R머신러닝 기초(02반)

레스토랑 수익 예측을 위한 기계학습 모델 연구: 터키의 데이터를 중심으로

ICT 공학부
김준석
202104168

I. 요약 Abstract

새로운 레스토랑을 운영하기 위해서는 시간과 자본의 큰 투자가 필요하다. 잘못된 위치를 선정할 시에는 18개월 이내에 폐쇄되고 영업 손실이 발생한다. 본 연구는 터키의 레스토랑 수익에 영향을 미치는 다양한 변수들을 분석하고, 수익을 예측하는 데 있어서 최적의 알고리즘을 제안한다. 따라서 투자의 효율성을 높이고 연간 레스토랑 판매를 예측하는 데 도움이 된다. 데이터의 불균질성과 상관관계에 있는 기능이라는 두 가지 문제를 해결하기 위해서 특별한 주의를 기울인다. 수익에 영향을 미치는 주요 변수는 레스토랑을 개장한 개장일, 대도시와 기타 도시로 나뉘는 도시 유형, 인구 통계학적 데이터와 부동산 데이터 및 상업 데이터를 담고 있는 P-변수로 구분된다.

필자는 레스토랑의 수익을 예측하기 위해 터키의 레스토랑 데이터를 활용했다. 분석에는 랜덤 포레스트(**Random Forest**)의 알고리즘이 적용되었고, 제시된 모형을 통해 레스토랑 수익 예측을 진행했다. 분석 결과, 개장일, 도시 유형 등과 함께 37개의 P-변수를 분류 및 선정하여 사용함에 따라 기계학습 모형의 예측 정확도가 증가되는 것을 확인할 수 있었다. 또한, 표준화 및 **PCA**와 **RBF**커널을 적용했을 때 예측 정확도가 높은 것을 확인할 수 있었다. **Random Forest**를 활용해 개장일이 수익을 예측하는데 가장 중요한 변수로 평가되었다. 본 연구를 통해 레스토랑 수익 예측에 필요한 변수들의 영향력을 파악하고, 이를 활용해 국내 레스토랑의 수익 예측에 활용해 투자의 효율성과 수익을 예측할 수 있다.

주제어: 회귀 분석, 기계학습, 수익 예측, 랜덤 포레스트

II. 서론 Introduction

2.1 문제의 정의 및 연구의 필요성

오랜 시간 전부터 레스토랑 즉, 음식업은 경제에 없어서는 안 될 존재가 되었다. 지난 10년간 자영업자 수는 감소하고 있지만 음식업종의 종사자 수는 오히려 증가하여 음식업의 생존경쟁은 더 치열해지고 있다. 자영업자는 2006년 758만 명에서 2016년 674만 명으로 11% 감소하였지만 음식점 및

주점업의 개인 사업체수는 2006년 57만개에서 2016년 65만개로 14% 증가하였고 종사자수도 143만명에서 172만명으로 20% 증가하였다. 이 같은 사실은 자영업자의 60%를 점유하고 있는 3대 자영업의 평균 생존 기간 분석 결과를 통해서도 확인할 수 있는데, 도소매업은 5.2년, 수리와 기타 개인 서비스업은 5.1년, 음식업 및 숙박업은 3.1년으로 음식업의 평균 생존 기간이 상대적으로 짧은 것으로 나타났다(남윤미, 2017). 생존 기간이 가장 짧은 음식업을 가벼운 마음으로 시작하기에는 시간과 자본의 소모가 크다. 세계보건기구(WHO)는 2020년 3월 11일에 신종 코로나바이러스(COVID-19)로 인한 세계적 대유행(펜데믹)을 선언하였다. 그리고 사람과 사람 사이에 확산되는 감염을 막기 위해 폐쇄나 사회적 거리두기, 집함금지 등의 과감한 조치가 내려졌다.(Lekfuangfu et al., 2020). 뿐만 아니라, 정부에서는 식당과 술집을 폐쇄하는 강력한 조치를 내리게 되었다(Ozili & Arun, 2020). 그 결과, 사람들은 가정에서 생활하는 시간이 늘어났고, 음식점 방문을 자제하게 되었다. 이로 인해 음식점들은 큰 타격을 받게 되었다(Kashif & Aziz-Ur-Rehman, 2020). 지난해 10월 19일부터 11월 5일까지 소상공인 1018개 사업체를 대상으로 코로나19 관련 소상공인 영향 실태조사를 했다. 그 결과, 10명 중 7명이 매출 감소를 겪은 것으로 보고했다(소상공인연합회, 2020). 이처럼 외식업계가 큰 어려움을 겪고 있다. 이로 인해 기존의 질서와 생활 방식은 무너지고 새로운 질서와 생활 방식으로 바뀌고 있다. 즉, 새로운 표준(new normal)과 가치관을 토대로 생존을 준비해야 하는 시대가 온 것이다. 이에 외식업 점포도 변화된 뉴노멀에 적응하지 않으면 안 되는 시대가 되었다(김옥현, 2020).

코로나19 발생 이전에도 경기침체로 인한 외식업 점포의 어려움은 가중되고 있었다. 그때에도 위기 상황을 해결하려는 방안이나 정책적 지원에 대한 중요성만 강조할 뿐, 그에 대한 실질적이고 구체적인 방안은 제시하지 못했다. 그나마 문제 해결을 위해 학계에서 외식업 점포의 성공 요인에 대한 연구가 있었다. 하지만, 연구의 대부분이 고객 측면에서의 선택속성에 관한 연구(안성식·박기용·양주환, 2005; 양위주·박희정, 2000; 유영진, 1999)와 외식업 점포의 경영자 특성에 관한 연구(김예주·김영국, 2013; 송경숙, 2013)였다. 음식 업종의 내부 경쟁이 심화되면서 창업에서 성공하기 위해서는 세부업종 선택과 입지 선택이 무엇보다 중요하게 되었다. 외식업은 위와 같이 여러 까다로운 조건에 따라 성공과 실패가 명확하게 갈린다. 일반적으로 음식점의 운영상태는 매출액을 통해 가늠해볼 수 있는데 매출액에 영향을 주는 요인을 규명할 경우 위험부담을 줄일 수 있으며 얼마나 유리한지를 알 수

있게 된다.

본 연구에서는 기계학습 기법을 활용해 터키의 레스토랑 수익 데이터를 분석하고, 수익을 예측할 수 있는 방법을 찾고자 한다. 기계학습 기법을 활용할 경우 일반적인 선형 예측모형에서 필요한 선형 가정의 제약에서 벗어날 수 있으며, 예측력이 높다는 장점이 있다(이태형, 전명진, 2018). 이에 본 연구에서는 기계학습 기법을 활용해 수익 데이터를 분석하고자 한다.

따라서 본 연구에서는 레스토랑 수익에 영향을 미칠 수 있는 변수들을 분석하여 각 변수의 영향력을 파악하고, 수익에 가장 큰 영향을 미치는 변수가 무엇인지 알아보고자 한다. 또한 기계학습 기법을 활용해 수익 예측 모델을 구축하고, 예측 모델 간 예측 정확도 비교를 통해 수익 예측에 가장 적합한 모델을 구축하고 수익에 대한 예측 결과를 제시하고자 한다.

본 논문의 구성은 다음과 같다. II.서론에서는 본 연구의 배경과 필요성에 대해 설명하고 연구의 범위 및 연구 방법을 정의하고, 이론적 배경과 수익 예측 관련 선행 연구에 대해 서술한다. 수익에 영향을 미치는 변수와 예측 알고리즘에 대한 내용이 포함되어 있다. III.방법론에서는 본 연구에서 활용된 예측 관련 방법론에 대해 서술한다. IV.결과에서는 연구에서 활용한 데이터에 대해 설명하고, 알고리즘의 성능을 확인한다. V.결론에서는 기존 연구와의 차이점과 향후 연구방향, 기대효과에 대해 서술한다.

2.2 수익에 영향을 미치는 변수

외식업 점포의 수익을 예측하기 위한 다양한 연구들이 진행되고 있다. 외식업 점포 성공 요인에는 많은 요인들이 있다. 좋은 입지조건, 쾌적한 분위기, 편리한 주차 공간, 효율적인 동선, 멋진 인테리어와 내부시설, 차별화된 메뉴, 합리적인 가격, 양질의 서비스, 매력적인 아웃테리어와 외부시설, 철저한 위생과 청결함 등이 간접적으로 영향을 미칠 수 있는 것으로 나타났다(권영상·동학림, 2021). 또 다른 선행연구 노은빈(2018)등에서는 유동인구와 업종, 주변 사업체, 상권 발달 수준, 높은 점포들끼리 군집한다는 의미의 공간자기상관 등이 수익에 영향을 미친다고 판단했다. 또한, 점포환경, 점포운영, 경영자특성을 꼽을 수 있다고 생각한다. 점포환경에는 입지와 시설을 하위변인으로 설정할 수 있고, 점포운영은 상품력, 서비스력,

위생환경을 하위변인으로 설정할 수 있다. 경영자 특성으로는 진취성과 위험감수성을 하위변인으로 설정할 수 있다.

요약하면, 수익에 영향을 미치는 요인으로는 위치, 유동인구 등으로 구분해 볼 수 있다. 이는 <표-1>에 정리되어 있다. <표-1>에서 확인할 수 있듯이 여러 연구들이 저마다 중요한 변수를 활용해 수익을 예측하고자 하였다. 하지만 국내에서 해외 외식업 수익 예측에 대한 연구는 거의 없었다. 따라서 본 연구는 위치, 유동 인구 등의 변수들을 포함하고 터키의 레스토랑 데이터 집합에 속해있는 추가적인 변수들을 활용하여 수익을 예측할 것이다.

<표-1> 참고문헌의 주요 활용 변수

참고문헌	주요 활용 변수
권영상, 동학림, 2021	입지조건과 합리적인 가격
노은빈,2018	유동인구와 업종 및 상권 발달 수준

2.3 수익 예측

기계학습에 대한 기법들이 다양하게 개발됨에 따라 외식업의 수익을 예측하고자 하는 연구가 많이 진행되고 있다. 노은빈 등(2018)은 다양한 예측 모델을 활용하여 수익을 예측하고자 했다. OLS(Ordinary Least Squares) 회귀모형과 함께 공간회귀모형(Spatial Regression Model)을 사용해 수익을 예측하고자 하였다. 필요한 입지 정보는 공간 빅데이터를 활용하는 소상공인진흥 공단의 "상권정보 시스템"과 서울시의 "우리 마을 가게 상권 분석 시스템"을 통해 쉽게 접근했다. 서울시 6개 구의 전체 음식업종 집계구 매출액 자료를 세부 업종별 집계국 매출액 자료를 재구축하여 분석에 이용했다.

본 연구에서는 선행연구에서 활용된 변수를 참고하여 변수를 선정한 뒤, 수익을 예측하고자 한다. 예측의 정확성을 높이기 위해 수익에 영향을 주는 다양한 변수들이 활용되었다. 특히, P-변수를 활용하기 위해 P-변수의 상관관계를 따져보고 P 변수 이외의 도시의 크기, 레스토랑의 나이 등을 주요 변수로 채택하여 활용한다. 본 연구의 목적은 레스토랑의 수익에 대한 예측력이 얼마나 되는지 확인하는 데 있다. 뿐만 아니라 기존에 성능 비교로 그쳤던

연구를 확장시켜 수익을 예측할 수 있는 모형을 제시한다. 즉, 개선된 수익 예측 모형을 활용해 국내 외식업 분야에 적용시킬 경우보다 안전하고 경제적인 외식업 선정과 투자를 할 수 있을 것으로 기대된다.

III. 방법론 Methods

최근 들어 인공지능의 위상이 떠오르면서 컴퓨터의 성능 향상을 기반으로 한 기계학습 모델이 주목받고 있다. 기계학습은 인간이 과거의 경험과 다양한 실험을 기반으로 새로운 것을 학습하는 모형을 컴퓨터 알고리즘에 적용한 것이다. 인간이 컴퓨터 알고리즘을 학습시키면서 컴퓨터가 스스로 문제를 해결해간다. 그 후 스스로 학습을 하여 문제에 대한 해답을 제시한다. 이를 기계학습이라고 일컫고, 기계학습은 정보를 활용한 많은 분야에 사용되곤 한다. 기계학습은 지도 학습과 비지도 학습으로 구분되고 지도 학습은 입출력 값이 존재하는 데이터를 강제로 학습시켜 새로운 데이터에 대한 예측이나 분류를 진행하는 것이다. 비지도 학습은 강제 학습이 아닌 컴퓨터가 스스로 데이터나 패턴 사이의 관계를 찾아내어 스스로 학습을 하는 것을 말한다.

본 연구에서는 지도 학습 기법 중 Random Forest 기법을 사용하여 레스토랑의 수익을 예측하고자 한다.

3.1 데이터 전처리

3.1.1 Open Days

pandas의 to_datetime을 사용하여 변경한다. 개장일(Open Date)로부터 나이(Open Days)를 계산하기 위해서는 기준일이 필요한데 본 연구에서는 기준일을 2018년 1월 1일로 설정하고 numpy의 repeat 메소드를 통해 반복 계산을 진행하여 레스토랑의 나이를 계산한다. 또한, 개장일 변수의 데이터는 object 타입의 변수이기 때문에 가게의 나이를 int형 변수로 변경한다.

3.1.2 Type

Type은 레스토랑의 유형으로 푸드코트(FC), 인라인(IL), 드라이브 스루(DT), 모바일 등과 같은 범주형 데이터이므로 삭제해도 무관하기 때문에 삭제한다.

3.1.3 Big Cities & Other

도시 유형(City Group)은 대도시(Big Cities)와 기타 도시(Other)로 나뉘는데 수치형 데이터로만 변환을 하게 되면 서로 간의 관계성이 생긴다. 그러나 관계성이 없기 때문에 관계성으로 인해 잘못된 학습이 일어나는 오류를 방지하고자 pandas의 get_dummies 메소드를 사용하여 더미로 만든 가변수로 변환한다.

3.1.4 Train / Test 분리

sklearn의 train_test_split을 사용하여 본 연구에서는 예측을 위해 원 데이터의 Train set과 Test set의 비율을 7:3으로 지정한다. 난수 발생기의 시드는 0으로 고정하여 코드가 실행될 때마다 동일한 난수 시퀀스가 생성되게 한다. 그리고 그 과정에 다른 임의성이 존재하지 않는 한, 생성된 결과는 언제나 동일하다.

<표-2> 하이퍼 파라미터

하이퍼 파라미터	파라미터 설명	값
test_size	test set 구성의 비율	0.3
random_state	셔플을 위한 시드값	0

3.2 Random Forest Classifier

sklearn의 RandomForestClassifier을 사용하여 많은 변수 중 가장 예측력이 강한 변수가 무엇인지 파악한다. 하지만 블랙박스 모형이기 때문에 설명변수와 반응변수의 설명력을 확보하기 어렵다. 따라서 이를 어느 정도 해결하기 위해, 변수 중요도(Variable Importance)라는 척도를 통해 어느 변수가 예측 성능에 중요한 역할을 하는지 추정한다.

<표-3> 하이퍼 파라미터

하이퍼 파라미터	파라미터 설명	값
n_estimatators	사용하는 결정트리의 개수	250
random_state	셔플을 위한 시드값	1

3.3 StandardScaler & PCA

표준화(standardization)는 기계학습 알고리즘을 훈련 시키는 데 있어서 사용되는 특성들이 모두 비슷한 영향력을 행사하도록 값을 변환해주는 기술이다. 수익을 예측하는 데 필요한 특성들의 단위도 다르고, 값의 범위에도 차이가 있다. 따라서 단위가 다른 특성들을 비교할 수 있도록 sklearn의 StandardScaler를 활용하여 표준화를 통해 0의 평균, 1의 표준편차를 갖도록 변환해준다.

<표-4> StandardScaler 메소드 하이퍼 파라미터

하이퍼 파라미터	파라미터 설명	값
with_std	True인 경우 데이터를 단위 분산 (또는 단위 표준 편차)으로 조정	True
with_mean	True인 경우 크기 조정하기 전, 데이터를 가운데에 맞춤	True

PCA는 특성 행렬의 차원을 축소한다. sklearn의 PCA, KernelPCA를 사용하여 선형적으로 구분되지 않아 선형 변환이 잘되지 않는 데이터를 커널 함수를 이용하여 선형적으로 구분되는 고차원 공간에서 주성분으로 투영된 결과를 반환하게 만든다.

<표-5> PCA 메소드 하이퍼 파라미터

하이퍼 파라미터	파라미터 설명	값
n_components	주성분의 개수 설정	2
svd_solver	full일 경우 전체 SVD를 실행하고 후처리를 통해 구성 요소를 선택	'full'

<표-6> KernelPCA 메소드 하이퍼 파라미터

하이퍼 파라미터	파라미터 설명	값
kernel	사용할 커널 선택	'rbf'
gamma	귤, poly, sigmoid 커널에 대한 커널 계수	1

3.4 RMSE(Root Mean Square Error)

RMSE는 회귀 모델 성능 평가 지표 중 하나이며 오차를 제곱해서 평균한 값의 제곱근으로 값이 작을수록 정밀도가 높은 평가 지표를 사용한다. 이 RMSE 지표 같은 경우 지나치게 높거나 낮은 값(outlier)에 영향을 적게 받는다. 말하자면 잘못 입력된 수익 항목에 페널티를 더 주는 RMSE의

특징은 수익을 예측하는 모델의 성능을 평가할 때 사용하면 적절한 평가 결과를 도출해 낼 수 있다. 본 연구에서는 회귀문제의 한 방법인 RMSE 오류율을 사용한다. RMSE는 MSE의 절대값인 표준 에러로 성능에 대한 에러율을 확인함으로써 회귀문제를 해결할 수 있다.

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

< Y_i = 결과값, $\hat{Y}_i$= 실제값>

Y_i = $\hat{Y}_i$ 일 경우 100%의 성능을 갖게 되며 이때 RMSE는 0이 된다. 즉 RMSE가 작을수록 성능은 향상된다.

3.4.1 RMSE를 이용한 회귀 분석 모델 비교

<표-7> 모델별 RMSE Value

Model	RMSE Value
Lasso Regression	1710660.3807717038
Ridge Regression	1710666.6082206795
Random Forest	1644100.1807500264

```
+----+------------------------+-------------+
|    | Models                 |  Train RMSE |
|----+------------------------+-------------|
|  0 | Lasso Regression - PCA |  1.71066e+06 |
|  1 | Ridge Regression - PCA |  1.71067e+06 |
|  2 | Random Forest - PCA    |  1.6441e+06  |
+----+------------------------+-------------+
```

3.5 Random Forest

랜덤 포레스트 기법은 여러 결정 트리 모델을 학습하는 앙상블 기법이다. 더 좋은 예측 성능을 갖기 위해 모델 한 개만 활용하는 것보다 다수의 모델을 종합해서 결정하는 기법을 앙상블 기법이라 한다. 앙상블 기법의 종류는 크게 배깅(Bagging), 페이스팅(Pasting), 부스팅(Boosting)이 있다. 배깅은 중복을 허용하여 표본을 추출하여 샘플링 하는 방식이다. 예를 들면 훈련 데이터 집합에서 30%만 추출하여 학습 모델을 생성하고 한 번 학습에 사용한 데이터를 다시 훈련 데이터 집합에 놓고 랜덤하게 복원 추출하여 여러 학습 모델을 만든다. 배깅 기법으로 표본의 피처(Freature)를 각기 다르게 무작위로 선정하여 다수의 결정 트리 분류기를 일괄적으로 학습하는 것을 랜덤 포레스트 모델이라 한다. 랜덤 포레스트는 결정 트리 기반의 하이퍼 파라미터를 사용한다. 수익 예측에 사용한 하이퍼 파라미터는

<표-8>과 같고 모두 동일하게 적용한다. 랜덤 포레스트는 편향과 분산의 균형을 맞추는 것이 중요한 기법이다. 편향은 모델이 예측한 값과 실제 값의 차이를 나타낸다. 분산은 학습 데이터의 결과를 얼마만큼 일반화할 수 있는가를 나타내는 것이다. 분산은 학습 데이터의 결과를 얼마만큼 일반화할 수 있는가를 나타내는 것이다. 편향과 분산은 서로 반비례 관계에 있는데, 예측률을 높이기 위해서 학습 데이터에 과적합할 경우 분산이 높아져 모델을 일반화하기 어렵고, 분산을 줄여 일반화 가능성을 높인다면 학습 오차율이 높아진다.

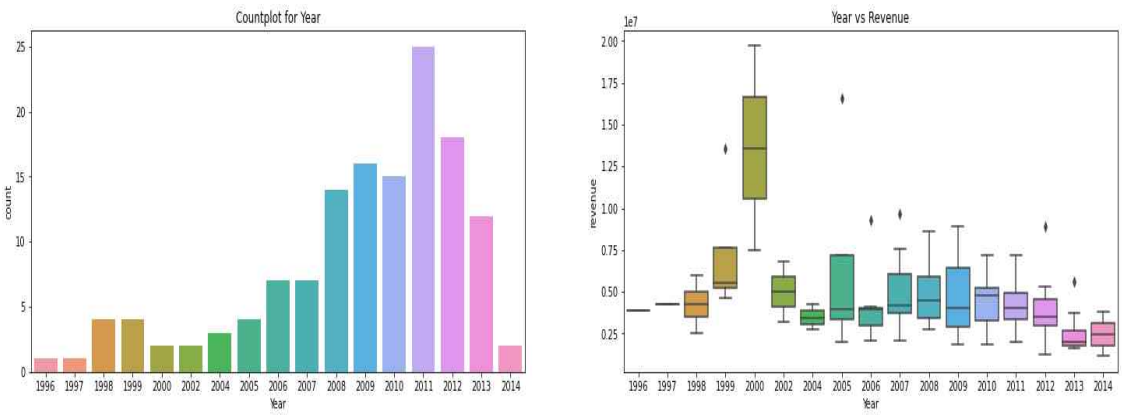
<표-8> 하이퍼 파라미터

하이퍼 파라미터	파라미터 설명	값
n_estimatators	사용하는 결정트리의 개수	250
criterion	분할 품질을 측정하는 기능	'mse'
max_depth	트리의 최대 깊이	30

IV. 결과 Results

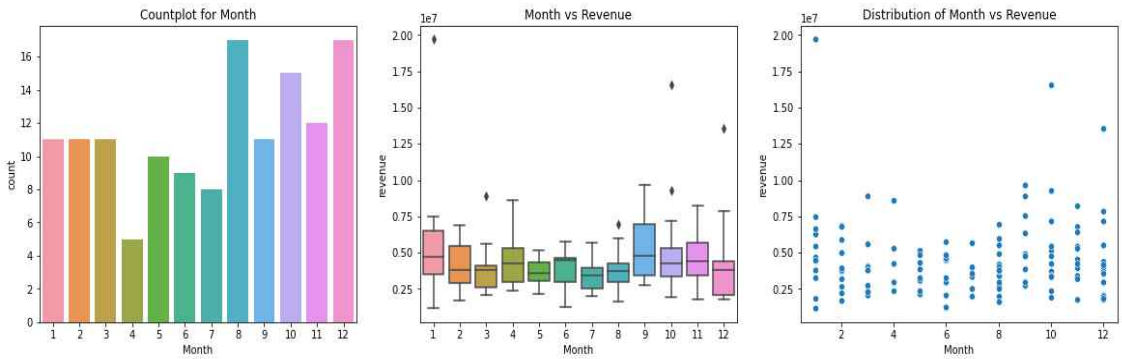
4.1 데이터

본 연구에서는 터키 레스토랑의 수익 데이터를 활용하여 레스토랑의 수익을 예측했다. 캐글(kaggle)에서 제공하는 터키 레스토랑 수익 예측(Restaurant Revenue Prediction) 데이터를 활용했다. 예측을 위한 터키 레스토랑의 수익 데이터는 개장일, 레스토랑 유형, 도시, 도시 유형, P-변수, 수익 데이터로 구성되어 있다. 데이터 시각화를 통해 P-변수를 제외한 수익에 높은 영향을 미치는 변수는 다음 그림들을 통해 알 수 있다.



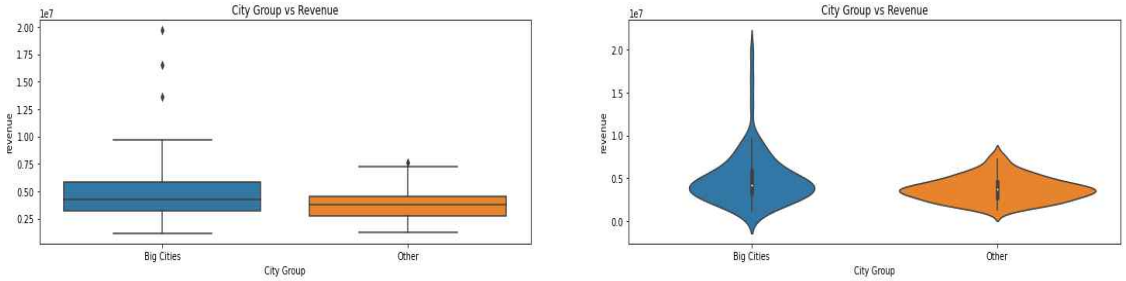
<그림 1> 연별 레스토랑 오픈 횟수와 연도와 수익의 상관관계

<그림 1>은 년도 데이터와 수익 데이터의 상관관계를 보여준다. 이를 통해 2008년부터 2013년 사이에 레스토랑 오픈 횟수가 잦았고, 2000년에 가장 높은 수입을 기록했음을 알 수 있다.



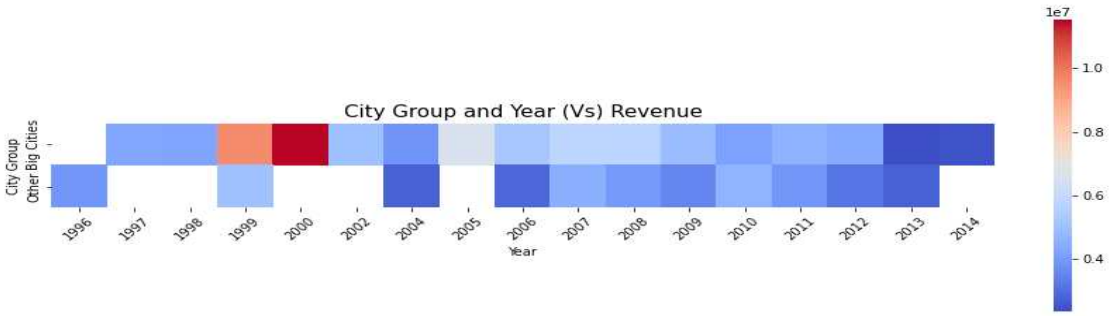
<그림 2> 월별 레스토랑 오픈 횟수와 연도와 수익의 상관관계

<그림 2>는 월 데이터와 수익 데이터의 상관관계를 보여준다. 이를 통해 대부분의 레스토랑이 8월과 12월에 오픈을 하고 있으며, 4월과 9월에 높은 수익을 기록했음을 알 수 있다.



<그림 3> 도시 유형과 수익의 상관관계

<그림 3>은 도시 유형 데이터와 수익 데이터의 상관관계를 보여준다. 이를 통해 대도시(Big Cities)에 있는 레스토랑이 더 높은 수익을 올리고 있음을 알 수 있다. 따라서 대도시에 새로운 레스토랑을 여는 것이 수익 면에서 이점을 갖는다는 것을 알 수 있다.



<그림 4> 도시 유형과 년도, 수익의 상관관계

<그림 4>는 년도 데이터와 도시 유형 데이터 그리고 수익 데이터의 상관관계를 보여준다. 이를 통해 1997년부터 매년 대도시(Big Cities)에 레스토랑이 오픈을 하고 있으며, 1997년부터 2012년 사이에 높은 수익이 있음을 알 수 있다.

4.2 변수 선정

RandomForestClassifier을 통해 변수 중요도를 따져보았을 때 레스토랑의 나이, 대도시, 기타 도시, P-변수 등의 상위 19개의 변수를 적용했을 때 가장 정확도가 높다고 판단했다. 도시의 이름이 적힌 도시 변수는 P-변수의 지리적 데이터에 포함되어 있기 때문에 사용하지 않는다. 선정된 변수는 다음 <표-9>의 상위 19개의 변수와 같다.

<표-9> 변수 중요도

01	Open Days	0.087800	14	P29	0.030229	27	P33	0.013801
02	P19	0.049030	15	P12	0.027642	28	P31	0.013405
03	P11	0.048616	16	P21	0.027468	29	P14	0.013398
04	P28	0.048073	17	Big Cities	0.021086	30	P34	0.012988
05	P6	0.047423	18	Others	0.019282	31	P16	0.012948
06	P2	0.046847	19	P13	0.018407	32	P15	0.012927
07	P20	0.044270	20	P9	0.018130	33	P25	0.012733
08	P22	0.041101	21	P7	0.017888	34	P37	0.012634
09	P5	0.038940	22	P10	0.017070	35	P24	0.011999
10	P23	0.035039	23	P32	0.015280	36	P30	0.011465
11	P4	0.033633	24	P27	0.015114	37	P36	0.010757
12	P8	0.032433	25	P26	0.014458	38	P35	0.010173
13	P3	0.031882	26	P17	0.014213	39	P18	0.009416

4.3 결과예측

다수의 변수를 활용해 수익을 예측한 결과 85.01070702666127%의 정확도, 37.6356215854514%의 평균 오차율을 기록하는 기계학습 모델을 만들어냈다. <표-10>에 수익에 대한 예측값과 실제값 그리고 오차율을 나타내었으며 결정 트리의 개수가 증가함에 따라 무조건적으로 오차율이 줄어드는 것은 아님을 <표-11>을 통해서 확인할 수 있다.

<표-10> 예측값과 실제값 그리고 오차율

예측값	실제값	오차율(%)	예측값	실제값	오차율(%)
3714896.756	2267425.0	63.837691	4705920.552	7495092.0	37.213305
4486442.688	4952497.0	9.410492	5180054.496	3956086.0	30.938875

4347229.640	3354383.0	29.714498	5105501.692	2156098.0	136.793582
4504207.520	3871344.0	16.347385	5220805.288	3752885.0	39.114449
4118482.116	2732645.0	50.714129	4712343.156	8904084.0	47.076609

<표-11> 결정 트리 개수에 따른 평균 오차율

결정 트리 개수(n_estimator)	평균 오차율(%)
1	58.984111557494074
10	62.5592519485054
100	43.33222656084231
200	39.91403430274294
300	40.439834917879175

V. 토의 및 결론 Discussion and Conclusion

본 연구는 다양한 수익 관련 변수를 활용해 레스토랑 수익 예측에 관한 연구를 수행했다. 기존 연구에서 주요 변수로는 입지 조건, 상권 발달 수준 등의 지리적 정보가 주요 변수로 작용했지만 본 연구에서는 RandomForest 기법을 통해 변수 중요도를 측정한 결과 레스토랑의 나이가 수익에 가장 큰 영향을 주는 변수로 나타났다.

본 연구에서는 레스토랑 수익 예측을 위해 다양한 변수들을 활용했다. 그러나 본 연구에서는 수익에 영향을 주는 P-변수들의 변수 중요도만을 고려하여 P-변수의 내용을 설명하지 못했다. 즉, P-변수는 지리적 정보와 인구 정보, 그리고 상업적 정보를 담고 있다는 것을 알고 있을 뿐 몇 번째의 P-변수가 어떠한 데이터를 담고 있는지 알 수 없는 P-변수를 사용함으로써 수익에 대한 영향력에 대해서 설명이 부족하다. 따라서 P-변수에 대한 정보를 자세히 파악할 수 있다면 수익 예측의 상관관계에 대해 파악하기 더 수월할 것이라고 생각한다.

본 연구에서는 Random Forest를 사용하여 레스토랑의 수익을 예측했지만 더 나아가 XG Boost나 LightGBM 모델을 사용하여 수익을 예측하면 더 높은 정확도를 관찰할 수 있을 것이라고 생각한다. 이를 통해 보다 안전한 외식업 투자가 가능할 것으로 기대된다.

[참고문헌]

남윤미(2017), "국내 자영업의 폐업율 결정요인 분석",『BOK 경제연구』.

Lekfuangfu, W. N., Piyapromdee, S., Porapakkarm, P., & Wasi, N.(2020). On Covid-19: New implications of job task requirements and spouse's occupational sorting. COVID Economics: Vetted and Real-Time Papers. Retrieved on July 5, 2020.

Ozili, P. K., & Arun, T.(2020). Spillover of COVID-19: Impact on the Global Economy(March 27, 2020).

Kashif, M., & Aziz-Ur-Rehman, M. K. J.(2020). Demystify the Covid-19 effect on restaurant. International Journal of Medical Science in Clinical Research and Review, 3(3), 281-289.

소상공인연합회(2020). "코로나19 관련 소상공인 영향 실태조사" 결과보고서.

김옥현(2020). "With COVID-19, 외식업의 길을 묻다".『외식경영학회』, 23(4), 343-365.

안성식·박기용·양주환(2005). "컨조인트분석을 이용한 패밀리레스토랑의 성공요인에 관한 연구".『외식경영연구』, 8(1), 87-104.

양위주·박희정(2000). "패스트푸드점 선택속성과 이용 행태 간의 관계에 관한 연구".『관광레저연구』, 12(2), 107-122.

유영진(1999). "패밀리레스토랑 이용 행태에 따른 선택속성에 관한 연구".『관광레저연구』, 11(1), 43-67.

김예주·김영국(2013). "외식업의 지식과 기술, 창업태도가 창업의도에 미치는 영향: 과잉자신감의 매개효과".『외식경영학회』, 16(2), 97-118.

송경숙(2013). "외식산업 선택속성이 고객지향성 및 창업성과에 미치는 영향". 한국콘텐츠학회논문지, 13(6), 481-495.

이태형・전명진 (2018), "딥러닝 모형을 활용한 서울 주택가격지수 예측에 관한 연구: 다변량 시계열 자료를 중심으로,"『주택도시연구』, 8(2), 33-56.

노은빈·이상경. (2018). "입지요인이 음식업 업종별 매출액에 미치는 영향 비교연구". 『부동산연구』, 28(4), 37-51.

권영산·동학림(2021). "외식업 점포의 성공요인이 경영성과에 영향을 미치는 요인에 관한 연구". 『한국외식산업학회지』, 17(2), 55-72.