



고객 리뷰 데이터 분석을 통한 방송 소개 식당의 폐업 요인 분석

B조

강현직 구민지
김다연 이채원

목차

1

프로젝트 개요

선행연구

2



3

분석과정

4

결과



GOAL

주제 선정 배경



미쉐린 스타 식당들도 폐업한다?!

경험을 중시하는
소비 트렌드

요리 예능
프로그램의 증가



또 간 집



방송 출연한 식당들도 비슷한 요인으로 폐업할 것으로 예상

주요 목적

폐업 예측모형 개발

- 실제 경험 데이터를 활용,
소비자의 감성 패턴과 폐업
여부 간 상관관계 탐구
- 폐업 위험도 예측할 수 있는
모형 구축

폐업 위험 군 식당 개선방안 제시

예측 모형에서 도출된 위험
요인을 분석하여 개별 식당 맞춤
개선 전략 제시

개선된 부분

01 방송 프로그램 출연 식당 선정 강화

- 주요 프로그램 출연 식당 데이터 확대
- 골목식당, 또간집, 맛있는녀석들, 수요미식회, 6시내고향, 생활의달인

02 데이터셋 규모 대폭 증가

- 초기 10개에서 400개 이상으로 확장
- 모델 성능 강화 및 신뢰성 향상

개선된 부분

03 폐업 영향 변수 추가

- 폐업에 영향 미칠 수 있는 변수들 기본 칼럼으로 추가
- 경제적, 비경제적 지역 특성 등 다양한 요인 포함

04 다양한 SNS 플랫폼 활용 리뷰데이터 수집

- Google Maps, 네이버 블로그, 유튜브 댓글에서 실제 소비자 경험 데이터 확보

2. 선행연구(1)

소진수, 신판섭. 2020. 음식점 리뷰 감성분석을 통한 세부 평가항목별 평점 예측

JKSCI

Journal of The Korea Society of Computer and Information
Vol. 25 No. 6, pp. 81-89, June 2020
<https://doi.org/10.5708/jksici.2020.25.66.081>

Rating Prediction by Evaluation Item through Sentiment Analysis of
Restaurant Review

Jin-Soo Son*, Pan-Seop Shin*

*Student, Dept. of Computer Engineering, Daejin University, Pocheon, Korea
*Professor, Dept. of Computer Engineering, Daejin University, Pocheon, Korea

[Abstract]

Online reviews we encounter commonly on SNS, although a complex range of assessment information affecting the consumer's preferences are included, it is general that such information is just provided by simple numbers or star ratings. Based on those review types, it is not easy to get specific information that consumers want and use it to make a decision for purchase. Therefore, in this study, we propose a prediction methodology that can provide ratings broken down by evaluation items by performing sentiment analysis on restaurant reviews written in Korean. To this end, we select 'food', 'price', 'service', and 'atmosphere' as the main evaluation items of restaurants, and build a new sentiment dictionary for each evaluation item. It also classifies review sentences by rating item, predicts granular ratings through sentiment analysis, and provides additional information that consumers can use to make decisions. Finally, using MAE and RMSE as evaluation indicators it shows that the rating prediction accuracy of the proposed methodology has been improved than previous studies and presents the use case of proposed methodology.

• **Key words:** Text Mining, Opinion Mining, Sentiment Analysis, Sentiment Dictionary, Rating Prediction

[요약]

우리가 SNS상에서 흔하게 접하는 온라인 리뷰에는, 소비자들의 선호도에 영향을 미치는 다양한 평가 정보가 복잡적으로 포함되어 있지만 이들 매우 간단한 형태의 수치(또는 별점)로 제공되는 것이 일반적이다. 이러한 리뷰에서, 소비자가 원하는 구체적인 정보를 얻고, 이를 구매를 위한 판단에 활용하기란 쉽지 않다. 따라서 본 연구에서는 한국어로 작성된 음식점 리뷰를 대상으로, 감성분석을 수행하여 평가 항목별로 세분화된 평점을 제공 가능한 예측 방법론을 제안한다. 이를 위해, 음식점의 주요 평가항목으로 '음식', '가격', '서비스', '분위기'를 선정하고, 평가항목별 맞춤형 감성사전을 새롭게 구축한다. 또한 평가항목별 리뷰 문장을 분류하고 감성분석을 통해 세분화된 평점을 예측하여 소비자가 의사결정에 활용 가능한 추가적인 정보를 제공한다. 마지막으로, MAE와 RMSE를 평가지표로 사용하여 기존의 연구보다 제안기법의 평점 예측 정확도가 향상되었음을 보이며, 제안 방법론의 활용 사례도 제시한다.

• **주제어:** 텍스트 마이닝, 오피니언 마이닝, 감성분석, 감성사전, 평점 예측

*First Author: Jin-Soo Son, Corresponding Author: Pan-Seop Shin
*Jin-Soo Son (jss21901@daejin.ac.kr), Dept. of Computer Engineering, Daejin University
*Pan-Seop Shin (pseshin@daejin.ac.kr), Dept. of Computer Engineering, Daejin University
• Received: 2020. 03. 10, Revised: 2020. 05. 25, Accepted: 2020. 05. 26.

Copyright © 2020 The Korea Society of Computer and Information
<http://www.jksici.kr> pISSN 1598-647X | eISSN 2389-9945

- 본 연구의 방법론적 근거에 참고가 된 연구
- 음식, 가격, 서비스, 분위기라는 4가지 주요 평가항목을 선정하여 각 항목별로 감성어휘를 포함하는 맞춤형 감성사전을 구축
- 한국어 리뷰를 기반으로 한 감성분석을 통해 세부 항목별로 평가할 수 있는 가능성을 제시

2. 선행연구(2)

방준아, 손광민, 이소정, 이현근, 조수빈. 2018. 서울 치킨집 폐업 예측 모형 개발 연구

한국빅데이터학회지
제3권 제2호, 2018, pp. 35-49

서울 치킨집 폐업 예측 모형 개발 연구

A Study on Predictive Modeling of Public Data:
Survival of Fried Chicken Restaurants in Seoul

방준아¹ · 손광민¹ · 이소정² · 이현근^{3*} · 조수빈¹
성균관대학교 통계학과¹, CJ올리브네트웍스 DT융합연구소², CJ올리브네트웍스 빅데이터센터³

요 약

대한민국에서 치킨집은 전 세계 맥도날드 매장 수보다 많은 정도로 자영업의 큰 비중을 차지하는 창업 업종이다. 치킨집은 꾸준히 생겨나고 있지만, 소상공인의 창업 후 폐업률은 3년 62%, 5년 71%에 육박하는 것으로 나타났다(4). 특히, 숙박 및 음식점업의 경우 70%가 3년을, 82%가 5년을 버티지 못하는 것으로 집계되었다(1). 이에 본 연구는 '서울 치킨집 폐업 예측 모형'을 개발하여, 예비창업자가 개업 후보지를 선정하는 의사결정 과정에 도움을 주고자 하였다. 먼저 행정자치부 지방행정 인허가 데이터의 업종별 개업 업 신고 일자를 중심으로 다양한 변수를 수집하였다. 이후 다양한 분류 알고리즘을 적용하고, 예측 모형의 성능을 비교하였다. 그 결과, 인공신경망(Neural Network)이 가장 높은 정확도를 보였지만 특이도와 민감도가 불균형적이었다. 이에 비해 유연판별분석(FDA)은 인공신경망보다 정확도는 낮지만, 상대적으로 균형적인 예측 성능을 보였다.

■ 키워드 : 자영업, 치킨집, 폐업, 예측 모형, 기계학습

Abstract

It seems unrealistic to say that fried chicken, often known as the American soul food, has one of the biggest markets in South Korea. Yet, South Korea owns more numbers of fried chicken restaurants than those of McDonald's franchise globally(4). Needless to say not all these fast-food commerce survive in such small country. In this study, we propose a predictive model that could potentially help one's decision whilst deciding to open a store. We've extracted all fried chicken restaurants registered at the Korean Ministry of the Interior and Safety, then collected a number of features that seem relevant to a store's closure. After comparing the results of different algorithms, we conclude that in order to best predict a store's survival is FDA(Flexible Discrimination Analysis). While Neural Network showed the highest prediction rate, FDA showed better balanced performance considering sensitivity and specificity.

■ Keyword : Entrepreneurship, Restaurant, Survival, Machine Learning, Predictive Model

- 서울시 치킨집의 폐업 예측 모형을 개발하여 자영업 창업자들이 개업 후보지를 선정할 때 도움을 주는 것을 목표로 한 연구
- 행정자치부의 인허가 데이터를 기반으로 업체 특성, 지역적 특성, 경쟁 업체 현황, 상권 특성, 경제적 변수 등을 포함한 다양한 입력 변수를 사용함
- 여러 머신러닝 알고리즘을 비교한 결과, 유연판별분석(FDA)이 민감도와 특이도에서 균형적인 성능을 보여 최적의 모델로 선정되었고, 주요 변수로는 유명 브랜드 여부, 시설 총 규모, 주변 경쟁 업체 수 등이 포함됨
- 특정 분류 알고리즘의 민감도와 특이도를 균형 있게 고려한 평가 방법론을 참고하여 적합한 모델을 선택하는 기준 설계
- 폐업에 미치는 주요 요인을 계량적으로 분석하여 방송 소개 식당의 폐업 요인을 해석하는 데 활용 가능

식당 선정 기준



구글맵 리뷰 50개 이상

- 충분한 데이터를 확보하기 위해 리뷰 50개 이상 식당들로 선별



서울에 위치

- 상권 간 차이를 최소화하기 위해 서울에 위치한 식당으로 제한



6개 방송 프로그램 중 하나에 출연

- 골목식당, 또간집, 맛있는녀석들, 수요미식회, 6시내고향, 생활의달인



Google Maps



리뷰 데이터

✓ restaurant & reviews로 column을 통일하여 데이터 수집

1



1) 구글맵 리뷰 크롤링 데이터

- 네이버지도와 달리 폐업한 식당들의 리뷰 보존

2



2) 블로그 리뷰 크롤링 데이터

- 식당 이름과 출연 프로그램 이름을 조합하여 제목을 검색
- 네이버 블로그 API를 호출하여 관련 포스트 크롤링

3



3) 유튜브 댓글 크롤링 데이터

- 관련 영상 댓글들에서 식당 방문 후기 수집

리뷰 감성분석 칼럼

01

한국어 KNU 사전을
활용한 감성분석

Features

구글 감성점수(KNU)
블로그 감성점수(KNU)
유튜브 감성점수(KNU)

02

X-ANew 사전을 활용한
영어 번역본 감성분석

Features

구글 번역본 감성점수(금부정)
구글 번역본 감성점수 (자극)
블로그 번역본 감성점수(금부정)
블로그 번역본 감성점수(자극)
유튜브 번역본 감성점수(금부정)
유튜브 번역본 감성점수(자극)

03

비정태적 CNN과 비지도
학습을 활용한 감성분석

Features

구글 딥러닝
블로그 딥러닝
유튜브 딥러닝

리뷰 감성분석

01 한국어 KNU 사전을 활용한 감성분석

- 한국어 KNU 사전을 기반으로 맛집 도메인에 특화한 한국어 감성 사전 구축
- 긍부정 리뷰를 그룹으로 분류 후 단어마다 score 부여
- 감성점수: 식당별 긍부정 점수를 합계 내어 긍부정 비율 계산, 높을수록 긍정적 리뷰가 많음.

Features

구글 리뷰 감성점수(KNU)
블로그 리뷰 감성점수(KNU)
유튜브 리뷰 감성점수(KNU)

리뷰 감성분석

02 X-ANew 사전을 활용한 영어 번역본 감성분석

- 구글 실시간 번역 Url 을 가져와 매크로 편집기에 VBA 코드를 추가하는 방식 사용
- 실시간 번역 함수 GoogleTranslate() 를 생성하여 크롤링한 블로그, 유튜브, 구글맵 리뷰 데이터 전부 **영어로 번역**
- 각 단어에 대해 score를 부여한 **X-ANew 사전**(단어에 대한 정서적 특성을 정량적으로 측정하기 위해 개발된 영어 단어 감정 사전)을 바탕으로 점수 부여
- 감성점수(금부정): 단어가 전달하는 감정의 긍정, 중립, 부정 판단 척도
- 감성점수(자극): 감정의 흥분도 혹은 자극 정도 산출

Features

구글 번역본 감성점수(금부정)
 구글 번역본 감성점수 (자극)
 블로그 번역본 감성점수(금부정)
 블로그 번역본 감성점수(자극)
 유튜브 번역본 감성점수(금부정)
 유튜브 번역본 감성점수(자극)

리뷰 감성분석

03

비정태적 CNN과 비지도 학습을 활용한 감성분석

배경

- 기존 모델은 맛집 리뷰에 특화되지 않아 한계 존재
- FastText 임베딩 벡터와 한국어 일반 담화 데이터셋 활용
- Non-static CNN(비정태적 CNN)으로 모델 성능 개선

전처리 과정

- 꼬꼬마(Kkma)로 문장 분리
 - 리뷰 내 여러 감성을 담은 문장 분리 (예: "맛있는데 비싸요~")
- 맞춤법, 띄어쓰기 교정 미수행
 - FastText의 OOV(out-of-vocabulary) 강점 활용

결과물

- 블로그, 구글, 유튜브 리뷰 데이터에 긍정(1), 부정(0) 라벨 부착

Features

구글 딥러닝
블로그 딥러닝
유튜브 딥러닝

상권 데이터 추가

“서울시 열린 데이터 광장”의 상권 분석 서비스 데이터를 통해 각 식당 주변 상권 단위의 행정동 데이터 수집

Features

- **길단위인구-행정동**: 총 유동인구 수, 연령대별 유동인구 수, 성별 유동인구 수, 시간대별 유동인구 수
- **상권변화지표-행정동**: 행정동별 운영 식당의 평균 영업 개월 수, 폐업 식당의 평균 영업 개월 수
- **상주인구-행정동**: 총 상주인구 수
- **직장인구-행정동**: 총 직장인구 수
→ 상주인구, 직장인구에서 연령대별, 성별 인구는 위와 중복되기에 사용하지 않음.
- **소득소비-행정동**: 총 지출 금액, 식료품 지출 금액, 유흥 및 여가 지출 금액,
- **아파트-행정동**: 아파트 평균 시가, 아파트 가구 수
- **집객시설-행정동**: 집객시설(약국, 종합병원, 학교, 행정기관 수)

데이터 전처리

1

SMOTE

- 불균형이 심한 데이터: 운영 식당 371개, 폐업 식당 35개
- train set에 SMOTE하여 불균형을 맞춰줌

2

ONEHOT ENCODING

- 출연한 방송이 폐업 여부에 미치는 영향을 관찰
- 출연 방송을 원핫 인코딩

3

COLUMN 선택

- 폐업여부와 상관관계가 높은 칼럼만 선택
- 65개 칼럼 → 47개 칼럼으로 축소

모델 선정

Logistic Regression

재현율과 F1 score을 기준으로 최적의 모델 선정

precision <-> recall trade off! 에서
폐업(1)을 폐업(1)으로 올바르게 분류하는 **재현율**이 더
중요하다고 판단!

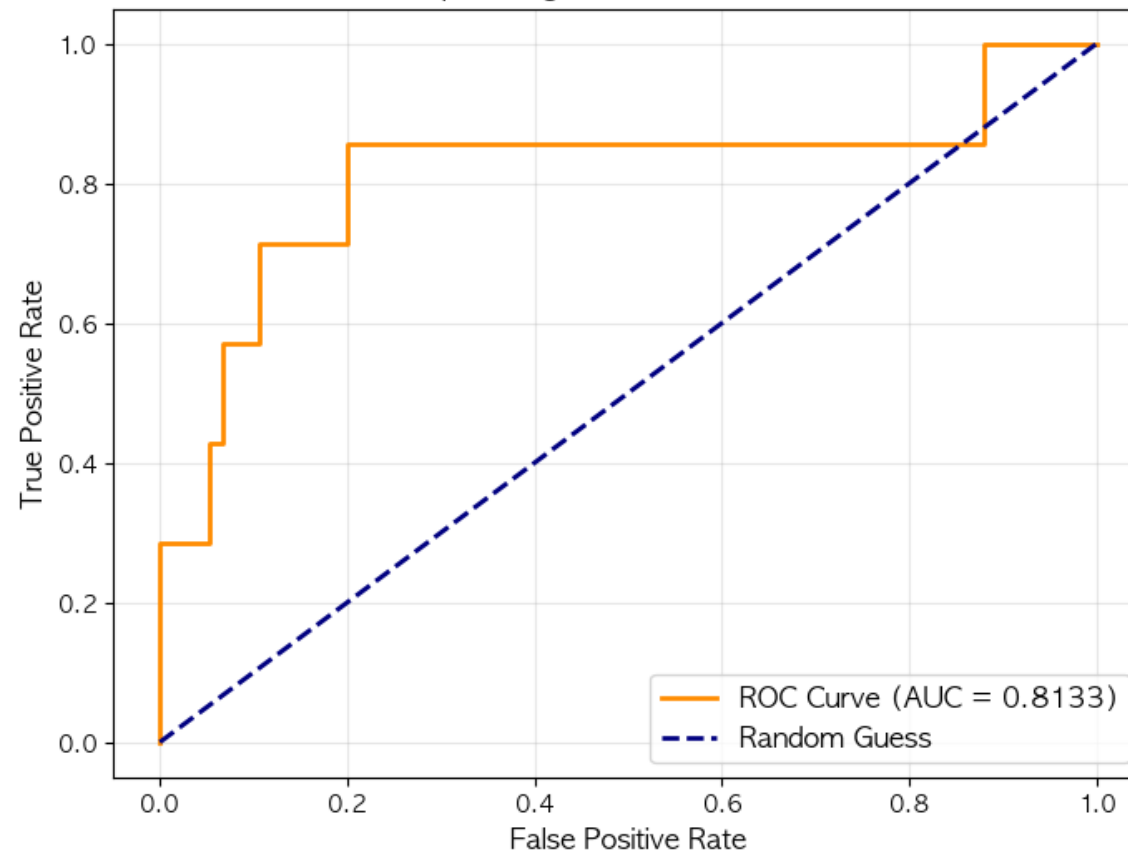
예측 성능

정확도 80.4%

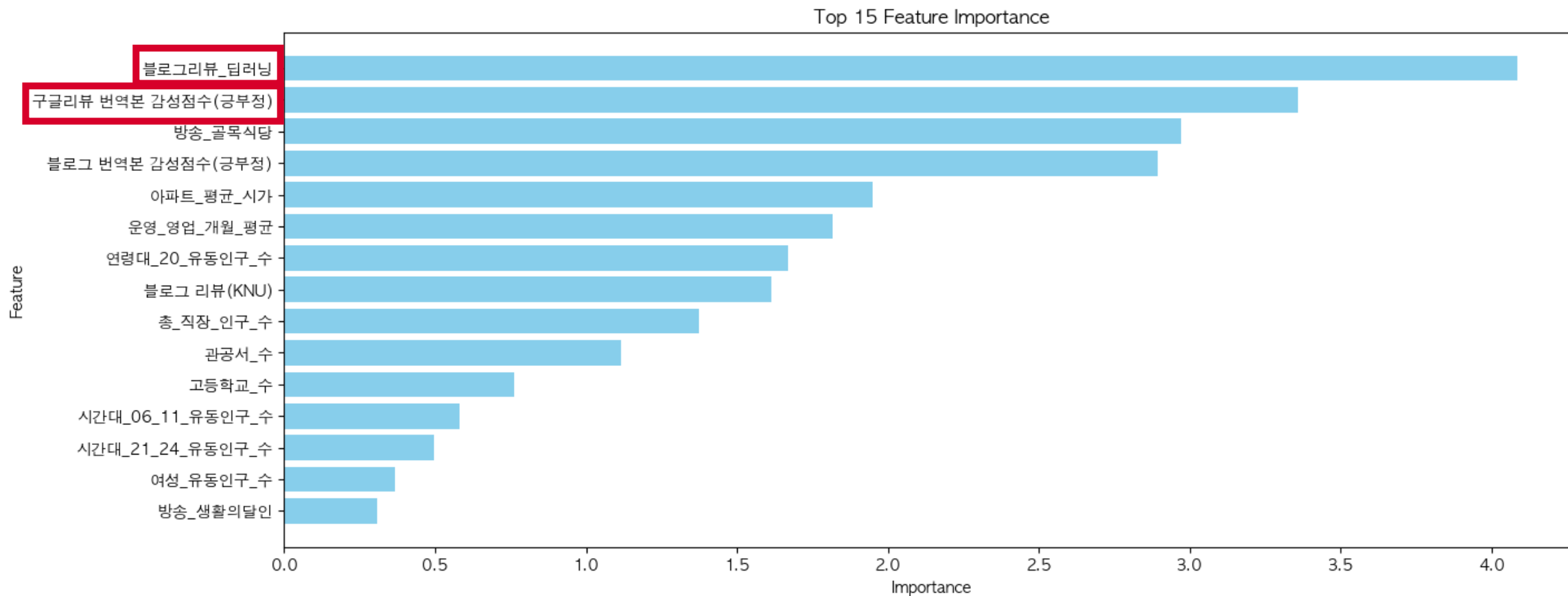
민감도, 재현율 85.7%

특이도 80%

Receiver Operating Characteristic (ROC) Curve



폐업 방지 solution –Feature Importance





폐업 방지 solution -<리뷰>

블로그 리뷰

1

✓ 긍정적인 블로그 리뷰가 가장 중요!

- 모델의 상위 15개 변수 중요도 중, 블로그 긍부정 점수를 나타내는 3개 지표가 모두 포함되어 있음.

→ 긍정적인 블로그 리뷰가 많아지도록 관리하는 것이 중요

ex. 블로그 체험단 활용 등

폐업 방지 solution -<리뷰>

구글 리뷰

2

✓ '구글리뷰 번역본 감성점수(금부정)'이 중요도 2위

→ **구글리뷰**도 식당 방문 의사결정에 중요한 영향

- 구글리뷰의 경우,
 - 한국어 기반 감성분석보다 영어사전을 활용한 감성분석의 폐업 여부와 상관관계가 높았음
- 외국인 이용자의 비율이 높은 구글맵 특성상, 영어 기반의 감성분석이 유의미한 분석 결과!

폐업 방지 solution -<리뷰>

유튜브 댓글 리뷰

3

- 유튜브에서 수집한 방송 원본 영상의 댓글, 식당 리뷰 영상에 대한 댓글 등은 폐업여부와 모두 상관관계가 낮은 변수로 예측 모델에서 제외됨
 - 식당 방문 과정에서 유튜브는 의사결정에 주요한 영향을 미치지 않을 것으로 해석
- 또한, 댓글에는 방송과 출연진에 대한 내용이 포함되어 있어 식당에 대한 관심이 분산됨.

폐업 방지 solution -<상권>

좋은 상권의 조건



교육

- 교육비 지출
- 초등학교 수
- 중학교 수

소비 연령대

- 30~40대 유동인구 수
- 20대보다 30~40대의
경제력이 중요함



지하철역

vs 버스 정류장

- 지하철역까지의 거리보다
버스 정류장의 밀도가 더
중요함!

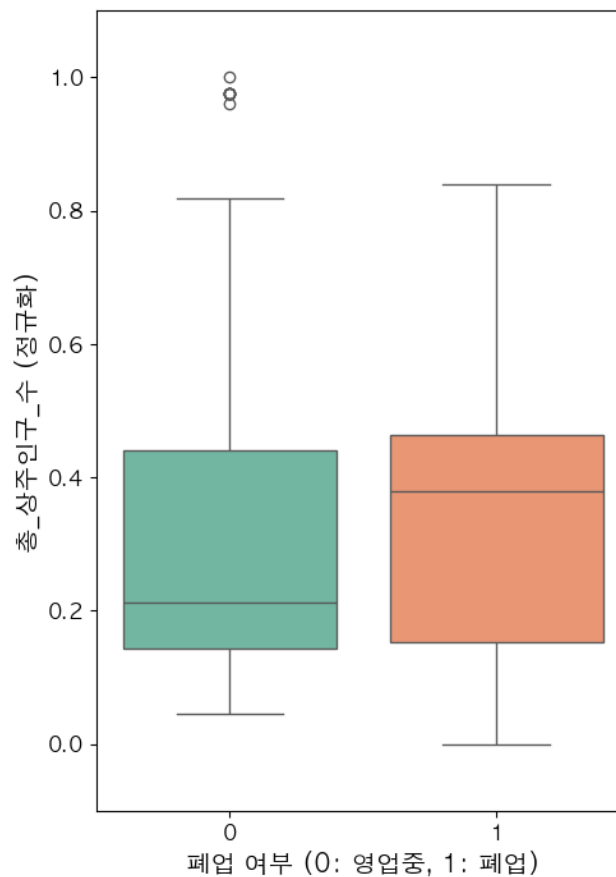
GOAL

폐업 방지 solution -<상권>

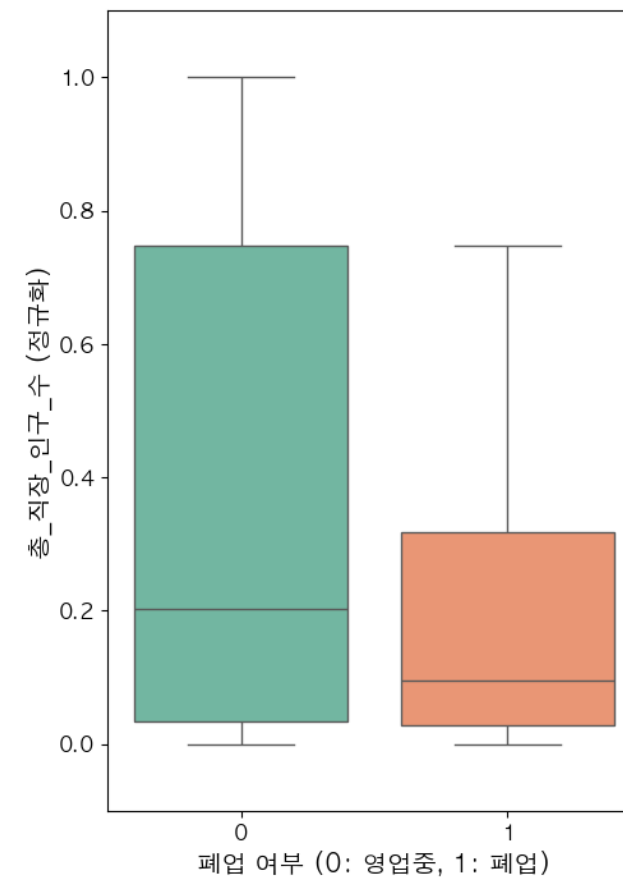
2. 상업지역 vs 주거지역 ?

- 폐업한 식당들이 운영 중인 식당에 비해 상주 인구가 많고 직장 인구가 적음
- 주거지역보다는 **직장 인구가 많은** 상업지역에서 영업을 지속될 확률이 높다.

폐업여부에 따른 총_상주인구_수 비교 (정규화)



폐업여부에 따른 총_직장_인구_수 비교 (정규화)



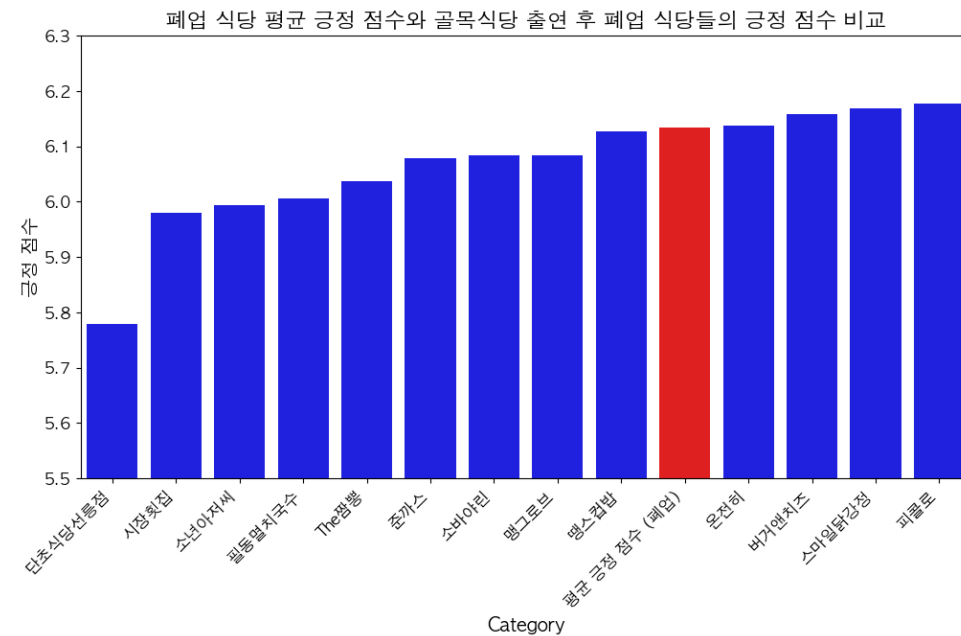
폐업 방지 solution -<방송>

분석 대상인 6개 방송 중 골목식당의 폐업 비율이 가장 높음.

- 식당의 장점을 더 많이 부각시키는 타 방송에 비해 **운영의 미흡한 점을 방송에 노출시켜야** 하는 방송의 특성상 방송 출연이 폐업으로 이어질 확률이 가장 높음.

폐업 식당들의 블로그 리뷰 긍정점수(빨간색)에 비해, 골목식당 출연 방송들의 긍정 점수가 대체로 낮음.

- 긍정 점수가 높은 4개 식당은 좋은 상권의 조건을 충족한 식당



기대효과

01 폐업 위기 관리

- 고객 리뷰 데이터와 폐업 여부의 유의미한 관계 확인
- 식당 개업 이후 고객의 목소리를 관리하여 폐업 위기 관리
- 폐업 위험군 식당 맞춤형 개선책을 통해 안정화된 운영과 생존율 향상에 기여할 수 있음

02 모델 확장

- 외식업을 넘어 숙박업 등 다른 서비스 산업에 적용 가능



감사합니다

DScover 메인프로젝트 B조