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

방준아[1] · 손광민[1] · 이소정[2] · 이현근[3†] · 조수빈[1]

성균관대학교 통계학과[1], CJ올리브네트웍스 DT융합연구소[2], CJ올리브네트웍스 빅데이터센터[3]

요 약

대한민국에서 치킨집은 전 세계 맥도날드 매장 수보다 많을 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 음식점의 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 업소별 개·폐업 신고 일자를 중심으로 다양한 변수를 수집하였다. 이후 다양한 분류 알고리즘을 적용하고, 예측 모형의 성능을 비교하였다. 그 결과, 인공신경망(Neural Networks)이 가장 높은 정확도를 보였지만 특이도와 민감도가 불균형적이었다. 이에 비해 유연판별분석(FDA)은 인공신경망보다 정확도는 낮지만, 상대적으로 균형적인 예측 성능을 보였다.

■ 중심어 : 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 Discriminant 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

2018년 11월 27일 접수; 2018년 11월 27일 수정본 접수; 2018년 12월 31일 게재 확정.
† 교신저자 contact@hyeongeun.com

Ⅰ. 서 론

대한민국에서 ‘은퇴 후 창업’이라 하면 곧바로 떠오르는 것이 치킨집일 것이다. 언론 보도에 따르면 전 세계 맥도날드 매장 수(3만 5천여 곳)보다 대한민국의 치킨집 수(3만 6천여 곳)가 더 많다고 한다[4].

우선 배달음식 하면 짜장면 혹은 치킨이 떠오를 정도로, 치킨은 매우 대중적인 음식이다. 국세청 부가가치세 신고 자료에 따르면 창업자 5명 중 2명은 치킨집이거나 편의점을 개업하는 것으로 나타나고 있다[9]. 박주영 숭실대 벤처중소기업 학과 교수는 이러한 사회적 현상에 대해서 “특히 화이트칼라 직종에 종사하다 은퇴한 창업자들의 경우 자신이 많이 소비했고 익숙한 업종을 중심으로 창업하려는 경향이 크다. 치킨, 술·유흥 업종의 경우 자주 이용했기 때문에 쉽게 생각하고 가게를 오픈하려고 하는 것일 수 있다.”라고 설명한다[5].

그러나 문제는 ‘폐업률’에 있다. 2018년 OECD 자료에 따르면, 우리나라 취업자 중 자영업 종사자의 비중은 25.4%로 OECD 평균인 14.8%에 비해 매우 높다[7]. 일하는 사람 4명 중 1명꼴로 자영업에 의존하고 있을 만큼, 자영업은 우리에게 매우 중요한 생계 수단이라고 볼 수 있다. 하지만 중소기업청 조사에 따르면, 소상공인의 62%가 3년을, 71%가 5년을 버티지 못하고 폐업하는 것으로 나타났다[4]. 특히, 통계청에 따르면 숙박 및 음식점업의 3년 생존율은 고작 30.2%, 5년은 17.9%에 불과한 것으로 집계됐다[11].

‘그렇다면 치킨집을 창업하기 이전에, 폐업 가능성을 미리 분석하는 방법은 없을까?’ 본 연구는 이러한 물음에서 출발하였다. 예비창업자들에게 그들이 선정한 창업 후보지의 조건에 따라 예상되는 폐업 시기를 예측하여 제공해줄 수 있다면, 창업하는 데 있어 조금 더 합리적인 의사결정을 내릴 수 있을 것이다. 따라서 본 연구는 치킨집의 3년 내 폐업 여부를 예측하는 모형을 수립하고자 하였다. 더불어 통계청의 자영업 현황분석에 따르면, 서울특별시의 인구 천 명당 사업자 수는 104개로 가장 높으므로 본 연구는 서울시를 분석 범위로 선정하였다[12].

Ⅱ. 연구 방법론

2.1 데이터 취득

치킨집의 개·폐업 관련 가용 데이터를 조사한 결과, 행정자치부에서 ‘호프/통닭’ 및 ‘통닭(치킨)’ 업태로 신고한 각 사업자의 ‘인허가 일자’와 ‘폐업 일자’를 포함하여 ‘영업상태(폐업 여부)’, 업소명, 주소 등의 자료를 제공하고 있었다. 여기서 소재지 주소가 서울특별시인 사업장을 대상으로 데이터를 취득하였다. 또한, 인허가 일자를 최초 영업 개시일로 가정하여 그날로부터 폐업한 날까지의 기간을 일 단위 영업 기간으로 가공하였다. 이를 통해 영업 기간이 3년 이상이거나 최근까지 폐업하지 않았을 경우 ‘Yes’, 영업 기간이 3년 미만이면 ‘No’로 산정하여 출력변수를 수립하였다.

이러한 기본적인 자영업 데이터 외에도 해당 업소의 폐업에 영향을 미칠 것으로 예상되는 요인들을 수집하였다. 각 데이터의 출처는 입력변수의 설명과 함께 소개하였다.

2.2 입력 변수

예측력이 높은 모형을 구축하기 위해서는 예측 대상의 변화를 잘 설명할 수 있는 입력 변수를 확보하는 것이 매우 중요하다. 따라서 본 연구는 치킨집의 폐업에 영향을 미칠 수 있는 다양한 요인을 입력 변수로 수집하고 가공하였다. 이를 <표 1>과 같이 ‘업체 특성’, ‘지역적 특성’, ‘경쟁 업체 현황’, ‘상권 특성’, ‘경제적 변수’ 등

〈표 1〉 예측 모형 입력 변수

구분	변수	비고
업체 특성	유명 브랜드 여부	범주형
	시설 총 규모	수치형 (㎡)
	다중 이용 업소 여부	범주형
	개업 장소의 제곱미터(㎡)당 개별 공시지가	수치형 (원)
	개업 장소의 총 지가	수치형 (원)
지역적 특성	용도지역 : 상업지역 여부	범주형
	용도지역 : 그린벨트 여부	범주형
	용도지역 : 공업지역 여부	범주형
경쟁 업체 현황	서울 전체 폐업 업체 수	수치형 (개수)
	서울 전체 개업 업체 수	수치형 (개수)
	서울 전체 폐업 비율	수치형
	영업 중이며 가장 가까운 경쟁 업체와의 거리(m)	수치형 (m)
	반경 1km 內 경쟁 업체 수	수치형 (개수)
	반경 1km 內 개업 업체 수	수치형 (개수)
	반경 1km 內 폐업 업체 수	수치형 (개수)
	반경 1km 內 폐업 비율	수치형
	반경 1km 內 최근 3개월간 개업 업체 수	수치형 (개수)
	반경 1km 內 최근 6개월간 개업 업체 수	수치형 (개수)
	반경 1km 內 최근 1년간 개업 업체 수	수치형 (개수)
	반경 1km 內 최근 3년간 개업 업체 수	수치형 (개수)
	반경 1km 內 최근 5년간 개업 업체 수	수치형 (개수)
	반경 1km 內 최근 3개월간 폐업 업체 수	수치형 (개수)
	반경 1km 內 최근 6개월간 폐업 업체 수	수치형 (개수)
	반경 1km 內 최근 1년간 폐업 업체 수	수치형 (개수)
	반경 1km 內 최근 3년간 폐업 업체 수	수치형 (개수)
	반경 1km 內 최근 5년간 폐업 업체 수	수치형 (개수)
상권 특성	반경 1km 內 학교 수	수치형 (개수)
	반경 1km 內 전철역 수	수치형 (개수)
	반경 1km 內 공공기관 수	수치형 (개수)
	반경 1km 內 관광명소 수	수치형 (개수)
	반경 1km 內 주민등록인구 수 : 합계	수치형 (명)
	반경 1km 內 주민등록인구 수 : 10대 미만	수치형 (명)
	반경 1km 內 주민등록인구 수 : 10대 초반	수치형 (명)
	(생략)	...
	반경 1km 內 주민등록인구 수 : 60대 초반	수치형 (명)
	반경 1km 內 주민등록인구 수 : 65세 이상	수치형 (명)
	반경 1km 內 모든 전철역 승하차량의 과거 6개월 평균	수치형 (명)
	최근접 전철역 승하차량의 과거 6개월 평균 (명) ÷ 해당 전철역과의 거리	수치형 (m)
경제적 변수	선행종합지수	수치형
	선행종합지수 전월 비 (당월 선행종합지수 - 직전월 선행종합지수)	수치형
	선행종합지수 전월 비교 – 감소 여부	범주형
	선행종합지수 전월 비교 – 증가 여부	범주형
	소비자기대지수	수치형
	소비자기대지수 6개월 평균	수치형
	가계대출금리	수치형

크게 5가지로 정리할 수 있다.

2.2.1 업체 특성

먼저 ‘유명 브랜드 여부’를 업소명 데이터를 활용해 가공하였다. 치킨집을 포함한 대다수 자영업의 창업 방식은 유명 브랜드의 프랜차이즈로 개업하거나, 그렇지 않거나로 나눌 수 있다. 소비자 입장에서는 잘 알려진 브랜드가 친숙할 것이기 때문에, 긍정적이든 부정적이든 해당 업소의 영업에 영향을 미칠 것으로 예상하였다.

유명 브랜드의 기준은 한국공정거래조정원의 보고서를 참고하여 아래 <표 2>와 같이 15개 프랜차이즈로 결정하였다. 15개 프랜차이즈의 브랜드 중 하나라도 업체별 업소명에 포함되어 있으면 ‘1’, 그렇지 않으면 ‘0’으로 가공하여 ‘유명 브랜드 여부’를 의미하는 범주형 변수를 생성하였다.

〈표 2〉 가맹점 수 상위 15개 브랜드 (한국공정거래조정원)

순번	상호	영업표지 (브랜드)
1	㈜제너시스비비큐	비비큐
2	㈜페리카나	페리카나
3	㈜혜인식품	네네치킨
4	교촌에프앤비(주)	교촌치킨
5	㈜한국일오삼	처갓집양념치킨
6	㈜지앤푸드	굽네치킨
7	(유)비에이치씨	비에이치씨(BHC)
8	㈜농협목우촌	또래오래
9	호식이두마리치킨	호식이두마리치킨
10	㈜멕시카나	멕시카나
11	해마로푸드서비스㈜	맘스터치
12	㈜훌랄라	훌랄라참숯바베큐
13	㈜비케이부어코리아	부어치킨
14	㈜맥시칸	맥시칸치킨
15	지코바	지코바양념치킨

행정자치부의 원천 데이터에는 ‘시설 총 규모(m^3)’ 정보를 포함하고 있으며, 이 변수는 업소의 규모를 나타낸다고 볼 수 있다. 규모가 클수록 인테리어와 집기 등 초기 시설투자 비용이나 관련 권리금을 크게 지급했을 것이라고 예상할 수 있다. 반면, 사업장 규모가 매우 작다면 홀 영업을 하지 않는 Take-out 또는 배달 전문업체일 가능성이 크다. 따라서 ‘시설 총 규모’ 또한 폐업 예측에 유의미한 변수가 될 것이다.

상가 임대료와 같은 고정비용은 영업의 지속 가능성에 큰 영향을 미칠 것이다. 정확한 임대료는 알 수 없으나, 해당 업소의 ‘개별 공시지가(m^3)’ 데이터를 국토교통부로부터 얻을 수 있었다. ‘개업 장소의 m^3당 공시지가’와 ‘시설 총 규모(m^3)’ 값과 곱함으로써 ‘개업 장소의 총 지가’ 파생변수를 생성하였다.

‘다중 이용 업소 여부’ 또한 원천 데이터에 포함되어 있다. 이는 불특정 다수가 이용하는 영업 중 화재 등 재난 발생 시 생명, 신체 그리고 재산상의 피해 발생 가능성이 높은 업소를 말한다[3]. 즉, 많은 사람이 이용할 수 있는 사업장을 의미하므로 매출 규모와 관계가 있을 것으로 예상하였다.

2.2.2 지역적 특성

행정자치부 자영업 데이터에는 사업장 소재지의 ‘용도지역’을 기본적으로 제공하고 있다. “용도지역이란 토지의 이용 및 건축물의 용도, 건폐율, 용적률, 높이 등을 제한함으로써 토지를 경제적·효율적으로 이용하고 공공복리의 증진을 도모하기 위하여 서로 중복되지 아니하게 도시·군 관리계획으로 결정하는 지역을 말한다[2].” 서울특별시와 같은 도시의 ‘용도지역’은 크게 4가지로 <표 3>과 같이 구분한다.

이러한 용도지역에 따라 매출 등 영업에 영향을 줄 것으로 예상할 수 있다. 이에 주거지역을 기준 범주로 설정하고 각 용도지역의 여부를 변수화하였다.

〈표 3〉 도시지역의 용도지역과 그 목적
(국토의 계획 및 이용에 관한 법률)

용도지역	지정목적
주거지역	거주의 안녕과 건전한 생활환경의 보호를 위하여 필요한 지역
상업지역	상업이나 그 밖의 업무의 편익을 증진하기 위하여 필요한 지역
공업지역	공업의 편익을 증진하기 위하여 필요한 지역
녹지지역	자연환경·농지 및 산림의 보호, 보건위생, 보안과 도시의 무질서한 확산을 방지하기 위하여 녹지의 보전이 필요한 지역

2.2.3 경쟁 업체

경쟁 업체 현황은 해당 사업장의 영업 지속력에 큰 영향을 미칠 것이다. 이에 해당 업소의 개업 인가일 기준으로 '서울 전체 개업 / 폐업 업체 수'와 '서울 전체 폐업 비율'을 변수로 가공하였다.

또한, 업소별 '소재지 주소' 데이터를 바탕으로 주변 경쟁 업체 현황 변수를 가공하였다. 방법은 다음과 같다. 첫째, 구글의 'Geocoding API'를 활용하여 모든 주소를 '위경도'로 변환한다. 둘째, 서로 다른 주소의 위경도 쌍에 'Haversine' 수식을 적용하여 모든 업체의 거리 행렬(distance matrix)을 산출한다. Haversine 수식은 <수식 1>과 같다[27]. 셋째, 거리행렬을 이용하여 각 업소(행)를 기준으로 다른 나머지(열)와의 거리가 1km 이하인 업체를 추출한다. 이렇게 추출한 업체의 '인허가 일자'와 '폐업 일자'를 바탕으로 '반경 1km 內 최근 3개월 / 6개월 / 1년 / 3년 / 5년간 경쟁 / 개업 / 폐업 업체 수', '반경 1km 內 폐업 비율' 등의 변수를 생성하였다.

$$\begin{aligned} \mathrm{hav}(\Theta) &= \mathrm{hav}(\varphi_2-\varphi_1)+ \\ &\quad \cos(\varphi_1)\cos(\varphi_2)hav(\lambda_2-\lambda_1) \end{aligned} \quad (1)$$

$$where\ \varphi_1,\varphi_2: \text{ latitude of point 1 and 2,}$$
$$\lambda_1,\lambda_2: \text{ longitude of point 1 and 2.}$$

주변 치킨집과의 거리가 가까울수록 경쟁이 치열하여 향후 폐업 여부에 큰 영향을 미칠 것이다. 이에 따라, 앞서 구한 거리행렬을 바탕으로 '영업 중이며 가장 가까운 경쟁 업체와의 거리(m)'를 또 하나의 입력 변수로 활용하였다.

2.2.4 상권 특성

치킨집의 폐업 여부는 그 소재지의 상권에 크게 좌우될 것이다. 따라서, 예측 모형에 상권의 특성을 반영할 수 있는 가용 데이터를 탐색하였다. 그 결과 '다음 지도 API'를 이용한 주변 상권 정보와 공공 데이터의 '주민등록인구', '전철역 승하차량' 등 크게 3종의 데이터를 수집할 수 있었다.

먼저 해당 업체 주소를 중심으로 '반경 1km 內 학교 / 전철역 / 공공기관 / 관광명소 수'를 수집하였다. '다음 지도 API'에는 이 외에도 대형마트, 편의점, 주유소 등 다양한 종류의 지도 데이터를 조회할 수 있으나, 아쉽게도 과거 이력은 제공하지 않는다. 따라서 기간에 따른 변동성이 낮을 것으로 판단되는 종류의 장소 데이터만 사용하였다.

잠재 고객이 많거나 적음에 따라 매출도 크게 좌우될 것이다. 사람들이 치킨을 주로 배달하여 먹는다고 생각한다면, 업체 소재지 인근 거주민 수를 잠재 고객으로 볼 수 있다. 따라서 서울특별시 통계정보시스템에서 제공하는 행정동별 '주민등록인구'를 활용하였다. 여기서 주의할 점은 소재지 주소는 법정동이지만 주민등록인구는 행정동 기준으로 제공된다는 것이다. 이러한 불일치 문제를 고려하기 위해 각 업체의 주소를 중심으로 '반경 1km 內 연령대별 주민등록인구수'를 수집하였다. 인구의 변화를 고려하여 6개월 평균으로 변수를 가공하였다. 경쟁 업체 현황과 마찬가지로 행정동의 중심지에 해당하는 위경도를 수집하여 각 업소와 거리를 산출하는 방법을 적용하였다.

배달 이외에도 치킨집의 또 다른 잠재 고객으로는 해당 업소 방문객이 있다. 이와 관련하여 주민등록인구 외에 가용한 '전철역 승하차량' 데이터를 활용하였다. 마찬가지로 전철역 주소의 위경도를 수집하여 '반경 1km 內 모든 전철역 승하차량'을 구성한 후, 변동성을 고려하여 6개월 평균을 적용하였다. 또한, 반경 1km 內 전철역 존재하지 않는 업소를 고려하여 '최근접 전철역 승하차량'을 추가하였다.

2.2.5 경제적 변수

치킨집의 폐업을 설명할 수 있는 요인으로 국가의 전반적인 경제 상황은 빠질 수 없는 정보이다. 이를 보여주는 경제적 변수로 '선행종합지수', '소비자기대지수' 그리고 '가계대출금리'를 선정하였다.

'선행종합지수'의 경우 통계청에서 미래 경제상황을 보여줄 수 있다고 판단한 9가지 지표(구인 구직 비율, 재고순환지표, 소비자기대지수, 기계류 내수출하지수, 건설수주액, 수출입물가비율, 국제원자재 가격지수, 코스피지수, 장단기금리 차)를 종합하여 작성한다. 본 연구에서는 이를 활용하기 위하여 '전월 선행종합지수와의 차이'를 파생변수로 추가했다. '전월 대비 증감 여부' 또한 범주형 변수로 가공하여 선행종합지수에서 최대한의 정보를 추출하려 노력했다[10].

'소비자기대지수'의 경우 통계청에서 발표한다. 이 지수는 소비자가 생각하는 6개월 후의 경기에 대한 예상 지표이다. 구체적으로는 경기상황과 소비지출, 생활 형편 등에 대해 매우 좋음부터 매우 나쁨까지 5단계로 조사한 후 이를 가중평균한 수치이다. 치킨은 대중과 친숙한 음식으로 소비자들의 경기에 대한 인식이 치킨집의 이익과 상관성이 높을 것으로 판단하여 변수로 활용하였다. 소비자기대지수변수 또한 그 자체를 변수로 활용하면서 이전 6개월 평균을 활용해 최대한의 정보를 추출하려고 노력했다[6].

한국은행에서 공표하는 '가계대출금리'는 현재 시점에서의 대출금리를 보여주는 것으로 미래 경제에 영향을 미치는 정보라고 판단하였다. 금리가 상승하면 대출에 의한 경제적 부담이 늘어 소비자들의 외식 빈도가 줄어들 것으로 예상하였다. 또한, 창업 비용이 금리가 낮을 때보다 많이 발생할 수 있으며 영업 지속의 어려움을 겪는 등 다양한 영향이 있을 것으로 판단했다.

2.3. 데이터 전처리 및 분할

초기 데이터를 살펴본 결과, 치킨집(호프/통닭 등) 업태를 선택했음에도 불구하고 업소명이 '홍어 전문점', '쭈꾸미 포차', '참이맛 감자탕' 등 전혀 관련 없는 사업장들이 다수 포함되어 있었다. 따라서 데이터 구축 단계에서 업소명을 확인하여 치킨집과 거리가 먼 데이터를 제외하였다. 다시 말해, 연구 목적과 범위에 맞지 않는 데이터를 확인, 제거하는 일련의 과정을 진행하였다.

주민등록인구와 전철역 승하차량의 수집 가용 기간과 개업 3년 후의 폐업 여부 등 전반적인 데이터 정합성을 고려하여 2009년 1월 1일부터 2014년 12월 31일 사이에 인가된 사업장을 연구 범위로 한정하였다.

마지막으로 변수별 결측 현황을 살펴본 결과, 46개 사업장에서 '시설 총 규모'가 누락되어 있었다. 또한, 소재지 주소와 '주민등록인구'의 주소가 일치하지 않은 사업장이 84개 존재하였다. 결과적으로 7,684건 중 1.7%인 130개 데이터에서 결측이 발견되었다. 본 연구에서는 발견된 결측 데이터 수가 큰 비중을 차지하지 않는다고 판단하여 이들을 제거한 후 예측 모형을 구축하였다.

최종적으로 분석에 사용할 데이터는 총 7,554개 사업장으로 구성되었다. 본 연구는 예측 모형의 성능 평가를 위하여 8:2의 비율로 학습 데이터와 검증 데이터로 나누었다. 추가로, 출력변

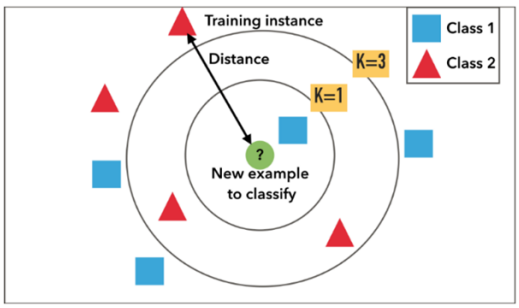
수의 비율을 살펴본 결과 'Yes'는 약 69.3%이고 'No'는 30.7%로 산출됐다. 이러한 출력변수의 불균형적 특성을 학습 데이터와 검증 데이터에도 유지하며 모형을 개발하기 위해, 데이터 분할 시 무작위 층화추출 방법을 사용하였다. 최종적으로 학습 데이터에는 6,044개 사업장이, 검증 데이터에는 1,510개 사업장이 추출되었다.

2.4 분류 예측 알고리즘

본 연구의 예측 대상이자 출력변수(Y)는 해당 치킨집의 계속 영업(Yes) 또는 폐업(No) 여부로 가공된 범주형 변수이다. 따라서 통계적 학습 또는 기계학습에서 지도 학습(Supervised Learning) 중 다양한 분류 예측(Classification) 알고리즘을 사용하였다.

2.4.1 k-Nearest Neighbor

kNN(k-Nearest Neighbor) 알고리즘은 k 개의 가장 가까운 이웃 데이터들과의 거리를 측정하여 분류하는 기계학습 기법이다. 비교하는 목표 데이터와 주변 이웃 데이터 사이의 유클리디안 거리를 측정하여, k=1일 경우 가장 가까운 데이터의 분류를 따르며, k=3일 경우 가장 가까운 상위 3개 데이터를 선별하여 가장 많은 수의 분류 결과를 따른다. 즉, 최적의 k값을 찾아서 데이터를 분류하는 것이 kNN의 성능을 최대화하는 방법이라고 볼 수 있다[23].



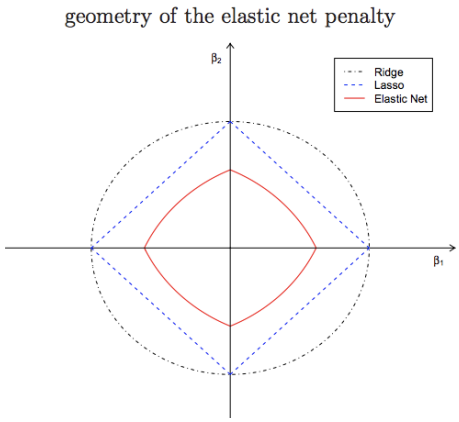
〈그림 1〉 k-Nearest Neighbor[14]

2.4.2 Ridge Regression & Elastic Net

Ridge 회귀모형은 평균 제곱 오차를 최소화하면서 L2 정규화를 적용한 모형이다. 하이퍼 파라미터 λ는 제약 조건의 정도를 결정해 주며, 이때 주의해야 할 점은 모형에서 자동으로 학습되지 않기 때문에 직접 값을 설정해야 한다는 것이다. λ가 0일 때, 제약 효과가 없어 일반 선형회귀 모형과 동일한 결과를 얻게 되고, 반면 λ의 값이 커질수록 불필요한 변수의 영향이 줄어 모형을 단순화하는 효과가 있다. 단, 계수가 정확히 0으로 수렴하는 것은 아니므로 변수 선택의 효과는 없다. 일반적으로 Ridge 회귀모형은 변수 간 상관관계가 높더라도 성능에는 큰 영향을 미치진 않는다는 장점이 있다[22].

Elastic Net 회귀모형은 L1, L2 정규화를 동시에 적용하는 모형으로 Ridge와 Lasso의 절충안이라고 생각할 수 있다. λ_1, λ_2 두 개의 하이퍼 파라미터를 가지며 $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$라고 정의하면, 다음과 같다.

$$\hat{\beta} = \operatorname{argmin}_{\beta} |\mathbf{y} - \mathbf{X}\beta|^2, \text{ subject to } (1-\alpha)|\beta|_1 + \alpha|\beta|^2 \le t \text{ for some } t. \quad (2)$$

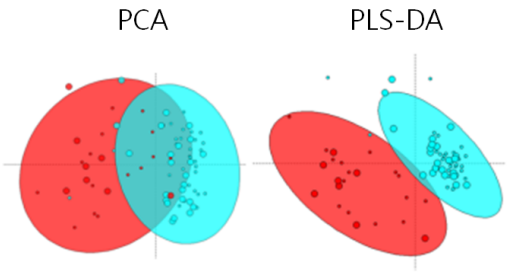


〈그림 2〉 Ridge, Lasso, Elastic Net 비교[30]

<수식 2>에서 $\alpha = 0$일 때 Ridge 회귀모형과 같고, 반대로 $\alpha = 1$일 때 Lasso 회귀모형과 같으며 0과 1 사이의 값을 가지면 Ridge와 Lasso의 결합인 Elastic Net이다. Elastic Net 회귀모형은 변수 간 상관관계를 반영하여 상관관계가 큰 변수를 동시에 선택하거나 배제한다. 일반적으로 Elastic Net 회귀모형은 데이터셋의 크기가 클 때 효율적으로 작동하는 장점이 있다[30].

2.4.3 Partial Least Square Discriminant Analysis

Partial Least Square Discriminant Analysis (PLSDA)는 Herman O. A. Wold에 의해 1980년대에 고안된 PLS에 기반한 분류모형이다. 우선 PLS는 PCA의 확장이라고 이해할 수 있다. PCA의 경우에는 단순히 입력변수만 고려해 분산이 가장 큰 선형결합을 찾는 것을 목표로 하지만, PLS의 경우 입력변수와 출력변수의 공분산을 고려한 선형결합을 찾는 것을 목표로 한다. 다시 말해, PLS는 차원을 축소하는 동시에 출력변수를 잘 설명하는 입력변수를 찾아주는 방식이다. 따라서 <그림 3>과 같이, 차원을 축소한 후 분류 예측 시 PCA보다 성능이 좋을 것으로 기대할 수 있다.



〈그림 3〉 PLSDA[13]

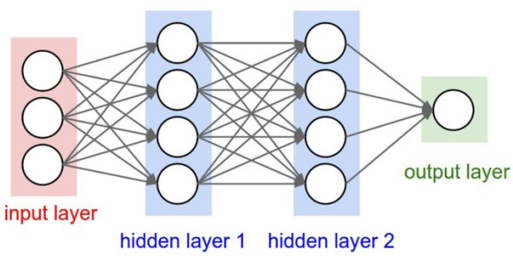
PLS는 기저의 입력변수와 출력변수의 관계를 찾아내는 데 쓰인다. 즉, 잠재변수를 찾은 후 모형에 입력변수로 사용하는 방식이다. 이때 찾아낸 잠재변수는 출력변수를 잘 설명하는 입력변수의 결합이라고 할 수 있다. 따라서 PLS는 변수가 많아 차원을 축소해야 할 필요가 있는 상황에서 좋은 방법이다[29].

2.4.4 Flexible Discriminant Analysis

Flexible Discriminant Analysis 즉, FDA는 쉽게 말해서 LDA의 확장이라고 생각할 수 있다. 두 모형 모두 데이터 분류를 위한 함수를 만든다는 것이 공통되지만, 두 모형의 차이는 함수의 형태에 있다. LDA는 선형함수를 가지지만, FDA는 굽은 형태의 함수를 가진다. 구체적으로, FDA는 각 클래스의 중심값을 정한 후 새로운 데이터에 적용하여 중심값에서 마할라노비스 거리가 최소화되도록 하는 비선형 판별함수를 구축하는 알고리즘이다. 이로써 FDA는 LDA의 유연성이 떨어지는 문제를 보완한 방법이다. 일반적으로 FDA는 데이터 분포가정이 불필요하며 Factor의 개수가 2개보다 더 많을 때 성능이 좋다고 알려져 있다[21].

2.4.5 Neural Networks

인공신경망(Neural Network)은 인간의 뇌에 있는 신경망 구조를 기반으로 만들어진 알고리즘으로, 가장 기본적으로 원하는 데이터를 넣는 입력층(input layer), 중간단계인 은닉층(hidden layer) 그리고 결과를 보여주는 출력층(output layer)으로 구성된다.



〈그림 4〉 Neural Networks[16]

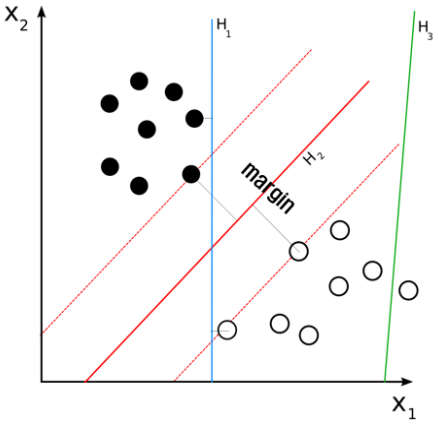
모든 레이어는 각각 다른 수의 노드들로 구성되어있으며 활성 함수를 통하여 임계치(threshold)를 넘는 정보들만이 다음 층으로 통과한다. 이를 통해 필요한 정보들만 도출한 후 마지막 출력층에서 결과를 예측할 수 있다[26].

인공신경망은 기존 기계학습 기법에 비하여 좋은 성능을 보이면서 주목받기 시작하였다. 특히 컴퓨터 비전의 분류 문제를 사람보다 더 좋은 성능으로 구분할 수 있게 되면서 다른 분야에서도 인공신경망은 빠르게 확산하고 있다.

2.4.6 Support Vector Machine

Support Vector Machine(SVM)은 이원 분류를 위한 지도학습 기반 기계학습 기법으로 주로 회귀분석과 분류 분야에서 활용되고 있다.

SVM은 1960년 Vapnik에 의해 제안된 통계적 학습방식으로, 패턴분류 문제를 해결할 수 있는 최적의 분류 경계면을 제공하는 알고리즘이다. 또한 커널 기법을 이용하여 비선형 분류 경계를 찾아낼 수 있다. 이 알고리즘은 주어진 데이터 집합을 바탕으로 하여 새로운 데이터가 어느 범주에 속할지 판단하는 비확률적 이진 선형분류 모형을 만든다.



〈그림 5〉 SVM 모형도[15]

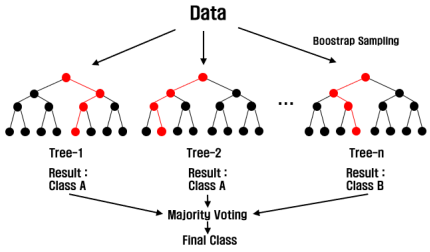
SVM은 데이터를 사상된 공간으로 표현하고, 경계를 통해 구분한다. <그림 5>와 같이 가장 큰 폭(margin)을 가진 경계를 찾는 것이 알고리즘의 목표이다.

특히, SVM은 인공신경망 기법의 문제점으로 지적되는 과적합 문제를 제약조건을 통해 피할 수 있으며, 다른 모형에 비하여 함수 근사에 있어서 이상값에 둔감하다고 알려져 있다[28].

2.4.7 Random Forest

Random Forest는 분류, 회귀분석 등에 사용되는 앙상블 학습방법의 일종으로 여러 개의 의사결정나무를 임의로 학습하는 방식이다. Random Forest는 다수의 의사결정나무를 구성하는 학습 단계와 입력변수가 들어왔을 때분류하거나 예측하는 검증 단계로 구성되어 있다.

Random Forest는 의사결정나무에 기반한 모형이다. 의사결정나무는 노드와 가지로 구성되며, 이는 입력변수의 값에 해당한다. 그리고 마지막 노드는 출력변수의 예측값을 결정한다. 이때, 주로 쓰이는 방식은 CART 알고리즘이다. CART는 의사결정나무의 마지막 노드를 제외한 모든 노드에서 입력변수를 순차적으로 할당하는 방식이다[25].



〈그림 6〉 Random Forest 모형도[8]

<그림 6>과 같이 Random Forest는 의사결정나무를 확장한 모형이며, 다수의 부트스트랩 샘플을 이용한다. 각 샘플마다 의사결정나무를 학

습한 후, 신규데이터가 입력되었을 때 의사결정 나무 결과 중 가장 많은 빈도를 보인 결과를 이용하여 최종적인 출력값을 예측한다.

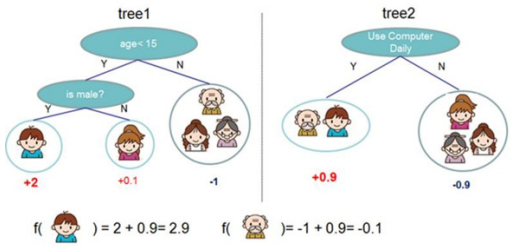
2.4.8 AdaBoost

AdaBoost는 성능을 향상시키기 위하여 약한 학습 알고리즘의 결과물들에 가중치를 두어 결합하는 방법이다. 잘못 분류된 결과들은 이어지는 약한 분류기들에 의해 재분류되어 결과적으로 좋은 성능을 보인다. 하지만, AdaBoost는 잡음이 많은 데이터와 이상값에 취약한 모습을 보이기도 한다.

AdaBoost는 인공신경망이나 SVM과는 다르게, 학습 과정에서 모형의 예측 능력을 향상시킬 것으로 생각되는 특성들만 선택한다. 따라서 차원 수를 줄이는 동시에 불필요한 특성들은 고려하지 않음으로써 수행 시간이 줄어든다는 장점이 있다[19].

2.4.9 XGBoost

XGBoost는 다른 부스팅 방법과는 다르게 병렬처리가 가능하여 속도가 빠르다는 장점이 있다. 또한 다양한 하이퍼 파라미터를 통해 모형에 대한 평가함수를 조절할 수 있으며, 병렬처리를 위한 프로세서 코어의 개수 등 여러 옵션을 조정하며 알고리즘에 적용할 수 있다. 그리고 가지치기를 통해 과적합을 방지할 수 있다고 알려져 있다.



〈그림 7〉 Boosting 알고리즘[17]

XGBoost는 나무 모형기반으로 CART 알고리즘 방식을 사용한다. XGBoost는 <그림 7>과 같이 구성된 나무 모형을 통해 예측된 결과를 종합하여 최종 결과를 도출하는 방식이다[17].

즉, XGBoost는 한 번에 여러 나무 모형을 구축하는 것이 아니라, 초기 나무 모형을 적합 시킨 후 생성된 잔차에 대하여 다음 나무 모형을 적합 시키는 방법을 반복함으로써 최종 모형을 결정한다. 특히, XGBoost는 가중치 학습 과정이 포함되는데, 이는 모형들의 단순평균보다는 가중평균을 적용하는 것이 더 좋은 결과를 나타낸다고 알려져 있기 때문이다[17].

Ⅲ. 모델링 결과

3.1 예측 성능의 검증 방법

분류 예측 문제에서, 범주형 출력변수의 특정 클래스 개수가 다른 클래스와 비교하여 매우 많을 수 있다. 이런 경우를 분류 비대칭 문제(Class Imbalance Problem)라고 한다[24].

본 연구의 데이터는 69.3%의 ‘Yes’ 클래스와 30.7%의 ‘No’ 클래스로 구성되어 있다. 즉, ‘Yes’ 클래스가 ‘No’ 클래스보다 2.26배 많았다. 이런 상황에서 예측 알고리즘은 학습 데이터를 ‘Yes’ 클래스로 분류하려는 경향이 있다. 예측 알고리즘은 오차를 최소화하는 함수를 만드는 과정이기 때문이다. 극단적인 예로, 이 데이터를 전부 ‘Yes’라 출력하면 69.3%의 예측 정확도(Accuracy)를 갖게 된다. 그러나 ‘No’ 클래스에 대한 예측 정확도는 0%이므로 예측 모형의 제 기능을 하지 못한다고 평가할 수 있다.

이러한 예측의 분류 비대칭 문제를 완화하고자, 본 연구에서는 ‘Sub Sampling’, ‘AUC’ 그리고 ‘Balanced Accuracy’를 고려하여 예측 알고리즘을 학습하고 그 결과를 검증하였다.

3.1.1 Sub-sampling

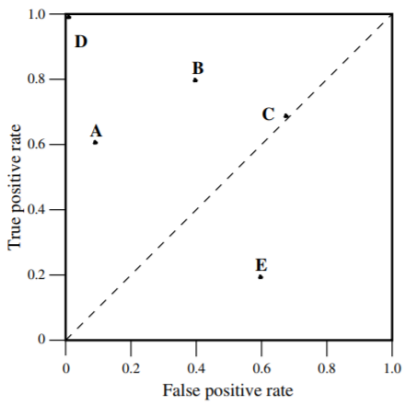
Sub-sampling은 예측 알고리즘이 더 큰 비중의 클래스로 예측하려는 경향을 줄이는 대표적인 방법이다. 일반적으로 Down-sampling과 Up-sampling이 많이 채택된다.

먼저 Down-sampling은 개수가 많은 클래스를 무작위 추출하여 두 클래스의 비중이 동등하도록 학습 데이터를 구성하는 방식이다. 가령 총 150개 데이터 중 큰 클래스가 100건이고 적은 클래스가 50건이면, 큰 클래스에서 무작위로 50건만 추출하여 모형 학습 시 총 100건의 데이터만 사용한다. Up-sampling은 이와 반대로, 적은 클래스가 큰 클래스만큼의 수를 갖도록 무작위로 복제하는 방법이다.

결과적으로 Sub-sampling은 무작위 추출을 활용해 학습 데이터의 모든 클래스가 같은 비율을 갖게 함으로써, 예측 알고리즘이 단순히 큰 클래스로 편향된 결과를 출력하지 않도록 유도할 수 있다.

3.1.2 Area Under the Curve (AUC)

AUC는 분류 예측 모형의 성능을 균형적으로 평가할 수 있는 지표 중 하나로, ROC(Receiver Operating Characteristic) 곡선의 하단 면적을 의미한다[18].



〈그림 8〉 ROC 곡선[18]

본 연구의 데이터를 고려했을 때, ROC 곡선은 'Yes' 케이스 예측력을 나타내는 민감도(Sensitivity, 또는 True Positive Rate)와 'No' 케이스의 예측력을 나타내는 특이도(Specificity, 또는 False Positive Rate)를 활용한 축으로 그릴 수 있다.

따라서, ROC 곡선의 하단 면적(AUC)이 큰 모형일수록 출력변수의 클래스를 균형적으로 예측한다고 판단할 수 있다. 본 연구에서는 각 예측 알고리즘의 하이퍼 파라미터 탐색 시 AUC를 최대화하는 기준으로 선택하였다.

3.1.3 민감도와 특이도

일반적으로 분류 예측 모형의 결과는 혼돈행렬(Confusion Matrix)로 정리한다. 혼돈행렬은 TP(True Positive), FP(False Positive), FN(False Negative), TN(True Negative)으로 구성되어 있다. 본 연구의 데이터를 고려했을 때 TP는 'Yes'라고 예측했을 때 실제로 'Yes'인 경우이며, TN은 'No'라고 예측했을 때 실제로도 'No'인 경우이다. 반면 FP는 'Yes'라고 예측했지만 실제로는 'No'인 경우이며, FN은 'No'라고 예측했지만 실제로는 'Yes'인 경우이다[20].

민감도는 TP/(TP+FN)으로, 실제 'Yes'인 데이터의 개수 대비 'Yes'라고 예측한 결과가 정답인 개수의 비율이다. 따라서 민감도를 'Yes' 케이스의 예측 성능이라고 볼 수 있으며, 특이도는 민감도의 반대이므로 'No' 케이스의 예측 성능을 나타낸다고 할 수 있다. 추가로, 균형 정확도(Balanced Accuracy)는 민감도와 특이도의 평균값으로 모형의 균형적인 성능을 나타낼 수 있다.

본 연구의 목표는 치킨집에 대하여 영업의 지속과 폐업 여부를 판단하는 모형을 구축하는 것이다. 따라서, 알고리즘의 전체 예측 성능(Accuracy)과 더불어 민감도(Sensitivity)와 특이도(Specificity)를 함께 고려하여 예측 모형 평가

를 진행하였다.

3.2 최종 모형 선정

본 연구에서 수행한 2가지 Sub-sampling 방법 및 각 알고리즘에 따른 검증 데이터의 예측 성능을 비교하기 위해 <표 4>와 같이 결과를 정리하였다.

먼저 전체 예측 성능(Overall Accuracy)을 기준으로, 인공신경망(NN)이 67%로 가장 높은 성능을 보였다. 하지만 민감도가 다른 모형에 비해 14%로 매우 낮고, 상대적으로 특이도는 90%로 매우 높았다. 이는 인공신경망이 'Yes'인 경우는 잘 예측하지 못하고, 'No'인 경우만 잘 예측한다는 것을 의미한다. 즉, 인공신경망은 영업의 지속에 대한 균형적인 예측 성능을 보이지 못하기 때문에 본 연구의 목적과 부합하지 않는다고 판단하였다.

본 연구에서는 'Yes'와 'No' 케이스를 모두 균형적으로 잘 예측하는 유연판별분석(FDA with Up-sampling)을 최종 모형으로 선정하였다. FDA의 전체 예측 성능은 60.9%로, 가장 높은 인공신경망의 67%보다 낮지만, 그 차이는 제한적(-6.1%p)이다. 또한, 민감도와 특이도는 각각 60.2%와 62.4%로 전체 예측 성능과 유사하다. 따라서 클래스 종류와 무관하게 안정적인 예측 결과를 얻을 수 있을 것이다.

3.3. 변수 중요도 해석

모형의 변수 중요도(Variable Importance)란 사용된 모든 입력변수 중 예측에 가장 유의미한 변수를 100으로 두고, 나머지 변수들은 상대적 중요도로 해석하는 지표이다[24].

〈표 4〉 모델링 결과 비교 (Overall Accuracy 기준 내림차순)

Sub-sampling	Algorithm	Overall Accuracy	Sensitivity	Specificity	Balanced Accuracy
Down	**NN**	**67.0%**	**14.0%**	**90.0%**	**52.0%**
Up	NN	67.0%	14.0%	90.0%	52.0%
Up	RF	66.2%	88.5%	18.5%	53.5%
Up	AdaBoost	63.5%	73.9%	39.8%	56.8%
Down	AdaBoost	63.4%	73.8%	39.8%	56.8%
Down	XGBoost	62.8%	66.3%	54.8%	60.5%
Up	XGBoost	62.8%	66.8%	53.7%	60.2%
Down	FDA	61.5%	62.7%	58.9%	60.8%
Up	**FDA**	**60.9%**	**60.2%**	**62.4%**	**61.3%**
Down	RF	55.2%	52.7%	61.1%	56.9%
Up	Elastic Net	55.0%	51.6%	62.6%	57.1%
Down	Elastic Net	54.1%	48.3%	67.4%	57.8%
Up	PLSDA	52.4%	44.7%	70.0%	57.3%
Up	KNN	51.7%	47.1%	62.2%	54.7%
Down	SVM	51.5%	40.3%	77.2%	58.7%
Down	PLSDA	50.9%	42.3%	70.7%	56.5%
Down	KNN	49.3%	42.6%	64.6%	53.6%
Up	SVM	47.0%	32.2%	80.9%	56.5%

〈표 5〉 최종 모형(FDA)의 변수 중요도

순번	변수	중요도
1	유명 브랜드 여부	100
2	시설 총 규모	78
3	다중 이용 업소 여부	43
4	반경 1km 內 최근 5년간 개업 업체 수	25
5	반경 1km 內 경쟁 업체 수	21
…	(생략)	…

<표 5>와 같이 최종 선정된 FDA의 변수 중요도를 살펴보면 '유명 브랜드 여부'가 가장 예측에 유의미한 것으로 나타났다. 결과적으로 비비큐, 페리카나, 네네치킨 등 유명 브랜드 여부는 창업의 성패에 가장 큰 영향을 미친다고 해석할 수 있다.

두 번째로 중요도가 높은 변수는 '시설 총 규모'였으며, 다음으로는 '다중 이용 업소 여부'로 나타났다. 이는 상가의 면적에 따라 임대료나 집기 등 운영비용이 다르므로, 서울 시내 치킨집의 영업 지속에 영향을 미쳤다고 볼 수 있다.

마지막으로는 '반경 1km 內 최근 5년간 개업 업체 수'와 '반경 1km 內 경쟁 업체 수'로 산출됐다. 따라서 주변에 경쟁 치킨집이 얼마나 새로 개업하는지에 대한 추세와 이미 주변에 경쟁하고 있는 치킨집의 수 역시 해당 치킨집의 성패를 결정짓는 주요 요인이라고 볼 수 있다.

Ⅳ. 결론 및 의의

본 연구의 목적은 서울 치킨집의 폐업 예측 모형을 구축함으로써, 창업의 의사결정 과정에 정량적 결과를 활용할 수 있도록 하는 것이다. 이를 위해 가용한 공공 데이터를 최대한 수집하고, Feature Engineering을 통해 다양한 파생변수를 생성하는 등 예측에 유의미한 입력변수를 확보하고자 상당한 노력을 기했다. 실제 모델링 결과 프렌차이즈 여부나 경쟁업체 현황 등의 파생변수들이 예측에 유의미한 변수로 나타났다.

또한, kNN을 비롯하여 회귀, RF, 부스팅, 판별분석, SVM, 인공신경망 등 다양한 알고리즘을 시도함은 물론, 분류 비대칭 문제 완화 방법론을 접목함으로써 최적의 예측 모형을 찾고자 하였다. 결과적으로 치킨집 '폐업 가능성'이라는 자영업 창업에 직접적이고 정량적인 의사결정 지표를 도출했다는 점에서 본 연구의 의의가 있다.

본 연구는 다양한 확장이 가능하다. 치킨집의 영업 기간을 3년에 한정하지 않고 1년이나 5년 등 다양한 기간에 대한 폐업 여부를 예측하는 방향으로 연구를 확장할 수 있다. 나아가 연구의 분석 범위인 서울 시내의 치킨집을 전국의 다양한 업태로 확장하여 예측 모형을 구축할 수 있을 것이다.

참 고 문 헌

[1] 고은지, "소상공인 71%, 5년내 문 닫아…식당·여관은 1년내 절반 폐업", 2016.09.28., https://www.yna.co.kr/view/AKR20160927179000003

[2] 국가법령정보센터, *국토의 계획 및 이용에 관한 법률*, 2018.06.12.

[3] 국가법령정보센터, *다중이용업소의 안전관리에 관한 특별법*, 2017.12.26.

[4] 김현우, "'우후죽순' 치킨집, 전 세계 맥도날드 매장보다 많아", 2015.10.05., https://www.ytn.co.kr/_ln/0102_201510052201553904

[5] 박수호, & 서은내, "3040 vs 5060 세대별 창업 특징 살펴보니 | "인생 이모작…내 노후는 내가" 5060 반란", 2016.08.12., http://news.mk.co.kr/newsRead.php?no=575619&year=2016

[6] 배정원, "소비자기대지수", 2012.11.24., http://biz.chosun.com/site/data/html_dir/2012/

11/24/2012112400390.html

[7] 윤효원, "자영업자 700만명, 절반으로 줄여야", 2018.09.03., http://www.labortoday.co.kr/news/articleView.html?idxno=153665

[8] 이진욱, 유국현, 문병민, 배석주, "감성분석과 Word2vec을 이용한 비정형 품질 데이터 분석", 품질경영학회지, 제45권, 제1호, pp.117-127, 2017.

[9] 이현, "창업자 5명 중 2명은 치킨집 · 편의점…이미 포화 상태", 2016.07.11., http://news.jtbc.joins.com/article/article.aspx?news_id=NB11269730&pDate=20160711

[10] 최현준, "선행종합지수", 2012.03.04., http://www.hani.co.kr/arti/economy/economy_general/521840.html

[11] 통계청, *기업생멸행정통계*, 2016.

[12] 통계청, *자영업 현황분석*, 2016.

[13] "Orthogonal Partial Least Squares (OPLS) in R", 2013.07.28., https://www.r-bloggers.com/orthogonal-partial-least-squares-opls-in-r

[14] "A Quick Introduction to K-Nearest Neighbors Algorithm", 2017.04.11., https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7

[15] "SVM Separating Hyperplanes", 2012.11.26., https://en.wikipedia.org/wiki/Support_vector_machine#cite_note-CorinnaCortes-1/512px-Svm_separating_hyperplanes_(SVG).svg

[16] "What is an artificial neural network? Here's everything you need to know", 2018.09.13, https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/

[17] Chen, T., & Guestrin, C., "XGBoost: A Scalable Tree Boosting System", *International Conference on Knowledge Discovery and Data Mining*, pp.785-794, 2016.

[18] Fawcett, Tom, "An Introduction to ROC Analysis", *Pattern Recognition Letters*, Vol.27, No.8, pp.861–874, 2006.

[19] Friedman J, Hastie T, Tibshirani R., "Additive Logistic Regression: A Statistical View of Boosting", *Annals of Statistics*, Vol.28, No.2, pp.337–374, 2000.

[20] Han, J., & Kamber, M., *Data mining: Concepts and techniques (3rd ed.)*, Amsterdam: Elsevier, Morgan Kaufmann, 2011.

[21] Hastie, T., Tibshirani, R., & Buja, A., "Flexible Discriminant Analysis by Optimal Scoring", *J. of the American Statistical Association*, Vol.89, No.428, pp.1255-1270, 1994.

[22] Hoerl, A., & Kennard, R., "Ridge Regression: Biased Estimation for Nonorthogonal Problems", *Technometrics*, Vol. 42, No. 1, pp.80-86, 2000.

[23] Keller, J. M., Gray, M. R., & Givens, J. A., "A fuzzy K-nearest neighbor algorithm", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol.SMC-15, No.4, pp.580-585, 1985.

[24] Kuhn, M., & Johnson, K., *Applied predictive modeling (2nd ed.)*, New York: Springer., 2016.

[25] Leo Breiman, "Random Forests", 2001., https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

[26] Schalkoff, R.J, *Artificial neural networks*, McGraw-Hill, 1997.

[27] Sinnott, R.W, "Virtues of the Haversine", *Sky and Telescope,* Vol. 68, Issue 2, pp.158, 1984.

[28] Vapnik, V. N., *The nature of statistical learning theory*, New York: Springer, 2010.

[29] Wold, S., Sjöström, M., & Erikssonb, L., "PLS-regression: A basic tool of chemometrics", *Chemometrics and Intelligent Laboratory Systems*, Vol.58, No.2, pp.109-130, 2001.

[30] Zou, H., & Hastie, T., "Regularization and Variable Selection via the Elastic Net", *J. of the Royal*

Statistical Society. Series B (Statistical Methodology), Vol. 67, No.2, pp.301-320. 2005.

저 자 소 개

방 준 아(Junah Bang)
- 2018년 : 세종대학교 수학통계학부 (학사)
- 현재 : 성균관대학교 통계학과 (석사 과정)
- 관심분야 : Data Visualization, Data Analytics

손 광 민(Kwangmin Son)
- 2018년 : 한국외국어대학교 통계학과 (학사)
- 현재 : 성균관대학교 통계학과(석사 과정)
- 관심분야 : 지리 정보 시스템(GIS), Data Analytics, 알고리즘 개발

이 소 정(So Jung Ashley Lee)
- 2016년 : Emory University BS in Quantitative Science
- 현재 : CJ올리브네트웍스 DT융합연구소 AI개발
- 관심분야 : Deep Learning, Computer Vision, Vertical AI Development

이 현 근(Hyeongeun Lee)
- 2017년 : 성균관대학교 통계학과 (학사)
- 현재 : CJ올리브네트웍스 빅데이터센터 데이터분석컨설팅팀
- 관심분야 : Data Analytics, Data Science, Machine Learning, Deep Learning

조 수 빈(Subin Jo)
- 2018년 : 성균관대학교 통계학과(학사)
- 현재 : 성균관대학교 통계학과 (석사 과정)
- 관심분야 : R 패키지, Machine Learning, Deep Learning, Data Science