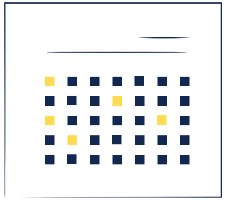


E-Commerce Customer Churn Prediction

Project by Erlando Febrian

About Me

I graduated of bachelor's degree from Bandung Institute of Technology, School of Business and Management, Business degree. I also graduated from Rakamin Data Science bootcamp with outstanding grade, awarded as best final project team, and also my role as team leader. I experienced in the following scope:

- Supervised & Unsupervised Learning
- Time Series Forecasting
- A/B Testing
- Deep learning using TensorFlow and Pytorch
- Recommender System
- Customer Lifetime Value
- SQL & Data Visualization (Tableau & Power BI)

Contact

github.com/erlndofebri

brian-insights.site/

erlandoregita99@gmail.com

0821-1000-4094

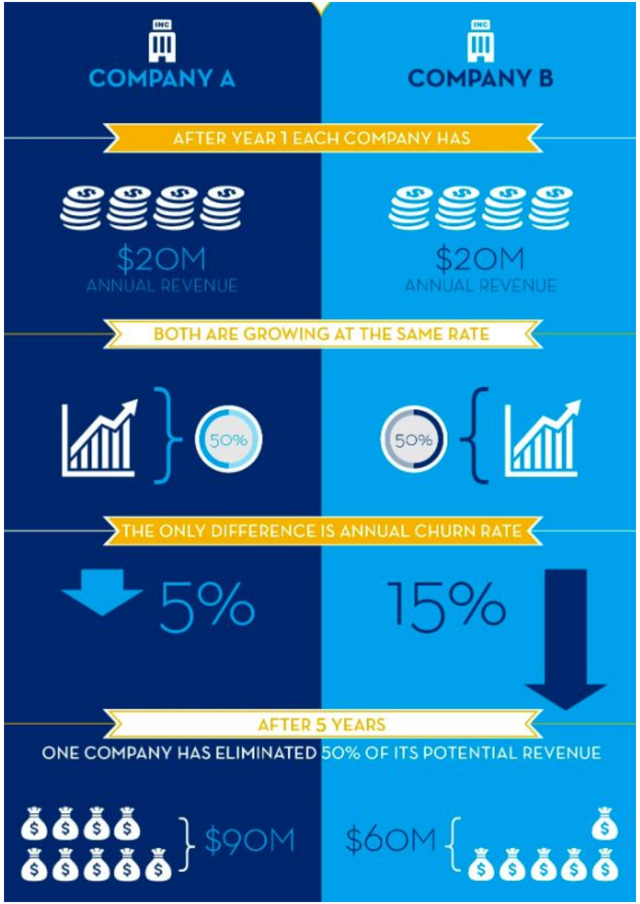
linkedin.com/in/erlandoregita/

Table of Content

01

Background

Background, Metric, Objective, and Goals



Why Customer Churn Rate is Big Problem?

01

Comparing 2 Companies With Same Annual Revenue

Company A has $20M annual revenue as well as Company B

02

They Have The Exact Same Growth Rate

The only difference is Churn Rate, and we will look forward 5 years later in the future

03

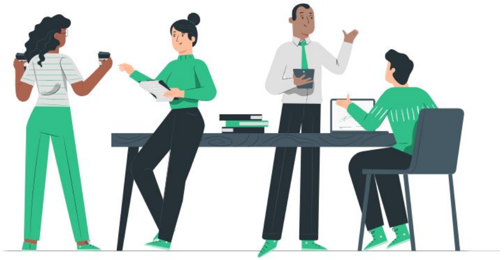
After 5 Years Company B Loss Their Potential Revenue

Look at the difference, company A gain $90M and company B only gain $60M

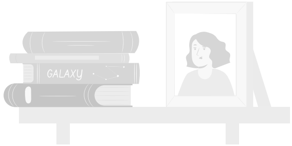
*) Reference : https://www.satrixsolutions.com/blog/customer-churn-cost

it cost 5x more to get a new customer than it did to keep an existing

- Rule of Thumb -



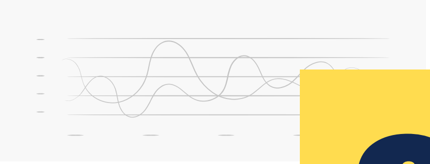
Goal, Objective, and Business Metric



Metric
Churn Rate

Goal
Reduce Churn Rate up to below **15%**

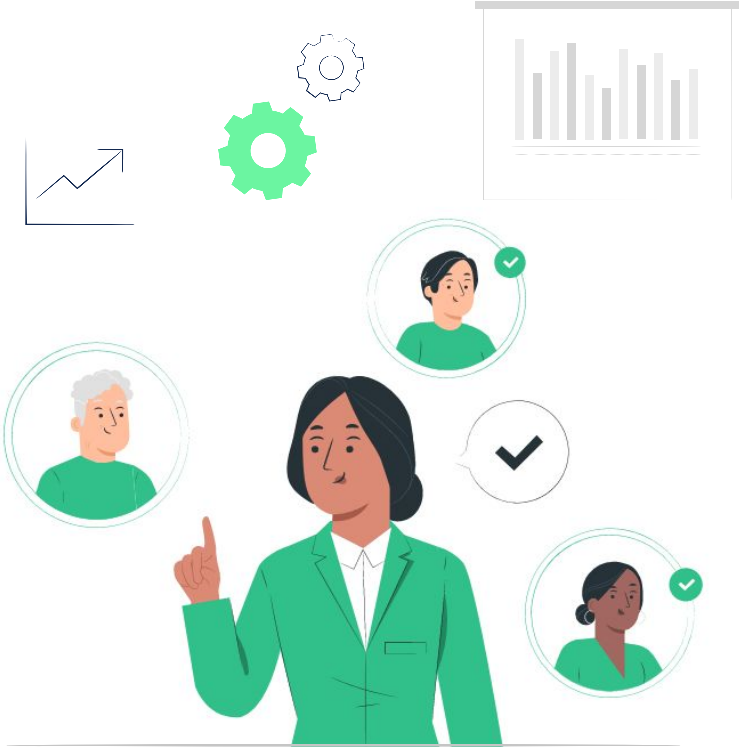
Objective
- Analyze factors that cause high churn rate
- Predict customer will churn or not



02

Exploratory Data Analysis

Data Exploration and Insights



Dataset Overview

1 Year Historical Data, contains of 5630 rows

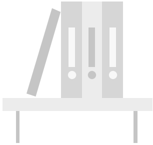
Numerical Features

- Customer ID
- DaySinceLastOrder
- Churn
- CashbackAmount
- CouponUsed
- Tenure
- CityTier
- OrderCount
- Complain
- WarehouseToHome
- OrderAmountHikeFromlastYear
- NumberOfAddress
- HourSpendOnApp
- SatisfactionScore
- NumberOfDeviceRegistered

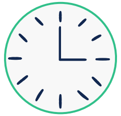
Categorical Features

- PreferredLoginDevice
- PreferredPaymentMode
- Gender
- PreferedOrderCat
- MaritalStatus

*) Detail Features Dictionary Written On Appendix



Target Feature Overview



Churn - Target Feature

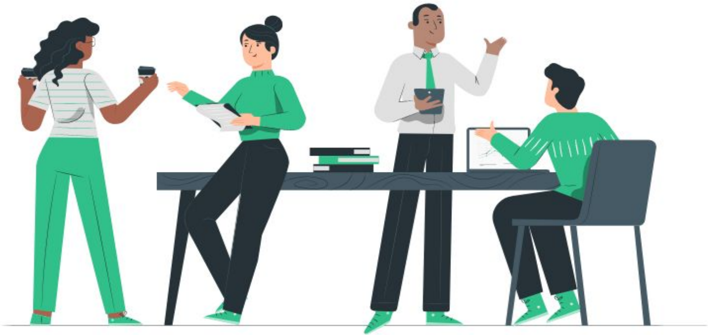
Is Customers will Churn?

Churn

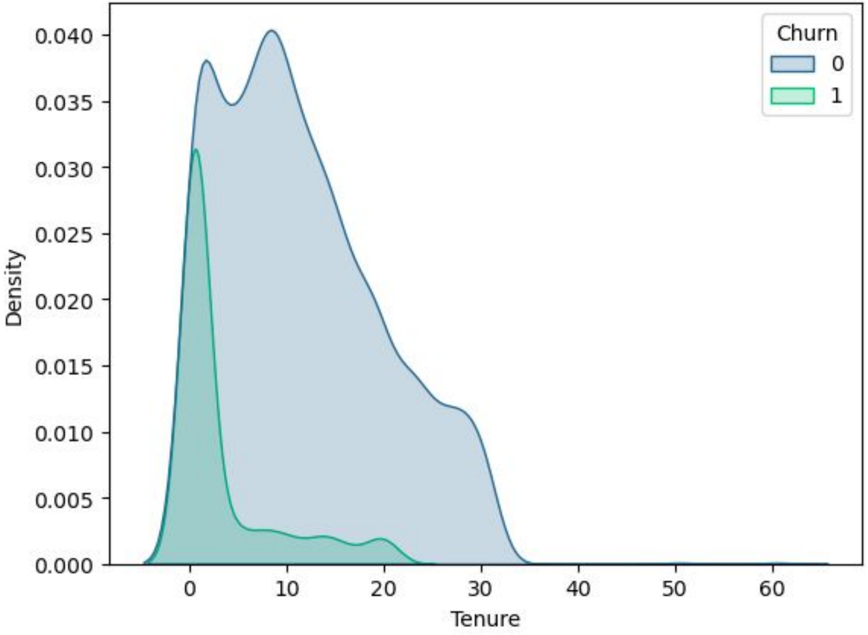
948 Customers

Not Churn

4682 Customers



Data Exploration - Tenure



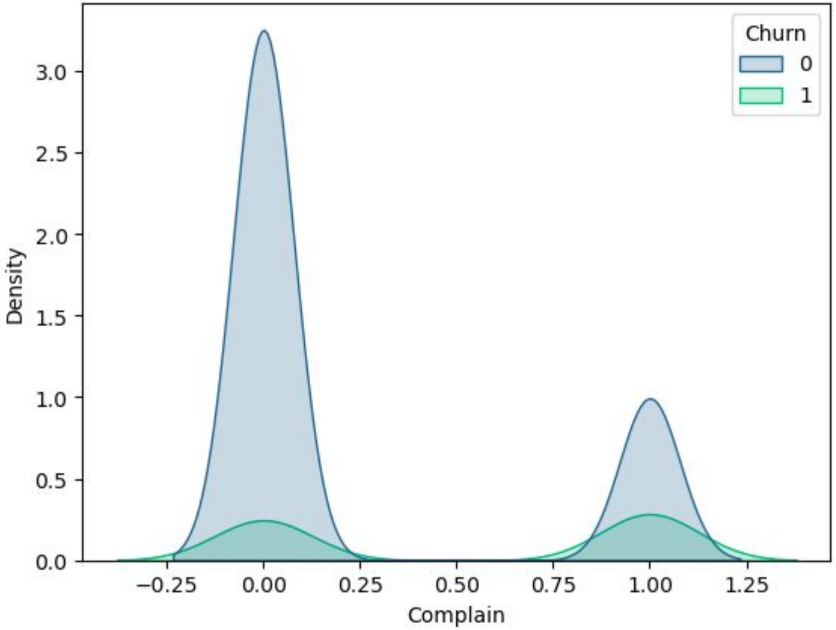
Avg Tenure for **Churn Customers is 3 days** and Avg Tenure for **Not Churn Customers is 11 days**

Insights:

- We need to know how to deal with customers that have high/low Tenure
- Find the behavior of high and low tenure customers

*) Tenure is the term used to describe the length of time (days) a customer remains a customer.

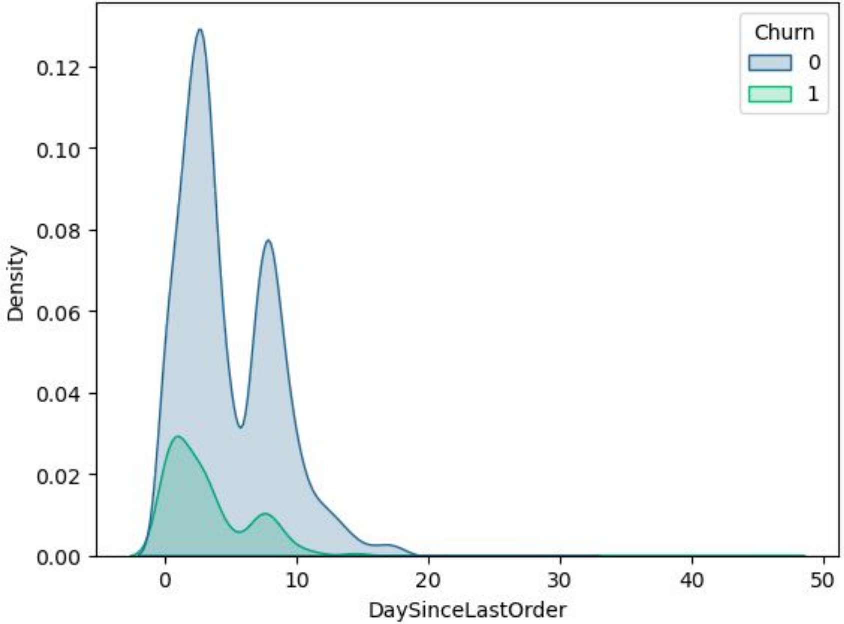
Data Exploration - Complain



Churn Customer have 1 median of complain and **Retain Customer have 0 median of complain.** It means that churn customers tend to be more complaining than retaining customer

*) Any complaint has been raised in last month

Data Exploration - Day Since Last Order



Churn customers have slightly lower day since last order than retain customers.

Insights:

- We need to know how to deal with customers that have high/low Recency (Day since last order)
- Find the behavior of high and low recency customers

*) Day Since last order by customer

03

Data Pre Processing

All Pre processing step

Data Pre-Processing

1 Handling Missing Values
Impute median and min values to missing values based on business case

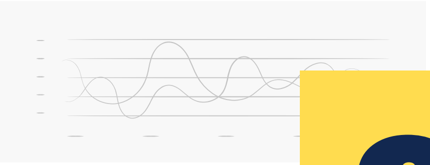
2 Feature Engineering
Engineer some features: PreferredLoginDevice, PreferredPaymentMode, and PreferedOrderCat

3 Feature Encoding
Using One Hot Encoding method

4 Feature Transformation
Using Log Transformation and Robust Scaler method

5 Outlier Handling
Using Z-score method, rows after handled : 5250 rows

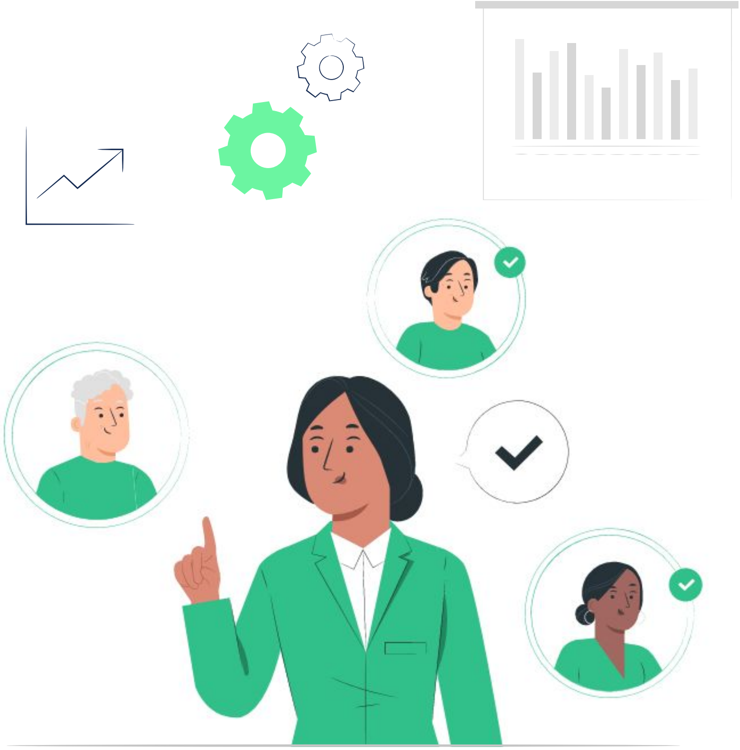
6 Imbalanced Target Handling
Using SMOTE, default sampling strategy



04

Modeling

Basic Model, Hyperparameter, and Feature Importance



Basic Modeling

	Model	Accuracy	Precision	Recall	F1 Score	F2 Score
0	Logistic Regression	0.810159	0.484277	0.813380	0.607096	0.834398
1	Decision Tree	0.897143	0.698052	0.757042	0.726351	0.931426
2	Random Forest	0.939683	0.882591	0.767606	0.821092	0.971965
3	Ada Boost	0.810159	0.484277	0.813380	0.607096	0.834398
4	Gradient Boost	0.893333	0.695946	0.725352	0.710345	0.932019
5	XG Boost	0.945397	0.905738	0.778169	0.837121	0.932019

We decided to do hyperparameter tuning on Random Forest and Decision Tree, because:

- High F2 score
- Less computational cost (Decision Tree)

*) pos_label = 0, F2 score to avoid high cost of False Negative (Predicted as Not Churn Customer, but actually it's Churn)

Hyperparameter Tuning

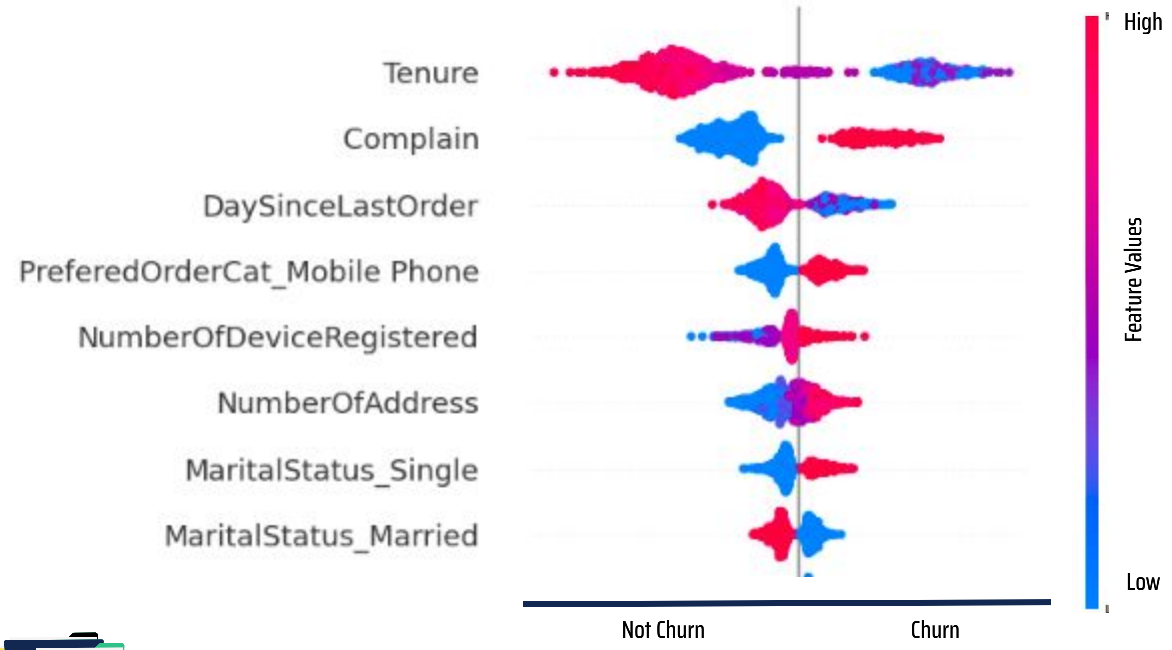
	Train F2 Score	Test F2 Score
Decision Tree	0.99	0.93
Random Forest	1	0.97

We decide to interpret Random Forest because:

- High F2 test score
- Not overfitting

*) pos_label = 0, F2 score to avoid high cost of False Negative (Predicted as Not Churn Customer, but actually it's Churn)

Feature Importance



We extracted 8 top importance features. For example, we can see that :

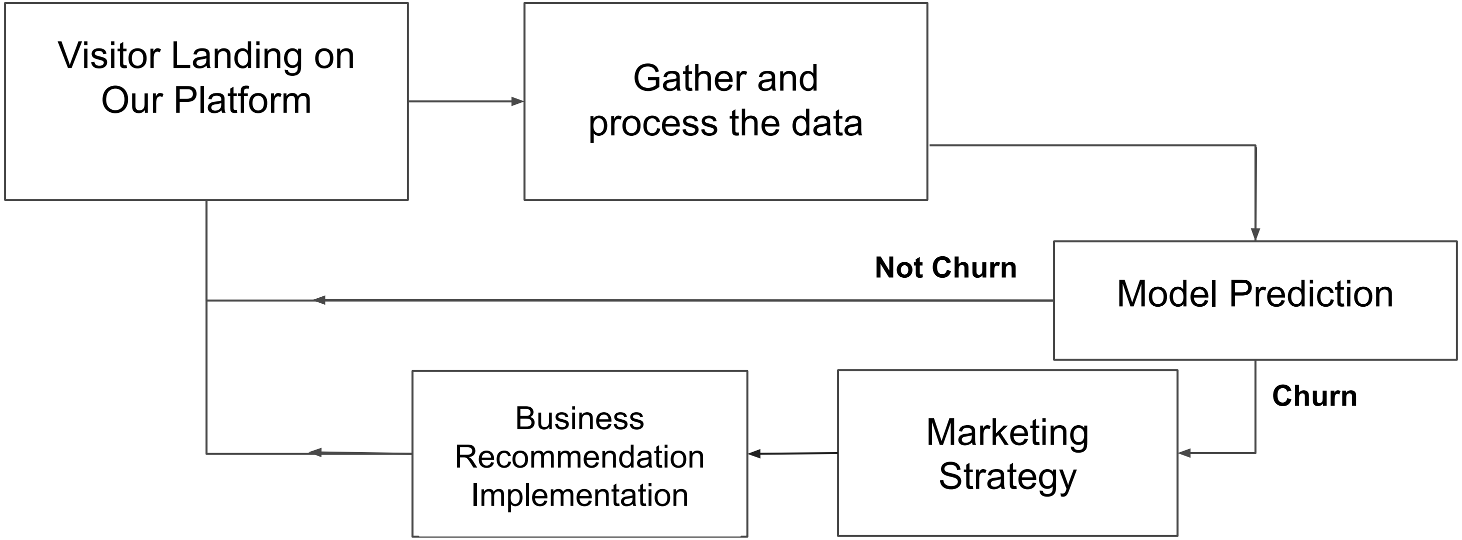
- Lower Tenure tend to be more churn than high Tenure
- Complaining customer tend to be more Churn rather than no complain customers
- etc

05

Business Insights and Recommendation

Data Interpretation, Insights, Simulation, and Business Recommendation

How Our Model Works?



Business Insights & Recommendation

Based On SHAP Values Feature Importance

Tenure

New customers tend to churn, the company must often **interact with new customers** through various marketing media

Complain

Customers who like to complain tend to be more churn, the **customer service and customer experience team must make extra effort** to handle customers who complain

Recency

New customers tend to churn (Low Recency), the company must often **interact with new customers** through various marketing media

Num Devices

Customers who have many registered devices tend to churn, **promo hunters and scammers** are a group of customers who churn and generally have many registered devices.

Num Address

Same as the number of registered devices, the large number of addresses creates a hunter and scammer promo. **Companies must reduce address slots and registered** devices to reduce potential customer churn due to scammers and promo hunters

*) Recency = DaySinceLastOrder

Strategy 1

Reallocate Cashback Amount

01

Reallocate Cashback

With the same Total Amount of Cashback $ **997K,** we reduce **3.6%** cashback from predicted non-churn customers then relocate to predicted churn customer (it's about **20%** increment amount of cashback for predicted churn customer)

02

Customer Churn Rate (Before - After Strategy Implementation)

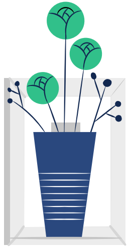
16.84 %
Churn Rate / Year

14.85 %
Churn Rate / Year

03

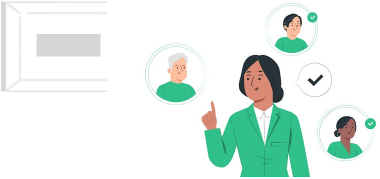
Customer Churn Rate Reduction

After implement the model and strategy, we can reduce Churn Rate up to **1.99%**



Strategy 2

Reallocate Cashback Amount and limit registered devices



01

Reallocate Cashback and limit registered devices

We tried to implement combine strategy, reallocate cashback amount and limit registered devices from **4 devices into 3 devices**

02

Customer Churn Rate (Before - After Strategy Implementation)

16.84 %
Churn Rate / Year

14.39 %
Churn Rate / Year

03

Customer Churn Rate Reduction

After implement the model and strategy, we can reduce Churn Rate up to **2.45%**

*) From SHAP values we know that the more number of devices registered tend to be more churn

Strategy 3

Reallocate Cashback Amount and limit registered address

01

Reallocate Cashback and limit registered address

We tried to implement combine strategy, reallocate cashback amount and limit registered devices from **4 address into 3 address**

02

Customer Churn Rate (Before - After Strategy Implementation)

16.84 %
Churn Rate / Year

→

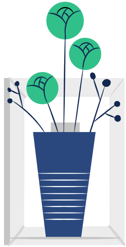
14.88 %
Churn Rate / Year

03

Customer Churn Rate Reduction

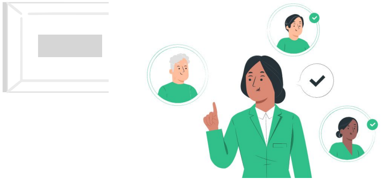
After implement the model and strategy, we can reduce Churn Rate up to **1.95%**

*) From SHAP values we know that the more number of address registered tend to be more churn



Combine All Strategy

Reallocate Cashback Amount + limit registered address and devices



01

Reallocate Cashback and limit registered address & devices

We tried to implement combine strategy, reallocate cashback amount and limit registered devices from **4 devices into 3 devices** and also limit registered address.

02

Customer Churn Rate (Before - After Strategy Implementation)

16.84 %
Churn Rate / Year

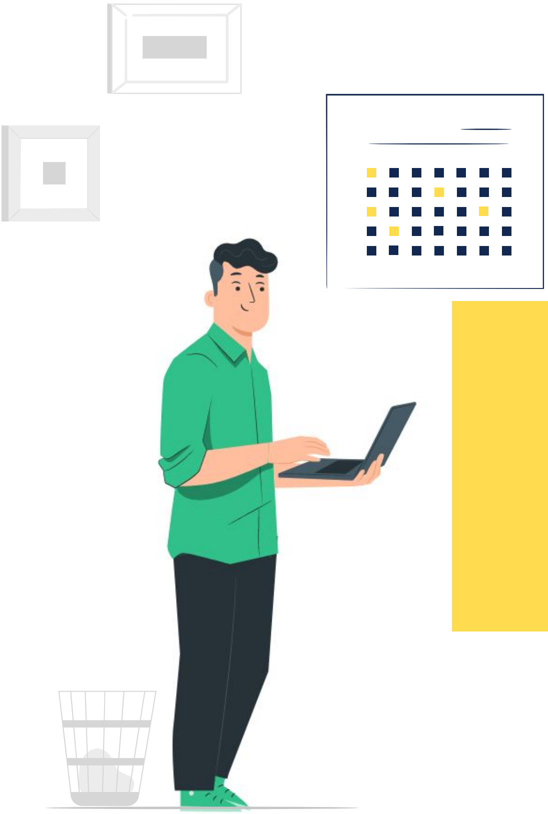
→

14.39 %
Churn Rate / Year

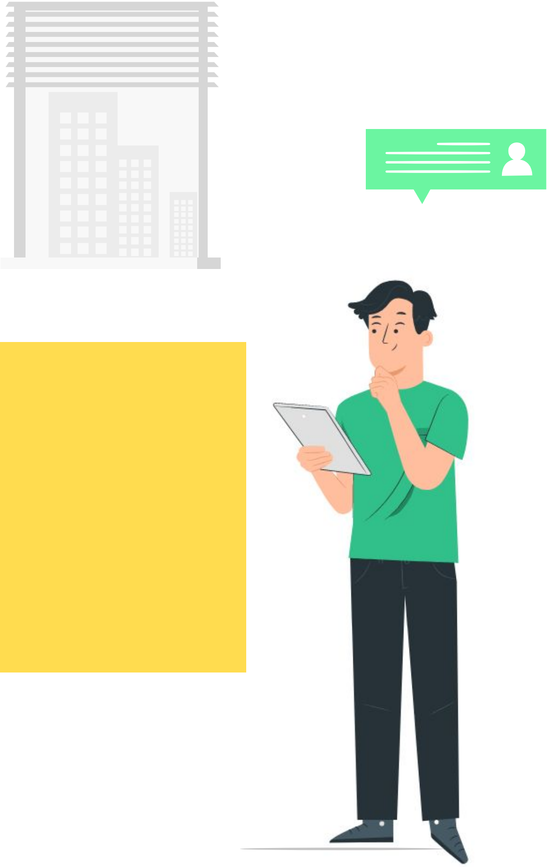
03

Customer Churn Rate Reduction

After implement the model and strategy, we can reduce Churn Rate up to **2.45%**



Thank You



Appendix

Features Dictionary

- Customer ID : Unique customer ID
- DaySinceLastOrder : Day Since last order by customer
- Churn : Churn Flag
- CashbackAmount : Average cashback in last month
- CouponUsed : Total number of coupon has been used in last month
- Tenure : Tenure of customer in organization
- CityTier : City tier
- OrderCount : Total number of orders has been places in last month
- Complain : Any complaint has been raised in last month
- WarehouseToHome : Distance in between warehouse to home of customer

Features Dictionary

- OrderAmountHikeFromlastYear : Percentage increases in order from last year
- NumberOfAddress : Total number of added on particular customer
- HourSpendOnApp : Number of hours spend on mobile application or website
- SatisfactionScore : Satisfactory score of customer on service
- NumberOfDeviceRegistered : Total number of devices is registered on particular customer
- PreferredLoginDevice : Preferred login device of customer
- PreferredPaymentMode : Preferred payment method of customer
- Gender : Gender of customer
- PreferedOrderCat : Preferred order category of customer in last month
- MaritalStatus : Marital status of customer