Churn Analysis & Prediction

Pongthawat Hanwathanawong

Agenda

Introduction

Data Exploration & Wrangling

Customer Churn Analysis

Customer Churn Prediction

Q&A

Introduction

Why we need Churn Analysis & Prediction?

- **Customer churn refer to lose of customer**

- To **retain current customer** often **cheaper than** to **acquire a new customer** also come up with loyalty benefit and better customer experience

- **It wasted money if** you target retention campaign to all the customer

  or customer who not going to churn

- **Target retention campaign to customer who have high risk of churning**

To identify risk of Churning → Churn analysis & prediction

Data Exploration and Wrangling

Data description

	Variable	Descripttion
0	CustomerID	Unique customer ID
1	Churn	Churn Flag
2	Tenure	Tenure of customer in organization
3	PreferredLoginDevice	Preferred login device of customer
4	CityTier	City tier
5	WarehouseToHome	Distance in between warehouse to home of customer
6	PreferredPaymentMode	Preferred payment method of customer
7	Gender	Gender of customer
8	Age	Age of customer
9	SizeofFamily	Gender of customer
10	HourSpendOnApp	Number of hours spend on mobile application or website
11	NumberOfDeviceRegistered	Total number of deceives is registered on particular customer

	Variable	Descripttion
12	PreferedOrderCat	Preferred order category of customer in last month
13	SatisfactionScore	Satisfactory score of customer on service
14	MaritalStatus	Marital status of customer
15	NumberOfAddress	Total number of added added on particular customer
16	Complain	Any complaint has been raised in last month
17	OrderAmountHikeFromlastYear	Percentage increases in order from last year
18	CouponUsed	Total number of coupon has been used in last month
19	OrderCount	Total number of orders has been places in last month
20	DaySinceLastOrder	Day Since last order by customer
21	CashbackAmount	Average cashback in last month
22	DayLogin	Day of Login on mobile app
23	QTY	Number of quantity
24	LastDate	Last date

	CustomerID	Tenure	PreferredLoginDevice	CityTier	WarehouseToHome	PreferredPaymentMode	Gender	Age	SizeofFamily	HourSpendOnApp	...
0	50001	4.0	Mobile Phone	3	6.0	Debit Card	Female	NaN	2	3.0	...
1	50002	NaN	Phone	1	8.0	UPI	Male	21.0	2	3.0	...
2	50003	NaN	Phone	1	30.0	Debit Card	Male	52.0	5	2.0	...
3	50004	0.0	Phone	3	15.0	Debit Card	Male	63.0	1	2.0	...
4	50005	0.0	Phone	1	12.0	CC	Male	23.0	1	NaN	...
...	...	...	...	...	...	...	...	...	...	...	...
5625	55626	10.0	Computer	1	30.0	Credit Card	Male	19.0	1	3.0	...
5626	55627	13.0	Mobile Phone	1	13.0	Credit Card	Male	44.0	5	3.0	...
5627	55628	1.0	Mobile Phone	1	11.0	Debit Card	Male	53.0	1	3.0	...
5628	55629	23.0	Computer	3	9.0	Credit Card	Male	72.0	3	4.0	...
5629	55630	8.0	Mobile Phone	1	15.0	Credit Card	Male	56.0	2	3.0	...

5630 rows × 25 columns

```
RangeIndex: 5630 entries, 0 to 5629
Data columns (total 25 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   CustomerID                  5630 non-null   int64
 1   Tenure                      5366 non-null   float64
 2   PreferredLoginDevice        5630 non-null   object
 3   CityTier                    5630 non-null   int64
 4   WarehouseToHome             5379 non-null   float64
 5   PreferredPaymentMode        5630 non-null   object
 6   Gender                      5630 non-null   object
 7   Age                         5629 non-null   float64
 8   SizeofFamily                5630 non-null   int64
 9   HourSpendOnApp              5375 non-null   float64
 10  NumberOfDeviceRegistered    5630 non-null   int64
 11  PreferedOrderCat            5630 non-null   object
 12  SatisfactionScore           5630 non-null   int64
```

- 5630 Observations
- 25 Attributes
- No duplicated CustomerID
- Some missing value

Categorize attribute name by characteristic

```
norminal = ['PreferredLoginDevice','PreferredPaymentMode','Gender','PreferedOrderCat',
            'MaritalStatus','Complain']

ordinal = ['CityTier','SatisfactionScore']

numeric = ['Tenure','WarehouseToHome','Age','SizeofFamily','HourSpendOnApp',
           'NumberOfDeviceRegistered','NumberOfAddress','OrderAmountHikeFromlastYear',
           'CouponUsed','OrderCount','DaySinceLastOrder','CashbackAmount','DayLogin','QTY']

datetime = ['LastDate']

target = ['Churn']
```

Spelling Mistake on categorical data

```
------
-- Gender --
Male          3382
Female        2242
ผู้หญิง          3
ชาย              2
หญิง             1
```

```
-- Gender --
Male       3384
Female     2246
Name: Gender, dtype: int64
```

```
------
-- MaritalStatus --
Married          2985
Single           1792
Divorced          848
โสด                 4
แต่งงานแล้ว          1
Name: MaritalStatus, dtype: int64
------
```

```
-- MaritalStatus --
Married     2986
Single      1796
Divorced     848
Name: MaritalStatus, dtype: int64
```

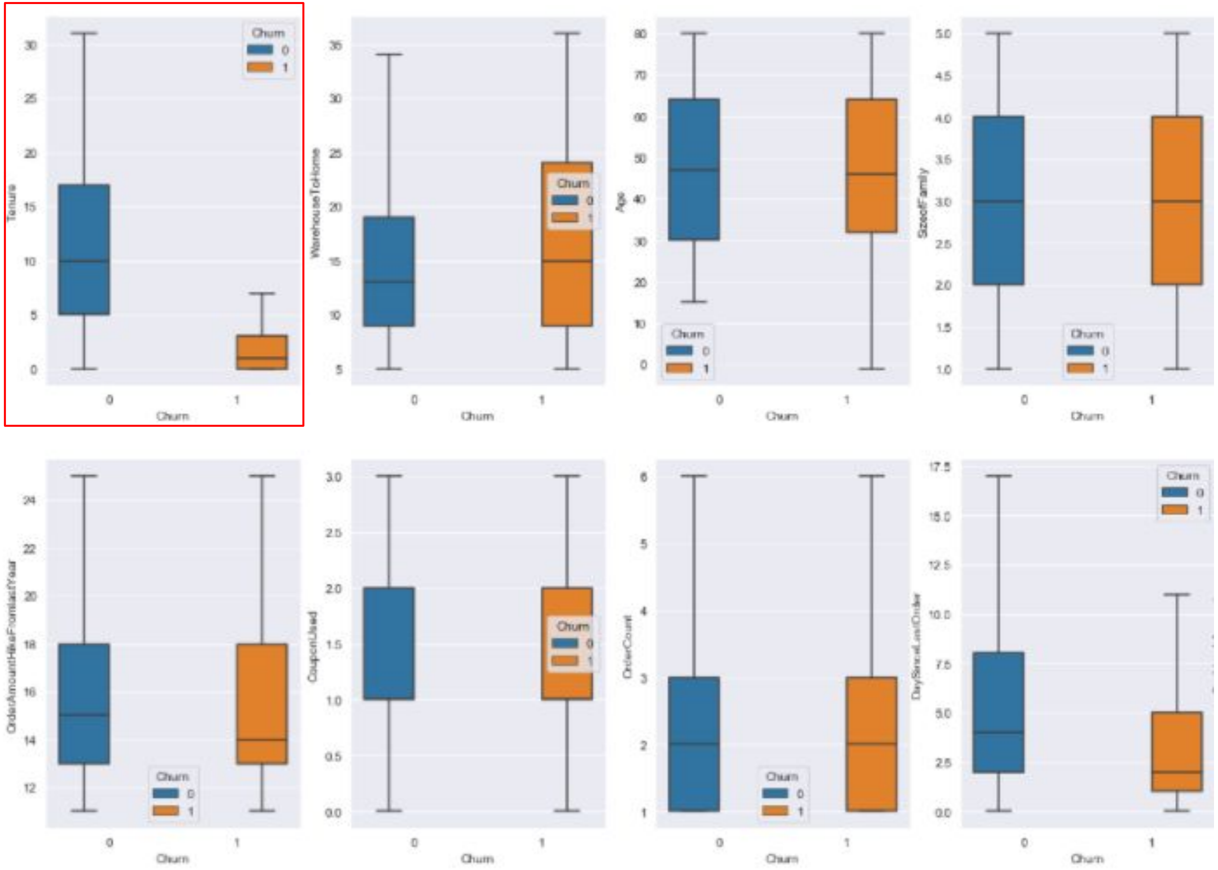
	count	mean	std	min	25%	50%	75%	max
Tenure	5366.0	10.189899	8.557241	0.0	2.00	9.000000	16.000000	61.00
WarehouseToHome	5379.0	15.639896	8.531475	5.0	9.00	14.000000	20.000000	127.00
Age	5629.0	47.283176	19.183838	-1.0	30.00	47.000000	64.000000	80.00
SizeofFamily	5630.0	3.019183	1.428707	1.0	2.00	3.000000	4.000000	5.00
HourSpendOnApp	5375.0	2.931535	0.721926	0.0	2.00	3.000000	3.000000	5.00
NumberOfDeviceRegistered	5630.0	3.688988	1.023999	1.0	3.00	4.000000	4.000000	6.00
NumberOfAddress	5630.0	4.214032	2.583586	1.0	2.00	3.000000	6.000000	22.00
OrderAmountHikeFromlastYear	5365.0	15.707922	3.675485	11.0	13.00	15.000000	18.000000	26.00
CouponUsed	5374.0	1.751023	1.894621	0.0	1.00	1.000000	2.000000	16.00
OrderCount	5372.0	3.008004	2.939680	1.0	1.00	2.000000	3.000000	16.00
DaySinceLastOrder	5323.0	4.543491	3.654433	0.0	2.00	3.000000	7.000000	46.00
CashbackAmount	5630.0	3249.088887	902.128997	0.0	2672.45	2993.466667	3600.529167	5958.15
DayLogin	5630.0	50.991119	29.112787	1.0	26.00	51.000000	77.000000	100.00
QTY	5630.0	5082.244760	7194.051053	-500.0	2519.00	4987.500000	7445.000000	500000.00

Abnormal on Numerical data

- Treat as a missing value (Replace with null)

  Dealing with it later

Customer Churn Analysis

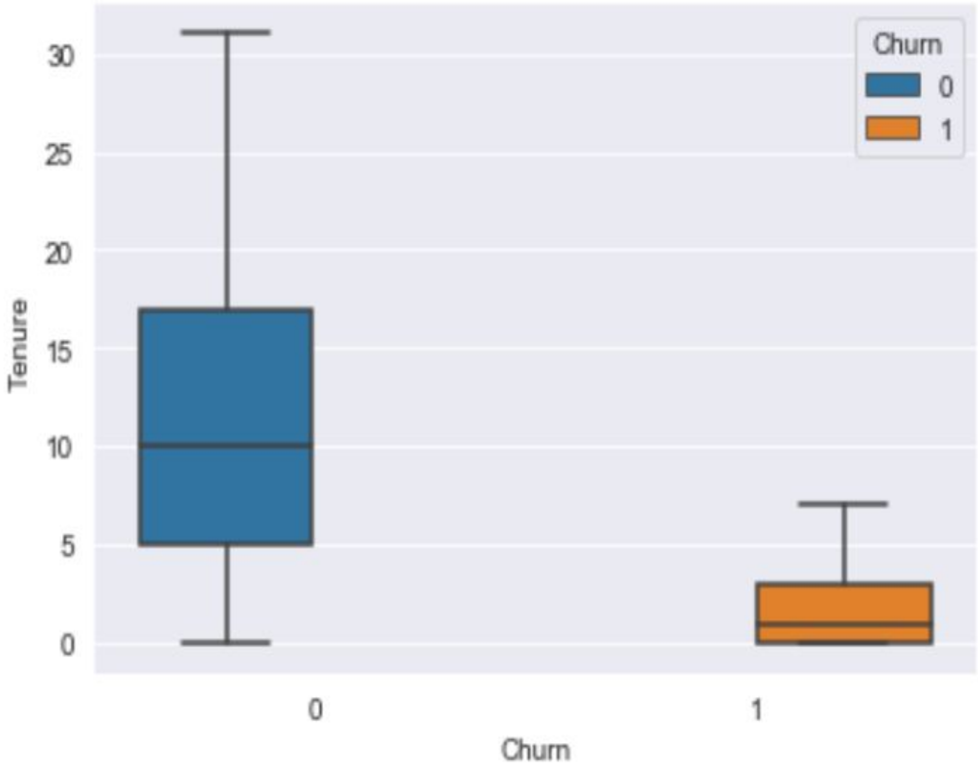


Relation between
Numeric Data and Target (Churn)

Box-plot

- Orange - Churn
- Blue - Not Churn

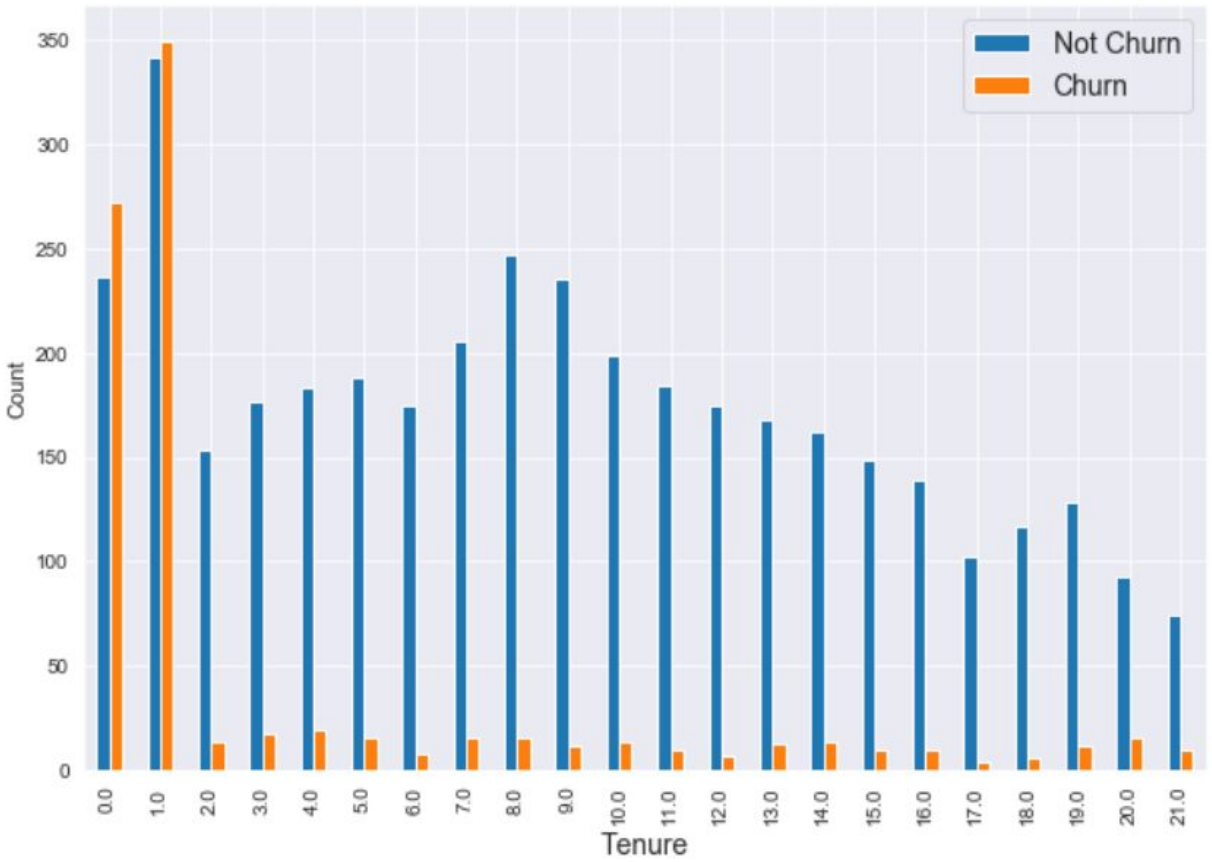
mostly no significant different that customer Churn or Not through each value of Numeric data



Tenure

Customer with **“short tenure”** have significantly “**higher churn rate”**

Short-term customer are much more likely to Churn



Bar chart

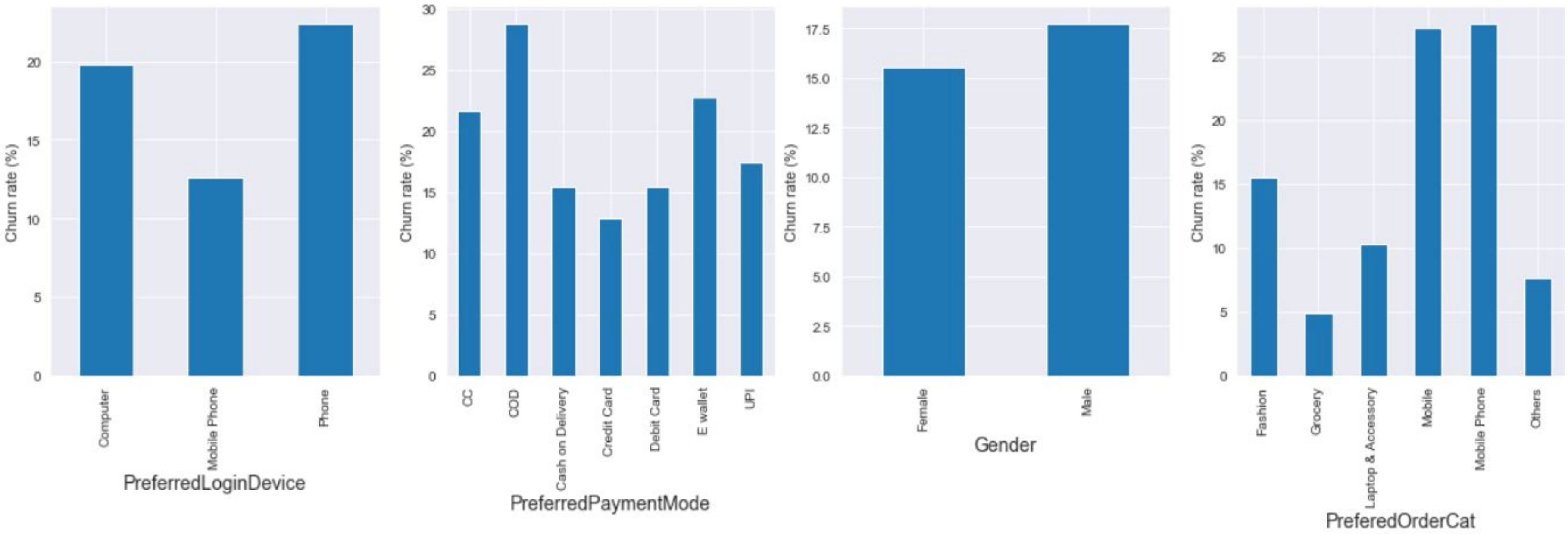
- **Tenure** vs Churn/Not churn

Customer with 0 - 1 tenure
have much higher Churn rate than customer with longer Tenure

If we can retain customer with 0, 1 long tenure to be 2 or more tenure the customer are more likely to become long-term customer

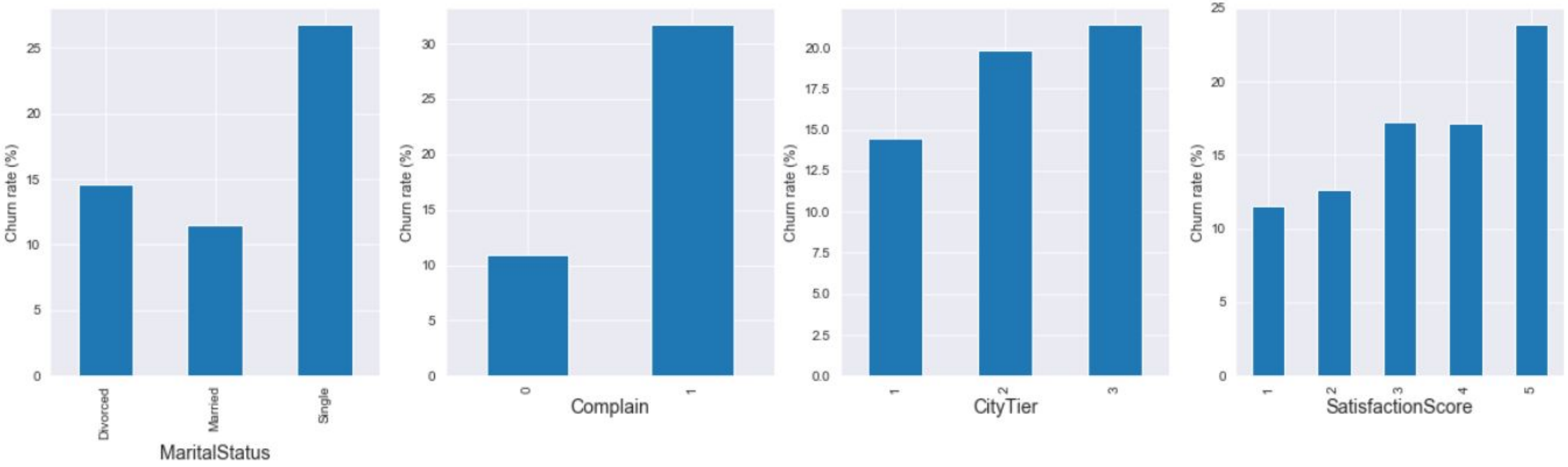
Relation between Category Data - Target (Churn)

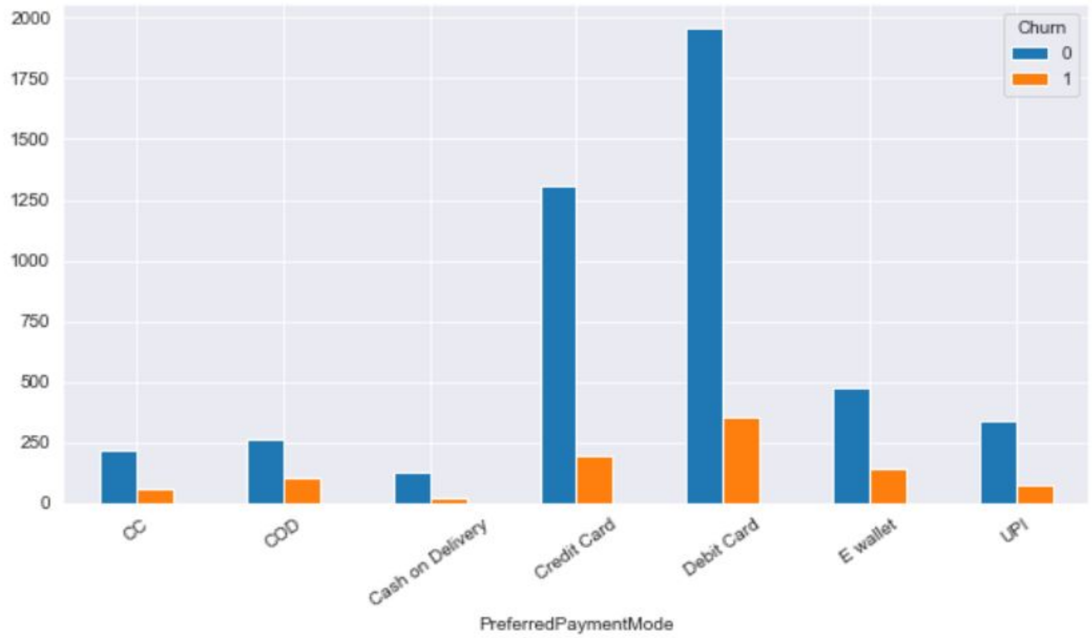
Bar Chart - Churn rate (%) of each value in each attribute



Relation between Category Data - Target (Churn)

Bar Chart - Churn rate (%) of each value in each variable (cont.)

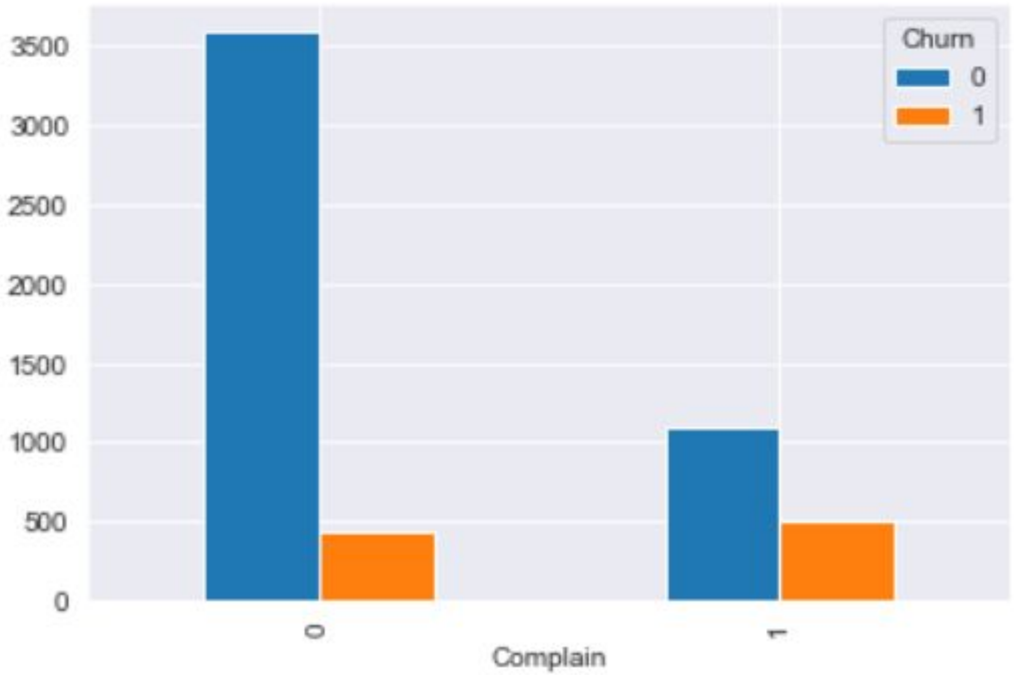




Preferred Payment Method

- More convenient payment method may cause lower risk of churning

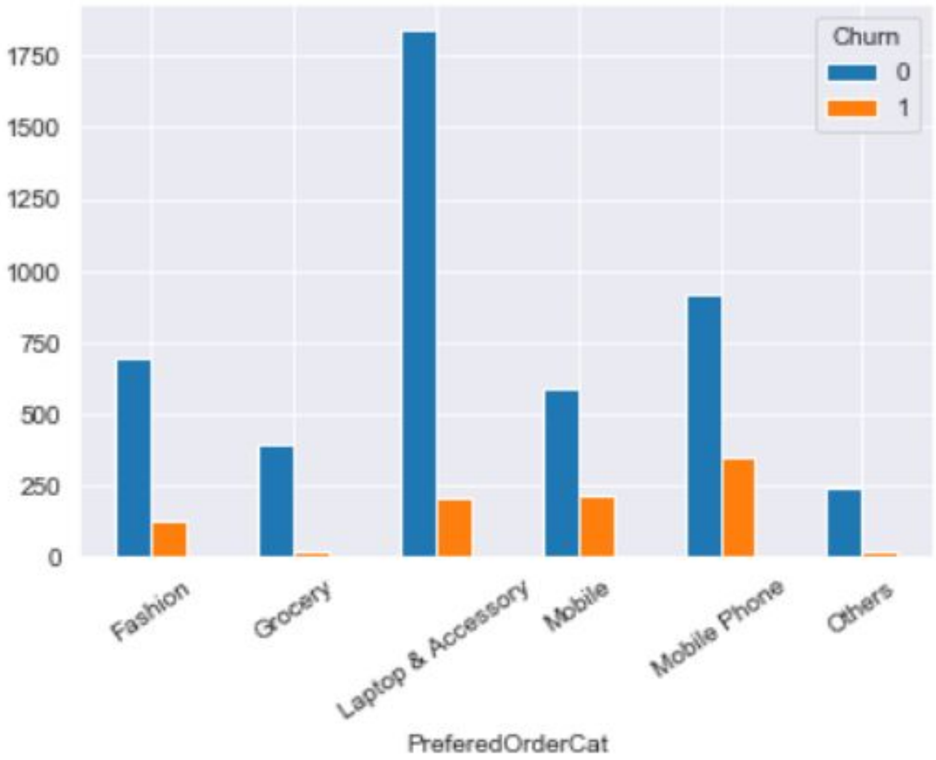
If you can make the customer connect with some payment method customer are more likely to have lower risk of churning



Complain

Obviously complaining refer to dissatisfaction of the customer which cause of churning

If there is any complaining occurs, you may need to suddenly handle it



PreferedOrderCat

Customer who preferred order in Mobile, Mobile Phone category in last month have higher Churn rate

It possible that the customer buy and switch to new mobile phone may not download and back to use the application

You may need to track the customer who buy a new phone and offer them to promotion of Mobile accessory

Customer Churn Prediction

Handle Outlier

Handle missing data

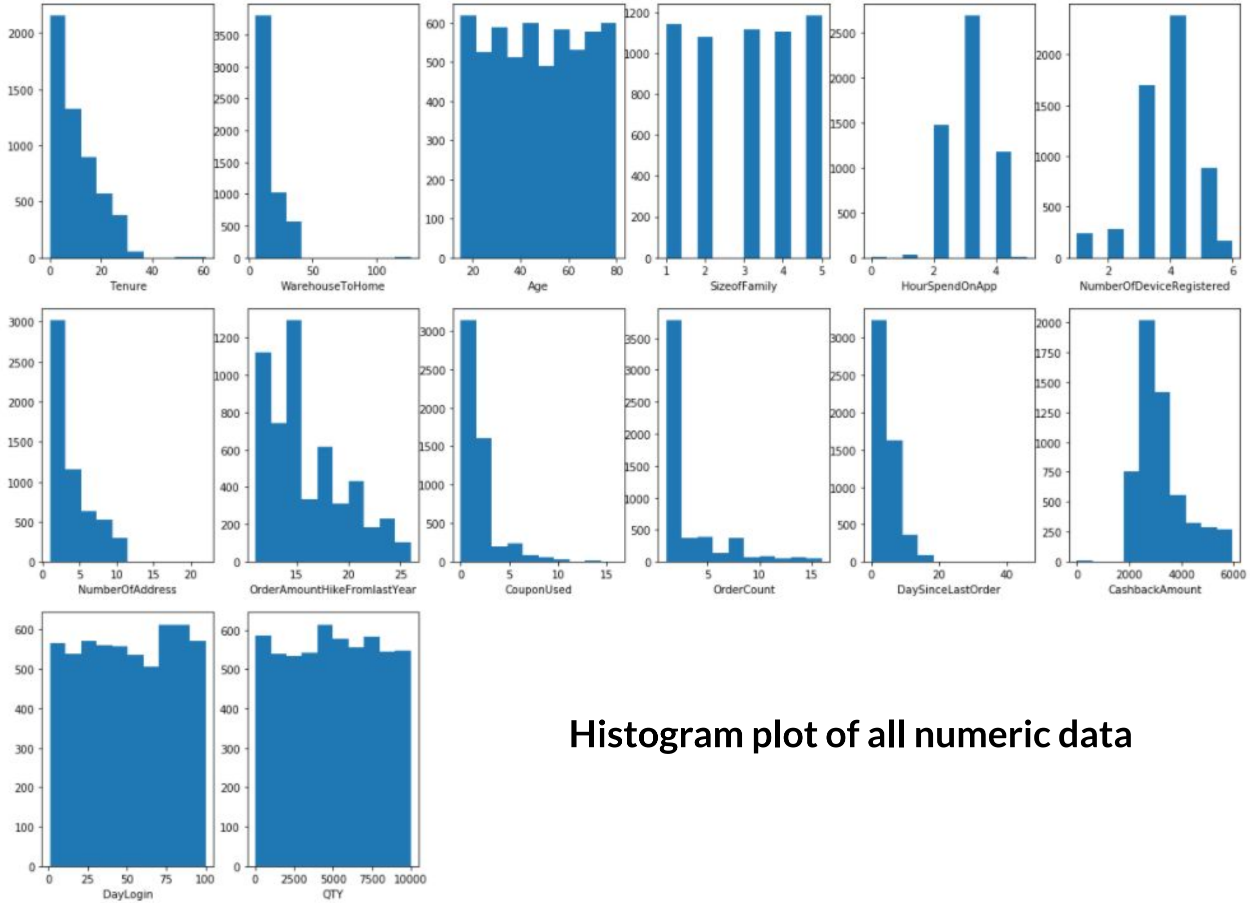
Modeling

Handle Outlier

- valid but extreme value

Some machine learning model are easily impacted by outliers

Removing → lost in information   “Trade-off”



Histogram plot of all numeric data

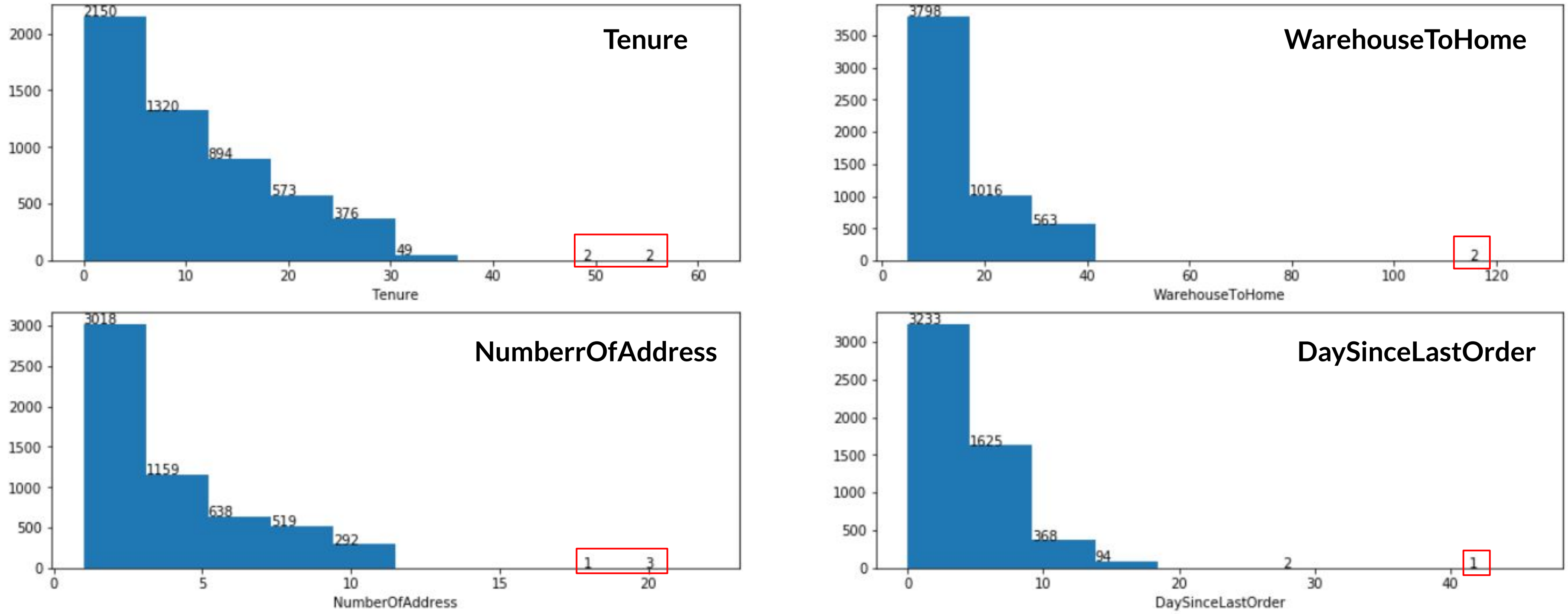
More than half of the variables have right skew distribution

Use Box-plot with **[Q3+3*IQR, Q1-3*IQR]** as upper and lower limit

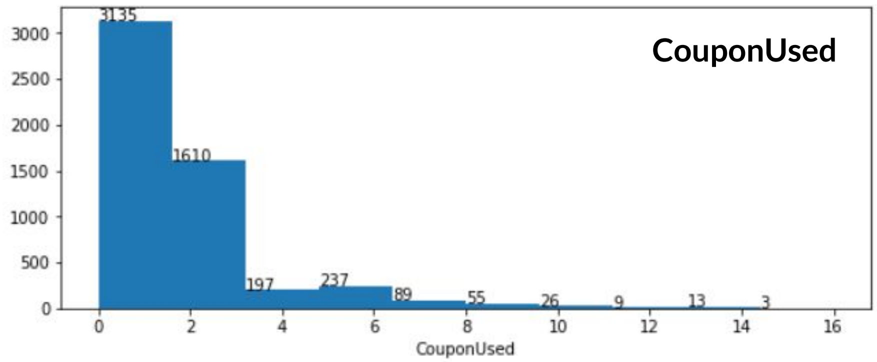
Number of observations below lower limit and above upper limit

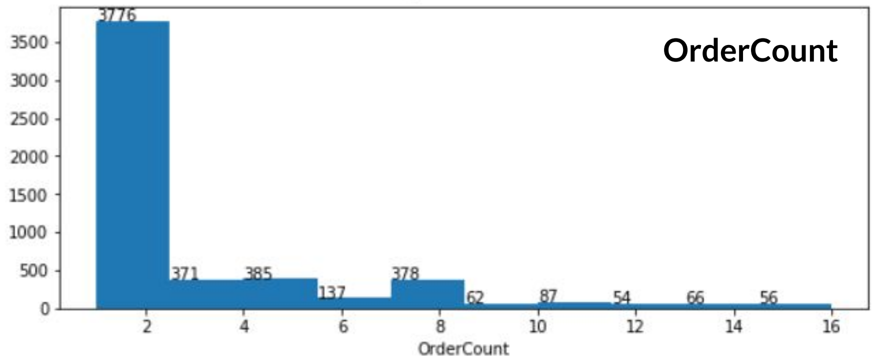
```
Tenure
2
WarehouseToHome
2
NumberOfAddress
4
CouponUsed
303
OrderCount
263
DaySinceLastOrder
3
```

Consider as Outliers

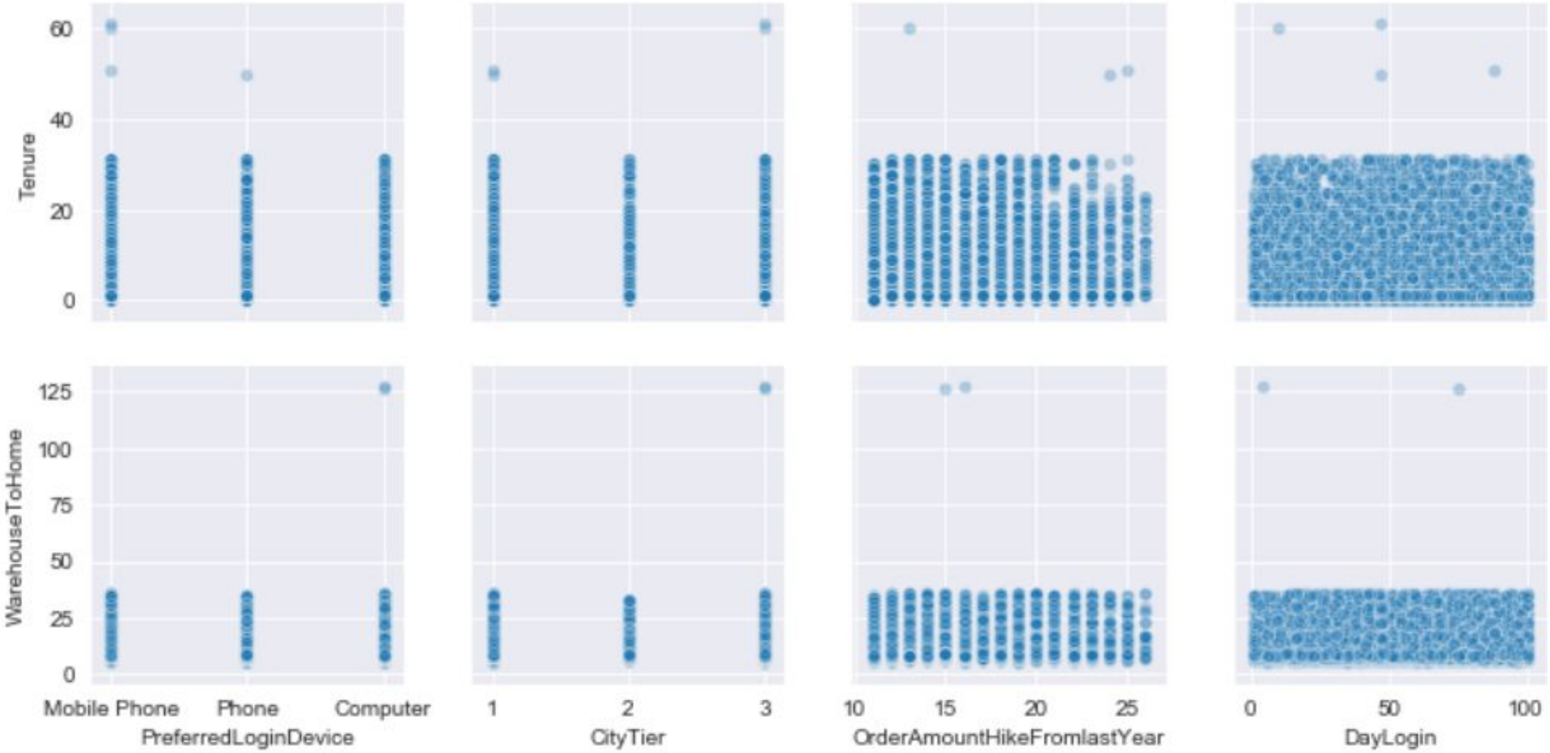


Not consider as Outlier

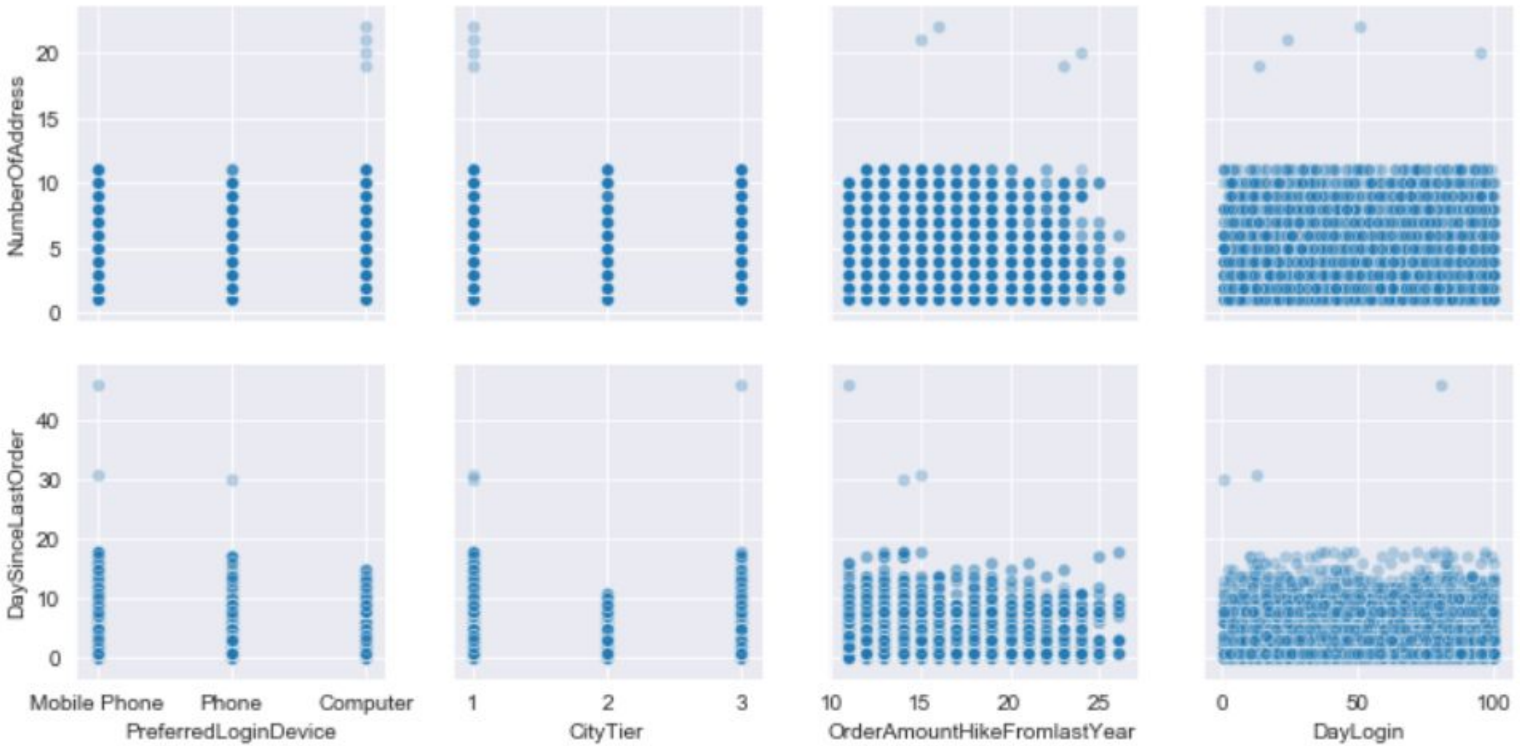




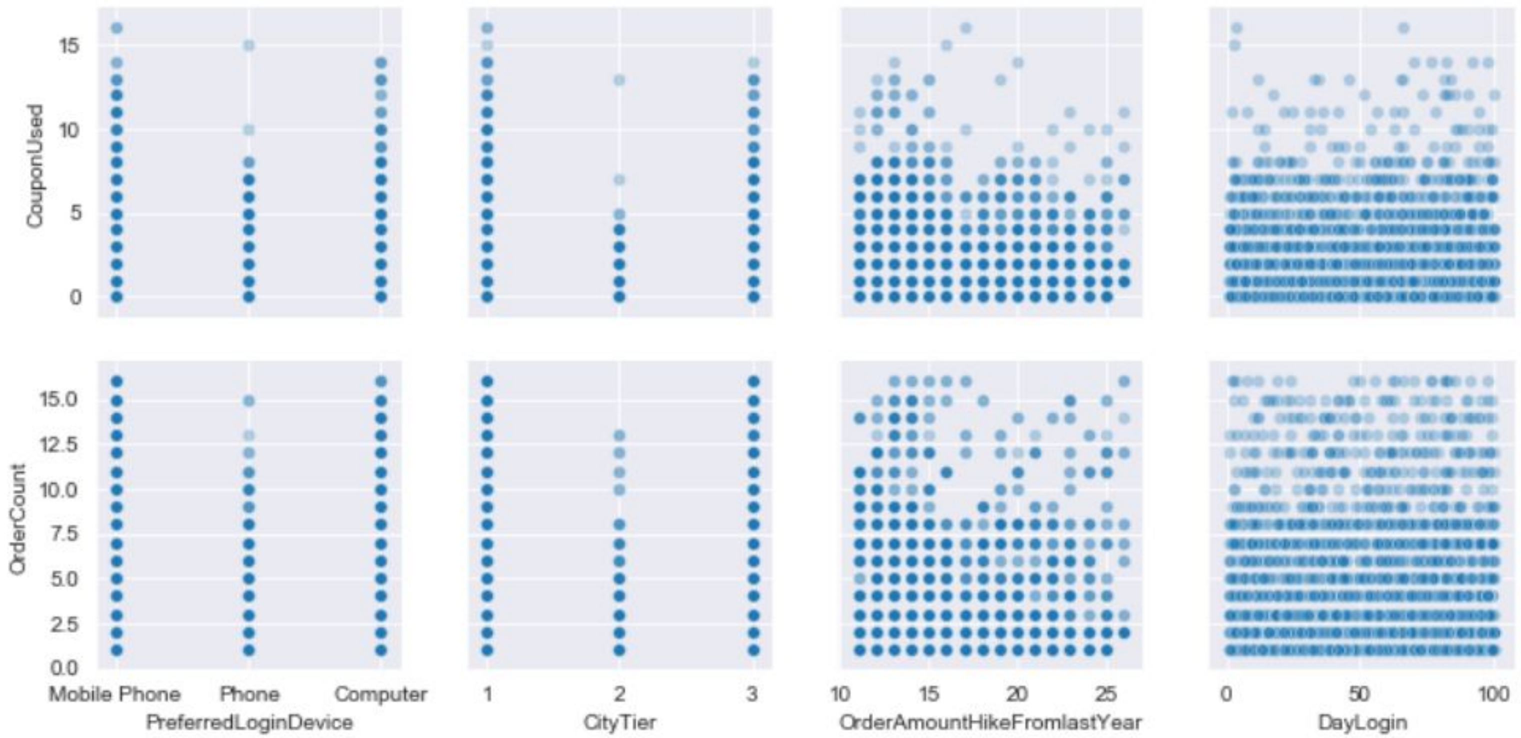
Consider as Outliers

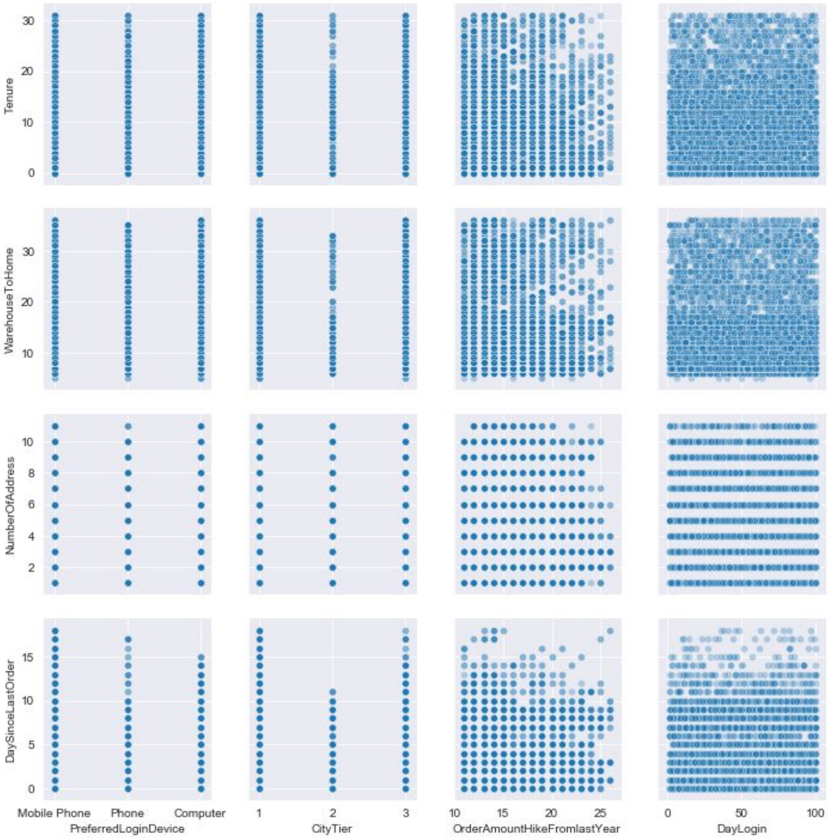


Consider as Outliers



Not consider as Outlier





After Drop the observations that I considered as a outlier

Handle Missing

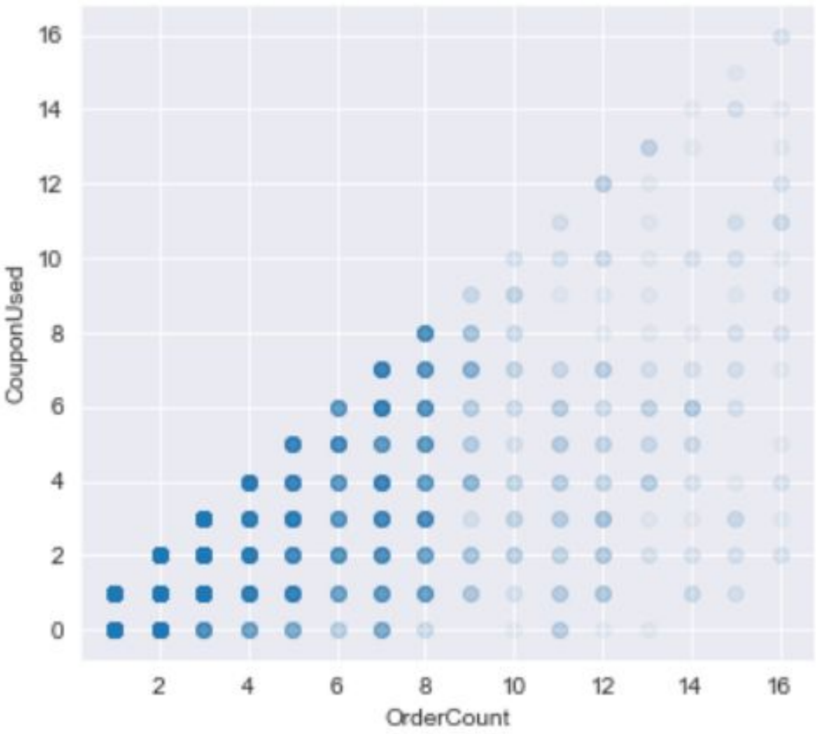
Dropping is not a good idea - lost too much information

Impute with a measure of central tendency: **Median**

```
Number of Observation
Before drop missing values: 5617
After drop missing values: 3761
```

```
-- Missing Values --
Tenure                          263
WarehouseToHome                 251
Age                               2
HourSpendOnApp                  255
OrderAmountHikeFromlastYear     264
CouponUsed                      255
OrderCount                      258
DaySinceLastOrder               307
QTY                               3
dtype: int64
```

Looking for relationship between variable with missing value and all other variable



Order >= Coupon

Use this constraint to impute both OrderCount and CouponUsed

	CouponUsed	OrderCount
467	NaN	3.0
782	NaN	3.0

Impute with median of CouponUsed which lower than 3

	CouponUsed	OrderCount
419	7.0	NaN
713	7.0	NaN

Impute with median of OrderCount which higher than 7

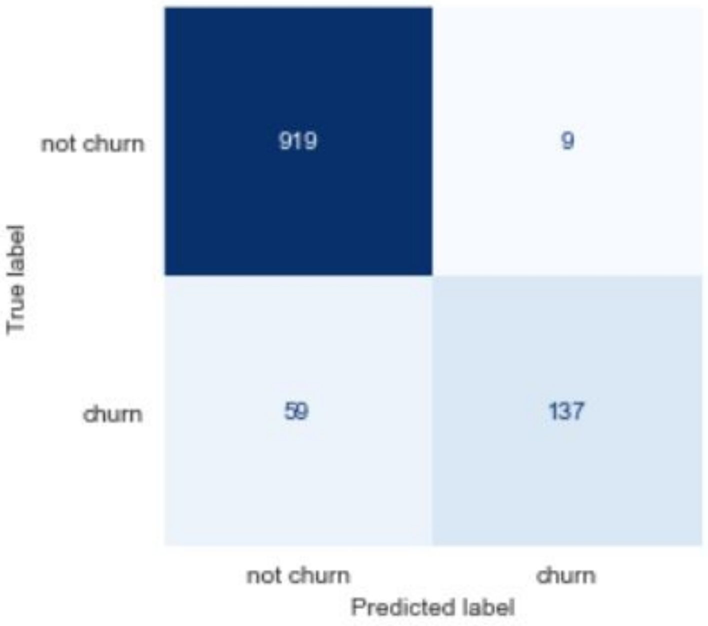
Modelling

Classification Method

- Logistic Regression
- Decision Tree
- Random Forest

Interpretable
well-know

Metric to evaluate the model



low precision - **Predicted** Churn, **True** is Not Churn

wasted money on retention target

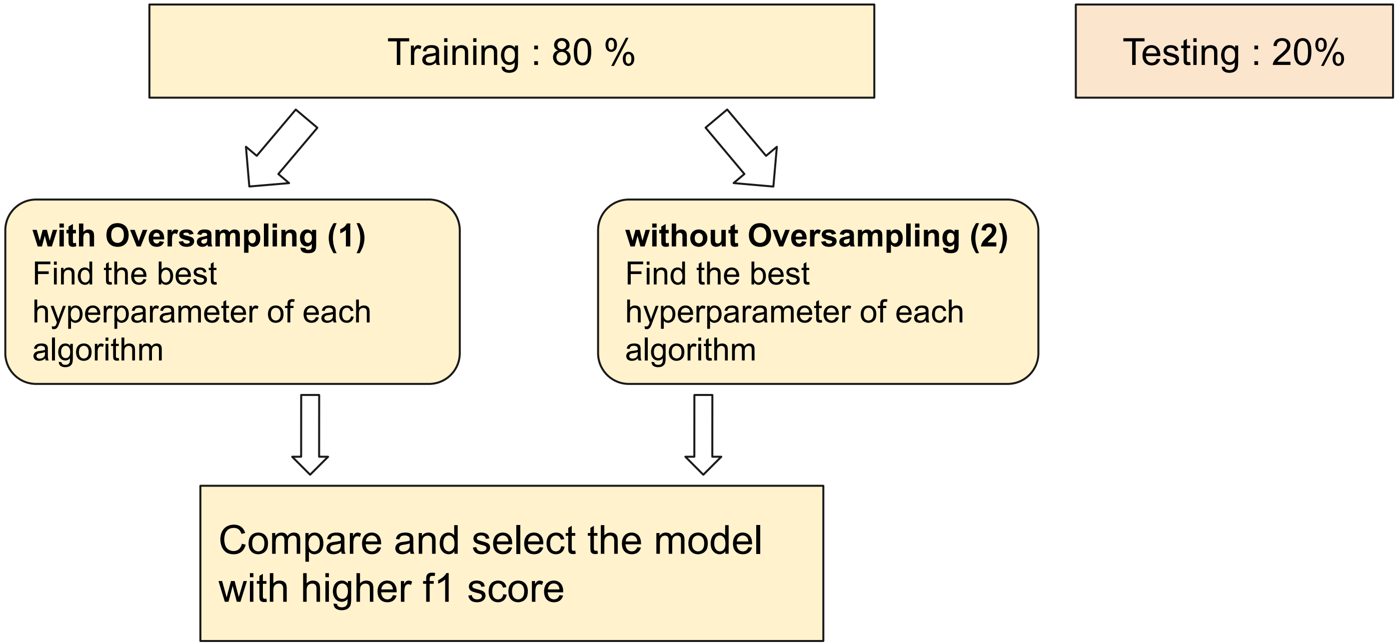
low recall - **Predicted** Not Chrun, **True** is Churn

lose customer

both are importance decide to use **f-1 score** as a metrics

F1 = harmonic mean ระหว่าง precision และ recall

Process


Training : 80 %
Testing : 20%
with Oversampling (1)
Find the best hyperparameter of each algorithm
without Oversampling (2)
Find the best hyperparameter of each algorithm
Compare and select the model with higher f1 score

with Oversampling (1)

Pipeline

One-hot encoding ⇩ Oversampling ⇩ Standard Scalar ⇩ gridsearchCV

- for categorical data
- for numeric data
- use **f1** to select the best hyperparameter

Imbalance dataset impact on some ML algorithm

https://www.researchgate.net/figure/Differences-between-undersampling-and-oversampling_fig1_340978368

without Oversampling (2)

Pipeline

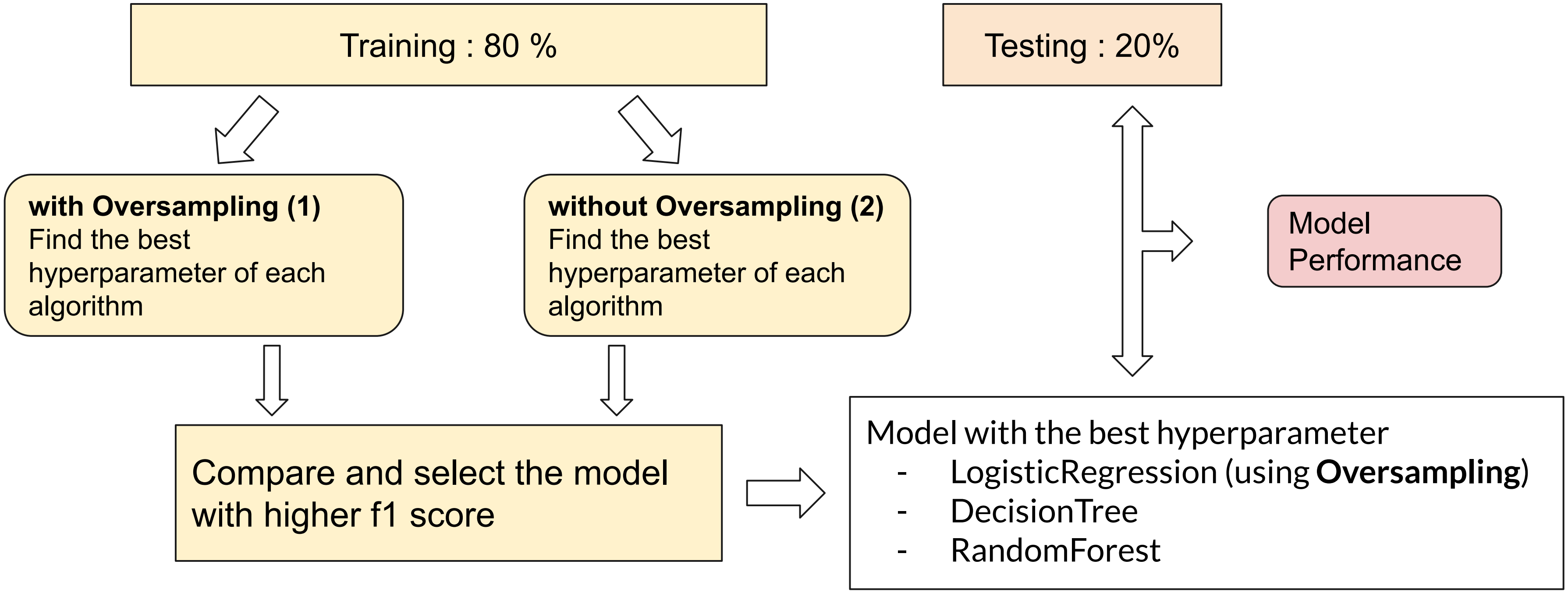
One-hot encoding - for categorical data

⇩

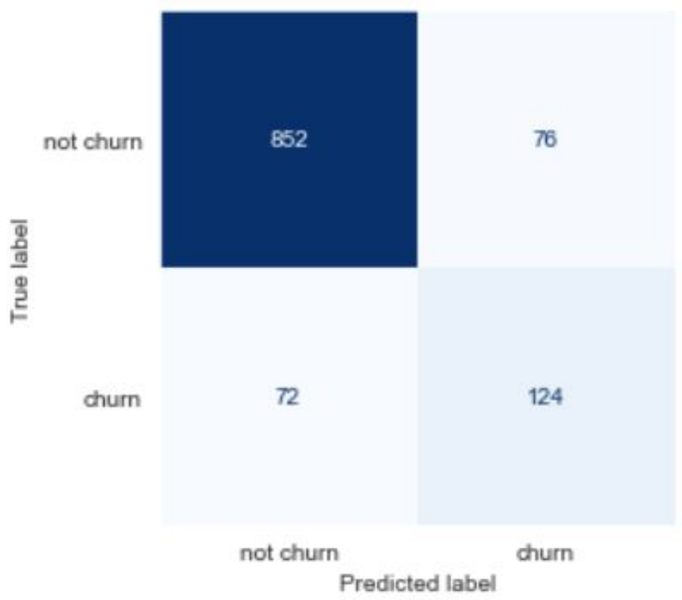
Standard Scalar - for numeric data

⇩

gridsearchCV - use **f1** to select the best hyperparameter

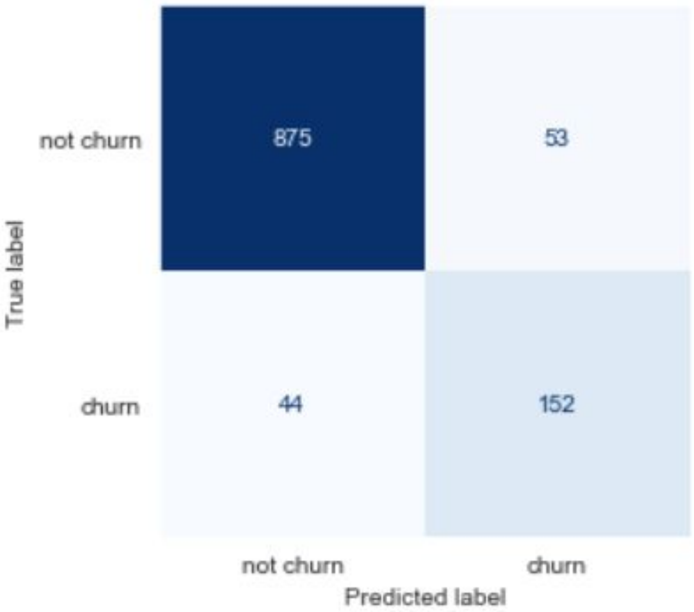

Training : 80 %
Testing : 20%
with Oversampling (1)
Find the best hyperparameter of each algorithm
without Oversampling (2)
Find the best hyperparameter of each algorithm
Model Performance
Compare and select the model with higher f1 score
Model with the best hyperparameter
- LogisticRegression (using **Oversampling**)
- DecisionTree
- RandomForest

Logistic Regression



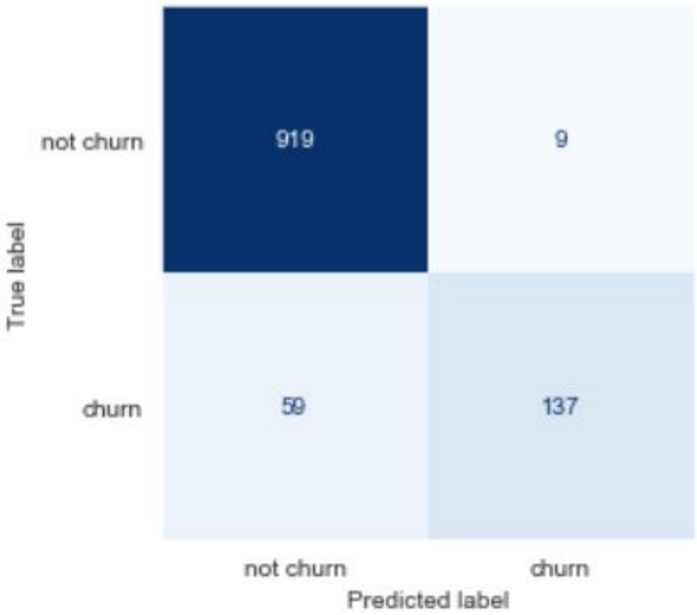
	precision	recall	f1-score	support
0	0.92	0.92	0.92	928
1	0.62	0.63	0.63	196
accuracy			0.87	1124
macro avg	0.77	0.78	0.77	1124
weighted avg	0.87	0.87	0.87	1124

Decision Tree



	precision	recall	f1-score	support
0	0.95	0.94	0.95	928
1	0.74	0.78	0.76	196
accuracy			0.91	1124
macro avg	0.85	0.86	0.85	1124
weighted avg	0.92	0.91	0.91	1124

Random Forest



	precision	recall	f1-score	support
0	0.94	0.99	0.96	928
1	0.94	0.70	0.80	196
accuracy			0.94	1124
macro avg	0.94	0.84	0.88	1124
weighted avg	0.94	0.94	0.94	1124

Q&A