TELECOM CUSTOMER CHURN PREDICTION

MileStone3

Content

1. Feature selection
2. Imbalance data Handling
3. Model Selection & Comparing Models

Feature selection

- Drop customer ID column
- For feature selection used chi2 test
- Gender, Multiplelines, Phone Servies have greater p-value, so we dropped them.
- Top 5 Most Influential Features:
    - Contract
    - TechSupport
    - OnlineSecurity
    - InternetService
    - PaymentMethod

```
customerID: p-value = 0.49439767459438705
gender: p-value = 0.48657873605618596
Partner: p-value = 2.1399113440759935e-36
Dependents: p-value = 4.9249216612154196e-43
PhoneService: p-value = 0.3387825358066928
MultipleLines: p-value = 0.0034643829548773
InternetService: p-value = 9.571788222840544e-160
OnlineSecurity: p-value = 2.6611496351768565e-185
OnlineBackup: p-value = 2.0797592160865457e-131
DeviceProtection: p-value = 5.505219496457244e-122
TechSupport: p-value = 1.4430840279999813e-180
StreamingTV: p-value = 5.528994485739024e-82
StreamingMovies: p-value = 2.667756755723681e-82
Contract: p-value = 5.863038300673391e-258
PaperlessBilling: p-value = 4.073354668665985e-58
PaymentMethod: p-value = 3.6823546520097993e-140
```

Imbalance data Handling

The dataset was imbalanced: fewer churners (Churn = Yes) than non-churners.

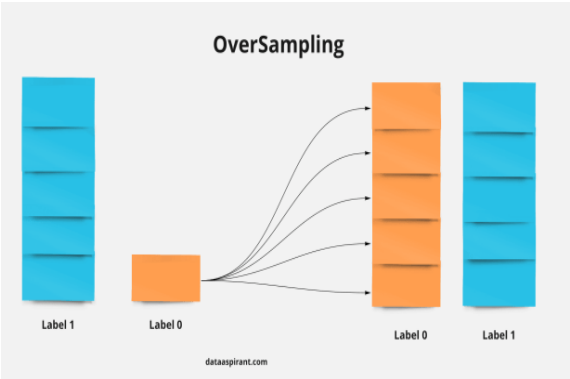
This can make models biased toward predicting "No" most of the time if not handled.

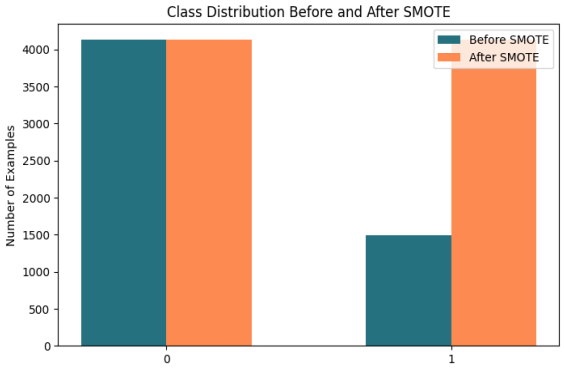
To address this, we applied oversampling techniques to increase the number of churners.

Smote:

SMOTE performs oversampling by creating synthetic samples for the minority class (Churn = Yes) in the feature space.

Applied inside the Pipeline during cross-validation to avoid data leakage.

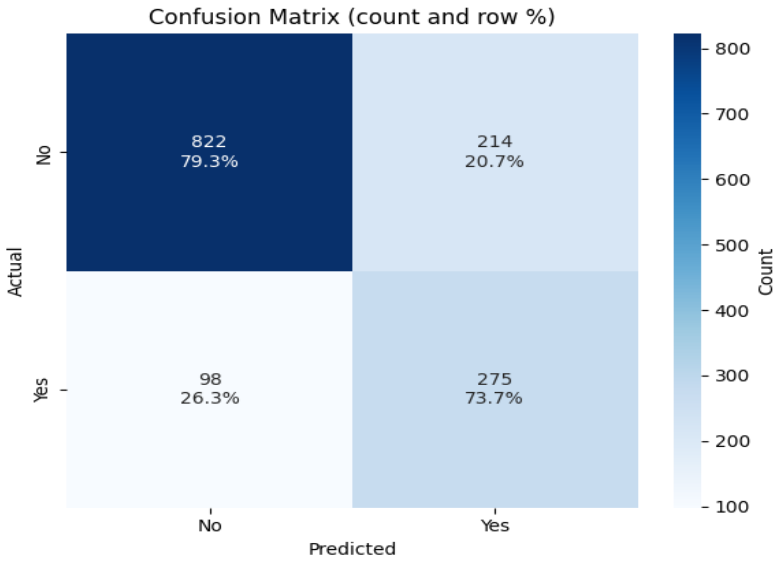


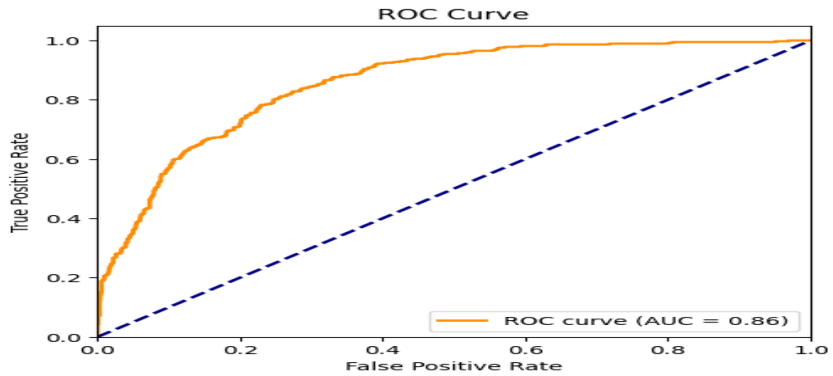


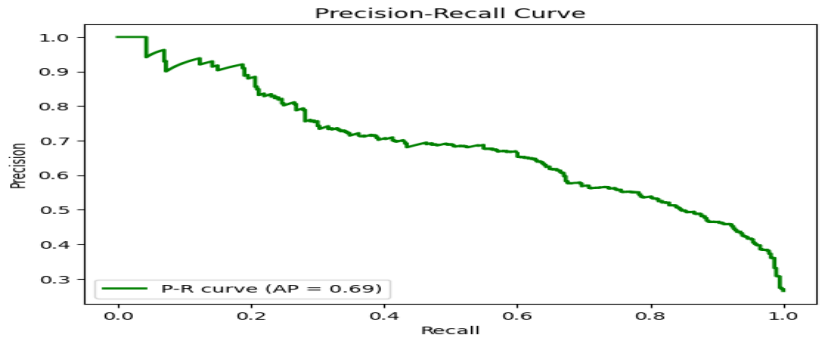
Model Selection & Comparing Models

Random Forest Classifier:

Select Random Forest classifier for model Building







Performance:

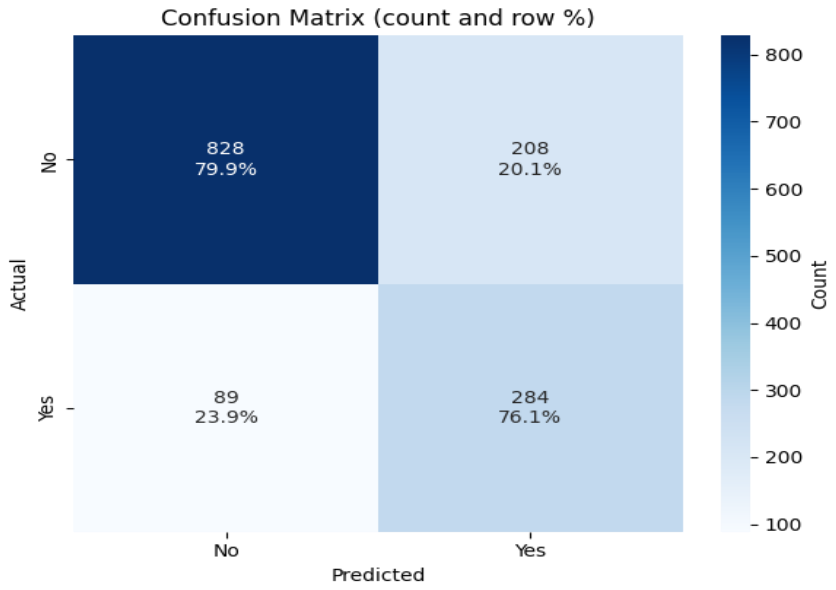
AUC Score (ROC): 0.8562668854223815

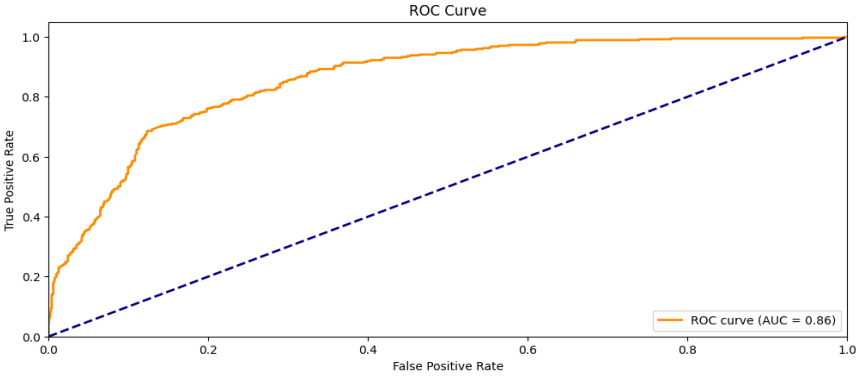
F1 score: 0.6380510440835266

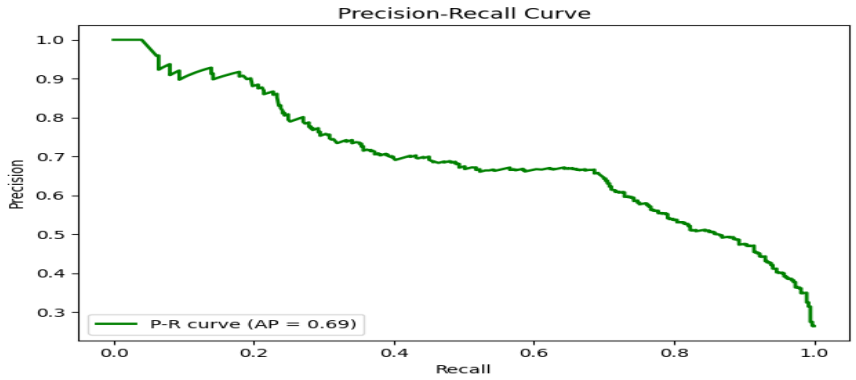
AUC Score (PR): 0.69

XGBoost Classifier

Select XGBoost classifier for model Building







Performance:

AUC Score (ROC):0.8610142122206464

F1 score:0.6566473988439306

AUC Score (PR): 0.69

Choosing the Best Model

After evaluating the two Models, we compared their performance on the test dataset using Accuracy, F1-Score, ROC-AUC, and the Confusion Matrix.

Why **XGBoost** was selected?

1-Highest Overall Performance:

XGBoost achieved the best F1-Score (0.6566) and highest ROC-AUC (0.8610), meaning it balances both accuracy and the ability to correctly identify churners.

2-Better at Detecting Churn Customers:

Compared to Random Forest, XGBoost had a higher recall for churn cases, which is crucial because missing churn customers leads to business loss.

3-Lower False Negatives:

It classified more actual churners correctly, which is the key objective in churn prediction.