「정보시스템연구」 제29권 제3호
한국정보시스템학회
2020년 9월, pp. 35~52

http://dx.doi.org/10.5859/KAIS.2020.29.3.35

머신러닝을 활용한 지역축제 방문객 수 예측모형 개발*

이인지** · 윤현식***

〈목 차〉

Ⅰ. 서 론

1.1 연구 배경

각 지역의 지방자치단체는 지역축제 개최를 통해 다양한 경제적 목적과 사회적 목적을 달성한다. 장기간 많은 비용을 지출하여 관광 인프라를 구축하는 것과 달리, 지역축제는 비교적 짧은 시간 내에 적은 비용으로 개최할 수 있으면서 지역 내 경제를 활성화하는데 효과적이기 때문에 많은 지방자치단체의 관광 전략으로써 활용되고 있다(오남현, 2012). 사회적인 측면에서는 지역축제를 통해 지역의 전통문화를 계승하고, 지역의 특성을 반영한 관광상품을 개발할 수 있으며, 지역 주민의 화합을 도모할 수 있다(김철원, 이석호, 2001). 국가적 관점에서는 외국 관광객을 유입시키고 지역 간 불균형을 감소시키기 때문에, 중앙정부 역시 다양한 정책을 통해 지역축제를 지원하고 있다(신현식, 2010).

우리나라에서는 1970년대, 지역 향토문화의 전승과 홍보를 목적으로 각지의 지역축제가 본격적으로 등장하기 시작하였다. 1980년 이후에는 중앙정부 주도하에 시행된 지역문화 활성화 정책과 문화체육부의 신설로 지원이 증대되고, 지방자치제도가 실시되면서 지역축제의 수가 지속해서 증가하였다. 문화체육관광부의 공시에 의하면 2019년에 기획된 전국 지역축제의 수는 총 884개에 이르는 것으로 나타났다(문화체육부, 1996; 문화체육관광부, 2019).

각 지역에서는 지역축제를 효율적이고 조직적으로 운영하기 위하여, 개별 축제에 대한 위원회 또는 사무국을 두고 있으며, 문화체육관광

* 본 논문은 제1저자의 석사학위 논문을 수정 및 보완하여 작성하였음을 밝힙니다.
** 전남대학교 일반대학원 전자상거래협동과정, as_990406@naver.com (제1저자)
*** 전남대학교 경영대학 경영학부, Dr.Yoon@jnu.ac.kr (교신저자)

부는 ‘문화관광축제 제도’에 따라 해마다 지역 축제의 성과를 평가하고 단계별 육성 정책을 시행하고 있다(오훈성, 2013). 2012년의 통계자료에 따르면 문화관광축제로 지정된 45개 지역 축제는 약 3,000만 명에 이르는 관광객을 유치하여, 약 1조 7,800억에 이르는 경제적 파급효과를 거둔 것으로 파악되었다(이한성, 2015).

반면, 일각에서는 다수의 지역축제가 지나치게 수익을 추구하거나 차별화된 콘텐츠가 결여되고 시설이 미흡하다는 등의 비판이 제기되고 있다(이준엽, 2018). 지역의 특색을 살리지 못하고 일시적 성과에 치중한 지역축제는 축제 방문객을 만족시키지 못하고 주민으로부터 예산 낭비라는 낙인이 찍히게 되어, 결국에는 폐지 수순을 밟게 된다(김영대, 이선영, 이환수, 2018). 문화체육관광부는 이러한 지역축제의 질적 퇴보를 방지하기 위하여 ‘문화관광축제 일몰제’를 2010년부터 실시하고 있다(김학용, 권호종, 2017). 이 제도는 지역축제에 대하여 일정 기간을 설정하고, 기간 만료 후 심사를 진행하여 축제에 대한 지원을 종료하는 제도이다(오훈성, 2013). 중앙정부뿐 아니라 지역축제를 개최하는 각 지방자치단체 역시 해마다 자체 평가를 수행하고 있으며, 지역축제를 평가하는 가장 객관적이고 보편적인 지표는 축제 방문객 수이다(오훈성, 2011). 축제 방문객의 수는 축제의 성과와 경제적 효과성을 확인하는 데 활용되는 기초자료로, 방문객 수의 파악은 지역축제의 평가에 필수적인 과정이다(이희찬, 문혜선, 2010).

축제 이후 성과평가를 위해 방문객 수를 집계하는 것뿐만 아니라, 축제 이전에 방문객 수를 예측하는 것은 축제의 기획단계에서 사전적으로 축제를 평가하고 운영 측면에서의 효율성을 높이는 데 매우 중요한 과정이다. 지역축제는 많은 수의 인원이 일정 기간 특정 장소를 방문한다는 특수성을 가지고 있으므로, 관련 부문의 의사결정권자를 지원하고 축제 이익을 극대화하기 위해 보다 정확한 방문객 수 예측이 요구된다(진이환, 2006; 김시중, 1993). 부정확한 방문객 수 예측은 축제 운영에 혼란을 야기하고 예산 낭비를 초래한다. 수요를 과소 예측하는 경우, 축제의 오락시설이나 주차시설 및 숙박시설이 부족하게 마련되어 방문객의 불만을 야기하며 최대 이익의 기회를 상실할 수 있다(이충기, 송학준, 신창열, 2007). 이는 더 나아가 방문객의 축제에 대한 만족도와 재방문 의도를 저하하며, 지역 이미지의 손상을 초래한다(이경모, 2005; 엄지영, 윤선영, 2016; 김선아 등, 2017). 반면, 방문객 수가 실제보다 과대 예측되는 경우에는 축제 시설의 과도한 공급과 불필요한 인력 고용으로 예산 낭비가 일어나게 된다. 따라서, 지역축제의 발전에 있어 정확성이 높은 방문객 수 예측 방법을 개발하는 것이 우선의 과제라 할 수 있다.

1.2 연구 문제

지역축제에 관한 연구는 관광학과 관련 학문 분야에서 적극적으로 진행되고 있다. 가장 큰 비중을 차지하는 연구는 축제와 관련된 다양한 요인이 방문객의 만족도나 재방문 의도와 같은 잠재변수에 미치는 영향을 실증적으로 분석하는 연구이다. 또한, 지역축제로 인해 발생한 경제적 파급효과를 추정하는 연구가 다수 이루어졌다. 이와 같은 연구들은 축제 기간 또는 그

이후에 자료를 수집하여 진행되었으며, 축제의 성과를 파악하고 향후 개선점을 도출하였다.

반면, 지역축제 방문객 수 예측에 관한 연구는 그 필요성에도 불구하고 거의 이루어지지 않았다. 진이환(2006)은 방문객 수 예측모형이 필요한 이유를 제시하고, 정량적 예측기법을 통해 특정 지역축제의 방문객 수 예측모형을 개발 및 비교하였다. 이어진 연구에서는 중력모형을 활용한 예측모형을 제시하였으며, 지역축제의 특성을 반영하는 방문객 수 예측 연구가 지속해서 진행되어야 한다고 강조하였다.

타 학문 분야 및 산업 분야에서 적극적으로 도입하고 있는 머신러닝(machine learnin: 기계학습) 기법은 다양한 예측 연구에서 탁월한 성과를 나타내고 있다. 컴퓨터가 데이터를 통해 모형을 개발하고 결과를 예측하는 머신러닝 기법은 타 연구 기법과 비교하였을 때, 시간과 비용은 감축시키면서 높은 예측 정확성을 보인다. 국내 관광 연구 분야에서 머신러닝 기법은 부분적으로 도입이 진행되고 있으나, 수요 예측에는 아직 선례가 존재하지 않는다. 따라서, 머신러닝 기법을 활용하여 지역축제와 관련된 관측 데이터만으로 방문객 수 예측모형을 개발하려는 시도는 문화 및 관광 학문 분야의 연구 방법을 확장하는 동시에, 지역축제를 위한 사전 평가지표와 실무적 시사점을 제시할 것으로 기대된다.

1.3 연구 목적

본 연구의 목적은 다음과 같이 정리할 수 있다.

첫째, 지역축제 관련 관측 변수 데이터를 수집한 후, 머신러닝 기법을 활용하여 방문객 수 예측모형을 개발하고자 한다. 머신러닝 기법을 통해 수집한 데이터만으로 방문객 수 예측을 수행하는 모형을 개발할 수 있다면, 관련 영역에서의 연구 방법을 확장하는 결과를 기대할 수 있을 것이다.

둘째, 방문객 수에 대한 지역축제 관련 특성 변수들의 영향력을 비교하고자 한다. 본 연구에서 고려한 특성 변수는 행정적 요소와 인구 및 접근성을 포함하는 지역 관련 변수 6개, 축제의 내용과 홍보, 예산 등을 포함하는 축제 관련 변수 15개, 총 21개이다. 모형 개발 과정에서 어떤 변수가 예측에 큰 영향을 미치는지 확인한다.

마지막으로, 연구의 결과에 따라 학문적 시사점과 실무적 시사점을 제시하고자 한다. 머신러닝 기법을 활용하여 예측 연구를 진행함으로써, 기존의 연구에서 도출하지 못했던 시사점을 찾을 수 있을 것으로 기대하며, 지역축제와 관련된 다양한 이해관계자를 위한 실무적 제안을 제시한다.

Ⅱ. 이론적 배경 및 선행연구

2.1 지역축제에 관한 실증연구

지역축제에 관한 연구는 1990년대의 지역축제의 증가와 함께 적극적으로 진행되기 시작하였다. 관련 연구 중 가장 큰 비중을 차지하는 연구는 통계적 기법을 통해 잠재변수 간 영향을 검증하는 실증연구이다. 다수의 연구자가 지역축제 방문객을 대상으로 실증연구를 수행하여 만족도나 재방문 의도 등의 변수에 영향을

미치는 요인을 탐색하고, 이를 통해 축제의 개선점과 차별화 전략 등 실무적 시사점을 제시하였다.

이장주와 박석희(1999)는 관광지 또는 관광시설의 이미지를 측정하는 도구는 많으나, 지역축제의 이미지 측정척도가 미비함을 지적하며, 이용성・유희성・신기성・향토성・전통성・체험성・교육성의 7가지 차원에서 30개 측정항목을 개발하였다. 이어진 이장주와 조현상(2000)의 연구에서는 개발된 측정척도를 활용하여 국내 6개 지역축제의 이미지를 분석하고 축제 차별화를 위한 개별 시사점을 제시하였다. 서휘석와 이동기(2000)는 물리적 환경요인이 축제 방문객의 만족과 재방문 의도, 구전 의도에 미치는 영향을 실증적으로 확인하고, 그에 따른 관리적 시사점을 도출하였다. 이환범과 송건섭(2002)은 지역축제의 서비스 측면에 집중하여 서브퀄(SERVQUAL) 요인이 방문객의 만족도 및 재방문 의도에 미치는 영향을 검증하였다. 김근우(2004)는 이벤트 신비감・흥분 스릴감・교육성・유흥성・문화체험성으로 구분되는 방문객의 방문 동기가 축제 만족도에 미치는 영향을 확인하고 비교하였다. 고승익 등(2007)은 지역주민의 지역사회 애착도에 따라 지역축제의 영향에 대한 인식이 달라지며, 이러한 차이가 축제에 대한 만족도・추천의도・재방문 의도에 영향을 미친다는 것을 실증적으로 검증하였다. 신현식과 김창수(2011)는 지역축제에서 스토리텔링[1]의 도입이 축제 매력성과 만족에 미치는 영향을 확인하고 그 필요성을 강조하였다.

지역축제를 대상으로 수행된 다수의 실증연구는 성과에 영향을 미칠 수 있는 다양한 요인변수를 파악함으로써 축제의 개선과 특화를 위한 시사점을 제시하였다. 그러나 대부분의 연구가 특정 지역축제의 방문객을 대상으로 이루어져 일반화가 어렵다는 점, 설문을 통해 측정하여 객관성이 다소 부족하다는 점 등의 한계가 있다. 이에 전국의 다양한 지역축제에 적용가능하면서 객관성을 지닌 지표의 필요성을 확인하였다.

2.2 지역축제의 경제적 파급효과에 관한 연구

지역축제의 개최는 관광 인프라 구축에 비해 적은 비용과 시간을 소모하지만, 지방자치단체의 예산에서 큰 비중을 차지하며 수익성이 모호한 측면이 있으므로, 축제의 타당성을 입증하기 위해 경제적 의미를 파악하는 것이 매우 중요하다(김상호, 2006; 신창열, 2019). 이는 단순히 축제 내에서만 이루어지는 거래만 고려할 것이 아니라, 축제 방문객이 축제 장소를 벗어나 지역 내에서 행하는 총체적인 경제 활동을 고려함으로써 파악할 수 있다(김연형, 2008). 이에 지역축제의 경제적 파급효과를 분석하는 연구가 지속적으로 진행되고 있다.

김상호(2006)는 산업연관표[2]에 입지상계수를 사용하여 전남 지역축제의 시군별 투입・산출 모형을 개발하였다. 이충기와 최영준(2010)은 보령머드축제를 대상으로 경제적 파급효과

1) 스토리텔링(storytelling): 이야기를 만들어 전달하는 과정에서 이야기의 내용과 전달 기술 및 매체를 아우르는 서사 방식(윤유석, 2010).

2) 산업연관표: 1년간 우리나라 상품과 서비스의 생산과 처분에 관련된 모든 거래를 종합 분석한 표(시사상식사전, 2015).

를 파악하기 위해, 충남지역 산업연관표를 토대로 분석을 수행하였으며 직접추계방식(조사법)과 간접추계방식(비조사법)을 비교하였다. 오남현(2012)은 농촌 지역의 환경자원을 활용한 곤충 바이오엑스포를 대상으로 단순특화계수법을 통해 지역투입·산출모형을 개발하였다. 이한성 등(2016)은 지역축제를 개최지역의 수출산업으로 간주하여, 수출승수를 추정하고 수출기반모형을 개발하여 경제적 파급효과를 분석하였다. 권재일과 김한주(2019)는 울진 워터피아 페스타의 경제적 파급효과를 파악하기 위하여, 표본조사를 통해 방문객의 총지출액을 추정하고 이에 관광승수를 대입하여 분석하였다. 신창열(2019)은 강원도의 관광승수를 지역 내 순수효과와 지역 간 누출효과로 구분하여 추정한 후 철원 한탄강 트레킹 축제의 경제적 파급효과를 산출하였다.

지역축제의 경제적 파급효과에 관한 다수의 연구는 축제의 경제적 의미를 제시하고 실무적 시사점을 도출하는 의의를 갖는다. 그러나 실증연구와 마찬가지로 특정 지역축제를 대상으로 연구가 진행되어 타 축제에는 일반화하기 어렵다는 점, 축제 종료 이후 진행되는 사후 평가 방법이라는 점 등의 한계가 있다.

2.3 관광수요 예측에 관한 연구

관광 관련 학문 분야에서는 다양한 관광수요를 예측하기 위해 여러 기법을 활용한 연구를 수행하고 있다. 관광수요 예측은 관련 기획자 및 운영자의 의사 결정을 지원하며, 관광자원을 효율적으로 분배하는 데 중요한 지표가 된다(안종윤, 1995; 김시중, 1993).

김시중(1993)은 외래관광객 수와 관광수출액을 예측하기 위하여 계량경제학적 접근방법을 활용하였다. 최영문과 김사헌(1994)은 박스-젠킨스 모형[3]의 계절 ARIMA를 활용하여 내국인의 월별 해외 관광을 예측하였다. 안종윤(1995)은 외래관광객 수와 내국민의 국내 관광 및 해외 관광 수요를 대상으로, 정량적 예측기법과 정성적 예측기법을 비교하고 복합적 예측모형을 정립하였다. 김경숙과 정의선(2002)은 강원 동해안 해수욕 방문객의 시계열 자료를 이용하여 추세 분석을 수행하였으며 곡선 추정 3차 모형을 개발하였다. 이충기 등(2007)은 Winters 지수평활법을 활용한 예측치에 추정된 방문의사율을 고려하는 결합기법을 고안하여 BIE Expo 방문객을 예측하였다. 이충기와 윤설민(2012)은 추정된 방문 의사율에 실현율에 해당하는 그루버 지수와 자기확신지수를 적용하여 여수 엑스포 방문 수요를 예측하고 비교하였다. 정철 등(2017)은 텍스트 마이닝[4] 기법을 활용하여 안동시 관광 수요를 예측하였다.

다양한 관광수요를 예측하고자 한 연구들은 예측 연구 방법을 확장하는 동시에, 실무적으로 매우 활용도가 높다. 그러나 지역축제는 다수의 방문객의 일정 기간 특정 장소에 유입되는 특수성을 가지고 있으며, 신생 지역축제의 경우 과거 자료가 존재하지 않아 시계열 모형 등을 적용하기 어렵다. 이에 존재하는 다양한 관측

3) 박스-젠킨스 모형(Box-Jenkins model): 단일변량 시계열 자료를 분석하여 예측을 수행하는 확률 과정 모형으로, AR·MA·ARMA·ARIMA·계절 ARIMA를 포함한다(최영문, 김사헌, 1994).

4) 텍스트 마이닝(text mining): 비정형 텍스트 데이터에서 새롭고 유용한 정보를 찾아내는 과정 또는 기술을 말한다(출처: 우리말샘).

변수 데이터를 통해 예측모형을 구축하는 머신러닝 기법을 활용하여 문제를 해소하고자 한다.

2.4 머신러닝 방법론

2.4.1 머신러닝의 개념

머신러닝(machine learning: 기계학습)은 인공지능(AI, artificial intelligence)의 한 분야이다. 인간이 학습을 수행하듯이 컴퓨터가 입력받은 데이터를 통해 스스로 학습을 수행하게 함으로써 새로운 지식을 얻어내는 기법이라고 할 수 있다(김덕현 등, 2019; 이동훈, 김태형, 2020). 본래 컴퓨터는 계산과 같은 직렬처리 작업은 인간보다 월등히 빠르고 정확하게 수행하지만, 다양한 정보가 제시되었을 때 어떠한 패턴을 발견하고 의미를 찾아내는 복합적인 작업은 수행하지 못하였다. 이에 뉴런 간 상호작용으로 학습을 수행하는 인간의 신경망을 컴퓨터에 구현한 것이 머신러닝으로, 복잡하고 무질서한 데이터 상에서도 분류와 회귀, 연관성 분석 등의 작업이 가능해졌다(조용준, 2018). 머신러닝의 종류는 매우 다양하지만 데이터 분석에 가장 많이 활용되는 학습 방법은 지도 학습과 비지도 학습이다.

지도 학습은 인간이 직접 컴퓨터의 학습 과정을 지도한다(조용준, 2018). 지도 학습을 위해서는 훈련 데이터 세트(training set)가 필요하며, 이는 특성 변수(feature variable)와 목표 변수(target variable)를 가지고 있어야 한다. 훈련 데이터 세트를 활용하여 컴퓨터는 일반화 모형을 구축하며, 새로운 데이터에 대하여 목표 변숫값을 유추해내는 것이 지도 학습의 목표이다(Marsland, 2016).

비지도 학습은 레이블 되지 않은 샘플 데이터 세트만 제공하며, 인간의 지도 없이 컴퓨터가 스스로 유사한 데이터를 그룹화하고 범주를 할당하는 작업을 수행하도록 한다(Bonnin, 2018; Marsland, 2016). 목푯값 없이 입력된 데이터 간의 상호 유사성과 차이를 분석하여, 군집을 만들거나 연관성 규칙을 찾아내는 것이 학습의 목표이다(조용준, 2018).

본 연구에서는 목표 변수에 해당하는 '지역축제 방문객 수'가 주어지므로 지도 학습이 적합하다. 지도 학습은 목표 변수의 유형에 따라 범주형(categorical)이면 분류(classification), 수치형(numericalal)이면 회귀(regression)로 구분된다(조용준, 2018). '지역축제 방문객 수'는 연속된 값을 갖는 수치형 데이터로, 회귀 문제에 해당한다. 회귀 문제에 적용이 가능한 지도 학습 알고리즘 중 선형 회귀(linear regression)와 랜덤 포레스트(random forest), 에이다부스트(adaboost)를 본 연구에서 사용하였다.

2.4.2 선형 회귀(linear regression)

선형 회귀는 데이터와 회귀 선 또는 초평면 사이의 오차 거리를 최소화하는 방정식을 찾아내는 것이 학습의 목표이다. 회귀식의 함수는 다음과 같다.

$$y_i = \beta x_i + \alpha + \epsilon_i$$

α는 식의 절편, β는 생성된 선의 기울기에 해당한다. 변수 x는 독립 변수이며, 종속 변수인 y는 회귀 변수 및 응답 변수라고도 한다. ϵ_i은 오차 또는 데이터에서 회귀한 선까지의 거리를 나타낸다. 답을 찾는 과정에서 손실 함수를 통해 이 거리가 계산되며, 가장 일반적인 손

실 함수는 최소제곱법이다(Bonnin, 2018). 최소제곱법의 함수식은 다음과 같다.

$$J(\beta_0,\beta_1)=\sum_{i=0}^{n}(y_i-\beta_0-\beta_1x_i)^2$$

2.4.3 랜덤 포레스트

랜덤 포레스트(random forest) 알고리즘은 의사결정 나무 중 CART(classification and regression tree)와 앙상블 기법(ensemble learning) 중 배깅(bagging) 알고리즘을 조합한 것으로, Breiman에 의해 2001년 제안되었다. CART 알고리즘을 채택하였기 때문에 데이터의 유형과 분포의 제약 없이 사용 가능하며, 과대 적합의 위험이 낮기 때문에 다수의 머신러닝 연구에서 채택되고 있다(조용준, 2018).

배깅은 부트스트랩(boot-strap)과 어그리게이팅(aggregating)이 합쳐진 말로 여러 모형을 결합하는 가장 단순한 기법이다(조용준, 2018; Marsland, 2016). 부트스트랩은 주어진 데이터 세트에서 무작위성을 원칙으로 데이터를 추출하여 여러 개의 데이터 세트를 생성한다(유진은, 2015). 어그리게이팅은 결합한다는 뜻으로, 부트스트랩으로 여러 개의 데이터 세트가 만들어졌을 때, 이를 각각 예측모형으로 만들어 결과를 예측하고 다수결 또는 평균 등의 기준을 통해 최종 결과를 도출한다. 랜덤 포레스트에서는 부트스트랩 과정에서 선택되지 않은 데이터인 OOB(out-of-bag) 데이터를 활용하여 스스로 모형의 정확도를 검증하므로, 모형의 생성 과정에서 스스로 안정성 확보가 가능하다(유진은, 2015; 조용준, 2018).

2.4.4 에이다부스트

에이다부스트(adaboost)는 Freund와 Shapiro에 의해 1995년 소개된 알고리즘으로, 앙상블 기법 중 부스팅(boosting)을 활용하는 기본적인 알고리즘이다(Marsland, 2016). 부스팅은 여러 개의 약한 분류기(weak classifier)를 활용해 강한 분류기(strong classifier)를 구축하는 기법을 뜻한다. 에이다부스트는 초기의 약한 분류기에 가중치를 할당하고 학습시키는 과정을 여러 번 반복하면서 최종의 강한 분류기를 생성한다(김재협 등, 2010). 처음에는 데이터에 대한 정보가 부족하여 각 데이터에 동일한 가중치를 부여하고, 이후의 학습에서는 오분류된 데이터일수록 높은 가중치를 준다. 학습의 차수가 높아질수록 오분류된 데이터가 점점 사라지면서 강한 분류기가 구축된다(김기상, 최형일, 2016). 다음은 에이다부스트의 가중치 함수와 생성된 강한 분류기의 함수이다.

$$w_n^{(t+1)}=w_n^t\exp(\alpha_t I(y_n\neq h_t(x_n)))/Z_t$$

$$f(x)=sign(\sum_{t=1}^{T}\alpha_t h_t(x))$$

Ⅲ. 연구설계

3.1 분석 대상

3.1.1 분석 대상 선정

본 연구는 문화체육관광부의 ‘2018년 전국 지역축제 개최계획’을 참고하여 분석 대상을 선정하였다. 해당 자료에서는 2018년에 개최된

국내의 886개 지역축제에 관한 정보를 제시한다. 886개 지역축제 중, 관광객 유치가 주목적인 타 지역과 달리 거주 시민을 위한 복지 목적으로 축제를 개최하는 서울특별시의 126개 축제와, 접근성 측면에서 큰 차이가 나타나는 제주특별자치도의 28개 축제를 제외하였다. 또한, 머신러닝 지도 학습을 수행하기 위해서는 각 축제에 따른 2018년 방문객 수가 제공되어야 한다. 이에 2018년 축제 방문객 수가 불분명한 139개 축제를 제외하여 총 593개 지역축제를 대상으로 분석을 수행하였다.

3.1.2 특성 변수 선정

예측모형을 개발하기 위하여, 지역축제 방문객 수에 영향을 미쳤을 것으로 사료되는 특성변수를 선정하였다. 지역 고유의 특성을 반영하

<표 1> 변수 표

구분	범주	변수명	유형	설명
특성 변수	지역 관련	지역 인구수	numerical	
		도 인구수	numerical	지역이 속한 도의 인구
		도 구분	categorical	경기도/강원도/경상남도/경상북도/전라남도/전라북도/충청남도/충청북도
		행정 구분	categorical	특별광역시/시/군
		ktx역 유무	categorical	유:1 무:0
		서울과의 거리	numerical	
	축제 관련	축제 종류	categorical	문화예술/생태자연/지역특산물/전통역사/ 주민화합/기타
		관련 뉴스 수	numerical	축제 개최 전 한 달간 게재된 뉴스 수
		관련 블로그 포스팅 수	numerical	축제 개최 전 한 달간 포스팅된 글 수
		전년도 포스팅 수	numerical	2017년 한 해 동안 포스팅된 글 수
		초대가수 수	numerical	
		대표 초대가수 화제성	numerical	축제 개최 전 한 달간 게재된 대표 초대가수 관련 뉴스 수
		날씨	numerical	축제 기간의 날씨를 점수화하여 입력
		기간	numerical	
		예산	numerical	
		횟수	numerical	
		입장료 유무	categorical	유:1 무:0 (주요 프로그램 유료:2)
		입장료 가격	numerical	
		개최 월	categorical	1/2/3/4/5/6/7/8/9/10/ 11/12
		개최 계절	categorical	봄/여름/가을/겨울
		전년도 방문객 수	numerical	2017년 해당 지역축제 방문객 수
목표 변수		2018년 축제 방문객 수	numerical	

는 변수 6개, 지역축제 자체의 특성 측면에서 15개, 총 21개 특성 변수가 선정되었다(표 1).

축제를 개최하는 지역의 인구수나 행정적 특징, 접근성 등이 지역축제 방문객 수에 영향을 주었을 것으로 예상하여, 지역 인구수・도(道) 인구수・행정 구분・도 구분・서울과의 거리・ktx역 유무의 6가지 관측 변수가 특성 변수로 포함되었다. 축제 자체의 홍보성이나 콘텐츠, 기후 요소 등 각 축제의 특징 역시 방문객 수에 영향을 미쳤을 것으로 판단하여, 축제 관련 뉴스 수・관련 블로그 포스팅 수・전년도 관련 블로그 포스팅 수・초대가수 수・대표 초대가수 화제성・날씨・기간・예산・횟수・입장료 유무・입장료 가격・개최 월・개최 계절・전년도 방문객 수를 변수로 채택하였다.

3.2 데이터 전처리

본 연구에서 특성 변수로 선정한 21개의 변수 중, 14개의 변수와 목표 변수에 해당하는 방문객 수는 데이터 유형이 수치형(numerical) 데이터에 해당한다. 수치형 변수는 데이터의 분포와 숫자의 규모(scale)가 큰 차이를 갖고 있어, 회귀가 제대로 수행되지 않아 결정계수 측정이 어렵고 오차값이 크게 나타난다. 이에 데이터 전처리 과정으로써 데이터 정규화를 진행하였다.

데이터 정규화는 데이터 관리를 용이하게 해주고 모형이 최적화되는 과정, 특히 반복적인 학습을 수행하는 알고리즘에서 모형이 데이터에 잘 수렴할 수 있도록 만든다(Bonnin, 2018). 본 연구에서는 각 수치형 특성 변수에 대해 엑셀 함수인 'Norm.dist'를 적용하여 정규화를 수행함으로써, 각 데이터가 0부터 1 사이의 정규분포 값을 갖도록 하였다(표 2).

3.3 예측모형 생성 및 선정

지역축제 방문객 수 예측모형을 개발하기 위하여 본 연구에서는 머신러닝 지도 학습 방법을 활용하였다. 목표 변수인 방문객 수는 연속적인 값을 갖는 수치형 데이터에 해당하므로 회귀분석이 가능한 선형 회귀, 랜덤 포레스트, 에이다부스트 알고리즘을 이용하여 모형을 생성하였다. 모형 생성에는 Orange(Ver. 3.23)를 사용하였으며, k-fold cross validation[5]을 수행하여 예측모형의 성능을 측정 및 비교하였다. 교차 검증을 통해 산출된 각 모형의 오차와 결정계수(R^2)를 비교하여 최종 예측모형을 채택하였다.

<표 2> 예산 변수의 기존 값과 정규화 진행 후 변환 값 일부

예산(단위:십만원)	150	200	360	500	1000	2140
예산 정규화	0.2319	0.2579	0.3503	0.4394	0.7506	0.9949

5) k-fold cross validation: 주어진 데이터의 규모가 상대적으로 작을 때, 이를 k개의 그룹으로 무작위 분할 후 k-1개의 데이터 세트로 모형을 생성하고 남은 1개의 세트로 모형의 오차를 구하는 작업을 반복하여 평균 오차를 최종적인 모형의 오차로 산출하는 검증 기법(김석범, 정한규, 황호연, 2018).

Ⅳ. 예측모형 평가 및 선택

4.1 예측모형 성과 검증 지표

회귀 문제 상황에서는 주로 오차값과 결정계수를 통해 모형의 성능을 평가한다. 실제 값과 모형의 예측값 사이의 거리가 오차(error)이며, 오차가 작을수록 예측모형의 정확도가 높은 것으로 판단한다. 평균제곱오차(MSE, mean squared error)에 제곱근을 씌운 평균제곱근오차(RMSE, root mean squared error)를 본 연구에서 생성된 모형의 평가지표로 선택하였다.

평균제곱근오차를 구하는 식은 아래와 같다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}$$

결정계수(coefficient of determination)는 회귀 문제 상황에서 생성된 회귀선이 주어진 데이터에 얼마나 적합한지를 보여주는 지표로, R^2로 표현된다(허명회 등, 1991). 실제 데이터의 변동량 중 회귀선을 통해 설명되는 정도를 나타내기 때문에 모형의 설명력으로 해석할 수 있으며, 예측을 수행하는 경우 예측 정확도로 볼 수 있다. 결정계수는 0부터 1 사이의 값을 가지며, 1에 가까워질수록 설명력 또는 예측 정확도가 뛰어난 유용한 모형이라고 할 수 있다(김석우, 2007). 결정계수를 산정하는 식은 다음과 같다.

$$R^2 = \frac{\sum(\hat{y_i} - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{회귀선에 의해 설명되는 변동}}{\text{전체 변동}}$$

4.2 예측모형 검정력 평가

선형 회귀, 랜덤 포레스트, 에이다부스트 알고리즘을 사용하여 생성된 세 가지 모형의 성능을 확인하고 비교하기 위하여, k-fold cross validation을 수행하고 각 모형의 오차와 결정계수를 측정하였다.

<표 3>에서는 선형 회귀, 랜덤 포레스트, 에이다부스트 순으로 각 모형의 평균제곱근오차(RMSE)와 결정계수 값을 보여준다. 세 모형 중 랜덤 포레스트와 에이다부스트 알고리즘을 통해 생성된 모형이 선형 회귀를 통해 생성된 모형보다 낮은 오차와 현저히 높은 결정계수를 기록하였다. 두 모형 중에서는 랜덤 포레스트의 결정계수가 0.800으로 에이다부스트의 결정계수인 0.789보다 높은 점수를 기록했으므로 최종 모형으로 랜덤 포레스트를 선정하였다.

<표 3> 생성된 모형의 오차와 결정계수

model	RMSE	R^2
Linear Regression	0.105	0.581
Random Forest	0.073	0.800
Adaboost	0.075	0.789

Ⅴ. 연구 결과

5.1 2018년 지역축제 방문객 수 예측

전국의 총 593개 지역축제에 대한 데이터를 수집하고, 이를 머신러닝 알고리즘에 대입하여 예측모형을 개발한 결과, 평균제곱근오차 0.073, 결정계수 0.800의 랜덤 포레스트 모형이 생성되었다. 이러한 결과는 예측모형이 작은 오차범위 내에서 주어진 데이터의 80% 정도를 설명할 수 있으며 높은 수준으로 예측이 가능함

을 입증하였다.

5.2 방문객 수 예측에 영향을 미친 변수 평가

RReliefF[6] 값은 회귀분석에서 각 특성 변수가 목표 변수를 예측하는데 미치는 영향을 점수로 표현한 것이다. 아래의 <표 4>를 보면, 개최지역의 도(道) 구분이 RReliefF 값 0.485를 기록하면서 지역축제 방문객 수 예측에 가장 큰 영향을 미친 것으로 나타났다. 그다음으로는 축제가 개최된 월(0.458)과 축제의 종류(0.403)의 영향이 크며, 대표 초대가수의 화제성(0.386)과 날씨(0.300) 또한 다소 영향을 미치는 것으로 확인되었다.

<표 4> 특성 변수의 RReliefF 값 순위

변수명	RReliefF 값
도 구분	0.485
개최 월	0.458
축제 종류	0.403
대표 초대가수 화제성	0.386
날씨	0.300
예산	0.285
행정 구분	0.272
초대가수 수	0.271
관련 블로그 포스팅 수	0.269
개최 계절	0.256
입장료 유무	0.254
기간	0.233
횟수	0.229
전년도 방문객 수	0.204
서울과의 거리	0.189

Ⅵ. 결과 토의 및 시사점

6.1 결과 논의

본 연구는 지역축제 방문객 수 예측모형 개발을 목적으로, 관련 특성 변수를 선정하고 데이터를 수집하여 머신러닝 기법을 적용하였다. 이를 통해 결정계수 값 0.800에 달하는 예측모형을 개발함으로써, 머신러닝을 기법을 활용하여 방문객 수 예측이 가능하다는 것을 확인하였다.

이는 문화 관광 산업 및 학문 분야에서 머신러닝을 통해 예측을 수행한 최초의 시도로, 관련 관측 데이터만으로 높은 성능의 예측모형 생성이 가능하며 이후 관련 연구 분야에서 활용될 수 있다는 가능성을 확인 할 수 있었다. 또한, 국내에서 미진하였던 지역축제 방문객 수 예측 연구를 진행함으로써, 지역축제의 특수성을 반영하면서 향상된 예측력을 가진 모형을 개발하였다. 덧붙여, 선행연구들은 특정 지역축제를 대상으로 연구를 수행하여 연구 결과를 타 지역축제에 적용하기 어렵다는 한계를 가졌으나, 본 연구에서는 전국의 지역축제 데이터를 활용하여 모형을 구축함으로써 다양한 지역축제에 적용이 가능한 일반화된 모형을 제시하였다는 점은 학문적으로 큰 의미가 있다.

실무적 관점에서는 지역축제의 사전 평가 도구를 개발하였다는데 첫 번째 의의를 갖는다. 특정 축제 개최 이전에 축제의 성과 지표인 방

6) RReliefF: Relief 평가를 회귀 모형에 적용하기 위하여 Marko Robnik-Sikonja와 Igor Kononenko가 제시한 알고리즘으로, 분류 모형에서 목표 변수와 특성 변수 간의 거리를 비교하여 변수의 영향력을 측정한 것과 달리 회귀선과 특성 변수 간의 MSE 값 비교를 통해 변수의 영향력을 도출하였다(Marko Robnik-Šikonja, Igor Kononenko, 1997).

문객 수를 예측함으로써 해당 축제의 효과성을 미리 파악할 수 있으며, 관련 정책 입안자 또는 축제 기획자의 의사 결정을 지원할 수 있다. 또한, 지역축제의 개최로 인해 증가하는 주차시설이나 숙박시설, 오락 시설 등의 수요를 축제 이전에 파악하여 대비함으로써 예산 운영의 효율성을 제고할 수 있다.

RReliefF 값의 비교를 통해 특성 변수들의 영향력을 살펴본바, 축제 개최지역의 도(道) 구분과 개최 월, 축제의 종류, 대표 초대가수의 화제성 등이 중요 요인으로 파악되었다. 개최지역의 도 구분이 가장 높은 영향력을 보인 것은, 도 차원에서 구축된 지역 인프라나 교통 체계, 사회적으로 자리 잡은 지역문화 등이 지역축제의 방문객 수에 미치는 영향이 컸다고 해석될 수 있다. 지역축제를 활성화하기 위해 개최지역의 환경만 고려할 것이 아니라, 광역적으로 효율적인 환경 조성이 필요하다고 판단된다.

특성 변수 중 기후를 대변하는 변수로 개최월을 포함하였는데, 연구 결과 방문객 수에 미치는 영향이 크다는 것을 확인하였다. 우리나라의 기후는 시기에 따라 뚜렷한 차이가 나타나기 때문에, 지역축제가 개최되는 월에 따라 축제에 대한 방문 의도에 영향을 준 것으로 추측한다. 축제의 소재가 기후에 중점을 두고 있는 생태자연축제를 제외한 다른 종류의 축제는 연구 결과를 고려하여 개최 시기를 조정하면 축제 방문객 수 증가시킬 수 있을 것으로 기대된다.

국내 지역축제는 주 소재에 따라 전통역사축제, 문화예술축제, 지역특산물축제, 생태자연축제, 주민화합축제, 기타축제로 구분되고 있다. 축제의 종류에 따라 방문객 수에 차이가 있음을 확인하였는데, 문화예술축제와 생태자연축제가 타 종류의 축제와 비교하였을 때, 방문객 수가 높은 것으로 나타났다. 이러한 결과를 참고하여 신규 지역축제를 기획하거나 기존의 지역축제에 관련 콘텐츠를 추가함으로써 더 많은 방문객을 유치할 수 있을 것으로 예상된다.

다음으로, 축제에 초청된 대표 초대가수의 화제성이 높은 영향력을 보인 것은 화제성 높은 유명 가수를 초청하는 것이 방문객을 유입시키는 데 효과적인 수단임을 나타내는 결과이다. 특성 변수 중 초대가수의 수는 RReliefF 값이 비교적 낮은 반면, 대표 초대가수의 화제성은 높은 점수를 기록하였으므로, 축제 기획 과정에서 화제성이 높지 않은 여러 명의 가수를 섭외하는 것보다 화제성이 높은 초대가수 한 명을 섭외하는 것이 더 효과적인 것으로 판단된다.

6.2 한계점과 향후 과제

본 연구는 지역축제 방문객 수 예측모형을 개발하기 위하여 머신러닝 기법을 활용한 높은 예측력의 모형을 생성하고, 연구 결과에 따른 학문적·실무적 시사점을 제시하였다. 하지만 데이터의 수집에 있어 문화체육관광부가 공시한 자료와 검색 엔진에 의존하였기 때문에, 부정확한 데이터와 결측치가 존재하여 모형 성능이 다소 저하되었을 것으로 예상한다. 또한, 지역축제 방문객 수에 영향을 미치는 최대한의 특성 변수가 고려되지 않았으며, 특히 축제 외부의 상황적 변수는 연구 과정에서 포함되지 않았다.

향후 문화체육관광부를 비롯한 지방자치단

체의 지역축제 관련 기관과의 협업이 가능하다면, 지역축제와 관련된 다양한 형태의 데이터를 확보하여 추가적인 특성 변수를 고려한 예측모형을 발전시킬 수 있을 것으로 기대된다. 또한, 본 연구에서는 각 모형을 사전에 설정해 놓은 train set으로 학습을 시킨 후, test set을 활용하여 각각 성능 비교를 하였다. 이에 <표 3>에서 제시한 바와 같이 각 모형별 결정계수를 비교하여 가장 성능이 우수한 모형(랜덤 포레스트)을 선정하였다. 본 연구와 같이 머신러닝을 활용하여 특정 값을 예측하는 연구에 있어서 왜 모형별로 각각 다른 목표 변수값을 나타내는지에 대한 기술적인 영역의 연구가 추가로 수행되어야 할 것이다.

참고문헌

권재일, 김한주, “지역산업연관모델을 이용한 지역축제의 경제적 파급효과 분석,” 관광레저연구, 제31권, 제1호, 2019, pp. 169-184.

김경숙, 정의선, “강원 동해안 해수욕 방문객의 수요예측과 정책적 시사,” 관광학연구, 제26권, 제1호, 2002, pp. 255-271.

김근우, “지역축제의 방문동기가 만족도에 미치는 영향 분석: 청도소싸움축제를 중심으로,” 관광학연구, 제27권, 제4호, 2004, pp. 203-218.

김기상, 최형일, “퍼지 Adaboost를 이용한 객체검출“, 한국콘텐츠학회논문지, 제16권, 제5호, 2016, pp. 104-112.

김덕현, 유동희, 정대율, “의사결정나무 기법을 이용한 노인들의 자살생각 예측모형 및 의사결정 규칙 개발”, 정보시스템연구, 제28권, 제3호, 2019, pp. 249-276.

김상호, “전남 지역축제의 경제적 파급효과 분석을 위한 시군별 투입·산출 모형”, 인문사회과학연구, 제14권, 2006, pp. 11-34.

김석범, 정한규, 황호연, “항공기 날개의 통계적 중량 예측식 도출 연구”, 한국항공우주학회지, 제46권, 제1호, 2018, pp. 32-40.

김석우, 기초통계학, 학지사, 2007.

김선아, 김정원, 원동연, 최예림, “무슬림 관광객 증대를 위한 머신러닝 기반의 할랄 푸드 분류 프레임워크”, 정보시스템연구, 제26권, 제3호, pp. 273-293.

김시중, “한국 국제관광수요의 계량경제학적 예측에 관한 연구”, 관광학연구, 제17권, 1993, pp. 57-80.

김연형, “지역문화축제의 지역경제파급효과에 관한 연구: 전주 국제영화제를 중심으로”, 응용통계연구, 제21권, 제1호, 2008, pp. 125-140.

김영대, 이선영, 이환수, “IT 거버넌스 요인이 지역축제 성과에 미치는 영향”, 디지털융복합연구, 제16권, 제12호, 2018, pp. 1-10.

김재협, 장경현, 이준행, 문영식, “아이다부스트(Adaboost)와 원형 기반함수를 이용한 다중표적 분류 기법”, 전자공학회논문지, 제47권, 제3호, 2010, pp. 22-28.

김종성, 이준형, 김동현, 최창현, 이명진, 김형수, “머신러닝 기반의 호우피해 발생 확

률 예측모형 개발”, 한국방재학회논문집, 제19권, 제6호, 2019, pp. 115-127.
김철원, 이석호, “문화관광축제 육성방안”, 한국관광연구원, 2001.
김학용, 권호종, “문화관광축제의 일몰제 도입에 따른 자생력 강화방안 연구”,인문콘텐츠, 제45권, 2017, pp. 173-189.
류상범, 박정수, “기상통계론”, 전남대학교 출판부, 2012.
문화체육부, “한국의 지역축제”, 문화체육부, 1996.
문화관광부, “문화관광축제 변화와 성과(1996~2005)”, 문화관광부, 2006.
문화체육관광부, “2018년 전국 지역축제 개최계획”, 문화체육관광부, 2018.
문화체육관광부, “2019년 전국 지역축제 개최계획”, 문화체육관광부, 2019.
박석희, 이장주, “지역축제의 이미지 측정척도 개발에 관한 연구: 진도 영등축제를 중심으로”, 관광학연구, 제22권, 제3호, 1999, pp. 243-261.
박종부, 이수범, “문화관광축제 서비스품질이 지역브랜드 자산, 지역태도, 지역 애호도에 미치는 영향: 축제유형에 따른 다중집단분석을 중심으로”, 관광레저연구, 제30권, 제9호, 2018, pp. 401-420.
서휘석, 이동기, “물리적 환경이 지역축제의 만족과 재방문 및 구전의도에 미치는 영향에 관한 연구”, 한국행정학보, 제34권, 제1호, 2000, pp. 229-243.
손종원, 나승화, “지역주민의 축제 참여동기와 만족도가 지지도에 미치는 영향”, 유통과학연구, 제12권, 제8호, 2014, pp. 103-112.
신창열, “지역산업연관모델을 활용한 춘천국제레저대회의 경제적 파급효과 분석”, 이벤트컨벤션연구, 제15권, 제1호, 2019, pp. 1-21.
신창열, “지역축제 개최로 인한 지역경제 파급효과에 관한 연구: 2019 철원 한탄강 얼음트레킹을 중심으로”, MICE관광연구, 제56권, 2019, pp. 199-218.
신현식, “문화관광축제 스토리텔링 속성 분석에 관한 연구”, 인문콘텐츠, 제19권, 2010, pp. 511-532.
신현식, 김창수, “지역축제 스토리텔링이 축제 매력성과 방문자 만족에 미치는 영향”, 관광연구, 제26권, 제3호, 2011, pp. 225-244.
안종윤, “바람직한 관광수요예측방법의 모형정립”, 관광연구논총, 제7권, 1995, 5-32.
엄지영, 윤선영, “축제 이미지가 도시브랜드자산 및 지역 애호도에 미치는 영향”, 관광연구, 제31권, 제2호, 2016, pp. 131-150.
오남현, “환경자원을 활용한 지역축제의 경제적 파급효과 분석: 예천군 곤충바이오엑스포 축제를 사례로”, 한국비교정부학보, 제16권, 제3호, 2012, pp. 363-382.
오훈성, “문화관광축제 평가체계 연구(보고서 번호: 기본연구 2011-56)”, 한국문화관광연구원, 2011.
오훈성, “문화관광축제 선정의 일몰제 적용에 따른 제도 운영 개선방안 연구(보고서 번호: 기본연구 2013-20)”, 한국문화관

광연구원, 2013.
유진은, “랜덤 포레스트”, 교육평가연구, 제28권, 제2호, 2015, pp. 427-448.
윤유석, “스토리텔링을 통한 지역 역사인물의 대중화”, 인문콘텐츠, 제19권, 2010, pp. 301-325.
이동훈, 김태형, “머신러닝 기법을 활용한 대졸 구직자 취업 예측모델에 관한 연구,” 정보시스템연구, 제29권, 제2호, 2020, pp. 287-306.
이경모, “축제시설이 만족도와 재방문의도에 미치는 영향연구”, 이벤트컨벤션연구, 제2권, 2005, pp. 137-152.
이준엽, “축제 성공요인에 대한 프리리스팅 연구”, 이벤트컨벤션연구, 제31권, 2018, pp. 1-18.
이충기, “관광응용경제학”, 대왕사, 2011.
이충기, 송학준, 신창열, “BIE Expo 방문객 수요예측”, 관광레저연구, 제9권, 제3호, 2007, 263-281.
이충기, 최영준, “지역산업연관모델을 이용한 보령머드축제의 경제적 파급효과 분석”, 관광연구, 제25권, 제5호, 2010, pp. 83-100.
이충기, 윤설민, “실현율(그루버지수, 자기확신지수)을 이용한 관광 수요 예측: 엑스포 잠재 방문객 사례”, 관광학연구, 제36권, 제2호, 2012, pp. 11-29.
이한성, “농어촌축제가 지역경제에 미치는 파급효과: 수출기반모형을 이용하여”, 지역개발연구, 제47권, 제1호, 2015, pp. 27-39.
이한성, 이상학, 윤상헌, “농촌축제가 지역경제에 미치는 파급효과의 추정: 하동군 악양대봉감축제를 사례로”, 한국지역경제연구, 제35권, 2016, pp. 49-60.
이환범, 송건섭, “서브퀄(SERVQUAL) 요인을 이용한 지역축제의 서비스질 평가”, 한국행정학보, 제36권, 제3호, 2002, pp. 249-268.
이희찬, 문혜선, “지역축제 시장규모추정 및 수요분석”, 관광학연구, 제34권, 제1호, 2010, pp. 277-294.
정철, 신진옥, 박수지, 송상헌, “텍스트 마이닝을 통한 관광지 수요 예측: 온라인 검색엔진을 중심으로”, 관광학연구, 제41권, 제1호, 2017, pp. 13-27.
조문수, 고승익, 오상운, 고경실, “지역주민의 지역축제 평가에 관한 연구: 지역사회 애착도, 지역축제 영향인식 차이를 중심으로”, 관광학연구, 제31권, 제4호, 2007, pp. 177-198.
조용준, “빅데이터 SPSS 최신 분석기법: 신경망, SVM, 랜덤포레스트 편”, 한나래출판사, 2018.
조현상, 이장주, “지역축제의 이미지 특성화에 관한 실증연구: 우리나라 6개 지역축제를 중심으로”, 관광학연구, 제24권, 제1호, 2000, pp. 205-224.
진이환, “축제방문자의 수요예측방법 비교”, 관광연구저널, 제20권, 제1호, 2006, pp. 49-61.
진이환, “중력모형을 이용한 축제 수요예측모형개발에 관한 연구”, 동국대학교 박사학위논문, 2008.
최영문, 김사헌, “연구노트 및 논평: 박스- 젠킨

스 방법을 이용한 관광수요 예측: 모형의 진단과 예측", 관광학연구, 제18권, 제1호, 1994, pp. 227-251.

허명회, 정진환, 이종한, "우도거리에 의한 결정계수 R2에의 통합적 접근", 응용통계연구, 제4권, 제2호, 1991, pp. 117-127.

Marsland, S., "머신러닝 가이드: 파이썬 코드 기반(2판, 강전형 역)", 제이펍, 2016.

Bonnin, R., "Machine Learning for developers", Packt, Birmingham 2018

Robnik-Šikonja, M.and Kononenko, K., "An adaptation of Relief for attribute estimation in regression", Machine Learning: Proceedings of the Fourteenth International Conference, ICML'97, 1997, pp. 296-304.

시사상식사전, 산업연관표 [웹사이트]. (2015). URL: https://terms.naver.com /entry.nhn?docId=69046&cid=43667&categoryId=43667

우리말샘, 텍스트 마이닝 [웹사이트]. (n.d.). URL: https://opendict.korean.go.kr/dictionary/view?sense_no=1003292&viewType=confirm

이 인 지 (Lee, In-Ji)

전남대학교 경영학사와 전남대학교 일반대학원 전자상거래협동과정 석사학위를 취득하였다. 주요 관심분야는 머신러닝, 빅데이터, 관광경영 등이다.

윤 현 식 (Yoon, Hyun Shik)

현재 전남대학교 경영학부 조교수로 재직 중이다. 미주리대(Columbia)에서 박사학위를 취득하였고, 동 대학교 경영대학, 택사스대(San Antonio), 경영대학과 오클라호마 주립대(Stillwater) 경영대학에서 강의하였다. 주요 관심분야는 머신러닝을 활용한 소비자행동분석, 데이터마이닝, 개인정보보호, 기술경영 등이다.

<Abstract>

Development of a Model to Predict the Number of Visitors to Local Festivals Using Machine Learning

Lee, In-Ji · Yoon, Hyun Shik

Purpose

Local governments in each region actively hold local festivals for the purpose of promoting the region and revitalizing the local economy. Existing studies related to local festivals have been actively conducted in tourism and related academic fields. Empirical studies to understand the effects of latent variables on local festivals and studies to analyze the regional economic impacts of festivals occupy a large proportion. Despite of practical need, since few researches have been conducted to predict the number of visitors, one of the criteria for evaluating the performance of local festivals, this study developed a model for predicting the number of visitors through various observed variables using a machine learning algorithm and derived its implications.

Design/methodology/approach

For a total of 593 festivals held in 2018, 6 variables related to the region considering population size, administrative division, and accessibility, and 15 variables related to the festival such as the degree of publicity and word of mouth, invitation singer, weather and budget were set for the training data in machine learning algorithm. Since the number of visitors is a continuous numerical data, random forest, Adaboost, and linear regression that can perform regression analysis among the machine learning algorithms were used.

Findings

This study confirmed that a prediction of the number of visitors to local festivals is possible using a machine learning algorithm, and the possibility of using machine learning in research in the tourism and related academic fields, including the study of local festivals, was captured. From a practical point of view, the model developed in this study is used to predict the number of visitors to the festival to be held in the future, so that the festival can be evaluated in advance and the demand for related facilities, etc. can be utilized. In addition, the RReliefF rank result can be used.

Considering this, it will be possible to improve the existing local festivals or refer to the planning of a new festival.

Keyword: Local Festival, Machine Learning, Supervised Learning, Random Forest, Adaboost

* 이 논문은 2020년 8월 2일 접수, 2020년 8월 15일 1차 심사, 2020년 8월 31일 게재 확정되었습니다.