EPL 경기 예측

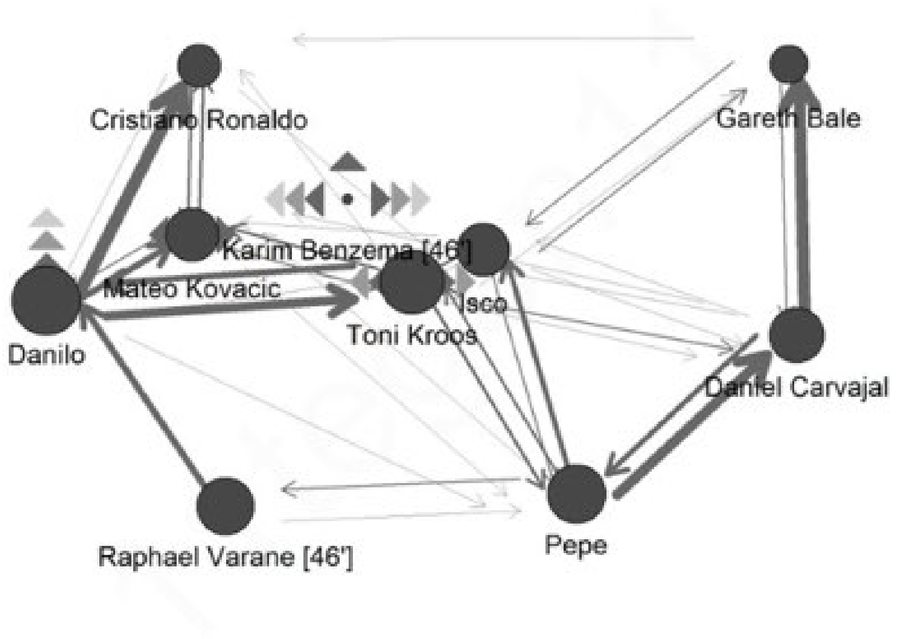
SOCIALMEDIA ANALYTICS
PROJECT

윤기태, 조일혜 황준선

PRESENTATION
CONTENTS

- Topic
- Motivation
- Model Design
- Data
- Modeling & Result
- Conclusion

PROJECT TOPIC



[그림] 선수 개인의 성향을 데이터로 분석한 전술판

프로스포츠와 데이터의 융합

- 최근의 프로스포츠는 **데이터 사이언스 분야**와 융합하여 가파른 성장세를 보이고 있다.
- 프로스포츠 관련 데이터 활용은 선수의 훈련, 선수에 대한 통계, 경기에 대한 통계, 승패 예측 등이 있다.
- 때문에 프로스포츠와 관련된 데이터가 범람하고 있는데 이를 **승패에 대한 예측**에 이용해보기로 하였다.

PROJECT TOPIC

슈퍼컴퓨터를 이용하여
프로 스포츠의 **승패를 예측**하는 시대

Home > 뉴스

슈퍼컴퓨터, 맨유 우승 예측.. 2위는 맨시티

기사입력 : 2016.08.11 기사보내기 : 트위터 페이스북

PROJECT TOPIC

PROJECT TOPIC

Who Win??

MOTIVATION

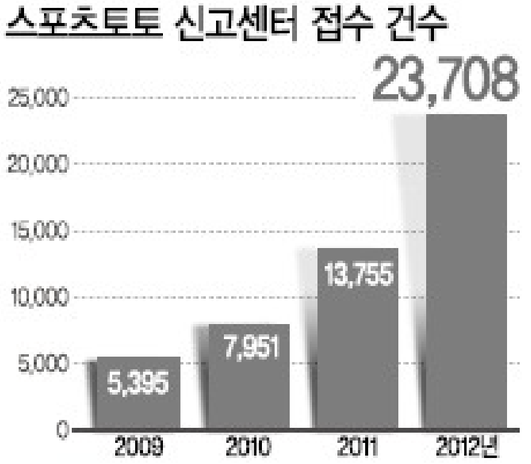
순수한 궁금증

- 0.02%의 희박한 확률로 우승
- **예상치 못한 변수**가 등장하는 프로스포츠에서 얼마나 정확하게 승패를 예측할 수 있을까?란 **호기심**을 자극함

MOTIVATION

불법 도박 근절

최근 스포츠 승패 예측과 관련된 **불법 도박 문제**가 대두되고 있는데, 프로스포츠에 대한 **승패 예측 모델**이 정확해진다면 이 문제가 근절될 것이다.



사설 스포츠토토 운영 사이트 추정치 (2012년)

구분	개수
대규모(가입자 2000~3000명)	10~15개
중규모(가입자 1000여명)	40~50개
소규모(가입자 500여명)	200~300개

MODEL DESIGN

MODEL DESIGN

Name	Season	Apps	Mins	Goals	Assists	SpG	KeyP	Drb	Fouled	Off	Disp	UnsTch	Rating	Team.x	Meta	Team.y	Cost
Andy King	2014	16(8)	1374	2	0	0.6	0.2	0.4	0.6	0	0.4	0.7	6.73	Leicester	AM(C)	leicester-city	1700000
Andy King	2015	9(16)	1050	2	1	0.8	0.4	0.3	0.3	0	0.2	0.5	6.56	Leicester	AM(C)	leicester-city	2130000
Andy King	2016	6(4)	624	0	0	0.6	0.4	0.3	0.9	0.1	0.6	0.7	6.53	Leicester	AM(C)	leicester-city	2980000
Aruna Dindan	2009	18(1)	1479	8	2	2.5	1	1.3	3.1	0.4	2.2	0	7.04	Portsmouth	AM(C)	NA	NA
David Dunn	2009	20(3)	1720	9	2	1.7	1.6	1.2	1.7	0.5	1.9	0	7.11	Blackburn	AM(C)	NA	NA
David Dunn	2010	17(10)	1389	2	1	0.9	0.6	0.7	0.6	0	1.1	1	6.5	Blackburn	AM(C)	NA	NA
David Dunn	2011	21(5)	1613	2	1	1.5	0.7	0.6	1.1	0.1	0.8	0.9	6.63	Blackburn	AM(C)	NA	NA
Geovanni	2009	16(10)	1539	3	0	1.8	0.7	0.8	1.9	0.2	1.5	0	6.63	Hull	AM(C)	hull-city	3830000
Giorgos Kara	2012	20(5)	1556	1	0	0.8	0.6	0.7	2.2	0	1	0.3	6.69	Fulham	AM(C)	NA	NA
Jonathan Gre	2009	15(8)	1502	1	1	0.3	0.5	0.2	0.6	0	0.7	0	6.68	Fulham	AM(C)	west-bromwi	3490000
Ricardo Gard	2009	11(10)	1107	1	1	0.7	0.7	0.4	0.8	0.2	1.2	0	6.61	Bolton	AM(C)	NA	NA
Robert Koren	2013	10(12)	1136	2	0	1.2	0.5	0.2	0.4	0.1	0.6	0.8	6.39	Hull	AM(C)	hull-city	1280000
Benik Afobe	2016	0(10)	176	0	1	0.9	0.8	0.1	0.2	0.6	0.8	0.3	6.21	Bournemouth	AM(C),FW	afc-bournem	7650000
Danny Ings	2014	35	3035	11	4	2.8	1	1.6	1.4	0.5	2.1	2.5	6.89	Burnley	AM(C),FW	burnley-fc	2550000
Dimitar Berba	2009	24(9)	2006	12	5	3	1.6	0.9	1.1	0.7	0.9	0	7.41	Manchester U	AM(C),FW	manchester-	2.81E+7

축구에 대한 직관적 판단을 통해, 과적합을 발생시키거나
의미 없는 계산을 하게 할 변수들을 미리 제거하였음.

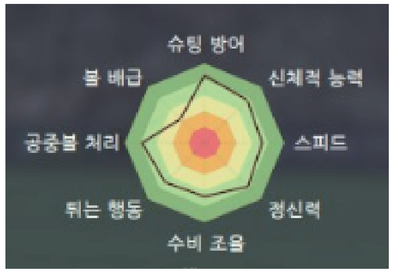
MODEL DESIGN

구단과 감독 각각의 승점으로 승부 예측변수

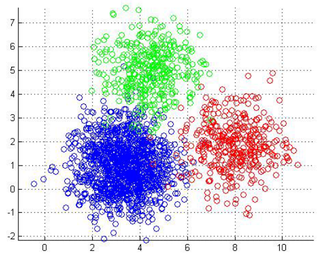
선수 데이터 변수를 어떤 방식으로 적용 할 것인가

MODEL DESIGN

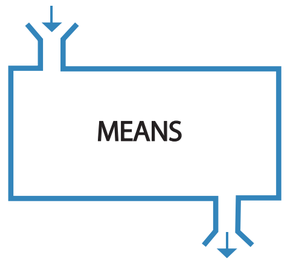
~~선수 데이터를 점수화~~



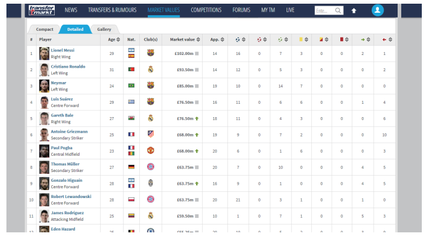
~~클러스터링 이용 선수 그룹핑~~



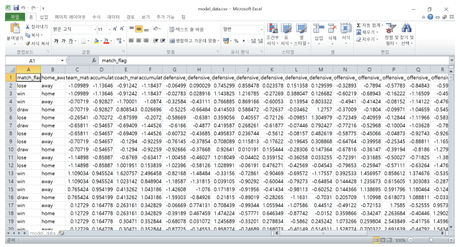
선수 스탯의 평균을 이용



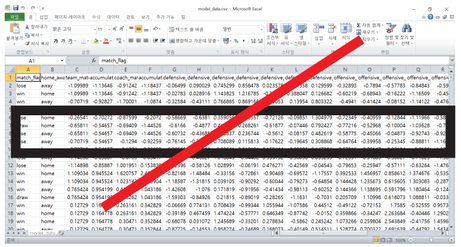
DATA



축구 통계 사이트에서
데이터 크롤링

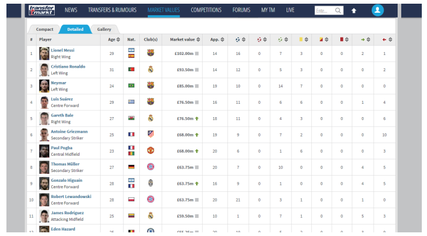


데이터를 합쳐서
Input Data 생성



결측값 제거등의
데이터 전처리 과정

DATA



축구 통계 사이트에서
데이터 크롤링

2009년부터의 EPL 데이터 수집

구단 데이터 - 모든 경기의 승/무/패
해당 경기 선수 라인업, 경기 감독 정보

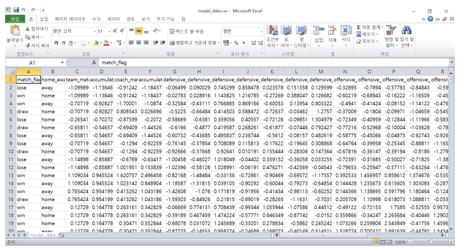
감독 데이터 - 모든 경기의 승/무/패
해당 시즌과 소속팀

선수 데이터 - 경기 관련 Stat 데이터
이적료와 소속팀, 평점과 해당 시즌

DATA

구단 데이터 - 모든 경기의 승/무/패
해당 경기 선수 라인업, 경기 감독 정보

감독 데이터 - 모든 경기의 승/무/패
해당 시즌과 소속팀



데이터를 합쳐서
Input Data 생성

선수 데이터 - 경기 관련 Stat 데이터
이적료와 소속팀, 평점과 해당 시즌

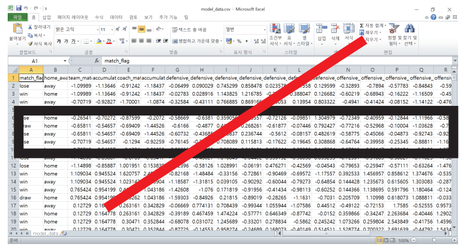
취합한 데이터를 Model에 적용하기
위해 Input Data 형태로 정리

DATA

전처리 과정

filter를 통해 데이터를 합치는 과정에서 중복되거나 에러가 나는 데이터를 정리, 분포조정을 진행.

어쩔 수 없는 결측값에 대한 처리 : 일단 모델에 적용해 본 뒤, 결측값을 제거한 것과 무시한 것 등의 결과를 비교하여 최종적으로 처리함.



결측값 제거등의
데이터 전처리 과정

RESULT

Multinomial
Logistic Regression

k-NN

Naïve Bayes

Decision Tree

RESULT

Multinomial Logistic Regression

k-NN

Naïve Bayes

Decision Tree

Logistic Regression

전역 탐색을 사용하려고 하였으나, 변수가 너무 많아서 노트북으로 감당하기 힘들어, step 함수를 이용하여 단계 선택 방식으로 **변수를 설정**하였음.

선택된 변수에 대해서 공선성 검증을 한 결과, 모두 10 이하로 나타났음.

RESULT

```
test_predicted_EPL_lm = predict(EPL_lm, type="class", newdata=test_data)
1-sum(test_data$match_flag != test_predicted_EPL_lm)/nrow(test_data)
```

```
## [1] 0.4900662
```

```
table(test_data$match_flag, test_predicted_EPL_lm)
```

```
##       test_predicted_EPL_lm
##        draw lose win
##   draw    8    7  26
##   lose    7   21  16
##   win    13    8  45
```

RESULT

Multinomial
Logistic Regression

k-NN

Naïve Bayes

Decision Tree

Naive Bayes

rfe 함수를 이용하여 변수를 설정하였음.
laplace 조정을 시도한 결과, 아무런 의미가 없었음.

RESULT

```
EPL_nb = naiveBayes(formula = match_flag ~ player_rating+team_match_point+accumulated_team_match_point+offensive_goals+
                      accumulated_coach_match_point+player_cost+passing_keyp+passing_assists+offensive_assists+offensive_spg+
                      offensive_keyp+coach_match_point+passing_thrb+home_away+offensive_drb,
                    data=train_data, laplace = 1)
test_predicted_EPL_nb = predict(EPL_nb, newdata = test_data)
1-sum(test_data$match_flag != test_predicted_EPL_nb)/nrow(test_data)
```

```
## [1] 0.4900662
```

```
table(test_data$match_flag, test_predicted_EPL_nb)
```

```
##       test_predicted_EPL_nb
##        draw lose win
##   draw    8   12  21
##   lose    6   26  12
##   win    12   14  40
```

RESULT

Multinomial Logistic Regression

k-NN

Naïve Bayes

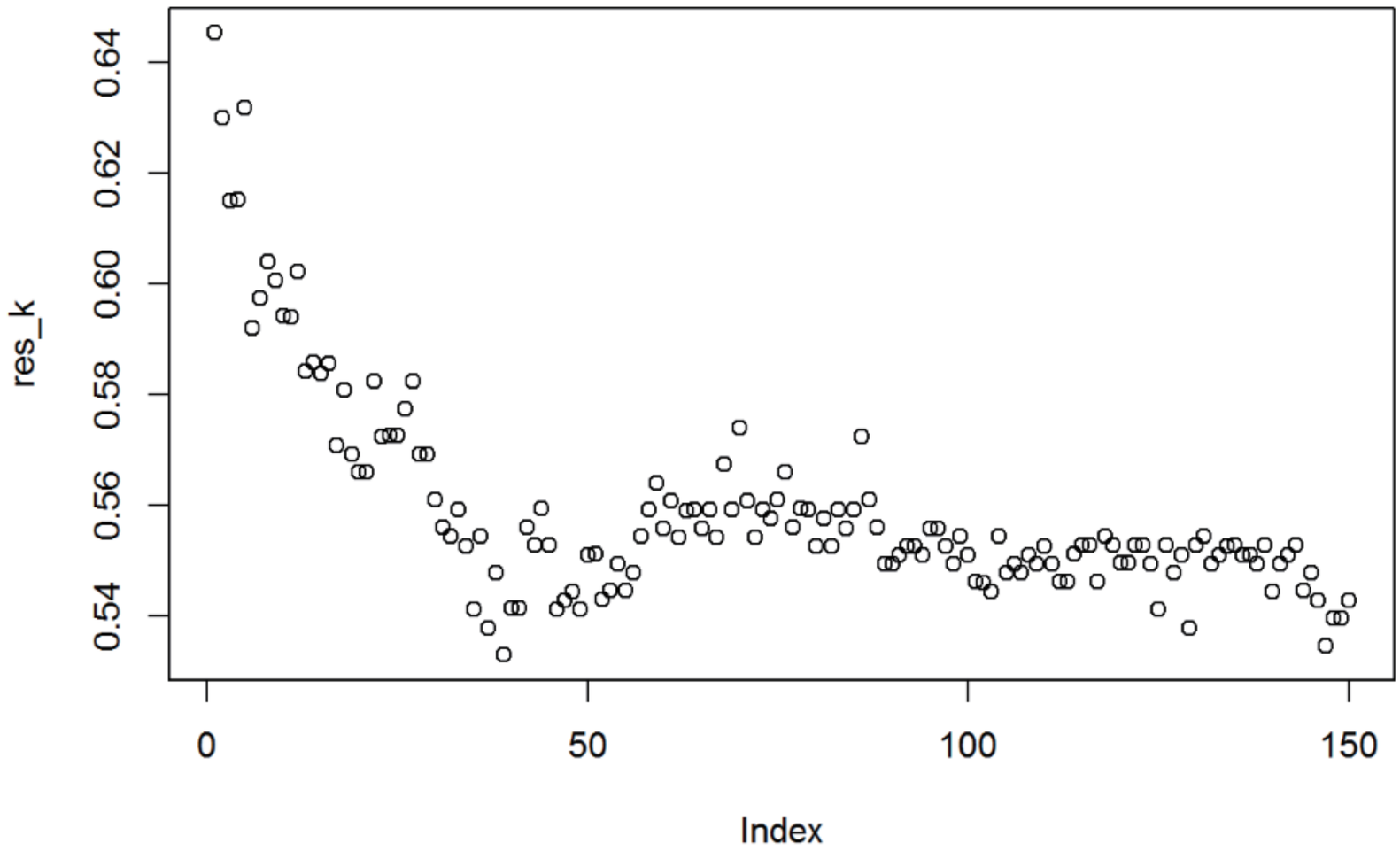
Decision Tree

k-NN

?forward.search() 함수를 이용하여 변수를 설정하였음.

k값은 10-fold Cross Validation을 이용하여 Overfitting을 해결하였음.

RESULT



RESULT

```
fit_k = which(res_k == min(res_k))

test_predicted_EPL_knn = knn(train = train_data, test = test_data, cl = train_data_cl, fit_k)
1-mean(test_data_cl != test_predicted_EPL_knn)
```

```
## [1] 0.5165563
```

```
table(test_data_cl, test_predicted_EPL_knn)
```

```
##             test_predicted_EPL_knn
## test_data_cl draw lose win
##         draw    6    9  26
##         lose    4   19  21
##         win     6    7  53
```

RESULT

Multinomial
Logistic Regression

k-NN

Naïve Bayes

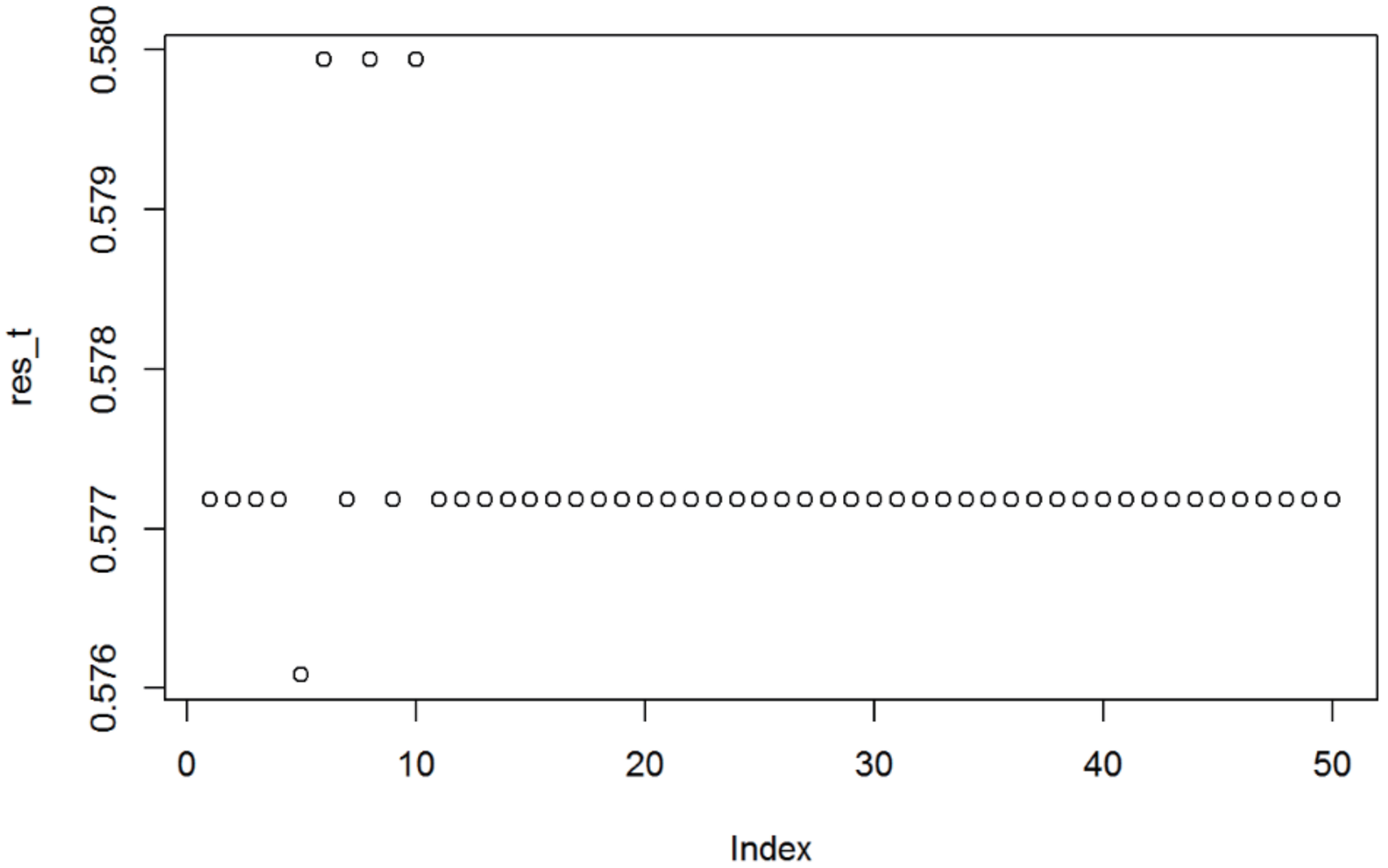
Decision Tree

Decision Tree

?forward.search() 함수를 이용하여 변수를 설정하였음.

1~50 횟수의 Boosting을 실행한 결과, 의미가 있었음.

RESULT



RESULT

```
fit_t = which(res_t == min(res_t))

train__EPL_tree <- C5.0(formula = f, data=train_data, trials = fit_t)
test_predicted_EPL_tree = predict(train__EPL_tree, newdata = test_data)
1-sum(test_data$match_flag != test_predicted_EPL_tree)/nrow(test_data)
```

```
## [1] 0.5364238
```

```
table(test_data$match_flag, test_predicted_EPL_tree)
```

```
##       test_predicted_EPL_tree
##        draw lose win
##   draw    0   13  28
##   lose    0   24  20
##   win     0    9  57
```

CONCLUSION

Multinomial Logistic Regression

k-NN

Naïve Bayes

Decision Tree

결과값이 가장 높은 Decision Tree로 결정하였다.

CONCLUSION

중요 변수

Decision Tree의 경우, Home/Away와 Player Rating 두 가지의 변수가 예측의 상관 관계가 가장 유의했다.

다른 모형들의 경우도 마찬가지로 구단이나 감독의 데이터보다는 선수의 시장가치나 경기 스탯에 많은 영향을 받는 것으로 나타났다.

CONCLUSION

한계점 및 소감

데이터를 어떻게 준비하고 처리할 지에 대해서 어려움이 있었다.

참조 문헌이나 자료가 전무한 모형을 만드는 작업이었기 때문에, 모형을 디자인하는 과정에서 난항을 겪었다.

Thanks!