

HAN JAEYOON

Data Enthusiast

Master of Engineering
Dept. Social Network Sciences
Kyung Here University

Phone : 010-3806-9224
E-Mail : otzslayer@gmail.com
Github : <http://github.com/otzslayer>
Blog : <http://otzslayer.github.io>

INTRODUCTION



Profile

한재윤 HAN JAEYOON

1992. 01. 06

서울특별시 동대문구 천장산로 46

경희대학교 일반대학원 소셜네트워크과학과 **공학석사**

(2016.03 - 2018.02)

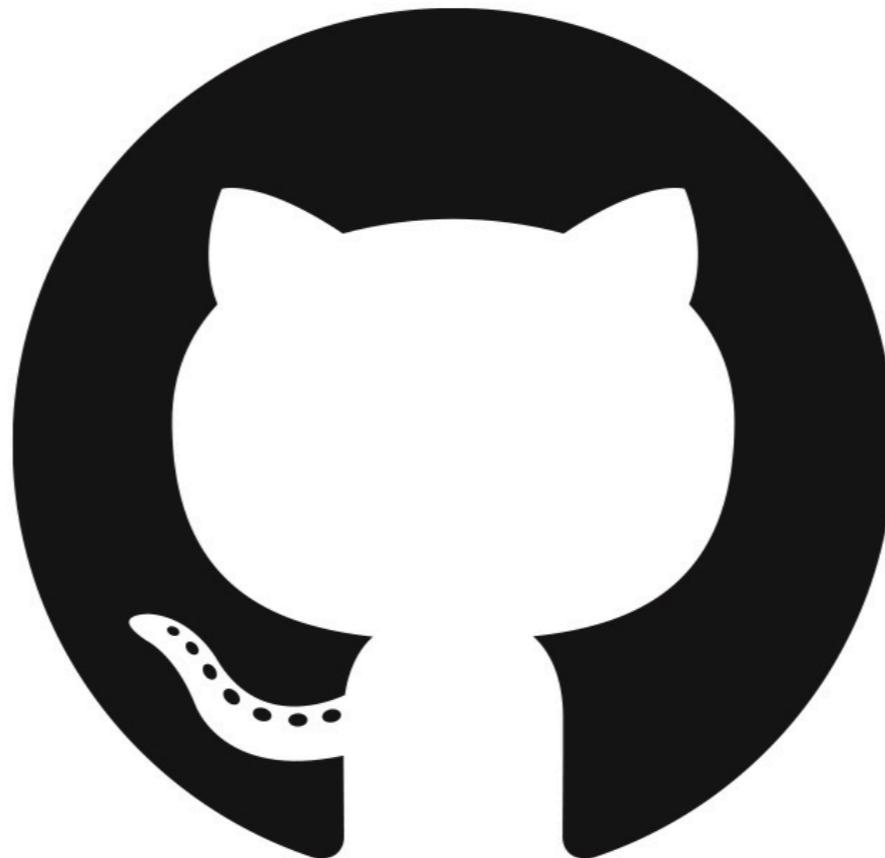
경희대학교 이과대학 수학과 **학사**

(2010.03 - 2016.02)

Research Interest

Machine Learning, Predictive Modeling, Data Visualization

PROJECTS



https://github.com/otzslayer/data_science_portfolio

본 프레젠테이션에 수록한 프로젝트를 제외한
나머지 프로젝트는 모두 **깃허브**에서 확인하실 수 있습니다.

WHAT STRATEGIES DO WE NEED TO SURVIVE FOR A LONG TIME?

PUBG Survival Time Prediction



Data

PlayerUnknown's BattleGround 상위 랭커 데이터

85,000명의 게임 플레이와 관련된 변수들로 구성
승률, 데미지, Top 10 달성을, 차량 이동 시간 등



Environment

Python 3.6+ / pandas, numpy, matplotlib, seaborn, scipy, scikit-learn, xgboost

Issues

1. 특정 타입의 게임만 고집하여 하는 **편향적인 플레이어** 존재
(솔로, 듀오, 스쿼드 중 하나만 하는 플레이어들)
2. 입력변수들과 출력변수 사이의 **데이터 누출** 의심
(오래 생존하였기 때문에 승률이 높다고 설명할 수 있지만,
승률이 높다고 하여 오래 생존했다고 추론할 수 없음)
3. 기존 데이터에 이미 **파생 변수**들이 많이 존재함
4. 각 변수들의 분포에서 **왜도(skewness)**가 크게 나타남



Purpose

게임 플레이 변수들을 이용하여 특정 플레이어의 **스쿼드 평균 생존시간을 예측**
변수 설명이 가능한 알고리즘을 사용하여 **생존에 중요한 변수를 탐색**

Methodology

알고리즘 성능의 비교를 위하여 다양한 알고리즘을 사용 / RMSE, R²로 모델 평가
최소제곱법(OLS), 능형 회귀(Ridge regression), Lasso,
랜덤 포레스트(Random Forest), 그라디언트 부스팅(Gradient boosting)

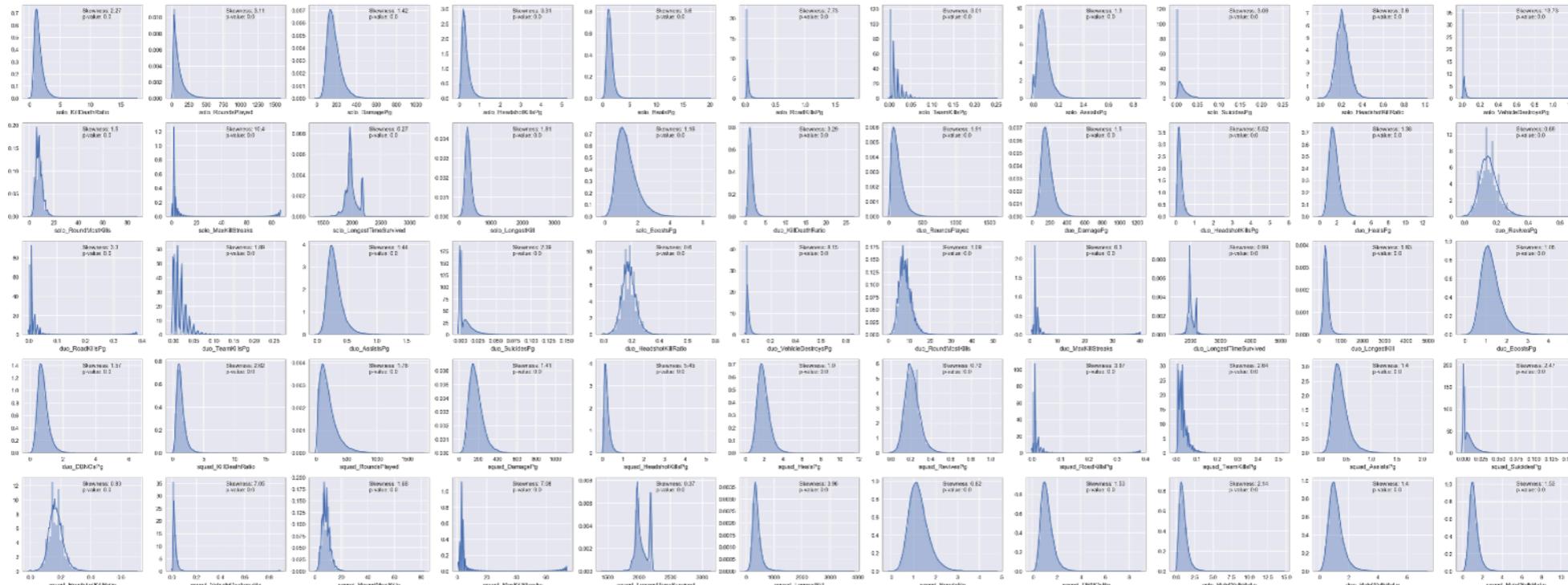
Solution

1. 각 게임 타입을 **20회 이상** 한 플레이어에 한해서 분석
(87,898명에서 57,593명으로 대상을 약 34% 줄임)
2. 승률, 레이팅 등의 **직접적인 변수들을 제거**
평균도보시간과 평균차량탑승시간은
도보와 차량탑승의 비율로 새로운 파생변수 생성
3. 누적값인 변수는 **총 게임수로 나누어** 파생변수 생성
4. 왜도가 2보다 큰 변수들은 **로그를 취하여** 왜도를 제거
(너무 큰 왜도를 갖는 경우, 이를 이용하여도 왜도가 제거되지 않는 경우 존재)

WHAT STRATEGIES DO WE NEED TO SURVIVE FOR A LONG TIME?

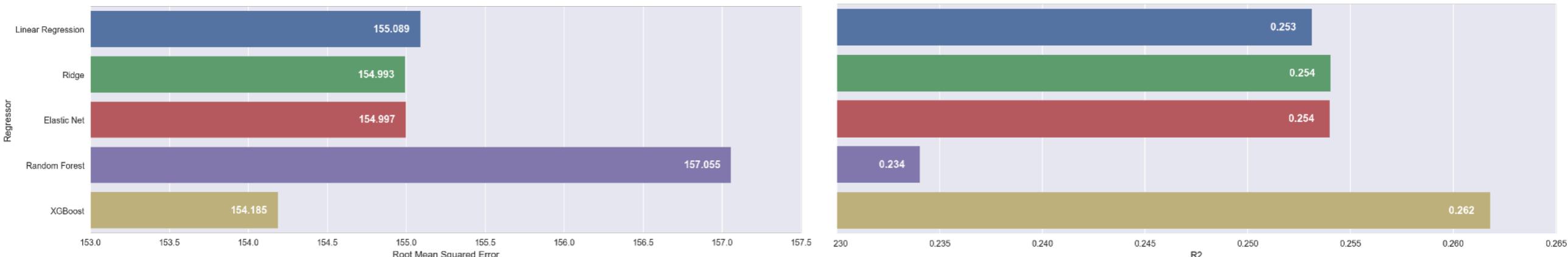
PUBG Survival Time Prediction

 [Github Link](#)



각 변수들의 분포를 나타내는 그래프

Prediction performance for each model

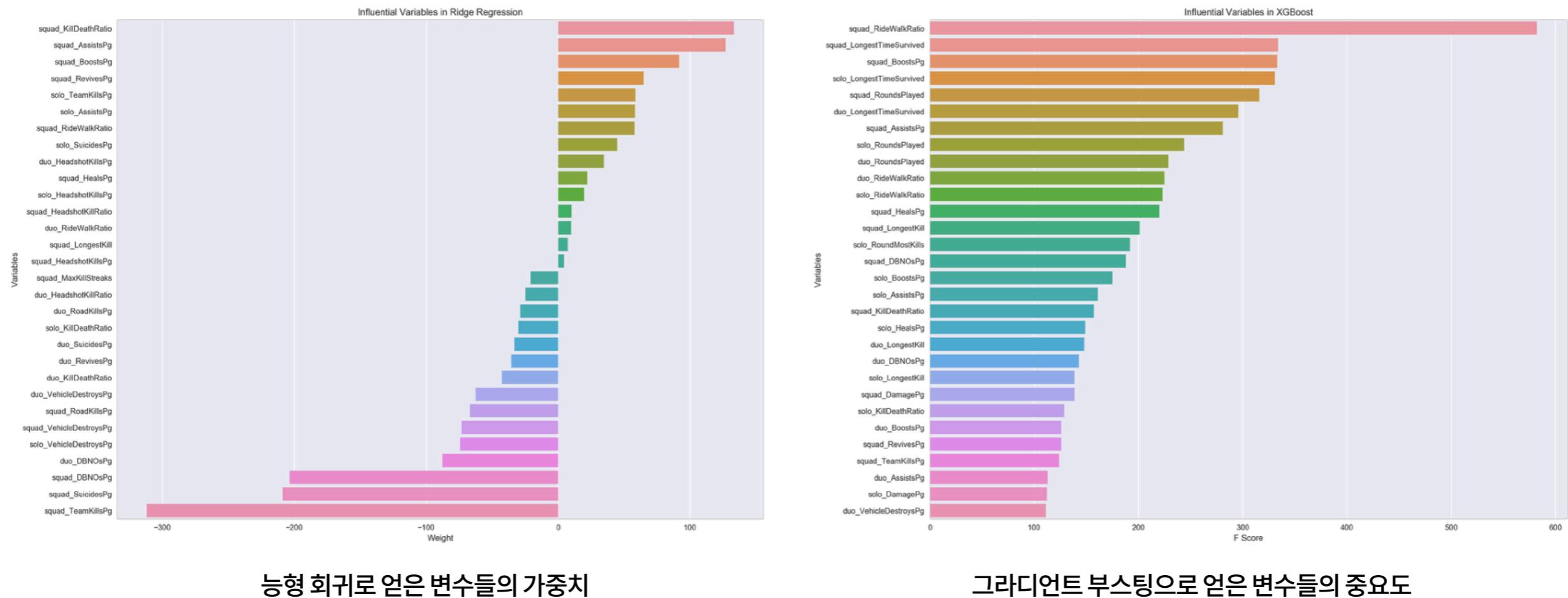


각 모델들의 성능 비교

그라디언트 부스팅의 성능이 가장 좋으며, 랜덤 포레스트의 성능이 가장 나쁨을 알 수 있다.

WHAT STRATEGIES DO WE NEED TO SURVIVE FOR A LONG TIME?

PUBG Survival Time Prediction



에너지드링크 등의 **부스팅과 어시스트**가 생존 시간을 크게 늘려주는 변수이며
걷는 시간보다 **차량을 이용한 시간이 긴 경우** 역시 생존 시간을 크게 늘려줄 수 있다.
추가적으로 **헤드샷 비율**이 높은 경우, 빠른 적 사살로 인해 생존 시간에 도움이 됐다.
반대로 **팀킬**이나 **DBNO(Down But Not Out)**가 많은 경우는 생존 시간을 줄이는 것으로 확인됐다.

HOW CAN WE ACCURATELY PREDICT THE FINANCIAL MARKETS?



Ensemble Approach for Financial Prediction

GRANT FUNDED AND SUPPORTED BY ANONYMOUS CONSORTIUM

Data

다양한 국가 기반의 주가 지수, 환율, 원자재, 국채선물 18종

연합인포맥스에서 2010년부터 2016년까지의 데이터 추출

입력변수의 형태를 로그 수익률로 변환하여 사용

Purpose

18개의 지수의 증감과 지수값을 상호 예측

로그 수익률을 입력 변수로 하여 까다로운 기술적 지표 계산을 피하고

양상별 기법 중 스태킹(stacking)을 이용하여 기존 모델의 성능을 발전시키고자 함

Environment

Python 3.6+ / pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost

Data Preprocessing

휴장일로 인한 결측값은 이전 타임스탬프의 지수값을 그대로 사용

입력변수의 형태를 **1일 전, 5일 전**의 로그 수익률로 변경 (**일간, 주간** 로그수익률)

국채선물의 경우, 음수값을 제거하기 위해 100에서 각 값을 빼서 사용

Features

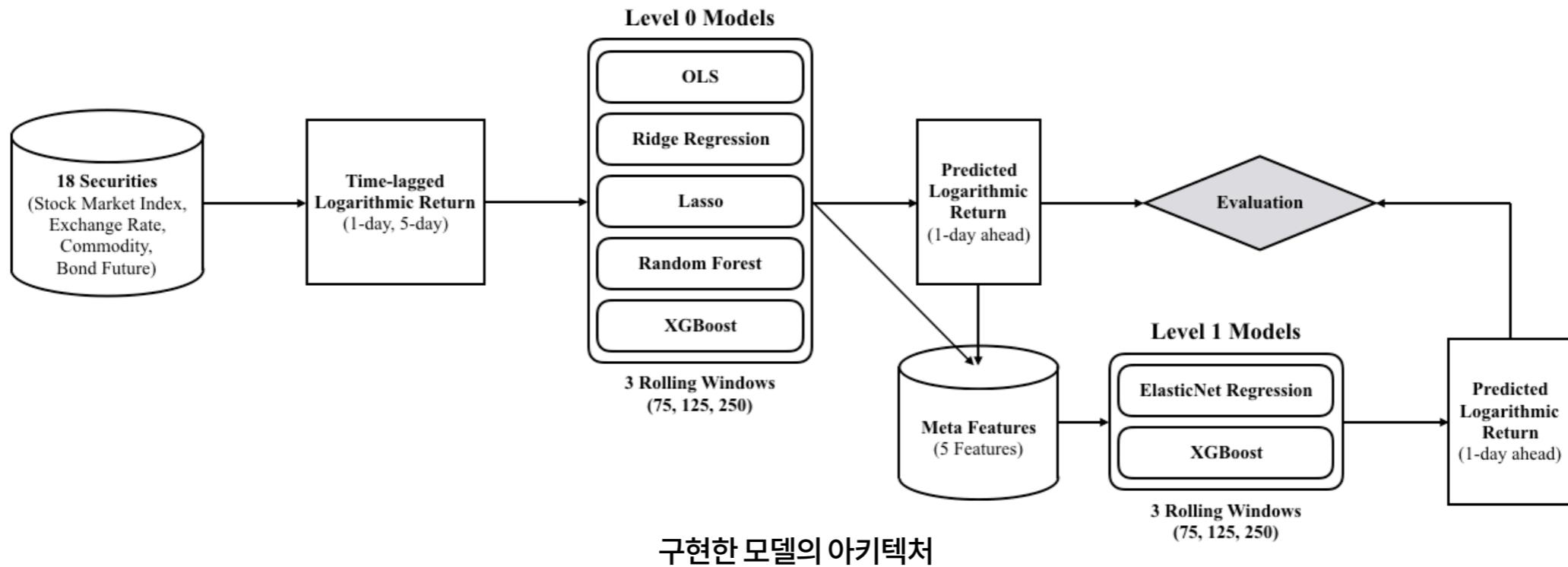
1. 장기간의 추세가 모델의 분산을 크게 만드는 문제를 해결하기 위해 **롤링 윈도우(rolling window)** 기법을 사용
2. 일간 수익률보다 **주간 수익률을 사용한 경우** 성능이 더 좋음
3. Lasso를 이용하여 **모델의 설명력을 극대화** 시킴
4. 기존 금융시장 관련 연구들을 **뒷받침**하는 결과를 보여줌

HOW CAN WE ACCURATELY PREDICT THE FINANCIAL MARKETS?



Ensemble Approach for Financial Prediction

GRANT FUNDED AND SUPPORTED BY ANONYMOUS CONSORTIUM



Models	Lagged	Measures			
		Accuracy	R ²	rRMSE	MAPE
OLS	1-Day	0.486	0.000739	0.00859	0.6116
	5-Day	0.824	0.674622	0.01079	0.8373
Ridge	1-Day	0.488	0.001003	0.00846	0.6020
	5-Day	0.823	0.675341	0.01069	0.8300
LASSO	1-Day	0.488	0.001596	0.00829	0.5837
	5-Day	0.814	0.643632	0.01079	0.8398
RF	1-Day	0.486	0.021385	0.00823	0.5843
	5-Day	0.810	0.605804	0.01129	0.8749
XGB	1-Day	0.486	0.007559	0.00851	0.6100
	5-Day	0.807	0.602628	0.01147	0.8776
XGB (Stacked)	1-Day	0.497	0.005446	0.00905	0.6603
	5-Day	0.828	0.692136	0.01055	0.8087
ELN (Stacked)	1-Day	0.484	0.002283	0.00821	0.5820
	5-Day	0.826	0.677381	0.01039	0.8096

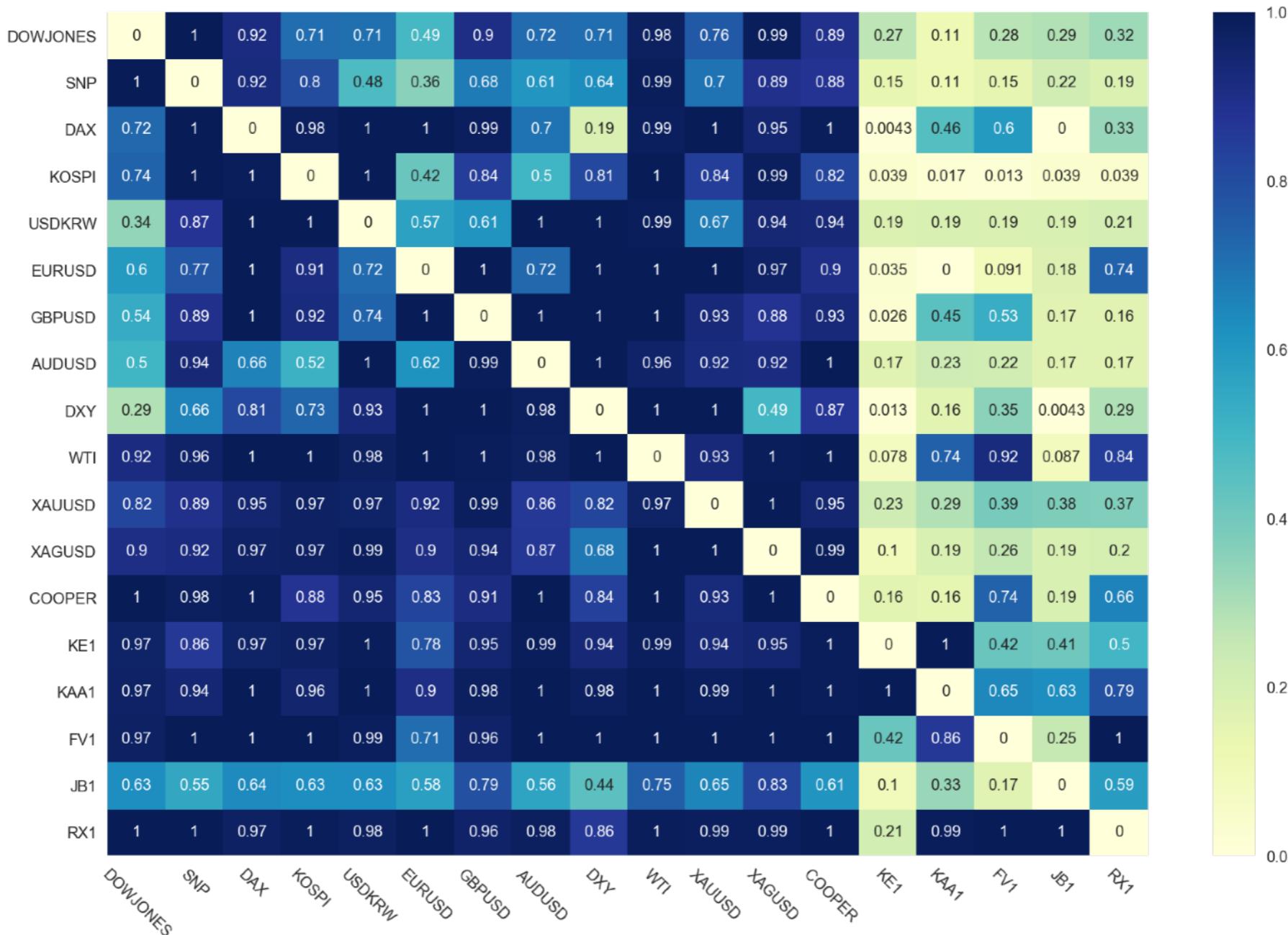
스태킹을 활용한 기법의 성능이 가장 뛰어남

HOW CAN WE ACCURATELY PREDICT THE FINANCIAL MARKETS?



Ensemble Approach for Financial Prediction

GRANT FUNDED AND SUPPORTED BY ANONYMOUS CONSORTIUM



Lasso를 이용하여 얻은 각 모델들의 중요 변수 히트맵

기존 연구들에서 언급된 각 변수간 관계를 재검증

HOW CAN WE CUT THE TIME A MERCEDES-BENZ SPENDS ON THE TEST BENCH?

Mercedes-Benz Greener Manufacturing



Data

메르세데스-벤츠 사의 차량 테스트 시간 및 관련 변수 데이터

모든 변수가 라벨링되어 있고, 변수명이 암명화되어 있음

Environment

Python 3.6+ / pandas, numpy, matplotlib, seaborn, scikit-learn, boruta, skmca, xgboost

Purpose

차량 테스트와 관련한 변수들을 이용해 테스트 시간을 예측

Methodology

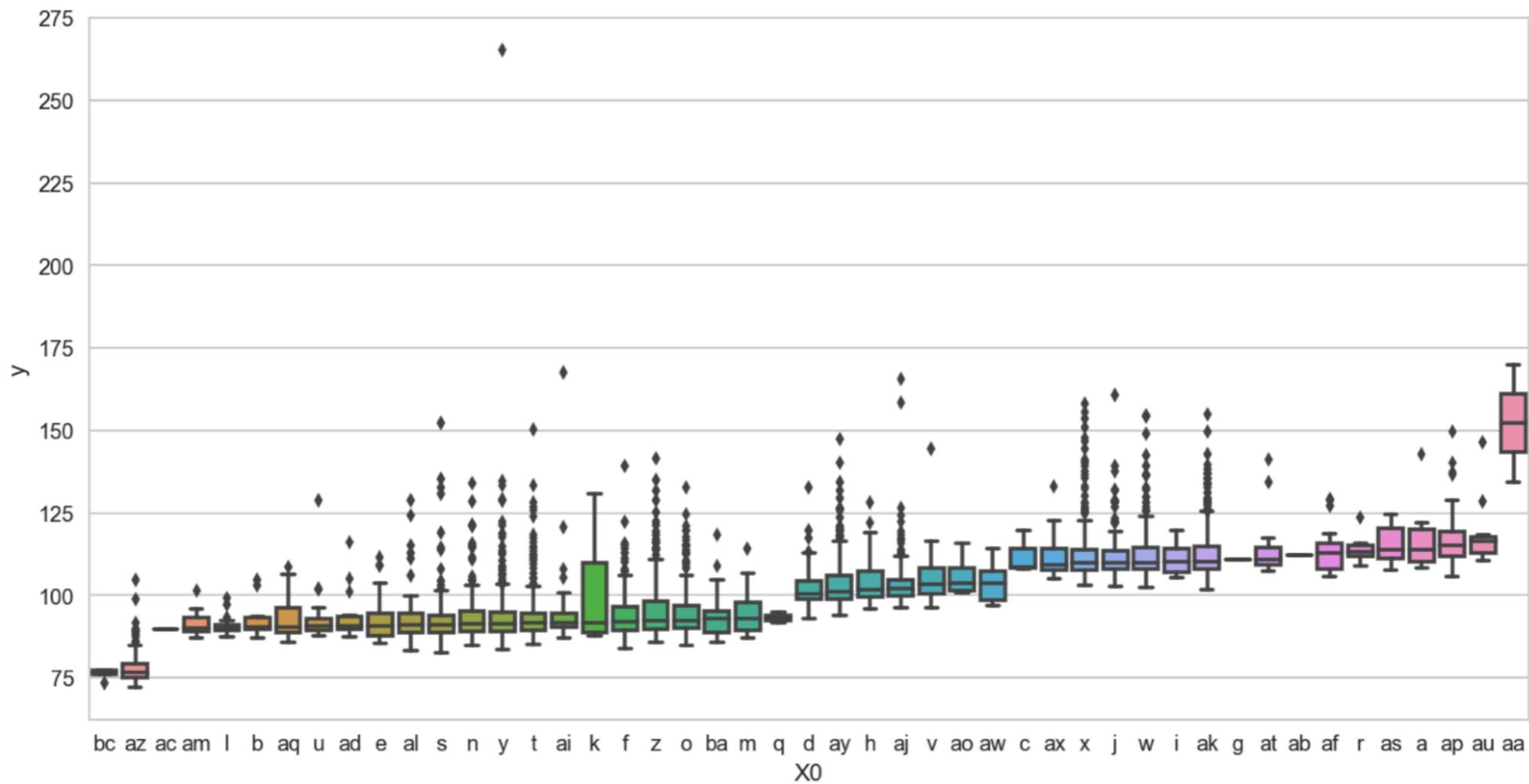
- 각 알고리즘에 대해 **5-fold Cross Validation** 으로 파라미터 튜닝 (그라디언트 부스팅, 능형 회귀, 랜덤 포레스트)
- 여섯 개의 **서로 다른 변수 집합**을 이용하여 18개의 단일 모델 구현 (Boruta 알고리즘을 이용해 변수 선택 / 차원축소법 / 타겟 인코딩 등)
- 단일모델을 스태킹하여 최종 모델 구현

What I Learned

1. 스몰 데이터에 대해서 예측을 진행할 때 **올바른 교차검증법을 사용**해서 파라미터를 튜닝할 것
(교차검증법을 이용할 경우, 각 Fold에 대한 모델의 중요 변수를 체크하여 변동성을 확인 / 특정 범주형 변수들이 잘 섞일 수 있게 수동적인 교차검증법을 이용)
2. 변수 집합을 구성할 때, 차원축소법 (PCA, MCA, ICA 등) 의 **결과 컴포넌트 개수 역시 하나의 파라미터로** 볼 것
3. 변수에 대한 **2-way 인터랙션, 3-way 인터랙션**이 정보가 없는 데이터에 대해서 좋은 전략이 될 수 있음
4. 특정 범주형 변수의 영향력이 크다면 **타겟 인코딩**으로 모델의 성능을 높일 수 있음

HOW CAN WE CUT THE TIME A MERCEDES-BENZ SPENDS ON THE TEST BENCH?

Mercedes-Benz Greener Manufacturing

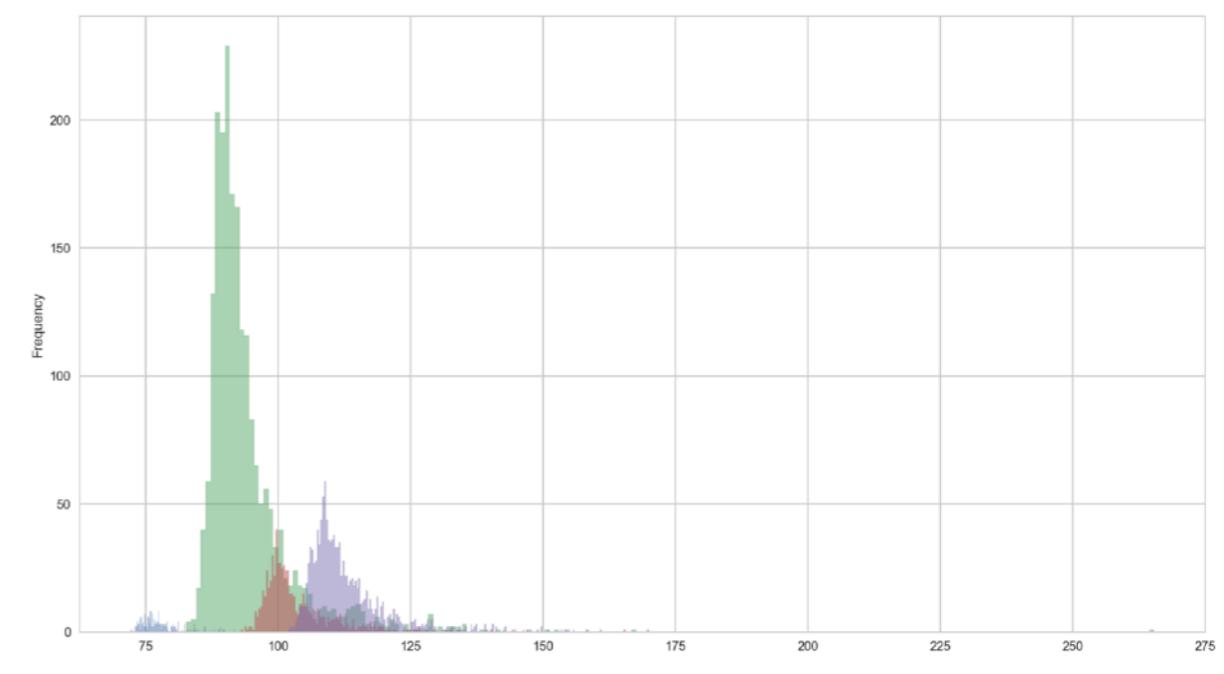
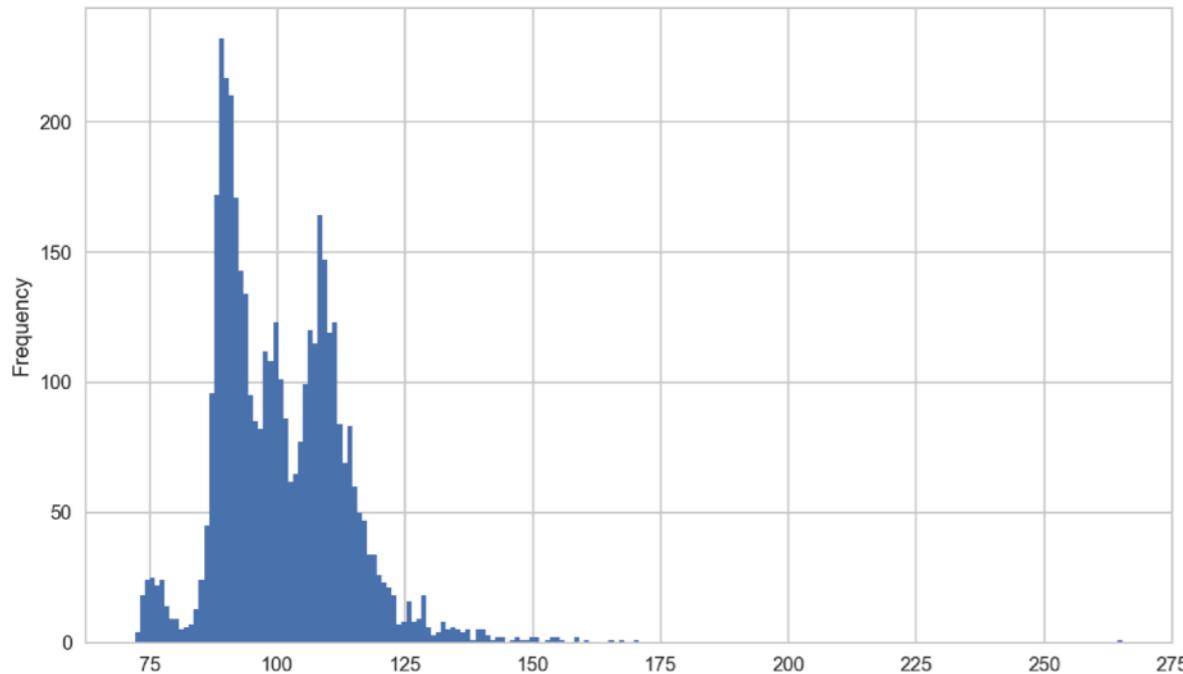


영향력이 큰 범주형 변수에 따른 타겟값의 분포

각 범주에 따라 값의 분포가 확실하게 나누기 때문에 타겟 인코딩을 이용해 예측하는 경우, 좋은 결과를 얻을 수 있음

HOW CAN WE CUT THE TIME A MERCEDES-BENZ SPENDS ON THE TEST BENCH?

Mercedes-Benz Greener Manufacturing



영향력이 큰 범주형 변수에 따른 타겟값의 분포

각 범주에 따라 값의 분포가 확실하게 나누기 때문에 타겟 인코딩을 이용해 예측하는 경우, 좋은 결과를 얻을 수 있음

RECOGNIZE THE VISITOR'S BEHAVIORAL PATTERN FROM IOT-BASED SMART EXHIBITION

Visitor Segmentation for Baik Nam June's Exhibition

WON THE BEST PAPER AWARD OF ICEC 2017

Data

DDP에서 열린 "백남준쇼"에 적용된 버튼 인터넷 데이터

각 작품마다 도슨트 서비스를 위한 버튼이 설치되어 있으며
버튼이 눌릴 때마다 트랜잭션 데이터가 실시간으로 적재됨

Environment

R 3.4.2+ / RStudio 1.1.383

Purpose

버튼을 누른 이력을 이용하여 서비스 사용 **방문객의 이탈률 분석**
방문객의 패턴을 군집 분석하여 추후 타겟 마케팅에 이용

Methodology

- 주어진 데이터를 정제하여 **User-item matrix**를 구성
- **t-SNE**를 이용하여 데이터를 변환
(시각화를 통해 적당한 파라미터 설정)
- k-평균 군집 분석을 통해 방문객 그룹 생성
(Elbow method로 군집 개수 설정)

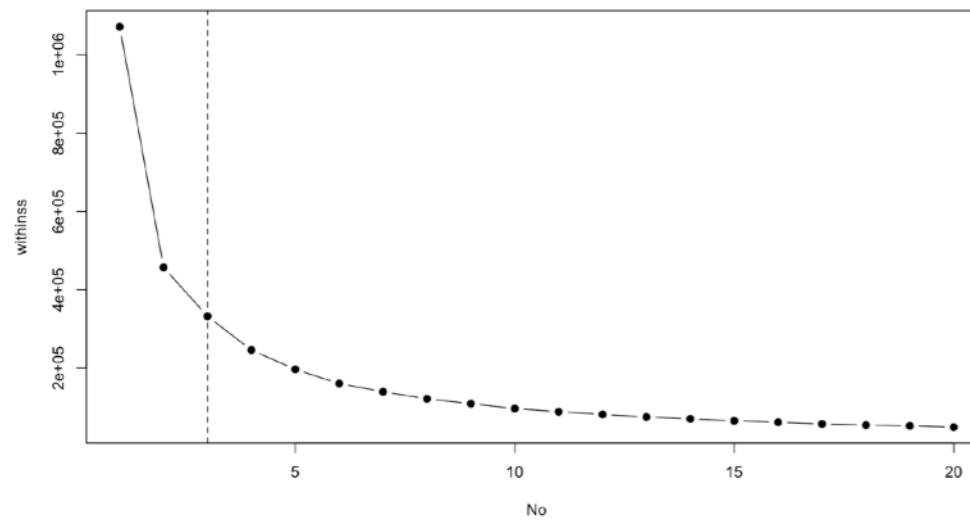
Issues and Comments

1. mongoDB에 적재된 데이터를 **JSON 형태로 추출**하였으나 R에서 자동 파싱이 되지 않아 **정규식으로 수동 파싱**
2. 데이터 정제 과정에서 **수많은 이상치가 존재**했고, 다음의 기준으로 제거함
(전시회 개장 시간 이외 데이터 제거 / 같은 사용자가 30초 이내에 다시 버튼을 누른 경우 제거)
3. 전체적으로 관람 시간이 흐를 수록 **버튼의 사용률이 현저히 줄어듦**
4. 인구통계학적 데이터의 부재로 보다 자세한 분석은 어려움
(가용한 외부 데이터를 추가적으로 결합할 필요성이 있음)

RECOGNIZE THE VISITOR'S BEHAVIORAL PATTERN FROM IOT-BASED SMART EXHIBITION

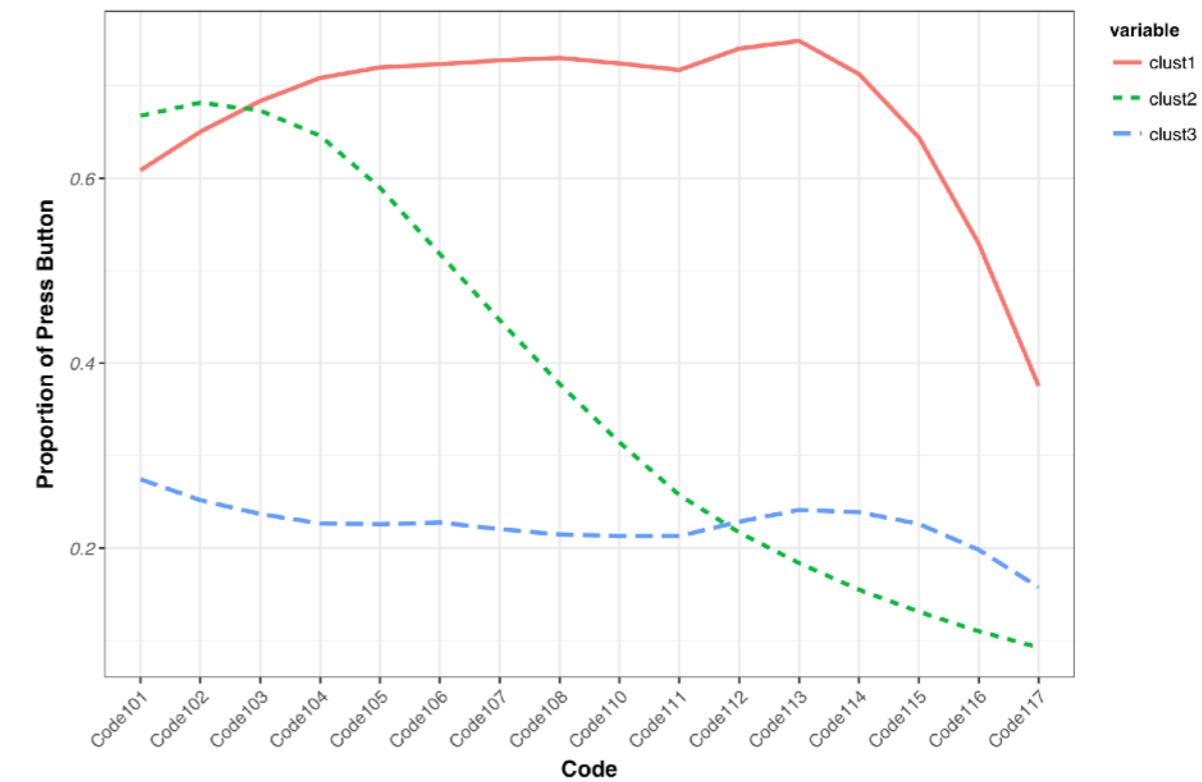
Visitor Segmentation for Baik Nam June's Exhibition

WON THE BEST PAPER AWARD OF ICEC 2017



Elbow Method

(withinss 값의 변화가 완만해지는 지점의 값을 선택)



방문객들의 특성

전체적으로 후반부 전시물에 대해서는 버튼 사용률이 낮음

세 번째 그룹(사용률이 낮은 그룹)보다 두 번째 그룹(초반에 관심을 갖는 그룹)에 대한
타겟 마케팅이 비용적으로 효율적일 것으로 예상

WHICH CLUB HAS AN ENORMOUS INFLUENCE ON THE FOOTBALL TRANSFER MARKET



Network Analysis of Football Transfer Market

Data

transfermarkt.de의 2016년 여름 이적시장 데이터

크롤링하려 했으나 테이블 구문의 데이터가 과도하게 중첩되어 있어
별도의 스크래핑을 통해서 데이터 구축

Purpose

유럽 축구 이적시장에서 각 리그의 비교 분석 및 특징 파악
실제 분석 결과를 이용해 각 리그의 현 상황을 설명하기 위함
추가적으로 한 눈에 확인할 수 있는 시각화 결과물을 구현

Environment

R 3.4.2+, RStudio 1.1.383 / igraph

How to Construct the Network

- 선수 이적료가 각 네트워크 연결의 가중치가 되며,
선수의 이적은 방향성이 있는 네트워크로 구성
- 소속이 없는 선수의 경우 이적료가 없기 때문에 시장가치로 대체
- 이적료가 비공개인 경우 시장가치로 대체
- 가중치의 단위는 백만 파운드 당 1로 기준

Comments

1. In-node degree와 Out-node degree의 분포는 **각각 1.5, 1.8의 power-law**를 따름
2. 매개 중심성보다는 **위계 중심성**을 통하여 네트워크를 분석하는 것이 적절함
3. 특정 팀들의 위계 중심성이 **매우 높게** 나타나는데, 이는 현재 유럽 축구 이적시장에서 **몇몇 팀들이 경제적 우위를 점하고 있음**을 반증함
4. 전반적으로 분석 결과가 현 리그의 상황을 잘 반영하고 있음

WHICH CLUB HAS AN ENORMOUS INFLUENCE ON THE FOOTBALL TRANSFER MARKET



Network Analysis of Football Transfer Market

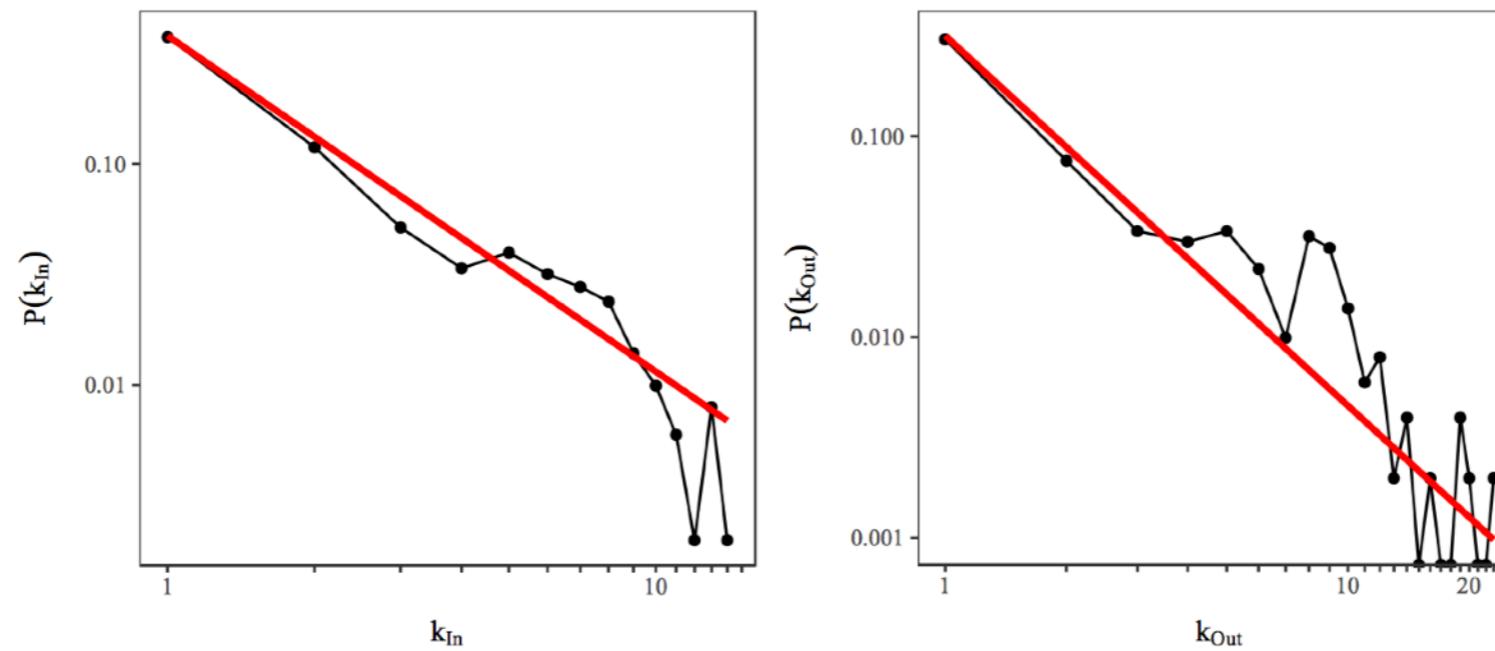


Figure 1: (Left) In-node Degree Distribution of Whole Network. Degree Exponent $\gamma_{in} \simeq 1.5$; (Right): Out-node Degree Distribution of whole Network. Degree Exponent $\gamma_{in} \simeq 1.8$.

Rank	Team	Degree	League	Team	Betweenness	League	Team	Eigenvalue	League
1	Sampdoria	0.062	Italy	Pescara	0.12	Italy	Manchester Utd.	1	England
2	Juventus	0.06	Italy	Chievo Verona	0.096	Italy	Manchester City	0.506	England
3	Torino	0.056	Italy	Sampdoria	0.086	Italy	Juventus	0.277	Italy
4	Udinese	0.056	Italy	Crotone	0.081	Italy	Crystal Palace	0.272	England
5	AS Roma	0.052	Italy	Sporting Gijon	0.071	Spain	Liverpool	0.238	England
6	Pescara	0.05	Italy	Genoa	0.071	Italy	Everton	0.214	England
7	Genoa	0.048	Italy	Ternana	0.061	Misc	Real Madrid	0.207	Spain
8	Chievo Verona	0.046	Italy	Real Zaragoza	0.059	Misc	AS Roma	0.206	Italy
9	Bologna	0.044	Italy	FC Augsburg	0.059	Germany	Spurs	0.182	England
10	Real Betis	0.044	Spain	Dep. La Coruna	0.056	Spain	Sunderland	0.171	England

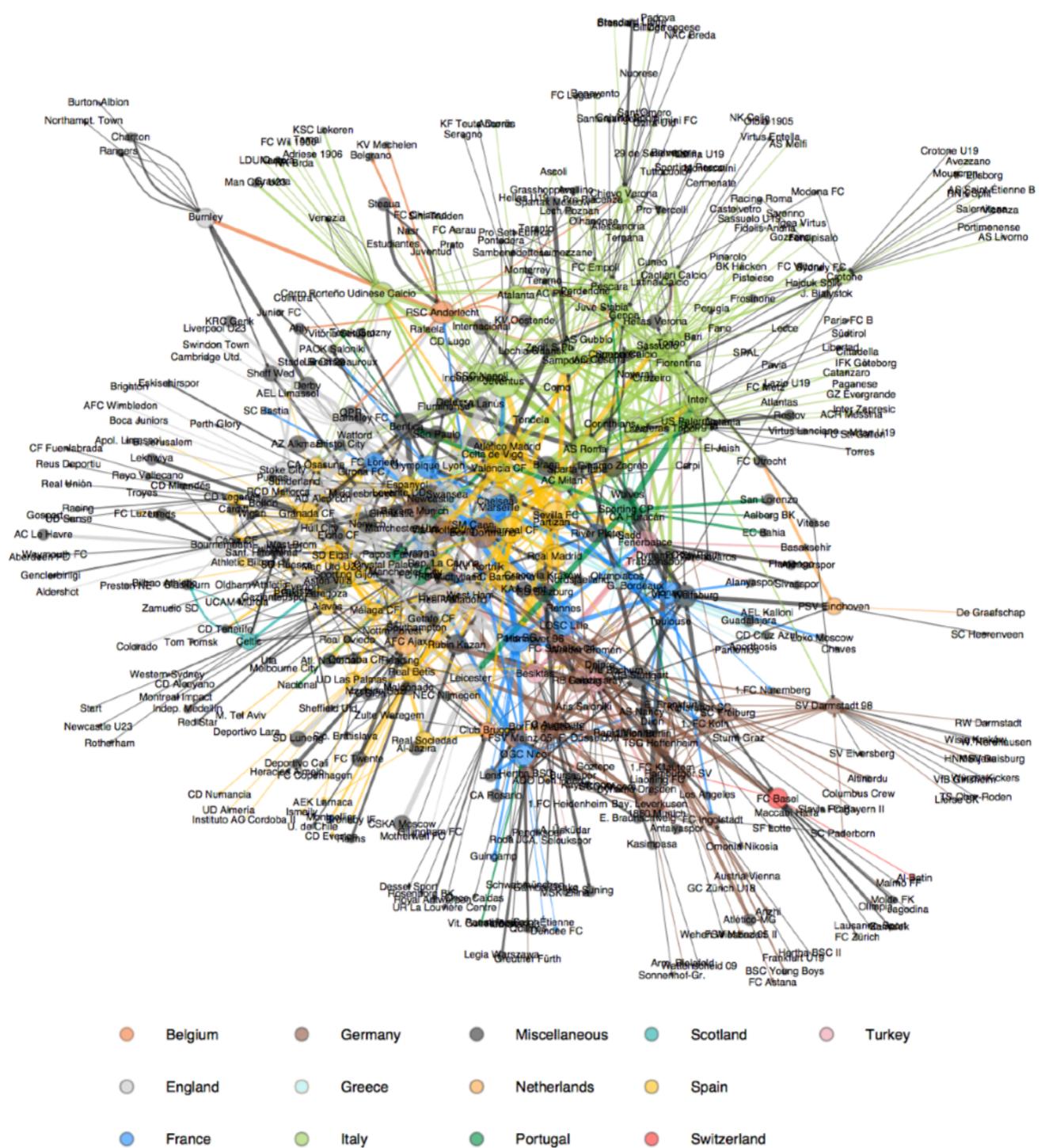
Table 2: Top 10 of Centralities of Whole Network

네트워크에서 In-node, Out-node degree와 각 리그 네트워크의 중심성값

WHICH CLUB HAS AN ENORMOUS INFLUENCE ON THE FOOTBALL TRANSFER MARKET



Network Analysis of Football Transfer Market

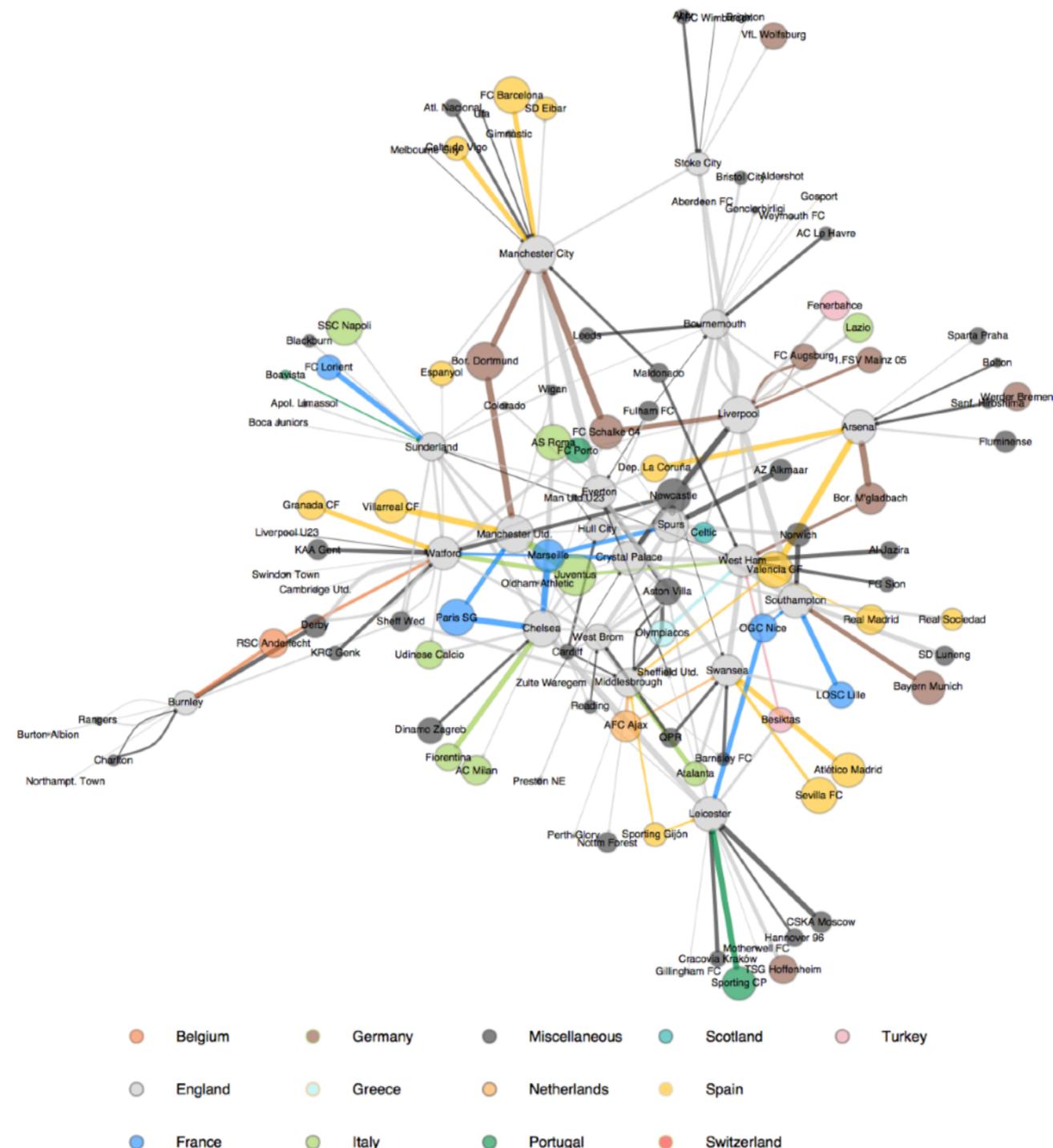


전체 이적시장 네트워크

WHICH CLUB HAS AN ENORMOUS INFLUENCE ON THE FOOTBALL TRANSFER MARKET



Network Analysis of Football Transfer Market



잉글리쉬 프리미어리그 이적시장 네트워크

THANK YOU

:)

Master of Engineering
Dept. Social Network Sciences
Kyung Here University

Phone : 010-3806-9224
E-Mail : otzslayer@gmail.com
Github : <http://github.com/otzslayer>
Blog : <http://otzslayer.github.io>