

# Churn Prediction Model

Emirhan Kayar, 513222

# Introduction

# Business Context

As given in the case study, problem is increasing churn rate. It is an indicator of the business experiencing a reduction in the customer purchases over time.

This trend has a negative impact concerning:

- AOV - Average order value - fewer purchases repeated

- CLV - Customer lifetime value - long-term value of each client declines

- Revenue Growth - slowing, due to reduced repeats

As for the analysis will be performed, ways to prevent customers from churning must be assessed and provided.

# Objectives

- Perform Explorative Analysis & Feature Selection

Inspection and understanding of the data

Followed by preprocessing and statistical analysis

- CHURN Model Development / Optimization

- XAI and Interpretation of the results

# Dataset

# Dataset

Consists of 5 categorical, 12 numerical features, excluding the target “Churn” and the “CustomerID”, total of 5630 entries

For this dataset, it is not necessary to define the time period since the features containing temporal information are already calculated, so it could be said that the time period is already predefined.

Dataset has class imbalance

Results obtained (post-preprocessing) - 0: 3739 (83.2%) , 1: 757 (16.8%)

In order to assess features to be used for training, statistical methods must be performed on the columns, so that significant values are filtered for the model.

# Key Variables

According to t and chi-square tests, most significant (p-value < 0.05) variables:

#	Feature	Test	P-Value
0	Tenure	T-test	4.000206e-150
1	Complain	T-test	4.644857e-81
2	PreferredOrderCat	Chi-square	6.797507e-60
3	MaritalStatus	Chi-square	9.650900e-41
4	DaySinceLastOrder	T-test	4.342190e-32
5	PreferredLoginDevice	Chi-square	1.358165e-16
6	NumberOfDeviceRegistered	T-test	6.333351e-16
7	SatisfactionScore	T-test	2.795835e-15
8	PreferredPaymentMode	Chi-square	7.685798e-15
9	CityTier	T-test	1.801763e-10
10	WarehouseToHome	T-test	2.284650e-07
11	NumberOfAddress	T-test	9.504937e-04
12	Gender	Chi-square	2.617454e-02

# Methodology



# Model Methodology

- 1 - Train-Test split (80/20%)
- 2 - SMOTE to handle imbalance
- 3 - XGBoost (Extreme Gradient Boosting) Model
- 4 - Hyperparameter Optimization - Bayesian, CV with 10 folds
- 5 - Plot the best fit model results

## Best Hyperparameters

```
'colsample_bytree': np.float64(0.659),  
'gamma': np.float64(0.159),  
'learning_rate': np.float64(0.149),  
'max_delta_step': np.int64(4),  
'max_depth': np.int64(12),  
'min_child_weight': np.float64(2.0),  
'reg_alpha': np.float64(0.051),  
'reg_lambda': np.float64(0.740),  
'subsample': np.float64(0.736)
```

# Interpretation (Results & XAI)

# Model Results

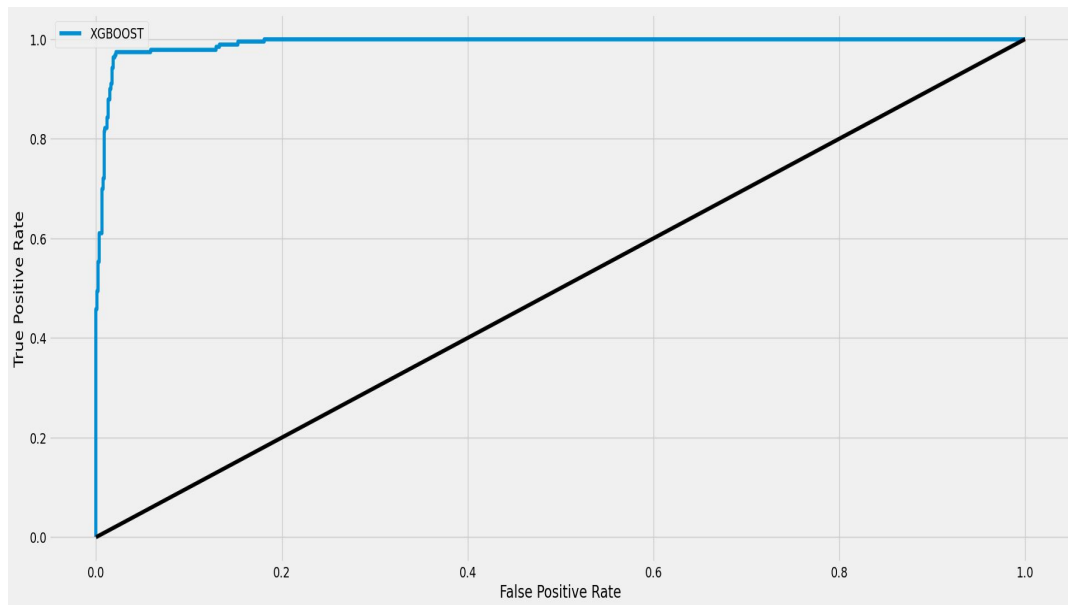


Figure 1

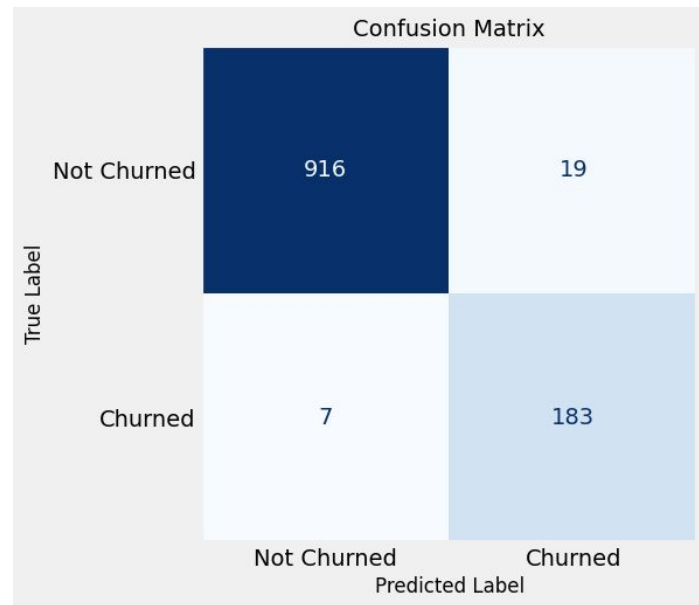


Figure 2

# Model Results

Referring to Table 1 and Table 2, accuracy might be misleading, other scores must be taken into consideration, but yet 98% is more than enough.

Also considering recall and precision, it is clear that model performs well on both classes on average, except for class “Churned” with precision score 91%, however, it is still a good score.

Even though there is class imbalance, proposed model has an excellent performance with F1 score, which is 98%.

AUC (Figure 1), graph showcases that, the model is able to generalize on a very high percentage (97%).

Confusion matrix (Figure 2) indicates that, the model is able to distinguish between the two classes almost perfectly.

roc_auc	recall	precision	accuracy
0.97	0.96	0.91	0.98

Table 1

	Precision	Recall	F1-Score	
Support				
Not Churned	0.99	0.98	0.99	935
Churned	0.91	0.96	0.93	190
accuracy			0.98	
1125				
macro avg	0.95	0.97	0.96	
1125				
weighted avg	0.98	0.98	0.98	
1125				

Table 2

# XAI

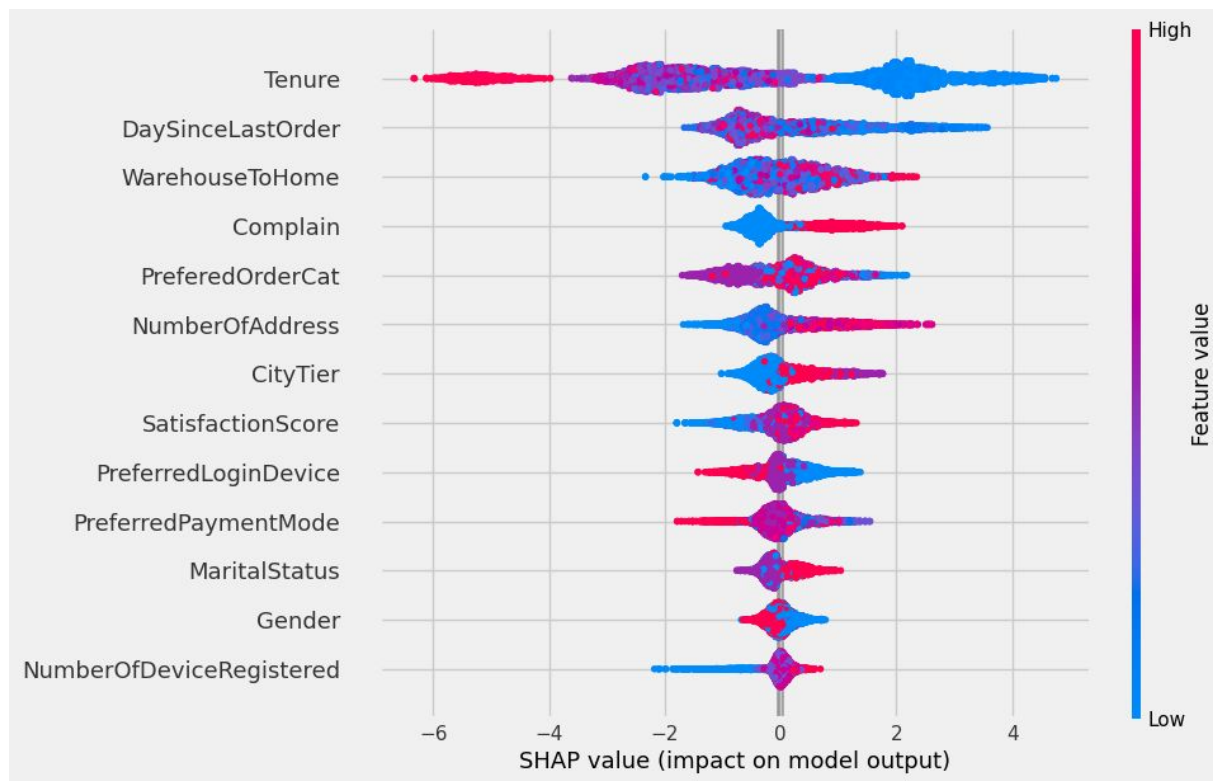


Figure 3

# XAI

Figure 3 explains, which features were the most important for model to make such decisions.

Tenure - Low tenure has higher contribution which indicates that it strongly increases churn, however high tenure is inversely proportional to low tenure such that it has less contribution. Loyal customers are less likely to churn.

Day Since Last Order - Has highly mixed effects, but, it is clear that lower amounts increase churn risk. This could suggest, if a customer orders frequently it does not mean the risk of churning is gone.

Warehouse to Home - Mixed as before but, overall, high values contribute positively, which indicates, if the warehouse is far from where the customer lives it is a reason to churn.

Complain - Intuitive and self explanatory, amount of complaints and churning risk is directly proportional.

Preferred Order Category - Can be said that it is the fourth most important feature in our table, but it is not highly interpretable since we can not understand what category impacts the decision mostly

Number of Addresses - Higher amounts have higher risk to churn, customers who has more addresses (they might be more active) added is more likely to churn.

The other features are not as impactful as the first six, for them overall, it can be said that, they are moderately impacting the decision making.

# Insights

# Recommendations

From what is observed from the explainability and the interpretation it is suggested that the company needs to consider,

- 1 - Focus on new customers, campaigns could be provided, discounts could be applied, some coupons could be given etc.
- 2 - Frequent orders does not mean loyalty, company should focus on keeping those who buy recently.
- 3 - Since distance from warehouse increases and churning risk increases, company should focus on providing better shipment service.
- 4 - Complaints always must be considered, maybe customer support could be added if it does not exist. If it exists, it must be fixed.
- 5 - Preferred order category must be tracked in order to understand what category of product causes the churn.
- 6 - When looking into number of addresses registered and frequent orders, we see that the company does not care about its most active customers, more accounts and frequent order means active users. However, usually the active users are churning.
- 7 - another indicator to the previous point, more devices the user has is more likely to churn again an indication of



Thank you, for attention.