KBO 우천취소 예측

Naive Bayes & KNN

한경민

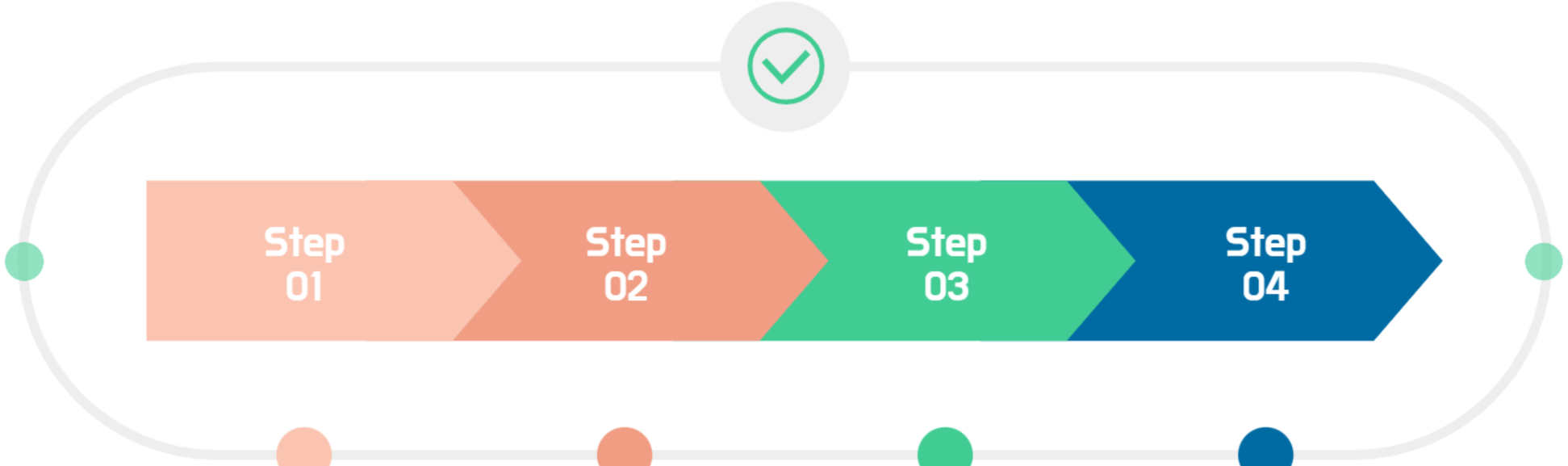
목차

분석목표

데이터 수집

시각화 및 모델링

분석목표



데이터 수집

-분석에 필요한 데이터를 확인하고, 웹크롤링을 이용하여 수집할 수 있다.

데이터 전처리

-수집한 데이터를 필요한 데이터로 변환할 수 있다.

시각화

-데이터를 시각화하여 데이터의 특성을 파악할 수 있다.

분류

-Naive Bayes & KNN을 이용하여 우천취소/정상경기 여부를 분류 및 예측할 수 있다.

Packages

데이터 수집

```
from bs4 import BeautifulSoup
from selenium import webdriver
import time
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
```

데이터 처리

```
import pandas as pd
import numpy as np
from pandas import Series, DataFrame
from collections import Counter
```

모델링

```
from sklearn.naive_bayes import MultinomialNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import GridSearchCV, tr
ain_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix, classific
ation_report
```

데이터 수집

경기일정/결과

홈 > 일정/결과 > 경기일정/결과

2021 08

KBO 정규시즌 일정

날짜	시간	경기			게임센터	하이라이트	TV	라디오	구장	비고
08.10(화)	18:30	SSG	0 vs 4	LG	리뷰	하이라이트	SPO-T		잠실	-
	18:30	두산	vs	삼성			MS-T	T-R	대구	우천취소
	18:30	롯데	5 vs 2	NC	리뷰	하이라이트	SPO-2T	KNN-R	창원	-
	18:30	한화	1 vs 4	KIA	리뷰	하이라이트	KN-T G-CMB		광주	-
	18:30	KT	1 vs 3	키움	리뷰	하이라이트	SS-T		고척	-

KBO 홈페이지에서 각 구장별 경기 일정 스크래핑

지점 광주(유) 선택 년도 2020년 선택 요소 강수량 선택

[일강수량(mm)] 156 광주 / 2020년

일자	1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
1일			0.0	0.0		0.0		2.9	0.1		1.7	
2일					0.3			0.7	49.4		0.3	
3일		0.0			25.4	0.0	7.4		17.0	0.5	0.0	
4일			0.0				0.0			0.9		
5일		0.0			0.0			6.6	0.0			
6일	15.5					0.2	0.3	38.8	11.3		0.0	
7일	32.0		9.1				0.0	259.5	30.1			

기상청 홈페이지에서 각 지역 일자별 강수량 스크래핑

미리케이터링 메뉴

데이터 수집(코드)

```
for k in range(3,8):
    driver.find_element(By.CSS_SELECTOR,'select#ddlYear > option:nth-child('+str(k)+')').click()
    time.sleep(0.7)
    for j in range(3,11):
        driver.find_element(By.CSS_SELECTOR,'select#ddlMonth > option:nth-child('+str(j)+')').click()
        year = soup.select_one('select#ddlYear > option:nth-child('+str(k)+')').text
        time.sleep(2)
        soup = BeautifulSoup(driver.page_source)
        table = [i for i in soup.select('tbody > tr > td')]
        date = []
        for i in table:
            try:
                if i.attrs['class'] == ['day']:
                    date_tmp = year+'.'+i.text
                elif i.attrs['class'] == ['time']:
                    date.append(date_tmp)
            except:
                pass
        base_time = [i.text for i in soup.select('tbody > tr > td.time')]
        for i in soup.select('td.day'):
            i.extract()
        stadium = [i.text for i in soup.select('tbody > tr > td:nth-child(7)')]
        cancellation = [i.text for i in soup.select('tbody > tr > td:nth-child(8)')]
        df = df.append(DataFrame({'일자':date,'시간':base_time,'경기장':stadium,'취소여부':cancellation}),ignore_index=True)
```

KBO 홈페이지에서 각 구장별 경기 일정 여부 스크래핑

```
stn_lst = {108:'서울',112:'인천',119:'수원',155:'창원',159:'부산',143:'대구',156:'광주',
           133:'대전',138:'포항',152:'울산',131:'청주'}
driver = webdriver.Chrome('c:/data_bigdata/chromedriver.exe')
result = DataFrame(columns=['일자','value','지역'])
for k in stn_lst.keys():
    for i in range(2016,2022):
        driver.get('https://www.weather.go.kr/w/obs-climate/land/past-obs/obs-by-element.do?stn='+str(k)+'&yy='+str(i)+'&obs=21')
        time.sleep(1)
        soup = BeautifulSoup(driver.page_source)
        month = [j.text for j in soup.select('tr.tablesorter-headerRow > th')]
        month.remove('일자')
        rain = [j.text.strip() for j in soup.select('#weather_table>tbody>tr>td')]
        day = [rain[j] for j in range(0,416,13)]
        for j in day:
            rain.remove(j)
        baseball_pivot = DataFrame(np.array(rain).reshape(32,12),columns=month,index=day)
        baseball_pivot = baseball_pivot.drop('합계')
        baseball_pivot = baseball_pivot.reset_index()
        baseball_pivot = baseball_pivot.replace('', np.nan).fillna(0)
        id_var = list(baseball_pivot.columns)
        baseball_pivot.iloc[:,1:12].astype('float')
        baseball_unpivot = pd.melt(baseball_pivot,id_vars=['index'],value_vars=id_var)
        baseball_unpivot.insert(0,'일자',str(i)+'년'+baseball_unpivot['variable']+baseball_unpivot['index'])
        baseball_unpivot = baseball_unpivot.drop(['index','variable'],axis=1)
        for j in soup.select('option'):
            if j.attrs.get('value') == str(k):
                baseball_unpivot['지역'] = j.text.split('(')[0]
        result = result.append(baseball_unpivot,ignore_index=True)
        time.sleep(2)
```

기상청 홈페이지에서 각 지역 일자별 강수량 스크래핑

데이터 전처리

KBO 데이터

	일자	시간	경기장	취소여부
0	2021.04.03(토)	14:00	잠실	우천취소
1	2021.04.03(토)	14:00	문학	우천취소
2	2021.04.03(토)	14:00	창원	우천취소
3	2021.04.03(토)	14:00	수원	우천취소
4	2021.04.03(토)	14:00	고척	-
..	...	...	..	...
115	2021.04.30(금)	18:30	잠실	-
116	2021.04.30(금)	18:30	사직	-
117	2021.04.30(금)	18:30	대구	-
118	2021.04.30(금)	18:30	수원	-
119	2021.04.30(금)	19:00	창원	-

일자 형식 변경

일자가 요일을 포함한 문자로 되어있어 이를 일자 형식으로 변환

POINT1

취소여부 수정

정상 경기 진행 여부를 입력 후 우천취소 제외 다른 취소 사유는 삭제함

POINT2

경기장 수정

마산이 추후 창원으로 변경되어 마산을 창원으로 변경 / 고척은 실내 구장이므로 데이터에서 삭제

POINT3

데이터 전처리

기상청 데이터

index	1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
0	1일	0	0	0	0	0	108.5	0	0	0	0	0
1	2일	0	0	0	0	16.5	0	4.0	2.0	4.5	18.5	0.1	0
2	3일	0	0	0	2.0	27.0	0	0	0	0.5	6.5	0.2	0
3	4일	0	0	0.1	0	0	0	29.5	0	0	0	0.0	0
4	5일	0	0	40.0	0	7.5	0	100.5	0	0	0	0	1.5
5	6일	0	0	0.2	3.0	9.5	0.0	0.0	0	0	0	0	0

일자	value	지역
0	2016년1월1일	0 서울
1	2016년1월2일	0 서울
2	2016년1월3일	0 서울
3	2016년1월4일	0 서울
4	2016년1월5일	0 서울

데이터 형식 변경

행에 일자 / 열에 월로 되어있던 Pivot 형식을 Unpivot 형식으로 변경

POINT1

일자 형식 변경

일자 형식을 KBO에서 추출한 데이터와 동일한 형식으로 변경

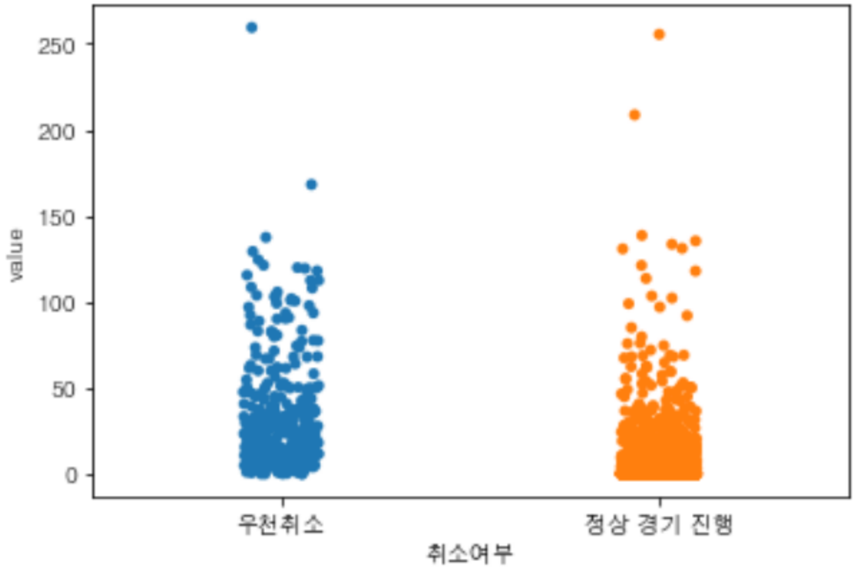
POINT2

경기장명 통일

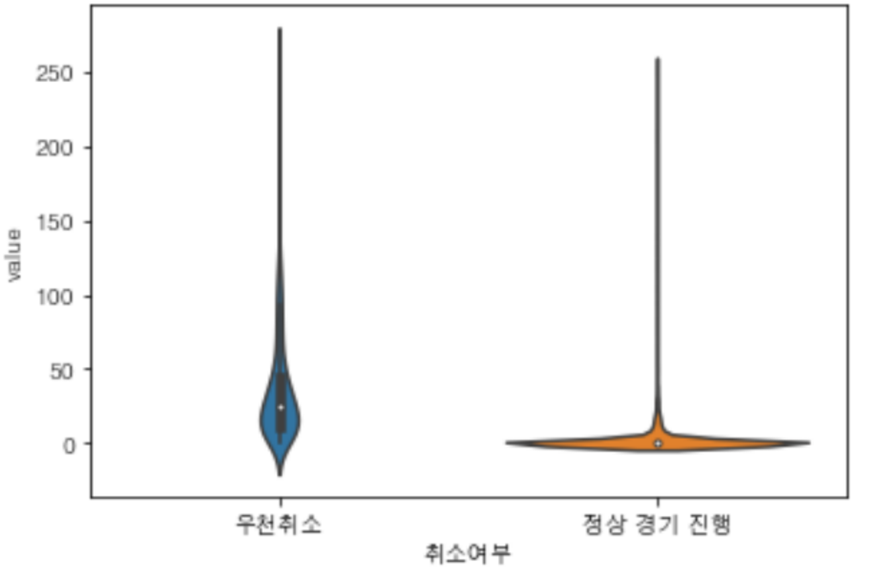
KBO자료는 구장이름, 기상청 자료는 지역이름이므로 이를 구장이름과 통일하여 변경

POINT3

시각화

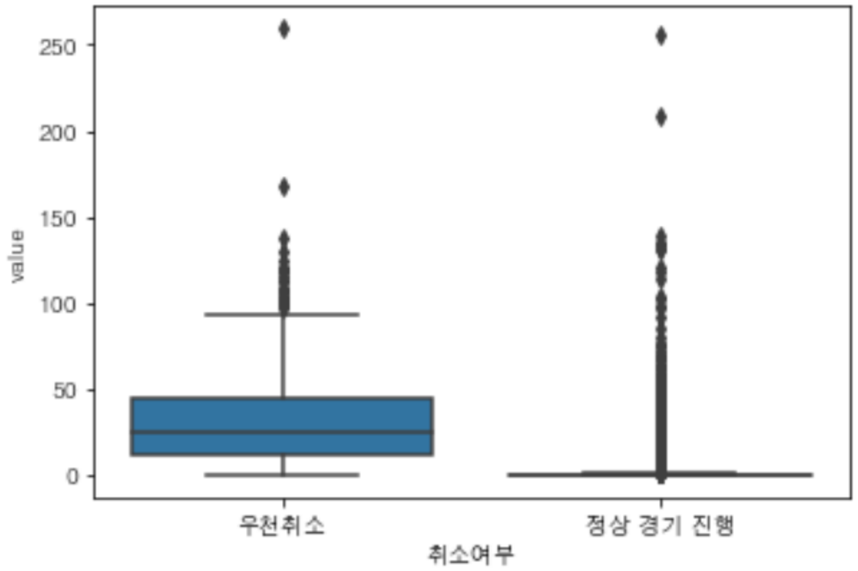


강수량별 경기 진행여부 확인
: Striplot 경우 우천취소의 분포가 크지만 차이가 크지 않음



강수량별 경기 진행 여부 확인
: Violinplot의 경우 정상 경기 진행에 강수량 적은 값이 많음.

시각화



강수량별 경기 진행여부 확인
: Boxplot의 경우 정상 경기 진행의 경우 이상값이 많음.

정상 경기 진행의 이상값을 확인해보니 수치는 맞음. (일 강수량은 높지만 야구 경기 진행 전에 비가 그친 경우)
반대로 우천취소인데 강수량이 0인 경우도 있었음. (예보대로 취소 했으나 비가 오지 않은 경우)

모델링

Naive Bayes 이용

TRAIN, TEST SET

80% / 20%로 훈련과 시험
테스트셋 분리함.

POINT1

정확도 도출

0.9196217494089834
의 정확도 도출

POINT2

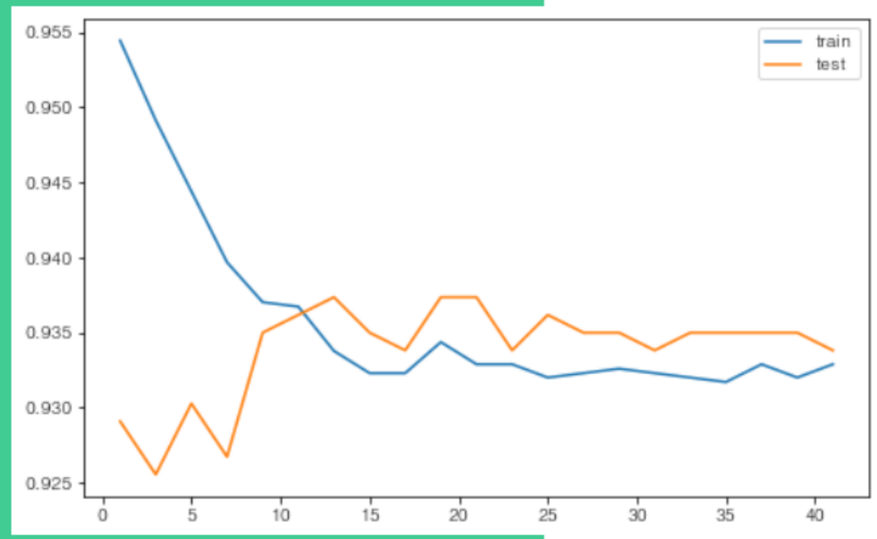
혼합행렬 확인

array([[0, 68],
[0, 778]]로
우천취소로의 분리가 약함

POINT3

모델링

KNN 이용



TRAIN, TEST SET

80% / 20%로 훈련과 시험
테스트셋 분리함.

POINT1

K값 및 정확도 도출

그래프로 최적의 K 확인,
과적합과 과소적합을
방지하기 위해 K를 9로 정함
정확도
0.9361702127659575
도출

POINT2

혼합행렬 확인

array([[29, 39],
[15, 763]]로
Naive Bayes 분류보다 우천
취소의 분류가 잘 됨.

POINT3

개선점

-정상 경기 진행 : 3892 / 우천취소 : 336 건으로 우천취소에 대한 데이터가 현저히 적음.
-시간대별 강수량이 아닌 일강수량으로 경기 전 비가 그치고, 정상 경기 진행될 경우 예측 오류 발생. 시간대별 강수량으로 데이터 수집이 필요.

감사합니다.

한경민