Transaction Behavior on Revoshop

Stanley Austin Susanto

Data Preparation for Propensity Model - 1

- First of all, we need to define **the target variable** which is **customers already activated and using PayLater** (at least, customers have done one transaction using PayLater). We make the table of target variable (suppose as **Target Table**).
- Now, we define past, present, and future period of the table, which is:

past period: 1 August 2022 to 31 January 2023 for **6 past promotion months** and **1 February 2023 to 30 April 2023**

present period: **May 2023**

future period: **June 2023 to December 2023**

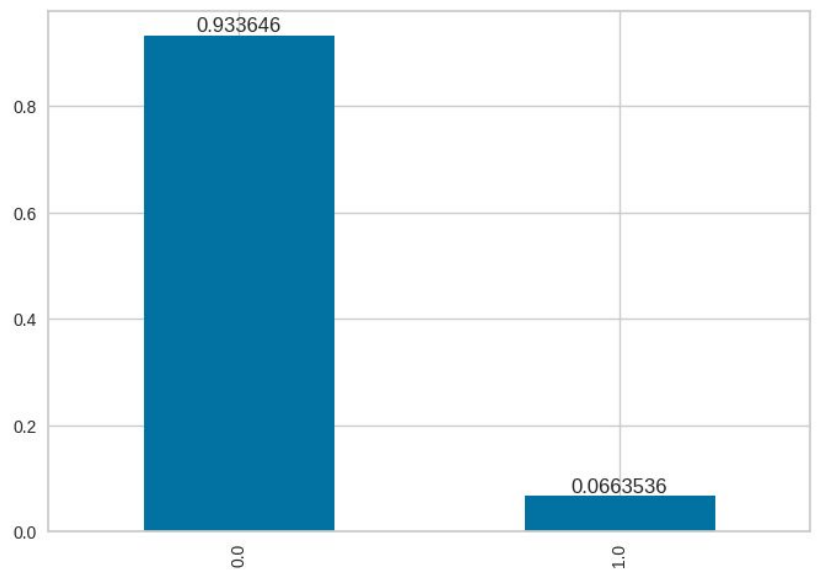
Data Preparation for Propensity Model - 2

- Next, we create features for propensity model, which is from table df2 that has been clustered (from Intermediate Assignment 2). Since the data of MOB are collected at 31 January 2023, we need to add the number of months on MOB by 3 (to 30 April 2023).
- Still the same topic, we can **achieve Total Sales** by **multiplying AVG_PROMO_TXN_AMT_L6M** and **PROMO_TXN_CNT_L6M**. We create this total sales as TOTAL_SALES_RESPECT_PROMO_6_MONTHS. Suppose we name this overall table as df3.
- The next step is, we **join Target Table with df3**. Since **the resulting table** has **missing values**, we need to **fill them with 0**.
- Dates are not related to the problem here, so we can drop the column of `PAYLATER_ACTIVATION_DT`.

Data Preparation for Propensity Model - 3

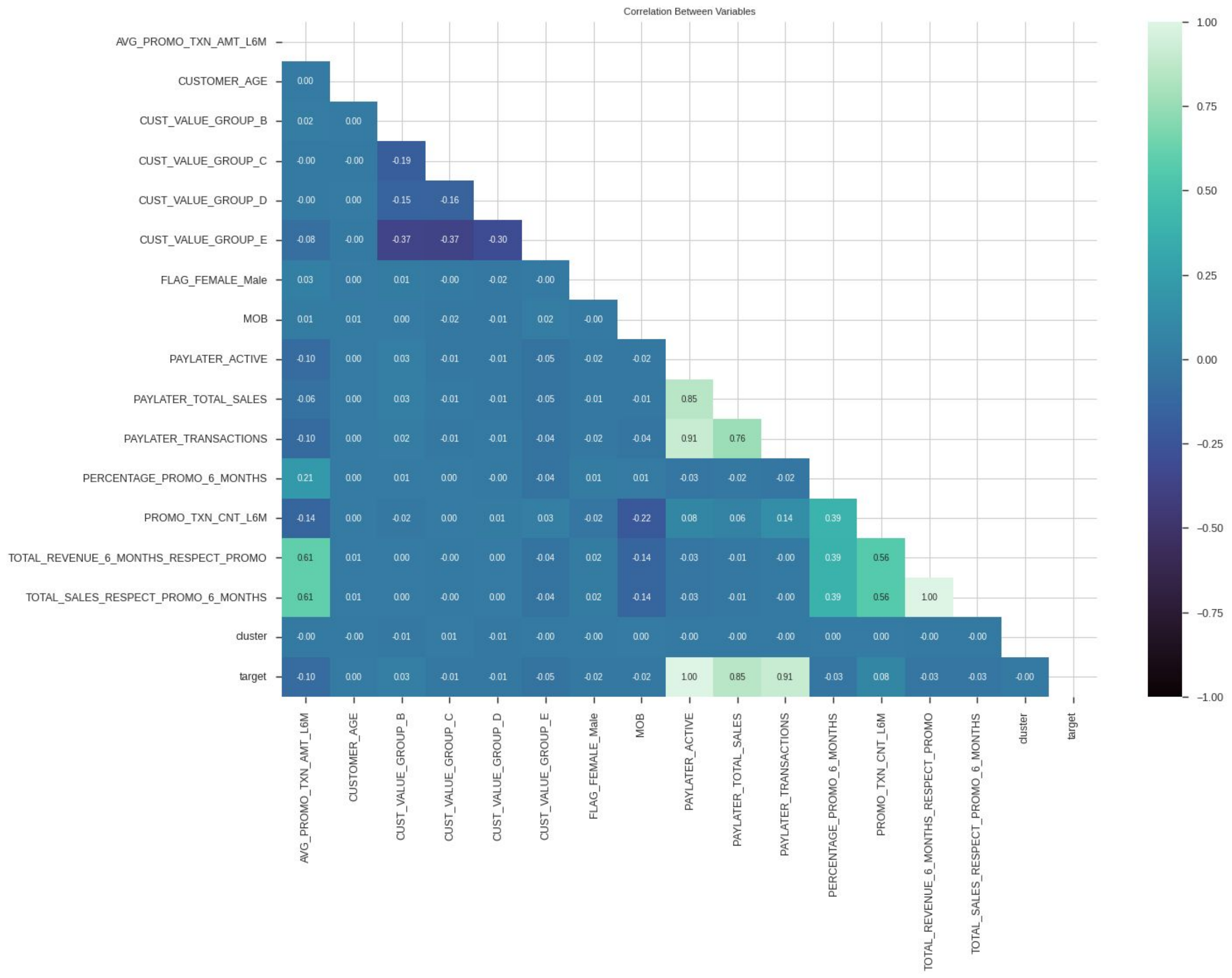
- After those steps explained in previous slides, we need to encode all object/string columns.
- Next, we need to create training and test dataset and do a correlation check of training and test dataset between targeted values and untargeted values.

- Since **the target** is only **6.6% of the population**, we can say that **the target is very unbalanced against population** (only 6.6% of population is using PayLater), so later on we will modify Logistic Regression to account for the difference.



Data Preparation for Propensity Model - 4

- Now, we are trying to remove some variables using correlation matrix. Bear in mind that we can only remove a variable if that **variable** has **correlation lower than -0.5** and **higher than 0.5** with **another variable** and that **variable has lowest correlation** with **target** value.
- Based on correlation matrix on the next slide, we can remove 6 columns such as: **'TOTAL_REVENUE_6_MONTHS_RESPECT_PROMO', 'CUST_VALUE_GROUP_C', 'CUSTOMER_AGE', 'CUST_VALUE_GROUP_D', 'cluster'**, and **'TOTAL_SALES_RESPECT_PROMO_6_MONTHS'.**

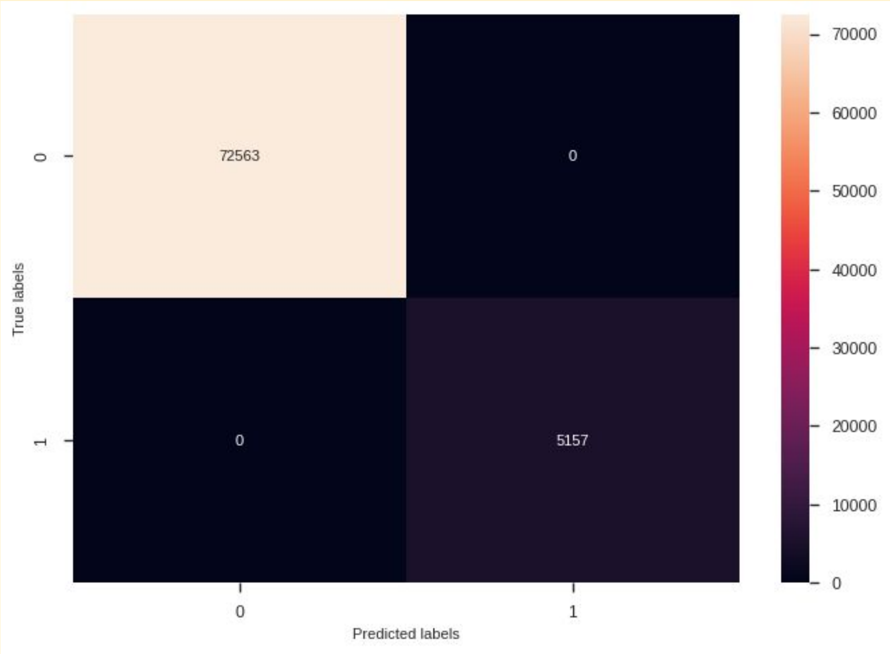

Correlation Between Variables
AVG_PROMO_TXN_AMT_L6M
CUSTOMER_AGE
0.00
CUST_VALUE_GROUP_B
0.02 0.00
CUST_VALUE_GROUP_C
-0.00 -0.00 -0.19
CUST_VALUE_GROUP_D
-0.00 0.00 -0.15 -0.16
CUST_VALUE_GROUP_E
-0.08 -0.00 -0.37 -0.37 -0.30
FLAG_FEMALE_Male
0.03 0.00 0.01 -0.00 -0.02 -0.00
MOB
0.01 0.01 0.00 -0.02 -0.01 0.02 -0.00
PAYLATER_ACTIVE
-0.10 0.00 0.03 -0.01 -0.01 -0.05 -0.02 -0.02
PAYLATER_TOTAL_SALES
-0.06 0.00 0.03 -0.01 -0.01 -0.05 -0.01 -0.01 0.85
PAYLATER_TRANSACTIONS
-0.10 0.00 0.02 -0.01 -0.01 -0.04 -0.02 -0.04 0.91 0.76
PERCENTAGE_PROMO_6_MONTHS
0.21 0.00 0.01 0.00 -0.00 -0.04 0.01 0.01 -0.03 -0.02 -0.02
PROMO_TXN_CNT_L6M
-0.14 0.00 -0.02 0.00 0.01 0.03 -0.02 -0.22 0.08 0.06 0.14 0.39
TOTAL_REVENUE_6_MONTHS_RESPECT_PROMO
0.61 0.01 0.00 -0.00 0.00 -0.04 0.02 -0.14 -0.03 -0.01 -0.00 0.39 0.56
TOTAL_SALES_RESPECT_PROMO_6_MONTHS
0.61 0.01 0.00 -0.00 0.00 -0.04 0.02 -0.14 -0.03 -0.01 -0.00 0.39 0.56 1.00
cluster
-0.00 -0.00 -0.01 0.01 -0.01 -0.00 -0.00 0.00 -0.00 -0.00 -0.00 0.00 0.00 -0.00 -0.00
target
-0.10 0.00 0.03 -0.01 -0.01 -0.05 -0.02 -0.02 1.00 0.85 0.91 -0.03 0.08 -0.03 -0.03 -0.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Propensity Model Training & Evaluation - 1

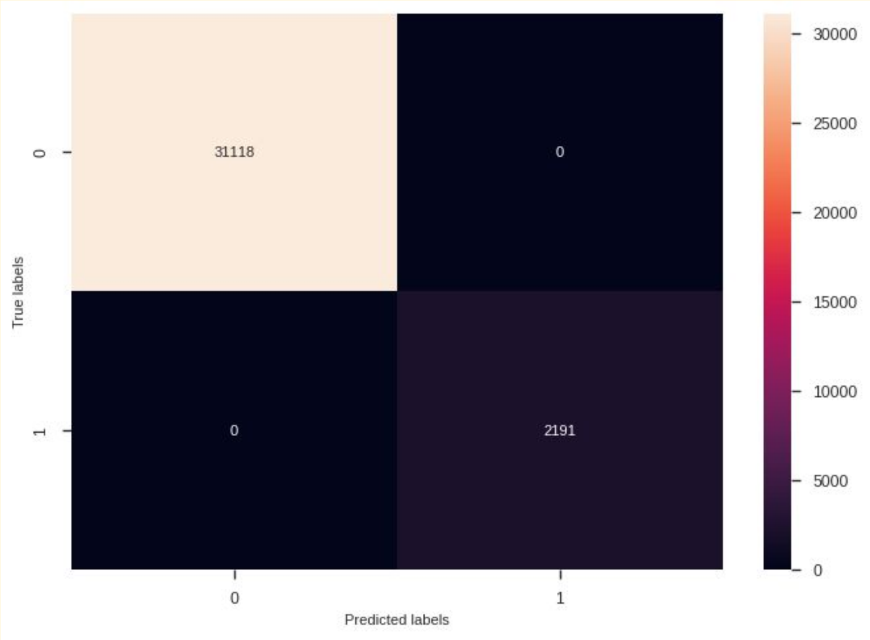
- We need to do fitting logistic regression model in order to find the probability and prediction of the model with the datasets from training and test datasets.
- Let us evaluate our model:

1. **accuracy** for **training dataset** is **100%**. Surprisingly enough that the **accuracy** for **test dataset** is also **100%**. Since the score of model training set and test set is same, then there is a **no noticeable difference** in **accuracy** between two.
2. **confusion matrices** for **training dataset** and **test dataset** can be shown on the next slide. It turns out that there is a **no noticeable difference** in **confusion matrix** between two.

Training dataset



Test dataset



Propensity Model Training & Evaluation - 2

We are going to evaluate the decile performance.

Based on the binning table here, we can say that there are **no noticeable rank breaks among the deciles**. If we broke them in percentiles, we might have some noticeable rank breaks.

target / binning	0.0	1.0	total_obs	pct_non_paylater	pct_paylater
(-0.000999999779, 9.01e-10]	7775.0	0.0	7775.0	0.107148	0.0
(9.01e-10, 9.49e-10]	7769.0	0.0	7769.0	0.107066	0.0
(9.49e-10, 9.72e-10]	7772.0	0.0	7772.0	0.107107	0.0
(9.72e-10, 9.91e-10]	7772.0	0.0	7772.0	0.107107	0.0
(9.91e-10, 1e-09]	7776.0	0.0	7776.0	0.107162	0.0
(1e-09, 1.01e-09]	7769.0	0.0	7769.0	0.107066	0.0
(1.01e-09, 1.03e-09]	7776.0	0.0	7776.0	0.107162	0.0
(1.03e-09, 1.06e-09]	7770.0	0.0	7770.0	0.107079	0.0
(1.06e-09, 1.09e-09]	7769.0	0.0	7769.0	0.107066	0.0
(1.09e-09, 1.0]	2615.0	5157.0	7772.0	0.036038	1.0

Propensity Model Training & Evaluation - 3

Since the table here shows only **one top decile** that have **larger count of positive instances** and the probability of finding positive instances) is 66.4%, we can conclude that **top decile effectively identify the most promising PayLater customers**. However, it is better to say that we are recommended to **use percentiles rather than deciles.**

target binning	0.0	1.0	total_obs	pct_non_paylater	pct_paylater
(-0.000999999779, 9.01e-10]	7775.0	0.0	7775.0	0.107148	0.0
(9.01e-10, 9.49e-10]	7769.0	0.0	7769.0	0.107066	0.0
(9.49e-10, 9.72e-10]	7772.0	0.0	7772.0	0.107107	0.0
(9.72e-10, 9.91e-10]	7772.0	0.0	7772.0	0.107107	0.0
(9.91e-10, 1e-09]	7776.0	0.0	7776.0	0.107162	0.0
(1e-09, 1.01e-09]	7769.0	0.0	7769.0	0.107066	0.0
(1.01e-09, 1.03e-09]	7776.0	0.0	7776.0	0.107162	0.0
(1.03e-09, 1.06e-09]	7770.0	0.0	7770.0	0.107079	0.0
(1.06e-09, 1.09e-09]	7769.0	0.0	7769.0	0.107066	0.0
(1.09e-09, 1.0]	2615.0	5157.0	7772.0	0.036038	1.0

Propensity Model Training & Evaluation - 4

By using the same technique for modeling our logistic regression for **customers who never used PayLater before**, we obtain segmentation table below. We can say that **4 top classes have highest probability** such as (0.99984, 0.99998], (0.99982, 0.99984], (0.99981, 0.99982], (0.9998, 0.99981]. The value of -1 is non-target values.

target binning	-1.0	1.0	total_obs	pct_non_paylater	pct_paylater
(9.7e-05, 0.99967]	5256.0	2517.0	7773.0	1.0	0.034734
(0.99967, 0.99973]	0.0	7772.0	7772.0	0.0	0.107253
(0.99973, 0.99976]	0.0	7774.0	7774.0	0.0	0.107281
(0.99976, 0.99977]	0.0	7771.0	7771.0	0.0	0.107239
(0.99977, 0.99979]	0.0	7770.0	7770.0	0.0	0.107226
(0.99979, 0.9998]	0.0	7774.0	7774.0	0.0	0.107281
(0.9998, 0.99981]	0.0	7770.0	7770.0	0.0	0.107226
(0.99981, 0.99982]	0.0	7772.0	7772.0	0.0	0.107253
(0.99982, 0.99984]	0.0	7774.0	7774.0	0.0	0.107281
(0.99984, 0.99998]	0.0	7770.0	7770.0	0.0	0.107226

PayLater Pilot Result - 1

Paylater Pilot Result:

Paylater Active = 0 --> Average Sales = 0, Count Customers = 103,548

Paylater Active = 1 --> Average Sales = 10,338, Count Customers = 7,481

Total Customers = 103,548 + 7,481 = **111,029**

Average Probability Using PayLater = (7,481/111,029) * 100 = **6.74%**

PayLater Pilot Result - 2

BENEFIT COST ANALYSIS

Total number of customer to call = 30000

Contact rate = 15%

Welcome bonus value (in euros) = €100.00

Cost / call (picked-up) = €2.75

Cost / call (dropped) = €1.15

Revenue margin (%) = 2.4%

Average PayLater sales / active customers = 10,338

Average probability of using PayLater = 6.74%

PayLater Pilot Result - 3

Takers calculation

Total number of contacted customers = 4500

Total number of dropped calls = 25500

Expected numbers of new users due to Project Contact = 303

Benefit

Expected numbers of new users due to Project Contact = 303

Expected total sales = €3,132,332.45

Expected revenue = €75,175.98

PayLater Pilot Result - 4

Cost

Cost of contacting customers = €12,375.00

Cost of contacting customers and dropped off = €29,325.00

Cost of welcome bonus = €30,300.00

Total cost = €72,000.00

R/E = 1.04