

데이터 분석을 통한 고객 이탈 예측

김진유: 팀장 & 발표

박경희, 정우영, 정인교

EDA, 머신러닝, 노션 & 문서작성, 시각화

목차

1. 프로젝트 목표
2. Dataset 선정 이유
3. 프로젝트 계획
4. EDA
5. 상관성 예측을 위한 가설 수립
6. XGBoost 로 최종 모델 선정
7. 하이퍼파라미터 최적화 출력 및 결과
8. Streamlit 구현
9. 프로젝트 결과 분석

1. 프로젝트 목표

E Commerce Dataset 을 분석하여 향후 고객 이탈 예측을 하고자 한다. 상관계수를 측정하여 Churn 과 가장 높은 연관이 있는 항목을 찾아내고 머신러닝(및 딥러닝)에 적용하여 고객 이탈을 막는 방안을 수립한다.

2. Dataset 선정 이유

한국 전자상거래 시장 규모는 2010 년 부터 2020 년까지 연평균 19.7% 증가하여 2020 년에는 그 규모가 131 조원에 달했다. 이렇게 치열한 한국 전자상거래 시장에 최근 쿠팡, 알리익스프레스, 아마존 등의 외국계 자본들이 공격적인 투자를 바탕으로 점유율 경쟁에 뛰어들었다. 점유율 경쟁이 심화됨에 따라 기존 이커머스 기업들의 이탈고객도 꾸준히 늘어날 수 밖에 없는 실정이다. 이에 따라 고객이탈 데이터를 이용하여 고객들의 이탈여부를 분류하고, 잠재적 이탈고객을 관리할 수 있는 시스템의 도입을 제안하기 위해 이커머스 데이터 셋을 선정하였다.

3. 프로젝트 계획

데이터 전처리 >> 머신러닝 모델 학습 및 시각화 >> 결과 평가 및 분석/검증

4. EDA

> 사용 데이터 : [E Commerce Dataset.csv](#)

> 요구사항 분석

1. 주어진 데이터의 열과 이탈 유무 간의 상관계수를 도출한다.
2. 주어진 데이터의 열을 가공해 이탈 유무와 상관계수가 높은 열을 만들어낸다.
3. 머신러닝 분류 모델 중 정확도와 AUC 가 높은 분류 모델을 찾아낸다.
4. 분석한 내용을 직관적인 시각화 자료로 표현한다.
5. 이후 주어진 신규 데이터에 대해, 해당 고객의 이탈 확률을 출력해주는 기능을 구현한다.

> Description of Columns

CustomerID	고객 식별 번호
Churn	고객 이탈 여부 (고객이 서비스를 더 이상 이용하지 않는 경우)
Tenure	서비스 이용 기간 (보통 월 또는 년 단위)
PreferredLoginDevice	선호하는 로그인 장치 (예: 모바일, 컴퓨터)
CityTier	도시 등급 (일반적으로 도시의 크기나 중요도에 따라 분류)
WarehouseToHome	창고에서 집까지의 거리 (배송에 영향을 줄 수 있음)
PreferredPaymentMode	선호하는 결제 방식
Gender	성별
HourSpendOnApp	앱 사용 시간 (시간 단위)

NumberOfDeviceRegistered	등록된 기기의 수
PreferedOrderCat	지난 달 선호한 주문 카테고리 (예: 음식, 전자 제품 등)
SatisfactionScore	만족도 점수
MaritalStatus	혼인 상태
NumberOfAddress	고객별 등록된 주소의 수
Complain	지난 달 불만 제기 여부
OrderAmountHikeFromlastYear	지난 해 대비 주문 금액 증가율
CouponUsed	지난 달 사용한 쿠폰 수
OrderCount	지난 달 주문 횟수
DaySinceLastOrder	마지막 주문 이후의 일수
CashbackAmount	지난 달 환불 금액

> head() 출력

	CustomerID	Churn	Tenure	PreferredLoginDevice	CityTier	WarehouseToHome	PreferredPaymentMode	Gender	HourSpendOnApp	NumberOfDeviceRegistered	...
0	50001	1	4.0	Mobile Phone	3	6.0	Debit Card	Female	3.0	3	...
1	50002	1	NaN	Phone	1	8.0	UPI	Male	3.0	4	...
2	50003	1	NaN	Phone	1	30.0	Debit Card	Male	2.0	4	...
3	50004	1	0.0	Phone	3	15.0	Debit Card	Male	2.0	4	...
4	50005	1	0.0	Phone	1	12.0	CC	Male	NaN	3	...

5 rows × 21 columns

> info() 출력

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5630 entries, 0 to 5629
Data columns (total 21 columns):
#   Column                Non-Null Count  Dtype
---  -
0   CustomerID            5630 non-null   int64
1   Churn                 5630 non-null   int64
2   Tenure                5366 non-null   float64
3   PreferredLoginDevice  5630 non-null   object
4   CityTier              5630 non-null   int64
5   WarehouseToHome       5379 non-null   float64
6   PreferredPaymentMode  5630 non-null   object
7   Gender                5630 non-null   object
8   HourSpendOnApp        5375 non-null   float64
9   NumberOfDeviceRegistered 5630 non-null   int64
10  PreferedOrderCat      5630 non-null   object
11  SatisfactionScore     5630 non-null   int64
12  MaritalStatus         5630 non-null   object
13  NumberOfAddress       5630 non-null   int64
14  Complain              5630 non-null   int64
15  OrderAmountHikeFromlastYear 5365 non-null   float64
16  CouponUsed            5374 non-null   float64
17  OrderCount            5372 non-null   float64
18  DaySinceLastOrder     5323 non-null   float64
19  CashbackAmount        5630 non-null   int64
20  Combined              5630 non-null   object
dtypes: float64(7), int64(8), object(6)
memory usage: 923.8+ KB
```

> 결측치 확인

```
df.isna().sum()

CustomerID          0
Churn               0
Tenure             264
PreferredLoginDevice 0
CityTier           0
WarehouseToHome    251
PreferredPaymentMode 0
Gender             0
HourSpendOnApp     255
NumberOfDeviceRegistered 0
PreferedOrderCat    0
SatisfactionScore  0
MaritalStatus      0
NumberOfAddress     0
Complain           0
OrderAmountHikeFromlastYear 265
CouponUsed         256
OrderCount         258
DaySinceLastOrder  307
CashbackAmount     0
Tenure_bins        264
dtype: int64
```

> 데이터 정제

- 선호하는 결제 수단(PreferredPaymentMode) 열에서 CC 와 Credit Card, COD 와 Cash on Delivery, Mobile Phone 과 Phone 범주 불일치 확인 이후 replace 로 통일

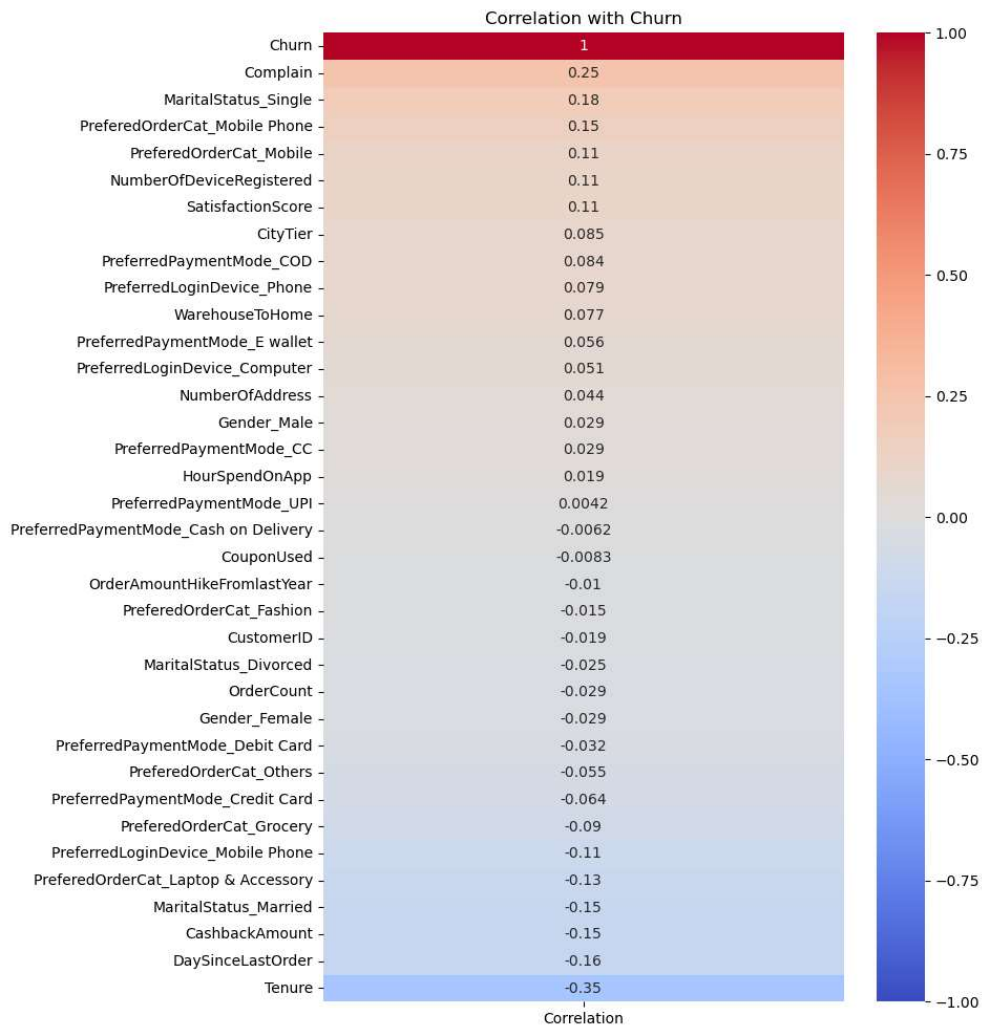
```
df["PreferredPaymentMode"].value_counts()

PreferredPaymentMode
Debit Card      2314
Credit Card    1501
E wallet        614
UPI             414
COD             365
CC              273
Cash on Delivery 149
Name: count, dtype: int64

df['PreferredPaymentMode'].replace('Cash on Delivery','COD', inplace=True)
df['PreferredPaymentMode'].replace('Mobile Phone','Phone', inplace=True)
df['PreferredPaymentMode'].replace('Credit Card','CC', inplace=True)

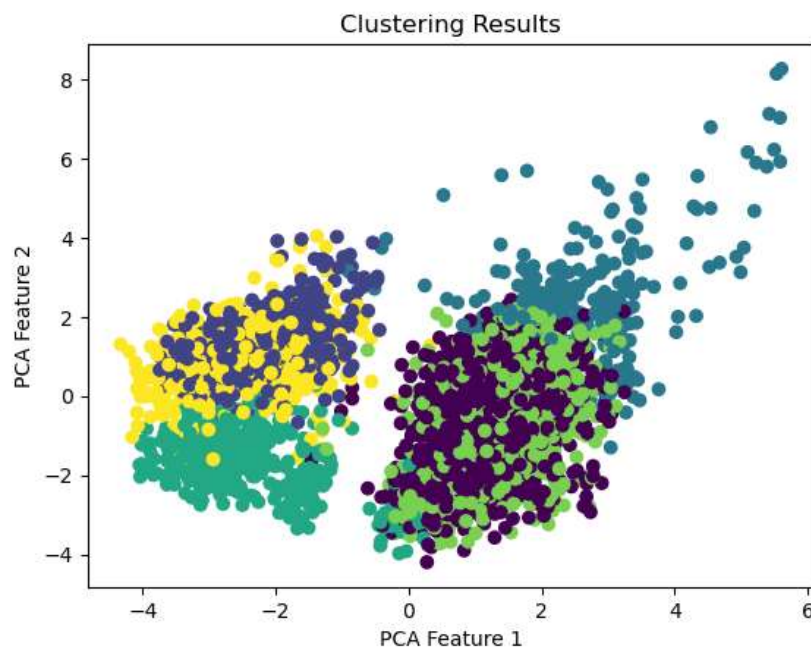
df['PreferredPaymentMode'].unique()
>> array(['Debit Card', 'UPI', 'CC', 'COD', 'E wallet'], dtype=object)
```

> 상관계수 확인



> 클러스터링

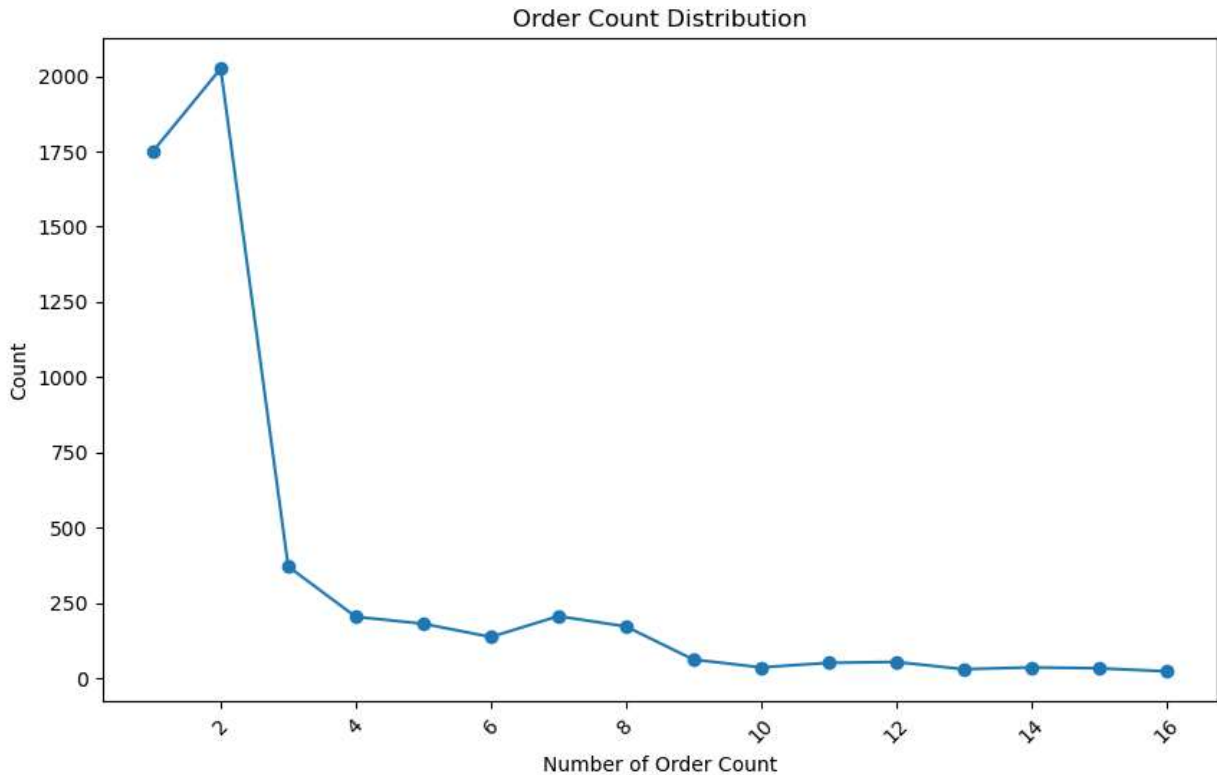
- PCA(n_components=5), KMeans(n_clusters=6)



5. 상관성 예측을 위한 가설 수립

분석을 통해 각 요인과 고객 이탈 사이의 상관성을 명확히 파악하면 전략적인 마케팅 및 운영 방안을 마련할 수 있다. 다음은 팀이 수립한 가설 목록이다.

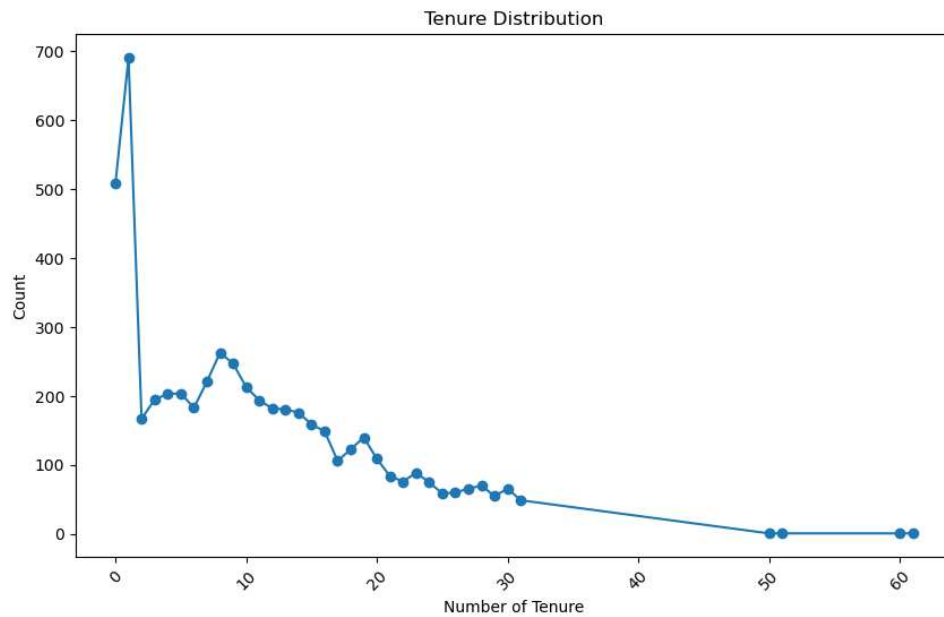
가설 1. 전월주문 횟수와 **churn** 은 높은 상관성을 가질 것이다.



전월 주문 횟수와 이탈 여부가 높은 상관관계를 보일 것이라고 예상하였다. 주문 횟수 '2'를 기점으로 구분했을 때 상관계수가 높을 것이라고 예상하여 python의 `corr()`함수를 사용하자 -0.03477로 낮은 상관계수값이 도출되었다.

주문횟수를 3, 4, 5 등 점차 기준을 높여가며 상관계수를 연산하였으나 의미있는 상관계수 값이 나오지 않았다. 이를 통해 주문 횟수와 이탈 여부와의 연관성은 높지 않다는 것을 알 수 있다.

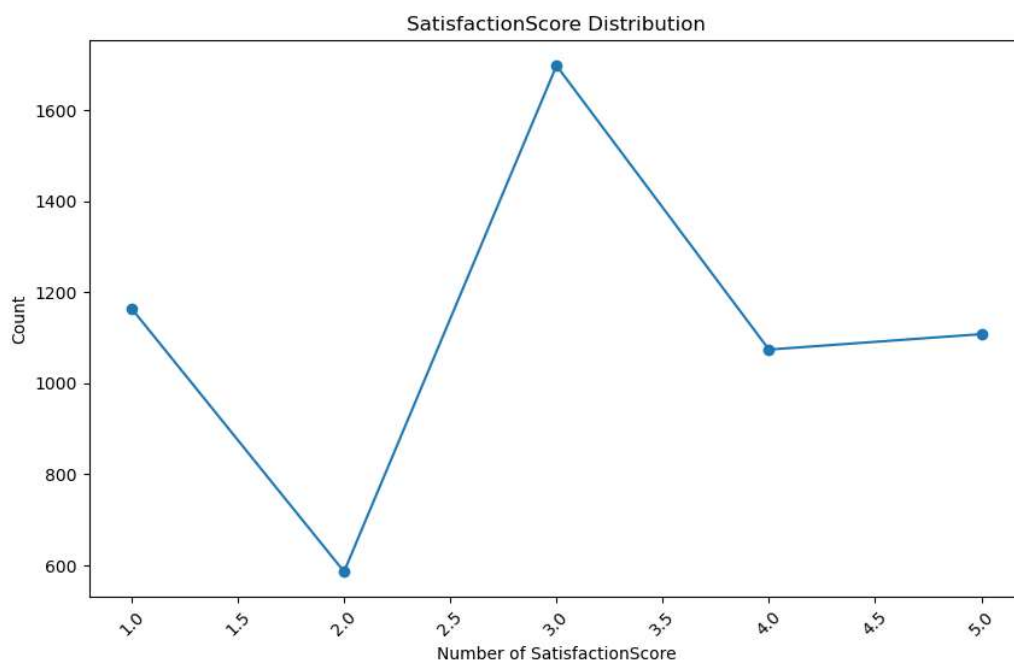
가설 2. 이용 기간(월)과 churn 의 상관성 예측



비교 기준을 이용 기간의 평균인 10 개월을 기점으로 가장 연관성이 클 것이라고 예측하였다. 숫자를 변경해가며 상관계수를 구하자, 비교 기준을 1로 잡았을 때 가장 큰 상관계수 값이 -0.5197 를 보였다. 음의 상관계수 값을 가진 것으로 보아 이용기간이 길어질수록 이탈 여부와는 반비례함을 알 수 있다.

1 보다 값이 커졌을 때 점점 상관계수가 작아지는 것을 확인하였다. 이용 기간이 1 보다 큰 값일 때 이탈 여부와 가장 큰 연관성이 있음을 알 수 있다.

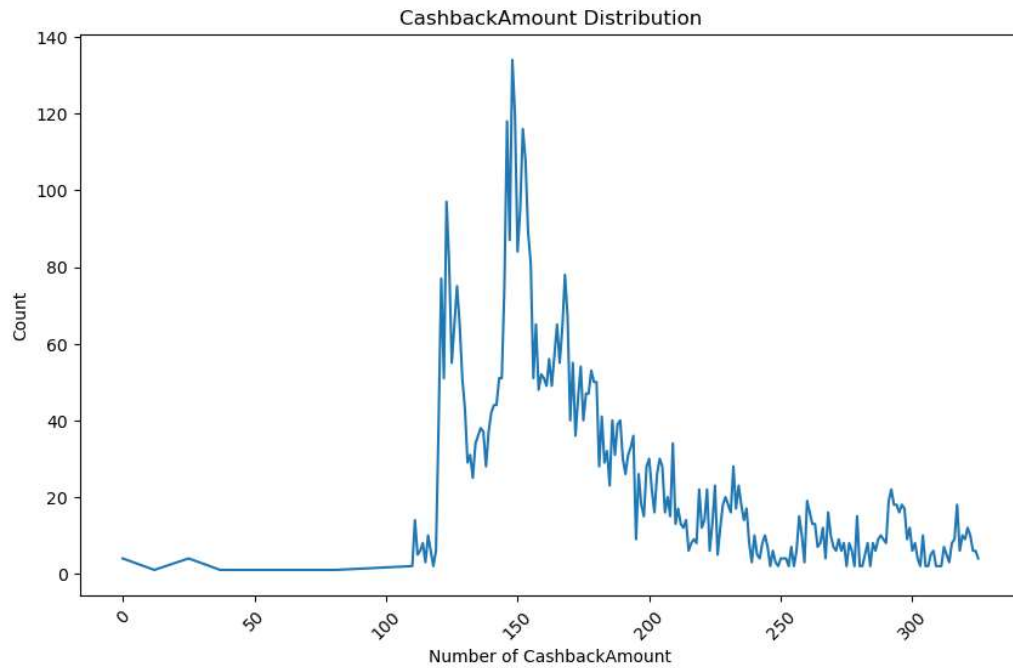
가설 3. 만족도 점수와 churn 의 상관성 예측



고객 만족도 점수와 이탈 여부가 상관성이 높을 것이라고 예측하여 상관계수를 도출해보았다.

만족도 점수에 대해 여러 상관계수를 도출해보았을 때 상관계수 값이 0.0924 정도로 전반적으로 매우 낮게 나온것으로 보아 만족도 점수와 이탈 여부는 낮은 연관성을 보인다는 것을 알 수 있다.

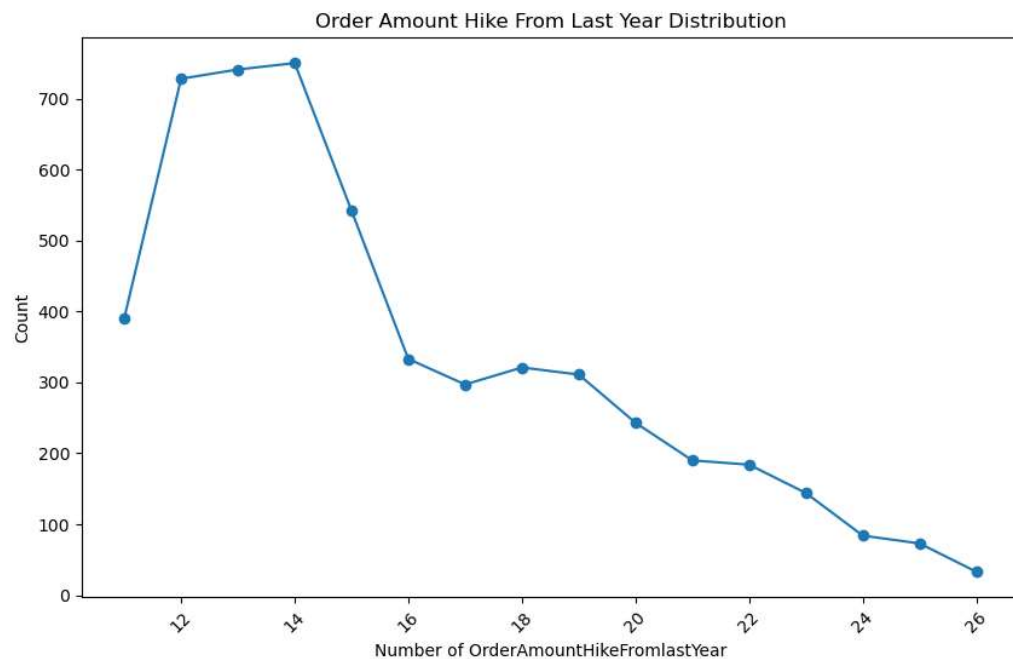
가설 4. 전월 환불 금액과 churn 의 상관성 예측



지난 달 환불 금액과 이탈 여부의 상관성을 알아보기 위해 상관계수를 도출하였다. 급격하게 도수가 변화하는 값을 기준으로 상관계수를 도출하자 -0.1599 이 나왔다.

따라서 지난 달 환불 금액과 이탈 여부는 큰 상관관계가 없음을 알 수 있다.

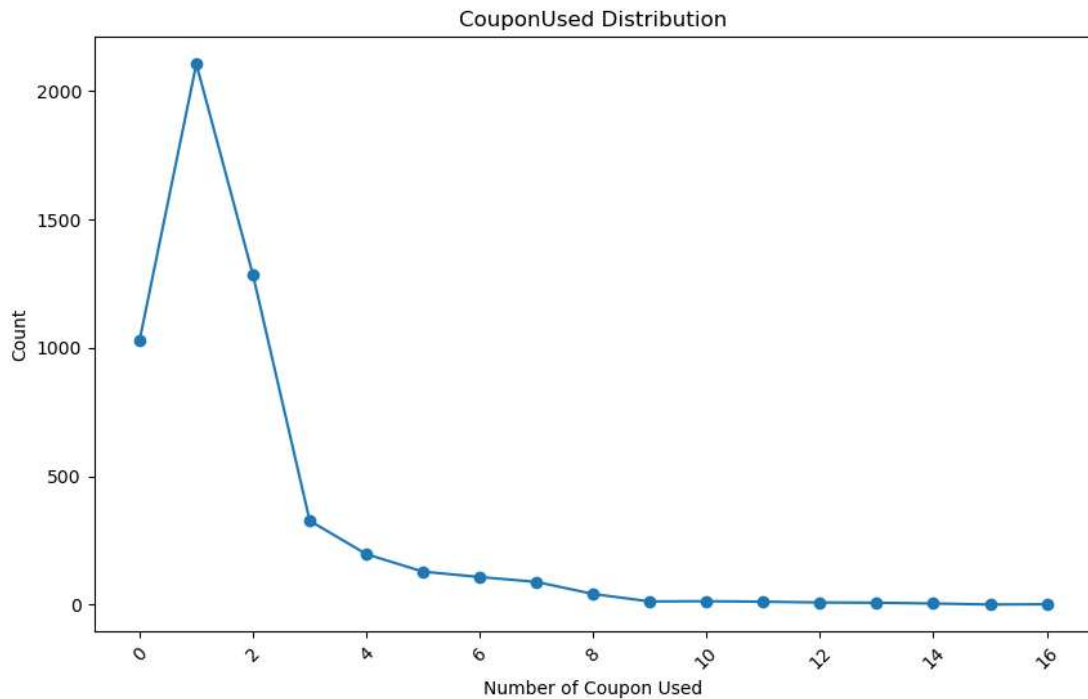
가설 5. 지난 해 대비 주문 금액 증가율과 churn 의 상관성 예측



전년 대비 주문 금액의 평균인 15 를 기준으로 높은 상관계수가 나올 것이라고 예상했으나, 최솟값과 최댓값의 전체 범위에 대해 상관계수를 구하더라도 아주 낮은 상관계수값(약 0.0184)이 도출되었다.

따라서 지난해 대비 주문 금액 증가율과 이탈 여부는 아주 낮은 상관관계를 보이는 것을 알 수 있다.

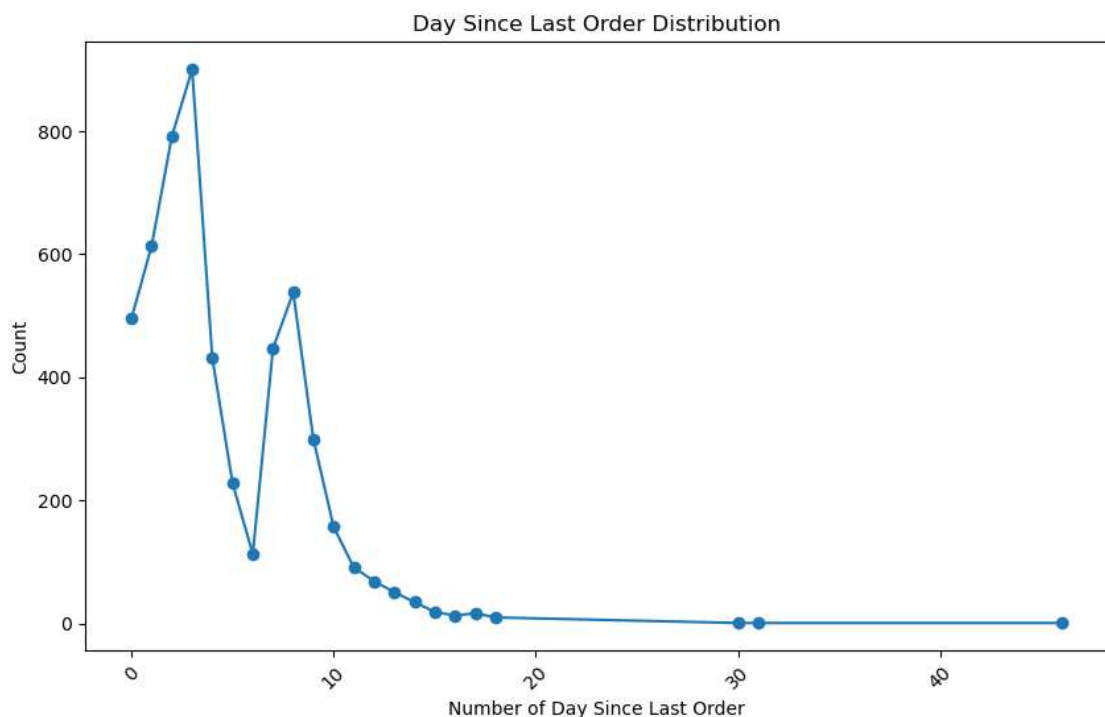
가설 6. 쿠폰 사용 횟수와 churn 의 상관성 예측



전월 쿠폰 사용 개수와 이탈 여부 사이에 상관관계가 있을 것이라 예상하고 상관계수를 도출하였다. 쿠폰 사용 개수가 1, 2, 3 이상일 때 각각의 상관계수를 도출하였으나 모두 매우 낮은 값(약 -0.0169)이 나오는 것을 확인했다.

따라서 쿠폰 사용 개수와 이탈 여부과의 연관성은 매우 낮다는 것을 알 수 있다.

가설 7. 마지막 주문 이후 경과 일수와 churn 의 상관성 예측



마지막 주문 이후 경과 일수와 churn 사이에 상관관계가 있을 것이라 예상하고 상관계수를 도출하였다. 10, 15, 20 일 등으로 일수가 늘어남에 따라 상관계수 값이 점점 낮아지는 것을 확인할 수 있다.

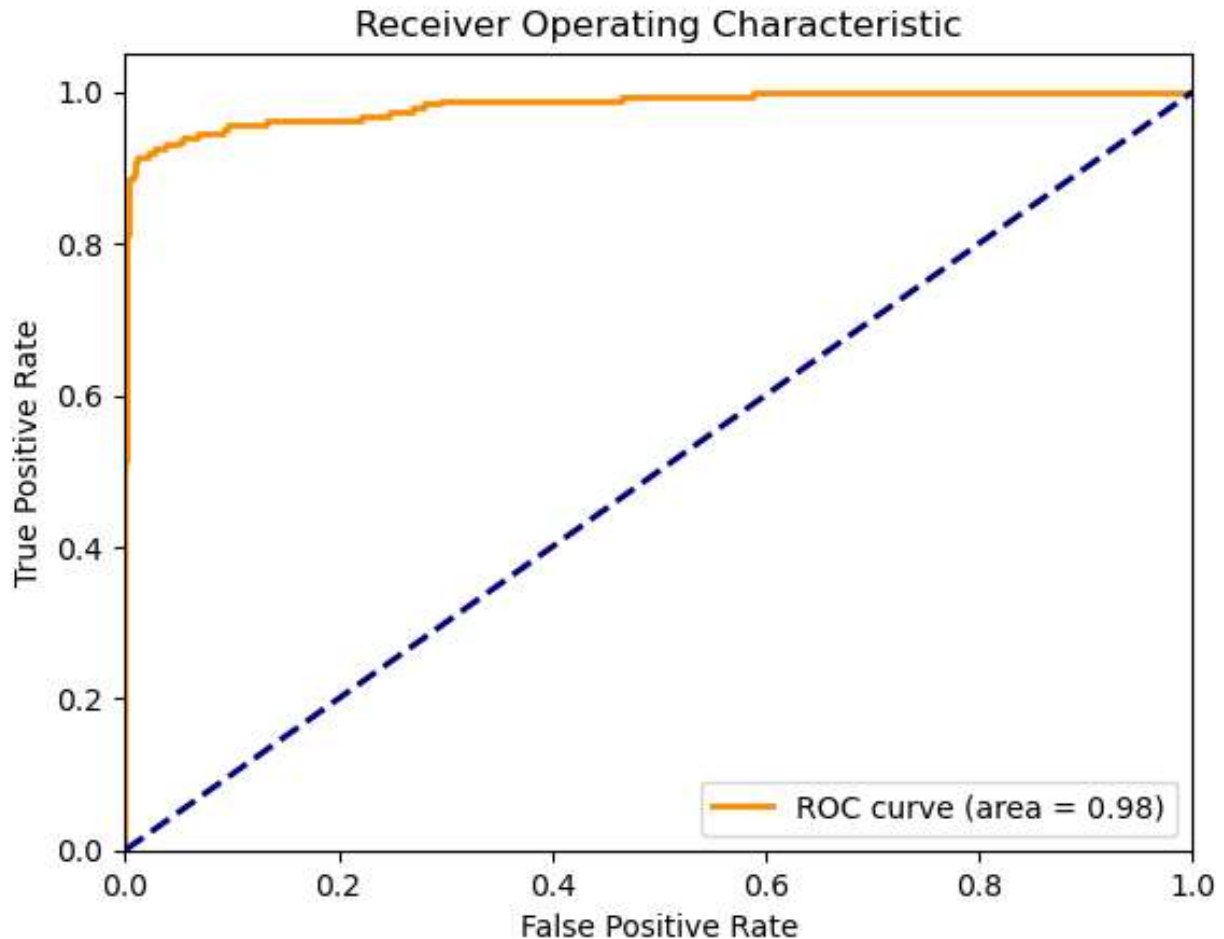
따라서 마지막 주문 이후 경과 일수와 churn 의 연관성은 낮음을 확인했다.

6. XGBoost 로 최종 모델 선정

가장 높은 정확도와, AUC 를 가진 XGBoost 를 선정하였다. XGBoost 는 Decision Tree 와 달리 병렬처리 가능하기 때문에 CPU 멀티 코어를 활용해 속도 개선이 가능하다.

Train accuracy: 100.0 %

Test accuracy: 97.60213143872114 %



7. 하이퍼파라미터 최적화 출력 및 결과

- GridSearchCV 를 활용한 하이퍼파라미터 최적화
- 최적 하이퍼파라미터 출력 및 결과


```
Best parameters: {'classifier__learning_rate': 0.1, 'classifier__max_depth': 7,
                  'classifier__n_estimators': 300}
Best score: 0.9567066222715501
Test accuracy with best model: 97.51332149200711 %
Test accuracy: 97.51332149200711 %
```

8. Streamlit 구현

> 메인화면에서 csv 파일 업로드

회원 이탈 확률 예측

csv 파일을 선택해주세요.

 Drag and drop file here
Limit 200MB per file • CSV

Browse files

CSV 파일을 업로드해주세요.

> 업로드한 고객 데이터에 대해 고객별 이탈 위험 비율과 위험도 표시

(Red 고객: 이탈 확률 50 퍼센트 이상, Yellow 고객: 이탈확률 30 퍼센트 이상)

회원 이탈 확률 예측

CSV 파일을 선택해주세요.

 Drag and drop file here
Limit 200MB per file • CSV

Browse files

 E Commerce Dataset.csv 471.1KB ×

Red 고객 비율: 16.46536412078153 %

Yellow 고객 비율: 0.2664298401420959 %

	CustomerID	Churn Probability (%)	Churn Risk
0	50,001	92.9981	RED
1	50,002	99.9221	RED
2	50,003	99.4874	RED
3	50,004	98.9837	RED
4	50,005	96.7011	RED
5	50,006	99.8993	RED
6	50,007	98.2652	RED
7	50,008	90.8084	RED
8	50,009	92.8986	RED
9	50,010	98.3452	RED

9. 프로젝트 결과 분석

- **결론** : feature 각각과 churn은 상관계수가 높지 않지만 기계학습 결과가 매우 좋은 것으로 나타났다. 상관계수가 낮아 직관적으로 이탈확률과 관계가 없을 것으로 사료된 feature들을 임의로 제거 후 머신러닝을 수행했을 때 결과가 오히려 기존보다 낮게 나오는 문제가 있었다.
- **기대효과** : 머신러닝 모델과 Streamlit을 연동해 실시간으로 신규 및 기존 고객의 이탈 위험성을 예측하여 이탈 가능성이 높은 고객들을 대상으로 추가 할인 쿠폰을 제공하거나, 더 나은 서비스를 제공하는 등의 방법으로 고객 이탈을 방지해 매출 하락을 방지할 수 있을 것으로 기대된다.
- **한계** : E 커머스 기업들의 대부분이 고객 관련 데이터를 기업 비밀로 관리하고 있기 때문에, 수업 때 배운 시계열로 추출할 수 있는 데이터나, 정확한 고객별 구매기록, 매출 등 BM과 직접적으로 연관된 데이터를 구할 수 없어 분석해보지 못했다. 우리가 선정한 데이터에서 유의미한 다른 특성을 도출해보려 했지만, 앞선 이유와 특성들이 이미 충분히 독립적인 이유로 새로운 특성을 찾는 데 한계가 있었다.
- **개선점** : 현재는 고객 데이터를 CSV 파일로 불러왔지만, DB에 저장된 고객 데이터를 직접 불러와 모델로 확률을 계산해 고객 이탈 리스크를 계산하는 방식으로 개선 가능하다.