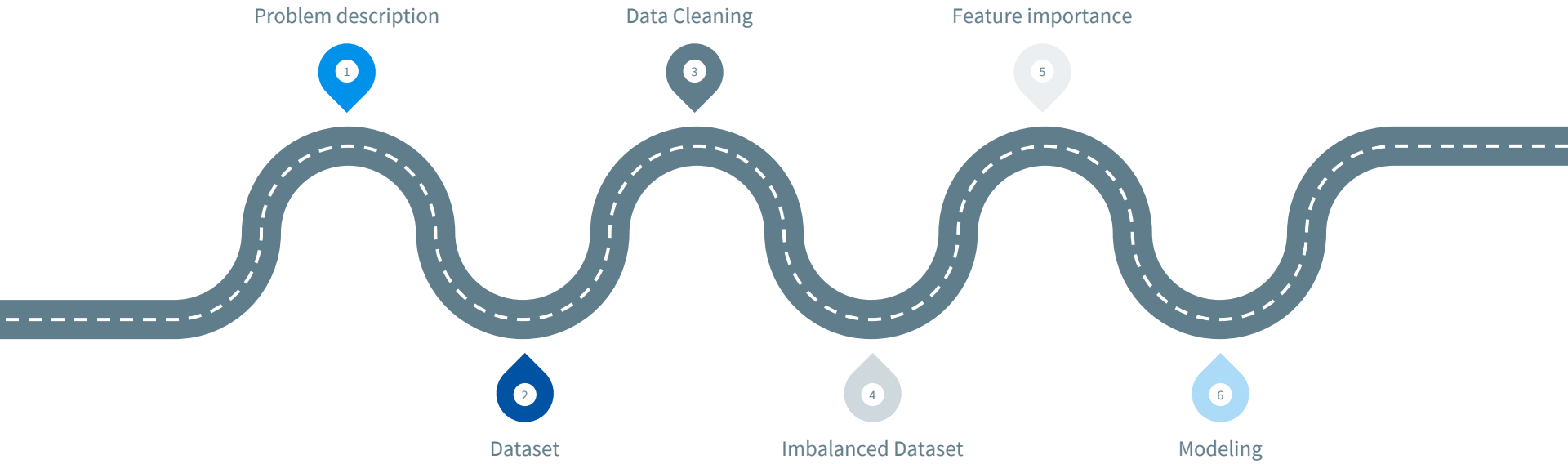
Predict customer chur using classification model

Roadmap


Problem description
1
Data Cleaning
3
Feature importance
5
2
Dataset
4
Imbalanced Dataset
6
Modeling

Problem description

Customer churn is one of the most vital data points for businesses to track. Customer churn analysis helps you identify key stages in the customer journey where people are falling off, allowing you to pinpoint specific strategies to improve their interactions with your brand and improve brand loyalty.

Dataset

kaggle

7043

Rows

21

Columns

Dataset

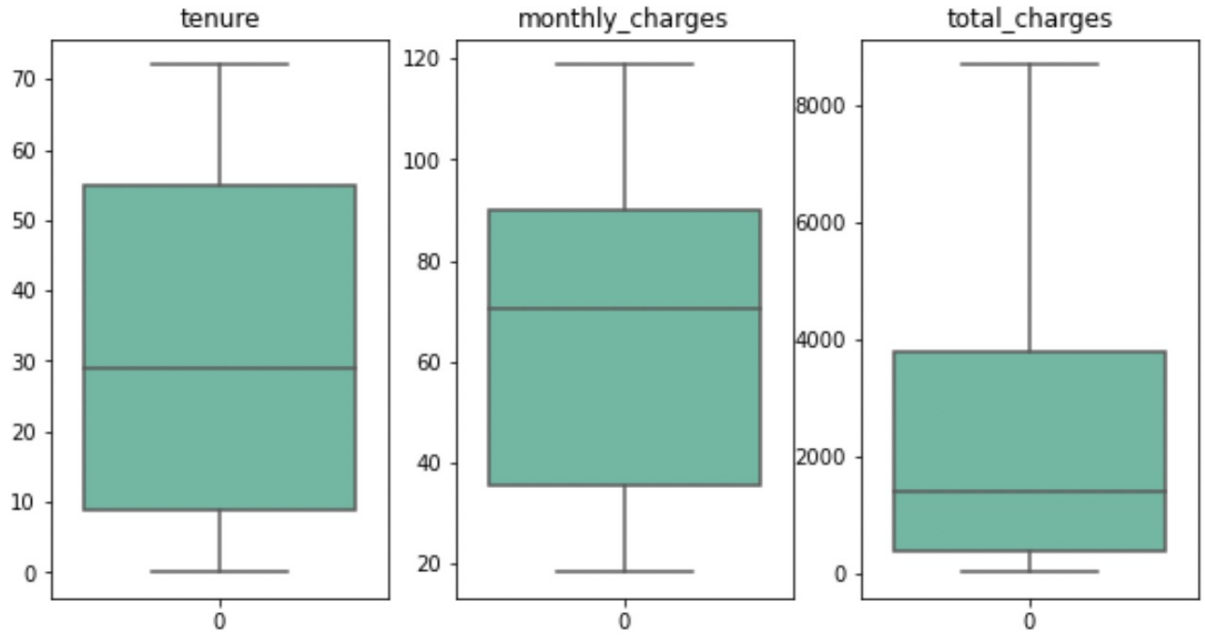
Train set: 55%

Validation set: 25%

Test set: 20%

1. Data Cleaning

1. Outliers


tenure
monthly_charges
total_charges
70
60
50
40
30
20
10
0
120
100
80
60
40
20
8000
6000
4000
2000
0
0
0
0

2. Null Values

1. Split the dataframe to train and validation
2. Assign mean to null values in the train

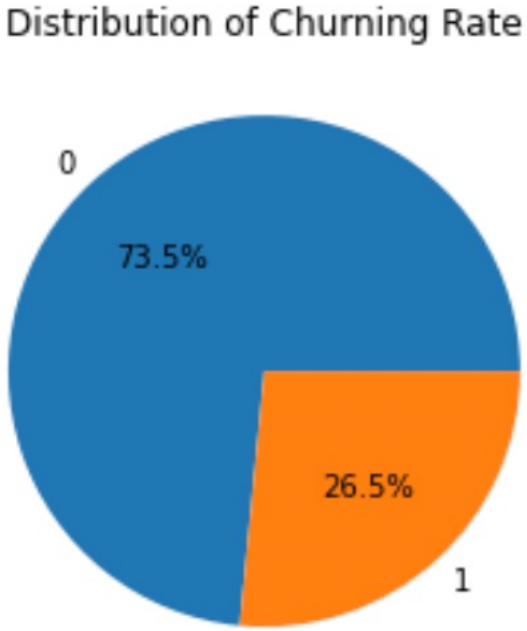
3. Drop duplicate rows

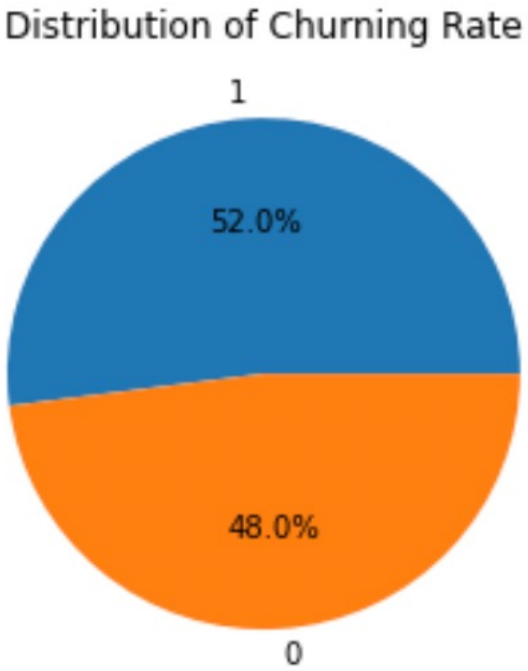
4. Change columns data type

5. Drop unique columns

2. Imbalanced Dataset

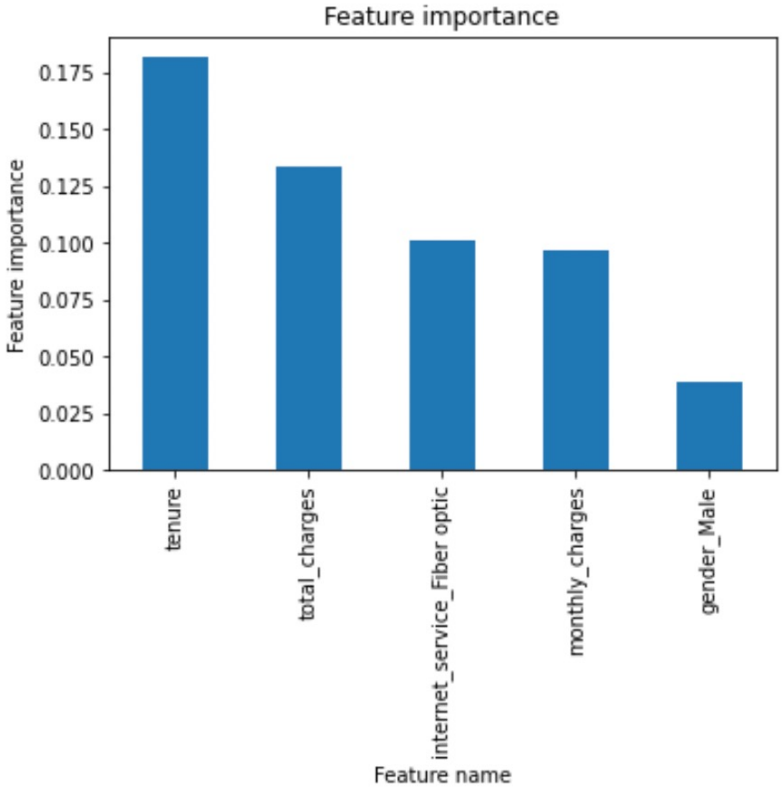
Over-sampling (by Random over sampling)





3. Feature Importance

Feature Importance



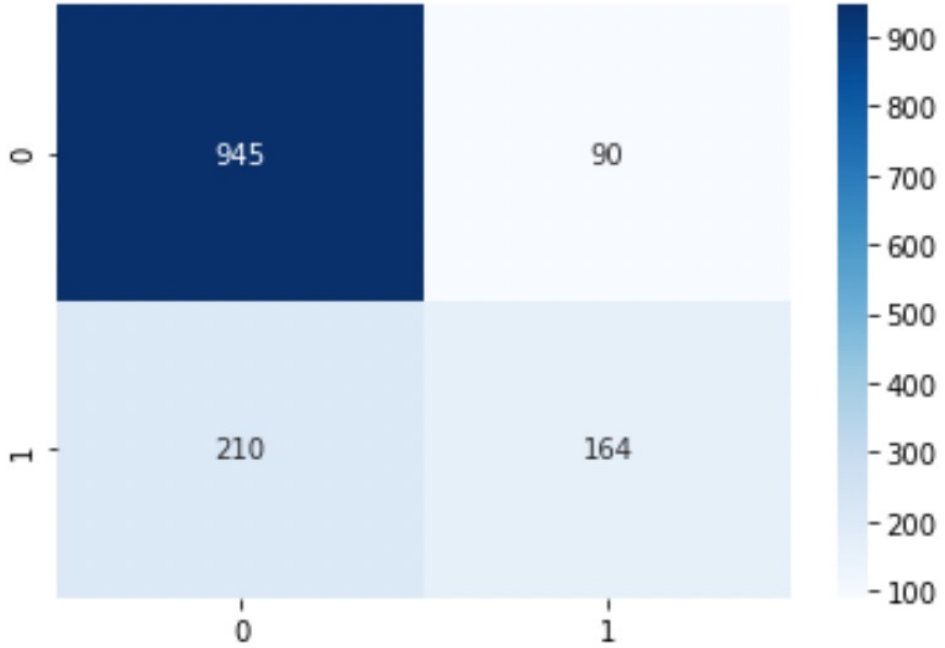
4. Modeling

1. Logistic Regression (baseline)

Accuracy: 79%

Recall: 43%

Precision: 64%

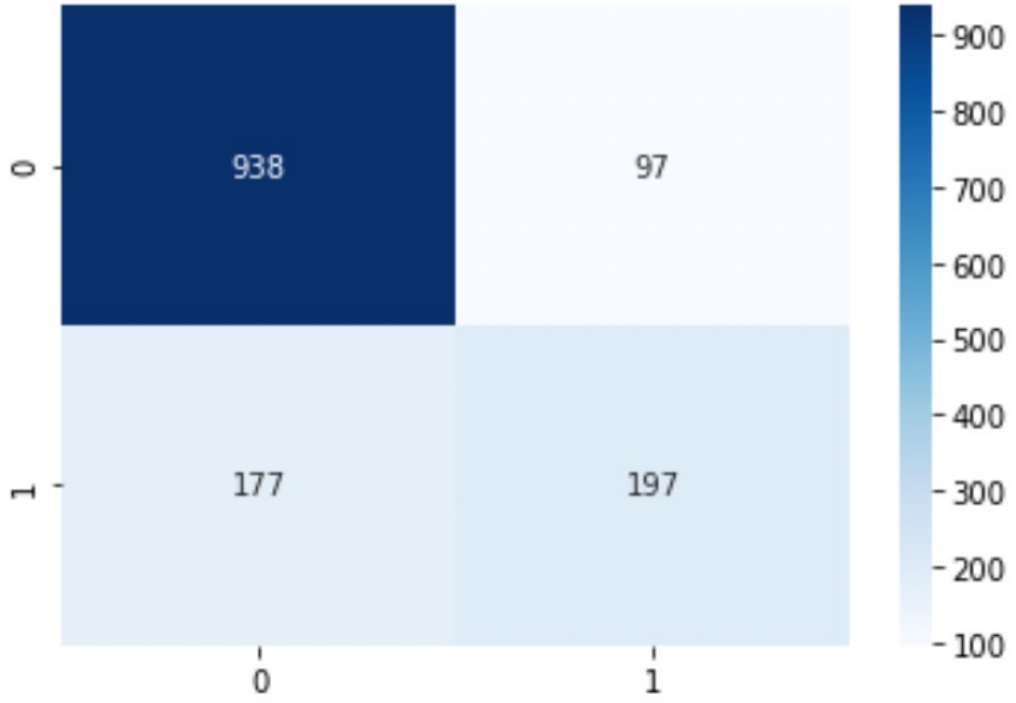


1. Logistic Regression (After Feature engineering)

Accuracy: 81%

Recall: 52%

Precision: 67%

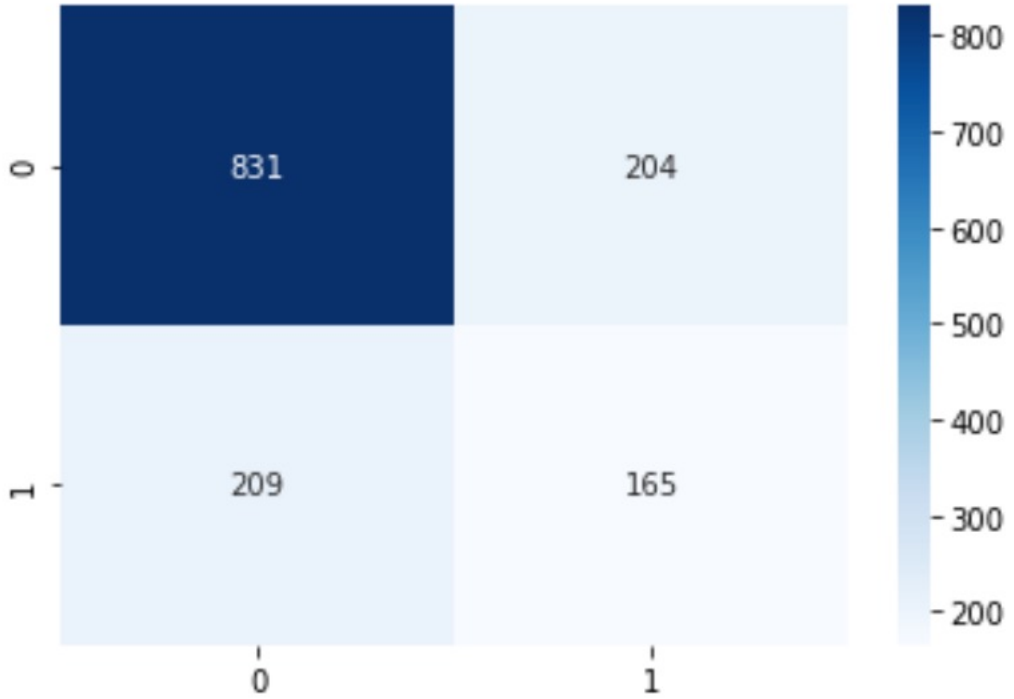


2. k Nearest Neighbors

Accuracy: 71%

Recall: 44%

Precision: 44%

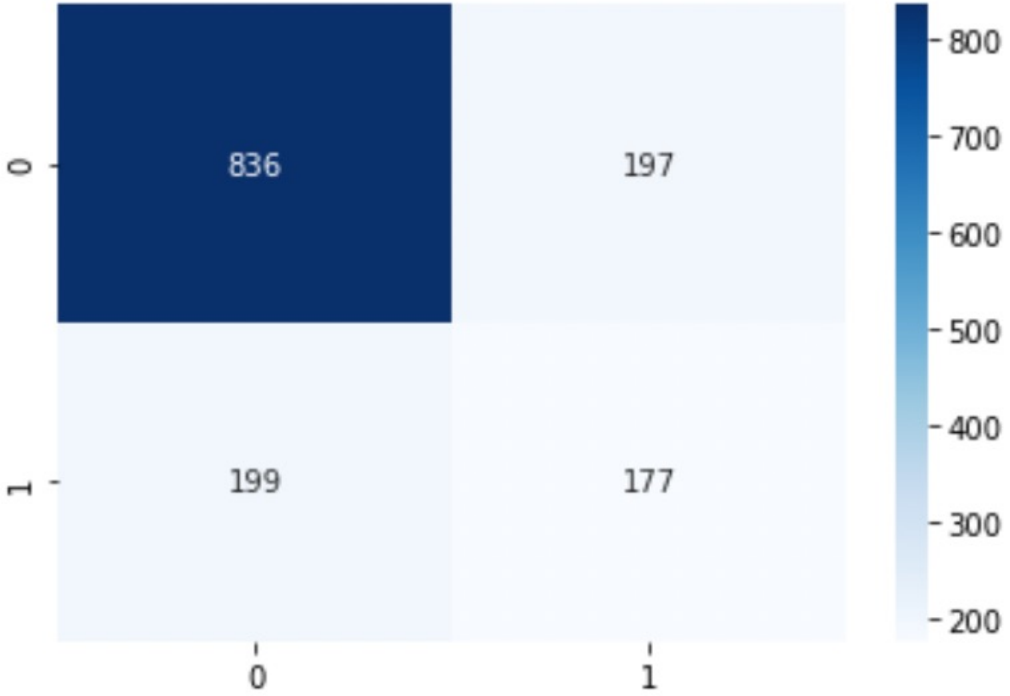


3. Decision Tree

Accuracy: 71%

Recall: 49%

Precision: 48%

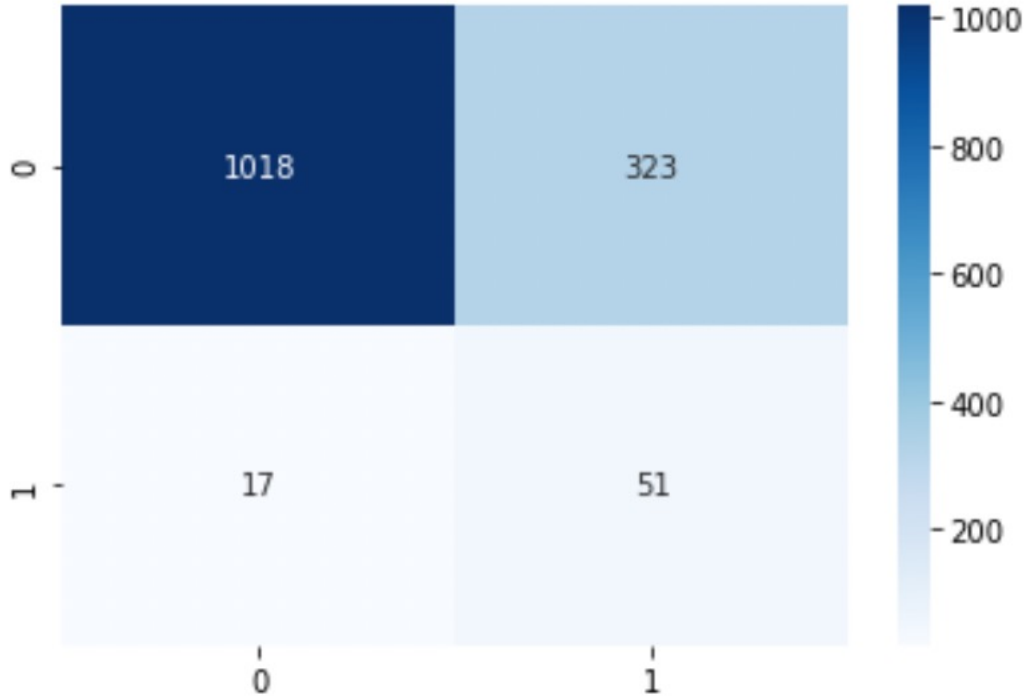


4. Random Forest

Accuracy: 81%

Recall: 13%

Precision: 75%

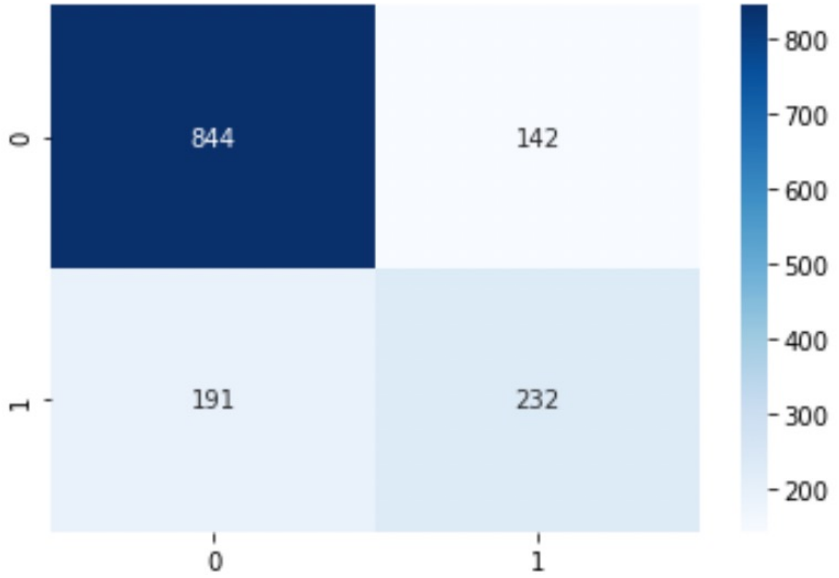


5. XGBoost (The best)

Accuracy: 81%

Recall: 62%

Precision: 54%



Thanks!

Any questions?