# 검색어 트렌드와 뉴스기사 수를 이용한 주간 경제심리보조지수 개발

Team: 박채린, 이혜승, {pcr0827, 2hyes}@sm.ac.kr

**Sookmyung Women's University**

## 1. Introduction

'침체', '금융위기', '불황', '폭락', '외환위기'의 키워드를 포함하는 뉴스 기사 수와 포털에서의 '경제' 키워드 검색량을 활용하여 주간 경제보조심리지수를 개발한다.

매 주 얻을 수 있는 주간 지수를 활용하여, 부정적인 경제 상황에서 빠른 피드백을 받아 탄력적인 대응을 할 수 있다. 또한, 주간 경제보조심리지수는 개발한 모형에 수집한 자료를 인입하여 지수를 바로 얻기 때문에, 속보성을 지닌다. 더 나아가 질문지법으로 자료 수집을 하는 소비자심리지수에 비하여, 표본 설계의 비용과 오류에서 비교적 자유로우며, 개인들의 현 경제상황에 대한 내면의 심리를 잘 파악할 수 있다.

## 2. Background

### ● 소비자심리지수

한국은행이 매월 전국 2,200가구를 대상으로 실시하는 소비자동향조사에서 산출하는 소비자동향지수 중 현재생활형편, 생활형편전망, 가계수입전망, 소비지출전망, 현재경기판단, 향후경기전망에 대한 6개의 개별지수를 표준화하여 합성한 지수로, 소비자심리를 종합적으로 판단하는 데 이용된다.

### ● 동행지수 순환변동치

경기동행지수에서 추세변동을 제거한 순환요인으로, 현재의 경기상황을 파악하는 보조 지표이다. 동행지수 순환변동치가 100보다 높을 경우 경기가 호황이고, 100미만이면 경기가 불황이라고 판단된다.

## 3. Data

### ● 데이터 소개

활용하는 자료는 소비자심리지수와 경기동행지수 순환변동치, 특정 키워드를 포함하는 뉴스 기사 수, 포털에서의 '경제' 키워드 검색 비율 정보이다.

첫째, 경제 지표인 소비자 심리지수과 경기동행지수는 한국은행에서 제공하는 경제통계 open API 서비스를 활용하여 수집한다. 소비자 심리지수는 주간 경제심리보조지수를 생성을 위한 반응 변수로 활용한다.

둘째, 포털에의 '경제' 키워드 검색 비율 정보는 구글 트렌드, 네이버 데이터랩, 카카오 트렌드에서 받아온다.

셋째, 특정 키워드를 포함하는 뉴스 기사 수는 한국언론진흥재단에서 운영하는 뉴스분석 서비스인 빅카인즈에서 json 파일로 데이터를 수집한다. 뉴스 키워드는 침체, 금융위기, 불황, 외환위기, 폭락, 금융위기를 고려한다.

### ● 데이터 전처리

반응변수인 CCSI는 소비자 심리지수는 한달 단위이고, 수집한 예측변수는 주간 단위이므로 지표 단위를 맞추어 준다. 주간 데이터들을 월간 데이터로 변환하는 방법, 월간 데이터들을  주간 데이터로 변환하는 방법을 고려한다.

실제 모델에 변수들을 적합할 때에는 StandardScaler을 통해 평균을 제거하고 데이터를 단위 분산으로 조정하여 모든 변수들이 같은 스케일을 갖게 한다.

### ● 평균 전처리를 적용한 데이터 셋

| id | year | month | keyword1 | keyword2 | keyword3 | keyword4 | keyword5 | google | naver | ccsi |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2016 | 1 | 895.00 | 403.50 | 449.00 | 610.00 | 161.25 | 61.00 | 25.83 | 99.5 |
| 2 | 2016 | 2 | 811.80 | 396.60 | 359.20 | 341.60 | 124.40 | 60.00 | 24.91 | 101.1 |
| 3 | 2016 | 3 | 667.25 | 267.00 | 293.00 | 98.50 | 107.50 | 67.75 | 28.30 | 98.6 |
| 4 | 2016 | 4 | 718.25 | 258.25 | 367.00 | 142.25 | 116.75 | 69.00 | 25.84 | 98.2 |

## 4. Methods and hypotheses

### ● 1 변수 상관관계 탐색

| id | coef | std err | t | P > \|t\| |
|---|---|---|---|---|
| intercept | 0.7699 | 0.076 | 10.161 | 0.000 |
| kakao | -0.3473 | 0.258 | -1.344 | 0.189 |

예측 *변수선택*을 위해, 포털 트렌드 변수와 경기동행지수 순환변동치와의 단순 회귀분석 계수 확인 및 상관 계수를 확인한다.

### ● 2 전처리 및 데이터 분할

- 데이터 전처리 방법으로는 크게 두가지를 고려한다.

  1. 주간 데이터를 월간 데이터로 변환

  수집한 데이터셋은 주간 데이터이고 예측 모델에서 반응 변수로 사용할 소비자심리지수는 월간 데이터이므로, *주간 변수 값들의 평균 혹은 중앙값*들을 해당 달의 데이터로 활용한다.

  2. 월간 데이터를 주간 데이터로 변환

  월간 데이터인 소비자심리지수에 *선형 보간법*을 적용하여, 주간 데이터에 y 참 값을 생성하여 주간 데이터셋으로 활용한다.

- 데이터 분할 방법으로는 임의로 섞어서 training set과 test set을 분할하는 방법과 시간의 흐름대로 2016~2019년까지는 training set, 2020년을 test set으로 활용하는 두 방법을 고려한다.

### ● 3 예측 모형

다중선형회귀 모형, 라쏘선형회귀 모형, 랜덤 포레스트 모형, 일반화 가법 모형을 각 전처리와 분할 방법으로 생성한 가공 데이터셋으로 적합한다.

### ● 4 T 검정을 통한 예측 성능 비교

$$\boldsymbol{H0}: RMSE\ of\ model\ 1 = RMSE\ of\ model\ 2$$
$$hence, there\ is\ no\ difference\ of\ performace\ between\ two\ models.$$
$$\boldsymbol{H1}: there\ is\ difference\ of\ performance.$$

10-fold Cross Validation을 통해 계산된 10개의 RMSE값으로 t검정을 진행한다. 0.05의 유의 수준에서 기각역을 설정하여, $|\boldsymbol{t-value}| > \boldsymbol{t_{0.025,9}} = \boldsymbol{2.262}$이면, $\boldsymbol{H0}$을 기각한다.

## 5. Experimental results

### ● 소비자 심리지수 예측 - T 검정

모든 전처리 및 분할 방법을 적용한 데이터셋에서 4가지의 모형군(다중회귀, 라쏘회귀, 랜덤포레스트, 일반화가법 모형)에 대해 t 검정을 하여, 최적의 모형을 선택한다.

우측의 표는 평균 전처리, 시간순서로 분할한 데이터셋의 일반화가법 모형과 라쏘회귀 모형의 RMSE 차이이다. $|T^*| = \frac{4.3380}{3.0806/\sqrt{10}} = 4.453 > t_{0.025,9}$ 이므로, 유의수준 0.05에서 기각된다. 따라서 일반화가법모형과 라쏘회귀 모형의 성능에는 차이가 있다고 할 수 있다. 또한 CV RMSE의 평균값이 양수이므로, 라쏘회귀모형의 성능이 더 좋다고 보여진다.

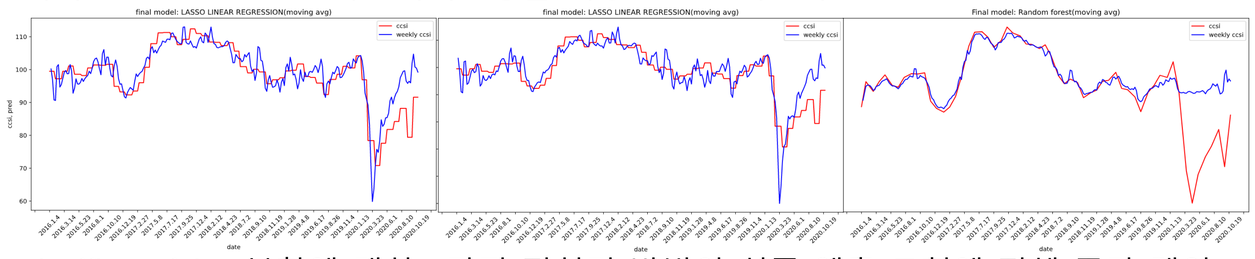
| | GAM − LASSO |
|---|---|
| 1-fold | 5.5585 |
| 2-fold | 3.9693 |
| ⋮ | ⋮ |
| 8-fold | 10.0049 |
| 9-fold | 3.3423 |
| 10-fold | -0.0671 |
| RMSE mean | 4.3380 |
| RMSE std | 3.0806 |

### ● 소비자 심리지수 예측 - 모형 비교 및 선택

| | Shuffle = True | | Shuffle = False | |
|---|---|---|---|---|
| 평균 | 다중선형회귀 | | 라쏘 선형회귀(alpha = 0.1) | |
| | Test RMSE: 4.006 | Test MAE: 3.122 | Test RMSE: 8.438 | Test MAE: 2.568 |
| 중앙값 | 다중선형회귀 | | 라쏘 선형회귀(alpha = 0.1) | |
| | Test RMSE: 6.203 | Test MAE: 3.664 | Test RMSE: 9.336 | Test MAE: 2.605 |
| 내삽 | 랜덤 포레스트(트리 수=32,최대 변수 수=4) | | 랜덤 포레스트(트리 수=64,최대 변수 수=3) | |
| | Test RMSE: 3.490 | Test MAE: 2.483 | Test RMSE: 13.889 | Test MAE: 3.381 |

전처리 및 분할 방법마다의 최적 모형들이다. RMSE와 MAE 오류 측도만으로 비교하면, 내삽을 적용한 후 임의로 섞은 데이터 셋으로 훈련시킨 랜덤 포레스트 모형이 가장 좋은 예측력을 보여준다.

그러나 2020년과 같이 예상하지 못한 경제 상황을 예측하고자 하는 목적을 고려하여 2016년부터 2019년의 데이터로  모델을 훈련시키고, 2020년 데이터로 테스트를 진행하는 방식이 적절하다. 따라서 데이터를 임의로 섞지 않고 평균 전처리 방법의 최적 모델인 α=0.1인 라쏘회귀모형을 최종으로 선택한다.



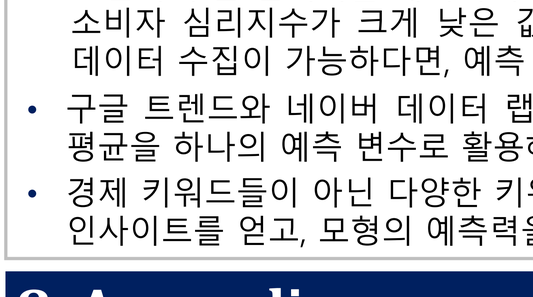

Shuffle = false 분할에 대한 3가지 전처리 방법의 최종 예측 모형에 전체 주간 데이터셋을 넣은 결과를 시각화한 것이다. 차례로 평균, 중앙값, 내삽을 전처리를 적용한 결과이며, 내삽 방법은 2020년에 대해 주간 경제심리보조지수가 소비자심리지수를 제대로 예측하지 못한다.

### ● 소비자 심리지수 예측 - 최종 예측 모형 해석

$$Y = 101.48 - 2.25\ X_1 - 0.316X_2 - 2.568X_3 + 0.907\ X_4 - 0.395\ X_5 - 1.954X_7$$

최종 예측 모형인 라쏘 회귀 모형은 위의 식과 같다. '침체', '금융위기', '불황', '외환위기' 키워드의 뉴스 기사 수와 네이버 검색량이 많은 경우 소비자심리지수가 낮았다. 즉, 부정적인 경제 상황을 다루는 뉴스 기사 수가 많을 수록, 사람들이 경제에 관심을 갖고 포털에 많이 검색해볼 수록, 소비 심리가 위축된다 해석해볼 수 있다.

### ● 2020년에 대한 과대 추정의 원인




예측 모형이 2020년에 대해서 과대 추정을 한다. y값이 높게 추정되는 2020년에 대해서 사후 분석을 한 결과, *2016년에 비해 상대적으로 불황 키워드(keyword3)에 대한 뉴스 기사 수가 적다.*

또한, '불황' 키워드의 계수 절댓값이 가장 크기 때문에, y 추정에 영향을 가장 크게 미친다. 따라서 2020년에 상대적으로 적은 '불황' 키워드 뉴스 기사수의 영향으로 y가 크게 추정된다.

## 6. Demonstrate Validity

| | date | weeklyCLI |
|---|---|---|
| 1 | 2016.1.4 | 100.11 |
| 2 | 2016.1.11 | 100.22 |
| 3 | 2016.1.18 | 96.92 |
| | ⋮ | |
| 250 | 2020.10.12 | 100.78 |
| 251 | 2020.10.19 | 100.45 |
| 252 | 2020.10.26 | 99.24 |

- CCSI의 추세를 잘 따르는 것을 보임으로써 개발한 주간 경제보조심리지수가 경기에 선행되는 지수의 역할을 하고 있음을 확인할 수 있다.
- 테스트 셋의 소비자심리지수 추정값과 소비자심리지수의 상관계수는 0.776 으로 강한 양의 상관관계이다. 즉, 2020년과 같이 경제 상황이 급격히 안 좋아진 시기에도, 소비자의 심리를 나타내는 지표로서 활용이 가능하다.

## 7. Future Work

- 포털의 주간 검색비율 데이터 수집이 2016년 1월 1일부터 가능하여, 2020년 이전에 소비자 심리지수가 크게 낮은 값을 갖는 경우가 없다. 2008년의 불황시기에 대한 데이터 수집이 가능하다면, 예측 성능이 더 좋아질 것으로 보인다.
- 구글 트렌드와 네이버 데이터 랩의 자료 간의 다중 공선성 제거를 위해, 두 자료의 평균을 하나의 예측 변수로 활용해보고자 한다.
- 경제 키워드들이 아닌 다양한 키워드를 포함하는 뉴스기사 수를 카운팅하여, 새로운 인사이트를 얻고, 모형의 예측력을 더욱 높이고자 한다.

## 8. Appendix

[1] SNS 데이터를 활용한 소비자성향 분석, 황영자, 2016, 통계개발원 연구보고서
[2] 빅데이터를 이용한 경기판단지표 개발: 네이버 검색 경기지수 작성과 유용성 검토, 이긍희, 황상필, 2014, 경제분석, 한국은행 경제연구원 제 20권 제 4호