Customer Churn Prediction In Telecom Industry

By : Ali-elin

Churn prediction detecting customers who are likely to cancel a subscription to a service. Churn is a problem for telecom companies because it is more expensive to acquire a new customer than to keep your existing one from leaving.

Procedure

- Identify Problem Statements.
- Data Collection
- Exploratory Data Analysis
- Revenue analysis
- Feature Engineering
- Feature Selection
- Handling Imbalance Data
- Model Selection
- Deployment

Problem Statement

- To predict Customer churn
- Factor Affecting Customer Churn
- Revenue analsysis

Data Collection

Data collected from Kaggle.

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age and if they have partners and dependents

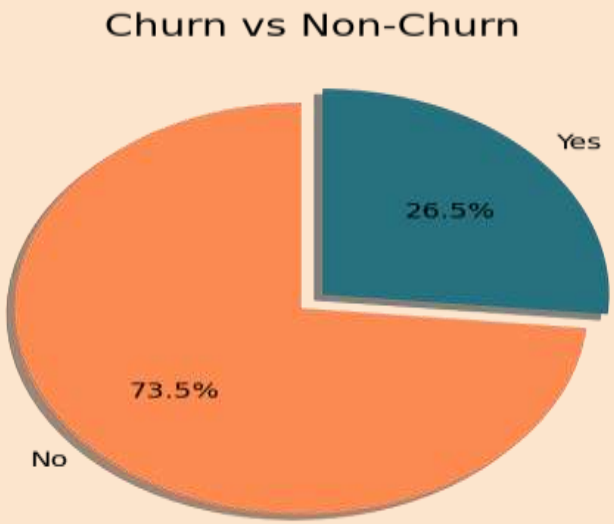
Libraries

- pandas
- numpy
- matplotlib
- seaborn
- scipy
- streamlit
- sklearn
- lifelines
- fastapi
- xgboost
- plotly
- pickle
- os
- Imblearn
- kagglehub
- Requests
- pydantic

EDA

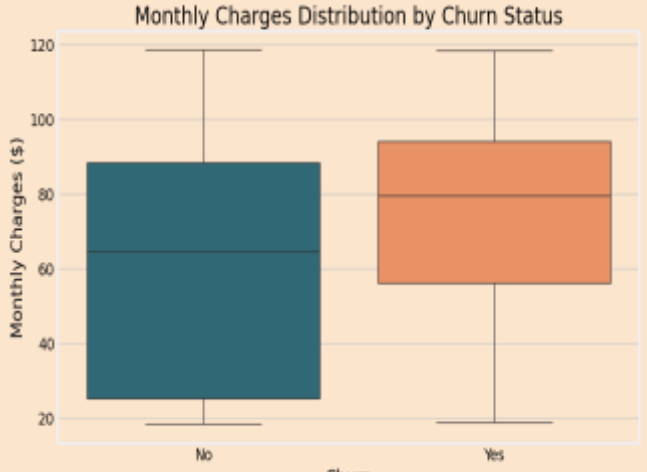
Our analysis shows that:

- Recent clients are more likely to churn.
- Clients with higher Monthly Charges are also more likely to churn.
- Tenure, Monthly Charges, and Contract type appear to be strong predictors of churn, based on the distribution patterns and group differences.

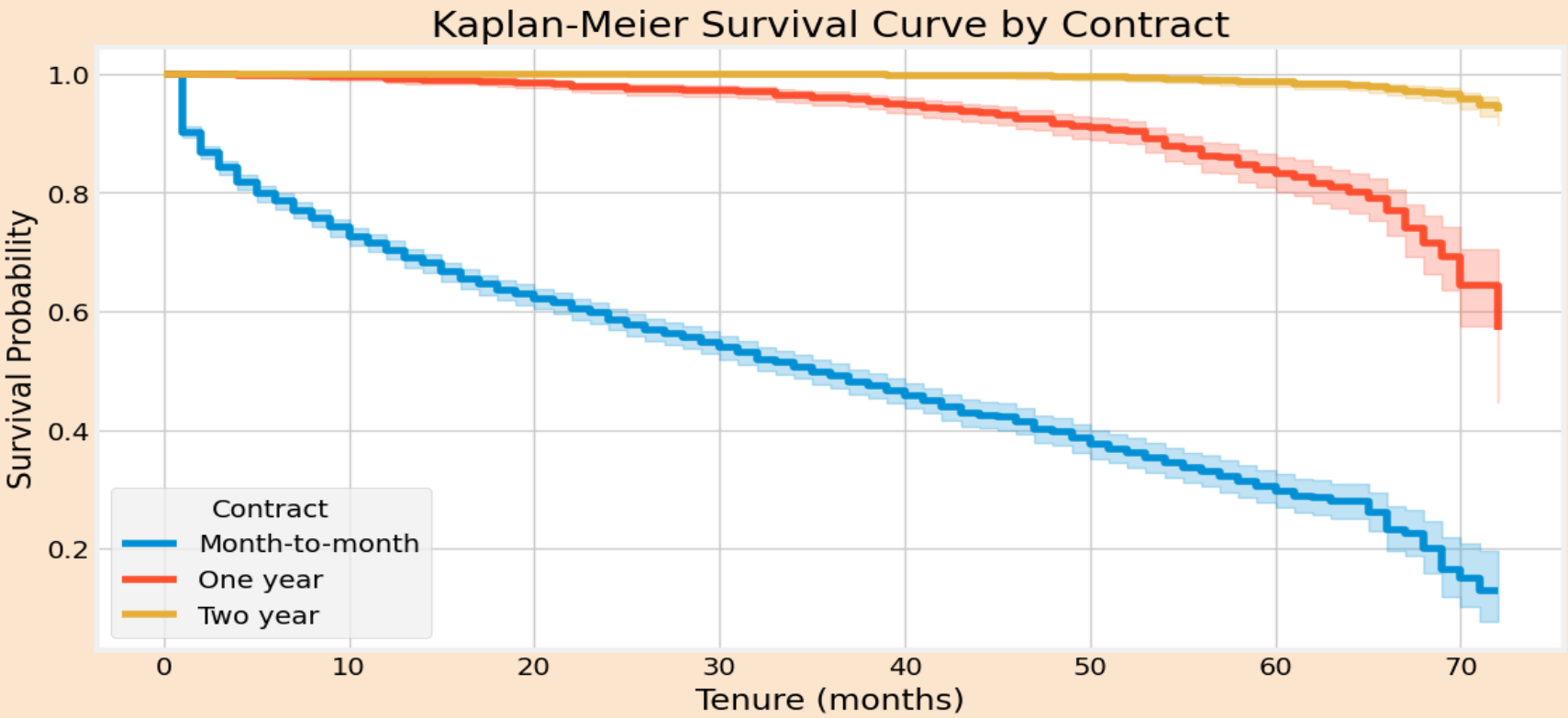


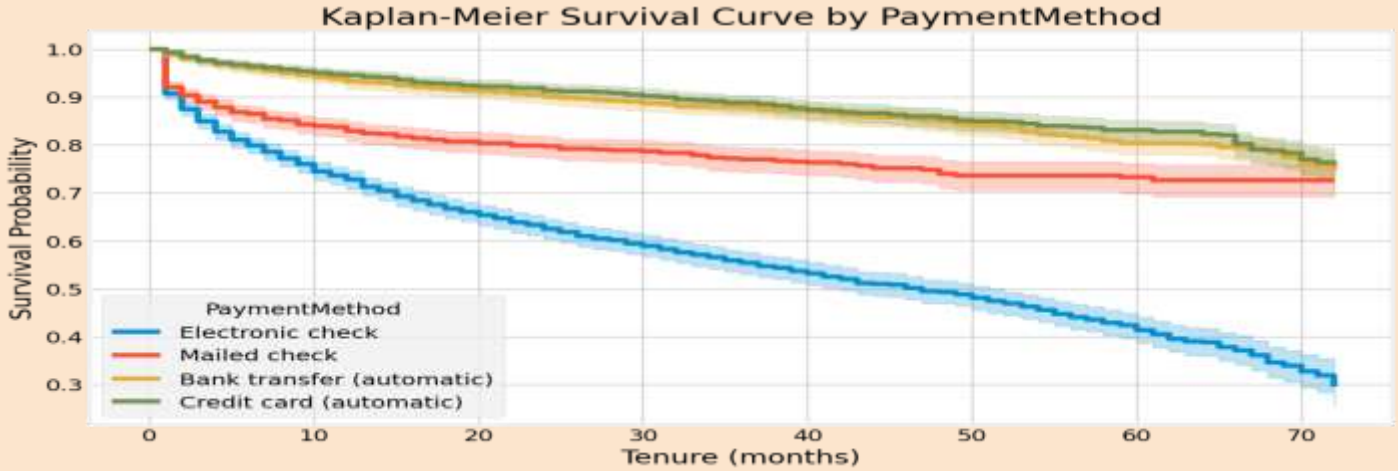
From boxplot

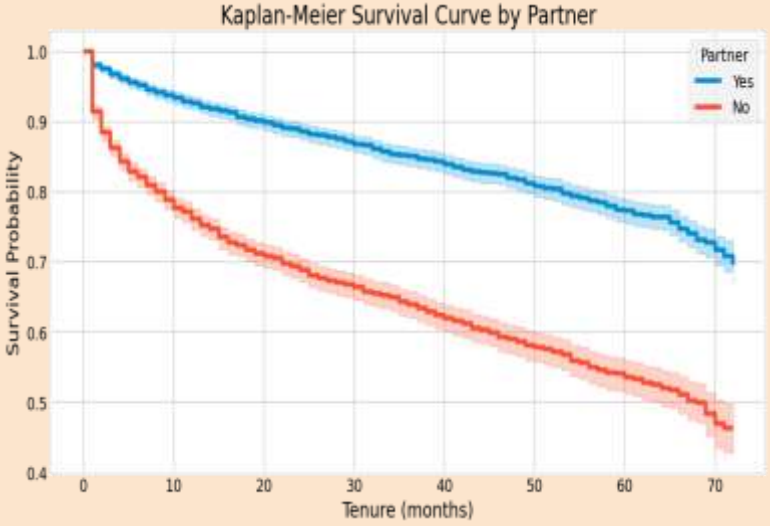
- 75% of churners pay between 60 and 100 dollars a month

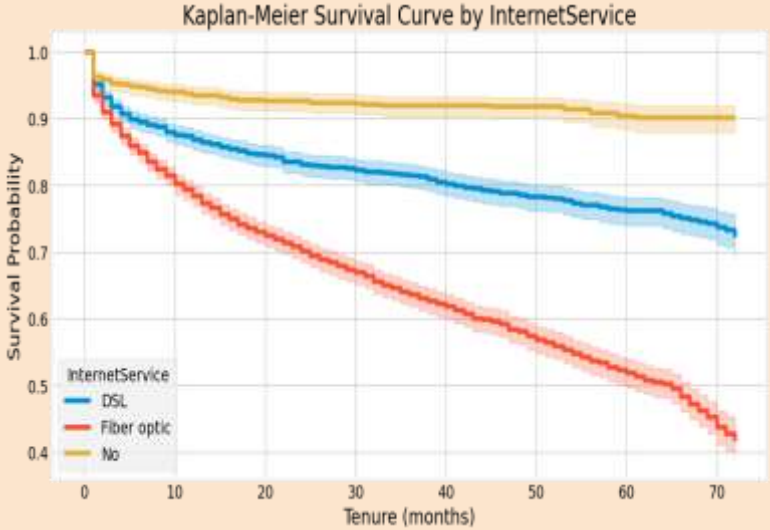


Survival Analysis (Kaplan Meier Curve)

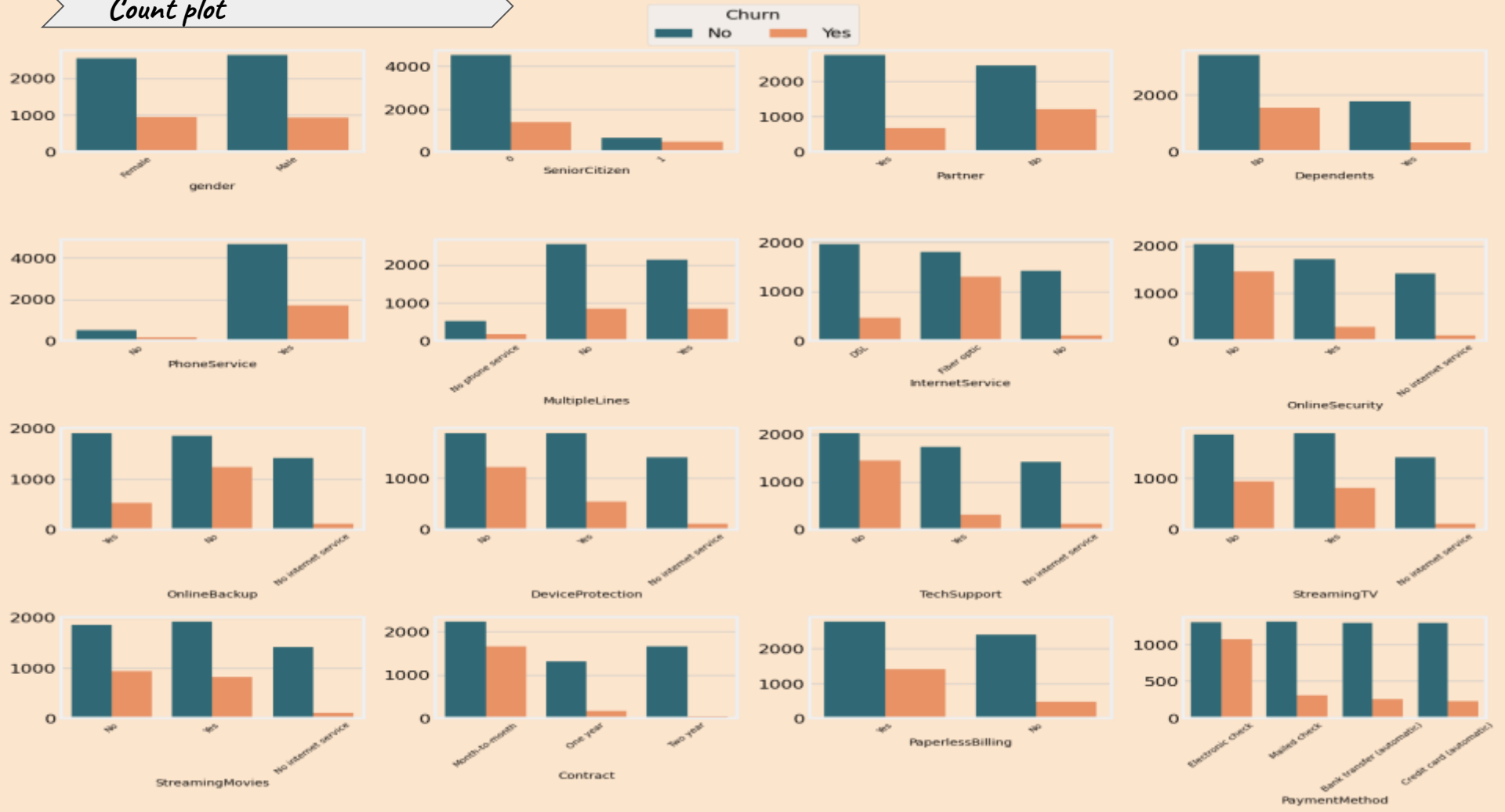



Kaplan-Meier Survival Curve by PaymentMethod
Survival Probability
Tenure (months)
PaymentMethod
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)


Kaplan-Meier Survival Curve by Partner
Survival Probability
Tenure (months)
Partner
Yes
No


Kaplan-Meier Survival Curve by InternetService
Survival Probability
Tenure (months)
InternetService
DSL
Fiber optic
No

Count plot


Churn
No
Yes
gender
Female
Male
SeniorCitizen
Partner
Dependents
PhoneService
MultipleLines
No phone service
InternetService
DSL
Fiber optic
OnlineSecurity
No internet service
OnlineBackup
DeviceProtection
TechSupport
StreamingTV
StreamingMovies
Contract
Month-to-month
One year
Two year
PaperlessBilling
PaymentMethod
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)

- **Gender** - The churn percent is almost equal in case of Male and Females
- The percent of churn is higher in case of **senior citizens**
- Customers with **Partners** and **Dependents** have lower churn rate as compared to those who don't have partners & Dependents.
- Churn rate is much **higher** in case of **Fiber Optic Internet Services.**
- Customers who do not have services like **No OnlineSecurity** , **Online Backup** and **TechSupport** have left the platform in the past month.
- A larger percent of Customers with **monthly subscription** have **left** when compared to Customers with one or two year contract.
- **Churn** percent is **higher** in case of customers having **paperless billing option**.
- Customers who have **Electronic Check** Payment Method tend to **leave** the platform more when compared to other options.

Revenue analysis

- 32% of total revenue was "lost" by people who left.
- Fiber optic is the most profitable service, contributing 53% ($283.28k) of the total monthly revenue.
- DSL contributes 26% ($140.67k) of total revenue.
- 87% of lost revenue comes from Month-to-Month contracts, highlighting that short-term customers churn the most.
- One year and Two year contracts have much lower loss percentages (15% and 4% respectively), showing that long-term contracts retain revenue better.

Feature Engineering

Cheak Outliers

Checked outliers by IQR method no outliers detected.

Missing Values Summary

During data checking, we found that almost all-important numerical fields were complete. Only the **TotalCharges** column had a small number of missing values. These were cleaned and replaced with valid numbers to keep the dataset consistent and ready for modeling.

New Feature

A new feature was added to mark customers who stayed longer than 12 months. This helps us better understand differences in behavior between short-term and long-term customers.

Feature Importance for chi2 test

```
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
```

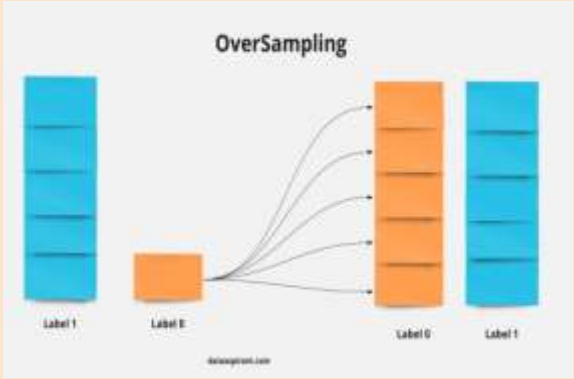
Feature selection

- ❑ **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

Imbalance data Handling

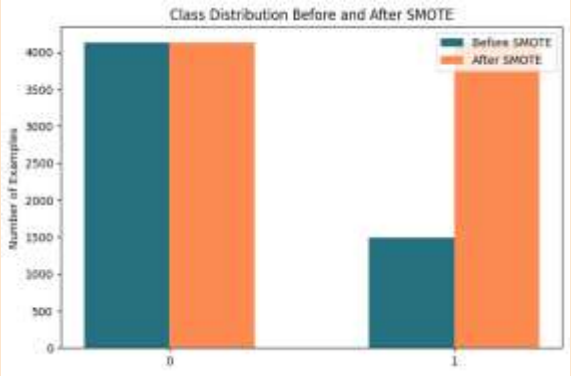
The dataset was imbalanced: fewer churners (Churn = Yes) than non-churners.
This can make models biased toward predicting "No" most of the time if not handled.
To address this, we applied oversampling techniques to increase the number of churners.



Smote

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

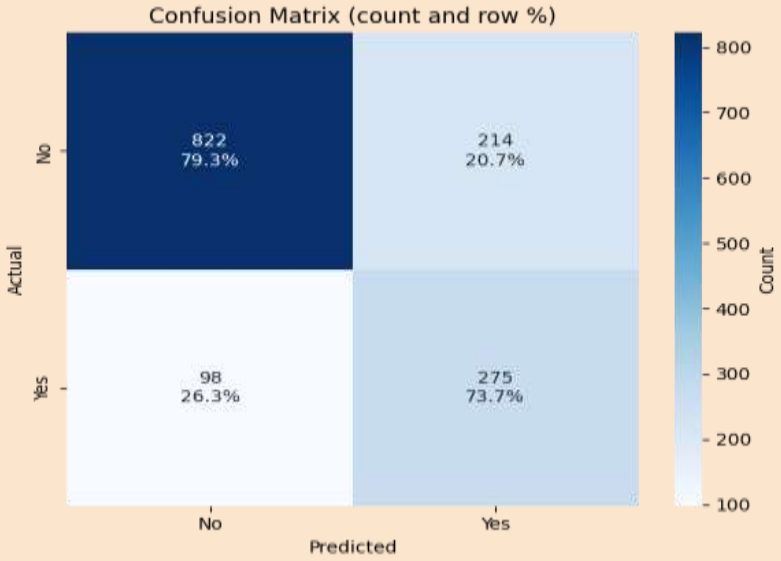
Applied inside the Pipeline during cross-validation to avoid data leakage.

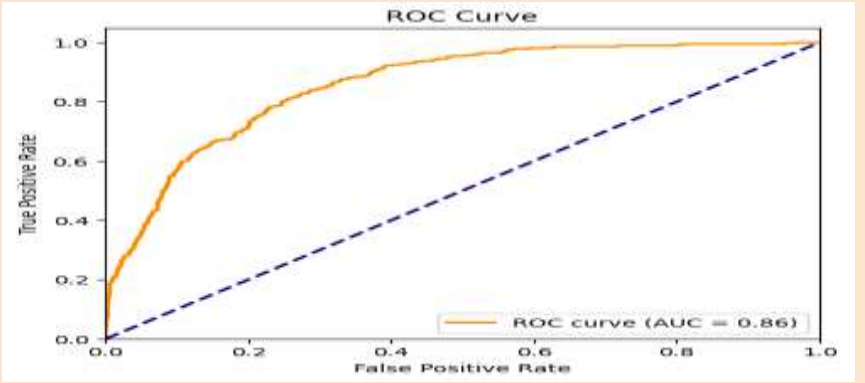


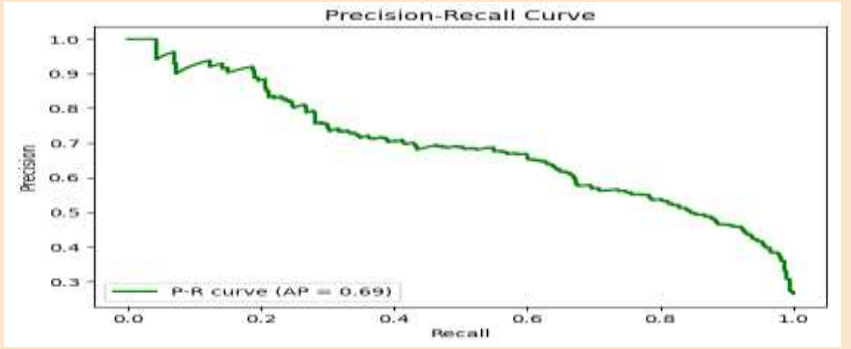
Model Selection&Comparing Models

Random Forest Classifier

Select Random Forest classifier for model Building







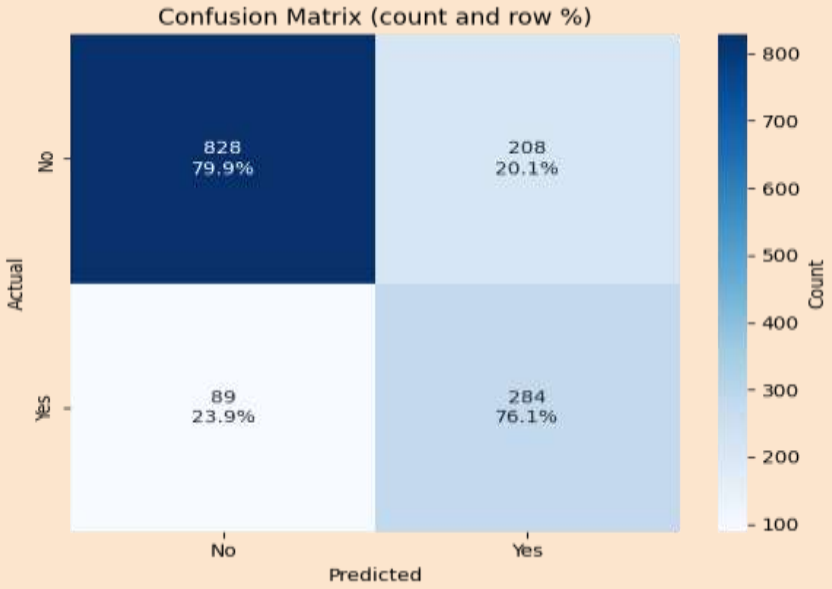
AUC Score (ROC): 0.8562668854223815

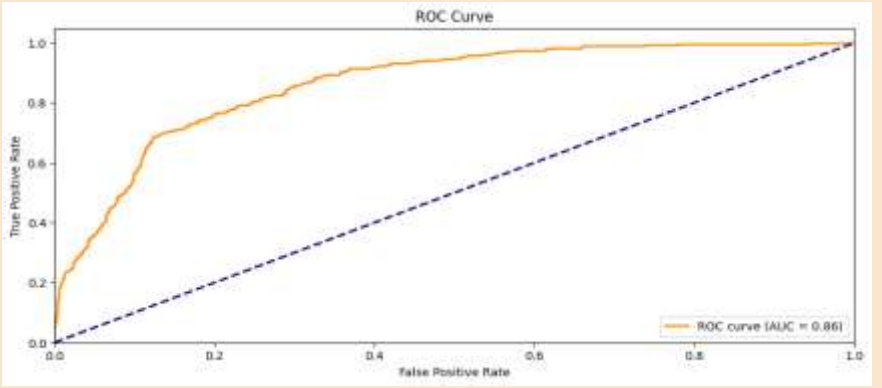
F1 score: 0.6380510440835266

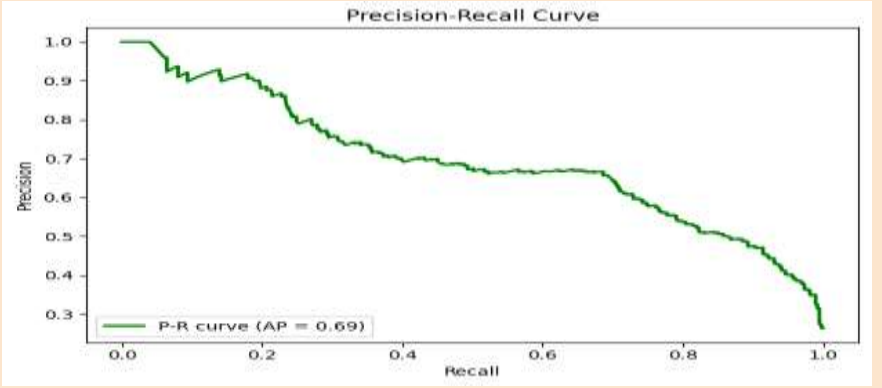
AUC Score (PR): 0.69

XGBoost Classifier

Select XGBoost classifier for model Building







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.

Deployment Summary

Model API Server:

- Built using FastAPI.
- Purpose: Receive customer data from the UI and return churn prediction and probability.
- Endpoints tested locally to ensure correct responses.

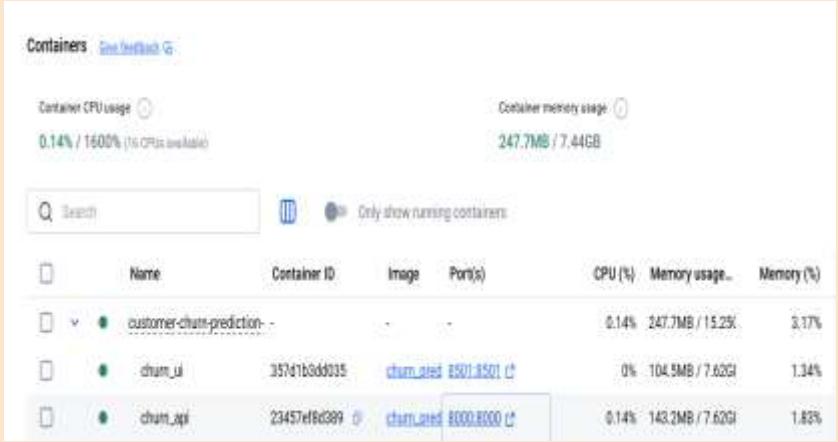
UI Dashboard:

- Built using Streamlit.
- Predict button to request prediction from the API.
- Dynamic gauge showing churn probability directly under the title after prediction.

Deployment Summary(Cont..)

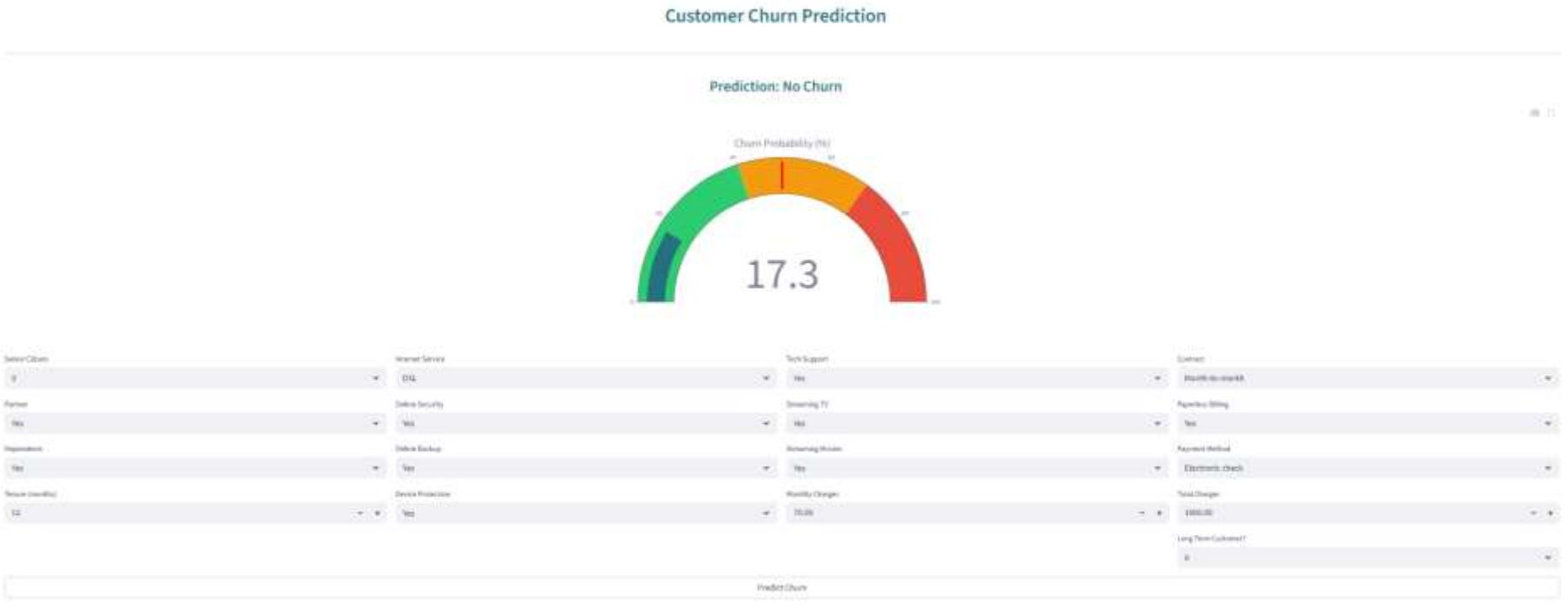
Docker:

- Two separate Docker containers
- API Container – contains FastAPI + trained model.
- UI Container – contains Streamlit dashboard.
- Ensures separation of frontend and backend and easy deployment on any environment.



Environment Management:

- Project environment managed with Poetry.
- All dependencies for model and UI are tracked for reproducibility.



Github - Link
Go to web App - Link

Reduce Telecom Customer Churn & Conclusion

- Use the App to predict churn probabilities
- Engage with your customers
- Define your most valuable customers
- Give better service
- Pay attention to complaints
- Make your best people deal with customers at risk
- Flaunt your competitive advantages
- Offer long term contracts


Thank you