Customer Churn Classification

Abdullah Sayyid Ayyash

Content

Background

The Sales Director is planning to implement a more personalized campaign for their customers to increase the company's customer retention rate. They plan to **invest in promotional activities for customers who are deemed valuable**. Furthermore, they intend to **double down on the amount spent on customers who have not been as active as before.**

Need to share insights and a pilot plan to increase the customer retention rate & creating **a model to identify which customers are predicted to have a high value and which customers are likely to churn**

Objective

- Customer Retention & Churn analysis from historical transaction
- Model to Classify Customers that will churn or not and strategy recommendation for Churn Customers to increase the retention

Summary

Monthly Summary

Number of Customer align with number of Order on every month, when the number of Customer increase, the number of order also increase in certain month

Number of Customer & Order Always increase Significantly on Q4 in every year (i.e. growth_n_customer in Q4 2010 = 30% & growth_n_order Q4 2010 = 38%), Possibly because of Holiday Season (New Year & Christmas) so People go shopping more often

Retention Monthly Cohort

Most of customer after first transaction, not purchased again in the next month, maximum only 36.4% customer that will re-purchase again in the next month (for customer that the first transaction on 2010-01-01)

Customers with first transaction on 2010-01-01 tend to increase from 2010-02-01 to 2010-10-01 (36.4% until 46.4%), decrease after that and increase again significantly on 2011-11-01 with value almost 40%

Customers with first transaction on 2010-12-01 have bad cohort monthly tend to always below 10% (except on 2012-11-01 with value 19%)

Days Distribution

Days between 2 consecutive orders per customer have left skewed distribution, median = 25 days & mean = 51.7 days, above 100 days have lower frequency, based on this distribution we can assume that customer that repurchased > 50 days can be classified as churn customer

Churn Customer

Churn Customer are dominating the population of customer with percentage 47.8% (2809). Non Churn Customer have 14.3% from all customer (1446 customers)

Churn Customer have the highest number in every month compared to non churn customer until end of the 2010, from the beginning of 2011 until 2011-09-01, non churn customers are dominating but end of the year 2011 churn customer dominating again.

Churn Customers always in have the highest number on the end of the year 2010 & 2011 up to 800-ish customers, possibly they are only purchased for holiday season needs

Number of customers in UK is significantly higher than other countries around 5000 ish customers. Number of churn customers in UK are higher, followed by one time order customer and last are non churn customers

Summary

Modelling

Use Model XGBoost for model to classify churn & non churn customer because it have Precision 0.75 and can predict TRUE POSITIVE Churn Customer higher than Logistic Regression & Decision Tree (421)

Recommendation

From all churn customer, 1670 customers should urgently campaigned. 109 of them should be recommend the 85123A Product (HANGING HEART T-LIGHT HOLDER) i.e for customer_id 15002 etc.

Data Context

Table historical transaction from 2009-12-01 until 2011-12-01

Features :

- Order_id
- Product_id
- Product_description
- quantity
- Order_date
- unit_price
- Customer_id
- country
- Revenue = quantity * price

ANALYSIS

Data Quality Check

Quantity & Unit Price have a negative value, because of order cancel (15% order).

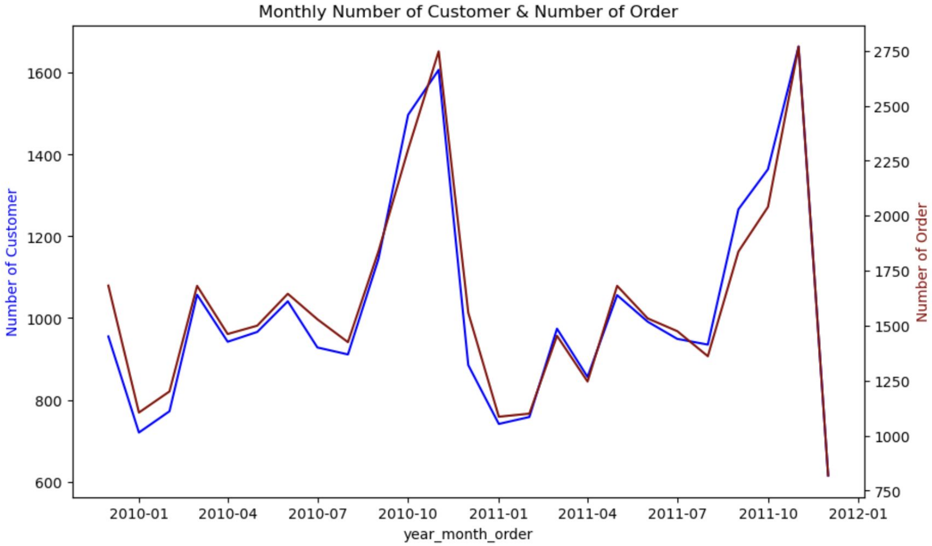
Exclude Order cancel from the analysis because we want to focused on demand analysis that can generating value/revenue.

	order_status	order_id	%
0	CANCEL	8292	15.462072
1	SALE	45336	84.537928

Metrics

1. Number of customer
2. Number of Transaction
3. Customer Retention
    a. Cohort analysis Monthly
4. Days between Orders

Monthly Number of Customer & Number of Transaction per month

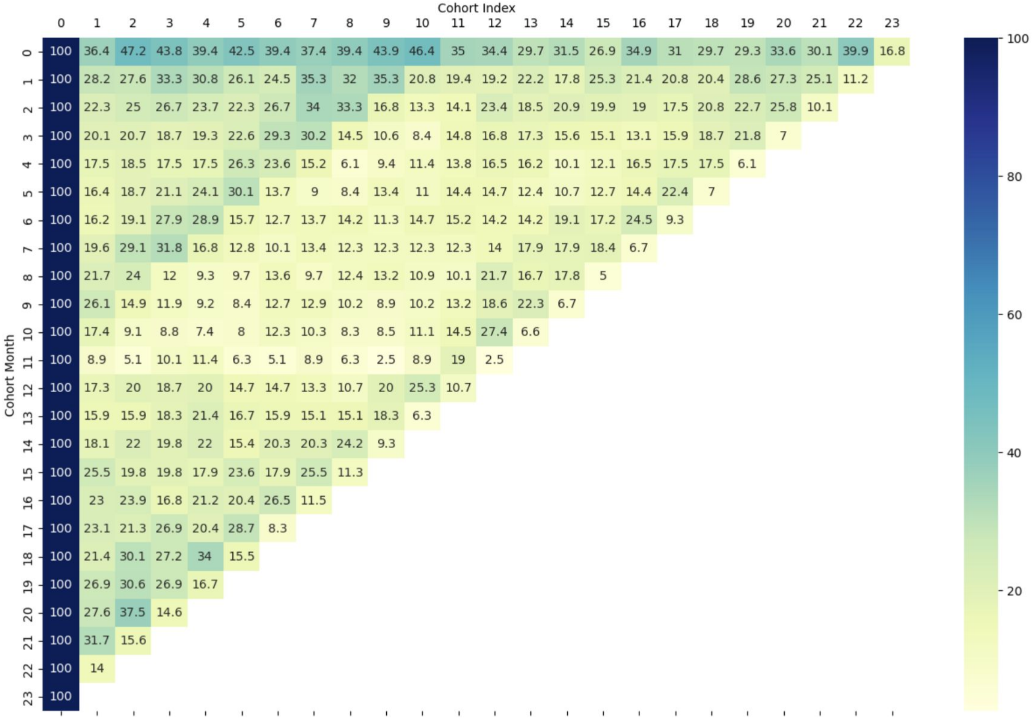


	year_quarter_order	n_customer	n_order	growth_n_customer	growth_n_order
3	2010-07-01	2060	4793	0.006351	0.040373
4	2010-10-01	2670	6607	0.296117	0.378469
7	2011-07-01	2161	4673	0.085384	0.047758
8	2011-10-01	2560	5628	0.184637	0.204366

Number of Customer align with number of Order on every month, when the number of Customer increase, the number of order also increase in certain month

Number of Customer & Order Always increase Significantly on Q4 in every year (i.e. growth_n_customer in Q4 2010 = 30% & growth_n_order Q4 2010 = 38%), Possibly because of Holiday Season (New Year & Christmas) so People go shopping more often

Customer Retention Cohort



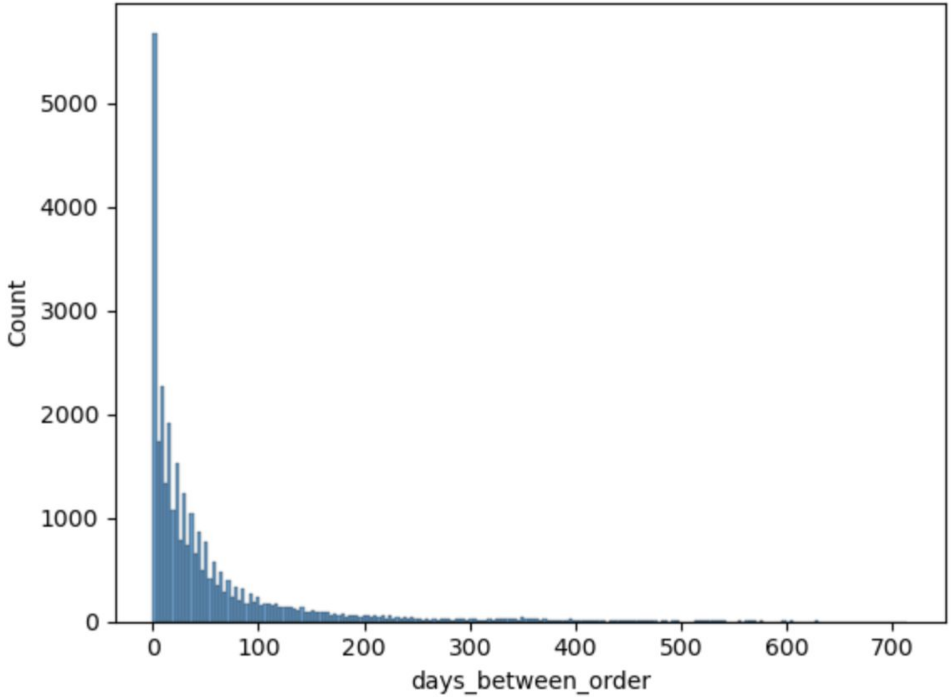
Most of customer after first transaction, not purchased again in the next month, maximum only 36.4% customer that will re-purchase again in the next month (for customer that the first transaction on 2010-01-01)

Customers with first transaction on 2010-01-01 tend to increase from 2010-02-01 to 2010-10-01 (36.4% until 46.4%), decrease after that and increase again significantly on 2011-11-01 with value almost 40%

Customers with first transaction on 2010-12-01 have bad cohort monthly tend to always below 10% (except on 2012-11-01 with value 19%)

Days Between 2 Consecutive Order per Customer Distribution

```
MIN : 0.0
MAX : 714.0
MEDIAN : 25.0
MEAN : 51.6872406805828
```



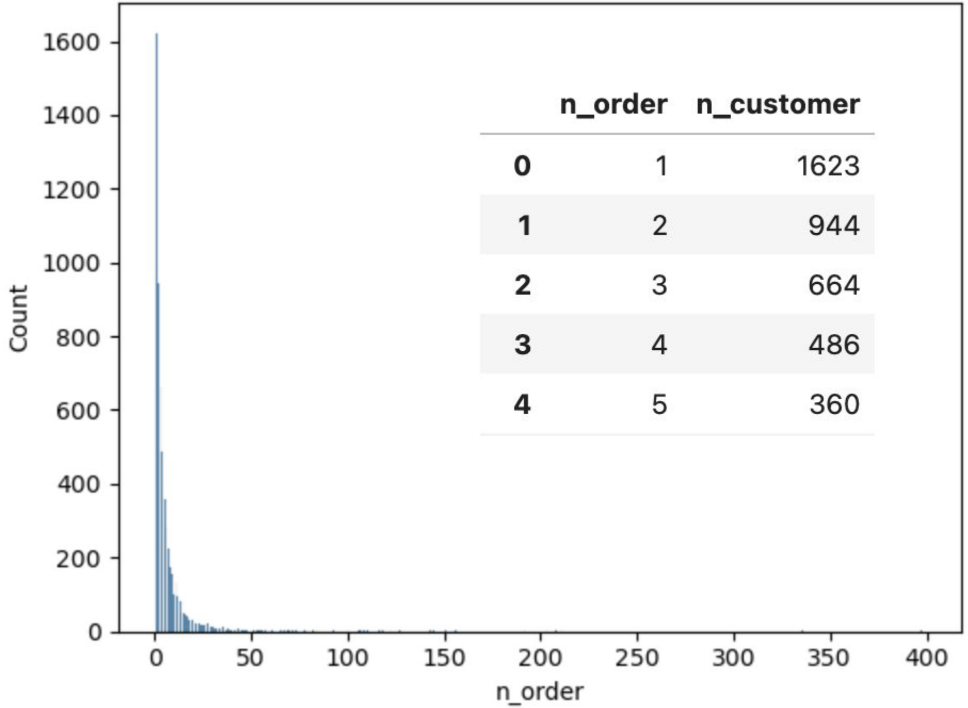
Days between 2 consecutive orders per customer have left skewed, median = 25 days & mean = 51.7 days

above 100 days have lower frequency

based on this distribution we can assume that customer that repurchased > 50 days can be classified as churn customer

Number of Customer per Count Order

```
MIN : 1
MAX : 398
MEDIAN : 3.0
MEAN : 6.289384144266758
```



	n_order	n_customer
0	1	1623
1	2	944
2	3	664
3	4	486
4	5	360

number of customer per n order have left skewed distribution

Customer that only have 1,2,3 order of all time are dominated with value 1623, 944 and 664 customers each

customer that only have 1 order would be excluded for models.

Customer Churn Analysis

Exclude the customer that only have 1 Order & the first order order from all customer because there is no delta days

CATEGORIZING CHURN

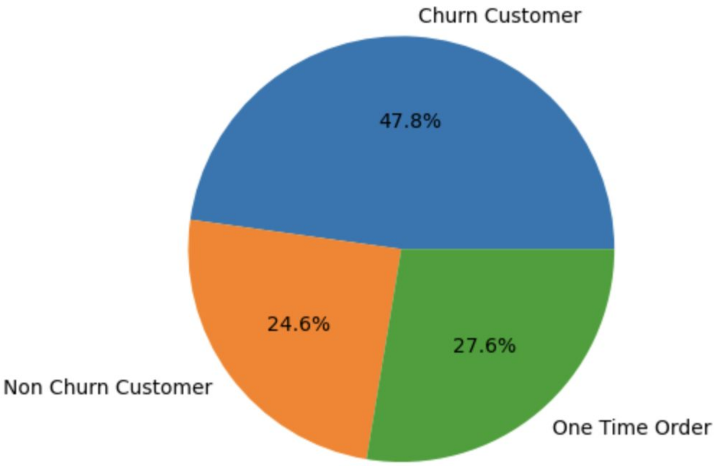
- 0-50 days : **Non Churn Customer**
- '>' 50 days : **Churn Customer**

Categorizing churn would have 2 ways :

1. categorizing directly from days between 2 consecutive order per customer
2. average-ing the days first than categorizing

using the method 2 because method 1 have possibility that churn category would not Mutually Exclusive per customer

Customer Churn Analysis - Number of Customer Contribution

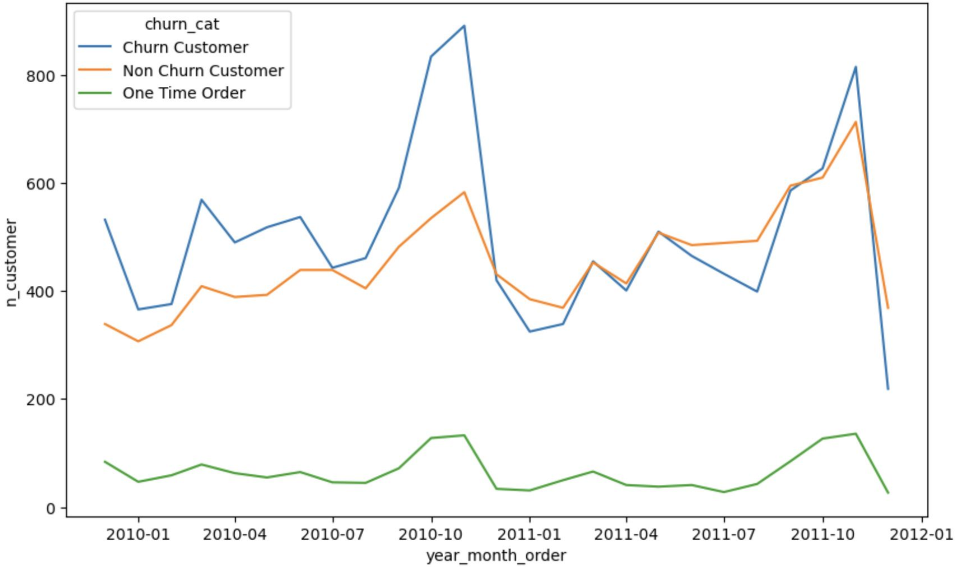


	churn_cat	customer_id
0	Churn Customer	2809
1	Non Churn Customer	1446
2	One Time Order	1623

Churn Customer are dominating the population of customer with percentage 47.8% (2809)

Non Churn Customer have 14.3% from all customer (1446 customers)

Monthly Number of Customer per Churn Category

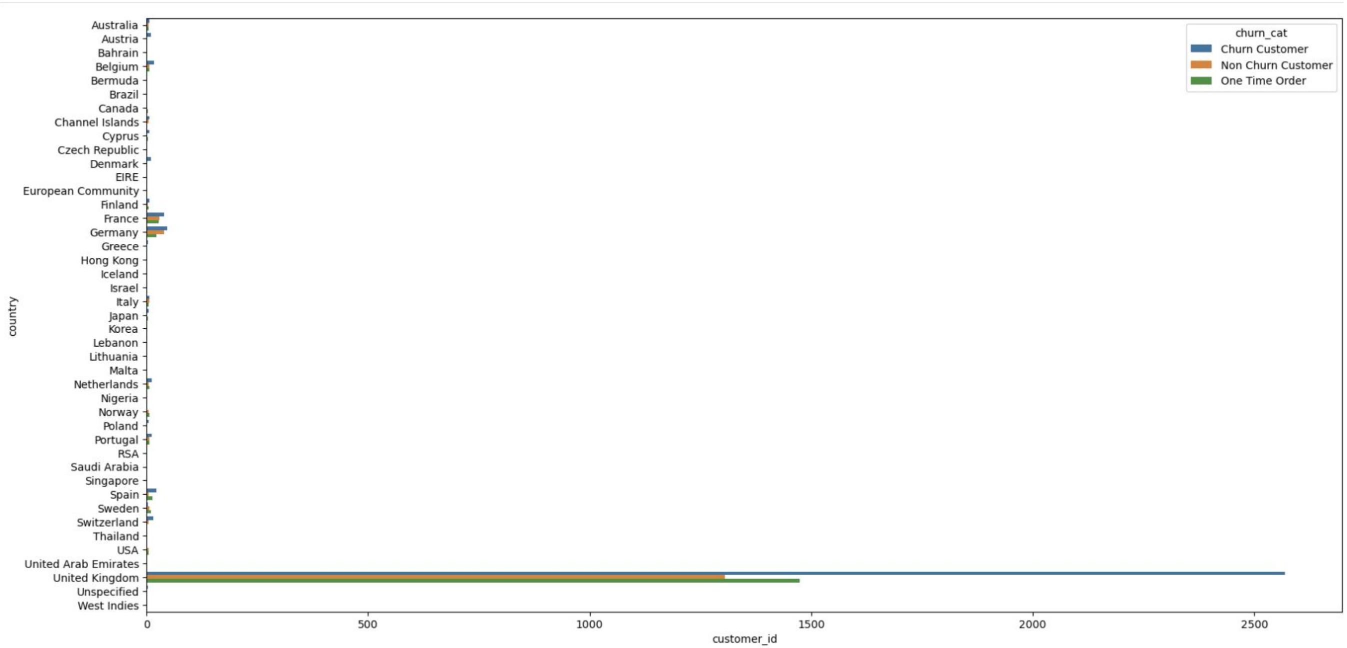


Churn Customer have the highest number in every month compared to non churn customer until end of the 2010, from the beginnig of 2011 until 2011-09-01, non churn customers are dominating but end of the year 2011 churn customer dominating again.

Churn Customers always in have the highest number on the end of the year 2010 & 2011 up to 800-ish customers, possibly they are only purchased for holliday season needs

Customer that only purchased one time of all time have the lowest number every month compared to other category. also have the highest number on every end of the year 2010 & 2011 (not significant)

Churn Customer per Country



Number of customers in UK is significantly higher than other countries around 5000 ish customers.

Number of churn customers in UK are higher, followed by one time order customer and last are non churn customers

Modelling (Classification)

Dataset Preparation for Model

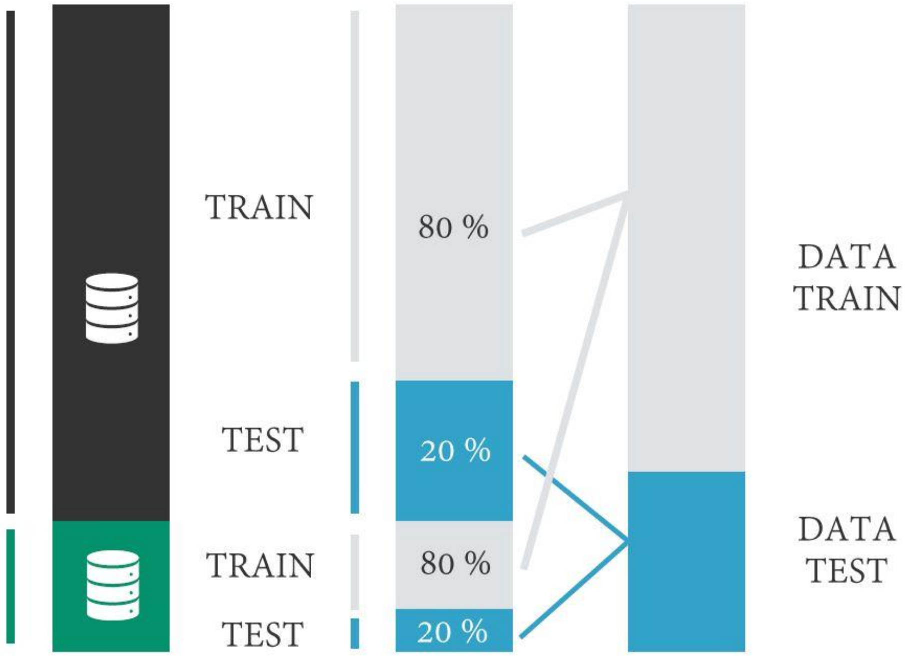
Based on EDA, Here is some Data that can be include for modelling :

1. GRANULARITY : customer_id
2. FILTER : order not cancel, Only customer that have total order > 1 of all time
3. FEATURES :
    a. customer_id
    b. country_big : country per customer that contribute the most quantity
    c. product_big : product per customer that contribute the most quantity
    d. Total_revenue
    e. Total_quantity
    f. Order_frequency
    g. churn_category

Modelling Methods

1. Train & Test data split
2. Numerical Columns handling
3. Categorical Columns handling
4. Machine learning

Modelling Methods - Train & Test data split



TRAIN : 80%

TEST : 20%

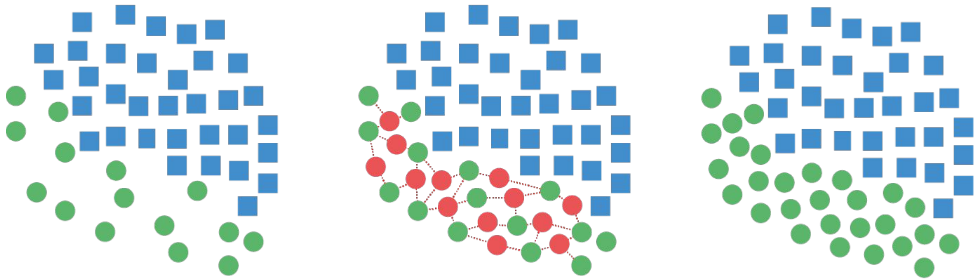
Modelling Methods - Oversampling

1 = Churn customer, 0=non churn customer

```
churn_category
1      66.01
0      33.99
Name: count, dtype: float64
```

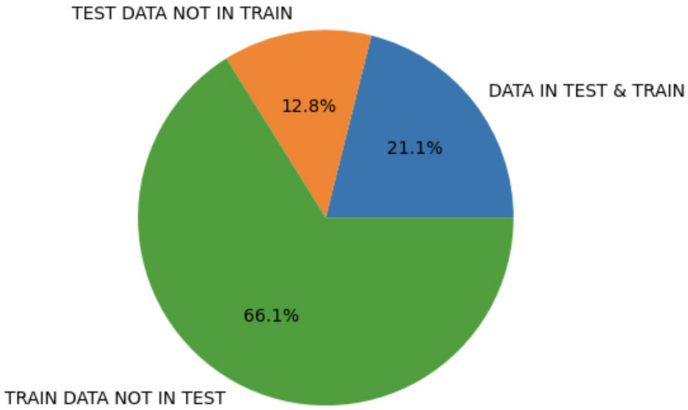
```
churn_category
1      2247
0      2247
Name: count, dtype: int64
```

Synthetic Minority Oversampling Technique



Original Dataset | Generating Samples | Resampled Dataset

Modelling Methods - Train & Test data split - Inclusivity of product & country



	category	n
0	DATA IN TEST & TRAIN	275
1	TEST DATA NOT IN TRAIN	167
2	TRAIN DATA NOT IN TEST	862

there are 12.8% (167) of product-country that in Test Data but not in Train Data. We should **exclude the data from test data** so that data not disturbs the process of evaluation model

Modelling Methods - Numerical columns handling

Numerical columns for model training are **'total_quantity', 'total_revenue', 'order_freq'**

Use Robust Scaling, because

- most of the numerical columns data are Skewed
- Robust scaler are median based data scaling

Robust Standardised Value

Original Value

Sample Median

$$x' = \frac{x - \text{median}(x)}{(Q3-Q1)}$$

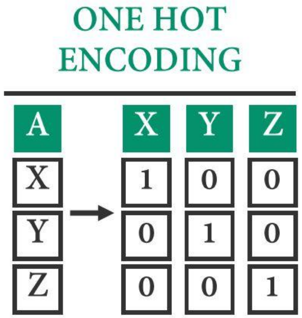
Interquartile Range = Q3 – Q1

Modelling Methods - Categorical columns handling

Categorical columns for model training are **'country_big', 'product_id_big'**

For categorical columns handling we can use many encoder methods, one of them are **One Hot Encoding** (pic)

But it depends on which machine learning we will use, because some of machine learning have parameters that can handling categorical columns directly (i.e. **XGBoost)**



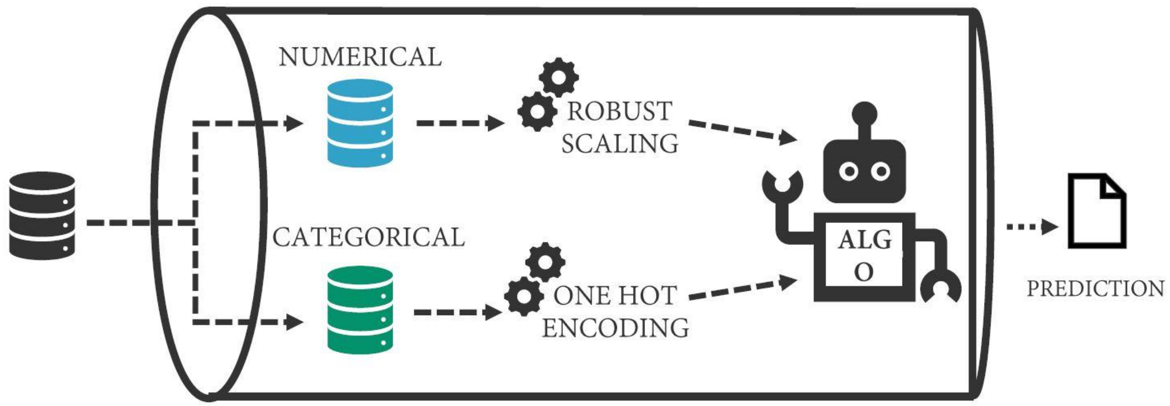
Evaluation Model

1. Metric to evaluate Model
2. Result Modeling - Evaluation Metric (decide which model we will use)
3. + & - of Model chosen
4. Feature Importance
5. Prediction Result

Modelling Methods - Machine Learning

For Machine Learning we use Logistic Regression, Decision Tree and XGBoost. Those three models are commonly use for classification problem

Also Use Pipelining Method to summarize preprocessing data & machine learning (pic below for example)



Evaluation Model - Metric to evaluate Model

PRECISION & TRUE POSITIVE

Because we want double down the cost to make churn customer more retain, so that we want the model that can predict Churn Customer more Precise there for the cost are more efficient to get the Churn Customers.

$$Precision = \frac{TP}{TP + FP}$$

Evaluation Model - Result Modeling

Classification report data TEST XGB Base

	precision	recall	f1-score	support
0	0.51	0.52	0.51	289
1	0.75	0.75	0.75	562
accuracy			0.67	851
macro avg	0.63	0.63	0.63	851
weighted avg	0.67	0.67	0.67	851

Confusion matrix data test XGB Base

	Pred1	Pred0
Akt1	421	141
Akt0	140	149

Classification report data TEST LR Base

	precision	recall	f1-score	support
0	0.53	0.50	0.51	228
1	0.75	0.77	0.76	444
accuracy			0.68	672
macro avg	0.64	0.64	0.64	672
weighted avg	0.67	0.68	0.68	672

Confusion matrix data test LR Base

	Pred1	Pred0
Akt1	342	102
Akt0	114	114

Classification report data TEST DT Base

	precision	recall	f1-score	support
0	0.50	0.54	0.52	228
1	0.75	0.72	0.74	444
accuracy			0.66	672
macro avg	0.63	0.63	0.63	672
weighted avg	0.67	0.66	0.66	672

Confusion matrix data test DT Base

	Pred1	Pred0
Akt1	320	124
Akt0	104	124

- XGBoost, Logistic Regression & Decision Tree relatively have same Precision = 0.75, but the True Positive of XGBoost significantly higher (421 customers)
- XGBoost Model is better than other model to predict Churn Customer
- Also XGBoost can predict customers that have product & country that not in train data, not like Logistic Regression & DT that should exclude those data from data test
- **Chose XGBoost**

Evaluation Model - [+ & - of Model chosen]

+

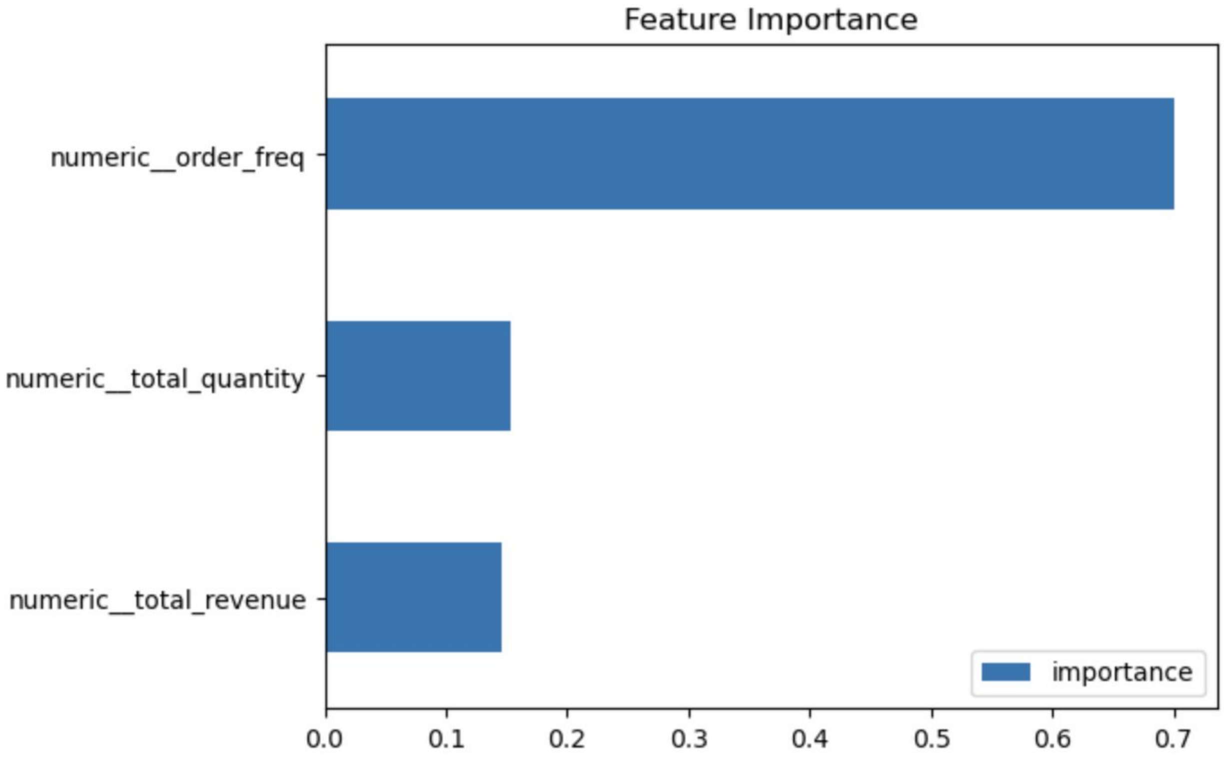
Can predict more Churn Customer so that the cost would be more efficient

Can predict all type of customers that have different Product & Country

-

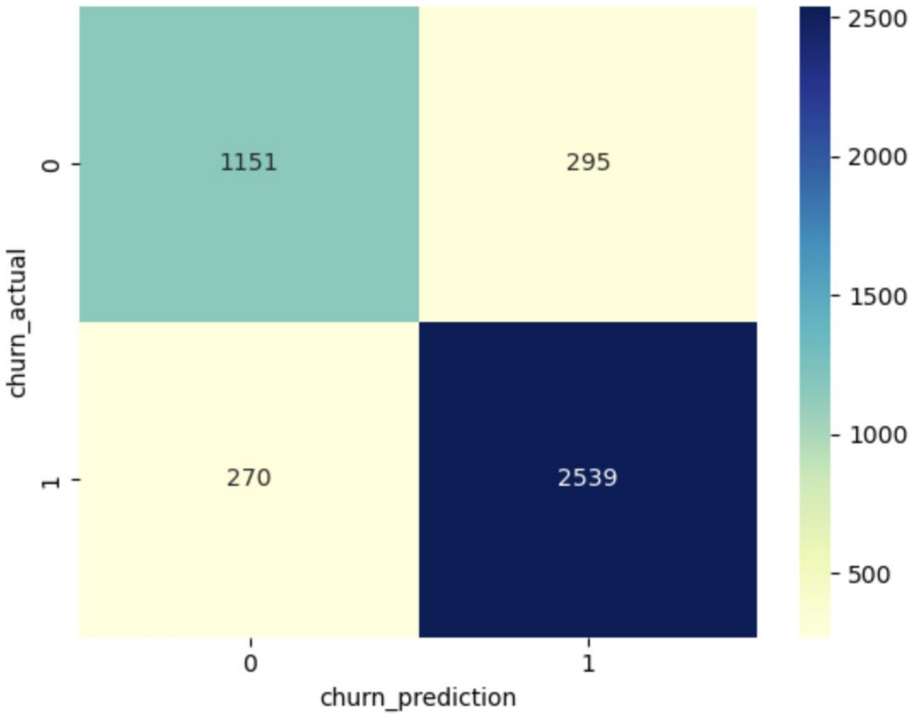
Have higher False Positive (Predict Churn, but actual not churn). **But** not really different from other models

Evaluation Model - Feature Importance



Evaluation Model - Prediction Result

Predict with all data



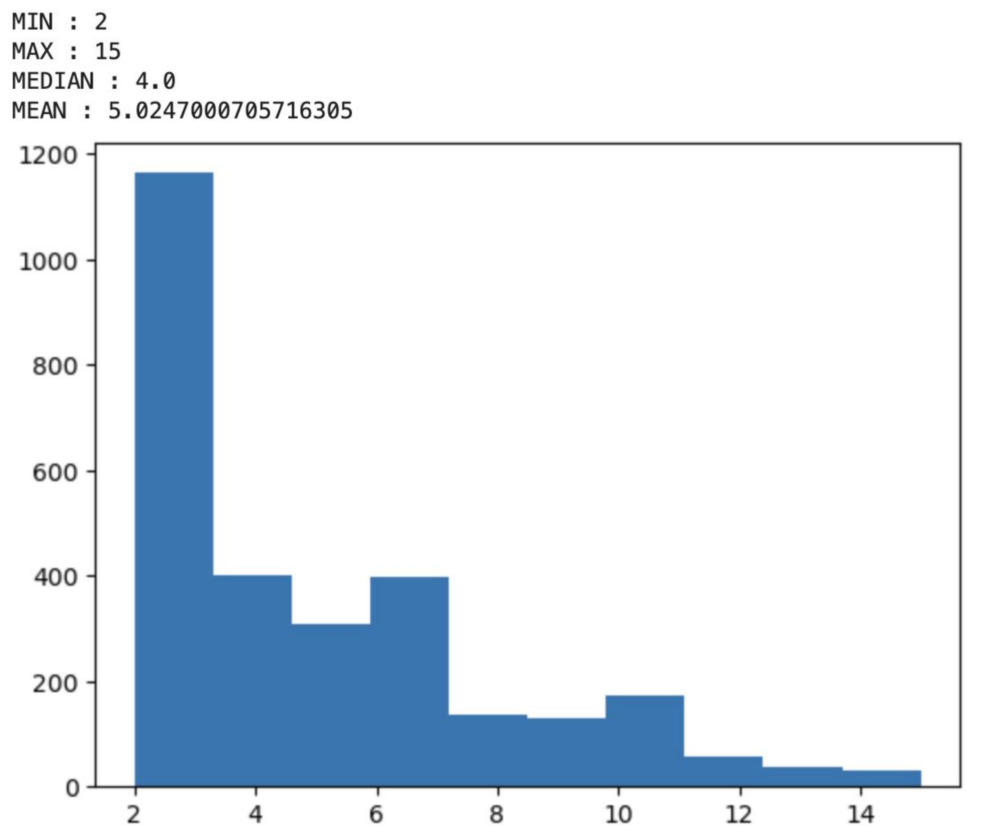
Strategy Recommendation

Based on prediction, order_freq have the highest importance with median 4 orders per customer (for churn customer) (pict)

We can set up urgency of campaign :

- >= 4 orders : urgent campaign
- < 4 order : not urgent campaign

For the campaign contents we can recommend the product that have **highest frequency order of all time**



Strategy Recommendation

campaign_urgency	n_customer
urgent	1670
not urgent	1164

	campaign_urgency	product_id_high_freq	n_customer	customer_id
0	urgent	85123A	109	15002.0, 17496.0, 16412.0, 17358.0, 13497.0, 1...
1	urgent	POST	84	12589.0, 12700.0, 12787.0, 12592.0, 12457.0, 1...
2	urgent	22423	46	12749.0, 14428.0, 14463.0, 15865.0, 17886.0, 1...
3	urgent	84879	44	13527.0, 13202.0, 15265.0, 13124.0, 12867.0, 1...
4	urgent	85099B	39	16086.0, 18257.0, 15394.0, 17031.0, 14127.0, 1...
5	not urgent	85123A	37	17344.0, 15469.0, 16406.0, 15385.0, 13733.0, 1...
6	not urgent	POST	35	12403.0, 12866.0, 12874.0, 12784.0, 12741.0, 1...
7	urgent	21034	31	17244.0, 14704.0, 18096.0, 14204.0, 17625.0, 1...
8	urgent	21232	19	16014.0, 12668.0, 15224.0, 16320.0, 14656.0, 1...
9	urgent	22086	18	14820.0, 12388.0, 15473.0, 16169.0, 16500.0, 1...

From all churn customer, 1670 customers should urgently campaigned. **109 of them should be recommend the 85123A Product (HANGING HEART T-LIGHT HOLDER)** i.e for customer_id 15002 etc.

Other than that, Picture in left are customers that need to urgently campaigned with their product recommendation

Conclusion & recommendation

EDA

1. Number of Customer align with number of Order on every month, when the number of Customer increase, the number of order also increase in certain month
2. Number of Customer & Order Always increase Significantly on Q4 in every year (i.e. growth_n_customer in Q4 2010 = 30% & growth_n_order Q4 2010 = 38%), Possibly because of Holiday Season (New Year & Christmas) so People go shopping more often
3. Most of customer after first transaction, not purchased again in the next month, maximum only 36.4% customer that will re-purchase again in the next month (for customer that the first transaction on 2010-01-01)
4. Customers with first transaction on 2010-01-01 tend to increase from 2010-02-01 to 2010-10-01 (36.4% until 46.4%), decrease after that and increase again significantly on 2011-11-01 with value almost 40%
5. Customers with first transaction on 2010-12-01 have bad cohort monthly tend to always below 10% (except on 2012-11-01 with value 19%)
6. Days between 2 consecutive orders per customer have left skewed distribution, median = 25 days & mean = 51.7 days, above 100 days have lower frequency, based on this distribution we can assume that customer that repurchased > 50 days can be classified as churn customer
7. number of customer per n order have left skewed distribution. Customer that only have 1,2,3 order of all time are dominated with value 1623, 944 and 664 customers each. customer that only have 1 order would be excluded for models
8. Churn Customer are dominating the population of customer with percentage 47.8% (2809). Non Churn Customer have 14.3% from all customer (1446 customers)
9. Churn Customer have the highest number in every month compared to non churn customer until end of the 2010, from the beginning of 2011 until 2011-09-01, non churn customers are dominating but end of the year 2011 churn customer dominating again.
10. Churn Customers always in have the highest number on the end of the year 2010 & 2011 up to 800-ish customers, possibly they are only purchased for holiday season needs
11. Number of customers in UK is significantly higher than other countries around 5000 ish customers. Number of churn customers in UK are higher, followed by one time order customer and last are non churn customers

Conclusion & recommendation

Modelling

Use Model XGBoost for model because have Precision 0.75 and can predict TRUE POSITIVE Churn Customer higher than Logistic Regression & Decision Tree (421)

Recommendation

From all churn customer, 1670 customers should urgently campaigned. **109 of them should be recommend the 85123A Product (HANGING HEART T-LIGHT HOLDER) i.e for customer_id 15002 etc.**

Other than that, Picture below are customers that need to urgently campaigned with their product recommendation :

	campaign_urgency	product_id_high_freq	n_customer	customer_id
0	urgent	85123A	109	15002.0, 17496.0, 16412.0, 17358.0, 13497.0, 1...
1	urgent	POST	84	12589.0, 12700.0, 12787.0, 12592.0, 12457.0, 1...
2	urgent	22423	46	12749.0, 14428.0, 14463.0, 15865.0, 17886.0, 1...
3	urgent	84879	44	13527.0, 13202.0, 15265.0, 13124.0, 12867.0, 1...
4	urgent	85099B	39	16086.0, 18257.0, 15394.0, 17031.0, 14127.0, 1...
5	not urgent	85123A	37	17344.0, 15469.0, 16406.0, 15385.0, 13733.0, 1...
6	not urgent	POST	35	12403.0, 12866.0, 12874.0, 12784.0, 12741.0, 1...
7	urgent	21034	31	17244.0, 14704.0, 18096.0, 14204.0, 17625.0, 1...
8	urgent	21232	19	16014.0, 12668.0, 15224.0, 16320.0, 14656.0, 1...
9	urgent	22086	18	14820.0, 12388.0, 15473.0, 16169.0, 16500.0, 1...