GURTEG SAWHNEY

STRIPE TAKE HOME PROJECT

Task

Data: The dataset has future merchant transaction activity, for merchants that start over a 2 year period (2033-2034). The data spans from 1/1/33 through 12/31/34. Although the data is made up, you can consider this to be a random sample of future merchants using Stripe. Each observation is a transaction amount in cents. If the merchant stops processing with Stripe, then they would no longer appear.

- takehome_ds_written.csv

Questions:

Part1

- 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. Using only the given data, how would you identify different kinds of businesses in the sample? Please generate assignments for each merchant.

Part2

- Sometimes a merchant may stop processing with Stripe, which we call churn. We are interested in identifying and predicting churn. Please a) come up with a concrete definition for churn b) identify merchants that have already churned in the dataset, and c) build a model to predict which active merchants are most likely to churn in the near future.

.

PART 1

EDA

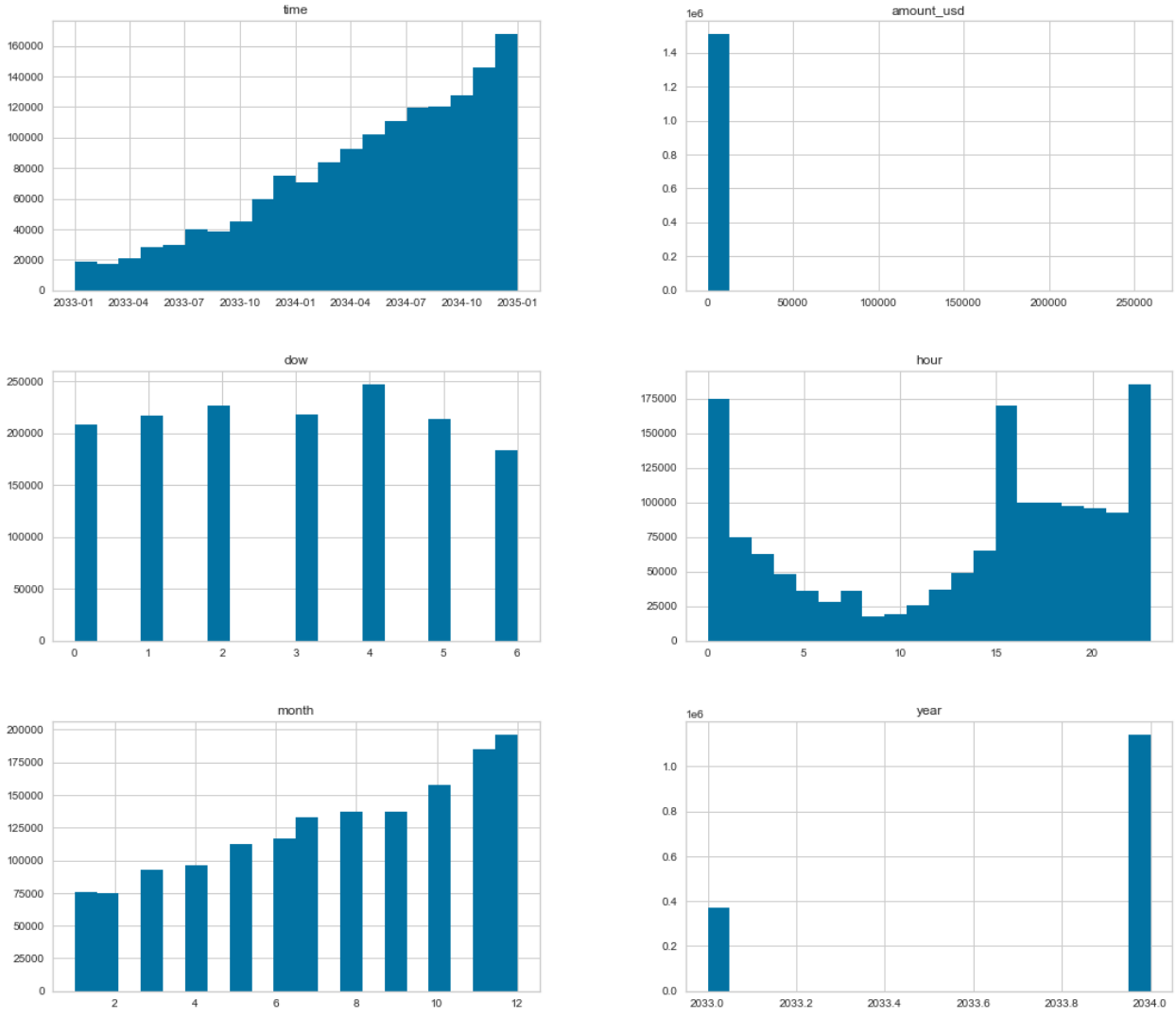
Exploratory Data Analysis of the data is performed.

	merchant	time	amount_usd_in_cents
0	faa029c6b0	2034-06-17 23:34:14	6349
1	ed7a7d91aa	2034-12-27 00:40:38	3854
2	5608f200cf	2034-04-30 01:29:42	789
3	15b1a0d61e	2034-09-16 01:06:23	4452
4	4770051790	2034-07-22 16:21:42	20203

The data has three columns merchant, time and amount_usd_in_cents, proiding ecah merchants transaction information. There are 14,351 different merchants in the dataset. We do not have any null values.

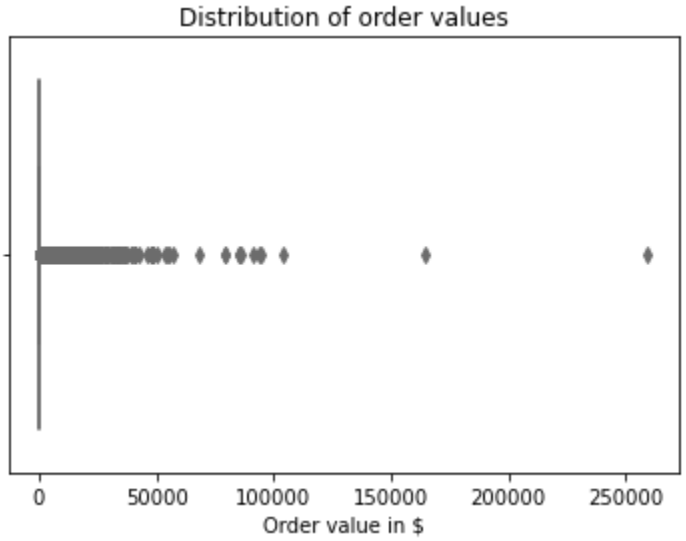
Extracting Date time features and converting cents to $

Cents were converted to $ for better intuition.Hour, Month,Year day of week were extracted from the time column and histogram plots were plotted.

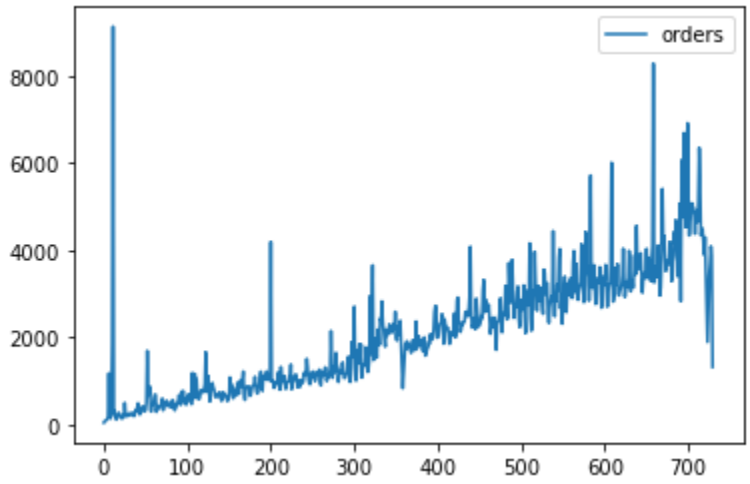


We can see that the transactions are spread across 2033 and 2034 starting from Jan 1 2033 to Dec 31 2034. The number of orders increased around late 2034.

Box plot Distribution of Amount Spent (in $) for each order



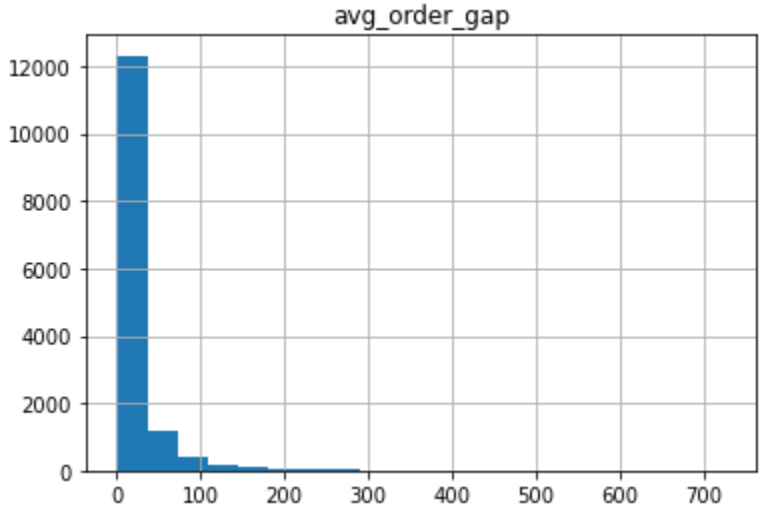
Orders over time



Repeat Merchants

We can see that around **86% of merchants are repeat merchants,** that is they placed an order on more than 1 day during the time period 2033-2034

Average Order Gap



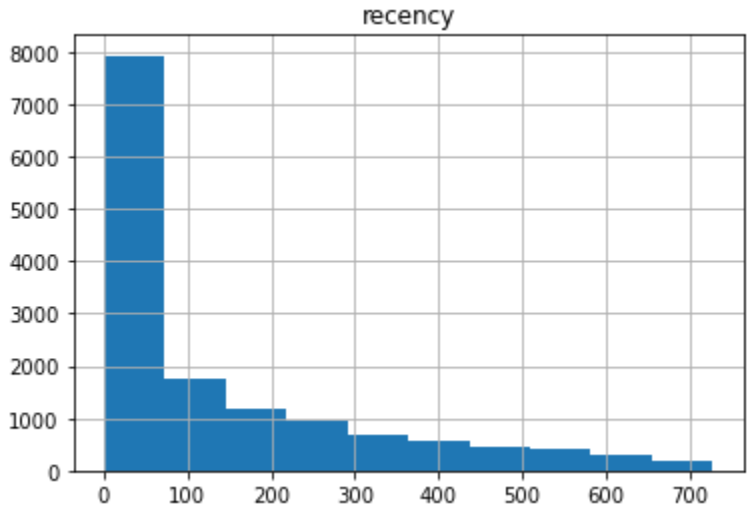
Order Gap = the gap in days between the consecutive orders for a merchant

Average Order Gap = Average of consecutive Order gaps for a merchant

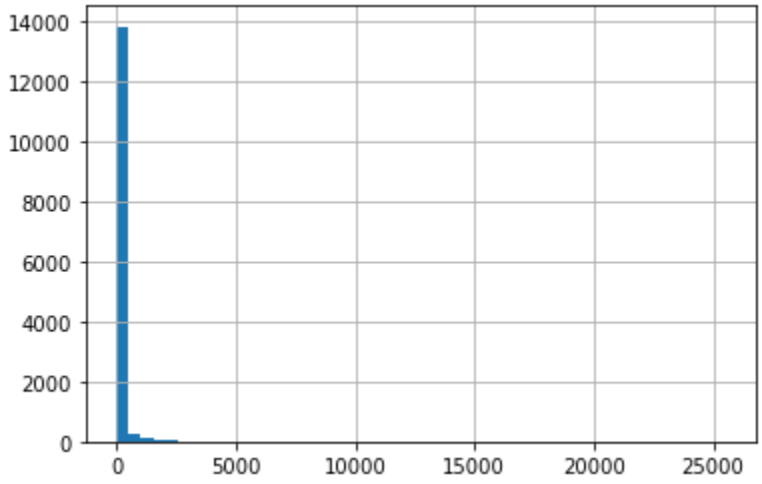
We can see from the histogram above that more than 85% of merchants have an average order gap within 60 days, which means 85% of the merchants are coming back and placing an order within 60 days from their first order on an average.

Recency, Frequency and Monetary Analysis (RFM)

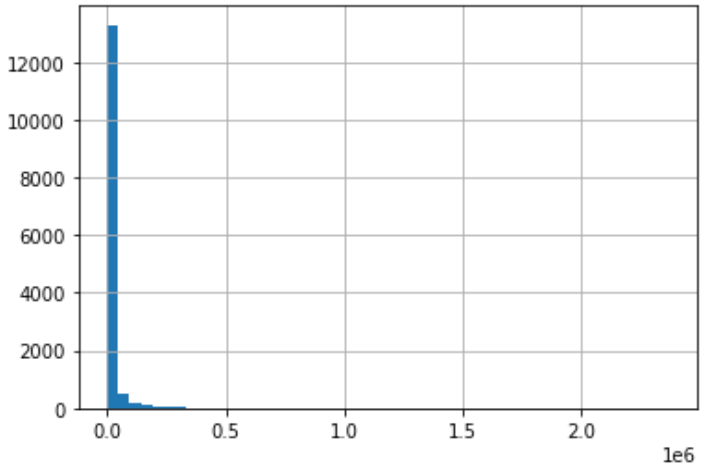
Recency = Days difference between last order date and 31 Dec 2034



Frequency = No of Orders placed by merchants



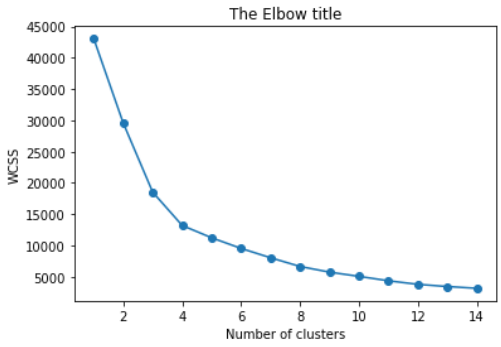
Monetary = TotalSpend distribution for merchants



Clustering - using RFM(Recency, Frequency and Monetary) Analysis and K Means

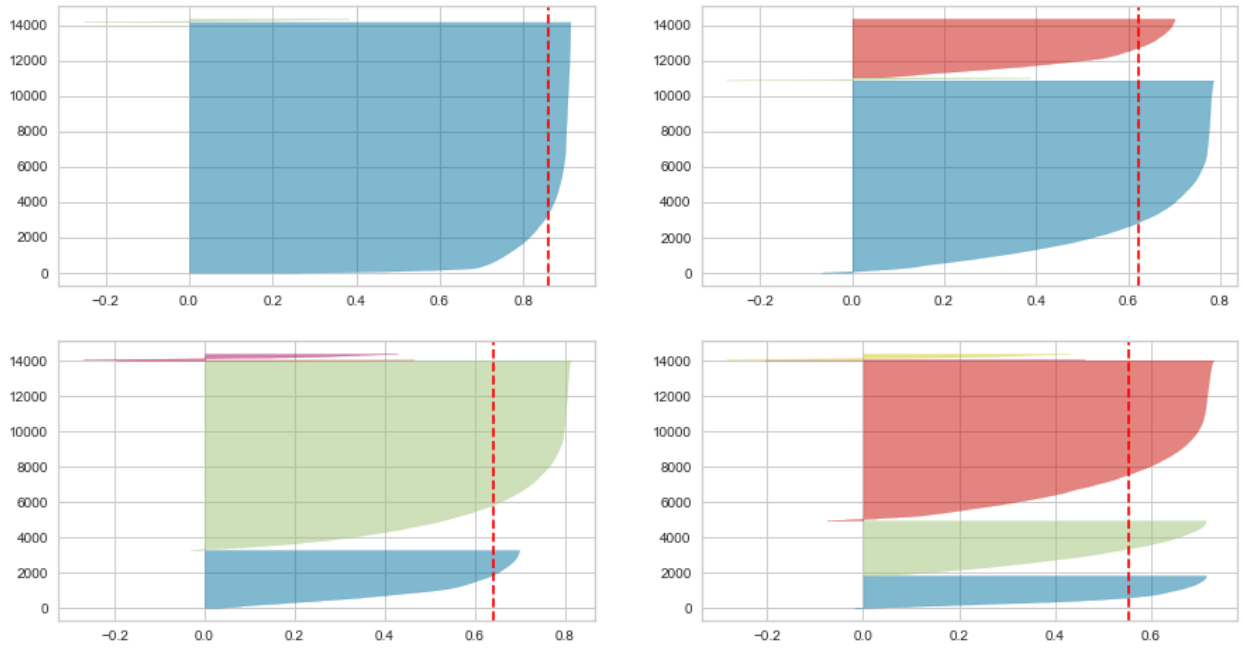
We performed K Means clustering on the dataset using RFM features to understand the customer segments.Based on Elbow Analysis and Silhouette Scores, we chose k=3 as the value.

Elbow Method



After k=3 Within cluster sum of squares does not decrease that much and so the variance explained would not increase much

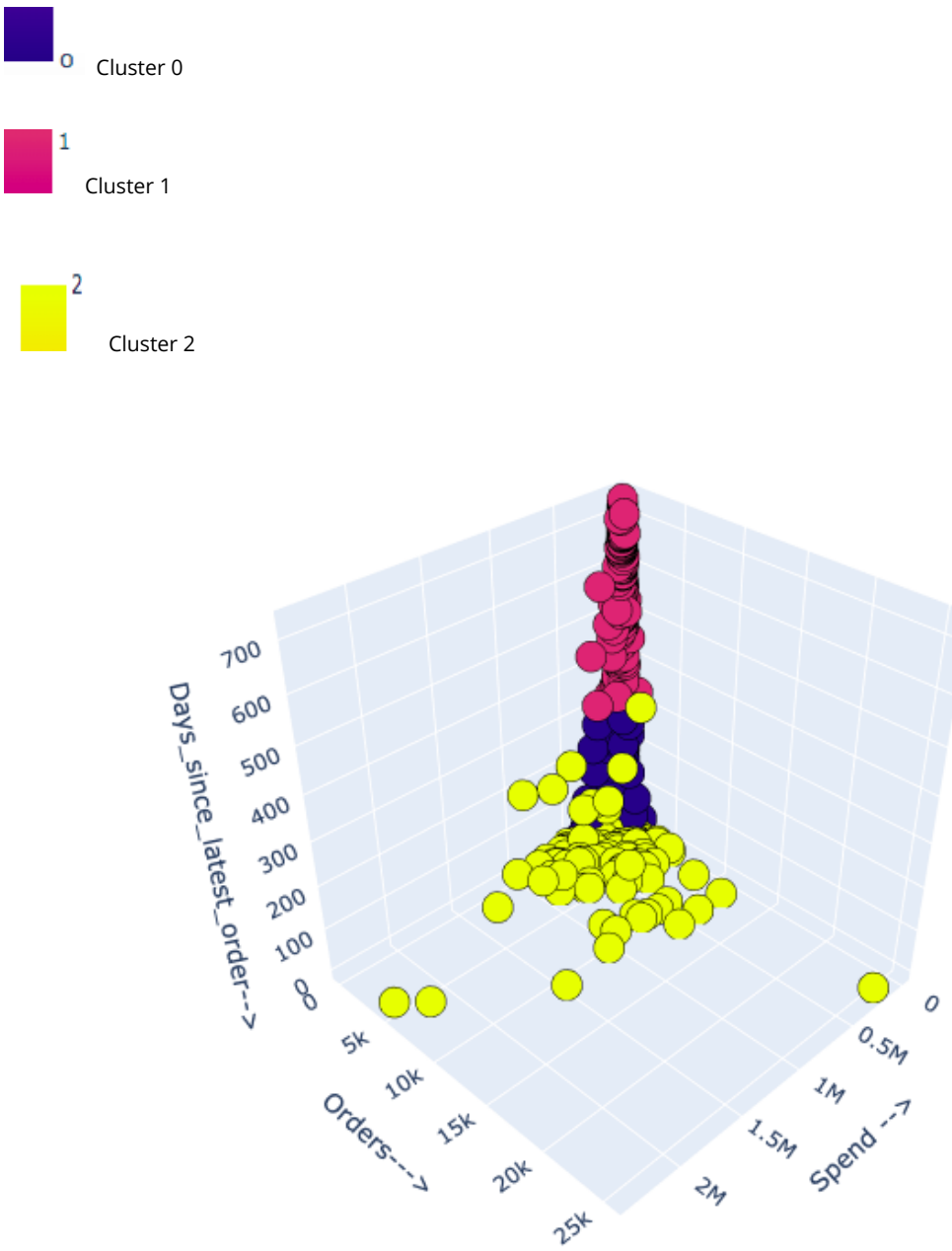
Silhouette Score Plots for k=2,3,4,5



Based on Elbow method and Silhoutte Score Plots k=3 seems optimal no of clusters, in the above silhoutte plot,k=4 and 5 have very different cluster sizes and k=2 has just one big cluster wheareas in k=3 we can see 2 distinct clusters and one very small cluster

Visualize Clusters based on RFM

The clusters obtained after K Means clustering are added to the original merchant data. We now have 3 different clusters based on Recency, Frequency and Monetary features


0
Cluster 0
1
Cluster 1
2
Cluster 2
Days_since_latest_order-->
700
600
500
400
300
200
100
0
Orders--->
5k
10k
15k
20k
25k
Spend -->
0
0.5M
1M
1.5M
2M

Cluster Summary

Based on cluster visualization above and the statistics below, we can conclude the following.

Metrics	Cluster 0	Cluster1	Cluster2
No of Merchants	10,870	3,327	154
% of total merchants	76%	23%	1%
% Total Gross Revenue	66%	5%	29%
Recency (avg days since last order per merchant)	53	414	13
Order Gap (avg consecutive orders gap days per merchant)	20	14	0.4
Monetary (avg spend per merchant)	$14k	$3.4K	$440k
AOV (Avg order value)	$391	$324	$908
Frequency(avg order per merchant)	82	22	3516

Cluster 0 - It is a large group identified by merchants who contribute to two thirds of gross revenue , repeat a purchase within 20 days, have not been active since end of Oct 2034

Cluster 1 - It is a group of merchants who do not contribute much to the revenue,are mostly active seasonally during a year, make lower valued transactions, and also the frequency of transactions is low.

Cluster 2 - It is a group of merchants which are the core merchants , 1% of these merchants contribute to 29% of gross revenue, they make bulk transactions with very high order values,have been active in the last two weeks and the frequency of the purchases is very high.

Best to worst - Cluster 2>Cluster 0> Cluster 1

Cluster 2 merchants should be retained. Cluster 0 merchants can be engaged more.

PART 2

In our data, what we have is merchant level transaction information. WE are assuming that this is not a contractual business and the merchants have no obligation based on a contract, in this case we do not have a clear churn definition. We will try to build a

probabilistic model instead of making it a binary Classification model, as this is acse of non contractual churn.

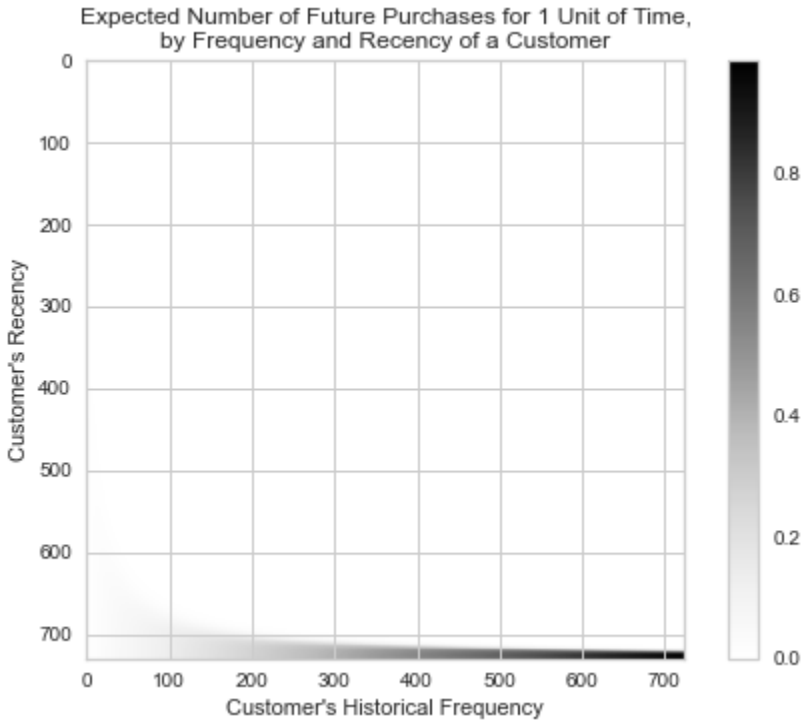
We will use the Beta geo Fitter model from Lifetimes package, based on Recency, Frequency and Monetary features.

- *frequency represents the number of repeat purchases the customer has made. It's the count of time periods the customer had a purchase in. So if using days as units, then it's the count of days the customer had a purchase on.*
- *T represents the age of the customer in whatever time units chosen (daily, in our dataset). This is equal to the duration between a customer's first purchase and the end of the period under study (31 Dec 2034)*
- *recency represents the age of the customer when they made their most recent purchases. This is equal to the duration between a customer's first purchase and their latest purchase. (Thus if they have made only 1 purchase, the recency is 0.)*
- *monetary_value represents the average value of a given customer's purchases. This is equal to the sum of all a customer's purchases divided by the total number of purchases.*

The dataset is converted into the format to reflect the above 4 features.

merchant	frequency	recency	T	monetary_value
0002b63b92	0.0	0.0	594.0	0.000000
0002d07bba	3.0	65.0	81.0	279.096667
00057d4302	1.0	66.0	580.0	91.350000
000bcff341	0.0	0.0	509.0	0.000000
000ddbf0ca	0.0	0.0	577.0	0.000000

Frequency Recency Matrix



We can see that if a customer has bought 600 times from us, and their latest purchase was when they were 700 days old , then they are our best customer (bottom-right). Our coldest customers are those that are in the top-right corner: they bought a lot quickly, and we haven't seen them in months.

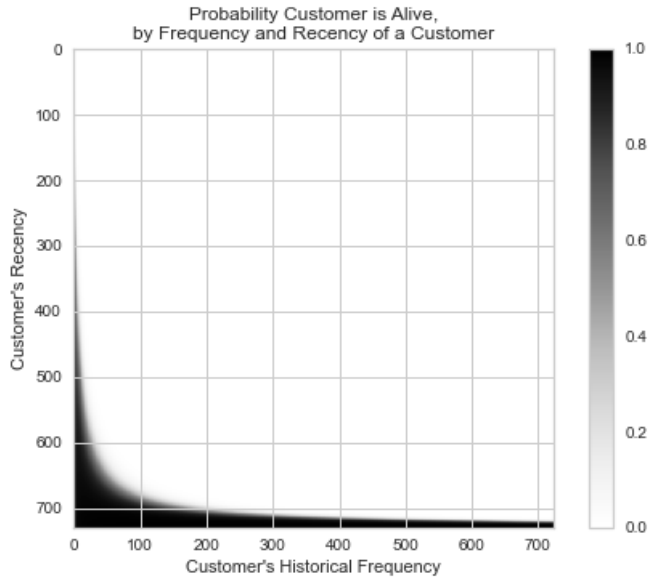
After Fitting the model, we come up with Probability of being alive as on 31 Dec 2034,
Churn Probability = 1- Probability of being alive

merchant	frequency	recency	T	monetary_value	probability_alive
0002b63b92	0.0	0.0	594.0	0.000000	1.000000e+00
0002d07bba	3.0	65.0	81.0	279.096667	9.086972e-01
00057d4302	1.0	66.0	580.0	91.350000	1.826430e-01
000bcff341	0.0	0.0	509.0	0.000000	1.000000e+00
000ddbf0ca	0.0	0.0	577.0	0.000000	1.000000e+00
...	...	...	...	...	...
ffd3e45675	4.0	23.0	726.0	139.997500	5.947391e-06
ffe1f6b51a	46.0	260.0	575.0	60.193478	3.881966e-14
ffe26b900d	65.0	334.0	374.0	143.492154	2.225668e-01
ffec05edb9	2.0	20.0	340.0	53.580000	1.192817e-02
fff1754102	19.0	504.0	510.0	266.602632	9.903620e-01

14351 rows × 5 columns

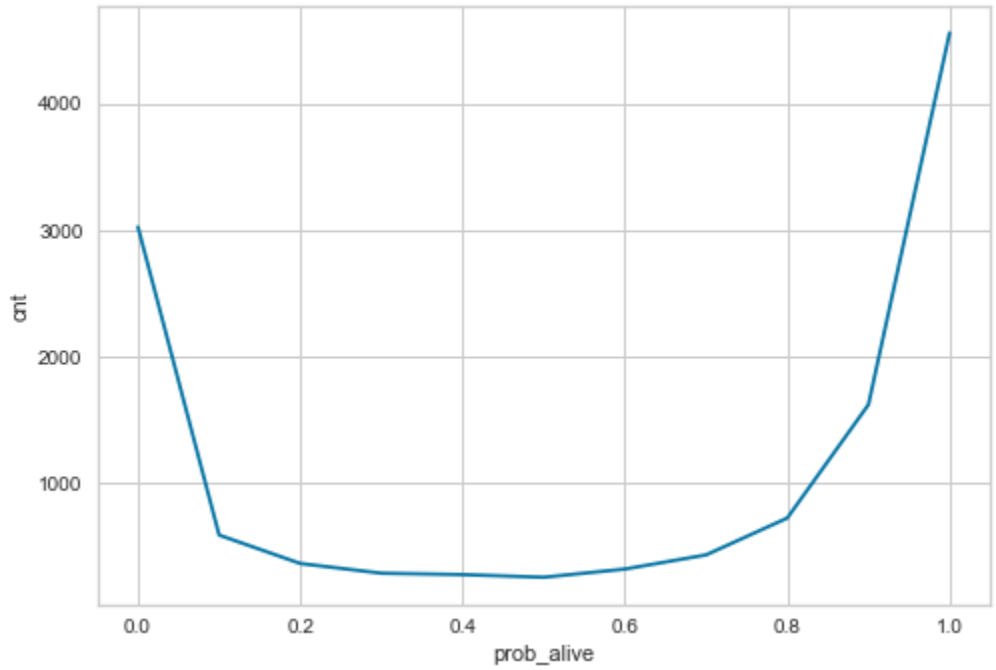
Note : We will remove all merchants with just one purchase, as the Beta Geo Fitter model predicts them to be alive with probability _alive =1 . We cannot be sure that they are alive. We are now left with 12676 merchants.

Probability Alive Matrix based on Frequency and Recency



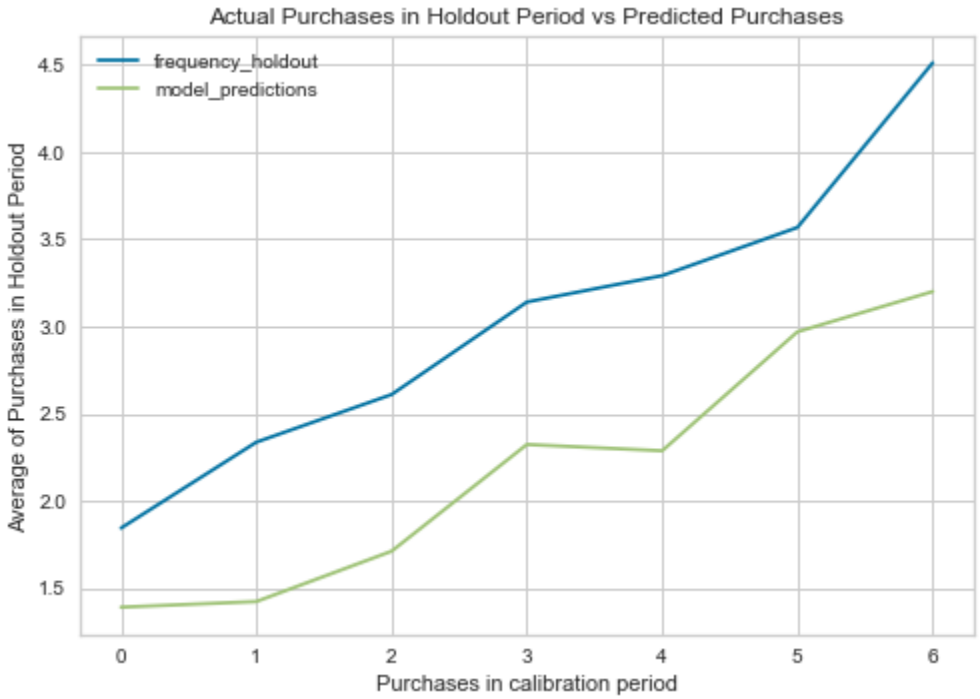
We see that if merchant's frequency is higher with longer age, he is more probable to not churn. Also, more recent merchants with around 100 orders are more probable to be active.

Line Plot showing Probability of being alive with the count of merchants



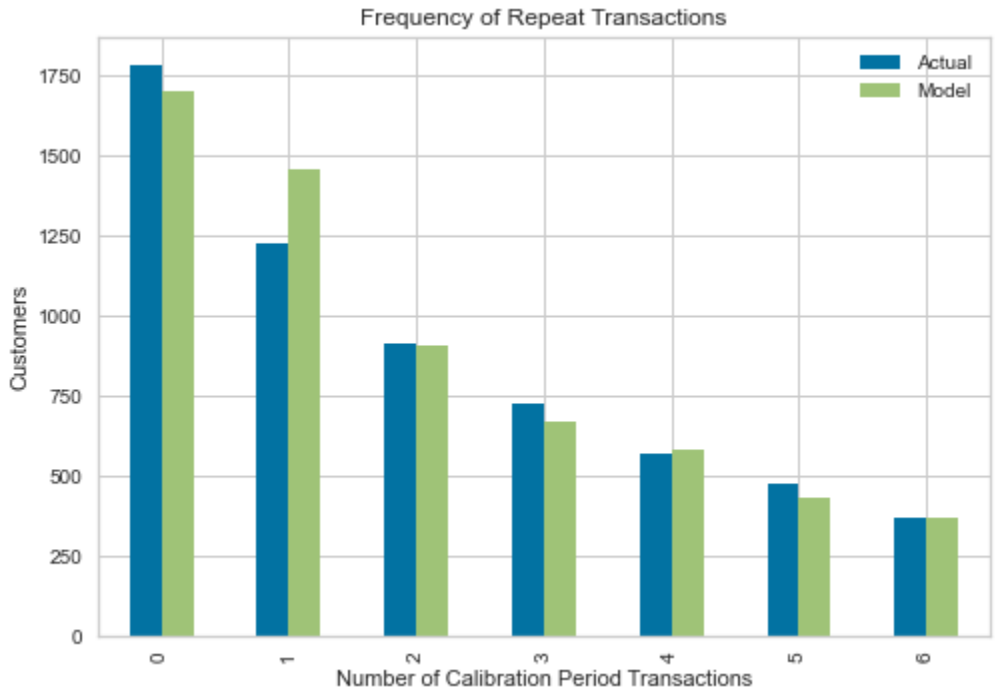
Model Evaluation

We have divided the period into calibration period and hold out period, 1 Jan 2033 till 1 Sep 2034 is the calibration period and the period after 1 Sep 2034 to 31 Dec 2034 is the hold out period. We will test the model's performance by training in calibration period data and testing it on the holdout period data



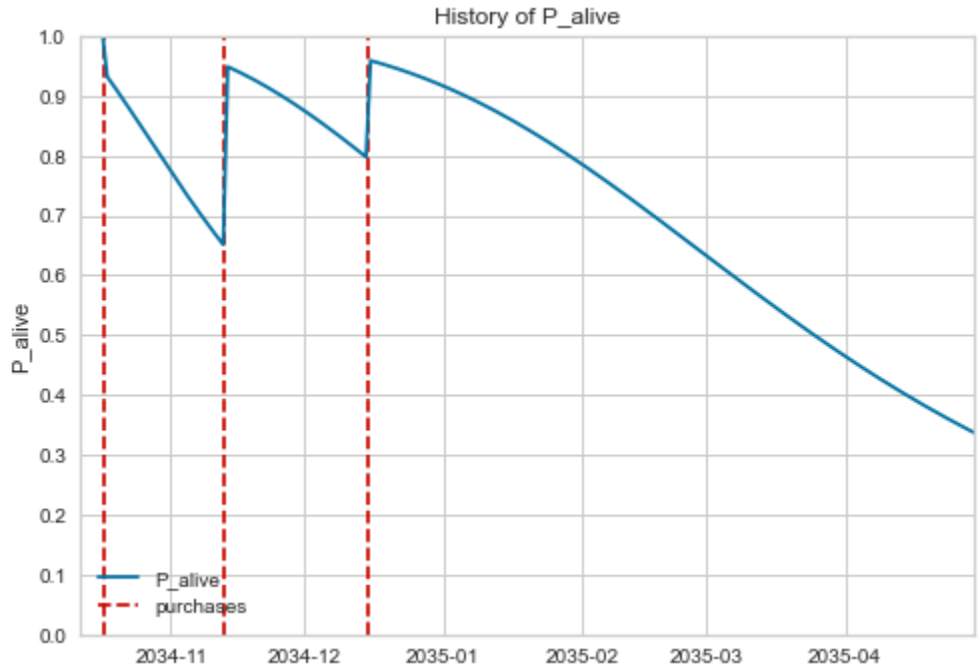
Model predicted purchases in the holdout period are close the actual purchases

Actual vs Predicted Frequency of repeat transactions



Plot for a Merchant, with history of Probability of being alive throughout his journey

Merchant='0002d07bba"



We see that this merchant is probable to be active at the end of the period and shall be highly probable to churn around April 2035

Already Churned Merchants

Based on the model analysis, we define churned merchants as the merchants with Probability of being alive =0 (rounding the probability to 1 decimal)

We find **3,021 merchants have already churned** as on 31 Dec 2034, which is 24% of all merchants.

Active Merchants likely to be churned

Based on the model analysis, we define likely to be churned merchants as the merchants with Probability of being alive greater than 0 but less than 0.5 (rounding the probability to 1 decimal)

We find 1,731 merchants are likely to churn in the near future (calculated as on 31 Dec 2034) , which is 17% of all currently active merchants.

Final_out dataframe in the Notebook attached has the merchant level information for churn. (1= churn, 0= not churn)