Bravo Team

E-COMMERCE CUSTOMER CHURN PREDICTION & ANALYSIS

By:

- Dicky Ardi

TABLE OF CONTENTS

BUSINESS OBJECTIVES

- Get a prediction system that can be used to find out which customers will churn and which factors that affect.
- Predict which customers will churn, and provide recommendations to the marketing team to marketing strategies in an effort to avoid losses in the form of losing customers.

Dataset

- **Source** : Ecommerce Customer Churn Analysis and Prediction from Kaggle
- **Data as of** : 26 January 2021
- **Number of rows (Customers)** : 5630 rows

Features

1. PreferredLoginDevice
2. PreferredPaymentMode
3. Gender
4. PreferedOrderCat
5. MaritalStatus
6. CityTier
7. Tenure
8. OrderCount
9. DaySinceLastOrder
10. NumberOfDeviceRegistered
11. SatisfactionScore
12. NumberOfAddress
13. Complain
14. CashbackAmount
15. WarehouseToHome
16. HourSpendOnApp
17. Average Cashback per Order
18. OrderAmountHikeFromlastYear

Algorithm Machine learning

1. Logistic Regression
2. Decision Tree
3. KNeighbors
4. AdaBoost
5. CatBoost

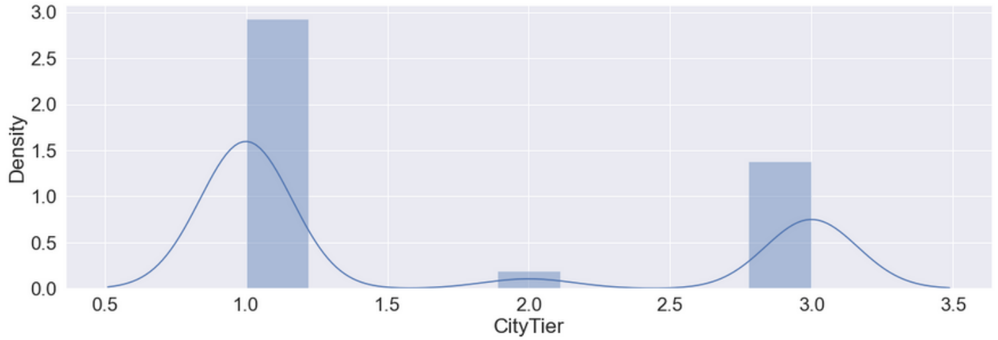
01
DATA UNDERSTANDING

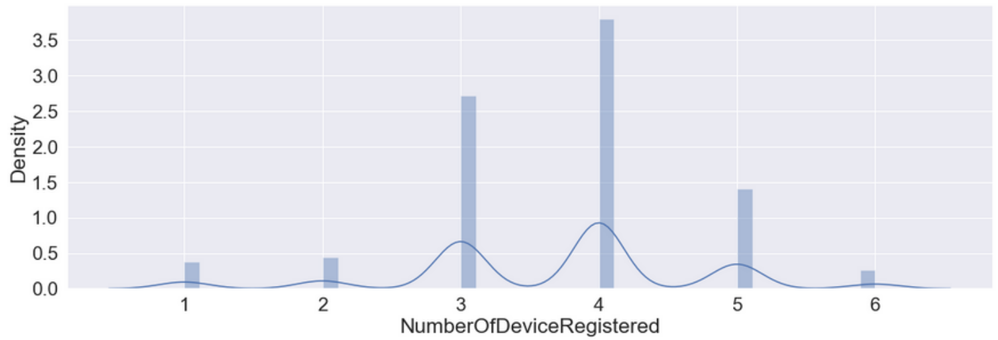
Dataset

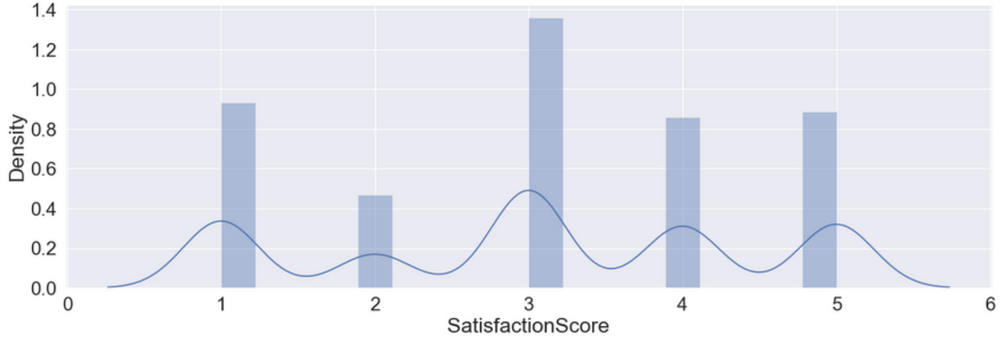
Target

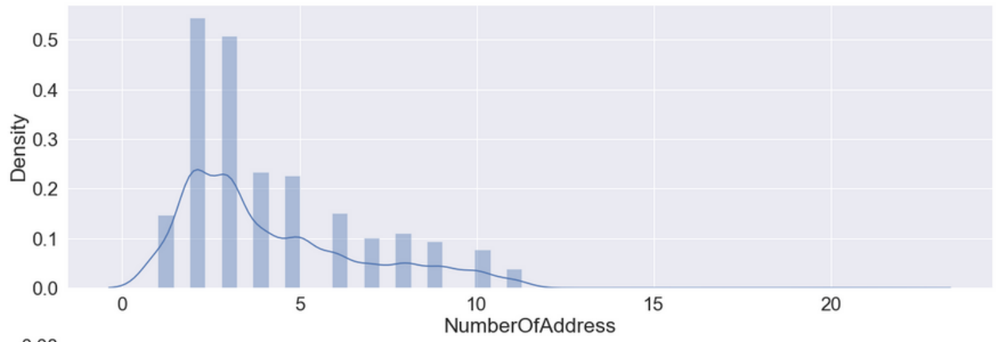
Columns	Description	Data Types	Remarks
CustomerID	Unique number for customer for identification.	Numerical	Min : 50001 Mean : 52815 Max: 55630 Median : 52815
Churn	0 : Customer who don't churn. 1 : Customer already churned.	Numerical	2 unique values Yes and No
Tenure	Tenure of customer in organization.	Numerical	Min : 0 Mean : 10.18 Max: 61 Median : 9
PreferredLoginDevice	Preferred login device of customer.	Categorical	3 unique values : Mobile Phone, Computer, Phone
CityTier	Column about city tier.	Numerical	Min : 1 Mean : 1.65 Max: 3 Median : 1
WarehouseToHome	Distance in between warehouse to home of customer.	Numerical	Min : 5 Mean : 15.63 Max: 127 Median : 14
PreferredPaymentMode	Preferred payment method of customer.	Categorical	7 unique values: Debit Card, Credit Card, E Wallet, UPI, COD, CC, Cash on Delivery
Gender	Gender of customer.	Categorical	2 unique values Female and Gender
HourSpendOnApp	Number of hours spend on mobile application or website.	Numerical	Min : 0 Mean : 2.93 Max: 5 Median : 3
NumberOfDeviceRegistered	Total number of deceives is registered on particular customer	Numerical	Min : 1 Mean : 3.68 Max: 6 Median : 4
PreferedOrderCat	Preferred order category of customer in last month.	Categorical	6 unique values: Laptop & Accessory, Mobile Phone, Fashion, Mobile, Grocery, Others
SatisfactionScore	Satisfactory score of customer on service.	Numerical	Min : 1 Mean : 3.06 Max: 5 Median : 3
MaritalStatus	Marital status of customer.	Categorical	3 unique values : Married, Single, Divorced
NumberOfAddress	Total number of added added on particular customer.	Numerical	Min : 1 Mean : 4.21 Max: 22 Median : 3
Complain	0 : Customer who have never complained in the last month. 1 : Customers who have complained in the last month.	Numerical	2 unique values Yes and No
OrderAmountHikeFromlastYear	Percentage increases in order from last year.	Numerical	Min : 11 Mean : 15.70 Max: 26 Median : 15
CouponUsed	Total number of coupon has been used in last month.	Numerical	Min : 0 Mean : 1.75 Max: 16 Median : 1
OrderCount	Total number of orders has been places in last month.	Numerical	Min : 1 Mean : 3 Max: 16 Median : 2
DaySinceLastOrder	Day Since last order by customer.	Numerical	Min : 0 Mean : 4.54 Max: 46 Median : 3
CashbackAmount	Average cashback in last month.	Numerical	Min : 0 Mean : 177.22 Max: 324.9 Median : 163.28

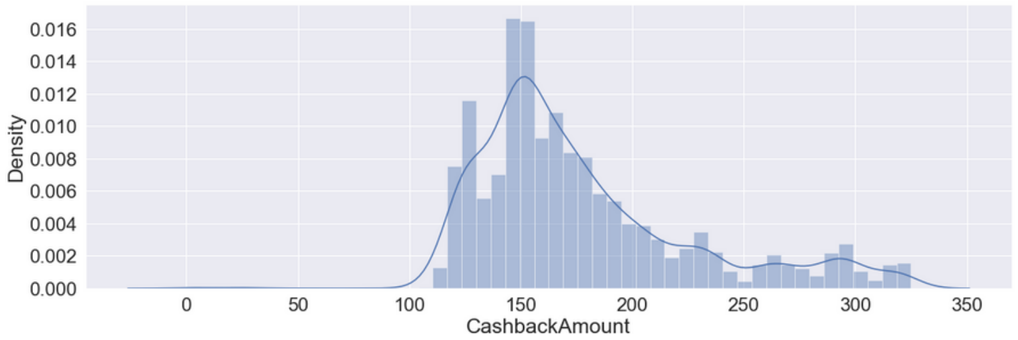
Distribution of Numerical variables

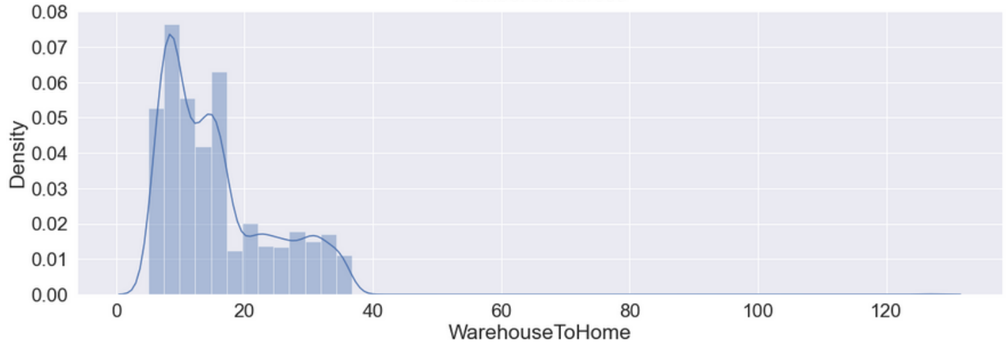




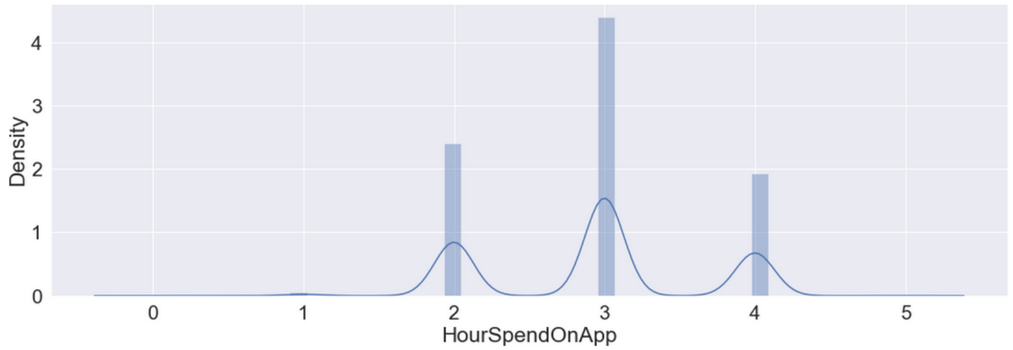


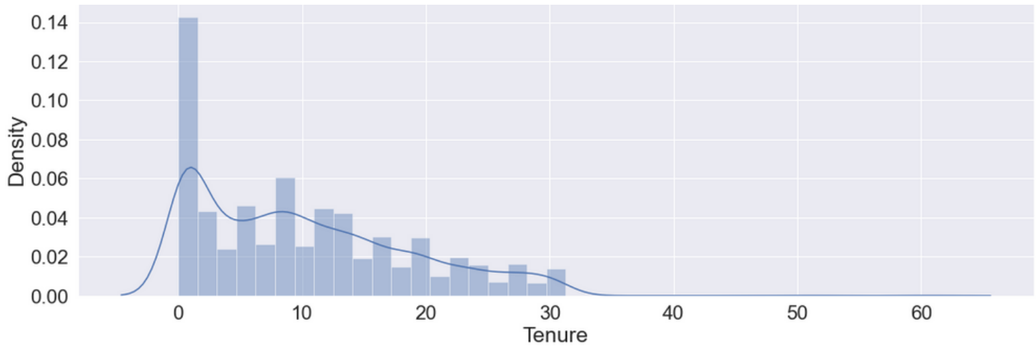


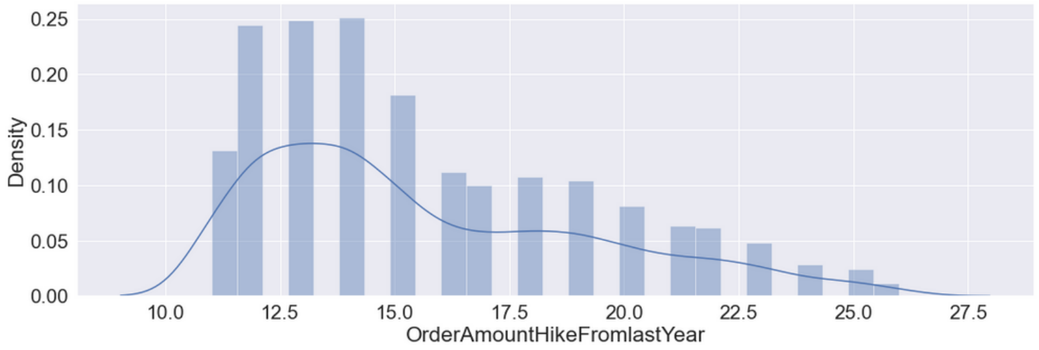


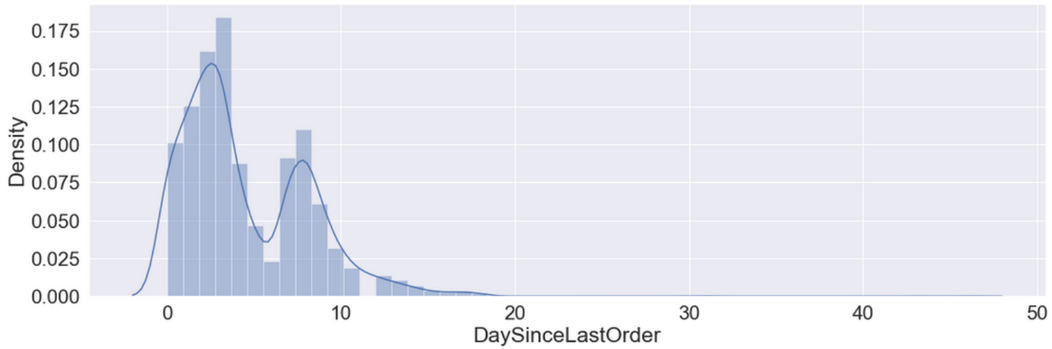


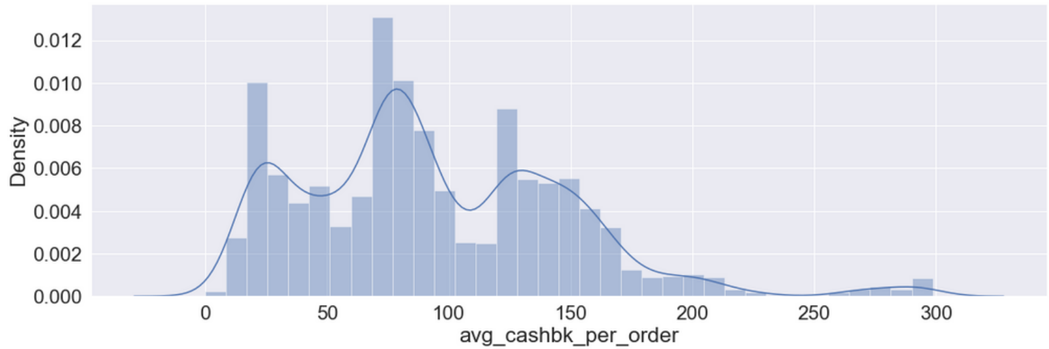
Distribution of Numerical variables







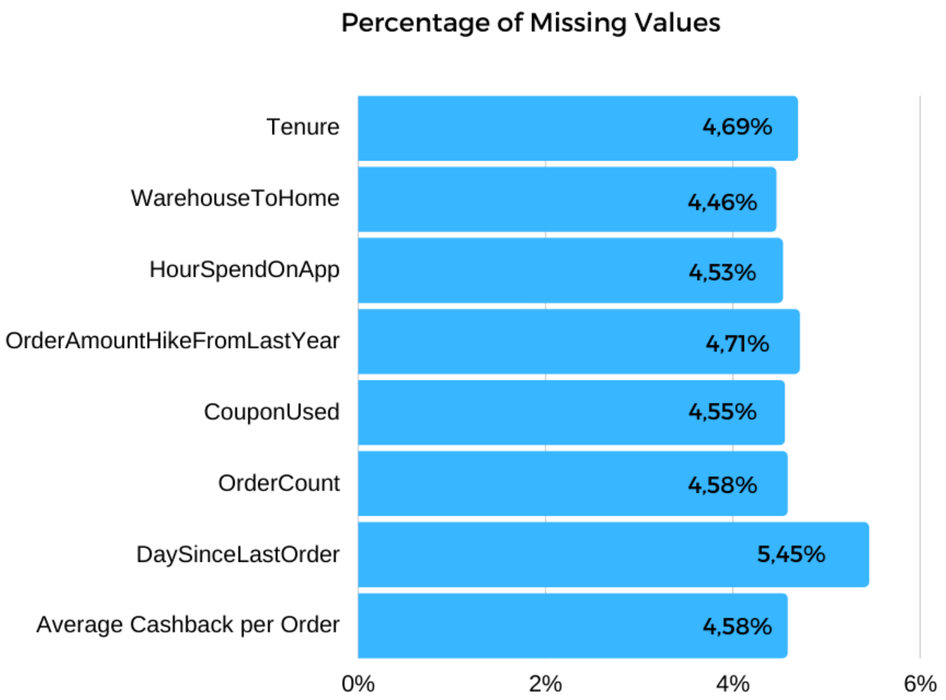




The distribution of numeric variables tends to **right skewed**

02
DATA PREPROCESSING

Missing Values



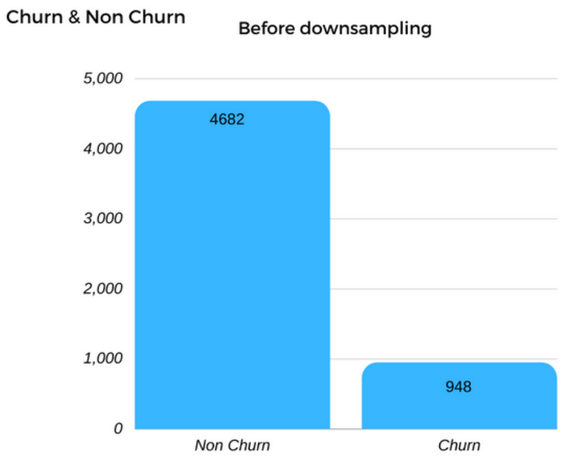
Description

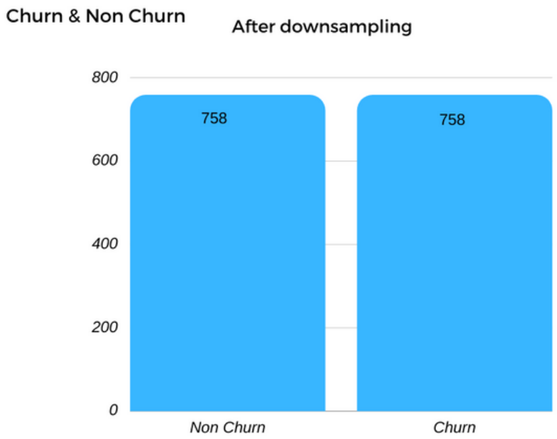
- There are **8 features** that have missing values.
- The percentage of missing data in the range of **4,46% - 5,45%**

Handling Missing Value

- Based on data, the distribution of data on numeric variables is **right skewed**, so we use the **median** to impute the missing value.

Imbalance Data





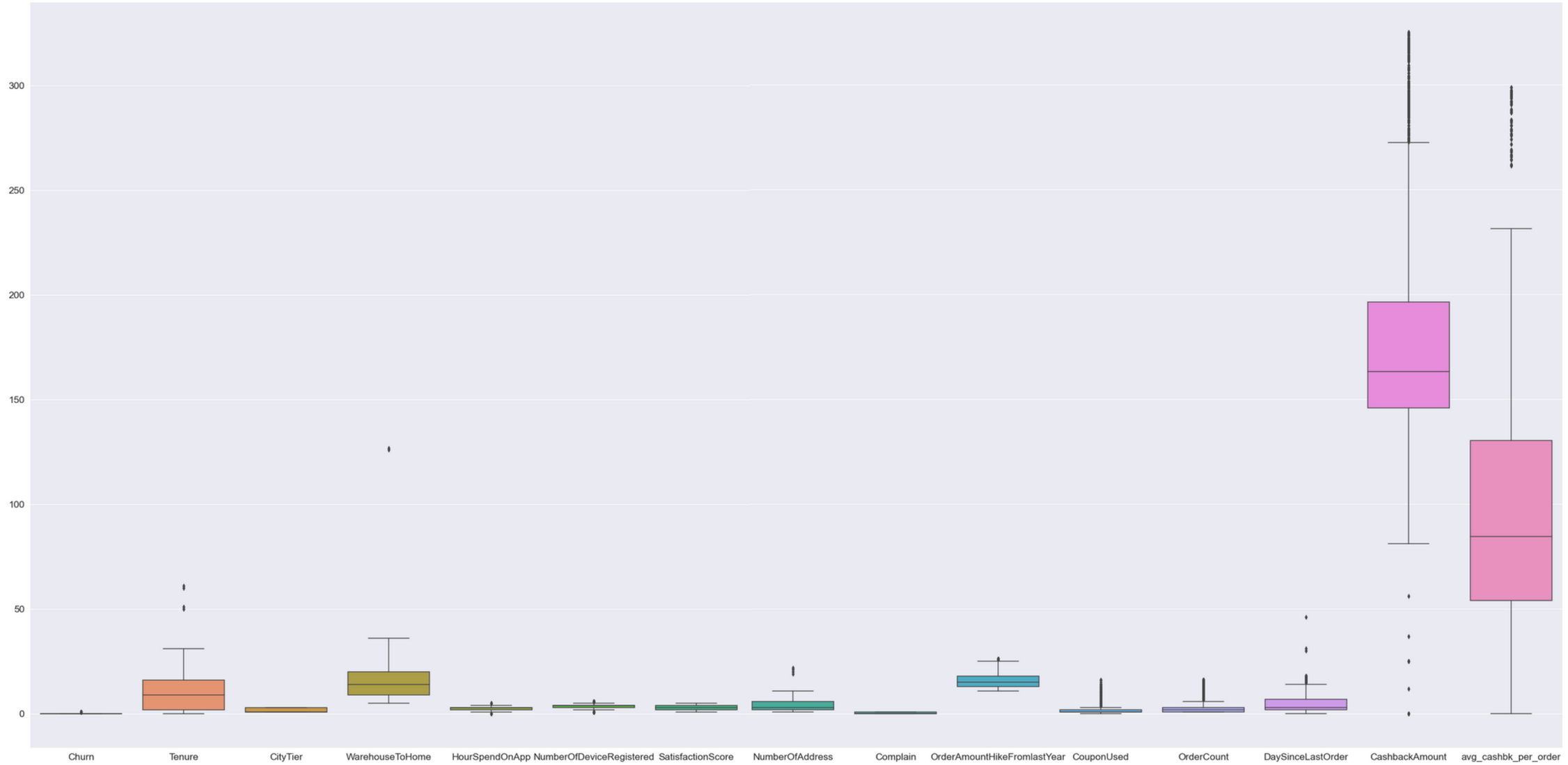
Description

- The proportion for the **minority class is 20.2%.**
- The degree of imbalance in the **mild category**.

Why downsampling ?

- To maintain **data purity.**
- The level of imbalance is **still at a low**.

Scalling



Bravo Team

There are a lot of outliers in the data, therefore we are using **Robust Scaler** to scale the data, which is more resistant to outliers.

Feature Engineering

Use **One Hot Encoding** for selected column

Features	After Encode						
PreferredLoginDevice	Mobile_Phone	Computer	Phone				
PreferredPaymentMode	Debit Card	Credit Card	E_Wallet	UPI	COD	CC	Cash on Delivery
Gender	Female	Male					
PreferedOrderCat	Laptop & Accesory	Mobile Phone	Fashion	Mobile	Grocery	Others	
MaritalStatus	Married	Single	Divorced				

03
EXPLORATORY DATA ANALYSIS

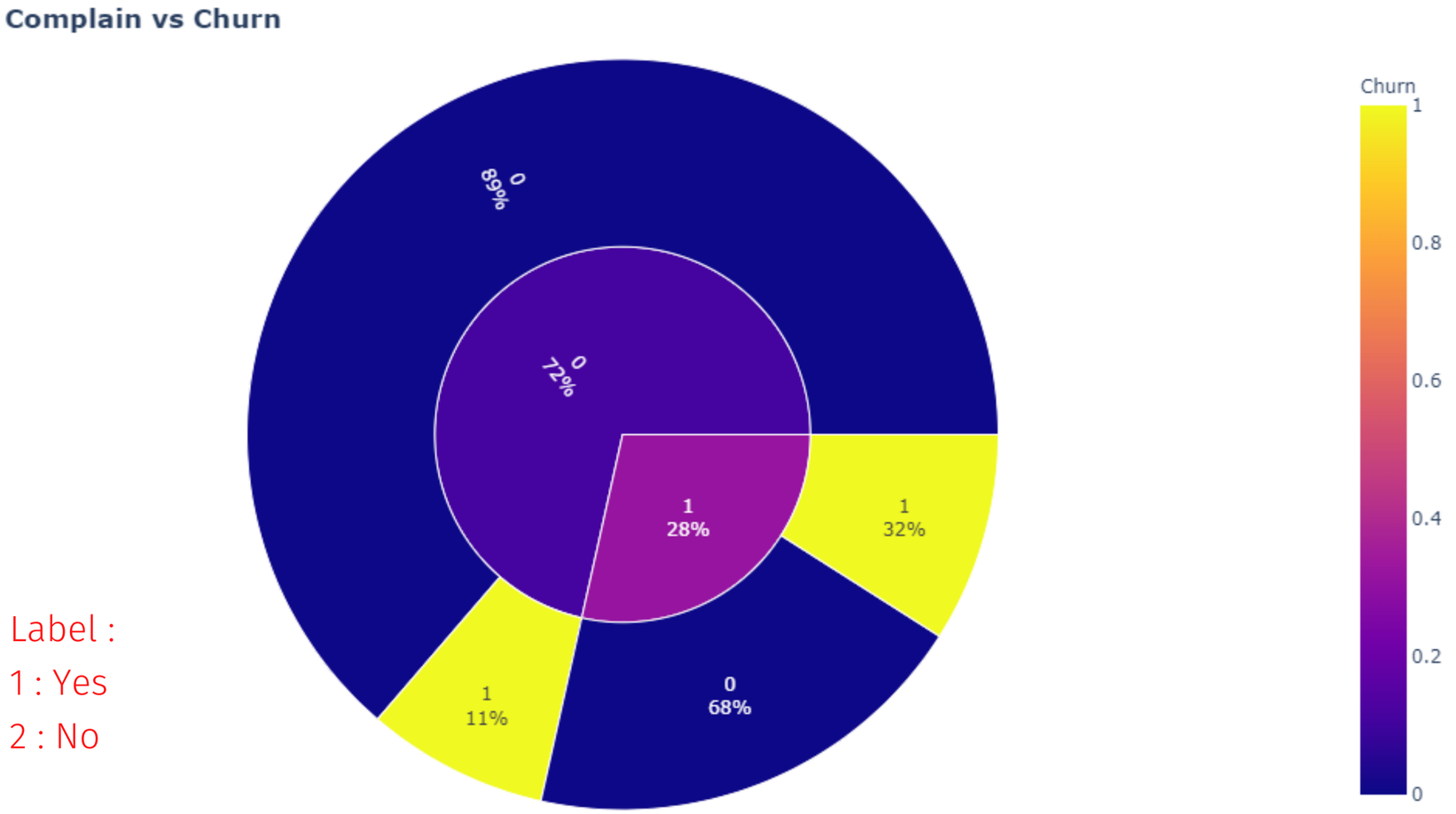
Correlation

Target

- All features are correlated weakly to each other
- **Highest correlation** for CouponUsed and OrderCount **at 0.75**

Bravo Team

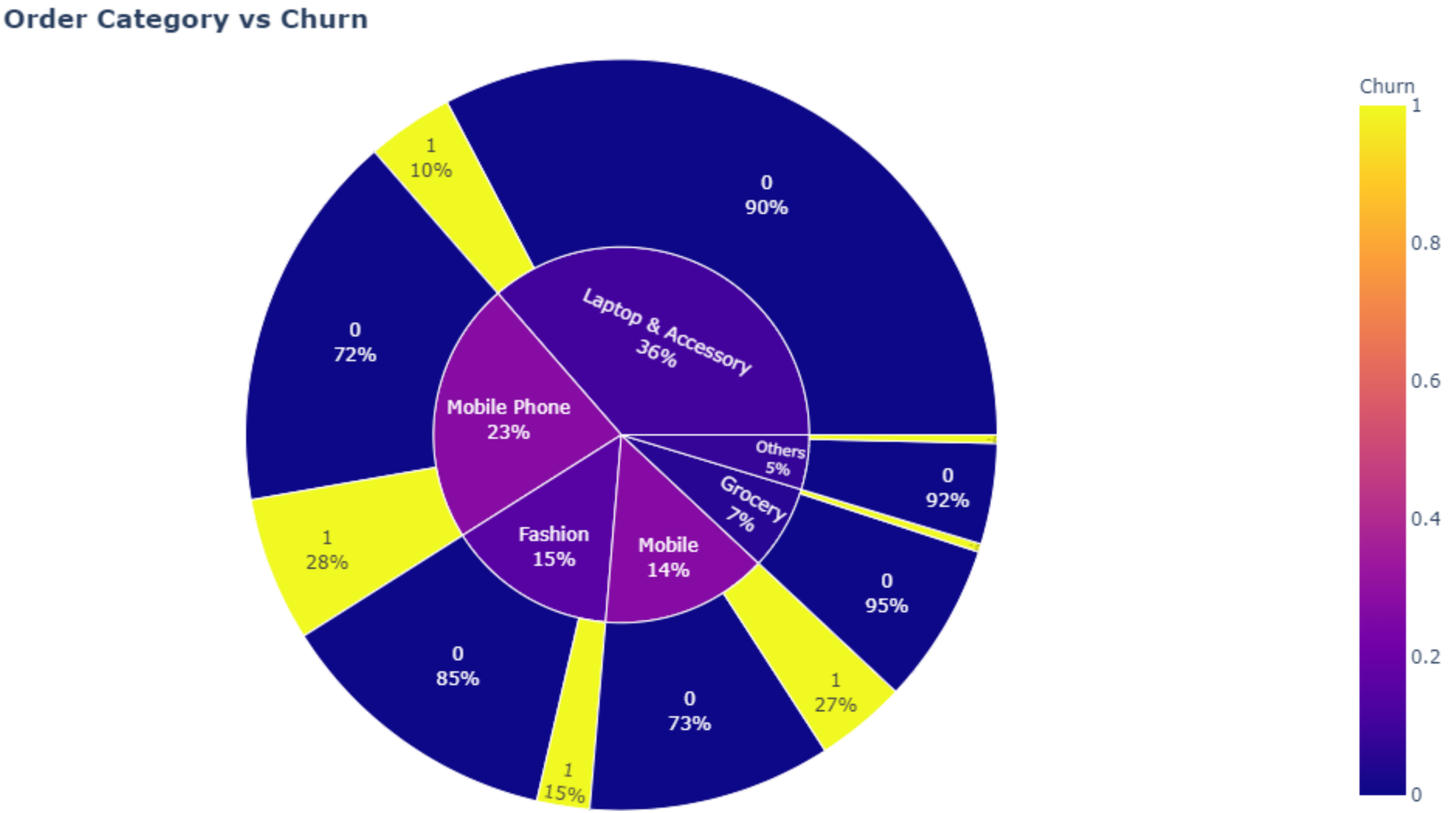
Insight



Label :
1 : Yes
2 : No

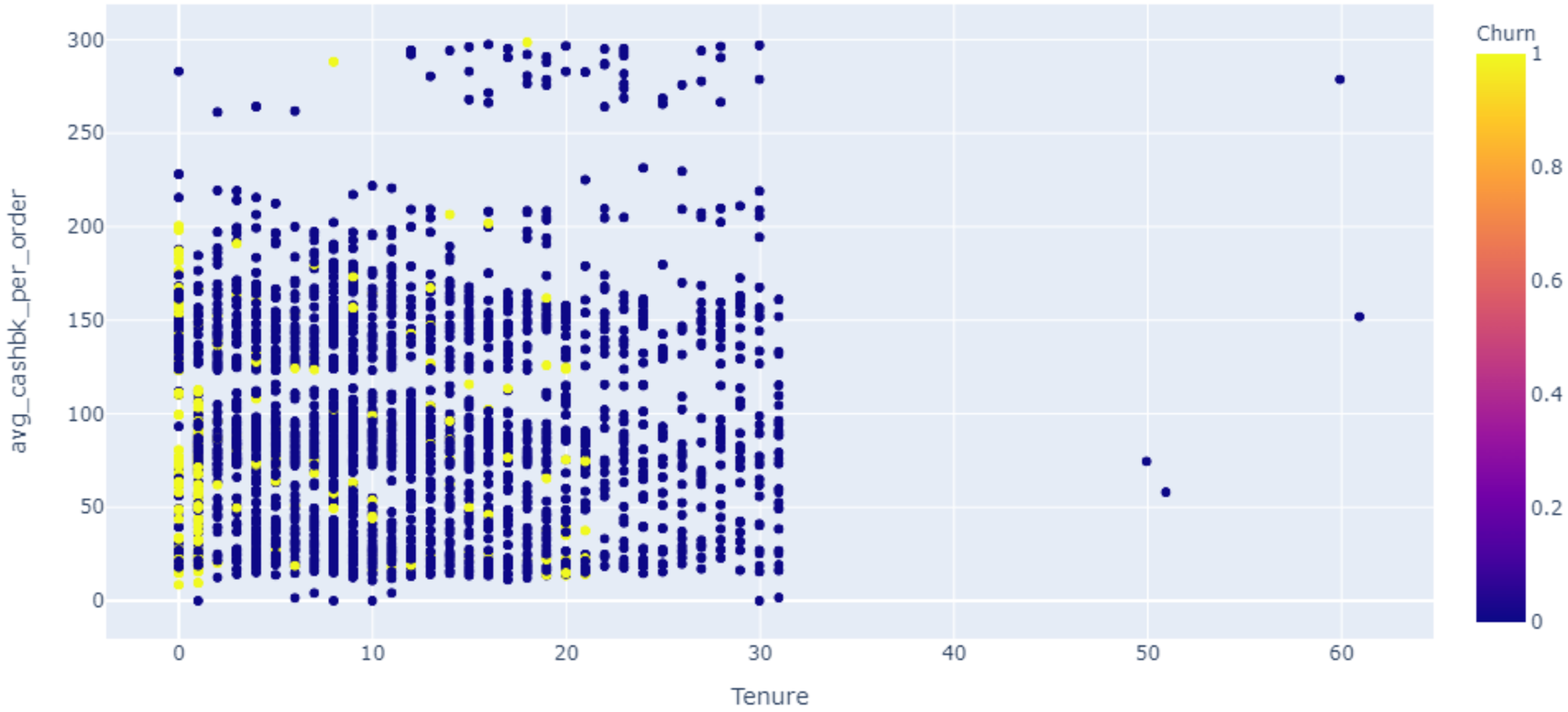
- This indicates that customers who **complained** in the last month tend to be **disloyal.**
- Otherwise, customers who do **not complain** tend to be **loyal.**

Insight



- Based on the ratio per order category, customers who ordered **Mobile Phone & Mobile** are indicated to be **disloyal**.
- Customers who ordered **Grocery** indicated to be **loyal**.

Insight



- Customers who have a **longer tenure & get more average cashback** tend to be **loyal**.
- Otherwise, Customers who have a **short tenure & get less average cashback** tend to be **disloyal.**

04
MODELING

ML Comparison - Classifier

CatboostClassifier outperformed all other ML models by AUC and Recall-score.

Model	Accuracy (train)	Accuracy (val)	Precision (train)	Precision (val)	Recall (train)	Recall (val)	F1 (train)	F1 (val)	roc_auc (train)	roc_auc (val)
LogisticRegression	0,828494	0,812001	0,821895	0,805198	0,838713	0,824521	0,830186	0,814257	0,90242	0,891793
DecisionTreeClassifier	1	0,844989	1	0,825804	1	0,874704	1	0,849426	1	0,844998
KNeighborsClassifier	0,871537	0,783653	0,849149	0,769757	0,903697	0,808757	0,875515	0,788482	0,941767	0,856254
AdaBoostClassifier	0,862963	0,825866	0,86072	0,821752	0,866098	0,832485	0,863383	0,827066	0,940366	0,906306
CatBoostClassifier	0,986478	0,907649	0,979216	0,899419	0,994063	0,918221	0,986578	0,908687	0,999156	0,959595

But we prefer **logistic regression** because it is more familiar to us. And tend not to overfitting like other models(DecisionTree, KNN, and CatBoost).

Description :

- **Accuracy** is percentage prediction were correct.
- **AUC (Area Under the Curve)** means model has a good measure of separability if it has AUC near to 1.
- **Recall** is actual positive rate.
- **Precision** is predicted positive rate.
- **F1-Score** combines Recall and Precision to one performance matrix.

Best Parameter Hyperparameter Tuned for the Model Logistic Regression

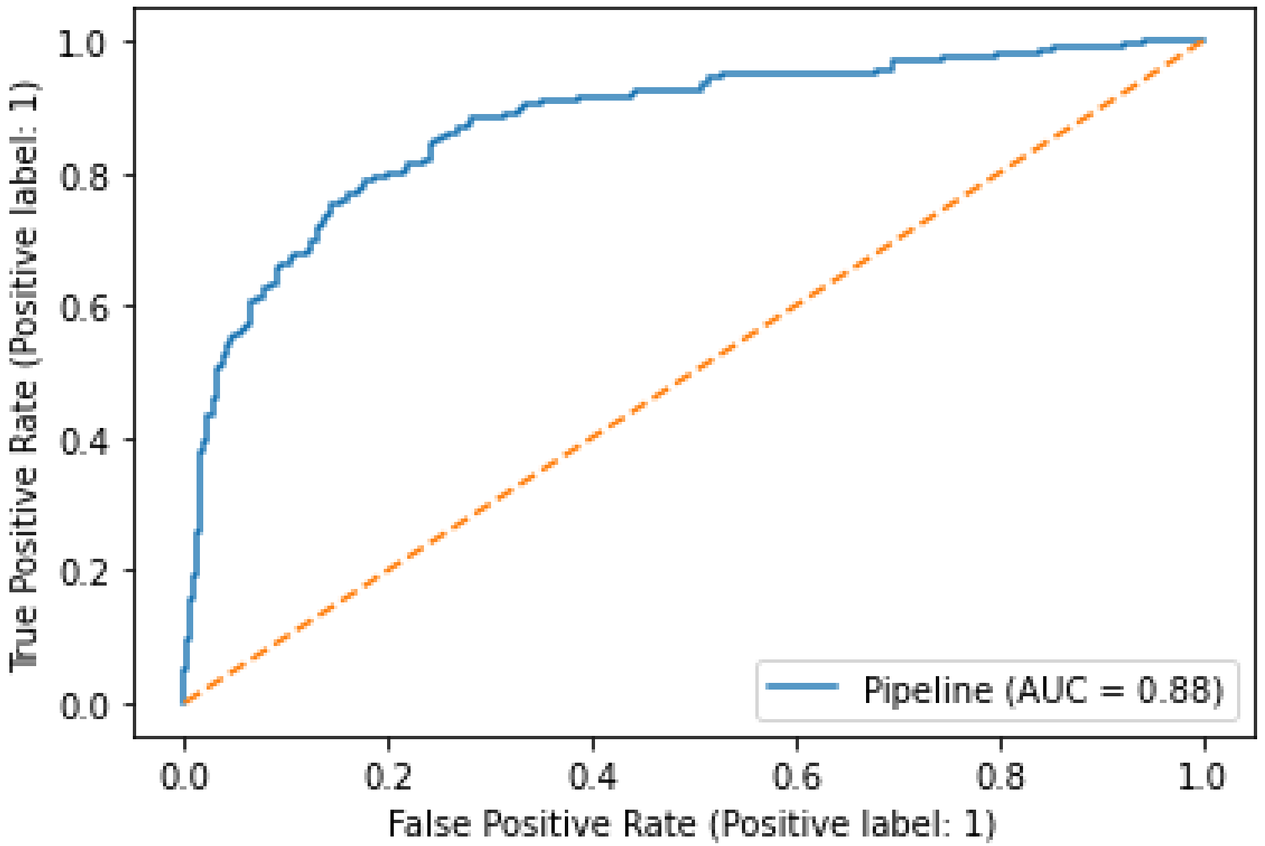
Parameters

```
Out[41]: GridSearchCV(cv=5,
                      estimator=Pipeline(steps=[('full_pipeline',
                                                 ColumnTransformer(transformers=[('ctg_transformer',
                                                                                  OneHotEncoder(),
                                                                                  ['PreferredLoginDevice',
                                                                                   'PreferredPaymentMode',
                                                                                   'Gender',
                                                                                   'PreferedOrderCat',
                                                                                   'MaritalStatus']),
                                                                                 ('num_transformer',
                                                                                  RobustScaler(),
                                                                                  ['CityTier',
                                                                                   'NumberOfDeviceRegistered',
                                                                                   'SatisfactionScore',
                                                                                   'NumberOfAddress',
                                                                                   'CashbackAmount',
                                                                                   '...houseToHome',
                                                                                   'HourSpendOnApp',
                                                                                   'Tenure',
                                                                                   'OrderAmountHikeFromlastYear',
                                                                                   'OrderCount',
                                                                                   'DaySinceLastOrder',
                                                                                   'avg_cashbk_per_order',
                                                                                   'Complain'])])),
                                                ('log_reg', LogisticRegression())]),
                      n_jobs=-1,
                      param_grid={'log_reg__C': [1, 2, 4, 8, 10, 20, 30, 40],
                                  'log_reg__penalty': ['none', 'l1', 'l2', 'elasticnet'],
                                  'log_reg__solver': ['newton-cg', 'lbfgs', 'liblinear',
                                                      'sag', 'saga']})
```

Best Parameters

```
Out[42]: Pipeline(steps=[('full_pipeline',
                          ColumnTransformer(transformers=[('ctg_transformer',
                                                           OneHotEncoder(),
                                                           ['PreferredLoginDevice',
                                                            'PreferredPaymentMode',
                                                            'Gender', 'PreferedOrderCat',
                                                            'MaritalStatus']),
                                                          ('num_transformer',
                                                           RobustScaler(),
                                                           ['CityTier',
                                                            'NumberOfDeviceRegistered',
                                                            'SatisfactionScore',
                                                            'NumberOfAddress',
                                                            'CashbackAmount',
                                                            'WarehouseToHome',
                                                            'HourSpendOnApp', 'Tenure',
                                                            'OrderAmountHikeFromlastYear',
                                                            'OrderCount',
                                                            'DaySinceLastOrder',
                                                            'avg_cashbk_per_order',
                                                            'Complain'])])),
                         ('log_reg',
                          LogisticRegression(C=1, penalty='l1', solver='liblinear'))])
```

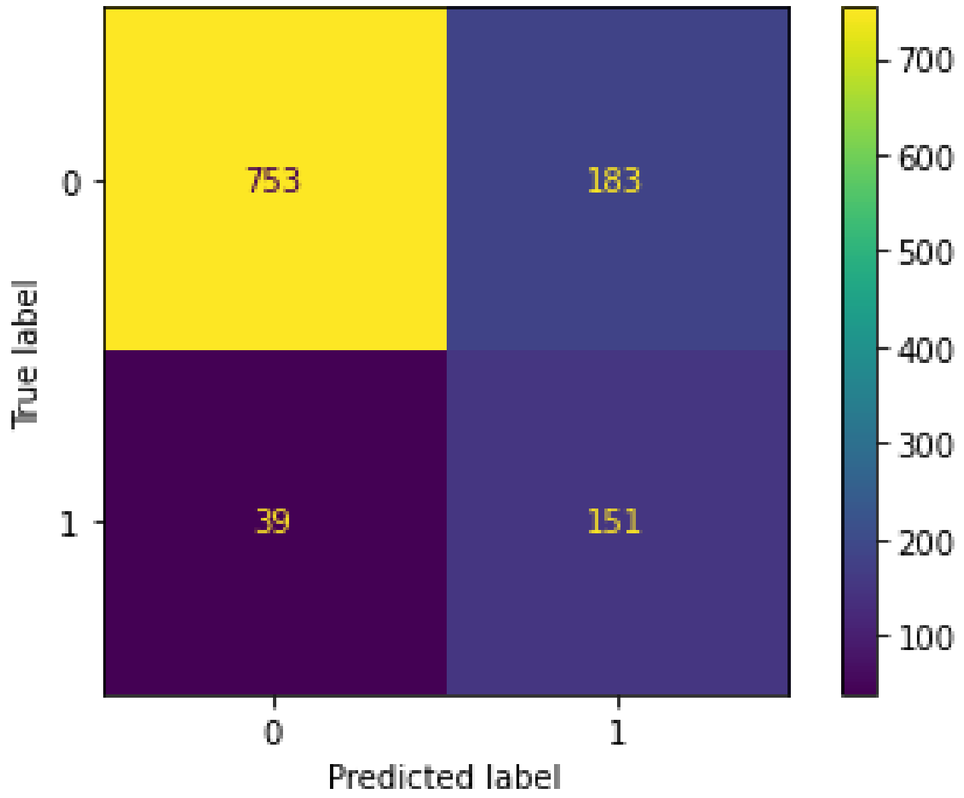
AUC - Logistic Regression



AUC - Score = 0.88

The model is good at **differentiate** which **will churn** and which will **not churn**

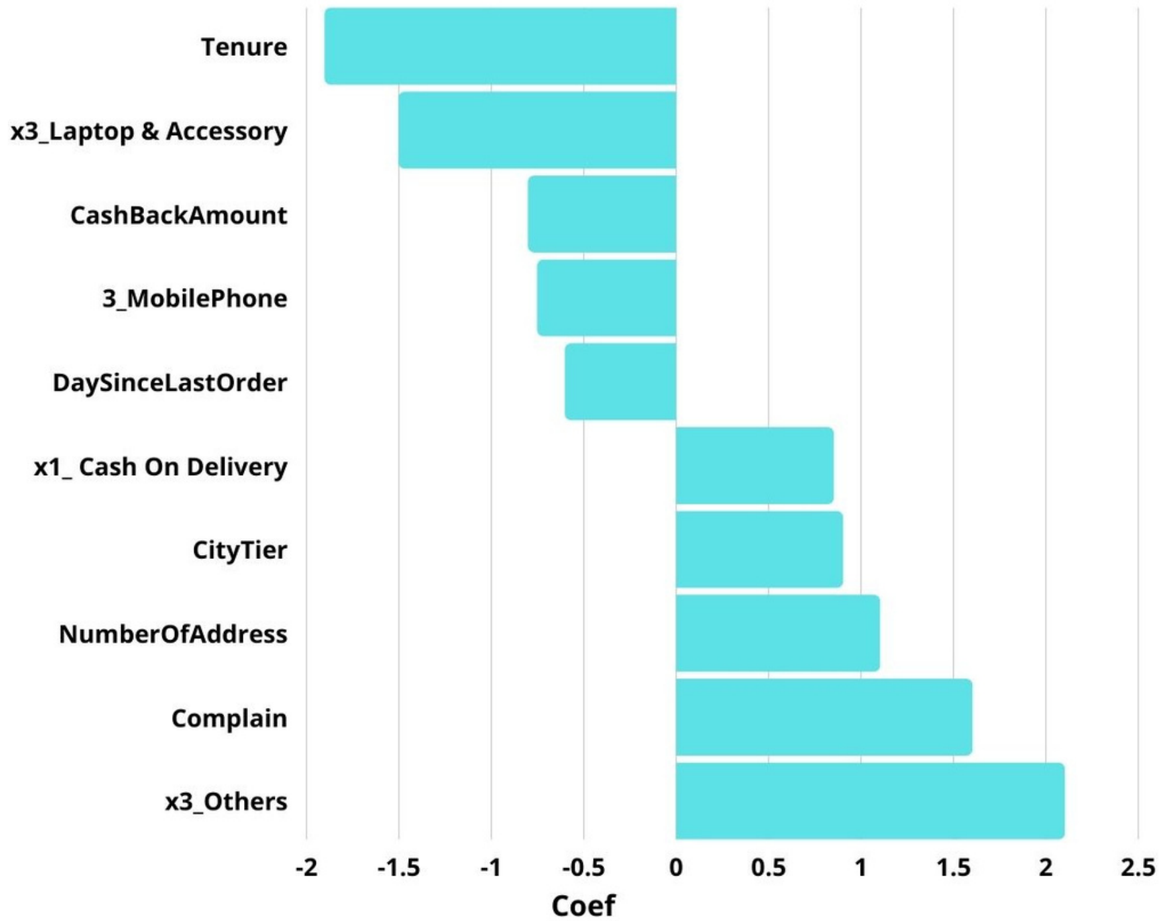
Confusion Matrix - Logistic Regression



We've got :

- **151 for True Positive**, that means 151 customers who were predicted to churn, actually will churn
- **39 for False Negative,** that means that 39 customers who were predicted not to churn, actually will churn.
- **753 for True Negative,** that means that 753 customers who are predicted to not churn, actually will not churn.
- **183 for False Positive,** that means that 183 customers who are predicted to churn, actually will not churn.

Feature Importance



Negative Class :

- Tenure features has **greatly affect** of **negative class.**
- Laptop & Accessory features **has affect** of **negative class.**

Positive Class :

- Others features has **greatly affect** of **positive class.**
- Complain features has **affect** of **negative class**.

05
BUSINESS INSIGHT & RECOMMENDATION

Key Finding

- Customers who **ordered laptop & accessories product categories** in last month tend to be more **loyal.**
- Customers who have a **longer tenure & get more average cashback** tend to be **loyal**.
- Customers who have a **short tenure & get less average cashback** tend to be **disloyal.**
- Customers who **ordered Mobile Phone & Mobile** are indicated to be **disloyal.**
- Customers who **ordered Groceries** indicated to be **loyal**.
- Customers who **complained** in the last month tend to be **disloyal.**
- Based on data from www.statista.com, "Average value of daily customer transaction" each customer has a value of Rp. 300,000.00. And if we assume that the cost of the promotion spends the marketing cost of Rp. 300/customer. Based on our calculations, the company can prevent an income loss of **Rp. 34.099.500**.

Recomendation

- Provide **promos** to customers who are predicted to churn.
- Improve the **quality** of service to customers.

Reference

https://www.kaggle.com/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction

https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

https://www.statista.com/statistics/1242651/indonesia-average-daily-e-commerce-transaction-value/

Bravo Team

THANK YOU !