html?no=54176

윤경희(2019), "선주문 후제작…크라우드 펀딩에 줄 선다," 중앙일보, 2019년 9월 17일자 중앙일보, https://news.joins.com/article/23578728

이강희, 이승훈, 김현철(2018), "멀티미디어 및 언어적 특성을 활용한 크라우드펀딩 캠페인의 성공 여부 예측," **멀티미디어학회논문지**, 21(2), 281-288.

이정은, 신형덕(2014), "크라우드펀딩 사이트의 게시글 정보가 펀딩 성공에 미치는 영향," **한국콘텐츠학회논문지**, 14(6), 54-62.

크라우드넷(2020), "펀딩성공현황," 크라우드넷. https://www.crowdnet.or.kr/statistics/success_outline.jsp

A Machine Learning Approach for the Success Prediction of Reward Crowdfunding Project

Dong-Ji Moon* · Sang-Hyeak Yoon** · Soobin Choi*** · Hee-Woong Kim****

Abstract

Crowdfunding has been recently rising as financing channel and showed rapid growth by integrating with social media. As of 2018, global crowdfunding market size was estimated as $ 9.37 billion, and Korea crowdfunding market size was about $ 110 million. However, the probability of crowdfunding failure showed more than 38%, which gives huge burden for participants (i.e., makers, investors, platforms). So, to prevent the failure and protect participants from their loss of time and money, predicting the success of the funding in the early step is crucial. Therefore, this study aims to build a model to predict whether the crowdfunding project will success or fail. Compare to the previous studies that they used data after the end of crowdfunding, we collected data seven days before the project ends. We used data from crowdfunding site 'Wadiz', by collecting comment data and funding information as predict variable. Then we applied machine learning methods such as Decision Tree, Support Vector Machine, Naive Bayes, AdaBoost, Gradient Boosting, Random Forest, and MLP. As a result, Gradient Boosting showed more than 90% accuracy, and Support Vector Machine showed the highest precision score (0.95). Also, this study has a practical implication of predicting funding success in the early stage of crowdfunding by developing a prediction model based on machine learning.

Key Words: Crowdfunding, Wadiz, Predictive Modeling, Machine Learning, Sentiment Analysis

* Staff, CJ Logistics, 1st Author
** Doctor. Graduate School of Information, Yonsei University, Corresponding Author
*** Master, Graduate School of Information, Yonsei University, 2nd Author
**** Professor, Graduate School of Information, Yonsei University, 3rd Author