Data Challenge
Customer Churn Prediction: A Machine Learning Classification Approach

By: Data Porter

Hi, we're Data Porter!

Alfian Syach

S1 - Management

Chantika Yahya

S1 - Management

Devina

S1 - Environmental Engineering

Report Content

Get to know the company

PT Telkom Indonesia (Persero) Tbk (Telkom) is a leading State-Owned Enterprise specializing in information and communication technology (ICT) services and telecommunications networks in Indonesia. Currently undergoing a transformation into a digital telecommunications company, TelkomGroup is implementing a customer-oriented strategy to become more agile and efficient.

Telkom Indonesia

the world in your hand

What problem is the company facing?

Customer Churn

What is it?

Customer churn is the percentage of customers who stopped using your company's product or service during a certain time frame.

What caused it?

Customer churn is often caused by factors like poor customer service, product dissatisfaction, competitive offerings, or changes in customer needs.

How to prevent it?

By using data analytics and predictive modeling, companies can proactively identify at-risk customers and take measures like improving customer service, addressing product issues, offering loyalty programs, and maintaining open communication to retain them.

What if we don’t handle this problem?

- Diminished customer base
- Churn rate escalation
- Immediate financial loses
- Negative impact on business sustainability
- Eroded customer trust
- Tarnished brand reputation
- Challenging competitive maintenance

PROBLEM

In the last quarter of the company's operation, customer churns has been occurring and reached

26.54%

GOALS

Develop a predictive model to proactively identify high-risk churn customers, enabling us to implement retention strategies effectively.

OBJECTIVES

- Key Churn Factors Identification.
- Build a predictive model.
- Feature Importance Analysis.
- Churn rate and predictive model performance monitoring and evaluation.

METRICS

Business:

- Churn Rate

Machine Learning Model:

- F1 Score
- Area Under ROC Curve
- Log Loss

Tools Used

Environment

Programming Language

Version Control

Web App Creation

Let's see our dataset

Categorical Features:

1. Location
2. Device Class
3. Games Product
4. Music Product
5. Education Product
6. Call Center
7. Video Product
8. Use MyApp
9. Payment Method
10. **Churn Label**

Numerical Features:

1. Tenure Months
2. Monthly Purchase (Thou. IDR)
3. CLTV(Thou. IDR)

Additional Features:

1. Customer ID
2. Longitude
3. Latitude

"Telco_customer_churn_adapted_v2"

- **18 features**
- **7043 rows of data**

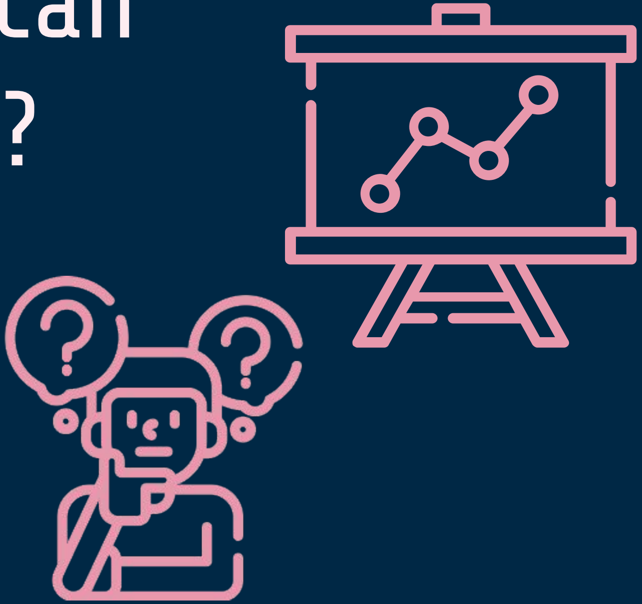
The numerical features are skewed.

No outliers.

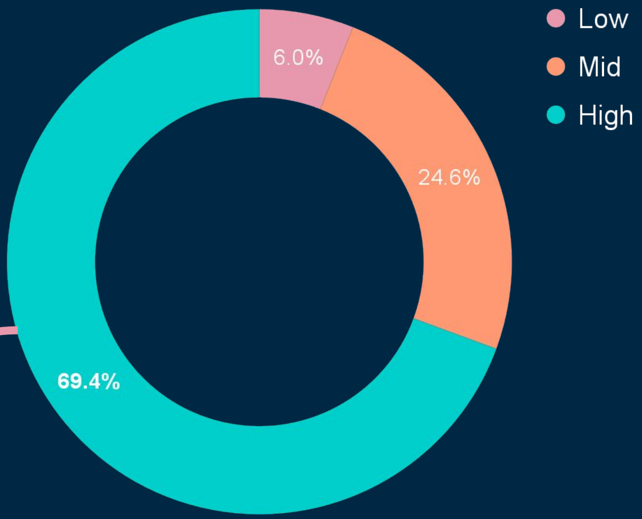
0% duplicate data.

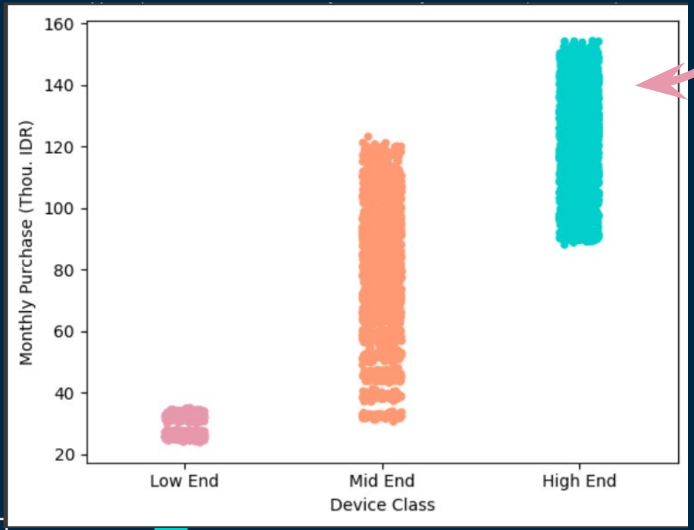
0% missing values.

What business insights can we get from the data?



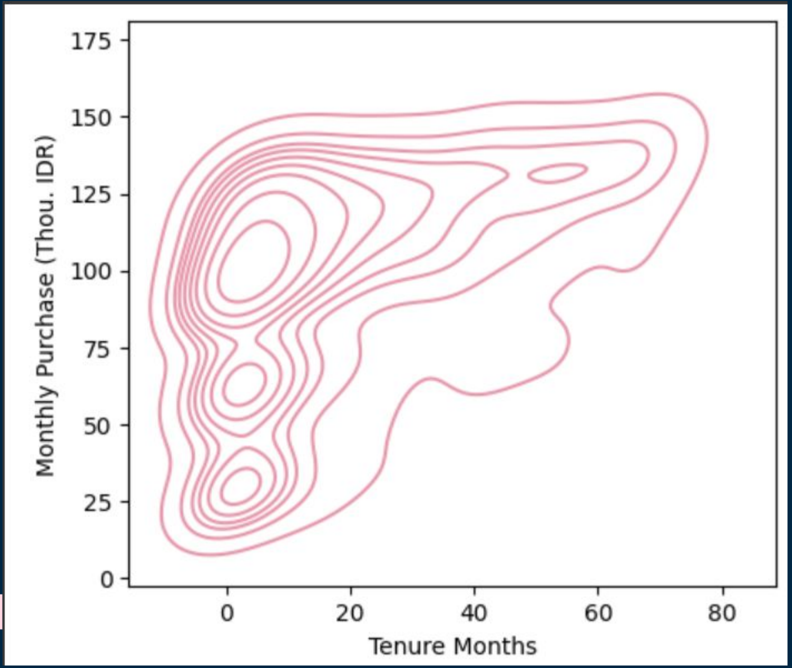
Let's begin by examining which device class experiences the highest churn rate.





It appears that the majority of churn customers belong to the high-end device class, indicative of the highest spending group.

Now, let's explore the duration of customer tenure before churn and their monthly spending patterns!

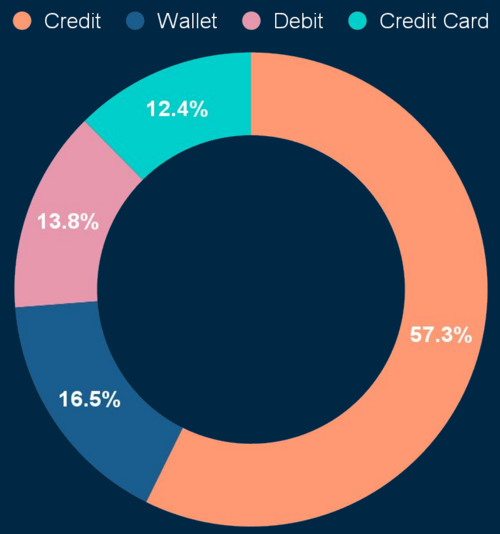


How many months, on average, have churned customers been with the company, and what is the range of their monthly spending?

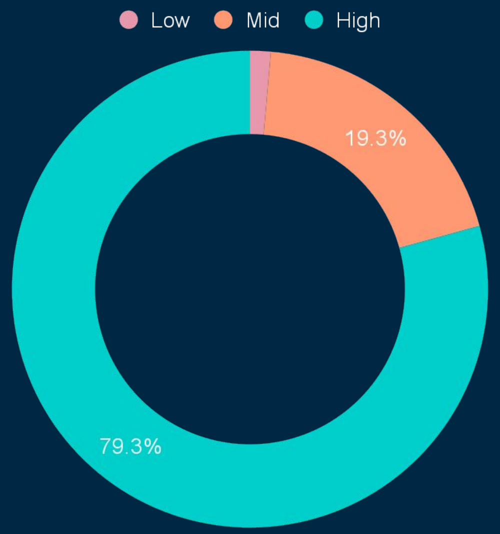
The graph illustrates a concentrated density in the upper-left region, suggesting that churn customers tend to be new customers with high spending habits a characteristic typically associated with the high-end device class..

There is a possibility that it might be caused by the customers' dissatisfaction with the products or services offered.

What payment methods does the churn customers used?



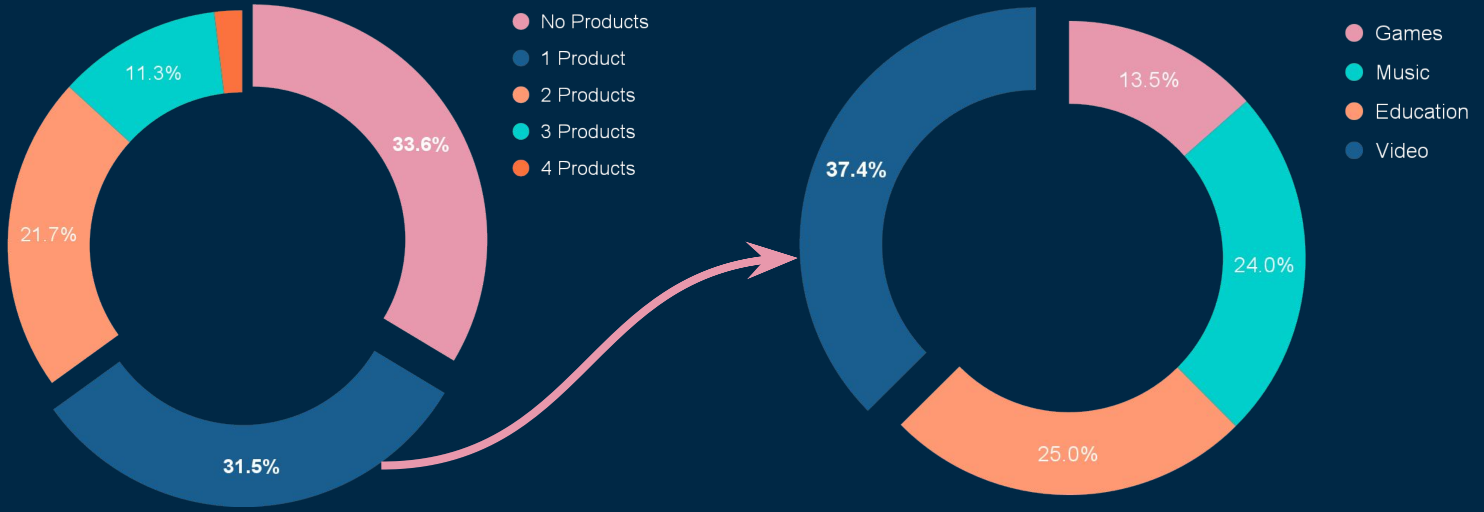
What's the majority of class device use credit as the payment method?



Most of the leaving customers are using the credit payment method. Interestingly, most of them are customers with high-end devices.

Further information is required to analyze the causality of this pattern, and it might be achieved by getting additional customer data, such as their age.

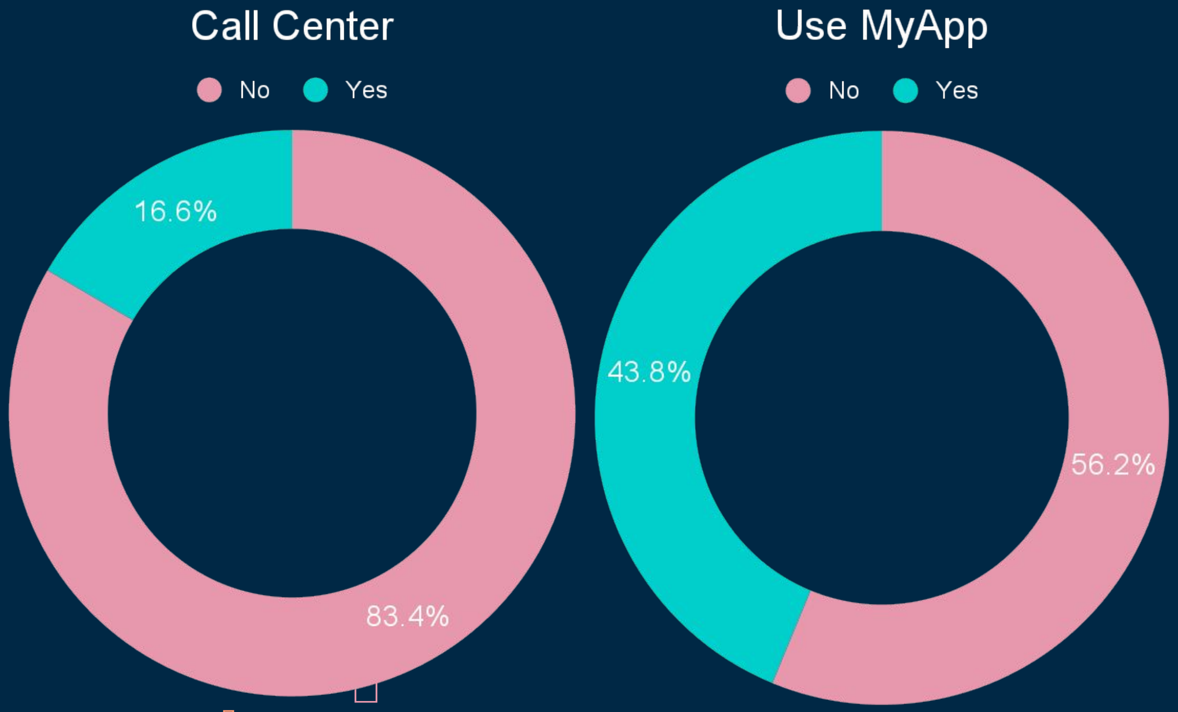
What is the number of products used by churned customers, and which product stands out as their most frequently used?



In the product usage distribution among churn customers, a significant portion did not subscribe to any product. However, for those who did, the video product emerged as the most preferred choice.

- There's a possibility that customers were initially attracted to our company due to the video product we offer, but it seems they may have been dissatisfied with their experience.

Did the churn customers ever called the call center or use MyApp?



The majority of departing customers never engaged with the Call Center, and slightly more than half did not utilize MyApp.

This trend may be attributed to a preference for simpler issue resolution, as switching to another telecom company is perceived as more convenient than navigating the complexities of contacting the call center or using MyApp for problem resolution.

Our expectation of the ML Model

Before Using ML Model

The company endeavors to provide treatment for **all its customers**, incurring **significant company costs** in the process.

After Using ML Model

The company strategically focuses its treatment efforts on **customers predicted to churn**, thereby **optimizing company costs**.

Data Preprocessing

Features Selected:

- Device Class
- Games Product
- Music Product
- Education Product
- Call Center
- Video Product
- Use MyApp
- Payment Method
- Tenure Months
- Monthly Purchase (Thou. IDR)
- CLTV(Thou. IDR)

- Convert data types from 'object' to 'float' for 'Monthly Purchase' and 'CLTV'.
- Adjust values from 'No Internet Service' to 'No' in 'Games Product', 'Music Product', 'Education Product', 'Video Product', and 'Use MyApp' features.

- Normalize the numerical features.

- Label encoding 'Games Product', 'Music Product', 'Education Product', 'Call Center', 'Video Product', 'Use MyApp', and 'Device Class' features
- Applying one-hot encoding to 'Payment Method.'

- Implement the **SMOTE** (Synthetic Minority Over-sampling Technique) for addressing class imbalance.

- Divide the data into three subsets: **70% for training**, **15% for testing**, and **15% for validation**. Dividing the data this will prevent any data leaking throughout ML modelling.

Now let’s get into the Machine Learning Modelling

The reasons behind the use of ML metrics

F1 Score	It considers both **precision** and **recall**, providing a balance between minimizing **false positives** and **false negatives**, which is particularly useful when dealing with **imbalanced datasets**.
AUC ROC	It provides a **comprehensive performance metric**, capturing the trade-off between **true positive and false positive rates** across various classification thresholds and offering a **single-value summary** of the model's **discriminative ability**.
Log Loss	It measures the **accuracy of probability estimates**, penalizing models more severely for **confidently incorrect predictions** and providing a detailed assessment of the model's **probabilistic performance**.

We evaluate several ML models using only our train data.

Model	Train F1 Score	Validation F1 Score	Train ROC AUC	Validation ROC AUC	Train Log Loss	Validation Log Loss
Decision Tree	1.000000	0.789688	1.000000	0.787401	2.220446e-16	7.668513
Random Forest	1.000000	0.835576	1.000000	0.917852	9.410066e-02	0.393621
SVM	0.842652	0.830493	0.927650	0.910016	3.615648e-01	0.384322
XGBoost	0.974662	0.839449	0.997159	0.922707	1.233475e-01	0.368049
Gradient Boosting	0.856224	0.835595	0.935591	0.920844	3.343977e-01	0.362643

Now, we're searching our best hyperparameter using OPTUNA

Why use optuna?

Optuna is a better choice than random search and grid search because instead of searching randomly or in a fixed way, Optuna learns from past attempts. It adjusts its search to focus on areas that seem more likely to have the best answers. This makes it find good solutions faster and makes hyperparameter tuning work better.

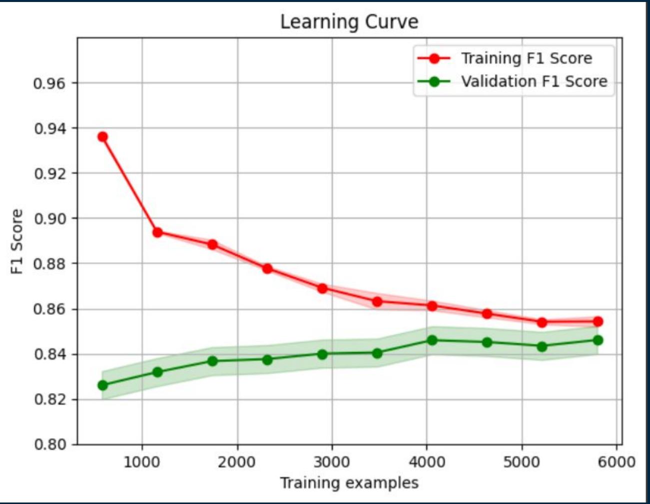
In hyperparameter tuning, we use the training data to optimize parameters and assess the learning curve. The validation data is then employed to analyze bias-variance trade-off and log loss.

After intensely doing hyperparameter tuning, we got…

Best Hyperparameter:

- 'max_depth': 5
- 'max_features': 'log2'
- 'learning_rate': 0.029
- 'subsample': 0.356
- 'n_estimators': 125

- Train F1 Score : 0.832
- Validation F1 Score : 0.846
- Log Loss : 0.383

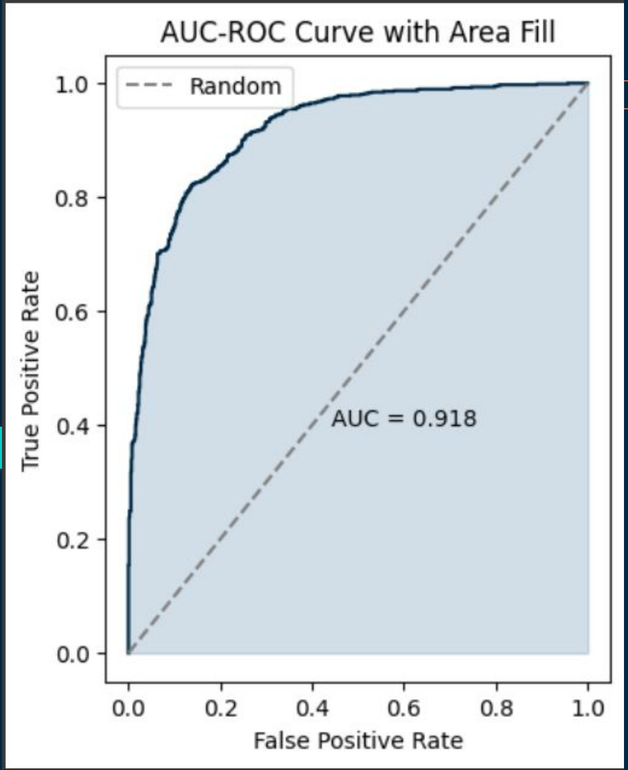


Bias-Variance Tradeoff Decomposition

- Average Bias: 0.155
- Average Variance: 0.031
- Average Expected Loss: 0.160

Our model robustly generalizes to new data, demonstrated by consistent performance on both training and validation sets. The low log loss highlights its efficiency in making accurate probabilistic predictions, while a stabilizing learning curve with more data indicates adaptability. Overall, our well-balanced model maintains moderate bias, low variance, and a low expected loss, showcasing strong performance.

Finally, the model evaluation



Here, we use the combined train and validation data for training the model, and test data for evaluating the model.

With a score of 0.913, we can confidently say that our model has a very strong ability to make accurate predictions, showcasing its excellence in separating positive and negative instances. This high value signifies a robust and reliable performance in our classification task.

Test F1 Score: 0.835
Log Loss: 0.381

The scenario of our business simulation

In our business simulation, we employ proactive measures to prevent customer churn, including personalized calls and broadcasts. Additionally, we enhance our retention strategy by introducing one-time offer discounts during these interactions. Initially, prior to the machine learning model, we assume a conservative approach, treating all customers as potential churn risks.
With the integration of the machine learning model, our strategy becomes more refined, focusing treatments on customers identified through predictions as likely to churn.

Before doing the business simulation, let's define our assumption factors first.

01 Cost

- **Marketing Cost per person → Rp10.000**
  The expenses incurred for reaching out to each customer through marketing calls and broadcast messages.
- **Discount Cost per person → Rp30.000**
  The expenditure associated with providing discounted prices to each customer.

02 Rate of Promo Acceptance & Rate of Promo Annoyance

- **Churn Customers' Rate of Promo Acceptance (RPA C) → 50%**
  The likelihood of customers who are at risk of churning accepting our offer.
- **Retain Customers' Rate of Promo Acceptance (RPA R) → 75%**
  The probability of customers who are likely to stay accepting our offer.
- **Rate of Annoyance (RoA) → 10%**
  The likelihood that customers experienced annoyance due to our call.

Let's assume the first scenario of our business simulation – the absence of a ML model.

This simulation is done by using the test data that consists of 1553 customers (776 Churn and 777 Retain)

In the test data, the median of CLTV is as follows:

- Churn Customer = Rp4.891.000
- Retain Customer = Rp5.991.000

Marketing Cost
= Total Customers x Marketing Cost
= 1,553 x Rp10.000 = **Rp15.530.000**

Discount Cost
Discount Cost for churning customers
= Churning Customers x RPA C x Discount Cost
= 776 x 50% x Rp30.000= **Rp11,640,000**

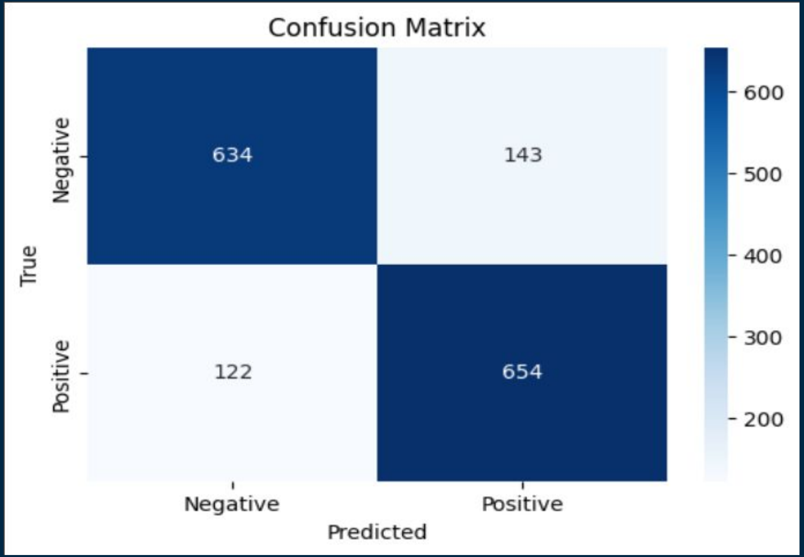
Discount Cost for retaining customers
= Retaining Customers **x** RPA R **x** Discount Cost
= 777 x 75% x Rp30.000 = **Rp17,482,500**

Total Cost = **Rp44.652.500**

CLTV Loss
= CLTV Loss from Churn Cust + CLTV Loss from Retain Cust
= (Total Churn Cust x RPA C x CLTV (Churn)) + (Total Retain Cust x RoA x CLTV (Retain))
= (776 x 50% x 4.891.000) **+** (777 x 10% x 5.991.000) = **Rp2.363.208.700**

CLTV Retain
= CLTV Retain from Churn Cust + CLTV Retain from Retain Cust
= ((Churn Cust **x** (1-RPA C) **x** CLTV (Churn)) **+** (Retain Cust **x** (1-RoA) **x** CLTV (Retain))
= (776 x 50% x 4.891.000) **+** 777 x 90% x 5.991.000
= **Rp6.087.214.300**

Now, let's see how our model perform in the business simulation.



Marketing Cost
= (TP + FP) x Marketing Cost
= (654 + 143) x Rp10.000 = **Rp7.970.000**

Discount Cost
Discount Cost for churning customers
= TP x RPA C x Discount Cost
= 654 x 50% x 30.000 = **Rp9.810.000**

Discount Cost for retaining customers
= FP x RPA R x Discount Cost
= 143 x 75% x 30.000 = **Rp3.217.500**

Total Cost = Rp20.997.500

CLTV Loss
= CLTV Loss from Churn Cust + CLTV Loss from Retain Cust
= (((TP x RPA C) + FP)x CLTV Churn) + (FP x RoA x CLTV Retain)
= (((654 x 50%) + 122) x4.891.000) + (143 x 10% x 5.991.000)
= **Rp2.281.730.300**

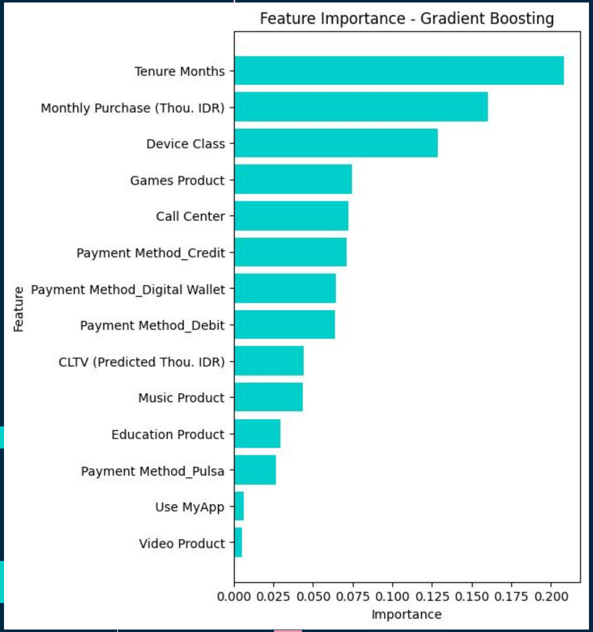
CLTV Retain
= CLTV Retain from Churn Cust+ CLTV Retain from Retain Cust
= (TP x (1-RPA C) x CLTV Churn) + (((FP x (1-RoA)) + TN) x CLTV Retain)
= (654 x 50% x 4.891.000) + (((143x 90%) + 634) x 5.991.000)
= **Rp6.168.692.700**

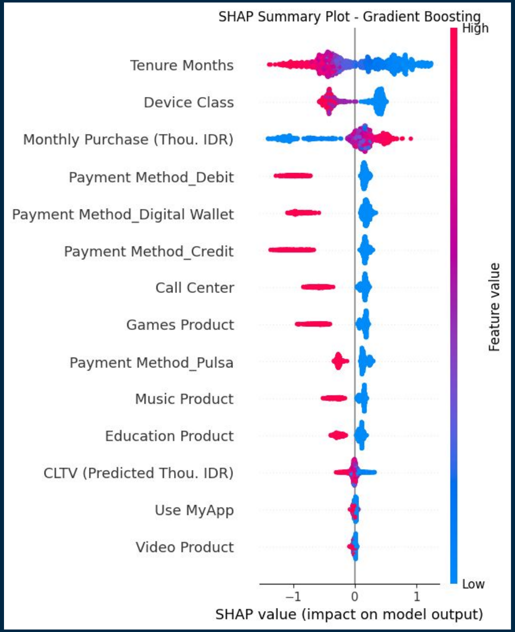
To what extent can our ML model decrease churn compared to before its implementation?

39.95% → **40.34%**

While the increase in customer retention is slightly higher after implementing the model, considering the costs, customer lifetime value loss, and the potential to retain customer lifetime value, utilizing the model proves to be a more advantageous approach.

Let’s see the what features are important and the SHAP values





In our analysis, we found that longer tenure, higher device class, and lower monthly purchase values are linked to reduced churn.

Yet, the impact of device class may differ due to the model's complex consideration of feature interactions, capturing non-linear relationships and potential confounding effects with other variables.

Business suggestions for the leaving customers

01 Offer a one-time discount during marketing calls and broadcasts for predicted churn customers.

02 Conduct satisfaction surveys for churned customers to identify reasons for their departure.

03 Launch win-back campaigns with enticing offers to encourage the return of churned customers.

What treat should we give to our loyal customers?

01 Conduct a biannual Customer Satisfaction Survey with rewards for feedback, including collecting additional customer information like age.

02 Implement product and service quality assurance based on feedback and company maintenance.

03 Monitor and analyze competitors to identify areas for improvement and strategic focus.

04 Introduce a Loyalty Program offering extra benefits to customers with tenure of 12 months or more.

We also built a simple web application that you can access it here.

If you’re interested to see our notebook, you can access our google colab notebook here.

Thank You!