Sejong AI Challenge

21011928
전주혁

Contents

Intro

1. 주제

통신사 고객 이탈 여부 예측 문제

* 통신 회사는 수익을 늘리기 위해 계약 해지(이탈)을 피해야함

* 고객 이탈 분석은 고객 이탈을 예측하고 이탈을 유발하는 근본적인 이유를 정의

* 이탈을 유발하는 원인을 파악해 대응조치를 취해 고객 이탈을 방지

Intro

2. Data Access

1. 시각화 후 불필요한 피쳐 판단

2. 가설 설정 후 피쳐를 Drop or 연관있다 생각되는 피쳐들을 이용해 파생 변수 생성

Part 2,

EDA

Part 2,

EDA

	index	Unnamed: 0	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines
0	0	1869	7010-BRBUU	Male	0	Yes	Yes	72	Yes	Yes
1	1	4528	9688-YGXVR	Female	0	No	No	44	Yes	No
2	2	6344	9286-DOJGF	Female	1	Yes	No	38	Yes	Yes
3	3	6739	6994-KERXL	Male	0	No	No	4	Yes	No
4	4	432	2181-UAESM	Male	0	No	No	2	Yes	No
...	...	...	...	...	...	...	...	...	...	...
4783	5981	3772	0684-AOSIH	Male	0	Yes	No	1	Yes	No
4784	5982	5191	5982-PSMKW	Female	0	Yes	Yes	23	Yes	Yes
4785	5983	5226	8044-BGWPI	Male	0	Yes	Yes	12	Yes	No
4786	5984	5390	7450-NWRTR	Male	1	No	No	12	Yes	Yes
4787	5985	860	4795-UXVCJ	Male	0	No	No	26	Yes	No

4788 rows × 23 columns

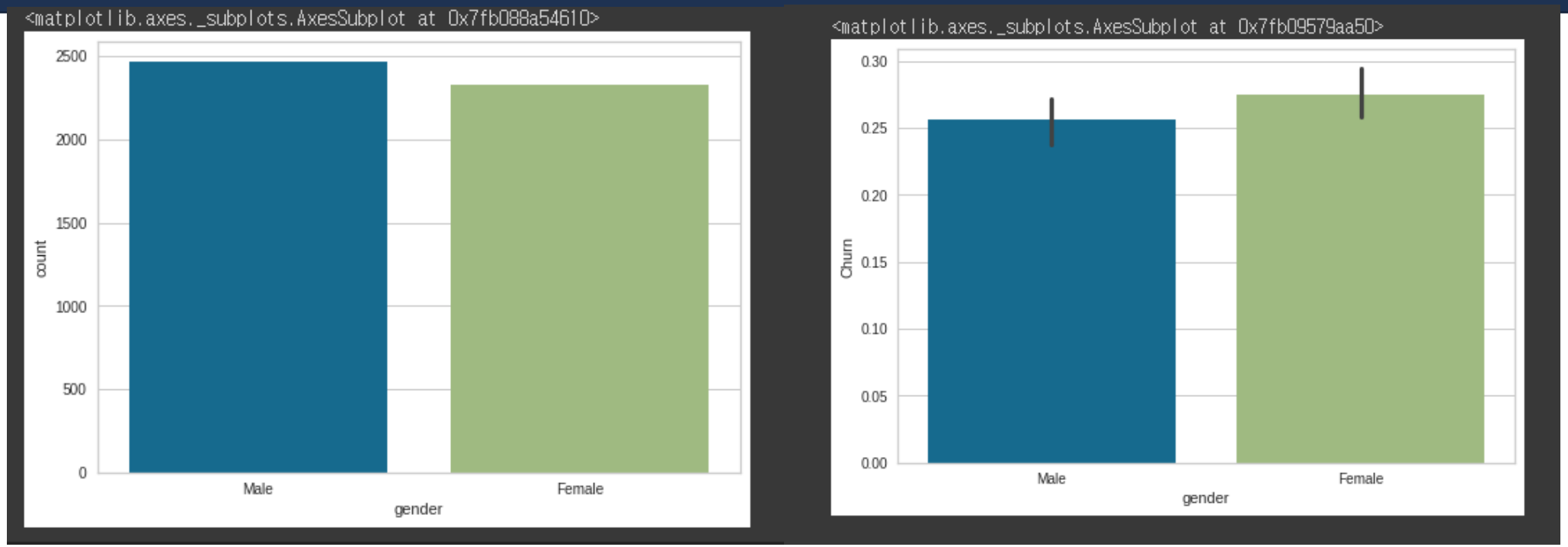
학습에 불필요한 피쳐

- index
- Unnamed: 0
- customerID

-> Drop

Part 2,

EDA

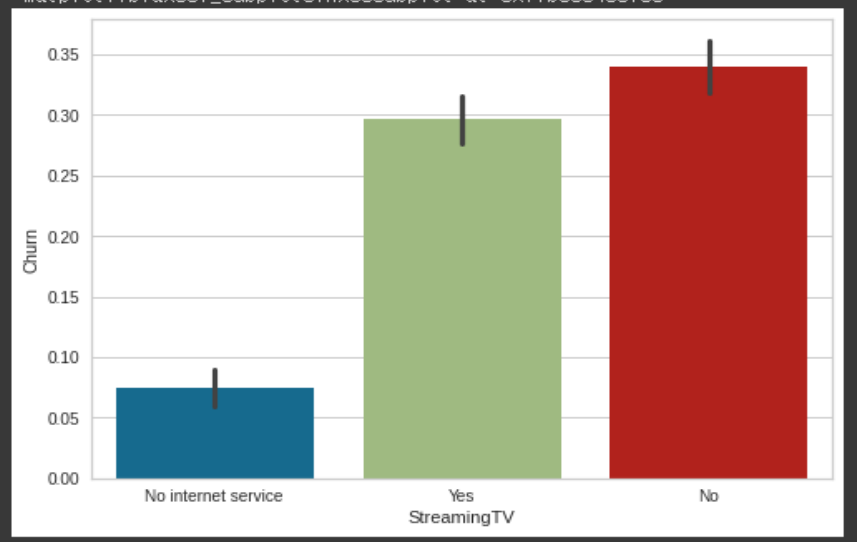


Barplot, countplot을 통해 학습에 불필요한 피쳐라 생각
-> Male, Female 이 타겟값
-> Drop

Part 2,

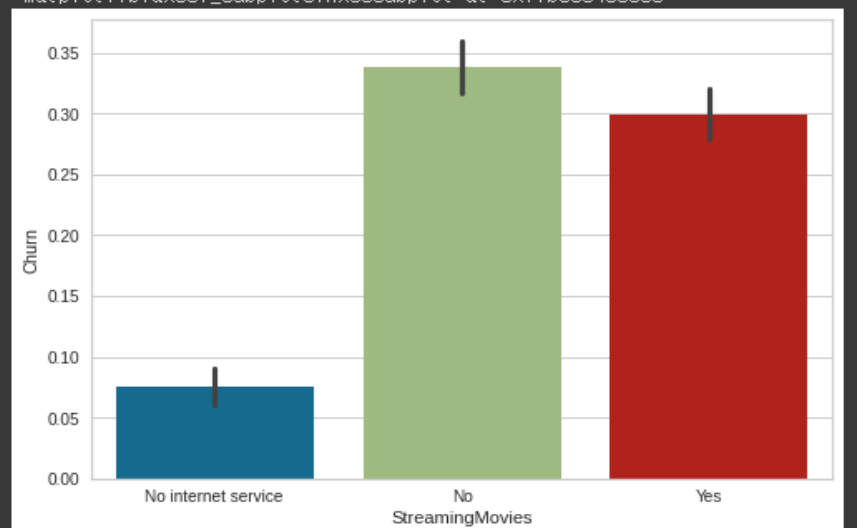
EDA

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb095466190>
```




```
4] sns.barplot(data = train, x = 'StreamingMovies', y= 'Churn')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb095466990>
```



StreamingTV, StreamingMovies,
InternetService 피쳐

-> TV 스트리밍 여부
-> 영화 스트리밍 여부
-> 인터넷 공급망 종류

피쳐들은 통신 회사와 관계 없이
고객 이탈 여부와 관련이 없을 것이라 가설 설정
-> 시각화 후 피쳐 drop 판단

Part 2,

EDA

```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4788 entries, 0 to 4787
Data columns (total 18 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   gender            4788 non-null   object
 1   SeniorCitizen     4788 non-null   int64
 2   Partner           4788 non-null   object
 3   Dependents        4788 non-null   object
 4   tenure            4788 non-null   int64
 5   PhoneService      4788 non-null   object
 6   MultipleLines     4788 non-null   object
 7   InternetService   4788 non-null   object
 8   OnlineSecurity    4788 non-null   object
 9   OnlineBackup      4788 non-null   object
 10  DeviceProtection  4788 non-null   object
 11  TechSupport       4788 non-null   object
 12  Contract          4788 non-null   object
 13  PaperlessBilling  4788 non-null   object
 14  PaymentMethod     4788 non-null   object
 15  MonthlyCharges    4788 non-null   float64
 16  TotalCharges      4788 non-null   object
 17  Churn             4788 non-null   object
dtypes: float64(1), int64(2), object(15)
```

15개의 Object Type 존재

-> 피쳐에 특성에 맞춰 인코딩

→ 하지만 이번 대회는 타임어택임

→ 따라서 모델로 pycaret을 사용하는 것이 적절하다 판단

Part 3,

Feature Engineering

Part 2,

EDA

```
train.drop(['index', 'Unnamed: 0', 'customerID'], axis = 1, inplace = True)
test.drop(['index', 'Unnamed: 0', 'customerID'], axis = 1, inplace = True)

train.drop(['StreamingTV', 'StreamingMovies', 'InternetService'], axis = 1, inplace = True)
test.drop(['StreamingTV', 'StreamingMovies', 'InternetService'], axis = 1, inplace = True)
```

1. EDA 중 가설로 설정했던
피쳐들은 통신 회사와 관계 없이
고객 이탈 여부와 관련이 없을 것이라는 판단으로
-> 3개의 피쳐 Drop

2. 학습에 불필요한
 - index, Unnamed: 0, customerID column Drop

Part 5,

Modeling

Part 5,

Modeling

pycaret

15개의 object 피쳐들을 빠르게 변환 하여
모델 별 평가지표를 쉽게 보기 위해 pycaret 사용

pycaret 이란?

- Machine Learning Workflow를 자동화하는 오픈소스 라이브러리

- Classification, Regression, Clustering 등의 Task에서 사용하는 여러 모델들을 동일한 환경에서 실행을 자동화한 라이브러리

- 여러 모델을 비교 가능

Part 5,

Modeling

Model Metrics

	Model	Accuracy
gbc	Gradient Boosting Classifier	0.8001
lr	Logistic Regression	0.7995
ada	Ada Boost Classifier	0.7947
ridge	Ridge Classifier	0.7941
rf	Random Forest Classifier	0.7905
lightgbm	Light Gradient Boosting Machine	0.7798
et	Extra Trees Classifier	0.7777
dt	Decision Tree Classifier	0.7649
dummy	Dummy Classifier	0.7440
svm	SVM - Linear Kernel	0.7120
nb	Naive Bayes	0.6938
knn	K Neighbors Classifier	0.6828
qda	Quadratic Discriminant Analysis	0.5503
lda	Linear Discriminant Analysis	0.5431

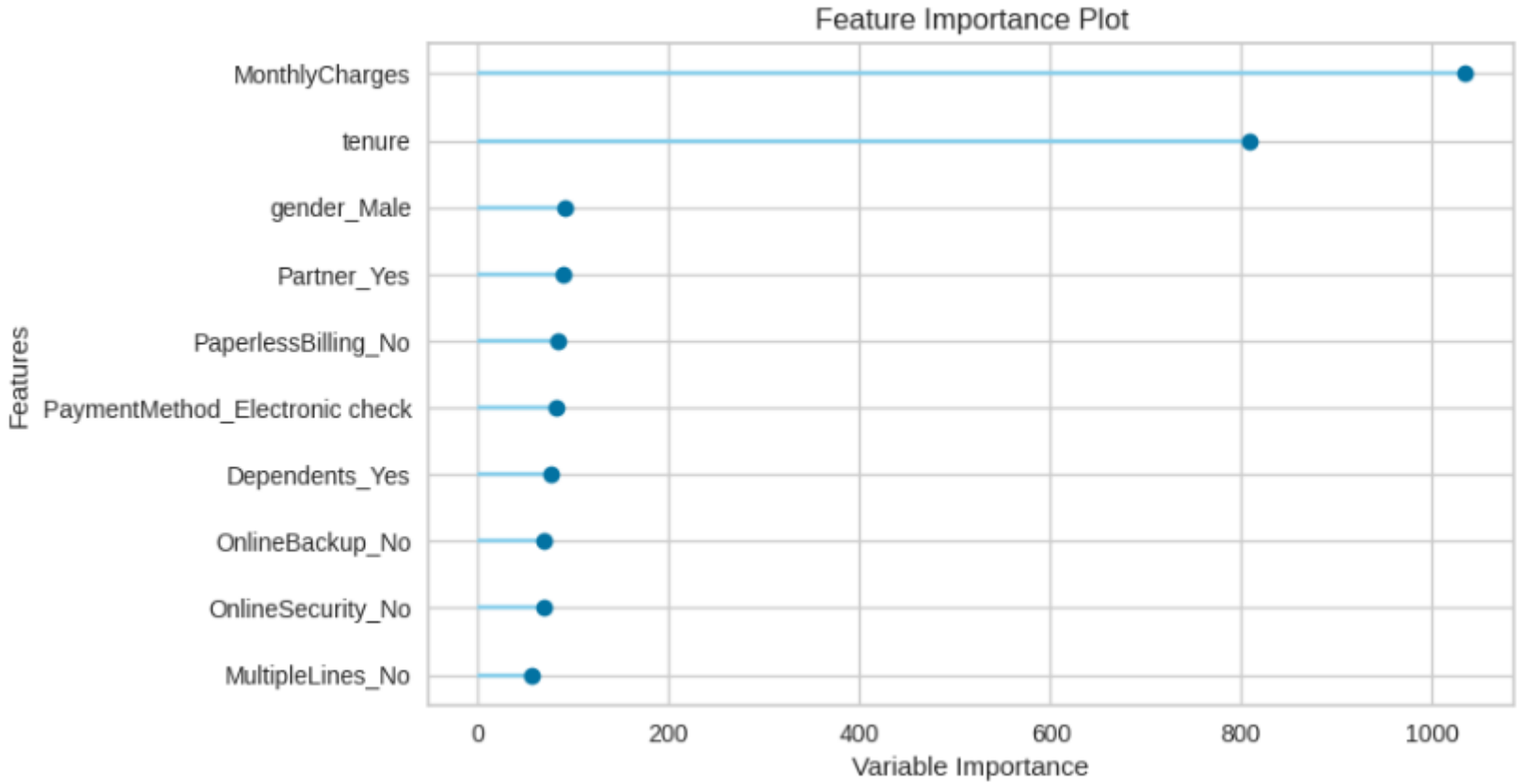
-> 평가지표인 Acc 값이 제일 높은 모델들 사용 결정

Fold	Accuracy
0	0.7589
1	0.8030
2	0.7761
3	0.7642
4	0.7761
5	0.7642
6	0.8030
7	0.8209
8	0.8358
9	0.7254
Mean	0.7828
Std	0.0312

GBC 모델
Stratified 10 Fold Acc

Feature Importance

Feature Importance



Feature Importance 결과

“ 감사합니다. ”