Churn Analysis

Present by Candidate of DS intern program

Project Overview

- Background: Stripe is a technology company that builds economic infrastructure for the internet. Businesses of every size—from new startups to public companies—use the software to accept payments and manage their businesses online.

- Data: The dataset has future merchant transaction activity, for merchants that start over a 2 year period (2033-2034). If the merchant stops processing with Stripe, then they would no longer appear. We have limited data on these merchants and their transactions, but we are still interested in understanding their payments activity to try to infer the types of merchants using Stripe and Customer retention and analysis is very important to Stripe, as well.

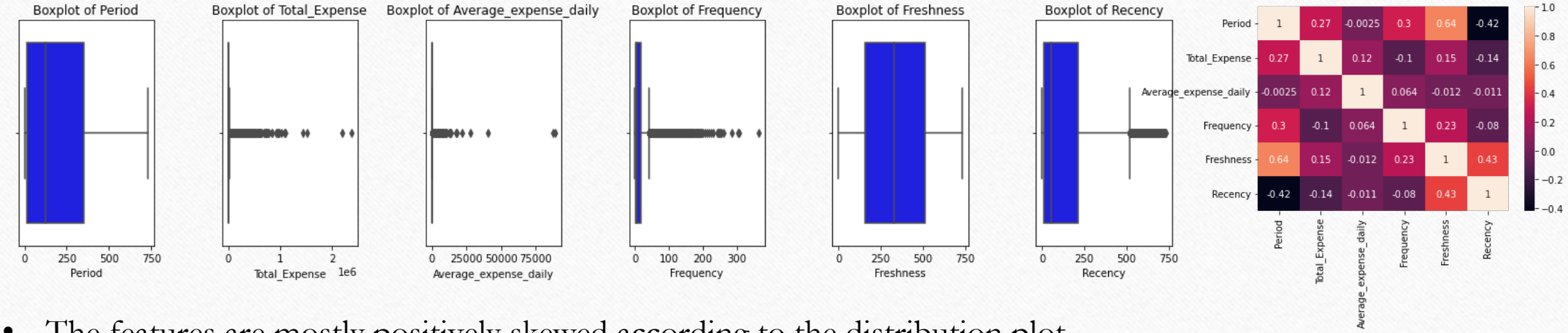
- Separate the task into two parts:
  - Data preprocessing (include: data clean and feature engineering)
  - Churn Analysis (include: Model selection and predict the future, conclusion)

First part: Data Preprocessing

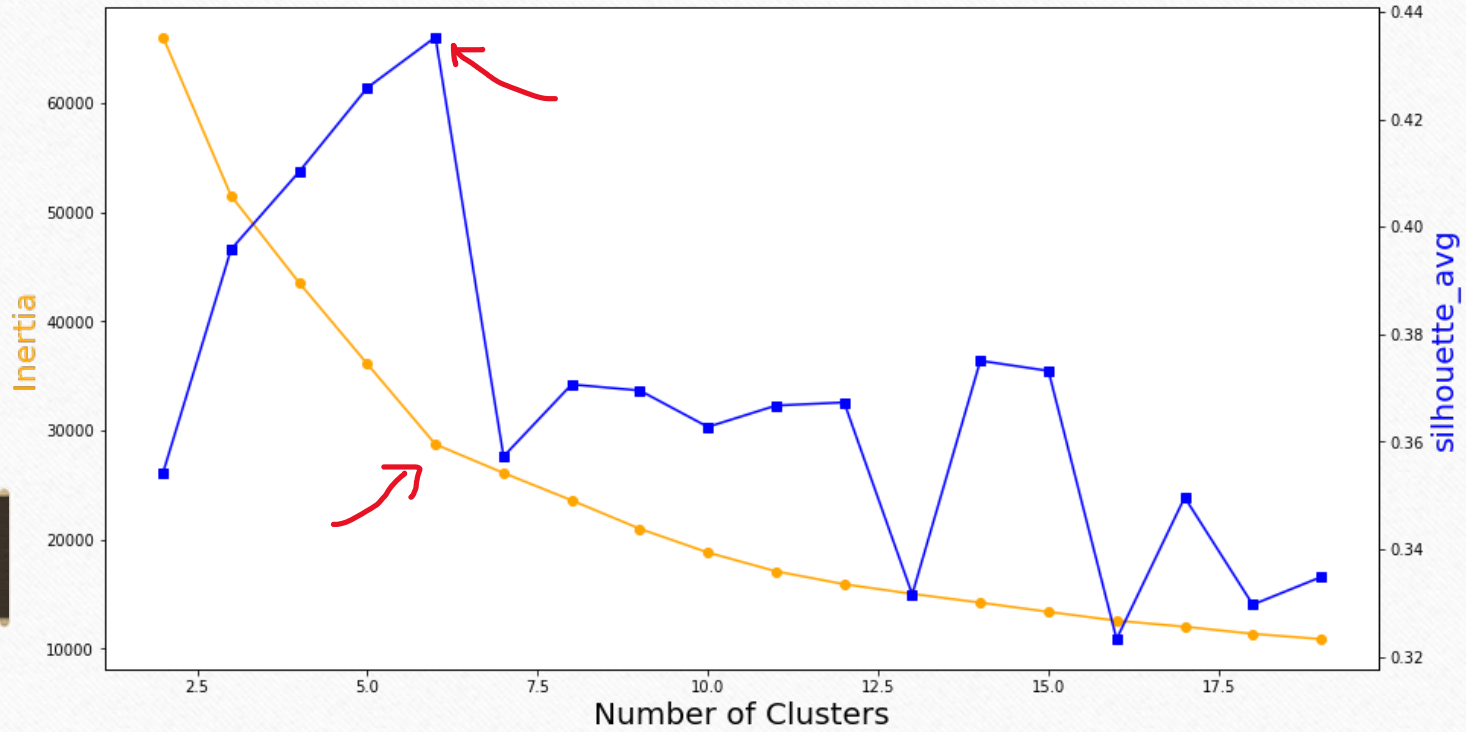
1.1 Build up the feature we need and do some overlook work us boxplot

merchant	Period	Total_Expense	Average_expense_daily	Frequency	Freshness	Recency
0002b63b92	1	33.79	33.790000	2.000	595	595

Frequency: One order every few days on average = (Period)/(total_days)
Freshness: The number of days since the merchant has been using the app = '2035-1-1' - Start_Date
Recency: The number of days since the last purchase time = '2035-1-1' - Last_Date



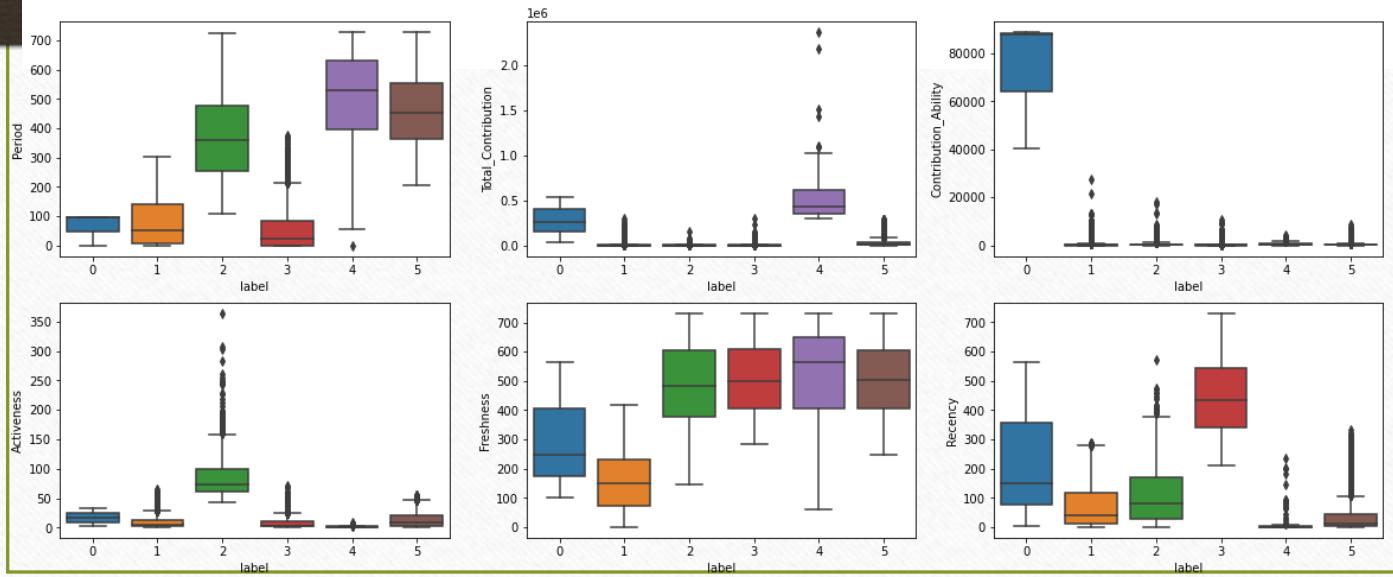
- The features are mostly positively skewed according to the distribution plot
- Most of the features are little correlated with each other according to the heatmap
- So for next step we need to use cluster method to cluster the data



label	
0	3
1	6786
2	957
3	2723
4	116
5	3766

- Use the KMeans unsupervised Learning algorithm to cluster the merchant
- Use "inertia" and "silhouette" to select the best value of parameter "n_clusters": which is 6.
- The smaller inertia is, the better; The larger silhouette is, the better.

label	Period	Total_Expense	Average_expense_daily	Frequency	Freshness	Recency	Characteristics	Active merchant
1	slightly shorter	small	mostly small	high-frequency	new	mostly short	New customers with high-frequency of use services and kind of good expense ability(Average_expense_daily)	Very active
4	mostly long	large	small	very high-frequency	mostly old	short	Old customers who keep using with high-frequency of use services and small expense ability(Average_expense_daily), but large total amount	active
2	long	small	mostly small some large	low-frequency	old	mostly short	Old customers with low-frequency of use services and kind of good expense ability(Average_expense_daily)	relative active
5	long	small	small	slightly low-frequency	old	short	Old customers who keep using with low-frequency of use services and small expense ability(Average_expense_daily)	relative active
3	mostly short	mostly small	small	low-frequency	old	long	Old customers who have not used for a long time with low-frequency of use and low expense ability(Average_expense_daily)	No
0	short	slightly larger	very large	high-frequency	relatively new	relatively long	Special Customers who only use service in a very short period with large expense amount	Special



Use the boxplot we can conclude the each merchant' preference

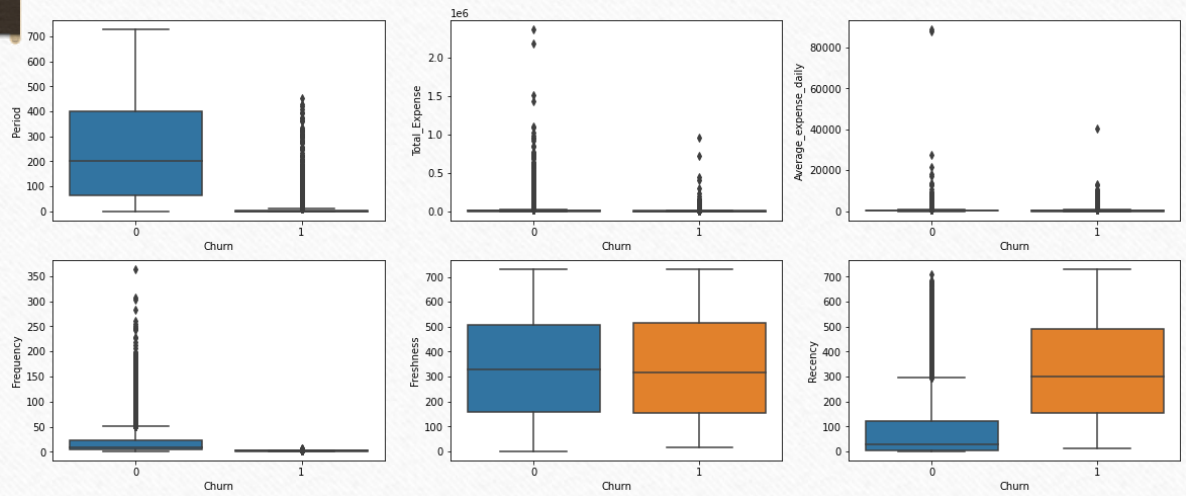
Second part : Churn Analysis

We make a initial guess here: From the conclusion previous two merchants may have not used Stripe for the same days , the merchant with "low-frequency " may be more likely to have been churned compared to the merchant with "high-frequency". The longer using period is, the more likely user rely on Stripe.

$$ChurnRate = \frac{Recency}{Frequency}$$

$$ChurnScore = \frac{1}{2}ChurnRate + \frac{1}{2}\frac{100}{Period}$$

Base on the
Recency,
Frequency,
Period



Use 80% percentile and above the 80% is churned, here is the distribution of the each feature, we can see 'Period', 'Frequency', 'Recency' are good to use as our churn varibales.

Churn	
0	11482
1	2869

1: Churn
0: No Churn

Model selection:

Consider Regression or Classification problem?

I choose binary Classification, because, if we use regression we will get for example person A will churn in 5 months and person B will churn in 3 months. It will have the same business impact, we class the A and B as churn

Random Forest Model

	precision	recall	f1-score	support
0	0.99528	0.98081	0.98799	3439
1	0.98115	0.99536	0.98820	3451
accuracy			0.98810	6890
macro avg	0.98821	0.98809	0.98810	6890
weighted avg	0.98820	0.98810	0.98810	6890

XGboost Model

	precision	recall	f1-score	support
0	0.99709	0.99767	0.99738	3439
1	0.99768	0.99710	0.99739	3451
accuracy			0.99739	6890
macro avg	0.99739	0.99739	0.99739	6890
weighted avg	0.99739	0.99739	0.99739	6890

If we just want to identtify as much as possible churned merchants and do not care about other factors(i.e. cost), we may need to select the model with the highest Recall

If the cost of saving a merchant is very high (i.e. sending too much marketing emails), we may need to select model with higher Precision.

Use XGboost to build the prediction model

merchant	Period	Total_Expense	Average_expense_daily	Frequency	Freshness	Recency	label	ChurnScore	Churn	Predict
09689dc3f0	69	1358.78	67.939000	4.450	550	482	3	54.881941	0	1
5519880667	17	11474.52	1434.315000	3.125	339	322	3	54.461176	0	1
77586d88d9	8	8457.55	291.639655	1.276	122	115	1	51.312696	0	1
826d94dbc5	20	568.31	71.038750	3.500	356	337	3	50.642857	0	1
9f0508694f	30	4947.91	72.763382	1.441	178	148	1	53.019894	0	1
e7f4855acd	28	920.45	131.492857	5.000	558	530	3	54.785714	0	1
eb6cafa141	388	248161.88	1204.669320	2.883	698	310	5	53.892307	0	1
f020b29644	55	2072.98	94.226364	3.500	423	368	3	53.480519	0	1

- we can see from the prediction, most of the merchants who are predicted to churn belong to "label 3", which is prove the initial I made before
- There are some merchants belong to "label 1" we still need to pay attention to them, or they may churn in the future.

Last but not least, with the result we can combine with the marketing team to do the marketing campaign, everything I do is for helping the company grow up so technical analysis need to base on the business first and content matters!!!

Thanks for watching!