# 목 차

## 서론

본 분석은 두산베어스의 관중 수 데이터를 이용하여, 두산베어스가 진행한 각종 프로모션을 효과를 분석하는데 목적이 있다. 파이썬 머신러닝을 활용하여 다중회귀분석을 실시해 관중 수와 프로모션의 상관관계를 파악한다. 분석대상 프로모션(변수)은 cap, shirts, fireworks이다. 이때 통제변수는 month, day_of_week, park이다.

## 본론

1. 활용 데이터 소개

81x12의 csv 데이터로 다음과 같이 이루어져 있다.

| | month | day | attend | day_of_week | opponent | temp | skies | day_night | cap | shirt | fireworks | park |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | APR | 10 | 56000 | Tuesday | Lotte | 67 | Clear | Day | NO | NO | NO | NO |
| **1** | APR | 11 | 29729 | Wednesday | Lotte | 58 | Cloudy | Night | NO | NO | NO | NO |
| **2** | APR | 12 | 28328 | Thursday | Lotte | 57 | Cloudy | Night | NO | NO | NO | NO |
| **3** | APR | 13 | 31601 | Friday | LG | 54 | Cloudy | Night | NO | NO | YES | NO |
| **4** | APR | 14 | 46549 | Saturday | LG | 57 | Cloudy | Night | NO | NO | NO | NO |

| | |
|---|---|
| month | 월 - 통제변수 |
| day | 일 - 고려대상에서 제외 |
| attend | 관중 수 - 구하고자 하는 종속변수 |
| day_of_week | 요일 - 통제변수 |
| opponent | 상대 팀 - 고려대상에서 제외 |
| temp | 온도 |
| skies | 하늘을 맑음(Clear), 흐림(Cloudy) 여부 |
| day_night | 낮경기(Day), 밤경기(Night) 여부 |
| cap | 모자를 나눠주는 프로모션 - 분석대상 |
| shirt | 셔츠를 나눠주는 프로모션 - 분석대상 |
| fireworks | 불꽃놀이 프로모션 - 분석대상 |
| park | 박철순 선수 인형을 나눠주는 프로모션 - 통제변수 |

2. 1차 회귀분석

원본 데이터에서 'attend'를 종속변수로 설정하고 고려대상이 아닌 'day' 와 'opponent' 변수를 제외하고 나머지 변수들을 독립변수로 설정하고 1차 회귀분석을 진행한다.

OLS Regression Results

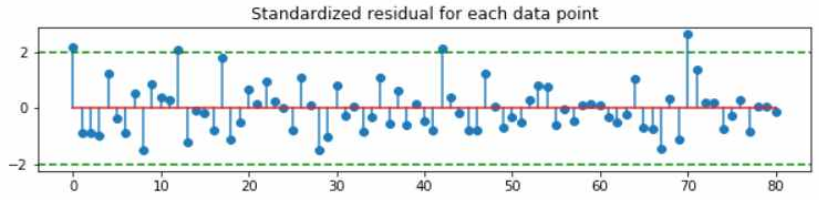
| | | | |
|---|---|---|---|
| **Dep. Variable:** | attend | **R-squared:** | 0.583 |
| **Model:** | OLS | **Adj. R-squared:** | 0.462 |
| **Method:** | Least Squares | **F-statistic:** | 4.824 |
| **Date:** | Sun, 17 Nov 2019 | **Prob (F-statistic):** | 1.62e-06 |
| **Time:** | 16:18:42 | **Log-Likelihood:** | -809.89 |
| **No. Observations:** | 81 | **AIC:** | 1658. |
| **Df Residuals:** | 62 | **BIC:** | 1703. |
| **Df Model:** | 18 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 3.128e+04 | 1.14e+04 | 2.734 | 0.008 | 8413.461 | 5.42e+04 |
| **month[T.AUG]** | 1732.4271 | 3505.718 | 0.494 | 0.623 | -5275.403 | 8740.257 |
| **month[T.JUL]** | 3158.0128 | 3136.903 | 1.007 | 0.318 | -3112.567 | 9428.593 |
| **month[T.JUN]** | 5844.5400 | 3058.636 | 1.911 | 0.061 | -269.587 | 1.2e+04 |
| **month[T.MAY]** | -2873.9771 | 2653.213 | -1.083 | 0.283 | -8177.674 | 2429.720 |
| **month[T.OCT]** | -1466.8018 | 5683.502 | -0.258 | 0.797 | -1.28e+04 | 9894.357 |
| **month[T.SEP]** | -1540.6983 | 4603.916 | -0.335 | 0.739 | -1.07e+04 | 7662.398 |
| **day_of_week[T.Monday]** | -396.8096 | 2615.282 | -0.152 | 0.880 | -5624.684 | 4831.064 |
| **day_of_week[T.Saturday]** | 7062.9253 | 2363.530 | 2.988 | 0.004 | 2338.296 | 1.18e+04 |
| **day_of_week[T.Sunday]** | 6310.0451 | 2923.161 | 2.159 | 0.035 | 466.730 | 1.22e+04 |
| **day_of_week[T.Thursday]** | 1472.6837 | 3013.590 | 0.489 | 0.627 | -4551.396 | 7496.764 |
| **day_of_week[T.Tuesday]** | 8399.6912 | 2418.101 | 3.474 | 0.001 | 3565.976 | 1.32e+04 |
| **day_of_week[T.Wednesday]** | 2984.2084 | 2550.325 | 1.170 | 0.246 | -2113.819 | 8082.235 |
| **skies[T.Cloudy]** | -1746.0631 | 2011.937 | -0.868 | 0.389 | -5767.868 | 2275.741 |
| **day_night[T.Night]** | -169.4715 | 3003.256 | -0.056 | 0.955 | -6172.895 | 5833.952 |
| **cap[T.YES]** | -5571.9439 | 4810.931 | -1.158 | 0.251 | -1.52e+04 | 4044.969 |
| **shirt[T.YES]** | 5782.3774 | 3893.811 | 1.485 | 0.143 | -2001.239 | 1.36e+04 |
| **fireworks[T.YES]** | 5452.0106 | 2303.971 | 2.366 | 0.021 | 846.438 | 1.01e+04 |
| **park[T.YES]** | 1.024e+04 | 2700.473 | 3.792 | 0.000 | 4843.337 | 1.56e+04 |
| **temp** | 47.5081 | 186.494 | 0.255 | 0.800 | -325.289 | 420.305 |

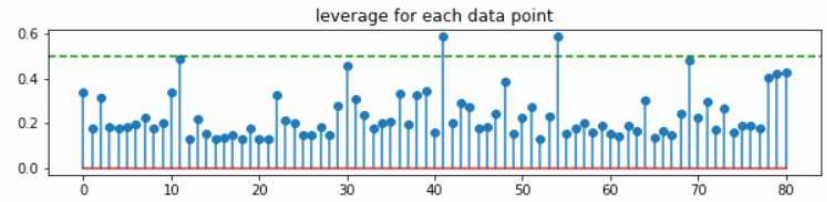
| | | | |
|---|---|---|---|
| **Omnibus:** | 9.424 | **Durbin-Watson:** | 2.233 |
| **Prob(Omnibus):** | 0.009 | **Jarque-Bera (JB):** | 9.256 |
| **Skew:** | 0.796 | **Prob(JB):** | 0.00977 |

R값과 Adj.R 값이 저조한 것을 확인할 수 있다. 이때 아웃라이어의 존재 때문일 것을 예상하고 표준화잔차와 레버리지를 확인한다.

3. 표준화잔차와 레버리지 확인 및 아웃라이어 제거




(표준화 잔차)




(레버리지)

표준화 잔차 : 0, 13, 42, 70번째 데이터가 아웃라이어일 것으로 예상
레버리지 : 11, 42, 54, 70번째 데이터가 아웃라이어일 것으로 예상

```
#cooks distance 및 fox 추천 판단
cooks_d2, pvals = influence.cooks_distance
K=influence.k_vars
fox_cr = 4/(len(doo['attend'])-K-1)
idx = np.where(cooks_d2 > fox_cr)[0]
idx
```

```
array([ 0, 41, 42, 54, 69, 70], dtype=int64)
```

그림과 같이 Cooks Distance 및 Fox 추천을 이용하여 아웃라이어를 판단한 결과
0, 41, 42, 54, 69, 70번째 데이터가 아웃라이어로 판단되었다. 이는 앞서 예상한 결과와 비슷하다. 이 6개의 데이터들을 제외하고 남은 75개의 데이터들을 이용하여 2차 회귀분석을 진행한다.

## 4. 2차 회귀분석

아웃라이어로 판단된 6개의 데이터를 제거하고 75개의 데이터를 이용하여 2차 회귀분석을 진행한다.

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | attend | **R-squared:** | 0.713 |
| **Model:** | OLS | **Adj. R-squared:** | 0.628 |
| **Method:** | Least Squares | **F-statistic:** | 8.335 |
| **Date:** | Sun, 17 Nov 2019 | **Prob (F-statistic):** | 5.00e-10 |
| **Time:** | 16:21:59 | **Log-Likelihood:** | -734.94 |
| **No. Observations:** | 75 | **AIC:** | 1506. |
| **Df Residuals:** | 57 | **BIC:** | 1548. |
| **Df Model:** | 17 | | |
| **Covariance Type:** | nonrobust | | |

R값과 Adj.R값이 현저히 상승한 것을 확인할 수 있다.

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 2.166e+04 | 9826.583 | 2.204 | 0.032 | 1982.504 | 4.13e+04 |
| **month[T.AUG]** | 2438.7512 | 3039.321 | 0.802 | 0.426 | -3647.383 | 8524.886 |
| **month[T.JUL]** | 3263.3712 | 2803.336 | 1.164 | 0.249 | -2350.213 | 8876.956 |
| **month[T.JUN]** | 6795.5257 | 2615.328 | 2.598 | 0.012 | 1558.423 | 1.2e+04 |
| **month[T.MAY]** | -1517.9311 | 2289.498 | -0.663 | 0.510 | -6102.571 | 3066.709 |
| **month[T.OCT]** | -1179.1911 | 4790.407 | -0.246 | 0.806 | -1.08e+04 | 8413.433 |
| **month[T.SEP]** | -3255.9311 | 4221.818 | -0.771 | 0.444 | -1.17e+04 | 5198.113 |
| **day_of_week[T.Monday]** | -1990.9897 | 2185.367 | -0.911 | 0.366 | -6367.112 | 2385.133 |
| **day_of_week[T.Saturday]** | 6210.4964 | 1971.039 | 3.151 | 0.003 | 2263.558 | 1.02e+04 |
| **day_of_week[T.Sunday]** | 6508.6933 | 2501.401 | 2.602 | 0.012 | 1499.724 | 1.15e+04 |
| **day_of_week[T.Thursday]** | 620.3745 | 2514.981 | 0.247 | 0.806 | -4415.787 | 5656.536 |
| **day_of_week[T.Tuesday]** | 6855.6350 | 2316.955 | 2.959 | 0.004 | 2216.013 | 1.15e+04 |
| **day_of_week[T.Wednesday]** | -1108.2009 | 2300.887 | -0.482 | 0.632 | -5715.647 | 3499.245 |
| **skies[T.Cloudy]** | -510.5413 | 1746.027 | -0.292 | 0.771 | -4006.900 | 2985.818 |
| **day_night[T.Night]** | 2339.7628 | 2669.901 | 0.876 | 0.385 | -3006.621 | 7686.146 |
| **shirt[T.YES]** | 1.031e+04 | 3934.469 | 2.620 | 0.011 | 2427.772 | 1.82e+04 |
| **fireworks[T.YES]** | 4563.8873 | 1917.618 | 2.380 | 0.021 | 723.923 | 8403.852 |
| **park[T.YES]** | 9799.9799 | 2270.521 | 4.316 | 0.000 | 5253.340 | 1.43e+04 |
| **temp** | 153.0028 | 160.112 | 0.956 | 0.343 | -167.615 | 473.621 |

하지만 1차 회귀분석 결과와 달리 cap(모자 프로모션) 변수가 사라진 것을 확인할 수 있다. 아웃라이어 제거 전 1차 회귀분석에서 cap변수의 p-value는 0.251 로 유의미하지 못하다고 판단되고, Fox판단 결과 cap변수 자체가 아웃라이어로 판단되었으므로 cap변수는 유의미한 프로모션이 아니라는 판단이 가능하다.
또한 날씨변수(Skies, day_night, temp)의 p-value가 각각 0.771, 0.385, 0.343으로 유의미하지 않다고 판단된다.
따라서 이후에는 남은 두 개의 프로모션(shirts, fireworks)만을 변수로 하여 attend를 예측하는 분석을 한다.

| | | | | | | |
|---|---|---|---|---|---|---|
| day_of_week[T.Wednesday] | 2984.2084 | 2550.325 | 1.170 | 0.246 | -2113.819 | 8082.235 |
| skies[T.Cloudy] | -1746.0631 | 2011.937 | -0.868 | 0.389 | -5767.868 | 2275.741 |
| day_night[T.Night] | -169.4715 | 3003.256 | -0.056 | 0.955 | -6172.895 | 5833.952 |
| cap[T.YES] | -5571.9439 | 4810.931 | -1.158 | 0.251 | -1.52e+04 | 4044.969 |
| shirt[T.YES] | 5782.3774 | 3893.811 | 1.485 | 0.143 | -2001.239 | 1.36e+04 |
| fireworks[T.YES] | 5452.0106 | 2303.971 | 2.366 | 0.021 | 846.438 | 1.01e+04 |
| park[T.YES] | 1.024e+04 | 2700.473 | 3.792 | 0.000 | 4843.337 | 1.56e+04 |
| temp | 47.5081 | 186.494 | 0.255 | 0.800 | -325.289 | 420.305 |

| | | | |
|---|---|---|---|
| Omnibus: | 9.424 | Durbin-Watson: | 2.233 |
| Prob(Omnibus): | 0.009 | Jarque-Bera (JB): | 9.256 |
| Skew: | 0.796 | Prob(JB): | 0.00977 |

(1차 회귀분석 결과에서는 cap 변수와 p-value를 확인할 수 있다.)

5. 프로모션 변수(shirt, fireworks) 회귀분석

앞서 cap은 효과가 없음을 판단했기 때문에, shirt와 fireworks 변수에 대한 분석을 진행한다.

| | p-value | Adj.R |
|---|---|---|
| shirt | 0.011 | 0.628 |
| fireworks | 0.041 | 0.590 |

(회귀분석 전체 요약 자료는 분량상 생략)

분석결과 shirt와 fireworks의 p-value가 일반적인 유의수준 0.05보다 낮으므로 두 프로모션 모두 관중 수의 증가에 유의미한 영향을 미친다고 판단할 수 있다.

조금 더 자세히 살펴보자면,
a)
fireworks의 경우 매주 friday에 시행되었는데, 이때 시행되지 않았던 friday에 대한 데이터가 존재하지 않는다.

b)
shirt 프로모션의 경우 Monday에 한번 Sunday에 한번 시행되었다. (Tuesday도 있으나, 아웃라이어 제거 과정에서 덜어냄). 이때 각 요일의 평균관중과 shirt 프로모션이 실행되었던 요일

의 관중 수를 비교해보겠다.

| | Monday | Sunday |
|---|---|---|
| 평균관중 수 | 34965 | 42550 |
| Shirt 프로모션 시 관중 수 | 50559 | 48753 |

확인 결과 shirt 프로모션 시 각 요일의 평균관중 수보다 훨씬 많은 관중을 유인하는 것을 알 수 있다. 회귀분석과 상응하는 결과이다.
특히 Monday의 관중 수 상승에 주목할 필요가 있다. 왜냐하면 Monday는 일주일 중 평균관중 수가 가장 작은 요일 중 하나인데, shirt 프로모션이 약 60%의 관중 수 상승을 이끌어냈기 때문이다.

| | |
|---|---|
| Monday | 34965 |
| Tuesday | 49013 |
| Wednesday | 34689 |
| Thursday | 40407 |
| Friday | 40116 |
| Saturday | 43072 |
| Sunday | 42550 |

(요일별 평균관중 수 )

하지만 이 역시 데이터의 수가 현저히 적으므로 데이터의 수를 늘려가며 단계적으로 프로모션을 진행할 필요가 있다.

## 결론 및 한계점

회귀분석 결과 3가지 프로모션 (cap, shirt, fireworks) 중 cap을 제외한 shirt와 fireworks는 관중 상승에 유의미한 효과를 줄 수 있을 것으로 예상된다.

Shirt 프로모션의 경우를 살펴보자면, Monday는 일주일 중 평균관중 수가 두 번째로 작은 34965를 기록하고 있는데(wednesday, 34689) 이때 shirt 프로모션을 시행했을 시 무려 50559의 관중을 불러왔다. 평균관중 수가 42550인 Sunday에 shirt 프로모션을 진행했을 시에는 48753의 관중을 불러왔다.
평균관중 수가 가장 낮은 Wednesday에도 shirt 프로모션을 시행하여 평균관중 수의 전반적인 상승을 기대할 수 있을 듯하다.
하지만 shirt 프로모션을 진행한 횟수가 너무나 적기에 시행횟수를 늘려가며 단계적인 프로모션을 진행할 필요가 있어 보인다.

이는 fireworks 또한 마찬가지이다. fireworks는 항상 Friday에 시행되었는데, 시행되지 않았던 Friday에 대한 데이터가 없다. 수치상 p-value값이 유의미한 것으로 나왔지만, fireworks를 시행하지 않은 Friday에 대한 데이터 수집이 필요해 보인다.