Eric Hwang’s Portfolio

데이터 사이언스

목록

hshun123@naver.com

- Name: 황승현
- Email: hshun123@naver.com
- Github: https://github.com/hshun123

Technical Skills

- Java/Spring Framework
- Restful API Design
- Python
- C++/C
- Ocaml
- MySQL
- MongoDB
- Kotlin

Project1 해외축구팀 분석
데이터소개및분석동기

컬럼 명	컬럼 의미
ID	고유의 번호
Name	이름
Age	나이
Overall	현재 능력치
Potential	잠재 능력치
Club	소속 팀
Value	예상 이적료 (유로)
Wage	주급 (유로)
Preferred Foot	잘 사용하는 발
Weak Foot	잘 사용하지 않는 발
Skill Moves	개인기
Position	포지션
Jersey Number	등번호
Joined	소속 팀 입단 날짜
Contract Valid Until	계약 기간
Height	키 (피트)
Weight	몸무게 (파운드)
LS ~ RB	포지션 별 능력치
Crossing ~ GKReflexes	세부 능력치
Release Clause	바이아웃

데이터 소개

축구 선수 기본 정보 데이터

(1) 선수 기본 정보

(2) 클럽 팀 / 포지션 / 포지션 별 능력치 / 계약 정보 포함

분석 설계

목표 : 해외 축구 선수 팀 맨체스터 유나이티드 선수 영입 제안

(1) 맨체스터 유나이티드 팀 부족한 포지션 분석

(2) 팀의 영입방침을 살펴보고 현재 이적 시장에 있는 인원 들의 능력치 비교 후 우수한 인원 영입 제안

프로젝트 개발 환경

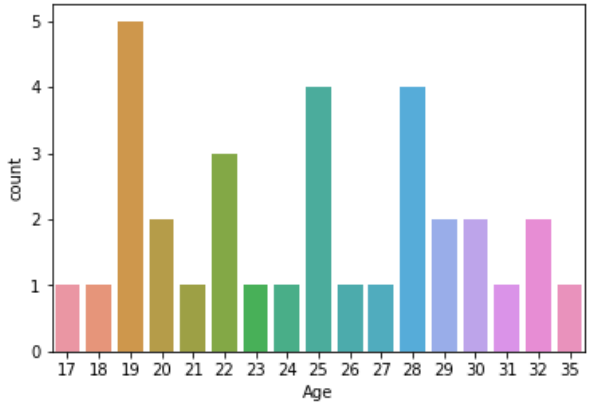
언어 : Python

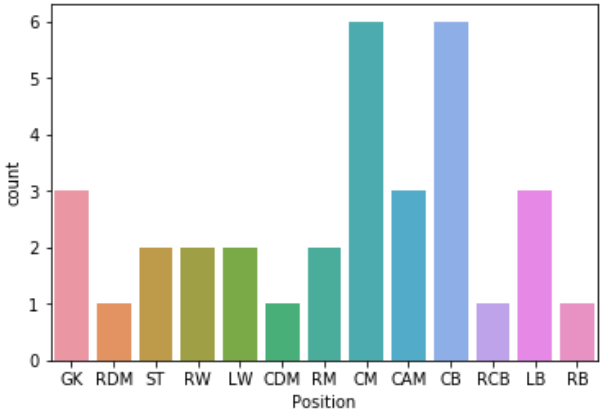
데이터 출처:
FIFA2018

활용 도메인 지식:
유럽 축구 데이터

분석 동기 :
4차 산업 시대에 맞게 축구 선수 영입시에도 데이터 분석이 필요함을 증명하기 위하여

Project1 해외축구팀 분석
데이터 특징 및 제안



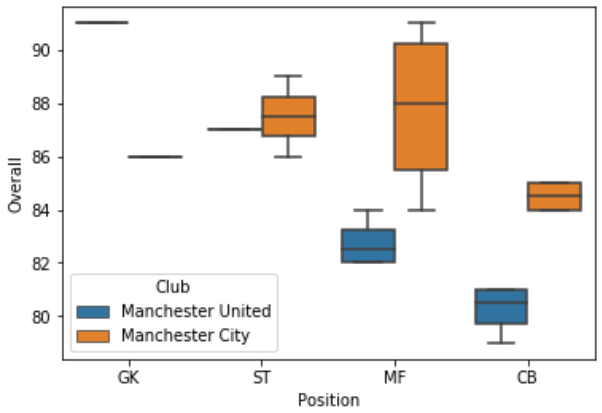


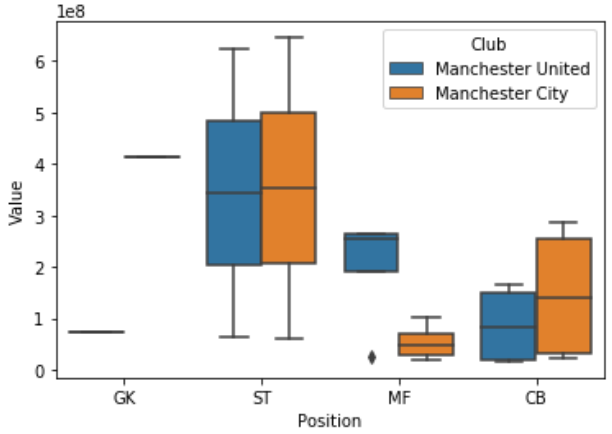
Countplot

맨유의 기본 정보를 확인

나이 및 포지션의 분포정도를

알 수 있음.





Boxplot

맨시티와 비교하여 각 포지션

능력치 및 몸 값을 비교

특징

라이벌 구단과 비교 시 이적료는 비슷하지만 능력치 부분에서 많은 차이가 보임

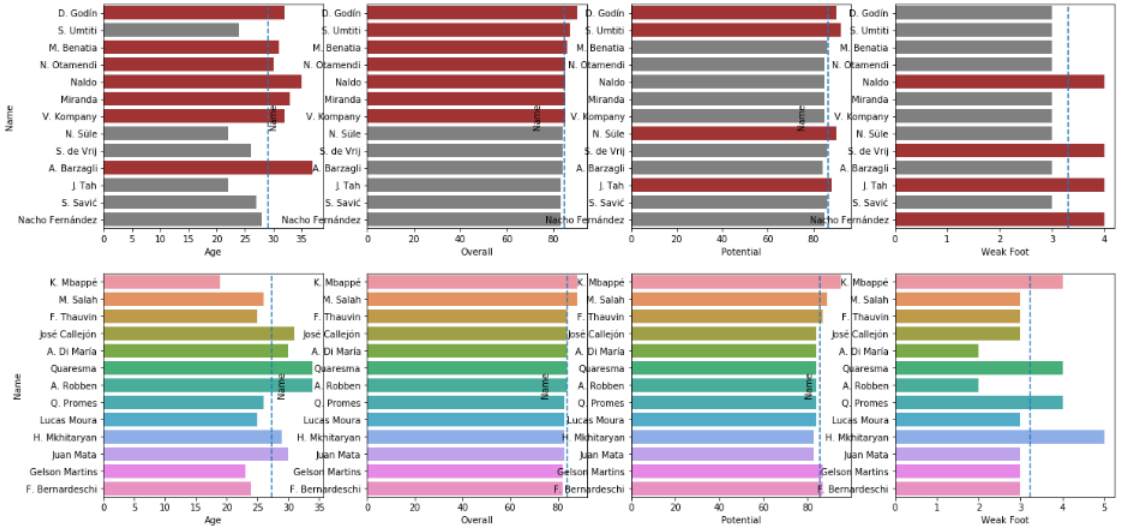
제안

라이벌 팀과 비교해서 특히 부족한 MF/CB를 영입

Project1 해외축구팀 분석
분석결과

	Name	Overall	Potential	Age	Joined	Point
327	E. Bailly	81.0	87.0	24	Jul 1, 2016	10.375000
377	C. Smalling	81.0	82.0	28	Jul 1, 2010	8.714286
454	L. Shaw	80.0	85.0	22	Jun 27, 2014	11.136364
584	V. Lindelöf	79.0	85.0	23	Jul 1, 2017	10.565217

	Name	Overall	Potential	Age	Joined	Point
132	N. Matić	84.0	84.0	29	Jul 31, 2017	8.689655
211	Juan Mata	83.0	83.0	30	Jan 25, 2014	8.300000
250	Fred	82.0	84.0	25	Jun 21, 2018	9.920000
254	J. Lingard	82.0	83.0	25	Jul 1, 2010	9.880000



결론 : 영입 방침

맨유의 영입 방침을 고려하여

잔류 Point라는 점수를 생성

잔류 Point가 낮은 선수 2명을 방출

이적시장 선수 중

잔류 Point가 높은 선수를

영입 제안

영입 방침에 따른 잔류 Poitnt :

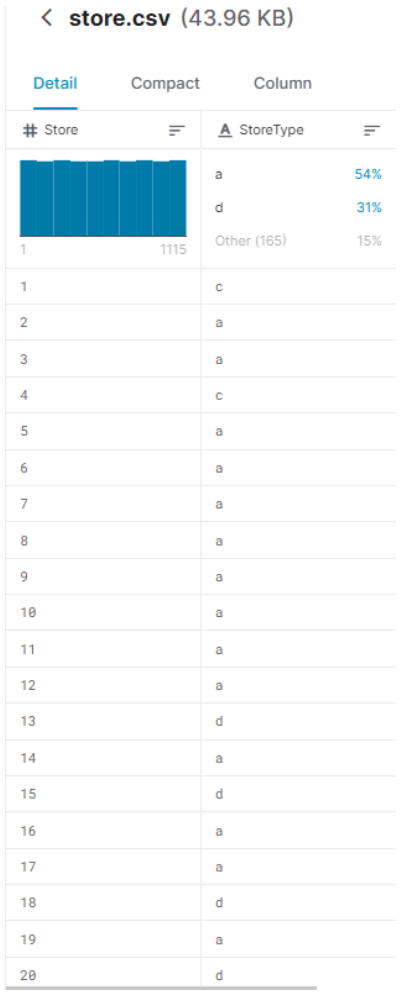
(Overall * 2 + Potential) / Age

잔류 Point 낮은 C.Smalling / Juan Mata 방출

이적시장 상위 13명의 Age, Overall, Potential, Weak Foot 비교 하여 영입 우선순위 판가름 후 영입

황승현

Project2 상점 매출 분석
데이터소개 및분석동기



데이터 소개

Rosmann 상점의 매출 데이터

(1) 상점별 상세정보(상점 위치, 기후, 경쟁 상점과의 거리 등)

(2) 날짜

분석 설계

(1) 베이스라인 모델링

(2) 피처 엔지니어링을 통해 여러 변수들 생성하기

(3) 궁극적인 매출 증대 방안 고려

프로젝트 개발 환경

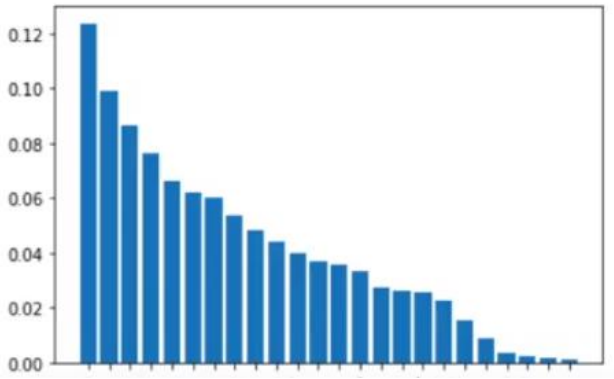
언어 : Python

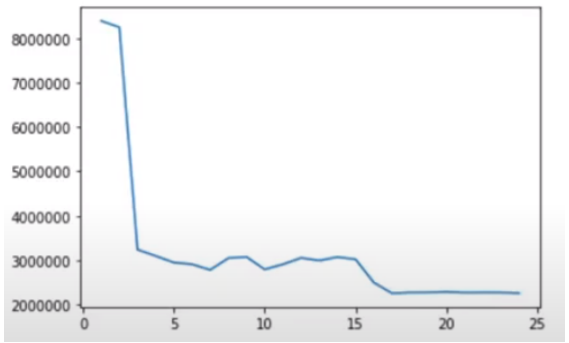
데이터 출처:
Kaggle / Rossmann Sales 데이터

활용 도메인 지식:
커머스

분석 동기 :
매출을 결정짓는 요인을 분석하고 그에 따른 매출 증대 계획 제안

Project2 상점 매출 분석
모델링결과





Submission and Description	Private Score	Public Score	Use for Final Score
submission3.csv 2 months ago by Geunil-Song n_estimators : 600	2003.63463	2003.63463	☐
submission2.csv 2 months ago by Geunil-Song add submission details	1801.85035	1801.85035	☐
submission1.csv 2 months ago by Geunil-Song xgb 'Promo','weekday'	2784.15955	2784.15955	☐
submission.csv 2 months ago by Geunil-Song xgboost 활용 'Promo','SchoolHoliday','StateHoliday_0','StateHoliday_a','StateHoliday_b','Sta	2903.38819	2903.38819	☐

기본 모델링+feature engineering + 변수선택

Q : 매출 예측 모델 제작 과정과 그 결과

A:

1. 보조 데이터의 feature engineering과 변수 중요도 개념을 활용한 변수선택으로 예측 rmse는 1800까지 떨어짐
2. 매출에 가장 중요한 요소는 프로모션 여부임
3. 경쟁업체와의 거리는 생각보다 덜 중요한 요소로 판명남

Project3 대출 상환 분석
데이터소개 및분석동기

col_name	설명
SK_ID_CURR	유니크한 아이디
TARGET	연체 혹은 문제가 생긴 경우)
CODE_GENDER	성별(0: 여성, 1: 남성)
FLAG_OWN_CAR	차 보유 여부(0: 없음, 1: 있음)
FLAG_OWN_REALTY	집 보유 여부(0: 없음, 1: 있음)
CNT_CHILDREN	자녀 수
AMT_INCOME_TOTAL	수입
AMT_CREDIT	대출금액
AMT_ANNUITY	1달마다 갚아야 하는 금액
NAME_TYPE_SUITE	신청을 할 때 누가 동행했는지
NAME_INCOME_TYPE	직업 종류
NAME_EDUCATION_TYPE	학위
NAME_HOUSING_TYPE	주거 상황
REGION_POPULATION_RELATIVE	지역의 인구
DAYS_BIRTH	나이
DAYS_EMPLOYED	얼마나 일했는지(365243는 결측치)
DAYS_ID_PUBLISH	신청한 ID 문서를 변경한 날짜
OWN_CAR_AGE	보유한 차의 나이
CNT_FAM_MEMBERS	가족 수
HOUR_APPR_PROCESS_START	언제 대출신청을 했는지 시간
ORGANIZATION_TYPE	일하는 조직의 종류
EXT_SOURCE_1	외부 데이터1로부터 신용점수
EXT_SOURCE_2	외부 데이터2로부터 신용점수
EXT_SOURCE_3	외부 데이터3로부터 신용점수
DAYS_LAST_PHONE_CHANGE	마지막 핸드폰을 바꾼 시기
AMT_REQ_CREDIT_BUREAU_YEAR	지난 한 신용정보를 조회한 개수

데이터 소개

Home Credit 기업 내부 데이터

(1) 채무자의 인적 정보 (나이, 성별, 사는 지역 등)

(2) 대출에 대한 상세 정보 (대출금액, 대출종류, 기간 등)

(3) 채무자가 성공적으로 대출 했는지에 대한 여부

분석 설계

(1) 모델링

(2) 모델링에 따른 피처들의 영향력 알아보기

(3) 영향을 많이 주는 5개의 변수와 대출금 상환 여부와의 관계 보기

프로젝트 개발 환경

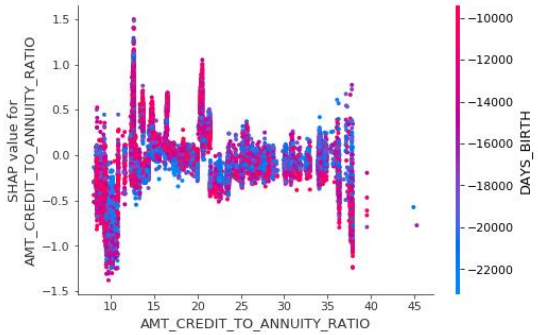
언어 : Python

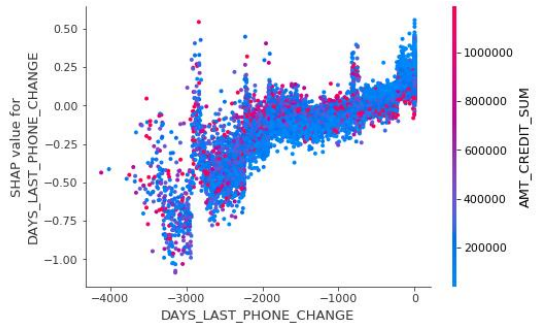
데이터 출처:
Kaggle / Home Credit Default Risk 데이터

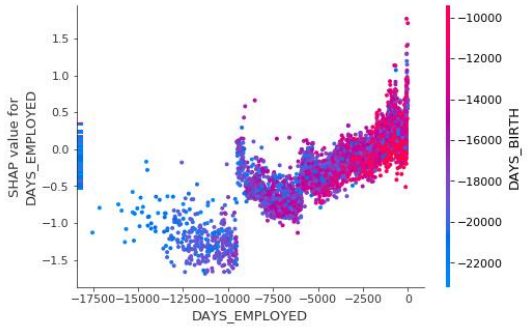
활용 도메인 지식:
금융 데이터

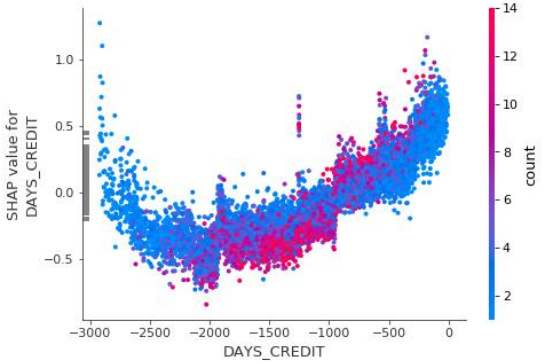
분석 동기 :
대출 상환 여부를 결정짓는 요인을 분석하고 그에 따른 대출 플랜 제안

Project3 대출 상환 분석
모델링









AMT_CREDIT_TO_ANNUITY_RATIO

총 대출 금액이 한달마다 갚아야 하는 금액의 12배~20배까지는 때는 비교적 상환을 못함
반면, 35배 이상부터는 상환을 잘 함.

DAYS_EMPLOYED

취업한지 오래될 수록 대출을 상환할 확률 상승.
특이점 : 대출일 기준 9000일보다 이전에 취업을 했을 때 대출상환 능력이 급격히 상승

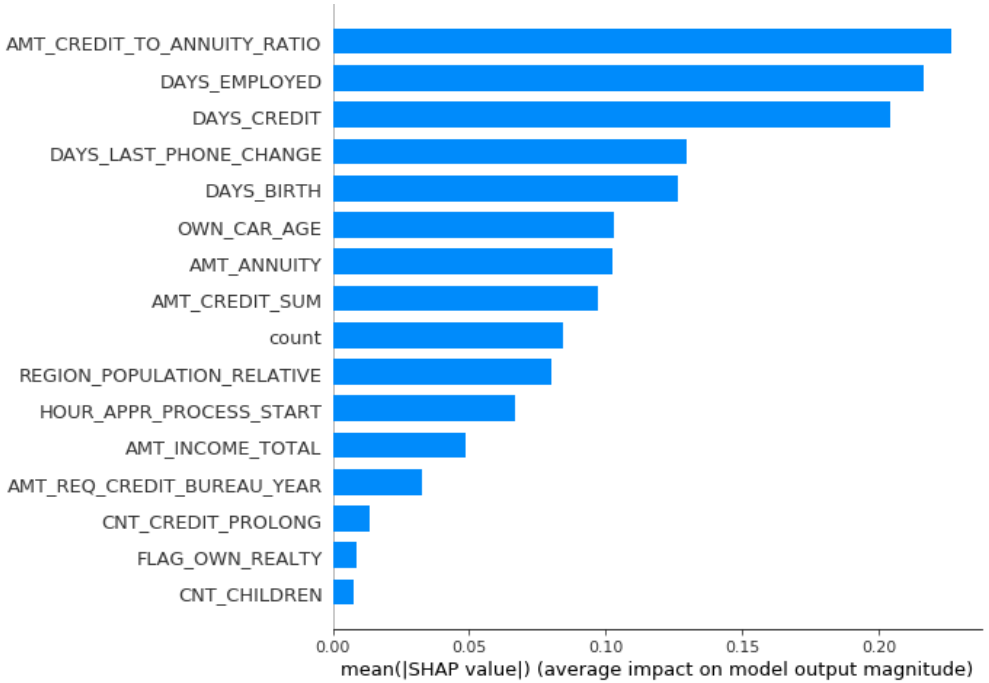
DAYS_CREDIT

DAYS_CREDIT 변수는 이전에 대출을 진행했을 시에 이전과 현재 대출일간의 차이의 평균DAYS_CREDIT은 -3000일부터 -2000일까지는 대출을 상환할 확률이 상승하며, 이후부턴 하락하는 비선형성을 보임.

DAYS_LAST_PHONE_CHANGE

핸드폰을 오래 전에 바꾸었을 수록 대출을 상환할 가능성 상승

Project3 대출 상환 분석
분석결과



Shap Value를 통한 상환 여부 중요도 표시

Q:대출상환여부영향요소

A:상환여부영향상위요소5

1. AMT_CREDIT_TO_ANNUITY_RATIO :
   대출 금액 대비 월별 상환금액의 비율
2. DAYS_EMPLOYED :
   취업한 시기
3. DAYS_CREDIT :
   다른 대출을 받은 시기
4. DAYS_LAST_PHONE_CHANGE :
   핸드폰을 바꾼 시기
5. DAYS_BIRTH : 태어난 시기

프로젝트 기술 스택

사용 모델링 :

Xgboost

모델링 사용 이유 :

1. Treeshap 밸류를 활용하기 위해 tree형 모델 선택

2. tree형 모델 중 속도가 빠르고, 평균 높은 성능율 유지하는 xgboost선택