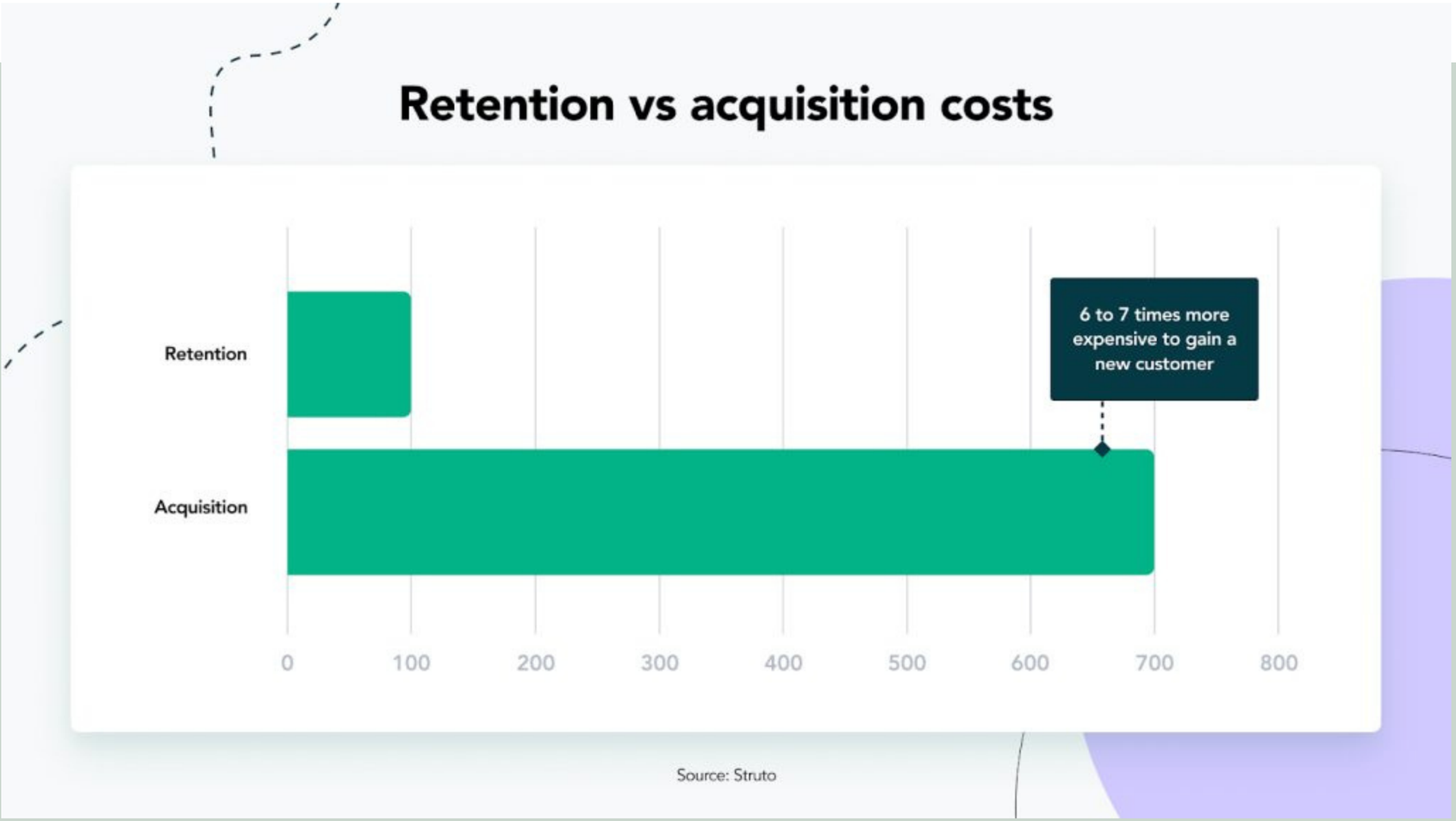
고객 행동 분석 & 매출 예측

AI 16기 최영조

Contents

01 - 1 프로젝트 수립 배경



신규 고객 확보 비용 6~7배

01 - 2 문제 정의 및 가설

"재구매율을 높여라"

가설1

소수의 로얄 고객이 전체 매출에 큰 영향을 가져온다.

가설2

6개월 이내에 재구매 하지 않은 고객은 이탈 가능성이 높다

02 - 1 데이터 설명

데이터 정보

온라인 커머스 회사 고객별 매출 정보
기간 : 2014년 2월 1일 - 2021년 10월 24일
구성 : 5000개의 고객 데이터, 41개의 컬럼

CustomerID	TOTAL_ORDERS	REVENUE	AVERAGE_ORDER_VALUE	CARRIAGE_REVENUE	AVERAGESHIPPING	FIRST_ORDER_DATE	LATEST_ORDER_DATE	AVGDAYSBETWEENORDERS
2354	124	11986.54	96.67	529.59	4.27	2016-12-30	2021-10-24	14.19
2361	82	11025.96	134.46	97.92	1.19	2018-03-31	2021-10-24	15.89
2415	43	7259.69	168.83	171.69	3.99	2017-11-30	2021-10-24	33.12
2427	44	6992.27	158.92	92.82	2.11	2019-04-09	2021-10-24	21.11
2456	55	6263.44	113.88	179.04	3.26	2020-10-23	2021-10-24	6.65

02 - 2 Feature Engineering

2개 모델

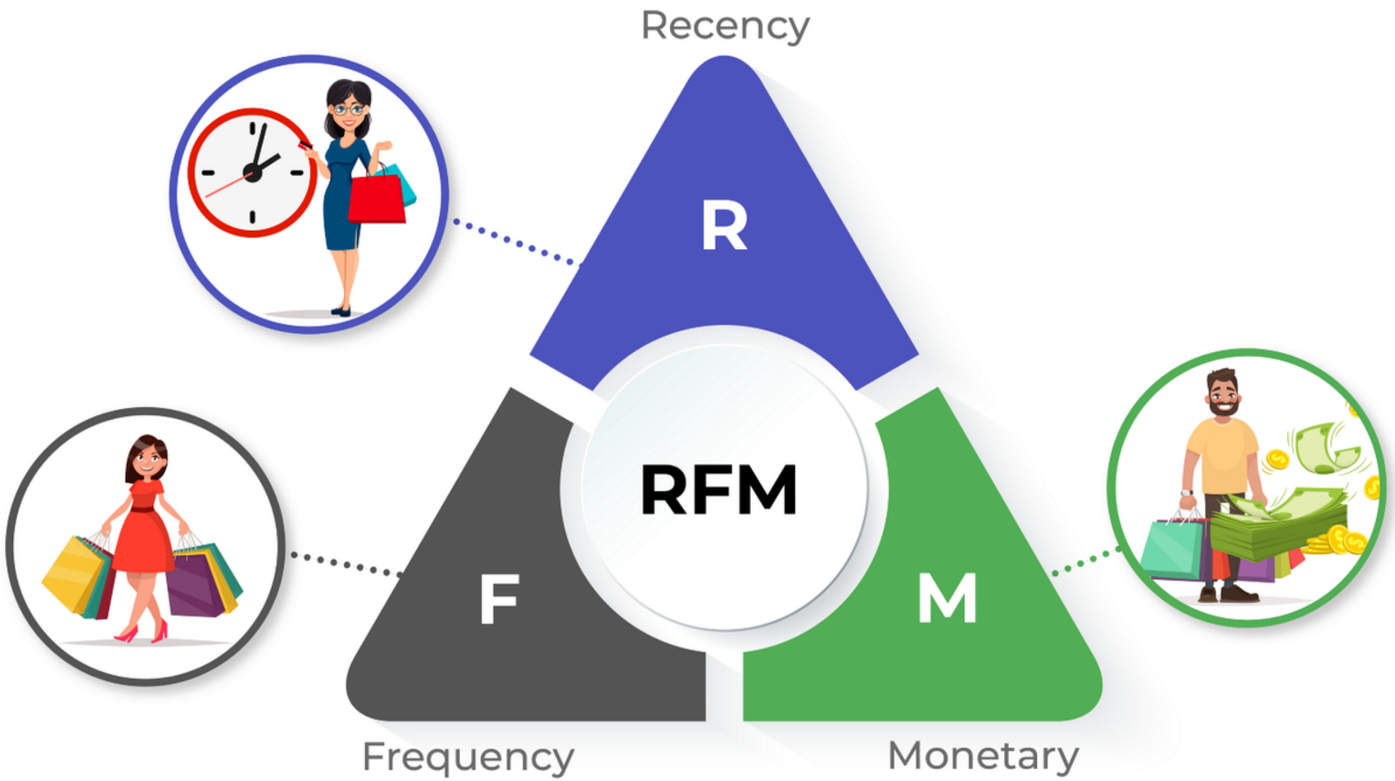
예상 매출액(연속형)

고객의 이탈(이진분류)

RFM Metrics

VIP 고객

RFM



Recency : 얼마나 최근에 구매했는가

Frequency : 얼마나 자주 구매했는가

Monetary : 얼마나 많은 금액을 지출했는가

03 모델링

목적

고객 매출 정보 -> 예상 매출과 고객 이탈 가능성

사용한 모델

dmlc XGBoost

주기별 데이터

- 트리기반 모델

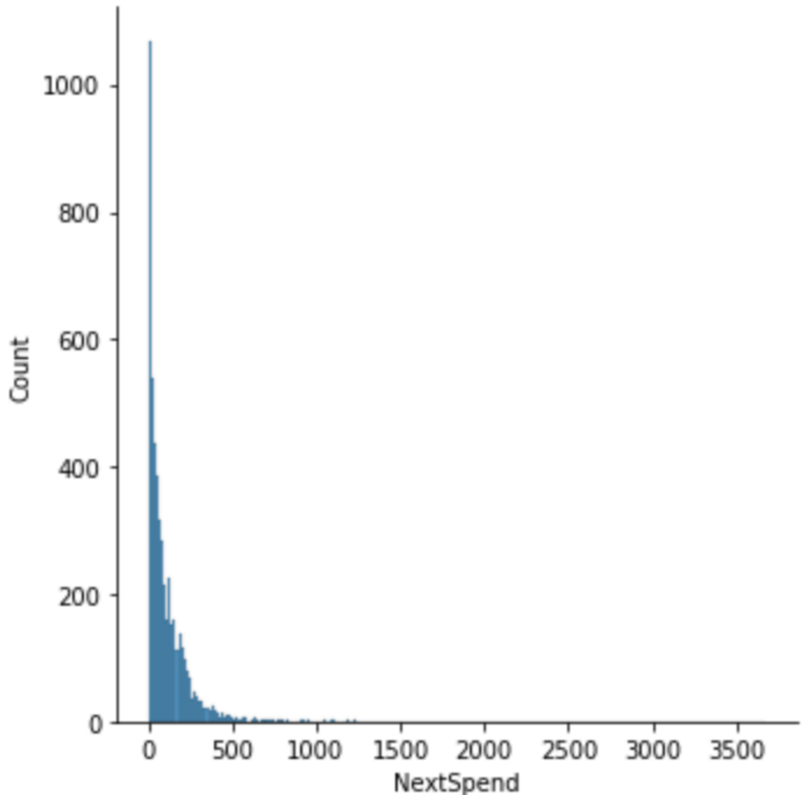
03 - 1 모델링 1

기준모델

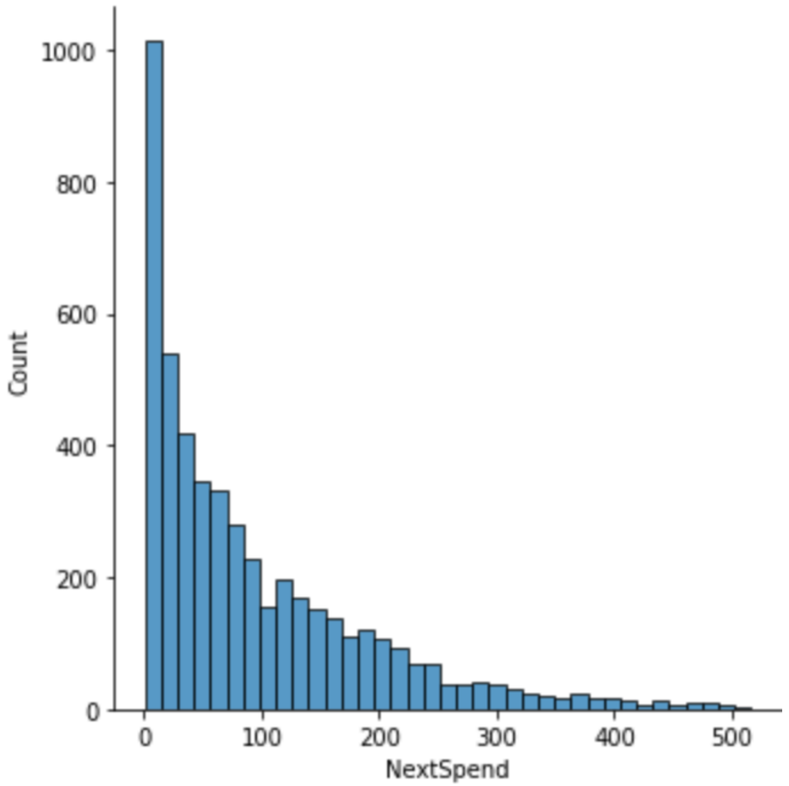
```
merged['NextSpend'].mean()
```

```
112.89721031949637
```

타겟 분포



->



03 - 1 모델링1

모델

```
tt = TransformedTargetRegressor(regressor=XGBRegressor(),func=np.log1p, inverse_func=np.expm1)

grid_params = {
 'regressor__n_estimators': [ 500, 1000 ],
 "regressor__learning_rate"    : [ 0.02, 0.05 ] ,
 "regressor__max_depth"        : [ 3, 5 ],
 "regressor__min_child_weight" : [ 2, 4 ],
 }

clf = RandomizedSearchCV(
    tt,
    param_distributions = grid_params,
    verbose = 1,
    n_iter=6,
    cv = 4,
    scoring='neg_root_mean_squared_error',
    n_jobs = -1
    )

results = clf.fit(X_train,y_train)
```

평가지표

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

```
#RMSE
results.best_score_
```

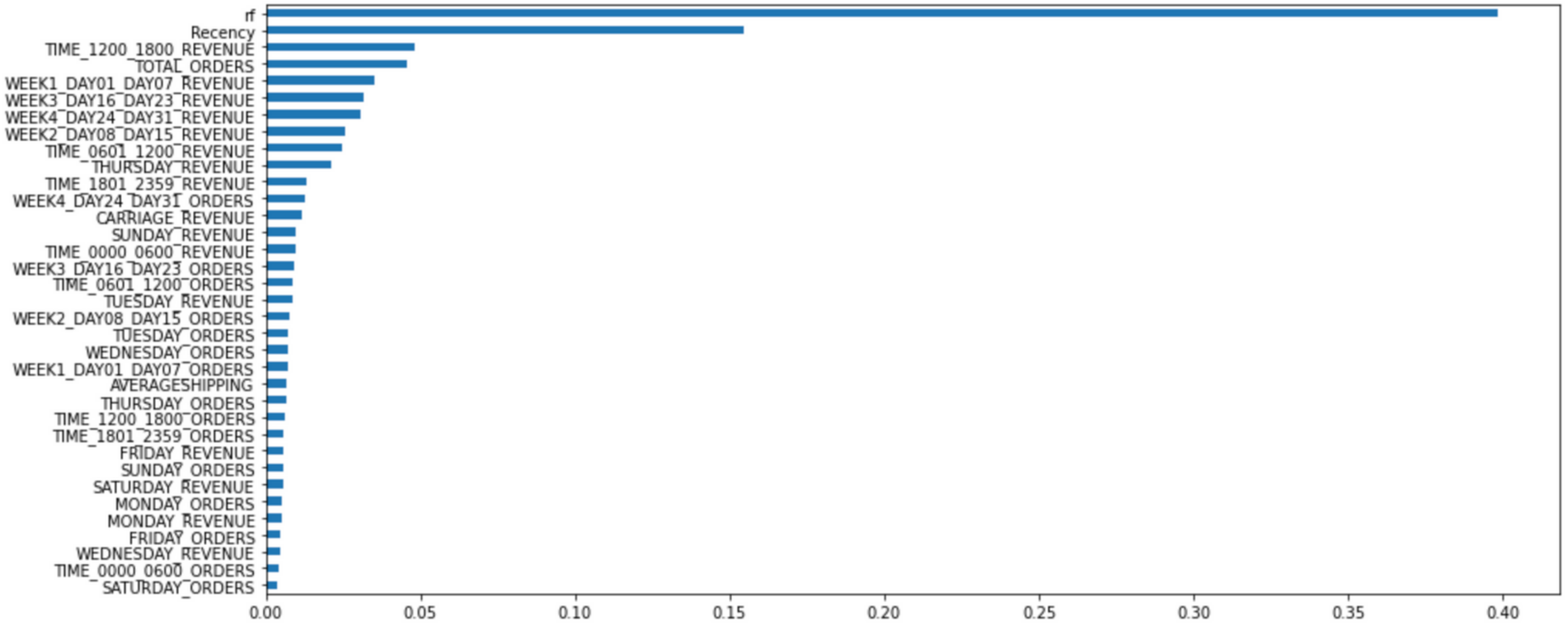
```
-44.17029268005997
```

```
#r2
results.best_score_
```

```
0.8284182941886379
```

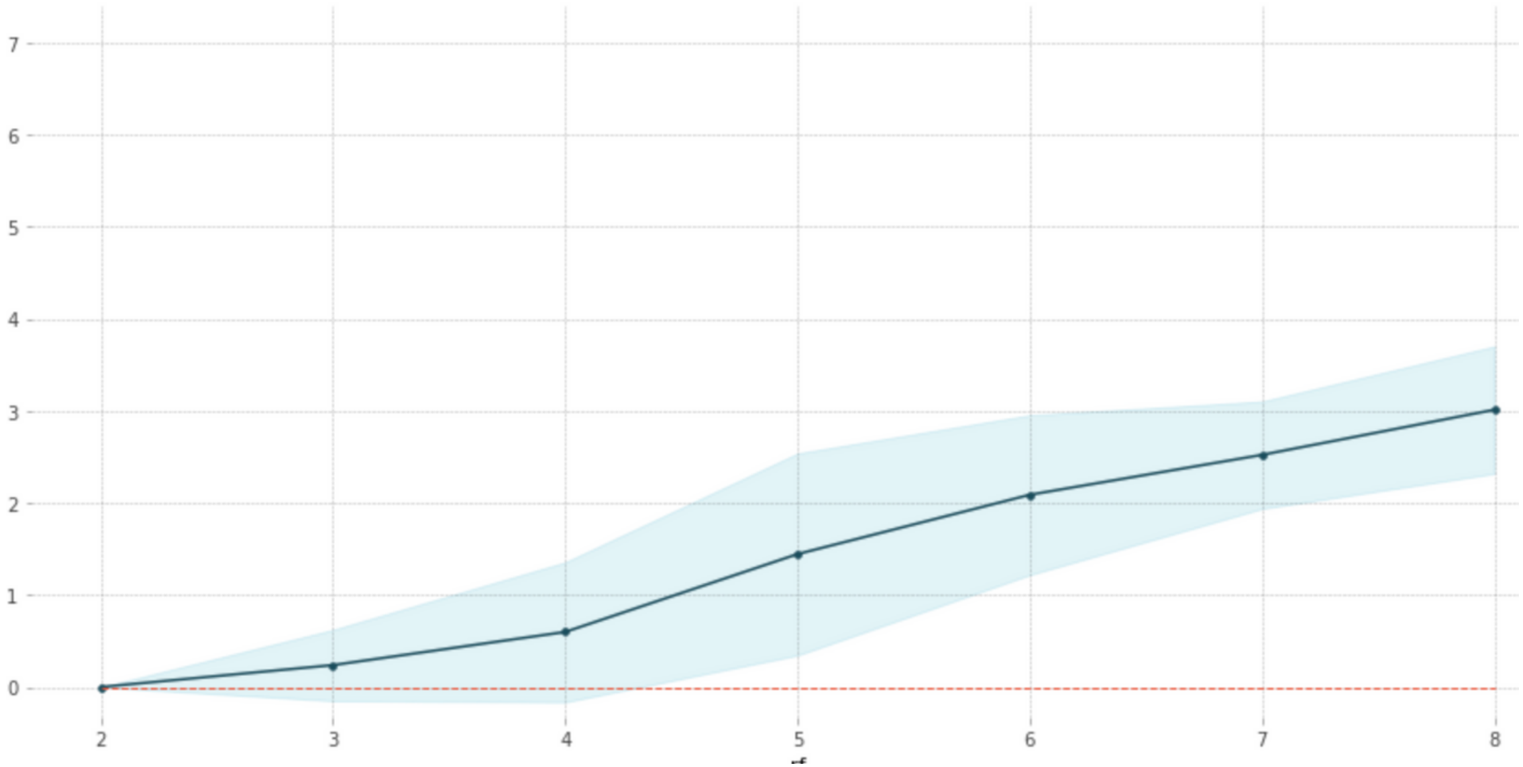
03 - 1 모델링1

특성 중요도



03 - 1 모델링1

PDP


PDP for feature "rf"
Number of unique grid points: 7
7
6
5
4
3
2
1
0
2
3
4
5
6
7
8

03 - 2 모델링2

기준모델

```
major = y2_train.mode()[0]
y2_pred = [major]*len(y2_train)
```

```
from sklearn.metrics import accuracy_score
accuracy_score(y2_train,y2_pred)
```

```
0.74925
```

평가지표

```
#roc_auc
clf2.best_score_
```

```
0.9234144935380861
```

모델

```
dists = {
    'xgbclassifier__max_depth': [5, 10, 15, None],
    'xgbclassifier__learning_rate' : [0.05,0.10,0.15,0.20,0.25,0.30]
}

pipeline = make_pipeline(
        OrdinalEncoder(),
        XGBClassifier(n_estimators = 200,
                      random_state=2,
                      n_jobs=-1,
                      max_depth=7,
                      learning_rate=0.2
                      )
    )

clf2 = RandomizedSearchCV(
pipeline,
param_distributions=dists,
n_iter=6,
cv=4,
scoring='roc_auc',
verbose=1,
n_jobs=-1
)


clf2.fit(X2_train, y2_train)
```

03 - 2 모델링2

순열중요도

Weight	Feature
0.0492 ± 0.0137	Frequency
0.0232 ± 0.0048	SUNDAY_ORDERS
0.0176 ± 0.0032	CARRIAGE_REVENUE
0.0164 ± 0.0084	WEEK4_DAY24_DAY31_ORDERS
0.0056 ± 0.0020	WEEK1_DAY01_DAY07_REVENUE

결론

활용방안

VIP 시스템

재구매 타이밍


감사합니다