Customer Segmentation and Customer Behavior Purchase Prediction At Olist Marketplace

Citra Diani Putri

Outlines

1 **Introduction**
- Biodata
- SWOT Analysis of Olist.com
- Background of Project
- Knowledge and Literature

2 **Problems and Goals**
- Defining Problems
- Goals Elaboration

3 **Data Preparation**
- Data Description
- Data Wrangling

4 **Exploratory Data Analysis**
- Data Visualization
- Retention Rate
- Churn Rate

5 **Machine Learning**
- Data Preprocessing
- Feature Engineering
- Feature Selection
- Data Modelization

6 **Conclusion**
- Conclusion of Modelization
- Suggestions for the company

Biodata

Name:
Citra Diani Putri

Education:
Bachelor of Science, Biomedical Engineering
Bandung, Institute of Technology (ITB)

Hobby:
Watching Netflix

Company SWOT Analysis

The Olist is a large online marketplaces, formed by thousands of stores throughout Brazil.
Industry: Marketplace
Founded: 2015
Headquarter: Curitiba, Latin America
Website: olist.com
Founder: Tiago Dalvi

OLIST.COM

Strengths

Have 3 Millions Customers, Exclusive contract with the post office, Have Artificial Intelligence System that does automatic categorization for Sellers.

Weaknesses

Percentage of customers who don't repurchase on Olist in 2 years is high. Search Engine ranking of the company is still in below the top 10 E-Commerce in Brazil.

Opportunities

Brazil E-commerce jumps 57% as a result of the Covid-19 outbreak [1]. Therefore, shopping behaviours of customers is changing (from bulk-buying to online shopping). People are changing what they’re buying, when, and how.

Threats

Offline stores are becoming online, therefore sellers become increasing. COVID-19 is changing how B2B buyers and sellers interact.

Competitors

Submarino:
https://www.submarino.com.br/

Submarine is a company Brazilian e-commerce . Created in 1999, it was one of the pioneers of Brazil in this segment. In 2006, he joined with Americanas.com .

Americanas:
https://www.americanas.com.br/

Lojas Americanas is a Brazilian retail chain founded in 1929 in the city of Niterói, Rio de Janeiro. Currently, the company has 1490 stores in all 26 Brazilian states and in the Federal District.

Mercado Livre:
https://www.mercadolivre.com.br/

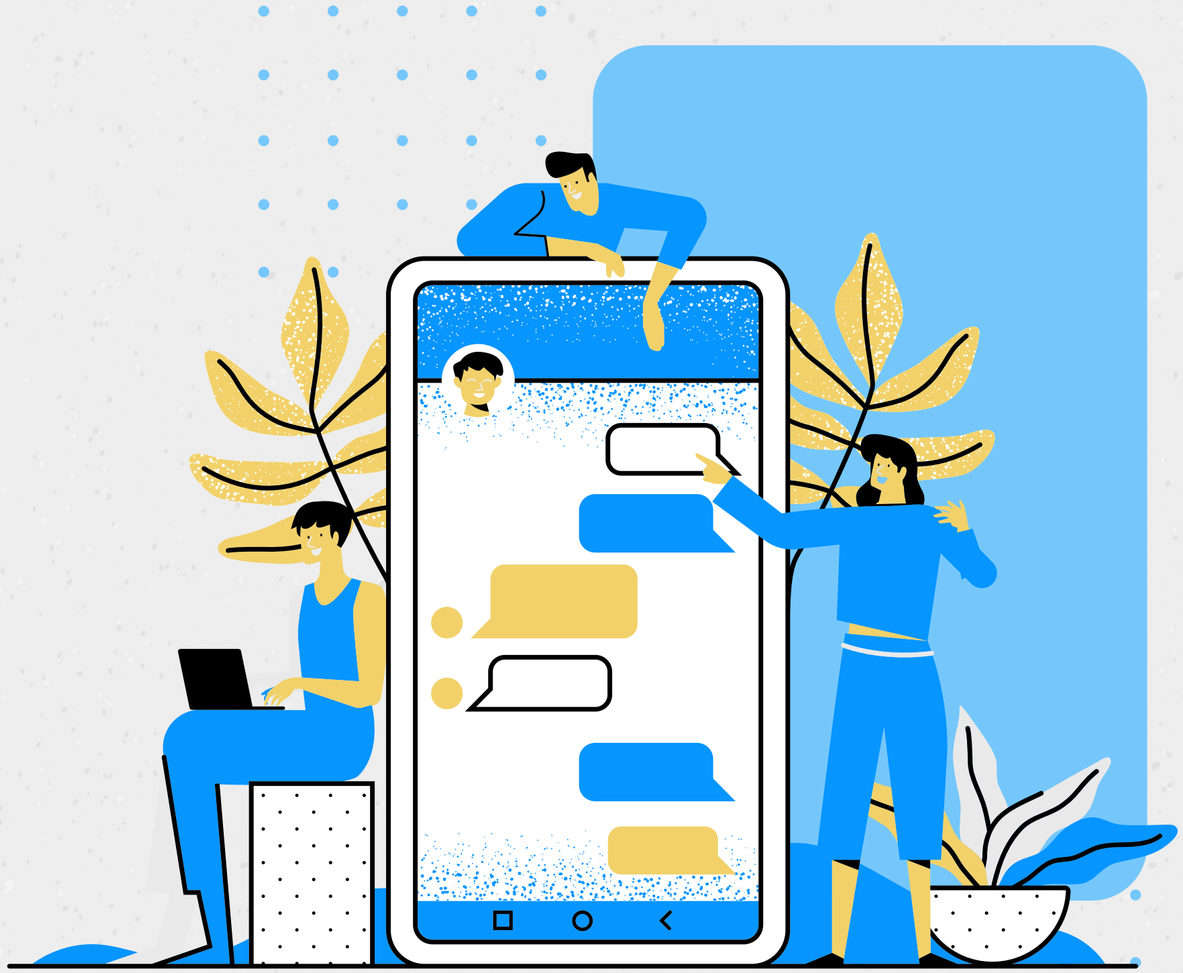
The Argentinian eCommerce site’s Cross Border Trading concept allows international sellers to reach all the regions in Latin America.

Customer Behavior in Brazil

60% of Brazilians won’t go back to a more expensive brand once they’ve made the switch.

Brazilian consumers often shop in popular discount chains and in shops combining retail and wholesale.

Brazilian consumers want to feel special.[5]



Problems

1. Customer's data regarding gender, location, etc might not be correct 100%, therefore, segmenting customers based on persona is not really accurate.
2. Olist has abundant customers who don't do repurchase in 2018.
3. To maintain sustainable growth of company, it's better to retain the existing customer than to get a new customer
4. It's harder to maintain customers than to get the new customers

Goals

1. Retarget/segmenting customers based on the customer's behavior using RFM method (Recency, Monetary and Frequency)
2. Making Retain Model by analyzing the customer behavior in one year and make a prediction of customer next purchase in the next 1 year. Classify the customers is conducted by dividing customer into chustomers who will do the next purchase < 4 months, > 4 month and > 8 months
3. Conducting Retention Rate and Churn Rate Analysis to be used as a metrics by Marketing Team

Article Insight

“Jika dulu perusahaan hanya mengandalkan impersona dalam melihat market, maka saat ini sudah tidak relevan. Impersona saja tidak menjamin bahwa target market sudah tepat. Terdapat kelompok-kelompok yang berbeda untuk satu kelompok impersona yang ada,” – Bayu Ramadhan (GO-JEK).[3]

Journal Insight

Dewasa ini, fokus perusahaan retail lebih kepada bagaimana menjual produk mereka kepada pelanggan yang sudah ada (existing customer) dibandingkan mencari pelanggan baru.

Studi yang menyatakan bahwa peningkatan sebanyak 2% dalam customer retention punya dampak terhadap laba seperti memangkas biaya sebesar 10%.[4]

To do lists

1) Segmenting Customers using RFM (Recency, Frequency and Monetary) Method using KMeans Clustering

2) Predicting the next purchase of customers using Classification Algorithm in Machine Learning

3) Data analysis regarding churn rate and retention rate

Marketing Calendar in Brazil

1. Valentine’s Day is June 12.
2. Mother’s Day is the second Sunday in May.
3. Father’s Day is the second Sunday in August
4. Children’s Day (October 12)

Datasets

- Source: https://www.kaggle.com/olistbr/brazilian-ecommerce

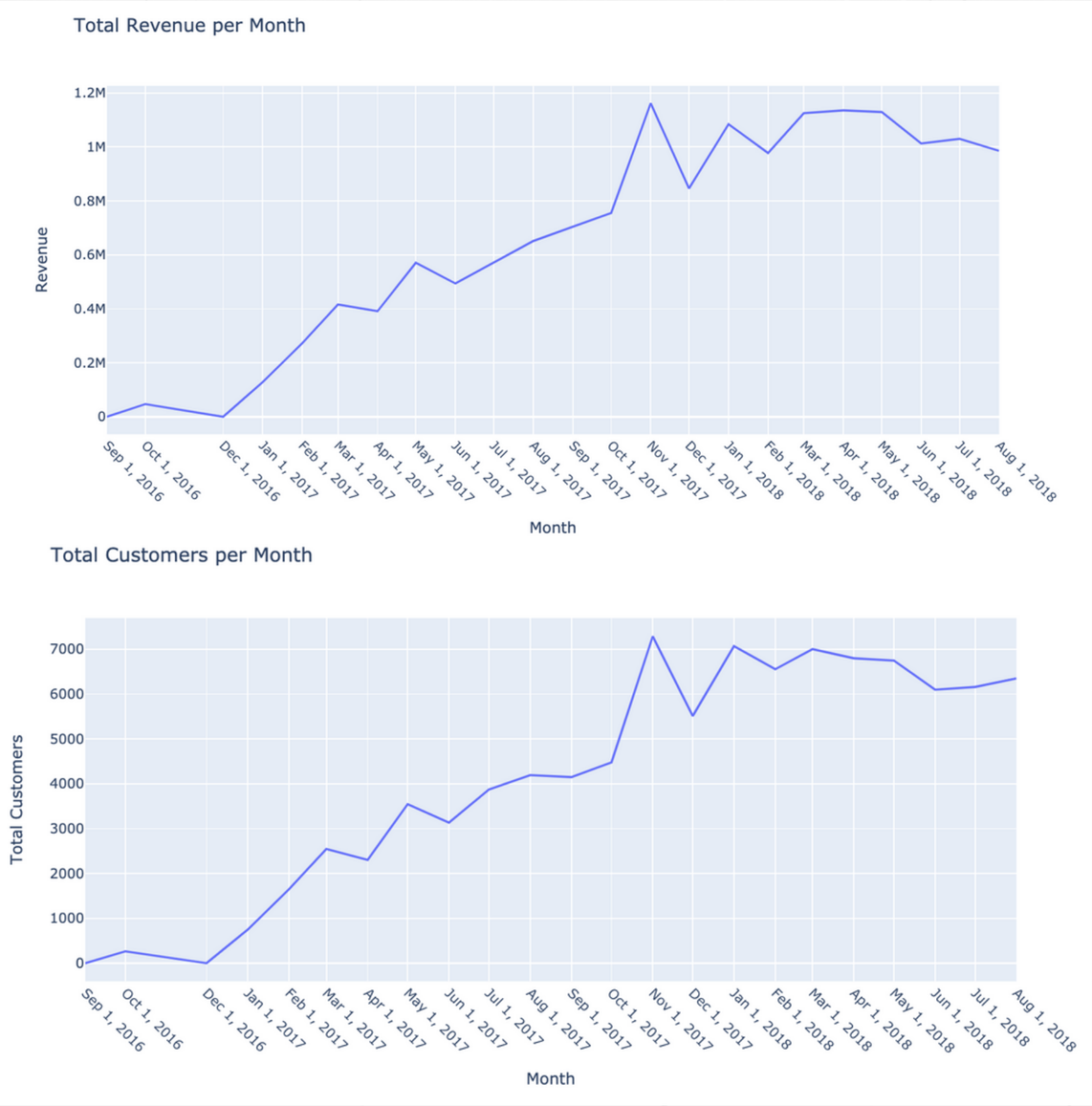
Period of transaction:
2016-09-15 12:16:38 to 2018-08-29 15:00:37

Merging Datasets:

1. Customers
2. Orders
3. Order_items
4. Products
5. Product_category_name
6. Geolocation
7. Order_review
8. Sellers

	columns	null_value_total	null_value_total_pct	unique_value_total	length	data_type
0	order_id	0	0.00	96478	110848	object
1	customer_id	0	0.00	96478	110848	object
2	order_purchase_timestamp	0	0.00	95956	110848	datetime64[ns]
3	customer_unique_id	0	0.00	93358	110848	object
4	customer_zip_code_prefix	0	0.00	14889	110848	float64
5	customer_city	0	0.00	4085	110848	object
6	customer_state	0	0.00	27	110848	object
7	order_item_id	0	0.00	21	110848	float64
8	product_id	0	0.00	32216	110848	object
9	seller_id	0	0.00	2970	110848	object
10	price	0	0.00	5859	110848	float64
11	freight_value	0	0.00	6924	110848	float64
12	product_category_name_english	1567	1.41	71	110848	object
13	customer_lat	293	0.26	14737	110848	float64
14	customer_lng	293	0.26	14737	110848	float64
15	review_score	0	0.00	5	110848	float64
16	seller_zip_code_prefix	0	0.00	2168	110848	float64
17	seller_city	0	0.00	595	110848	object
18	seller_state	0	0.00	22	110848	object
19	geolocation_zip_code_prefix	253	0.23	2161	110848	float64
20	geolocation_lat	253	0.23	2161	110848	float64
21	geolocation_lng	253	0.23	2161	110848	float64
22	order_date_day	0	0.00	31	110848	float64
23	order_date_day_name	0	0.00	7	110848	object
24	order_date_month	0	0.00	12	110848	float64
25	order_date_month_name	0	0.00	12	110848	object
26	order_date_year	0	0.00	3	110848	float64
27	order_date_monthyear	0	0.00	23	110848	object
28	order_date_hour	0	0.00	24	110848	int64
29	revenue	0	0.00	26639	110848	float64

Company's Condition



Total Revenue per Month

- Highest Total Revenue: November 2017

R$ 1,161,920.53

- Lowest Total Revenue: December 2017

R$ 19.62

- Highest Total Revenue Increase: Oct 2017 - Nov 2017

53.85%

Total Customers per Month

- Highest Total Customers: November 2017

7289

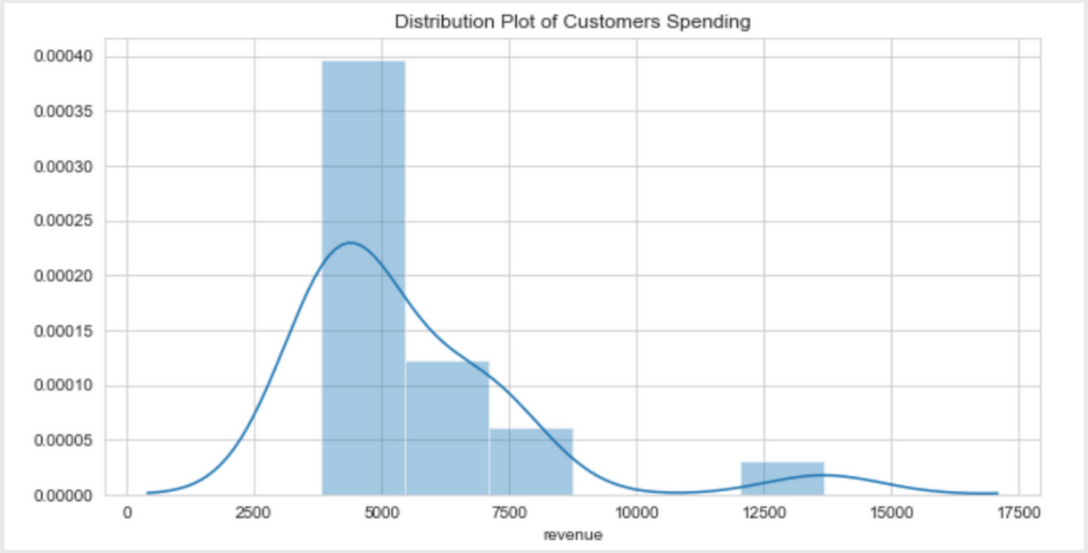
- Lowest Total Customers: December 2017

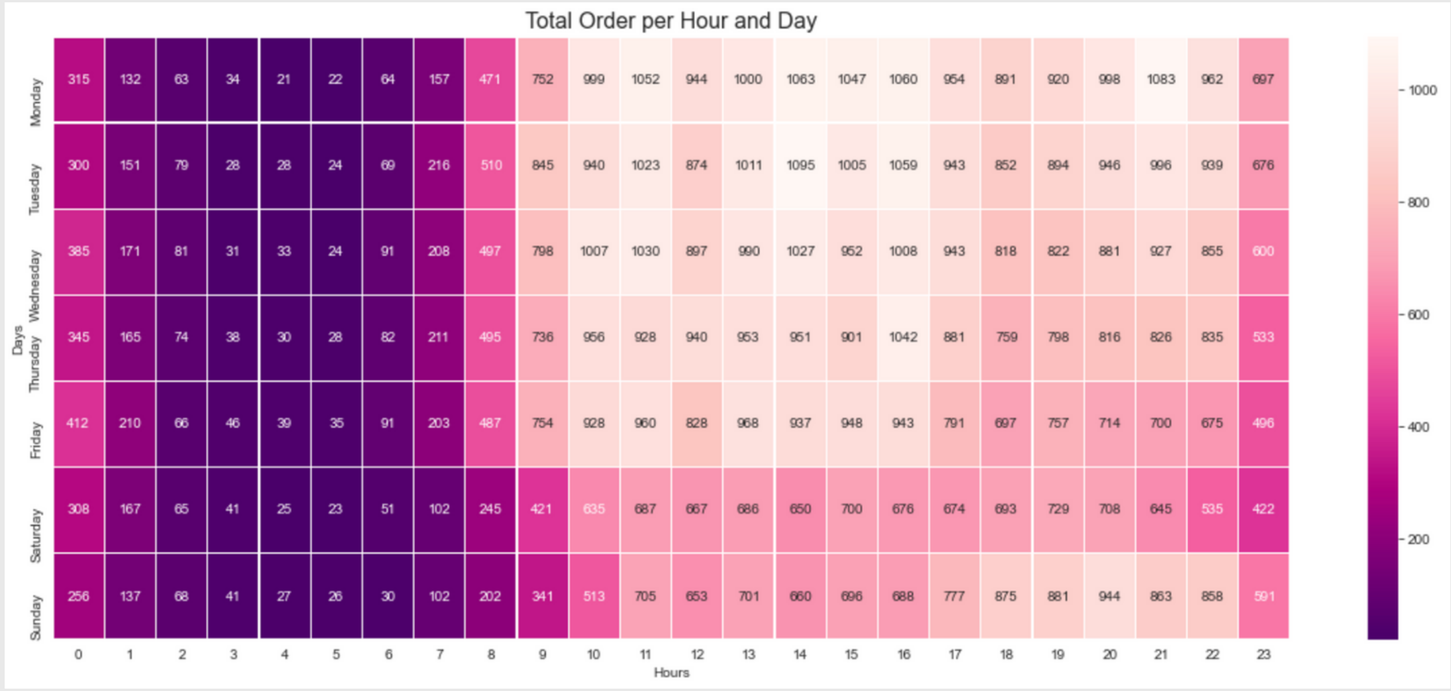
1

- Highest Total Customers Increase: Oct - Nov 2017

62.42%

Customer's Behavior





- Mostly customers spend money in 2 years

~R$ 4,000

- Active Order Hour:

10:00 AM - 22:00 PM

- Active Order Day:

Monday to Friday

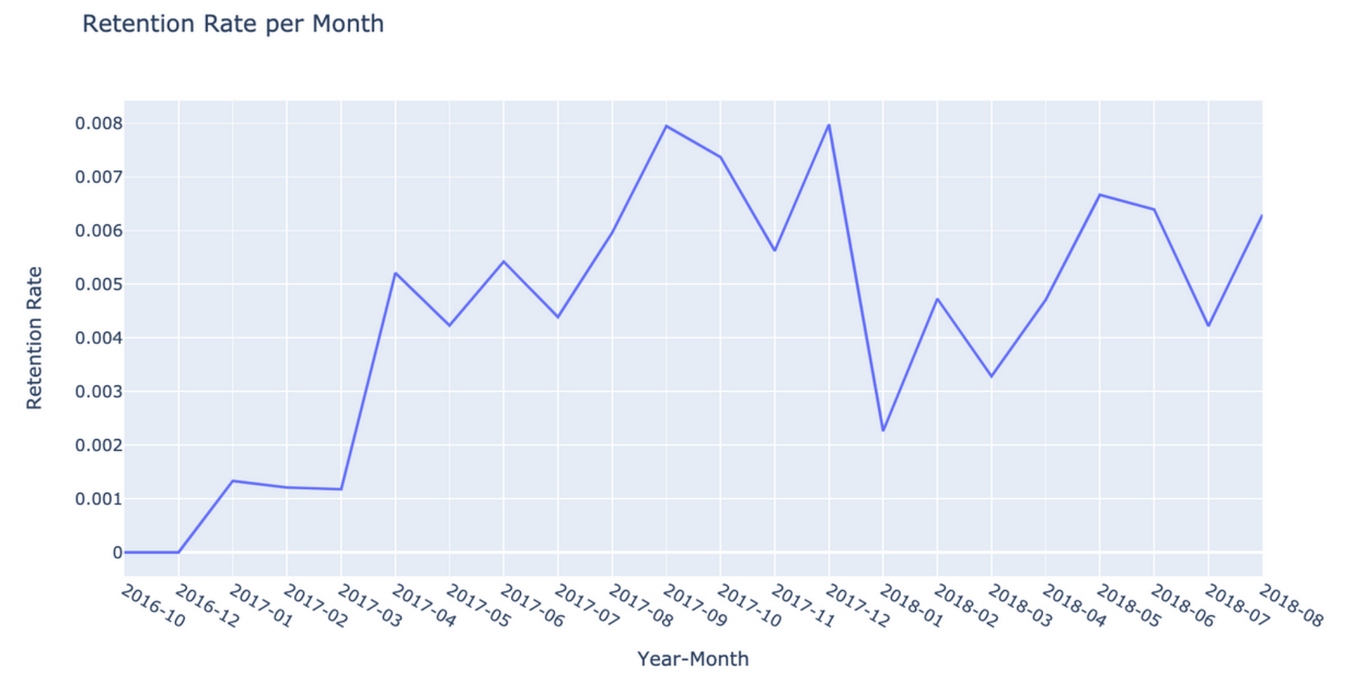
- Most Active Order Hour and Day:

Tuesday at 14:00 PM

Retention Rate

- Retention Rate is calculated by dividing the retained customer in the selected month and the previous month with the total customer in the selected month.
- Knowing how loyal our customers.

- Utilized as Marketing metric, to produce marketing strategy therefore customers could buy again in our store.
- It is hard to maintain old customers, than to get the new customers



- The largest Retention Rate:

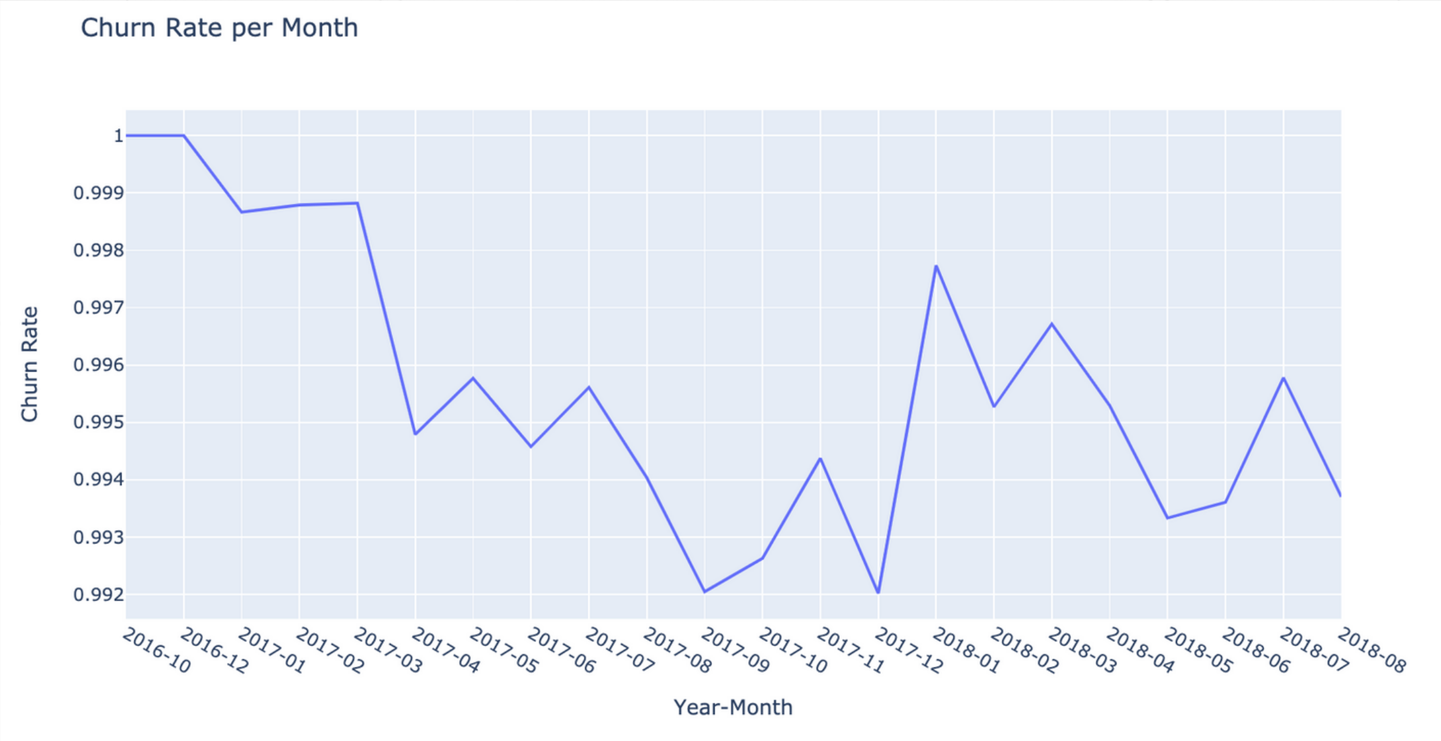
September 2017: 0.008%

December 2017: 0.008%

Churn Rate

Evaluating a company's customer loss rate in order to reduce it.
Churn Rate is calculated by subtracting 100% to retention rate.

- The lowest churn rate -> the better
- The high churn rate makes the company hard to grow



- Highest Churn Rate:

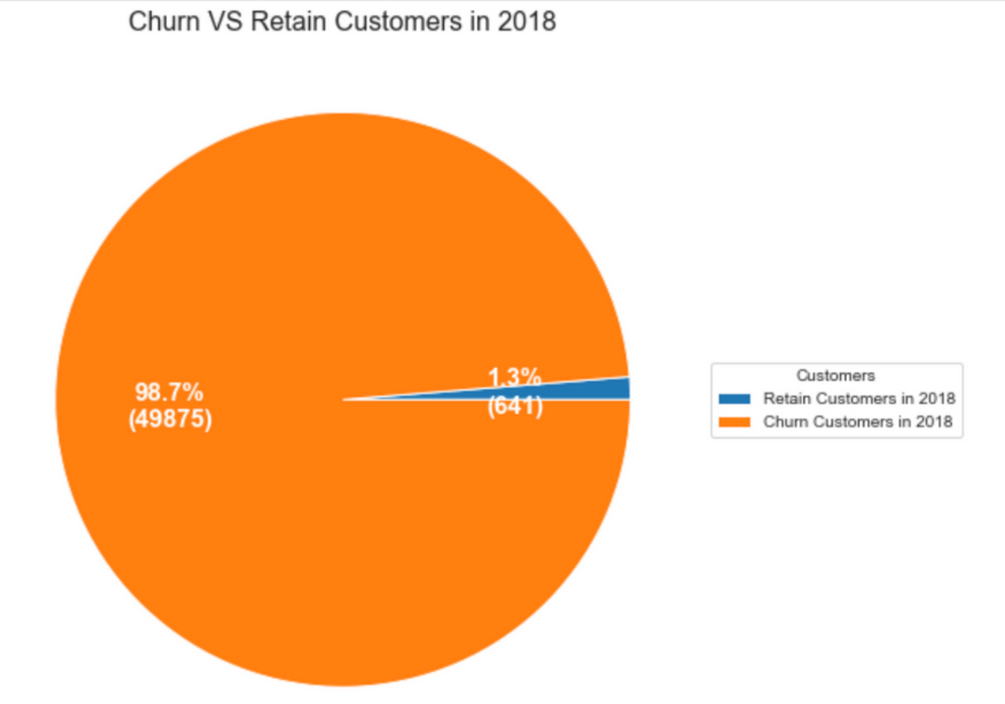
March 2017: 99.8822%

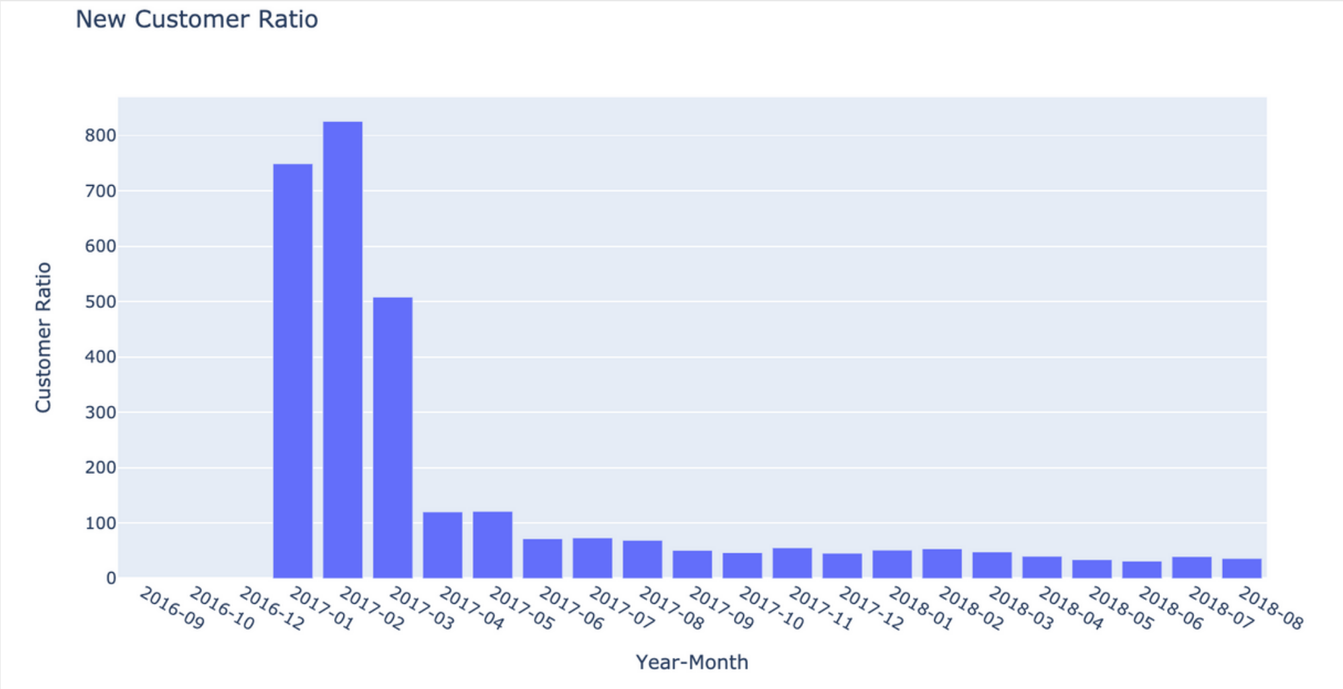
- Lowest Churn Rate Revenue:

September & December 2017
~99.2048%

New Customer VS Existing Customer

Comparison of customers who retain purchase in 2018 and the customers who only order in 2017.





- The Largest New Customer Ratio:

  February 2017: 825.50

- The lowest New Customer Ratio:

  June 2018: 31.44

Customer Segmentation

Dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing.

It is likely time to redefine engagement criteria in a segment that targets your most active customers who continue to interact with your brand amid these difficult times. To redefine engagement, consider adding conditions around purchase history that prompt engagement.

Benefits:

- Generate much higher rates of response, plus increased loyalty and customer lifetime value
- Each brand appears to sell effectively to only certain segments of any market and not to the whole market.

RFM Method

Recency

Recency is simply the amount of time since the customer's most recent transaction (most businesses use days, though for others it might make sense to use months, weeks or even hours instead).

Frequency

How often has a customer transacted or interacted with the brand during a particular period of time? Frequency is the total number of transactions made by the customer (during a defined period).

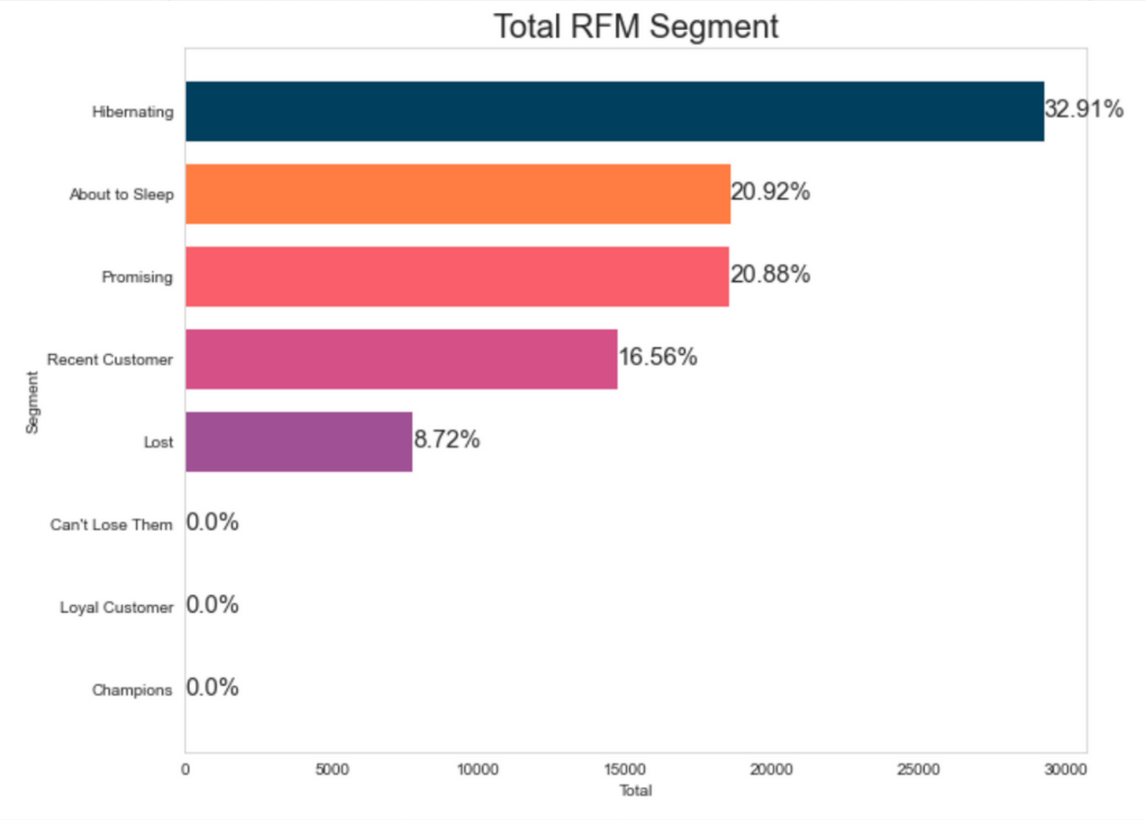
Monetary

Reflects how much a customer has spent with the brand during a particular period of time. Big spenders should usually be treated differently than customers who spend little. Monetary is the total amount that the customer has spent across all transactions (during a defined period).

Segmen	RFM Score	Deskripsi
Champions	453, 455, 552, 553, 555	Pelanggan baru saja membeli barang, sering membeli barang dan membelanjakan uang dengan jumlah besar
Potential Loyalist	451, 452, 551	Pelanggan baru saja membeli barang dengan frekuensi membeli diatas rata-rata
Recent Customer	512, 513, 515	Pelanggan baru saja membeli barang, tetapi tidak sering membeli barang
Promising	411, 412, 413, 415	Pelanggan baru saja membeli barang tetapi tidak berbelanja dalam jumlah besar
Loyal Customer	351, 352, 353, 355	Pelanggan selalu membeli barang, responsif terhadap promosi
Needs Attention	-	Recency, Frequency dan Monetary diatas rata-rata, bisa saja tidak baru membeli barang
About to sleep	311, 312, 313, 315	Recency dan Frequency dibawah rata-rata, akan kehilangan apabila tidak aktif kembali
At Risk	-	Sudah lama tidak membeli barang dan harus segera diberi promosi agar kembali lagi berbelanja
Can't Lose Them	151, 152, 153, 155	Pernah belanja sering namun nilai recency kecil
Hibernating	112, 113, 115	Terakhir membeli sudah lama dan jumlah barang yang dibeli sedikit
Lost	111	Terakhir membeli sudah lama, barang yang dibeli sedikit dan jumlah yang dibelanjakan sedikit

Recency, Monetary and Frequency Analysis

RFM is calculated by dividing data into 5 equal part (Quintiles) 20% of data is below Q1 , 40% of data is below Q2, 60% of data is below Q3, 80 % of data is below Q4.



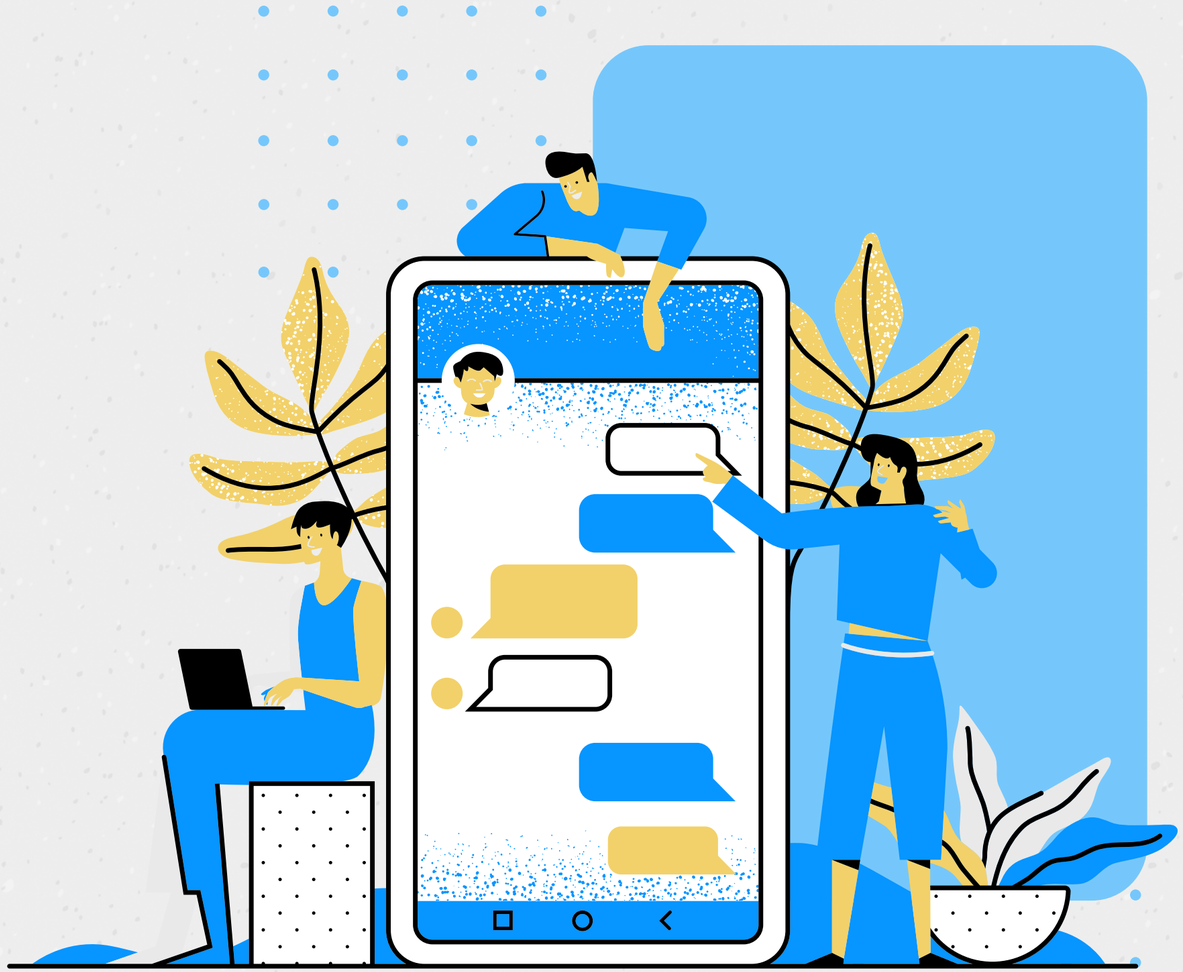
	Segment
Hibernating	29242
About to Sleep	18588
Promising	18557
Recent Customer	14720
Lost	7753
Can't Lose Them	4
Loyal Customer	1
Champions	1

Treatment for Segmented Customers

Segmen	RFM Score	Treatment
Champions	453, 455, 552, 553, 555	Reward them. Can be early adopters for new products. Will promote your brand.
Potential Loyalist	451, 452, 551	Offer membership / loyalty program, recommend other products.
Recent Customer	512, 513, 515	Provide on-boarding support, give them early success, start building relationship.
Promising	411, 412, 413, 415	Create brand awareness, offer free trials
Loyal Customer	351, 352, 353, 355	Upsell higher value products. Ask for reviews. Engage them.
Needs Attention	-	Make limited time offers, Recommend based on past purchases. Reactivate them.
About to sleep	311, 312, 313, 315	Share valuable resources, recommend popular products / renewals at discount, reconnect with them.
At Risk	-	Send personalized emails to reconnect, offer renewals, provide helpful resources.
Can't Lose Them	151, 152, 153, 155	Win them back via renewals or newer products, don’t lose them to competition, talk to them.
Hibernating	112, 113, 115	Offer other relevant products and special discounts. Recreate brand value.
Lost	111	Revive interest with reach out campaign, ignore otherwise.

Customer Purchase Prediction

1. Using the behavior of customers in the year of 2017 to predict the next purchase in 2018
2. Class 0: Customers who will purchase after 8 months
3. Class 1: Customers who will purchase in between 4 to 8 months
4. Class 2: Customers who will purchase below 4 months
5. All Class have done transaction min. 2x at Olist



Data Preparation

Data Wrangling

1. Creating transaction data between January, 1st 2017 until December, 31st 2017
2. Creating transaction data between Januari, 1st 2018 until December, 31st 2018
3. Creating all unique customer who made purchase in 2017
4. Creating Repurchase Day by subtracting minimal purchase day in 2018 and maximal purchase day in 2017
5. Merging Repurchase Day and Customers of 2017

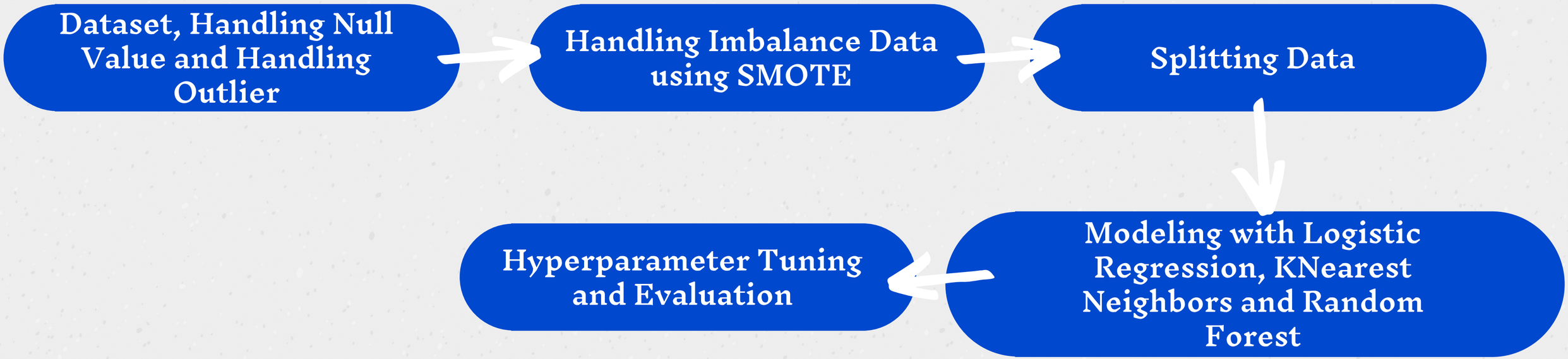
Feature Engineering

1. RFM Score, RFM Clusters, RFM Total Score and RFM Segment (and Binning the segment)
2. Creating data pervious purchase, in this project, 1 day of previous purchase is selected
3. Creating the difference day of purchase from maximal purchase in 2017 to customer's previous purchase

Dataset Summary

	columns	null_value_total	null_value_total_pct	unique_value_total	length	data_type
0	diff_day_1	0	0.0	212	1041	float64
1	RepurchaseDays	0	0.0	49	1041	float64
2	Recency	0	0.0	299	1041	int64
3	Frequency	0	0.0	6	1041	int64
4	Monetary	0	0.0	1026	1041	float64
5	R_Score	0	0.0	4	1041	int64
6	F_Score	0	0.0	5	1041	int64
7	M_Score	0	0.0	4	1041	int64
8	RFM_Total_Score	0	0.0	12	1041	int64
9	Segment_High-Tier	0	0.0	2	1041	uint8
10	Segment_Low-Tier	0	0.0	2	1041	uint8
11	Segment_Mid-Tier	0	0.0	2	1041	uint8
12	Target	0	0.0	3	1041	int64

Machine Learning Process



Model Evaluation

The selected algorithm of a model is Random Forest with Hyperparameter Tuning, because the result of train and test data is less overfit than Random Forest without tuning. Additionally, the precision score is still exceptional.

	Score Logistic Regression Default (%)	Score Logistic Regression Tuning (%)	Score KNN Default (%)	Score KNN Tuning (%)	Score Random Forest Classifier Default (%)	Score Random Forest Tuning (%)
Model Score in Data Train	54.109303	73.383396	95.577806	100.0	100.000000	97.329996
Model Score in Data Test	54.833333	74.333333	91.500000	91.0	95.333333	94.333333

	Score LR Micro Default (%)	Score LR Tuning Micro (%)	Score KNN Default Micro (%)	Score KNN Tuning Micro (%)	Score Random Forest Classifier Default (%)	Score Random Forest Tuning (%)
accuracy	54.833333	74.333333	91.500000	91.000000	95.333333	94.333333
recall	66.000000	82.500000	96.000000	97.500000	96.000000	95.500000
precision	77.192982	73.991031	91.866029	91.981132	97.461929	96.464646
f1_score	71.159030	78.014184	93.887531	94.660194	96.725441	95.979899

Conclusion

- Olist is presumed as a company with the low retention rate, Olist has to find selling point to make the customers loyal to the brand
- RFM Segmentation could be used to Increased customer retention as well as revenue by classifying customer and treat them accordingly
- This model is conducted to predict the customers next purchase under 4 months (Class 2) with predictor that consist of 'diff_day_1', 'RepurchaseDays', 'Recency', 'Frequency', 'Monetary', 'R_Score', 'F_Score', 'M_Score', 'RFM_Total_Score', 'Segment_High-Tier', 'Segment_Low-Tier', 'Segment_Mid-Tier'
- The company will not give the discount/promo to Class 2 Customers because we predicted that they will repurchase in before 4 months
- We determined the Precision Class 2 as a focus in this model because the company is better to reduce False Prediction Class 2 as an actual Class 0 or 1
- Failed prediction of Class 2 will result the actual Class 0 or 1 missing the promotion/discounts that leads to the potential of churn

References

[1] https://www.nasdaq.com/articles/brazil-ecommerce-jumps-57-in-first-five-months-of-2020-fueled-by-covid-19-2020-06-23
[2] https://www.zdnet.com/article/e-commerce-sales-reach-all-time-high-in-brazil/
[3] https://marketeers.com/go-jek-andalkan-rfm-dalam-tentukan-segmentasi-pelanggan/
[5] https://www.wordbank.com/us/blog/market-insights/brazilian-consumer-behaviour/#:~:text=Although%20Brazilians%20are%20typically%20brand,and%2014%25%20wait%20for%20sales.&text=In%20general%2C%20Brazilians%20have%20a,up%20rating%20of%20just%205%25.