최준영 portfolio

목차

자기소개

- 프로필
- 약력

프로젝트

- 해외축구 데이터 분석(맨체스터 유나이티드 팀 성적 하락 이유)
- 로스만 데이터 활용 로스만 상점 매출 예측
- Home Credit 데이터 활용 대출의 상환여부에 영향을 주는 요인 분석

자기소개

프로필

최준영
1993.01.21

연락처
- 이메일 cjy1705@naver.com
- Phone 010-6287-9951
- Github https://github.com/choijy1705

사용가능 기술
- Java
- JSP
- HTML/CSS/JavaScript
- Python

자기소개

약력

학력

- 동인고등학교(2008.03 ~ 2011.02)
- 울산대학교(2011.03 ~ 2018.02 - 기계공학과

경력

- 삼성전자 설비 엔지니어(2018.06 ~ 2019.07)

외부교육

- 머신러닝 플랫폼을 활용한 빅데이터 개발 및 분석(2019.09 ~ 2020.03)
  - Java, Javascript, Python, R 등 백엔드와 빅데이터 분석 및 머신러닝 기술 학습

프로젝트

- Google Playstore Install 수 예측
- 맨체스터 유나이티드가 부진에서 탈출하기 위한 방법
- 로스만 상점 매출 예측
- HomeCredit 데이터활용 대출 상환 가능성 분석

프로젝트

해외축구 데이터 분석

데이터 소개 및 분석 동기

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

데이터소개

축구선수의 기본정보 데이터

- 선수 이름 나이 등의 기본정보
- 클럽이름/포지션/주발 등의 축구관련 데이터

분석 동기

- 퍼거슨 감독 시절 최고의 팀 중 하나였던 맨유가 최근까지도 이전의 모습을 보여주지 못하는 이유와 영입을 통한 해결점을 찾아보기 위해

분석 설계

- 라이벌 팀 맨체스터 시티와 비교하여 부족한 포지션이 무엇인지 분석
- 다른 팀의 선수들 중 부족한 포지션에 해당하는 선수 중 적합한 대체 선수는 누가 있을 지 확인
- 2018년 데이터를 통해 분석한 결과와 2020년 현재 선수단과의 비교

사용언어

- Python

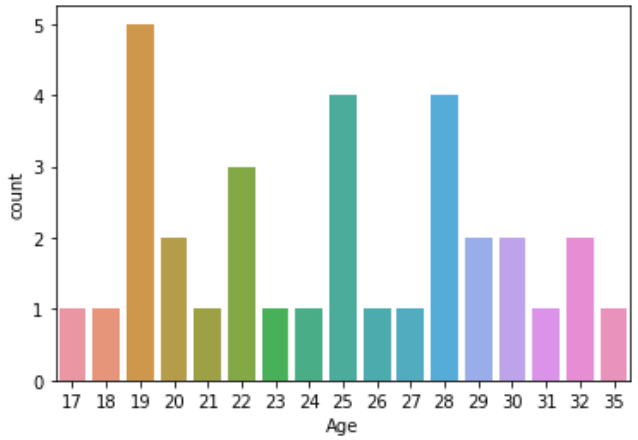
데이터 출처

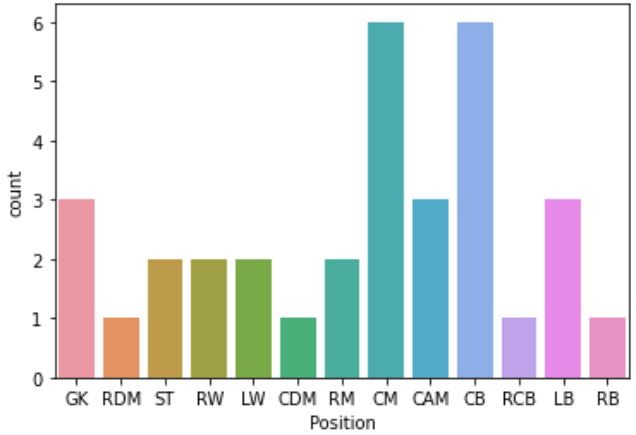
- Kaggle FIFA2018

프로젝트

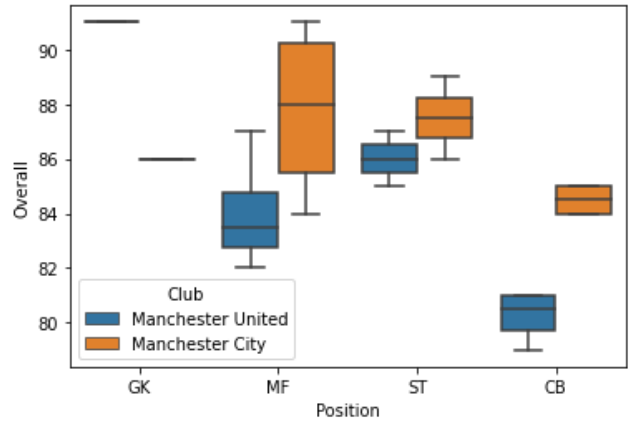
해외축구 데이터 분석

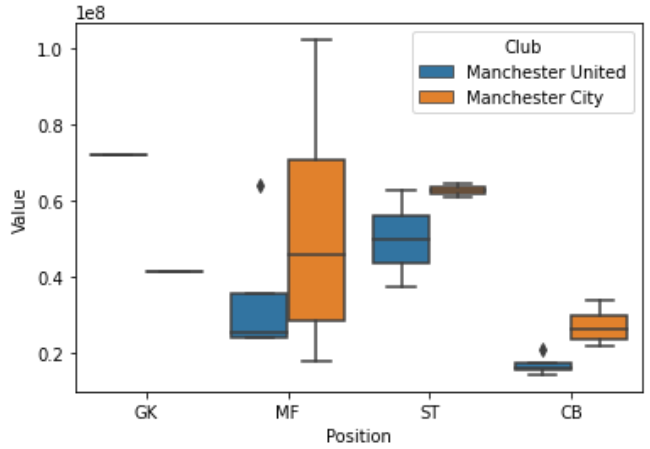
데이터 특징 및 부족한 포지션 확인





Countplot
맨유 선수들의 구성 및 선수들의 기본정보 파악
포지션 별로 선수 수 등을 파악





Boxplot
맨시티 선수단과 Position별 overall과 급여 등을 비교하여 부족한 부분 파악

선수단 특징
이적료와 선수의 실제 실력에서 라이벌구단과 비교 시 많은 부분이 부족한 것을 알 수 있음

해결방안
그 중에서도 RM, CB 포지션의 보강이 가장 시급해 보이는 것으로 생각 되어 짐

프로젝트

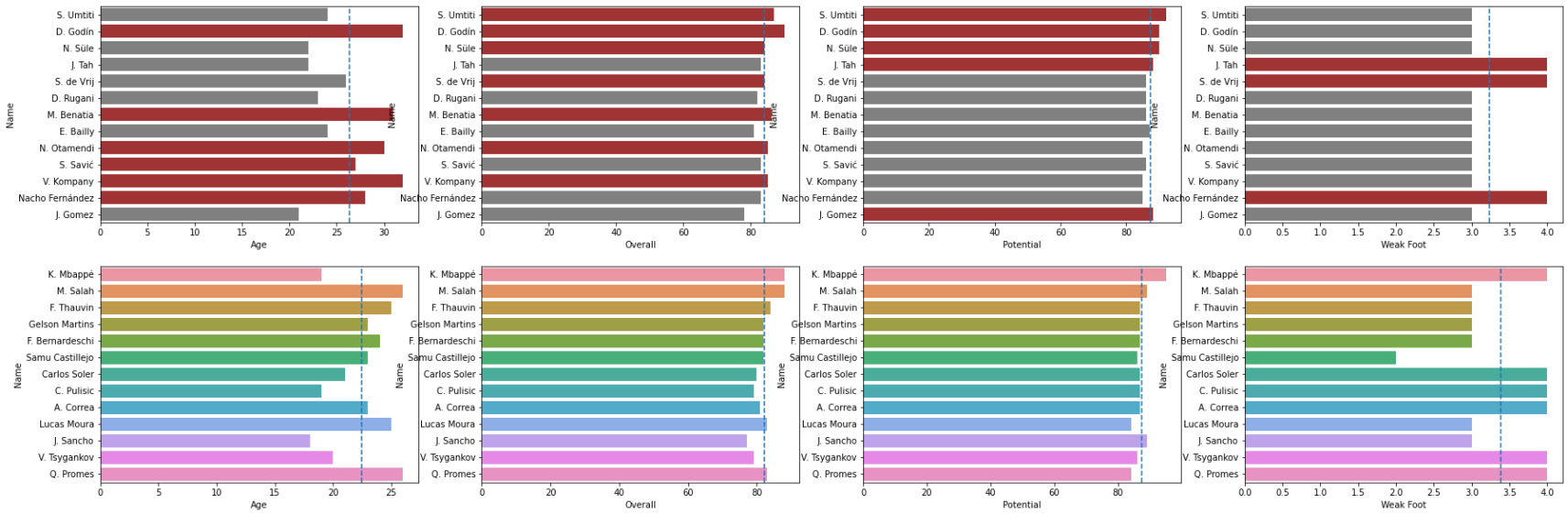
해외축구 데이터 분석

대체선수 확인 및 결론

	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage
25	231747	K. Mbappé	19	France	88.0	95.0	Paris Saint-Germain	€81M	€100K
26	209331	M. Salah	26	Egypt	88.0	89.0	Liverpool	€69.5M	€255K
122	204970	F. Thauvin	25	France	84.0	87.0	Olympique de Marseille	€39M	€72K

	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage
25	231747	K. Mbappé	19	France	88.0	95.0	Paris Saint-Germain	€81M	€100K
26	209331	M. Salah	26	Egypt	88.0	89.0	Liverpool	€69.5M	€255K
122	204970	F. Thauvin	25	France	84.0	87.0	Olympique de Marseille	€39M	€72K

포지션 별 가장 Point가 높은 선수 Top3



영입대상 선정
선수들의 특성을 종합하여 포인트화 하여 점수계산 후 정렬

Point
(Overall * 2) + Potential
Age

Point가 가장 낮은 C.Smalling, Juan Mata 방출

Point가 가장 높은
S.Umtiti, K.Mbape 영입제안

결론
단순히 선수의 능력과 급여 등 수치화만 할 수 있는 것을 고려하였고 구단의 재정과 국가 간의 상황 등은 고려되지 않아서 현재 영입된 상황과 많은 차이가 있음을 알 수 있음

프로젝트

로스만 상점 매출 예측

데이터 소개 및 동기

	id	Store	Date	Sales	Promo	StateHoliday	SchoolHoliday
0	14929	85	2015-05-01	11360	1	a	0
1	14930	512	2015-05-01	10534	1	a	0
2	14931	1097	2015-05-01	17039	1	a	0
3	14932	1	2015-04-30	6228	1	0	0
4	14933	9	2015-04-30	9717	1	0	0

변수 설명

- id : 상점의 id
- Store : 상점의 고유번호
- Date : 상점의 오픈 날짜
- Sales : 상품의 판매량 (Target 값)
- Promo : 프로모션의 진행 여부
- Stateholiday : 주의 휴무여부
- SchoolHoliday : 학교의 휴무여부

데이터 소개

로스만 상점들의 데이터

Train.csv, test.csv, store.csv로 구성된 데이터 셋 들로서 train 과 store데이터를 통하여 test 데이터 셋의 판매량(Sales)를 예측 해볼 수 있는 데이터

분석동기

실제 캐글에서 대회로 진행되었던 데이터 셋으로 이를 통하여 캐글 대회 형식에 익숙 해져보고자 진행

분석 설계

- 베이스라인 모델링
- 피쳐 엔지니어링 진행 후 새로운 변수 생성
- 매출을 증대 시키는 요인이 무엇인지 파악

사용 언어

Python

데이터 출처

Kaggle Rossmann Sales 데이터

평가 방식

RMSE

$$\sqrt{\frac{1}{N}\sum(yt - ypr)^2}$$

프로젝트

로스만 상점 매출 예측

베이스라인 모델링

```
xgb = XGBRegressor( n_estimators= 300 , learning_rate=0.1 , random_state=2020 )
xgb.fit(train[['Promo','SchoolHoliday','StateHoliday_0','StateHoliday_a','StateHoliday_b','StateHoliday_c','weekday','year','month']],
        train['Sales'])
```

XGBoost 를 활용한 베이스라인 모델 구현

submission.csv 3052.14204

a day ago by junyoungchoi

'Promo','SchoolHoliday','StateHoliday_0','StateHoliday_a','StateHoliday_b','StateHoliday_c','weekday

XGB

XGBoost 를 이용한 모델링 구현 이유

일반적으로 다른 머신 러닝보다 뛰어난
예측 성능과 빠른 속도이기 때문에 선택

베이스라인 모델을 통한
RMSE 결과 3052.14204

프로젝트

로스만 상점 매출 예측

피쳐 엔지니어링 진행

	id	Store	Date	Sales	Promo	SchoolHoliday	StateHoliday_0	StateHoliday_a	StateHoliday_b	StateHoliday_c	month	year	weekday
0	14929	85	2015-05-01	11360	1	0	0	1	0	0	5	2015	4
1	14930	512	2015-05-01	10534	1	0	0	1	0	0	5	2015	4
2	14931	1097	2015-05-01	17039	1	0	0	1	0	0	5	2015	4
3	14932	1	2015-04-30	6228	1	0	1	0	0	0	4	2015	3
4	14933	9	2015-04-30	9717	1	0	1	0	0	0	4	2015	3

날짜 데이터를 이용하여 년,월,주 컬럼 생성

	Store	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval
0	1	c	a	1270.0	9.0	2008.0	0	NaN	NaN	NaN
1	2	a	a	570.0	11.0	2007.0	1	13.0	2010.0	Jan,Apr,Jul,Oct
2	3	a	a	14130.0	12.0	2006.0	1	14.0	2011.0	Jan,Apr,Jul,Oct
3	4	c	c	620.0	9.0	2009.0	0	NaN	NaN	NaN
4	5	a	a	29910.0	4.0	2015.0	0	NaN	NaN	NaN

Store데이터와 train merge 후 상점의 프로모션 진행기간과 주변 상점의 오픈 일자 등 새로운 변수를 생성 후

submission3.csv

a day ago by junyoungchoi

XGB Store data merge후 진행

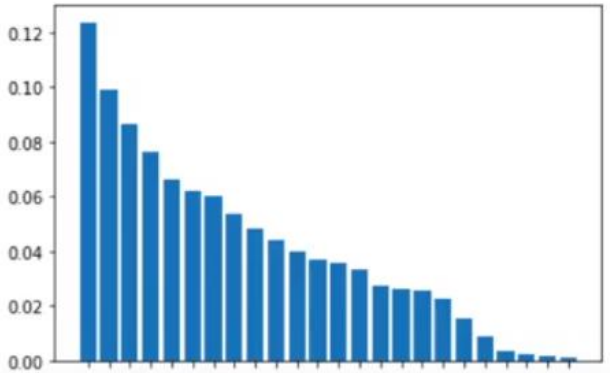
1634.69165

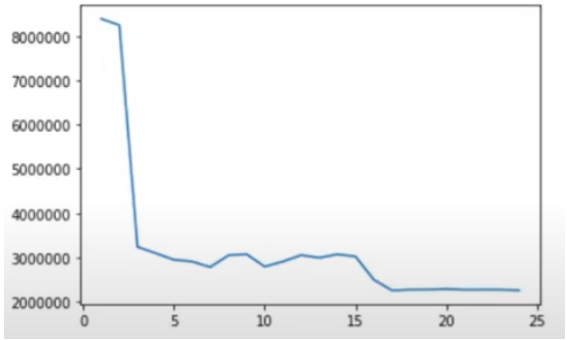
Store 데이터와 병합 후 모델링 구현 결과 1634.69 로 RMSE값 하락 모델의 성능이 좋아짐을 확인

프로젝트

로스만 상점 매출 예측

모델링 결과





submission3.csv a day ago by junyoungchoi XGB Store data merge후 진행	1634.69165
submission1.csv a day ago by junyoungchoi 'Promo','weekday','month' XGB	3803.78233
submission.csv a day ago by junyoungchoi 'Promo','SchoolHoliday','StateHoliday_0','StateHoliday_a','StateHoliday_b','StateHoliday_c','weekday XGB	3052.14204

결론

- Store 데이터 셋과 Train 데이터 셋을 통한 피쳐 엔지어링을 통해 좀더 의미 있는 변수들을 찾을 수 있었음.
- 결과적으로 매출에 가장 큰 영향을 끼치는 요소는 프로모션을 진행 하는 지의 여부
- 경쟁업체와의 거리는 중요할 것이라 생각하였지만 생각보다 덜 중요했음

프로젝트

대출상환 가능성 분석

데이터 소개 및 분석 동기

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	취업했는지(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	대한 신용정보를 조회한 개수

데이터 소개

Home Credit 기업 내부 데이터

- 채무자의 인적사항
- 대출에 대한 정보
- 채무자가 성공적으로 대출했는지 여부

분석 동기

대출 상환 여부를 결정짓는 요인을 분석하고 그에 따른 대출 플랜 제안

분석 설계

- 모델링
- 모델링에 따른 각 피쳐들의 영향력
- 영향을 많이 주는 5개 변수의 대출금 상환여부 파악

사용 언어 : Python

데이터 출처

Kaggle Home Credit Default Risk

프로젝트

대출상환 가능성 분석

모델링

모델링 진행전 shap value 해석을 위해 상관관계가 높은 변수를 삭제

	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	REGION_POPULATION_RELATIVE	DAY
FLAG_OWN_REALTY	1.000000	0.008244	0.003243	-0.042446	-0.001448	0.010826	-0.1
CNT_CHILDREN	0.008244	1.000000	0.029879	0.006465	0.023275	-0.033326	0.33
AMT_INCOME_TOTAL	0.003243	0.029879	1.000000	0.366717	0.441573	0.185047	0.06
AMT_CREDIT	-0.042446	0.006465	0.366717	1.000000	0.770938	0.092177	-0.0
AMT_ANNUITY	-0.001448	0.023275	0.441573	0.770938	1.000000	0.127204	0.01
REGION_POPULATION_RELATIVE	0.010826	-0.033326	0.185047	0.092177	0.127204	1.000000	-0.0
DAYS_BIRTH	-0.110930	0.332123	0.066875	-0.047089	0.017106	-0.023276	1.00
DAYS_EMPLOYED	-0.015164	0.068807	-0.041696	-0.085049	-0.048381	0.013870	0.34
DAYS_ID_PUBLISH	0.004217	-0.029581	0.029519	0.000988	0.013662	0.000946	0.26
OWN_CAR_AGE	0.019393	-0.010951	-0.126551	-0.111244	-0.108185	-0.088270	-0.0
CNT_FAM_MEMBERS	0.014595	0.883051	0.029342	0.066847	0.073912	-0.025638	0.28
HOUR_APPR_PROCESS_START	-0.105580	-0.009661	0.092505	0.047472	0.047113	0.182730	0.08
DAYS_LAST_PHONE_CHANGE	0.026066	-0.006102	-0.040823	-0.070924	-0.058709	-0.051167	0.08
AMT_REQ_CREDIT_BUREAU_YEAR	0.090058	-0.036431	0.031593	-0.037907	0.000270	0.015725	-0.0
AMT_CREDIT_TO_ANNUITY_RATIO	-0.083920	-0.022026	0.077303	0.656337	0.111694	0.003524	-0.0
AMT_CREDIT_SUM	-0.002745	0.035864	0.241929	0.135435	0.128144	0.077984	0.05
DAYS_CREDIT	0.000174	0.026285	-0.013266	-0.068411	-0.052613	-0.010819	0.20
CNT_CREDIT_PROLONG	-0.009790	-0.012065	0.016117	-0.000384	-0.005724	0.003701	0.02
count	0.008414	0.002649	0.116635	0.046902	0.013588	-0.034289	-0.0

XGBoost 를 이용한 모델링 진행

```
from xgboost import XGBClassifier
```

```
model = XGBClassifier(n_estimators=100, learning_rate=0.1)
model.fit(train[input_var],train['TARGET'])
```

상관관계 높은 변수 삭제 이유

변수 간의 상관관계가 높다면 shap value 해석을 진행하는데 있어서 설명력이 떨어지는 경향이 있기 때문

XGBoost 사용 이유

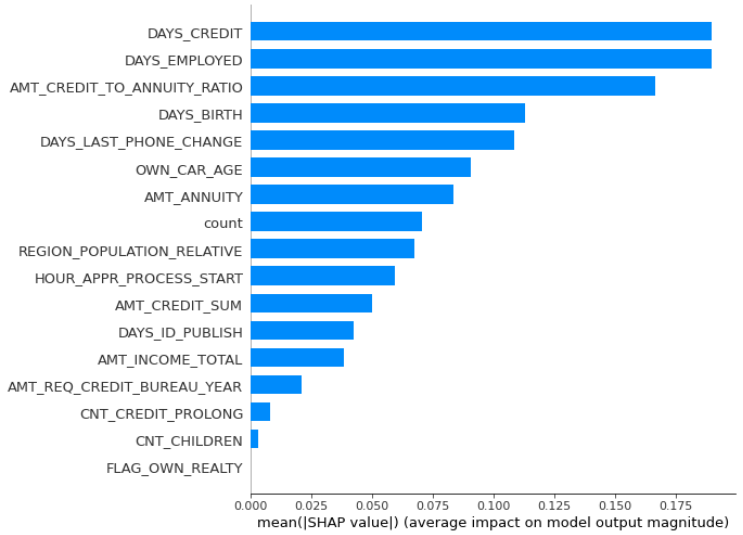
Shap Value를 사용하기위해서는 Tree 형 모델이여야 하고 속도가 다른 머신러닝 보다 빠르고 성능이 좋음

프로젝트

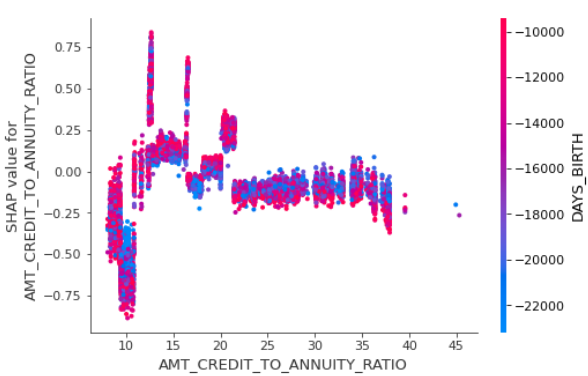
대출상환 가능성 분석

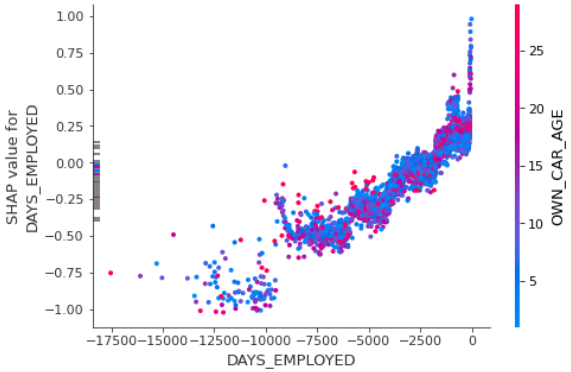
Shap Value 분석 및 결론

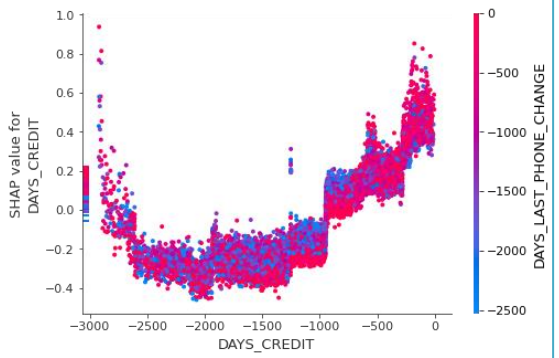
Shap Value 활용 타겟값의 영향력 확인



영향력 Top5 의 shap value 분석 진행







상환 여부 영향 상위 요소 5

- AMT_CREDIT_ANNUITY_RATIO
  - 대출 금액대비 월별 상환금액
- DAYS_EMPLOYED
  - 취업 시기
- DAYS_CREDIT
  - 대출을 받은 시기
- DAYS_LAST_PHONE_UPDATE
  - 핸드폰을 바꾼 시기
- DAYS_BIRTH
  - 태어난 날

감사합니다.