딥러닝 데이터 분석을 통한 최적의 상권 입지 추천 기술 개발

박형빈 [1], 김소희 [1], 남지수 [1], 조윤빈 [1], 전희국 [2], 임동혁 [1*]
[1] 광운대학교 정보융합학부
[2]㈜오픈업
idolphin.kr@gmail.com, sohee3055@naver.com, anm0307@naver.com, codo47@naver.com,
hgjun@openub.com, dhim@kw.ac.kr,

Commercial location recommend system using deep learning data analysis

Hyeong-Bin Park[1], So-Hee Kim[1], Ji-Su Nam[1], Yoon-Bin Cho[1]
Hee-Gook Jun[2], Dong-Hyuk Im[1*]
[1]School of Information Convergence, Kwangwoon University
[2]Corporation OpenUb

요 약

본 연구는 대량의 상권 데이터를 바탕으로 머신 러닝과 딥러닝 분석을 이용하여 최적의 상권 입지를 추천하는 시스템 개발을 목표로 한다. 자영업자들의 오프라인 창업에 있어 개개인의 매장 정보에 기반한 입지 조건 판단은 앞으로의 매출에 중요한 시작점이다. 따라서 상권 정보를 기반으로 미래 매출을 예측하여 최적의 상권 입지를 추천하는 기술이 필요하다. 이를 위해 기존에 선행된 다수의 회귀 기법과 더불어 강하게 편향된 데이터를 레이블링 하여 다중 분류 기법으로도 문제를 접근한다. 최종적으로 딥러닝 모델과 합성하여 더 높은 성능을 이끌어내고 이로부터 편향 데이터 처리 방법과 딥러닝 모델과의 앙상블 중요성에 대해 논의하고자 한다.

1. Introduction

자영업, 창업과 관련된 많은 선행 연구들에 따르면 오프라인 매장의 입지 선택은 매출에 가장 높은 영향을 주는 요인이라고[1] 말한다. COVID 19 이후 이런 경향성은 더욱 높아졌으며 이는 많은 소상공인과 자영업자들이 원하는 매출에 도달하기 위해서는 자신들의 매장에 맞는 적절한 상권 입지를 선택해야 할 중요성을 시사한다. 그러나 수많은 부동산, 입지와 관련된 정보들을 개인이 혼자서 판별하여 선택을 내리기에는 다소 어려움이 존재한다.

따라서 본 논문은 대량의 상권 데이터와 딥러닝 분석을 통하여 최적의 상권 입지를 추천하는 시스템을 개발한다. 이와 관련하여 이전에 선행된 주가, 부동산 등 가격 예측과 관련된 연구들은[2] 대부분 회귀분석이나 시계열 분석의 기법들을 통해 수행되었다. 그렇기에 개발 과정에서 일반적인 회귀 방법론과 더불어 분류의 방식으로도 문제를 해결해 보고 이에 대해 비교하여 최종적으로 구현된 머신 러닝모델을 딥러닝 모델과 조합해 예측 정확도 향상을 목표로 한다.

2. Problem Statement

2.1 Exploratory Data Analysis

본 연구에서는 매장의 속성 정보와 월평균 단가를 이용하여 월 매출을 예측하기 위해 <표 1>과 같은 변수들로 이루어진 데이터 셋을 이용하였다. 일반적인 정형 데이터 형태로 대략 300 백만 개에 달하며 데이터 분석 업체인 ㈜오픈업으로부터 제공받아 연구를 진행하였다.

<표 1> 매장 월 매출 예측을 위한 변수 속성

구분	변수
매장 속성 정보	매장 코드, 매출 날짜, 매장명, 업종 대분류, 업종 소분류, 위도, 경도
매장 매출 정보	월평균 단가, 월 매출

위 데이터 셋에서 주목할 만한 특징을 보인 것은 월평균 단가, 매출 변수이다. 각 변수의 분포 비대칭성을 확인하기 위해 왜도(Skewness)와 첨도(Kurtosis) 지표를 측정하였다. 예측할 y 변수인 월 매출은

Skewness 165.58, Kurtosis 37,638 이며, 월평균 단가는 Skewness 307.67, Kurtosis 238,001 로 왼쪽으로 쏠린 형태의 분포를 나타낸다. 이렇게 강하게 치우쳐진 데이터에 대해서는 정규성을 따르도록 전처리 하는 것이 중요하다. 특히 월 매출은 예측할 y 변수가 될 것이기 때문에 이를 고려한 전처리 과정과 문제 접근 방식이 시스템 설계의 핵심 주안점이 된다.

2.2 Data Preprocessing

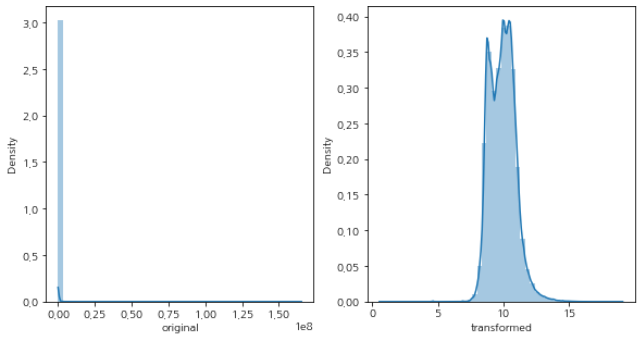
매출 코드는 각 행에 해당하는 매장을 구분 짓는 식별자 속성이며 매출 날짜는 시계열 분석 기법을 시도하지 않으므로 제거하였다. 위도와 경도 변수는 K-Means Clustering 기법을 활용하여 공간/위치 정보를 학습하기 위한 새로운 변수로 변환하였다. 업종 대분류와 소분류는 범주형 변수에 속하므로 Label Encoding 방식보다는 Binary Encoding 방식을 채택하였다. Binary Encoding 방식은 Label Encoding 방식보다 최종 입력 변수의 개수가 12 개가량 늘어나는 단점이 존재한다. 하지만 두 방식 모두 실험하였을 때 전반적으로 Binary Encoding 방식이 알고리즘 관계없이 정확도가 상승하는 것을 확인하여 최종 채택하게 되었다.

앞서 데이터 분석 과정에서 확인하였듯이 매출 정보 데이터들의 분포가 상당히 치우쳐져 있다. 이로부터 오는 영향을 줄이기 위해 두 가지 기법을 적용하였다. 첫 번째는 월평균 단가 변수에 Log Transformation 을 적용하였다. Log Transformation 은 극단적인 값이 몰려 있는 곳을 정규화하고 편차가 적은 곳을 퍼뜨려 최대한 정규화 된 분포를 따르도록 만든다. 두 번째는 각 업종 대분류에 해당하는 사분위수 값을 이용하여 이상치를 제거하였다. Q1, Q3 외의 값을 이상치로 보았을 때 약 2 만 개의 데이터가 탐지되었고 이를 모두 제거함으로써 총 270 만 개의 데이터가 정제되었다.

전처리 된 결과를 토대로 모든 알고리즘에 기본적으로 사용되는 변수는 6 개이며 각 접근 방식에 따라 세부적으로 사용되는 변수의 개수는 달라진다. 추가적으로 매출 정보 데이터에 대한 변환은 <표 2> 와 (그림 1) 에서처럼 정리할 수 있다.

<표 2> 매출 정보 데이터 전처리 결과

구분	월 평균 단가		월 매출	
	변환 전	변환 후	변환 전	변환 후
Skewness	165.58	1.68	307.67	0.44
Kurtosis	37,638.37	3.27	238,001.91	0.77



(그림 1) 월 평균 단가 변수에 대한 Log Transformation 결과

3. Our Method

3.1 Regression

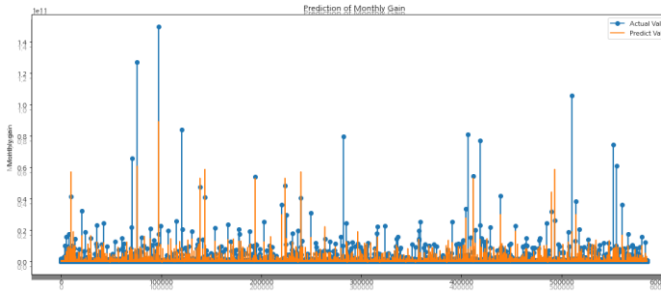
서론에서도 언급하였듯이 전통적으로 가격이라는 연속적인 값을 예측하는 문제들은 회귀, 시계열 분석 기법을 사용하였다. 본 논문에서도 4 가지의 대표적인 회귀형 모델을 사용하였고 이를 비교하였다. Gradient Boosting 종류의 XGB, LGBM 과 Linear Model 종류인 Ridge, Lasso Regression 을 사용하였다. 입력으로 사용된 변수는 '업종 명'을 제외한 총 5 개를 사용하였다. 모델의 평가 척도로는 MSE, RMSE, R2 를 사용하였고 각각의 결과는 아래의 <표 3>에서 확인할 수 있다.

<표 3> 회귀 모델 별 월 매출 가격 예측 정확도

구분	MSE	RMSE	R2
XGB	167,212,972,899,818	12,931,085	0.3
LGBM	170,001,344,508,730	13,038,456	0.2
Ridge	218,298,021,558,504	14,774,911	0.14
Lasso	218,298,022,158,726	14,774,911	0.14

RMSE 는 실제 가격과 예측 가격 사이의 차이를 제곱근으로 표현한 것으로 값이 클수록 오차가 크다. <표 3>의 값들은 이상치를 제거하였을 때 값으로 제거하기 전보다 모든 모델들의 오차가 60%가량 감소한 값이다. 후에 서술되는 방법론들과는 달리 회귀 방법론은 이상치 제거 효과를 가장 많이 보았다. 그럼에도 불구하고 오차의 크기가 상당히 커 보이는데 이와 대조적으로 가격 예측의 흐름을 보여주는 (그림 2)를 보면 전반적으로 가격의 추이를 비슷하게 예측하는 것을 알 수 있다. 이는 이상치를 제거했음에도 불구하고 여전히 월 매출 가격대가 업종, 매장에 따라 큰 편차를 가지고 있기 때문이다. 실제로 가격대의 분포를 확인하였을 때 적게는 1 원 단위부터 많게는 100 억 단위까지 나오는 것을 확인할 수 있었다. 이로부터 모델이 1 원 단위까지 예측하는 것에는 분명한 한계가 있고, 사용자들에게 서비스로 제공할 때에도 예상 매출을 자세하게 예측하여 제공할 필요가 없음을

인식하여 해당 문제를 회귀 방법론이 아닌 분류 방법론으로 전환하여 풀어보고자 하였다.



(그림 2) XGB Regressor 결과 그래프

3.2 Multi-Class Classification

분류 방법론으로 접근하고자 하기 때문에 최대한 많은 가격대의 레이블(Label)을 만들어 예측하는 다중 분류(Multi-Class Classification) 문제로 보아야 한다. 회귀 방법론으로 접근한 문제를 분류의 문제로 전환하기 위해서는 예측 변수 y 를 어떻게 레이블링(Labeling) 할지에 달려있다. 다시 말하자면 다중 분류 문제의 목표는 '월 매출' 변수를 어떤 방식으로 레이블링 하였을 때 최대한 많은 레이블로 나누면서도 높은 정확도를 보여주는 것이라고 할 수 있다. 그리하여 본 논문에서는 통계적 기법을 기반해 총 4 가지의 레이블링 방법을 시도하였다. 각 레이블링 방식은 순서대로 숫자를 붙여 지칭하겠다.

Labeling 1 방식은 '월 매출' 데이터에 대해서 순서대로 나열한 후 차례대로 10%, 20%, 25%, 33%씩 잘라서 레이블을 나누는 방식이다. 만약 10%씩 분할한다면 최종적으로 10 개의 레이블이 나오는 것이며 33%씩 분할할 경우 3 개의 레이블이 나온다. 최종 분할된 레이블의 개수가 적을수록 예측 정확도가 높아진다. 대표적으로 10%씩 분할하여 10 개로 나누었을 때 결과값을 <표 4>에서 확인할 수 있다.

Labeling 2 방식은 사분위수를 이용하여 분할한다. '월 매출' 데이터를 업종 대분류 변수에 따라 Q1 이하의 값은 lower fence 값으로, Q3 이상의 값은 upper fence 값으로, Q1 과 Q3 사이의 값은 중앙값인 Q2 값으로 레이블링 한다. 해당 방법의 한계점은 모든 업종 대분류에 대한 '월 매출' 데이터의 lower fence 값이 0 원이었기에 최종적으로 15 만 개의 데이터가 0 값으로 레이블링 된다. 이러한 이유로 총 업종 대분류 15 개 당 2 개의 레이블(upper fence, Q2 값) 과 lower fence(0 원) 1 개의 레이블로 최종 31 개의 레이블이 생성된다.

Labeling 3 방식 또한 사분위수를 이용하여 분할하지만 레이블링 되는 값을 달리하였다. '월 매출' 데이터를 업종 대분류 변수에 따라 Q1 이하의 값은 Q1 값으로, Q3 이상의 값은 Q3 값으로, Q1 과 Q3 사이의 값은 중앙값인 Q2 값으로 레이블링 한다. 이렇게 할 경우 Labeling 2 와는 달리 최종 레이블의 개수가 업종 대분류 15 개당 3 개의 레이블로 분류되어 총 45 개가 된다.

Labeling 4 방식은 '월 매출' 데이터의 최소값과 최대값 사이를 임의적으로 분류하여 레이블링 한다. 앞선 방법들을 통해 '월 매출' 데이터의 분포가 특히나 1 천~3 천만 원 사이에 집중되어 있음을 확인하였다. 이를 토대로 100,000 원 이상부터 150 억 사이로 레이블링 하되 천만 원 구간을 더 촘촘하게 분류하였다. 최종 레이블의 개수는 12 개가 된다.

총 4 가지의 방식에 대해서 간단한 설명과 최종 레이블 개수를 아래의 <표 4>와 같이 정리하였다.

<표 4> 레이블링 방식 요약 표

구분	Description	Label
Labeling 1	작은 값부터 큰 값까지 순서대로 특정 범위만큼(10%) 분류	10
Labeling 2	사분위수를 이용하여 lower fence, Q2, upper fence 값으로 분류	31
Labeling 3	사분위수를 이용하여 Q1, Q2, Q3 값으로 분류	45
Labeling 4	데이터의 최소값~최대값 사이를 임의적으로 분류	12

기본적으로 모델은 회귀에서도 쓰였던 2 가지 모델 XGB, LGBM 의 Classifier 를 사용하였다. 입력으로 사용된 변수는 '업종 명'을 제외한 총 5 개를 사용하였다. 모델의 평가 척도로는 Accuracy 를 사용하였고 모델과 레이블링 방법에 따른 각각의 정확도 결과는 아래의 <표 5>에서 확인할 수 있다.

<표 5> 분류 모델, 레이블링 방식 별 예측 정확도

구분	Label	XGB	LGBM
Labeling 1	10	0.22	0.21
Labeling 2	31	0.57	**0.30**
Labeling 3	**45**	**0.58**	0.11
Labeling 4	12	0.35	0.33

<표 5>에서 볼 수 있듯이 최종 레이블 수와 예측 정확도를 고려하였을 때 Labeling 3 방식의 실험 결과가 가장 좋은 것으로 알 수 있다. 매출 데이터를 최저 금액부터 최고 금액으로 줄 세웠을 때 평균값으로 데이터를 분류할 경우 높은 금액대의 가격들에 의해서 쉽게 영향을 받는다. 그러나 사분위수와 같은 중앙값을 이용하여 분류할 경우 이러한 영향을 덜 받을 수 있다. 특히 Labeling 3 의 방식은 이와 동시에 업종

대분류 변수를 이용하여 각각의 업종별로 가지는 편차에 맞게 적용하였기에 가장 좋은 결과를 얻을 수 있었다고 여겨진다.

3.3 Natural Language Processing

머신 러닝 모델을 기반으로 데이터를 분석할 수 있지만 딥러닝 모델을 기반하여 다른 시각으로도 문제를 분석할 수 있다. 이에 본 논문에서는 딥러닝의 한 분야인 자연어 처리 방법 NLP(Natural Language Processing)을 활용하고자 한다.

NLP 에서 학습하는 변수는 매장명, 업종 대분류, 업종 소분류 총 3 가지로 앞선 방법론들과는 다른 전처리 과정을 거쳤다. 세 가지 변수에 대해서 하나의 자연어 변수를 새로 생성하였다. 해당 변수를 토대로 약 12 만 개의 단어 데이터 셋을 얻었고 이로부터 자연어 변수를 임베딩 벡터로 변환하여 딥러닝 레이어의 입력으로 넣어주었다.

모델은 Keras Library 의 Sequential Model 을 이용하여 구축하였다. 최초의 입력 레이어에서 워드 임베딩을 학습한다. 그다음 Flatten 레이어부터 분류를 학습하게 된다. 이후 각각의 Dense 레이어의 Activation Function 은 Relu 와 Softmax 를 사용하였다. 모델의 학습 방식에 대한 손실 함수(loss function)는 Sparse Categorical Crossentropy 를, 정규화기(optimizer)는 Adam, 평가 척도는 분류 문제이기에 동일하게 Accuracy 로 설정하였다. 학습은 실험을 통해 최적의 Epoch 수인 10 으로 선정하였고 한 번의 학습에 대한 Mini-Batch 크기는 64 로 지정하였다. 최종 구축된 모델에 대한 요약은 (그림 3)과 같다.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 13, 10)            1169430
_________________________________________________________________
flatten (Flatten)            (None, 130)               0
_________________________________________________________________
dense (Dense)                (None, 64)                8384
_________________________________________________________________
dense_1 (Dense)              (None, 36)                2340
=================================================================
Total params: 1,180,154
Trainable params: 1,180,154
Non-trainable params: 0
_________________________________________________________________
```

(그림 3) Sequential Model 요약

2.4 장의 실험을 토대로 가장 결과가 좋았던 Labeling 3 방식으로 분류된 월 매출 데이터에 대해서 NLP 를 학습한 결과 예측 정확도는 0.80 으로 측정되었다.

3.4 Ensemble

머신 러닝 모델은 공간, 금액, 업종 등 다양한 변수를 학습하였고 딥러닝 모델은 자연어적 의미를 학습하였다. 서로 다른 시각에서 학습된 두 모델을 적절히 앙상블 하였을 때 예측 정확도를 확인해 보았으며, 머신 러닝 모델과 딥러닝 모델 그리고 두 모델의 앙상블에 대한 각각의 결과를 <표 6> 과 같이 정리하였다. 앙상블 모델은 두 모델 조합의 비율이 0.5 씩일 때 가장 최적의 성능을 보였으며, 따라서 XGB Classifier (머신러닝 모델)와 Sequential Model (딥러닝 모델)을 일대일 비율로 앙상블한 방법을 최종 모델로 구축하였다.

<표 6> 앙상블 결과

Model	Accuracy
XGB Classifier (ML)	0.58
Sequential Model (DL)	0.80
0.5 ML + 0.5 DL	**0.81**

4. Conclusion

본 연구의 목적은 대량의 상권 데이터를 기반으로 딥러닝, 머신 러닝 분석을 통해 최적의 상권 입지 추천 시스템을 개발하여 자영업자, 소상공인의 창업과 매출 증진에 도움을 주는 것이다.

시스템 개발을 통해 각 상권 데이터의 속성, 매출 변수를 활용하여 미래의 매출을 예측하였다. 이를 위해 머신 러닝의 대표적인 두 가지 방법론을 적용하였는데, 일반적으로 가격과 같은 실수형 변수를 예측하기 위해 회귀 방법론이 많이 쓰였으나 본 연구에서는 분류 방법론 또한 제시하였다. 이는 강하게 편향된 데이터에 대해 통계적 분석을 기반하여 예측 변수를 적절히 분류하는 것은 가격 예측 문제의 또 다른 대책이 될 수 있음을 보였다. 또한 이에 그치지 않고 매장명과 같은 자연어 데이터를 딥러닝 기법으로 학습하여 예측 정확도를 0.81 까지 향상시켰다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학 지원사업의 연구결과로 수행되었음(2017-0-00096).

참고문헌

[1] M Spann, Molitor, S Daurer, Using Location Data to Improve Marketing Decisions', NIM Marketing Intelligence Review, 2016, 32

[2] Martin J. Bailey, Richard F. Muth &Hugh O. Nourse, A Regression Method for Real Estate Price Index Construction, Journal of the American Statistical Association, 2012, 933