Sejong AI Challenge

문제 4번(기타 트랙) 통신사 고객 이탈 여부 예측 문제

19011818 정찬호

목차

1. 데이터 분석
2. 전처리
3. 모델 선택
4. 모델 평가 및 최적화

1. 데이터 분석

Numeric
index
Unnamed: 0
SeniorCitizen
tenure
MonthlyCharges

String
TotalCharges
custumerID/ gender
Partner/ Dependents
PhoneService/ MultipleLines
InternetService/'OnlineSecurity OnlineBackup/DeviceProtection TechSupport/StreamingTVStreamingMovies/Contract',PaperlessBilling/PaymentMethod /Churn

2. 전처리

방법 1. 제외

- 학습 시 혼란을 줄 수 있는 데이터(ID, Unnamed : 0, customerID)
- 처리 복잡한 데이터(TotalCharges)

```
x_train=train.drop(['index','Unnamed: 0','customerID','Churn','TotalCharges'],axis=1)
y_train=pd.DataFrame(train['Churn'],index=train['Churn'].index,columns=['Churn'])
x_test=x_test.drop(['index','Unnamed: 0','customerID','TotalCharges'],axis=1)
```

2. 전처리

방법 2. 변환

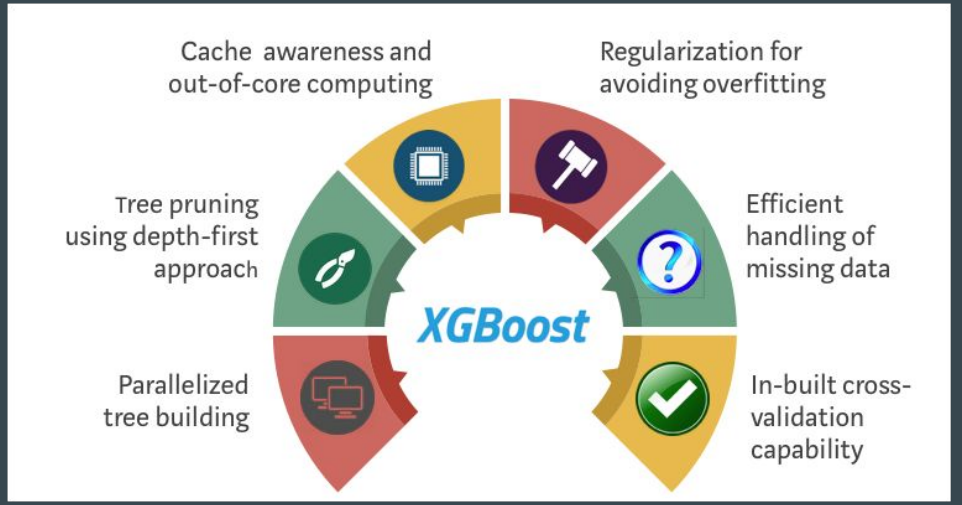
- LabelEncoder를 이용해 범주형 변수를 수치화

```
for c in cat_cols:
    if(c in ps):

        continue
    else:
        le=LabelEncoder()
        le.fit(x_train[c])
        x_train[c]=le.transform(x_train[c])
        x_test[c]=le.transform(x_test[c])
```

3. 모델 선택

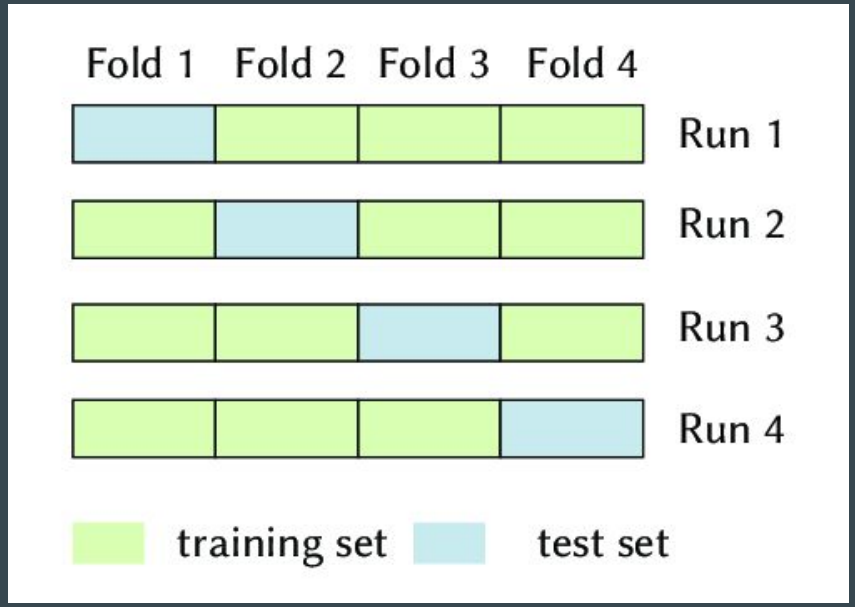
1. 결측치 고려
2. 빠른 속도
3. 높은 성능



4. 모델 평가 및 최적화

1. K-Fold Cross-Validation

- 모든 데이터를 가지고 train
- Accuracy=(acc1+acc2+acc3+acc4)/4



4. 모델 평가 및 최적화

2. Hyperparameter tuning

- learning_rate = 0.291
- n_estimator = 100
- gamma = 10

Parameters for Tree Booster

- `eta` [default=0.3, alias: `learning_rate`]
  - Step size shrinkage used in update to prevents overfitting. After each boosting step, we can directly get the weights of new features, and `eta` shrinks the feature weights to make the boosting process more conservative.
  - range: [0,1]
- `gamma` [default=0, alias: `min_split_loss`]
  - Minimum loss reduction required to make a further partition on a leaf node of the tree. The larger `gamma` is, the more conservative the algorithm will be.
  - range: [0,∞]

Thank you !