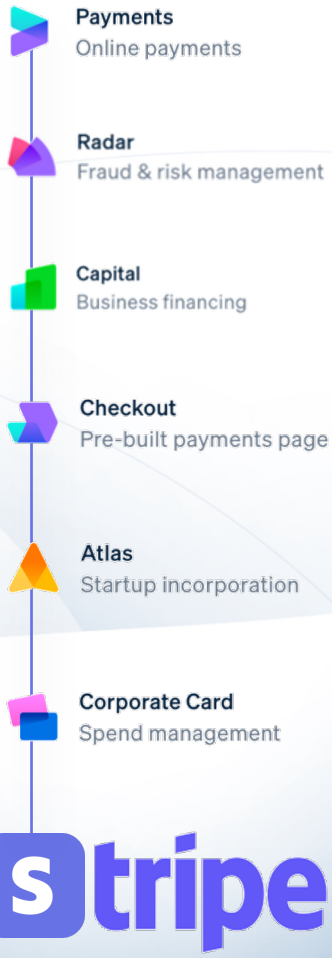
Merchant Segmentation & Churn Analysis

——Yanan Zhou | Takehome DS Written Project



Introductions:

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.
Customer retention and analysis is very important to Stripe, so it is curious about the customer types and churn level within 2033-2034. With dataset including future merchant transaction activity, we try to figure out the features and churn levels of the customers.

Assunptions:

Date: The date for "now" is "2035-1-1" .

Independence: Each merchant is independent from each other, which means one merchant's behavior will not influence whether another merchant will use or not.

No Returned Merchants: The first date appears in the dataset is the date the merchant used Stripe for the first time. Because some of the merchants might be churned for a long time and then come back to use Stripe. However, we do not have previous data, so we regard the earliest date appearing in the dataset as the first using time.

Content

1 Merchant Segmentation

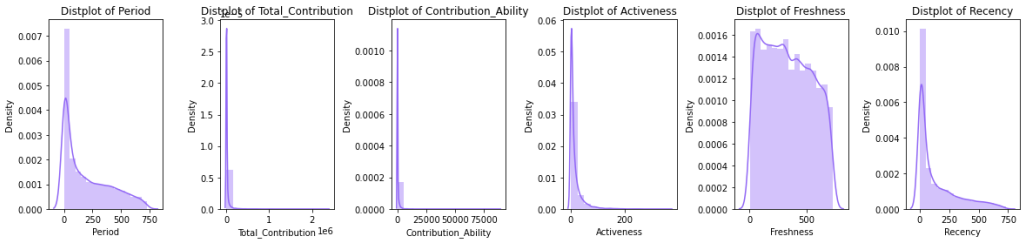
1.1 Generate Features

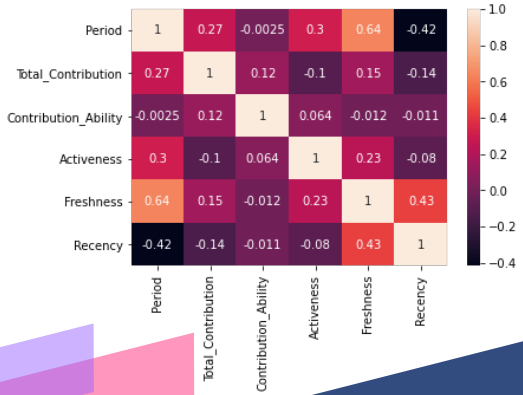
- **Contribution Ability:** Average amount per transaction
- **Total Contribution:** Total amount of all the transaction
- **Period:** The number of days between the first purchase and the last purchase
- **Activeness(Frequency):** One order every few days on average = (Period)/(total transaction times)
- **Freshness:** The number of days since the merchant has been using the app = '2035-1-1' - Start_Using_Date
- **Recency:** The number of days since the last purchase time = '2035-1-1' - Last_Using_Date

1.2 Simple Exploratory Data Analysis

According to the distribution and Correlation map:

- The generated features are mostly **right skewed**
- Some generated features are extremely **centered**
- Most of the generated features are **little correlated** with each other



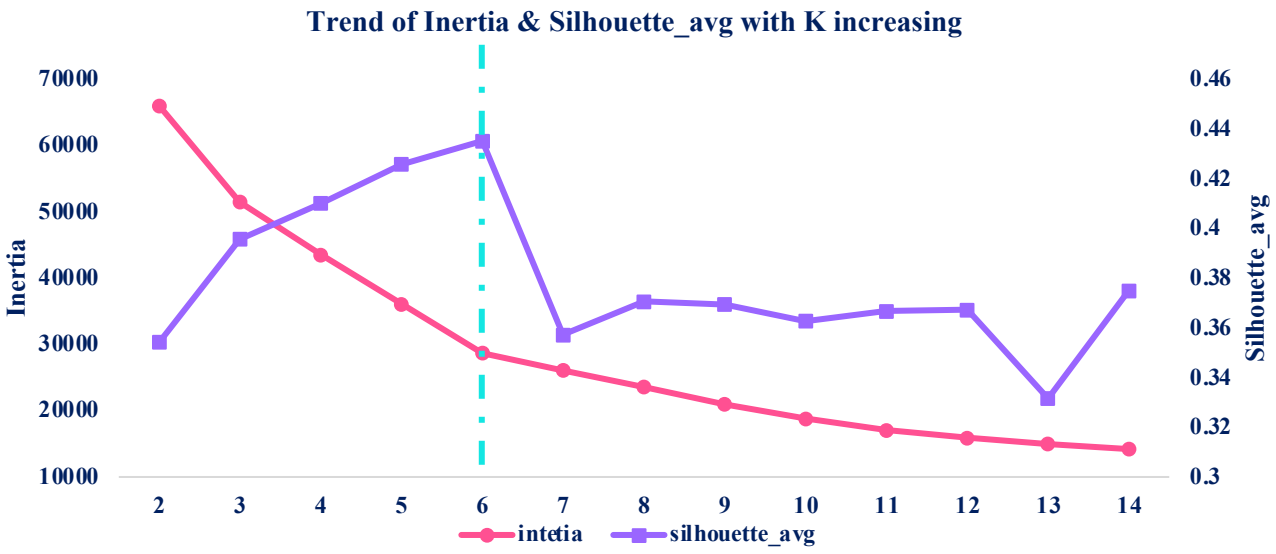


1 Merchant Segmentation

1.3 KMeans Clustering

Using "inertia" and "silhouette" to select the best value of parameter "n_clusters":

- The smaller inertia is, the better;
- The larger silhouette is, the better.



Thus, choose "n_clusters"=6 as the parameter.

Then, count to see how many merchants are in each type.

Type	Number
0	3
1	6786
2	957
3	2723
4	116
5	3766

1 Merchant Segmentation

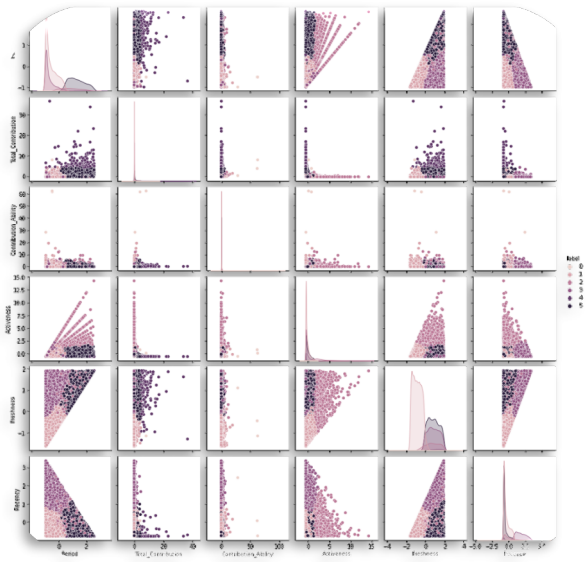
1.4 Characteristics of each type

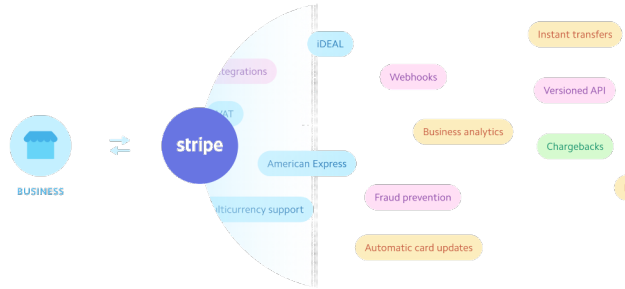
- Active Merchants
  - Very Active/Main Customers
    - **Type 4:** Use very frequently, old & stable, main contributor
    - **Type 1:** Use very frequently, new users, former type of "Type 4"
  - Relatively Less Active
    - **Type 2:** Has potential to be both active or churn, relatively good contributor
    - **Type 5:** Has potential to be both active or churn, less contribution ability
- Non-active Merchants
  - **Type 3:** Has the most possibility to churn
- Special Merchants
  - **Type 0:** Makes extremely large amount of transactions during very short period

Initial Guess and Treatment:

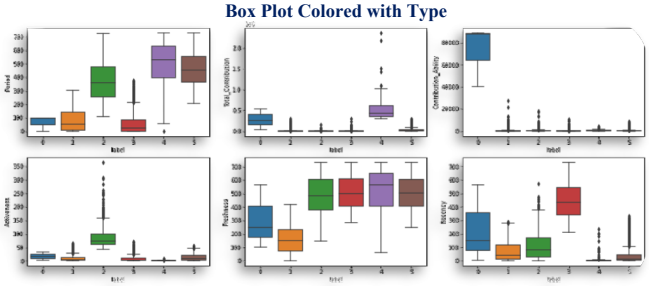
- Type1 → **maintain it and transform to Type4**
- Type2&5 → **save and transform to Type4, especially some of Type2 have potential contribution ability, we should grab them**
- Type 3 → **try our best to save them in case they churn quicly**
- Type 0 → **depend on the business they do (they might launch promotion for their customers at specific time, we may maintain the corporation)**

Pair Scatter Plot Colored with Type





Box Plot Colored with Type



Each feature's characteristics of different types according to the plot:

Type	Period	Total Contribution	Contribution Ability	Activeness	Freshness	Recency	Characteristics
0	short	slightly larger	very large	high-frequency	relatively new	relatively long	Customers who only use in a very short period with extremely large number of transaction
1	slightly shorter	small	mostly small	high-frequency	new	mostly short	New customers with high-frequency of use and relatively good contribution ability
2	long	small	slightly large	low-frequency	old	mostly short	Old customers with low-frequency of use and relatively good contribution ability
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 contribution ability
4	mostly long	large	small	very high-frequency	mostly old	short	Old customers who keep using with high-frequency of use and small contribution ability, but large total amount
5	long	small	small	slightly low-frequency	old	short	Old customers who keep using with low-frequency of use and small contribution ability

2 Churn Analysis

2.1 Definition of Churn

- **Although the "Recency" of some merchants are large, they may have used Stripe less often than others.** In other words, two merchants may have not used Stripe for the same days(i.e. 6 days), the merchant with "Activeness=2" may be more likely to have been churned compared to the merchant with "Activeness=12".
- **The longer using period is, the more likely user rely on Stripe**.

Thus, we can give "Churn" a definition with considering "Recency", "Activeness" and "Period":

$$Relative\ Leaving\ Possibility = \frac{Recency}{Activeness}$$

$$Churn\ Score = \frac{1}{2} \cdot Relative\ Leaving\ Possibility + \frac{1}{2} \cdot \frac{100}{Period}$$

Calculate the Churn Score of each merchant,
merchant with "Churn Score" > 80% quantile are defined as "Churn" (> 55.25)

The Number of each class:

Churn	Number
0	11482
1	2869

2 Churn Analysis

2.2 Merchants Who Have Already Churned

Similar to the initial guess:

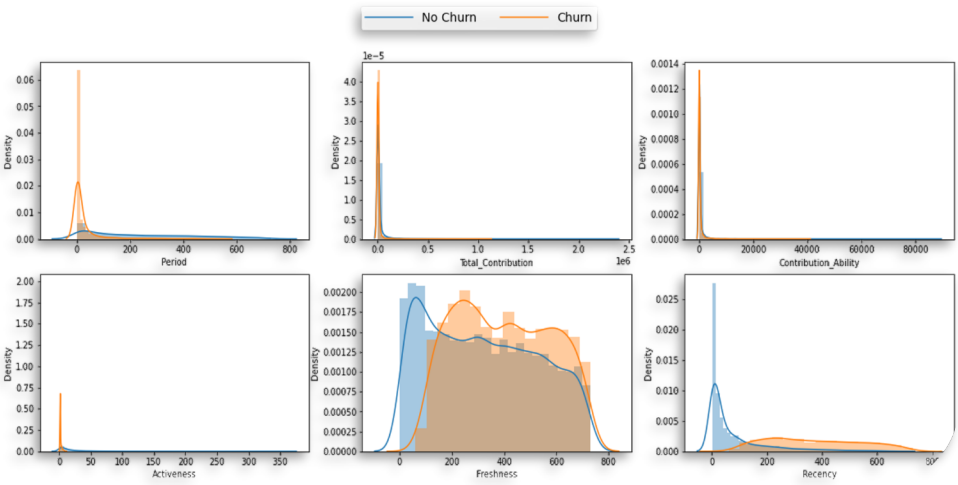
- Type 3 merchants have the highest probability of churning
- Type 0 merchants are unstable and unsure
- Type 1 merchants have the
- Type 1/2/4/5 merchants have lower rate of churning

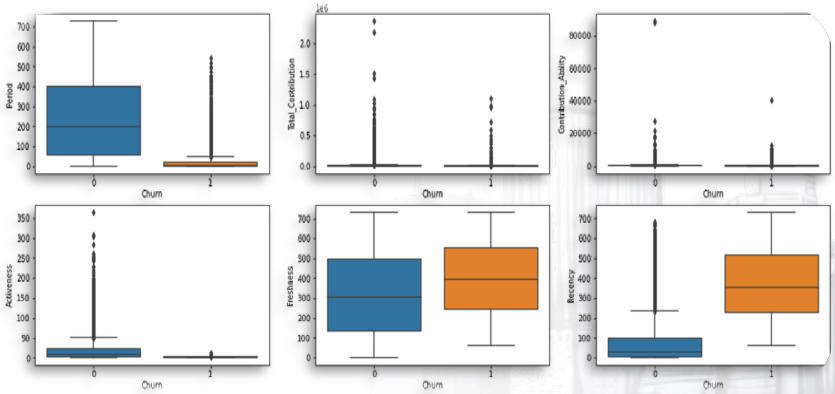
However, Type 2/5 perform better than Type 1/4 , we need more information to figure out the reason

Type	Number	Number of Churn	Ratio
0	3	1	0.333333333
1	6786	1328	0.195697023
2	957	0	0
3	2723	1523	0.559309585
4	116	4	0.034482759
5	3766	13	0.003451938

From the distribution of "Churn" and "Non-Churn" maerchants:

Period, Freshness, Recency seem to be the good predictors.





2 Churn Analysis

2.3 Churn Prediction Model

- Deal with imbalanced class: upsampling the "churn" class
- Scale the data
- Split train and test set (test set=0.3)
- Train and tuning parameters with 3 kind of models

According to the "Accuracy", "Recall", "Precision" above, we choose the XGBoost model. Because these scores of this model perform best among all the models.

Logistics Regression

	precision	recall	f1-score	support
0	0.99157	0.92323	0.95618	3439
1	0.92842	0.99218	0.95924	3451
accuracy			0.95776	6890
macro avg	0.95999	0.9577	0.95771	6890
weighted avg	0.95994	0.95776	0.95771	6890

Random Forest

	precision	recall	f1-score	support
0	0.99499	0.98081	0.98785	3439
1	0.98114	0.99507	0.98806	3451
accuracy			0.98795	6890
macro avg	0.98806	0.98794	0.98795	6890
weighted avg	0.98805	0.98795	0.98795	6890

XGBoost

	precision	recall	f1-score	support
0	0.99796	0.99738	0.99767	3439
1	0.99739	0.99797	0.99768	3451
accuracy			0.99768	6890
macro avg	0.99768	0.99768	0.99768	6890
weighted avg	0.99768	0.99768	0.99768	6890

In real situations, we may need to choose the measurement based on different need, for example:

- 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 or the way we saving a merchant affect other merchants' interest (i.e. sending too much subscribtion emails), we may need to select model with **higher "Precision".**

2 Churn Analysis

2.4 Make Prediction on Active Merchants

- Extract active merchants
- Scale the data / Get dummy variables
- Predict “Churn” by the XGBoost model before

Conclustions:

- **We can see from the prediction, most of the merchants who are predicted to churn belong to "label 3", which is similar to the initial guess we made before.**

- **There are some merchants belong to "label 1", who are regarded as the former type of "label 4". This means that although some merchants seem active, we still need to pay attention to them, or they may churn in the near future.**

Merchant	Period	Total_Contribution	Contribution_Ability	Activeness	Freshness	Recency	label	Pred
09689dc3f0	69	1358.78	67.939	4.45	550	482	3	1
5519880667	17	11474.52	1434.315	3.125	339	322	3	1
77586d88d9	8	8457.55	291.639655	1.276	122	115	1	1
826d94dbc5	20	568.31	71.03875	3.5	356	337	3	1
9f0508694f	30	4947.91	72.763382	1.441	178	148	1	1
bd6be9c660	81	192.08	8.003333	4.375	550	470	3	1
d719835aa8	3	794.87	66.239167	1.25	86	83	1	1
f020b29644	55	2072.98	94.226364	3.5	423	368	3	1
f0d46b1cd9	8	1399.38	87.46125	1.5	141	134	1	1