E-Commerce Evolution: A Comprehensive Analysis of Olist's Brazilian Dataset

PRESENTED BY

Group 7:

Nhan Yen Trang

Nguyen Thanh Long

Vu Mai Dung

Pham Xuan Loc

Nguyen Le Binh

OUTLINE

01
BACKGROUND

02
DATA OVERVIEW

03
DATA ANALYSIS

04
PREDICTING CUSTOMER SATISFACTION

05
CUSTOMER SEGMENTATION

06
CUSTOMER LIFETIME VALUE

07
RECOMMENDATION SYSTEM

01

BACKGROUND

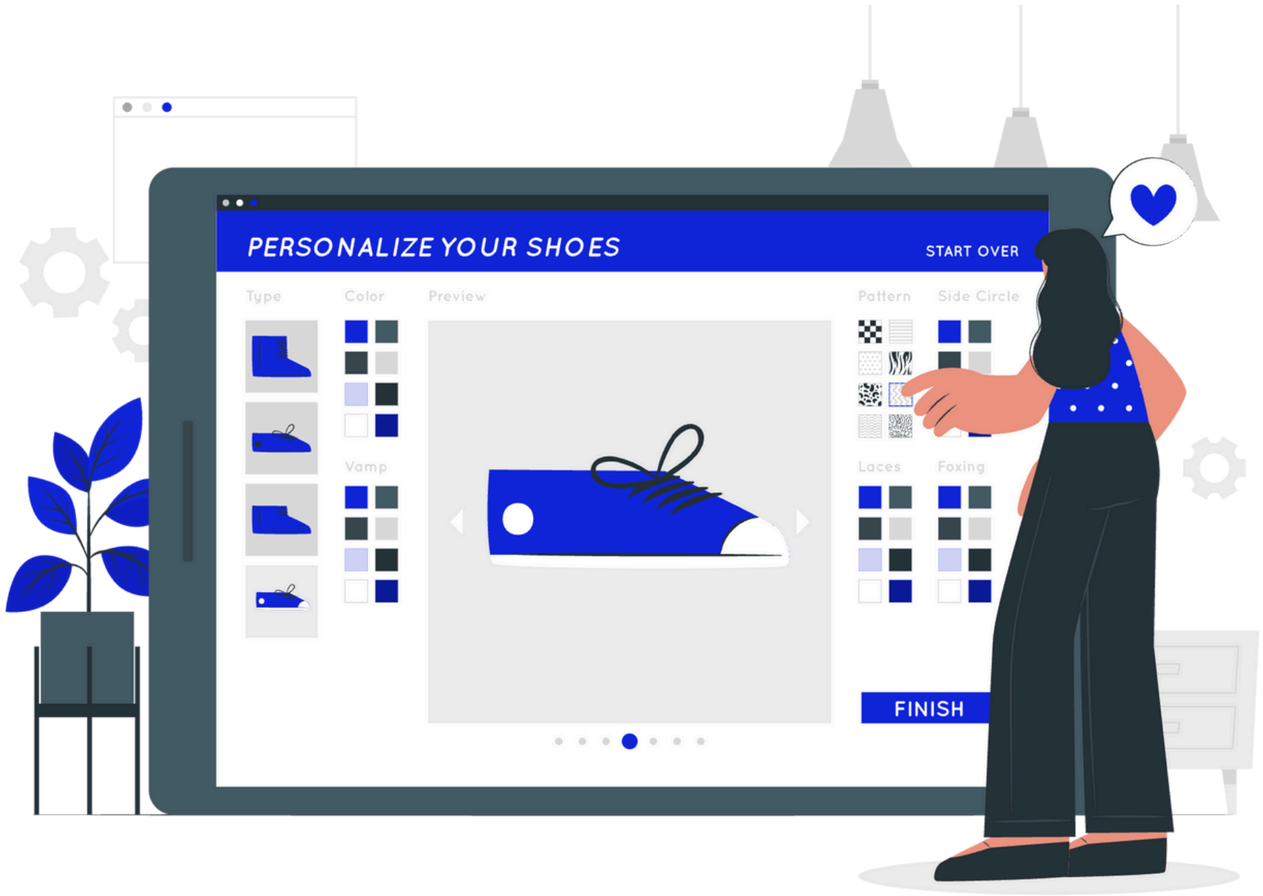
TOO MANY PRODUCTS MAKES PEOPLE CONFUSES

IN OTHER HAND, THE SELLER HAVE OBLIGATORY TO PROVIDE WHOLE PRODUCTS FOR ALL DAILY NEEDS

People is tending to following the * MAJORITY OF CHOICE

which end up with popular products


If a systemonly recommends
the popular product, customer
will get too boredwith the
same productsand might lose
interest



Because of that, we need to make

Personalized Buying Experience

by recommending their own preferences

which the tools called recommendation system

OBJECTIVES

Predict Review Score

Customer Lifetime Value Prediction

Customer Segmentation

Product recommendation system

DATA OVERVIEW

DATA SOURCE

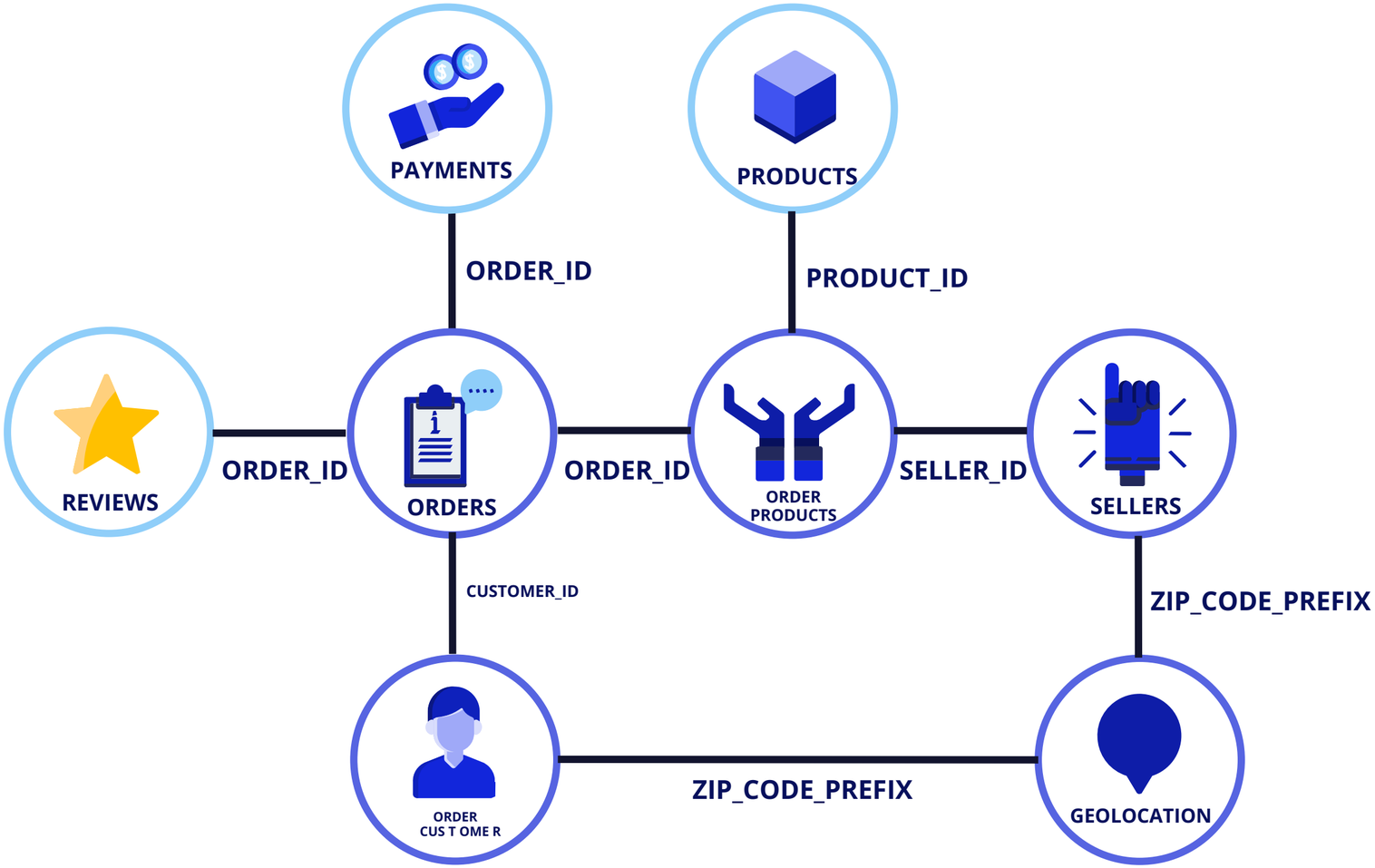
100k
ORDERS

2016 until 2018
PERIOD

8
DATASETS

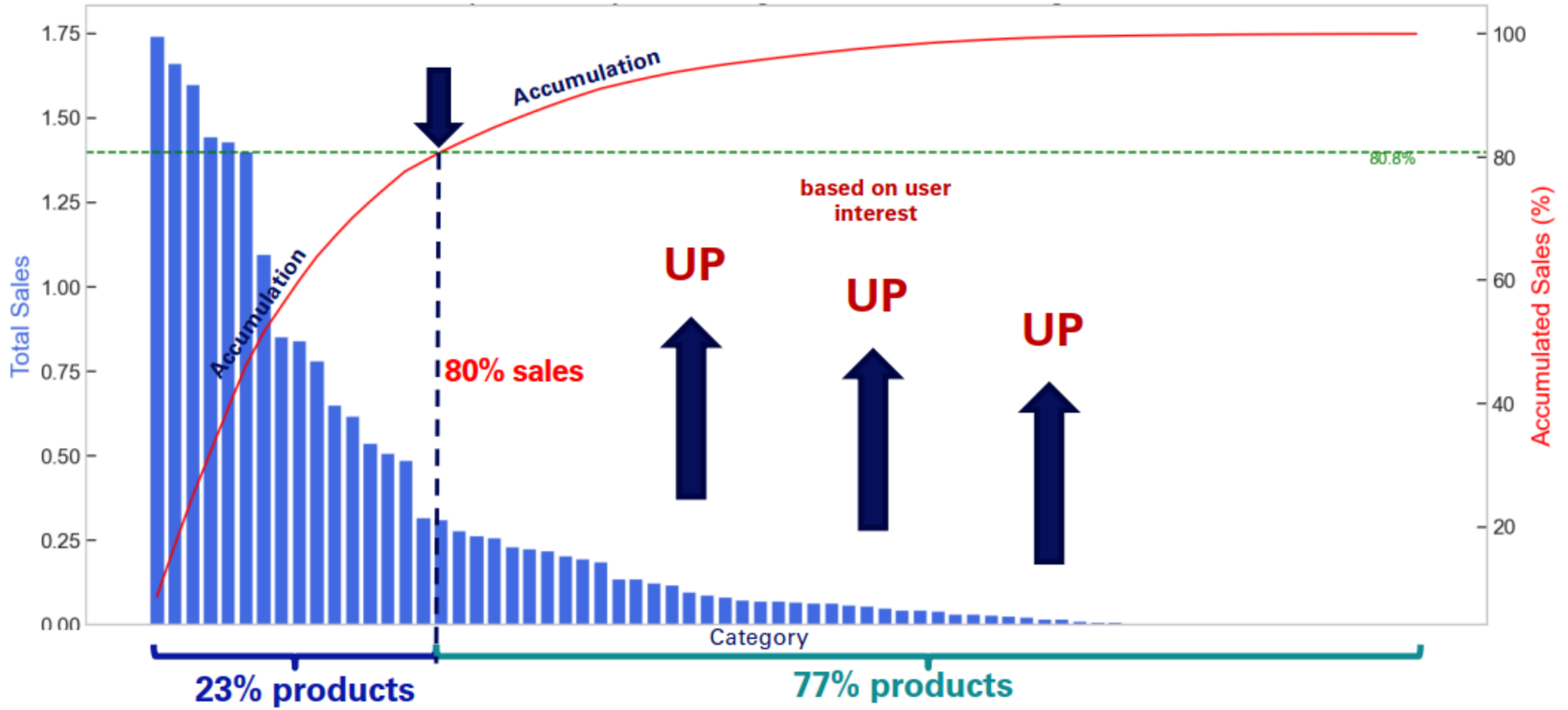
Olist is the largest department store in Brazilian E-commerce. Olist **connects small businesses** from all over Brazil to channels without hassle and with a single contract. Those merchants are able to **sell their products** through the OlistStore and **ship them directly to the customers** using Olist logistics partners.

OLIST DATA SCHEME



DATA ANALYSIS

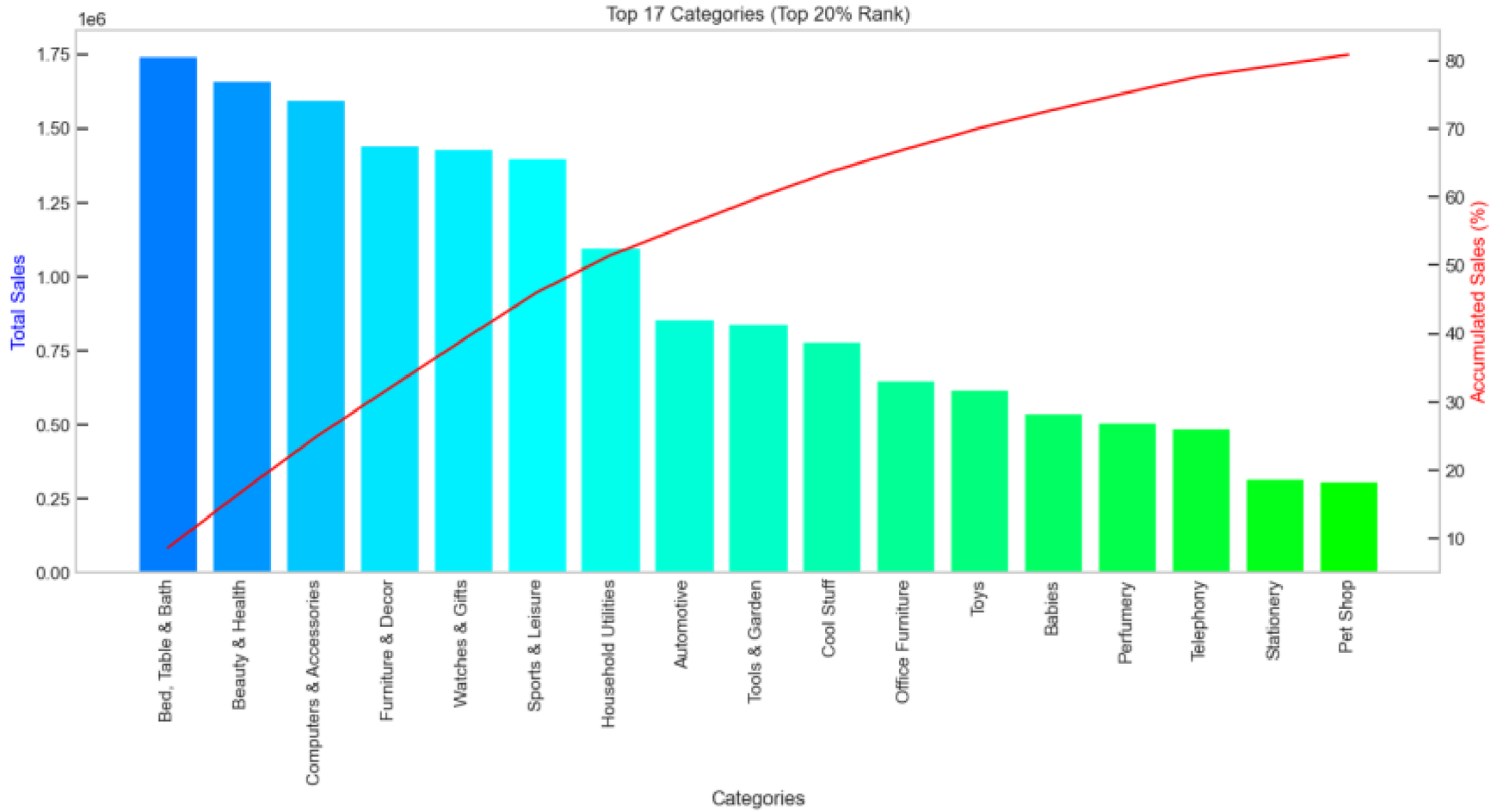
TOP 17 CATEGORIES FROM 72 CATEGORIES



23% of products contribute to 80% of total sales, which aligns with the Pareto principle and the long-tail phenomenon

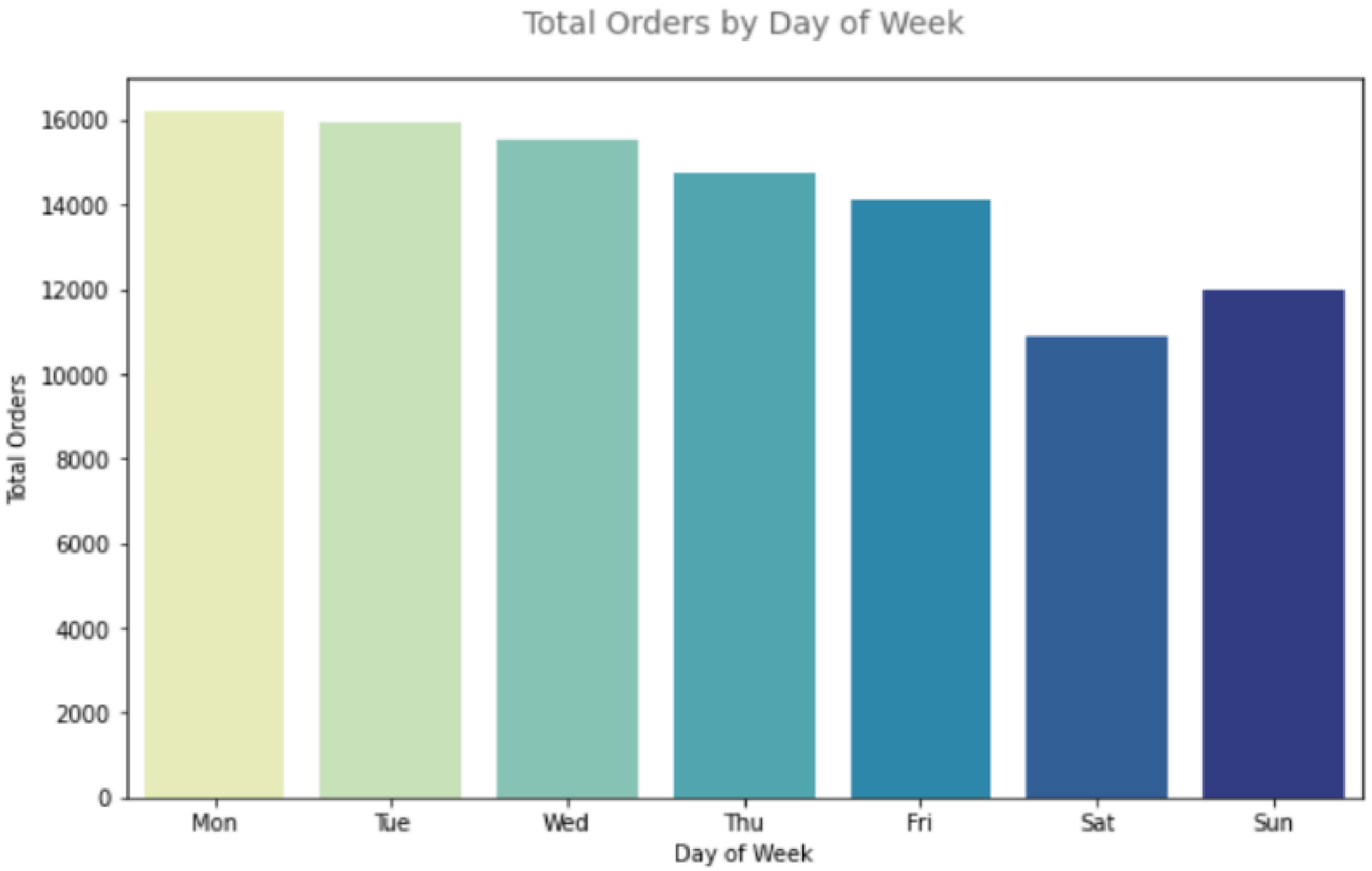
=> There are still many opportunities to recommend other products that may not have been fully exploited.

TOP 17 CATEGORIES (TOP 20% RANK)

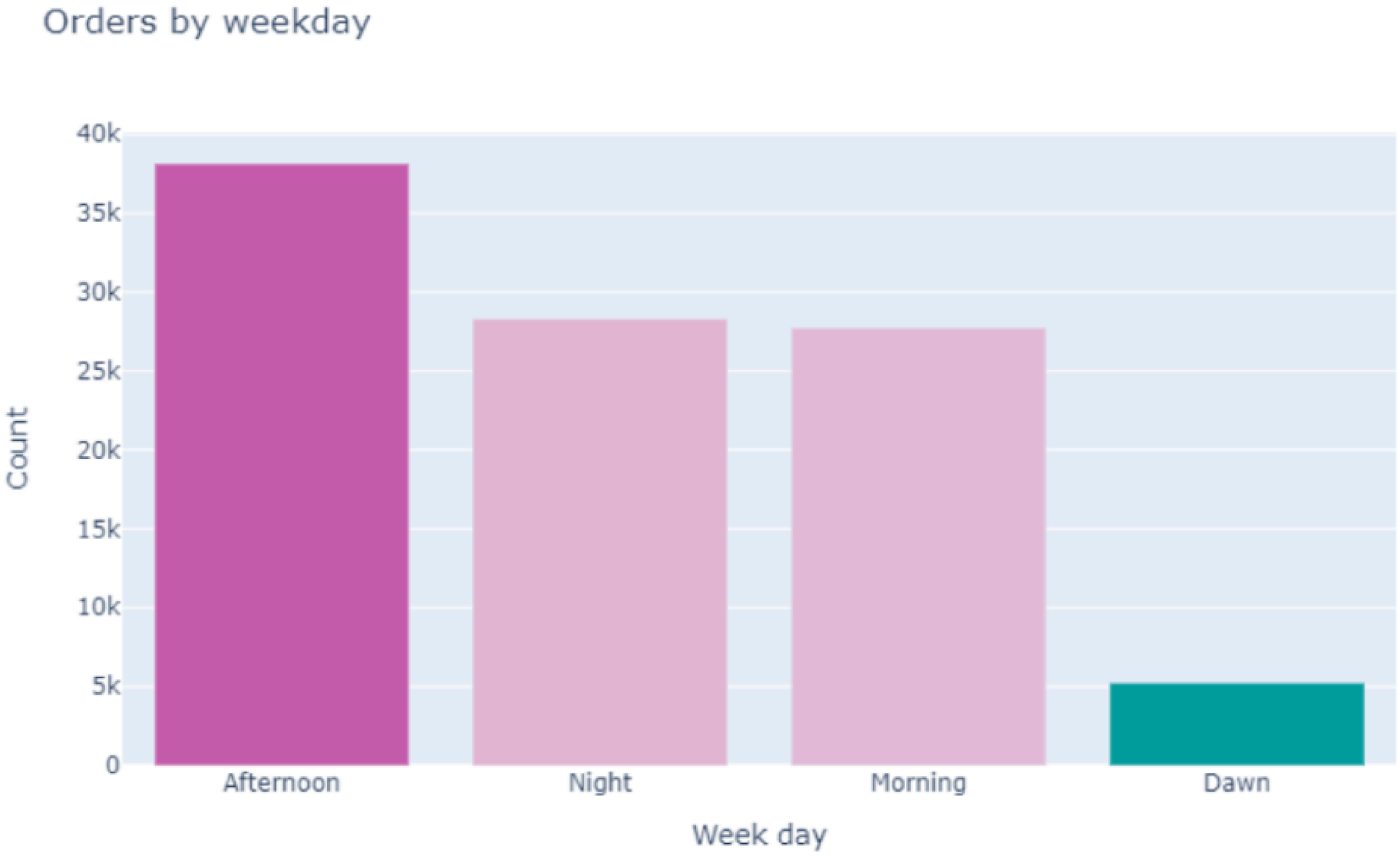


The top three categories on the chart are **Bed, Table & Bath; Beauty & Health; and Computer & Accessories,** while the next top three are Furniture & Decor; Watches & Gifts; and Sports & Leisure. Meanwhile, the remaining categories are significantly lower, with around 860,000 products

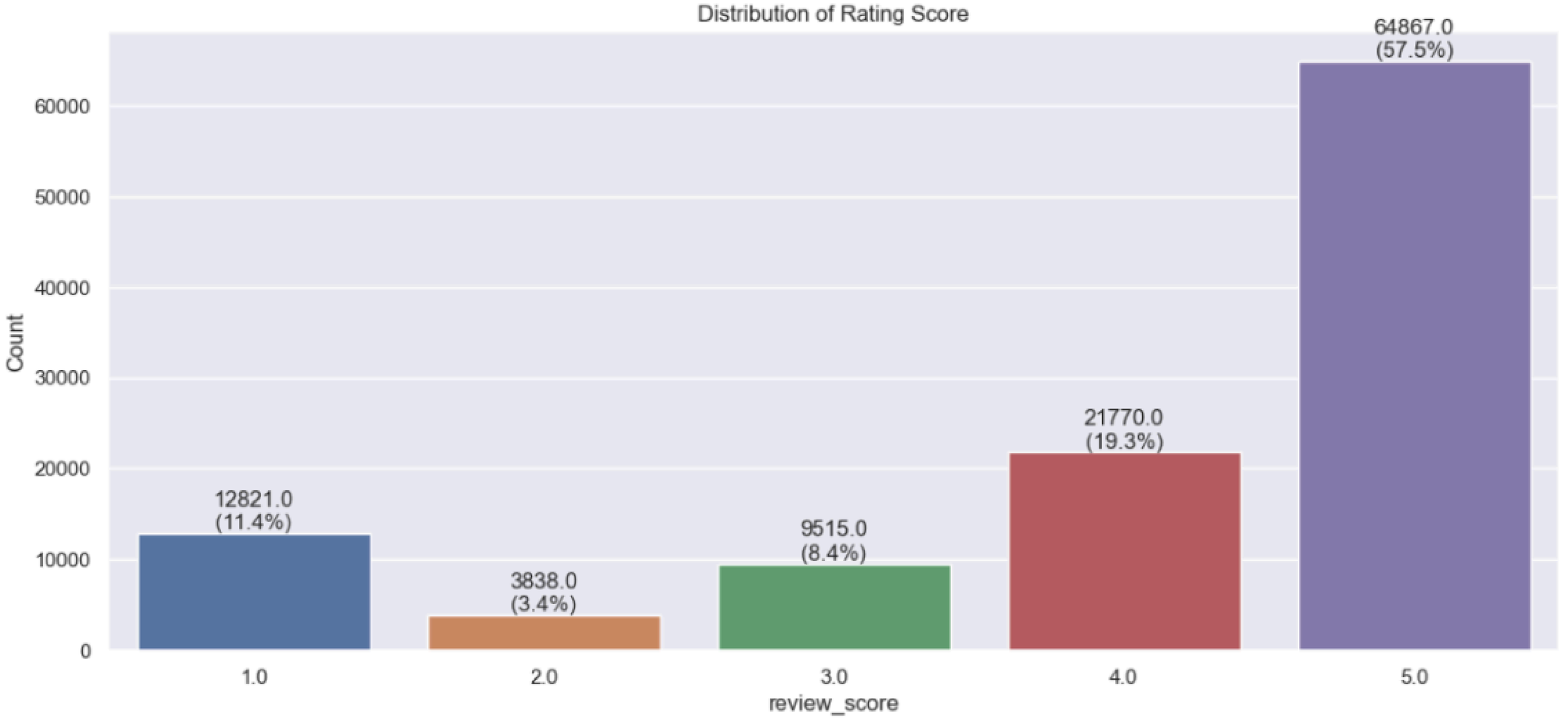
Most of the orders are getting **delivered before they are expected.**



Customers seem to most likely buy on **Monday** then any other day of the week.

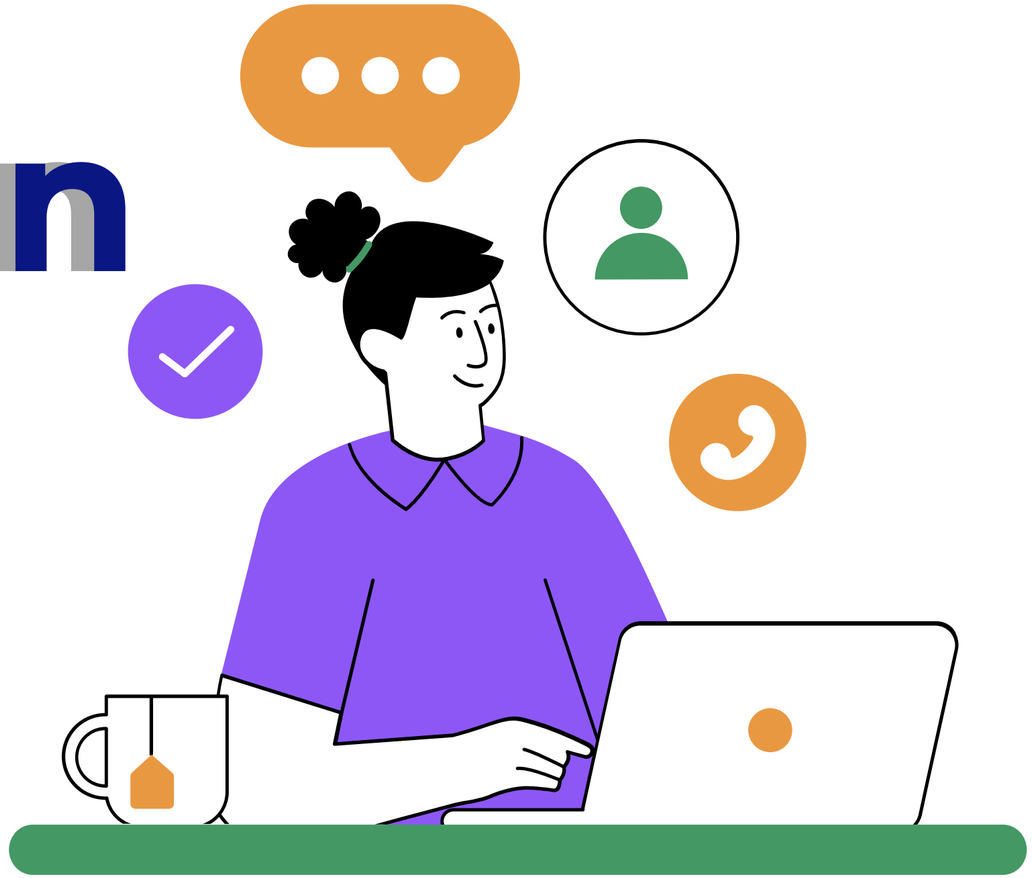


Afternoon boredom is making customers buy more items then any other time of the day.



Customer given review score mostly on 5 star score, and for others have gradually lower score except for 1 star score had high score

Predict Customer Satisfaction



Understand

The level of satisfaction that customers had with a product, service, or overall experience

MAIN GOAL

Adjust the strategies

To meet customers needs and expectation

product_name_lenght	1695
product_description_lenght	1695
product_photos_qty	1695
product_weight_g	20
product_length_cm	20
product_height_cm	20
product_width_cm	20
product_category_name_english	1720
review_score	0
review_comment_message	67650
customer_id	0
order_status	0
order_purchase_timestamp	0
order_approved_at	15
order_delivered_carrier_date	1235
order_delivered_customer_date	2471

HANDLING MISSING VALUES

- **review_comment_message** has 80% missing values
  ⇒ replaced it with **“no_review”**

- Filled the missing value in **order delivery date** and **order approve date** from the corresponding estimated **delivery date** column and **order purchase time** column.

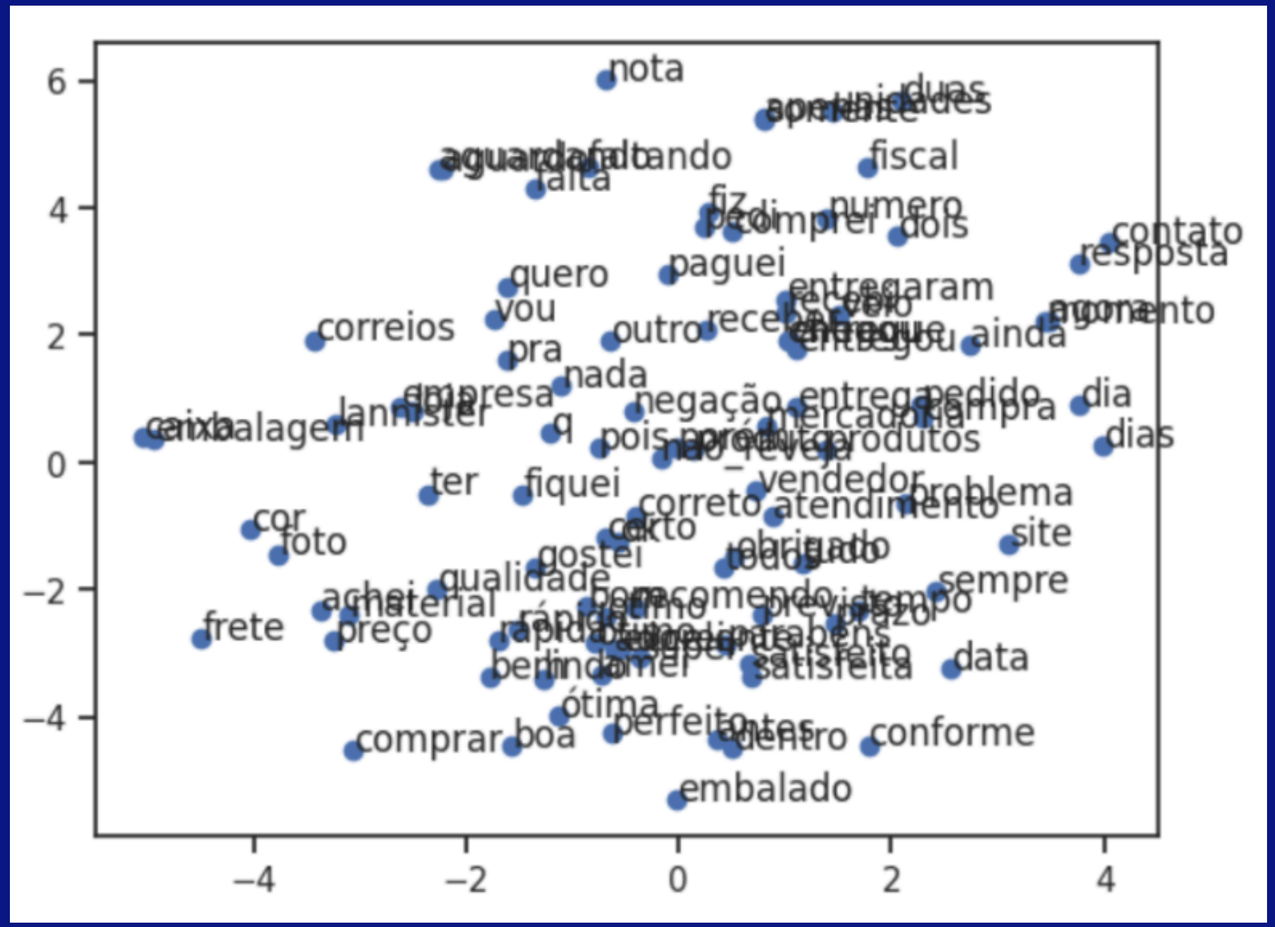
- **features has less than 1% missing values**
  ⇒ we dropped

- We dropped **order_delivered_carrier_date** column

FEATURE ENGINEERING

We created 11 new features and we analyzed them.

Besides, “word2vec” for the feature ‘Review_comment_mesage’ is also used



2D scatter plot where each point represents a word, and words with similar contexts or meanings are expected to be closer to each other in the plot.

11 NEW FEATURES

Seller & Order Details:

- **Sellers Count:** Explores competition by indicating the number of sellers offering each product.
- **Products Count:** Analyzes order composition by revealing the number of items included per order.

Delivery Time Analysis:

- **Estimated Delivery Time:** Captures the timeframe promised by the seller.
- **Actual Delivery Time:** Records the actual duration it took to deliver the order.
- **Difference in Delivery Days:** Calculates the gap between estimated and actual delivery times, potentially indicating delays or early arrivals.
- **Is Late:** A binary variable that simplifies analysis by showing if an order arrived after the estimated timeframe.

11 NEW FEATURES

Customer Experience Indicators:

- **Average Product Value:** Allows exploration of potential links between product price and customer satisfaction.
- **Total Order Value:** Investigates the relationship between order value and customer experience.
- **Order Freight Ratio:** Analyzes the cost of shipping relative to the total order value.
- **Purchase Day of Week:** Helps identify potential trends in buying behavior based on the day purchases are made.

Review Information:

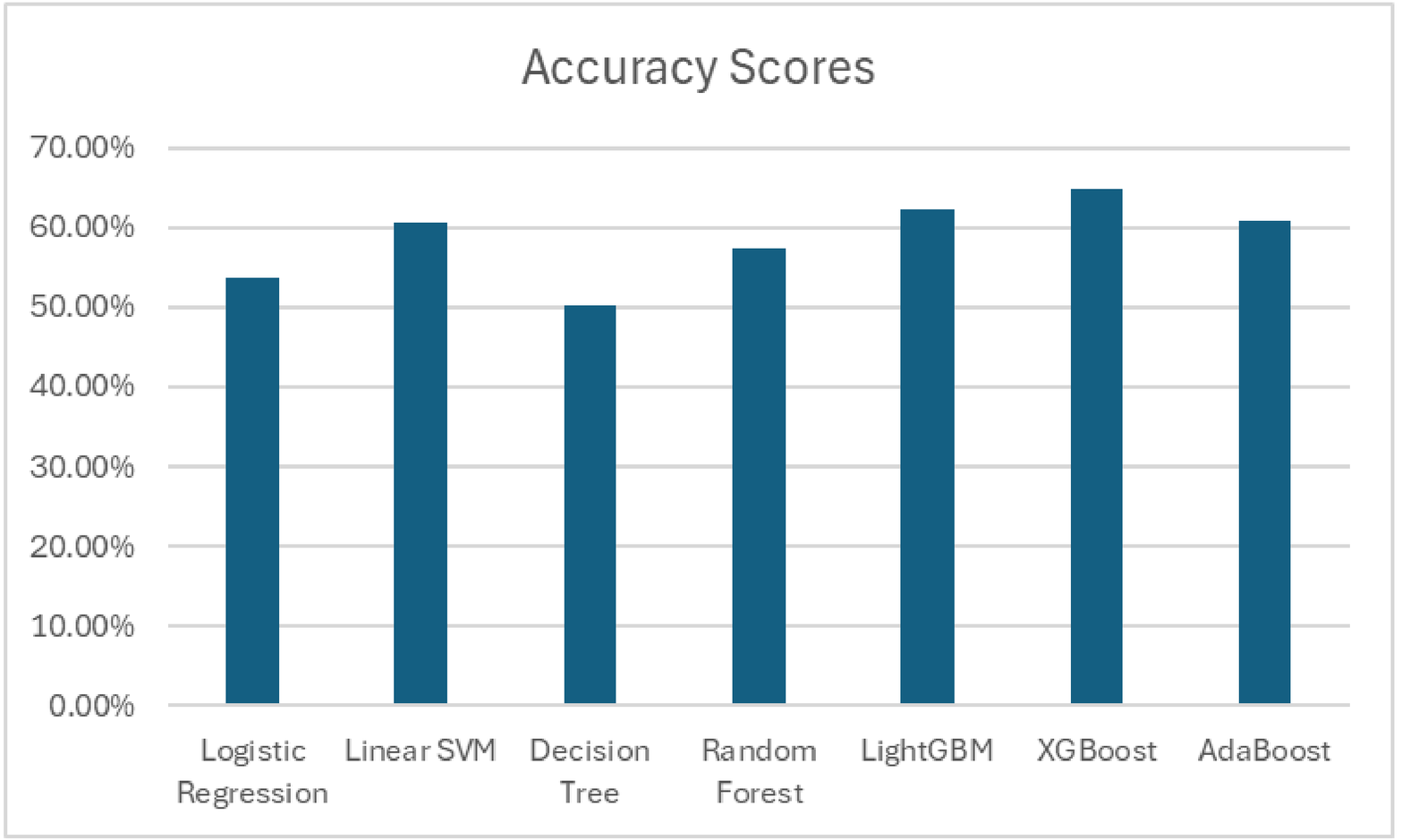
- **is_reviewed:** Indicates whether a customer left a comment in their review, providing additional insights into customer sentiment.

Word2Vec & Tensordflow - Pad_sequences

In TensorFlow, pad_sequences is a function used to ensure that sequences of variable lengths have the same length by padding them with zeros or truncating them.

Word2Vec is an algorithm that learns vector representations of words from large text datasets to capture semantic relationships and improve natural language processing tasks.

MODEL BUILDING



XGBoost has the highest specific performance score of 64.7%

Customer Segmentation

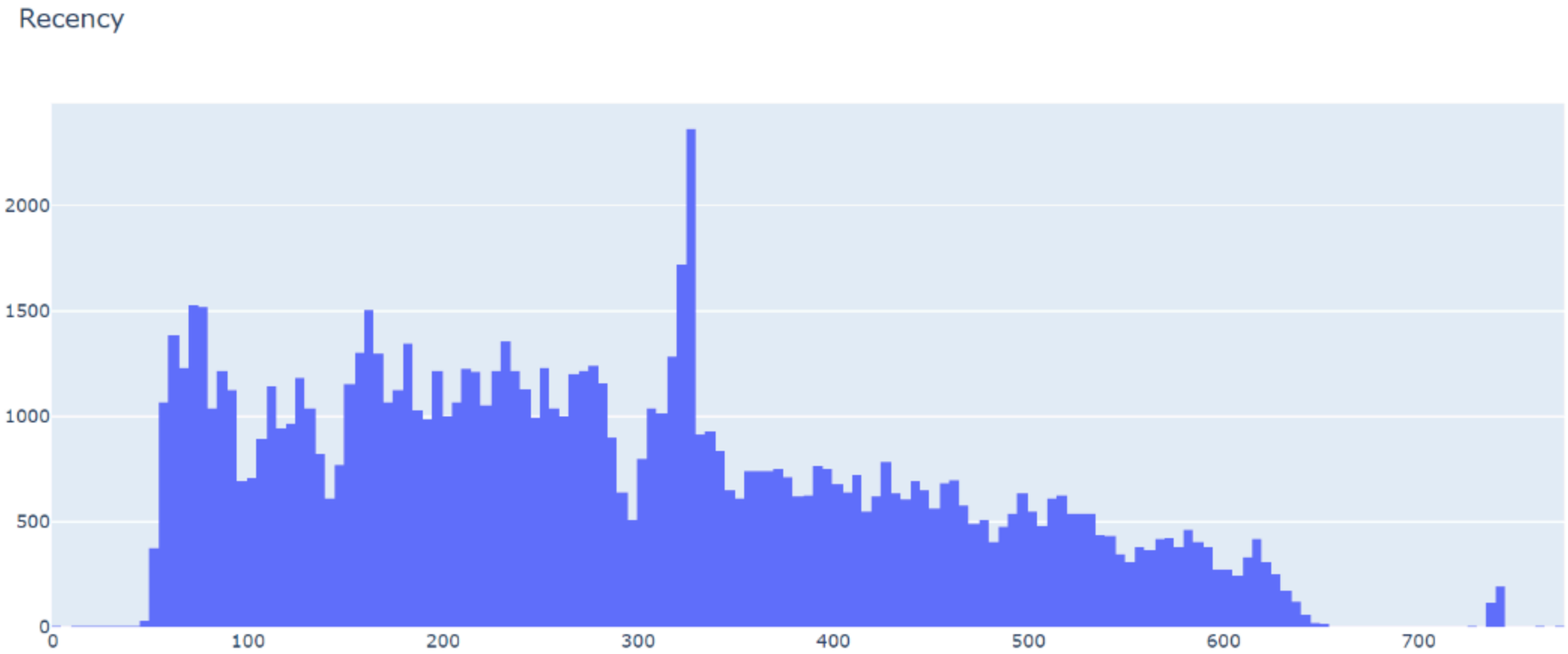
RFM

RFM stands for Recency - Frequency - Monetary Value (Revenue).

- **“Low Value”** group is less active than others and may generates very low to zero or maybe negative revenue.
- **“Mid Value”** group uses the platform fairly frequently and generates moderate revenue.
- **“High Value”** group is the center of the party because of their high revenue, frequency along with low inactivity.

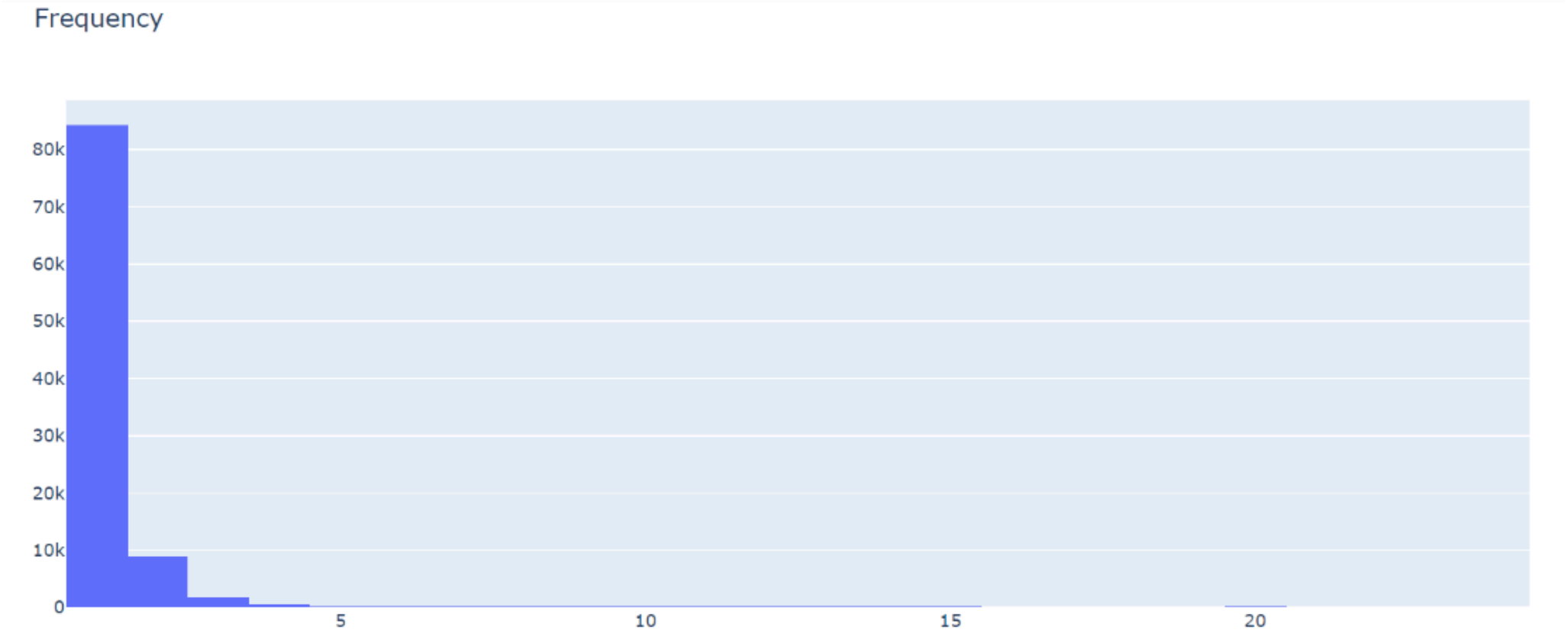
Recency

After checking how many inactive days through the most recent purchase date, applying K-means clustering assigns each client a recency score.



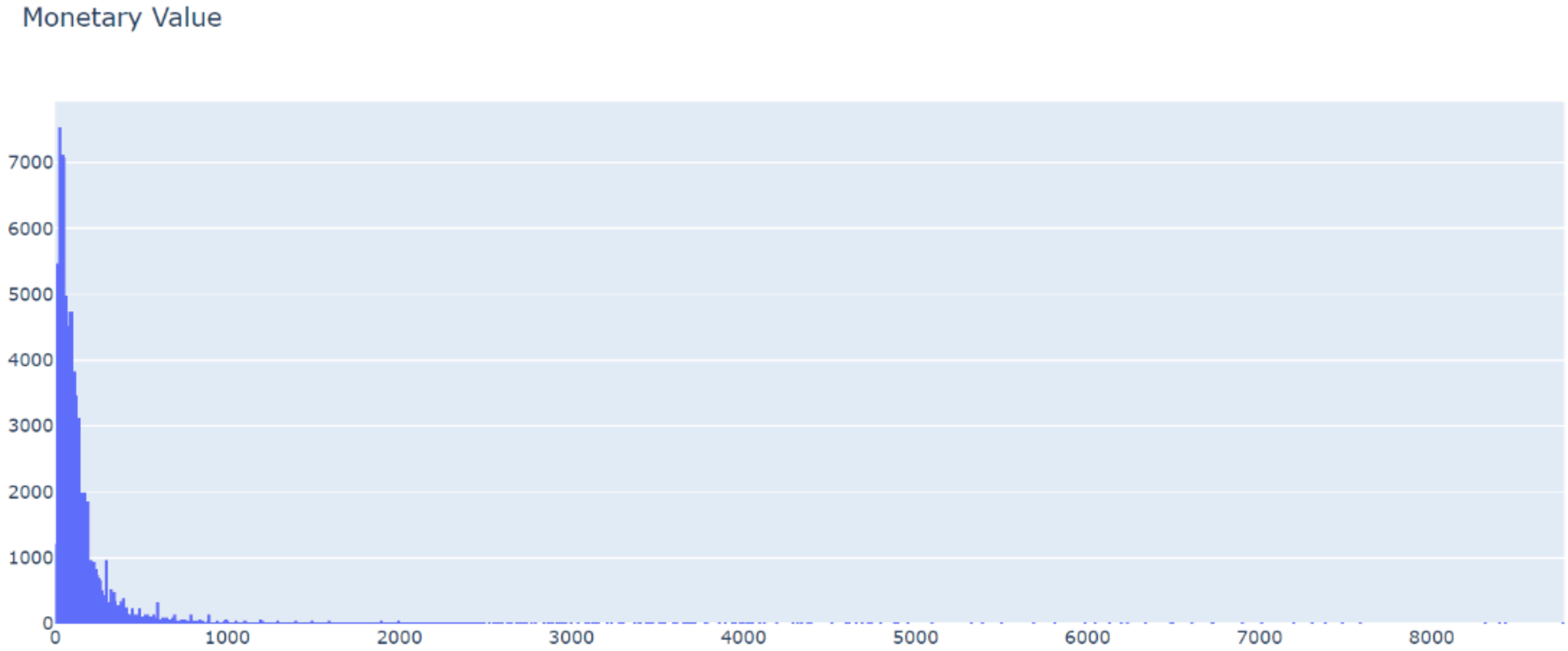
Frequency

Frequency clusters can be identified with the total number of orders for each customer.



Revenue

The plot has the number of products on x-axis and revenue in Brazilian Reals on y-axis.

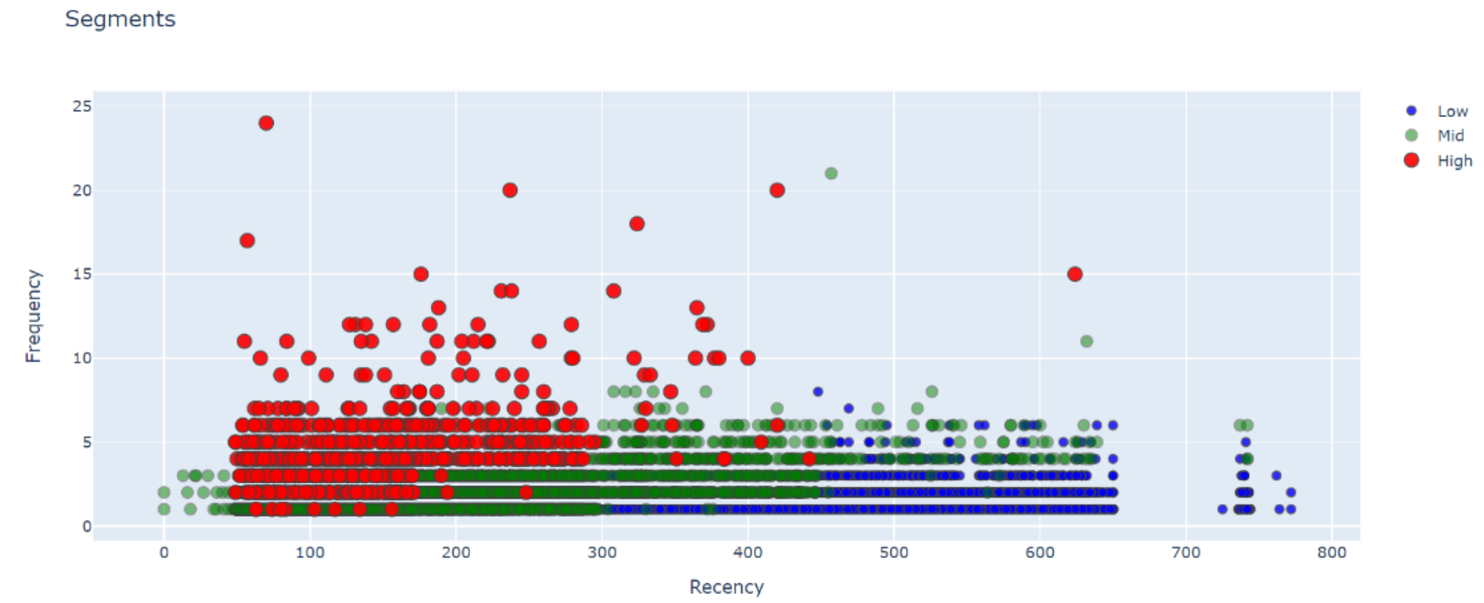


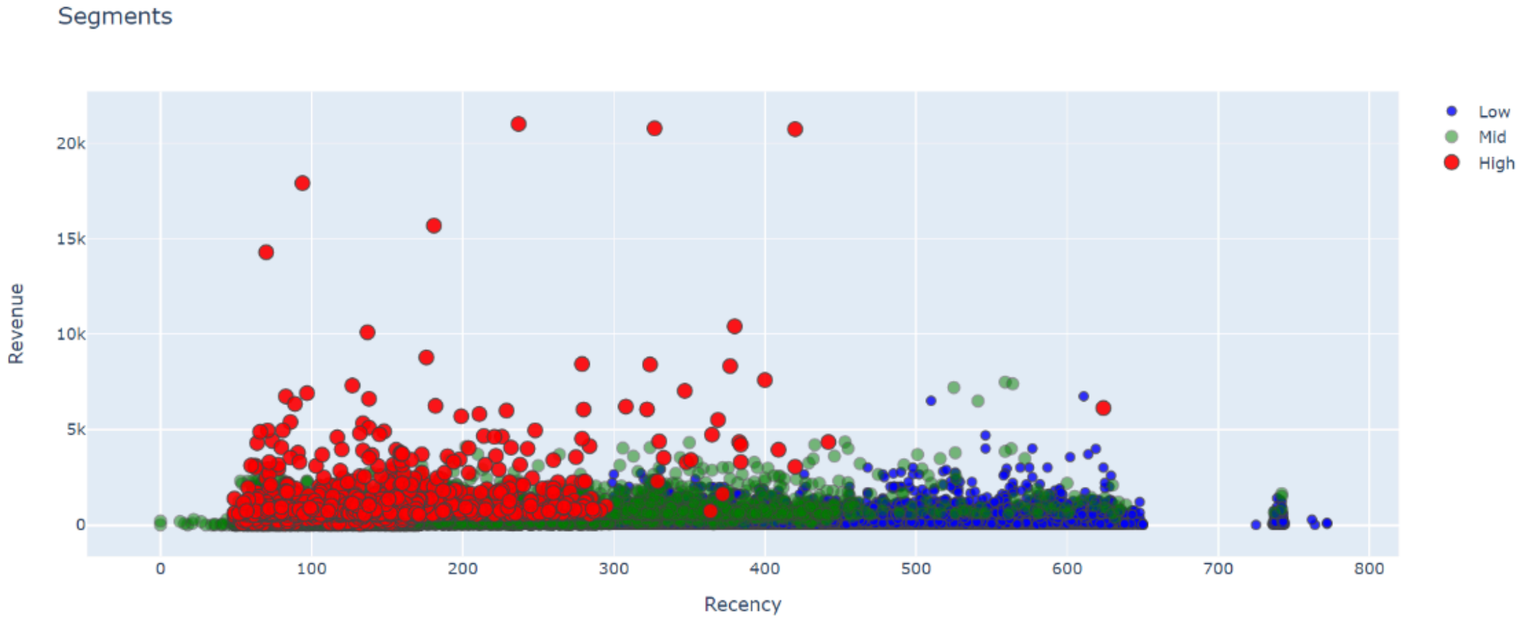
Overall Score

OverallScore	Recency	Frequency	Revenue
0	532.960070	1.000000	102.031549
1	378.206171	1.075928	127.369977
2	251.421429	1.128567	141.494018
3	132.791356	1.171811	151.887764
4	136.360532	2.138589	459.170201
5	149.474886	3.646880	962.015967
6	142.864734	5.521739	2149.913623
7	149.000000	9.204545	4734.113636
8	91.333333	14.333333	7554.533333

The scoring above clearly shows us that customers with score 8 are our best customers whereas 0 is the worst.

0 to 2: Low Value
3 to 4: Mid Value
5+: High Value





The segments are clearly differentiated from each other in terms of RFM

- **High Value:** Focus on retention
- **Mid Value:** Increase engagement and boost frequency
- **Low Value:** Reactivate interest to buff transaction frequency and value

Customer Lifetime Value

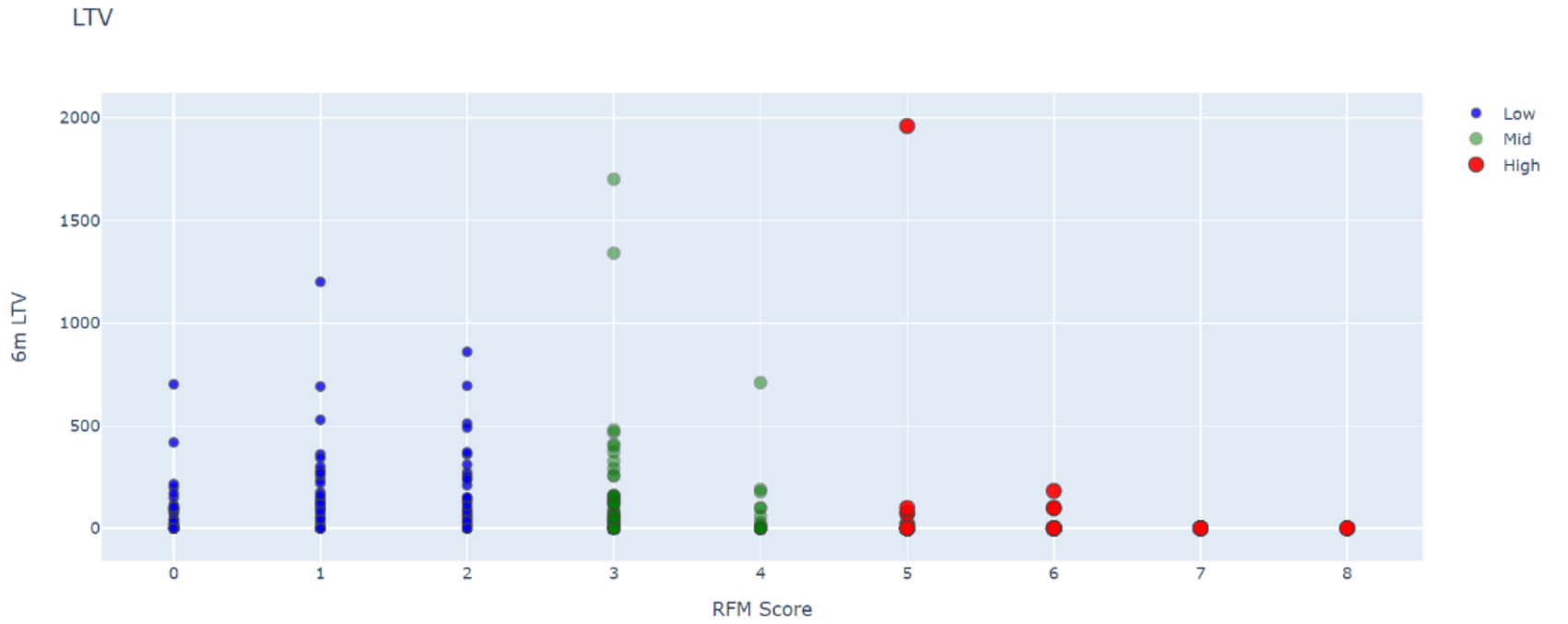
Customer Lifetime Value

To calculate Customer Lifetime Value (CLV) first we need to select a time window. It can be anything like 3, 6, 12, 24 months. By the equation below, we can have Lifetime Value for each customer in that specific time window:

Lifetime Value = Total Gross Revenue - Total Cost

To address negative lifetime values and anticipate future trends, we're implementing a ML model utilizing RFM scores derived from three months of data to predict CLV over the subsequent six months.

	customer_unique_id	Recency	RecencyCluster	Frequency	FrequencyCluster	Revenue	RevenueCluster	OverallScore	Segment
0	861eff4711a542e4b93843c6dd7febb0	15	3	1	0	124.99	0	3	Mid-Value
1	7f3a72e8f988c6e735ba118d54f47458	20	3	1	0	89.90	0	3	Mid-Value
2	2e6a42a9b5cbb0da62988694f18ee295	16	3	1	0	29.99	0	3	Mid-Value
3	fd2d5fdb84e65fa6b54b98b0e2df5645	9	3	1	0	44.90	0	3	Mid-Value
4	8728c766c84eeda24b3e54fe6e632051	1	3	1	0	117.30	0	3	Mid-Value



There is no cost specified, that's why Revenue becomes our Lifetime Value directly. After RFM scoring, the feature set looks like this.

A positive correlation is quite visible in the plot. A high RFM score means a high Lifetime Value.

LTV CLuster

Feature engineering, converted categorical columns to numerical columns. We checked the correlation of features against our label, LTV clusters and split feature set and label (LTV) as X and y.

LTVCluster	count	mean	std	min	25%	50%	75%	max
0	8452.0	0.007508	0.295046	0.00	0.0000	0.0	0.00	13.99
1	36.0	32.378611	11.175493	16.30	20.7475	34.9	39.99	49.95
2	27.0	72.509259	11.796669	56.99	59.9000	72.0	82.60	93.00

Then used XGBoost to do the classification. Since there are 3 groups, it is a multi classification model.

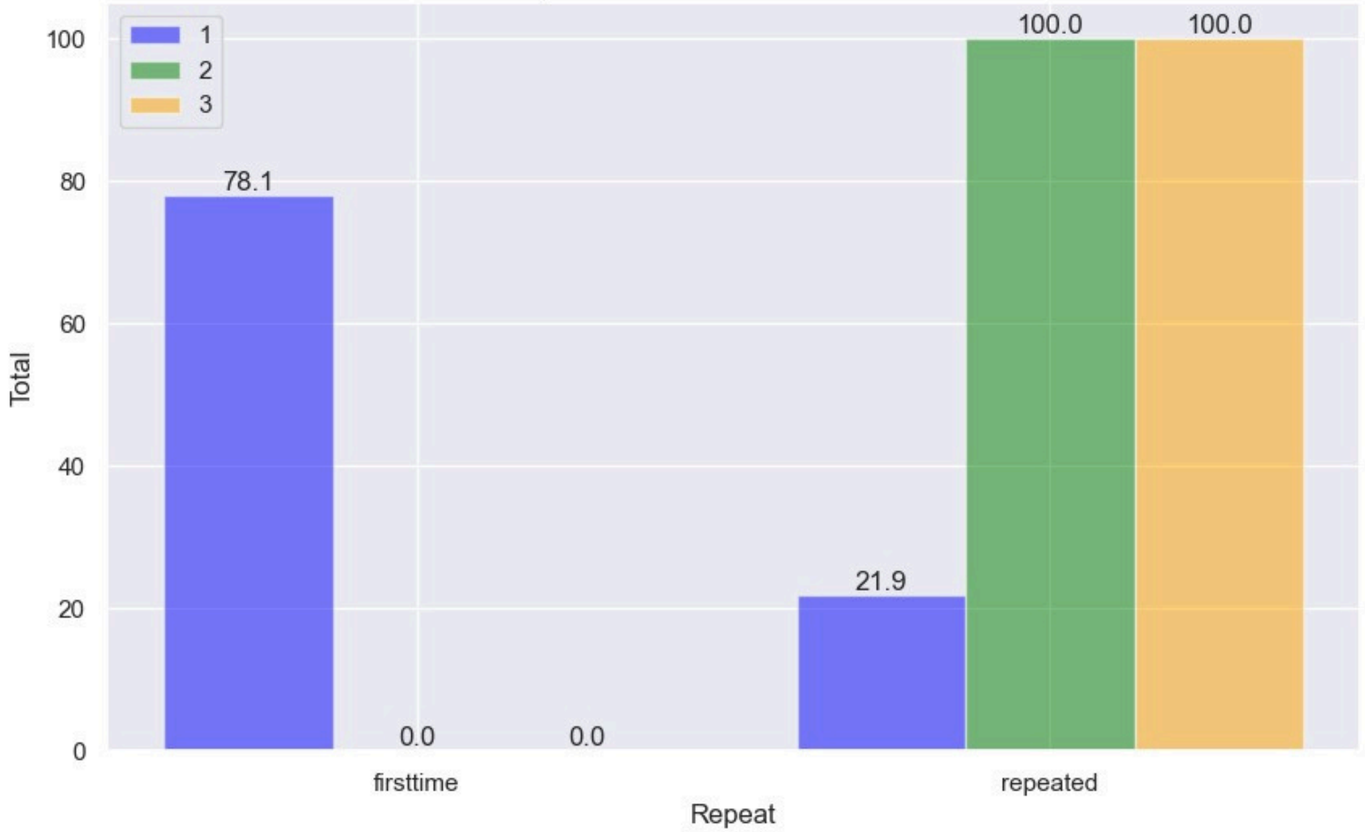
```
Accuracy of XGB classifier on training set: 0.99
Accuracy of XGB classifier on test set: 0.99
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       422
           1       0.00      0.00      0.00         3
           2       0.00      0.00      0.00         1

    accuracy                           0.99       426
   macro avg       0.33      0.33      0.33       426
weighted avg       0.98      0.99      0.99       426
```

Some future actions should be:

- Adding more features and improve feature engineering
- Try different models other than XGBoost
- Apply hyper parameter tuning to current model
- Add more data to the model if possible

RECOMMENDATION SYSTEM

First-Time vs Repeater:
HOW MANY PRODUCT PER ORDER?

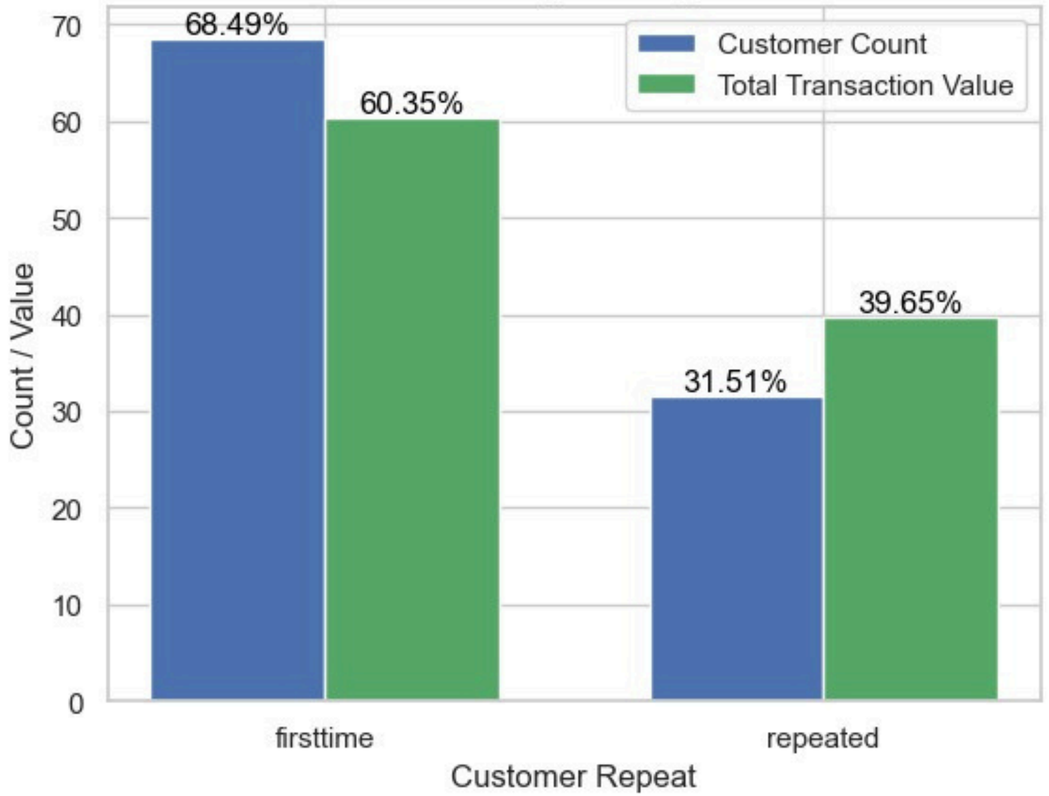


78% of First-Time Customers only buy 1 product

100% Repeat Customers buy 2-3 products

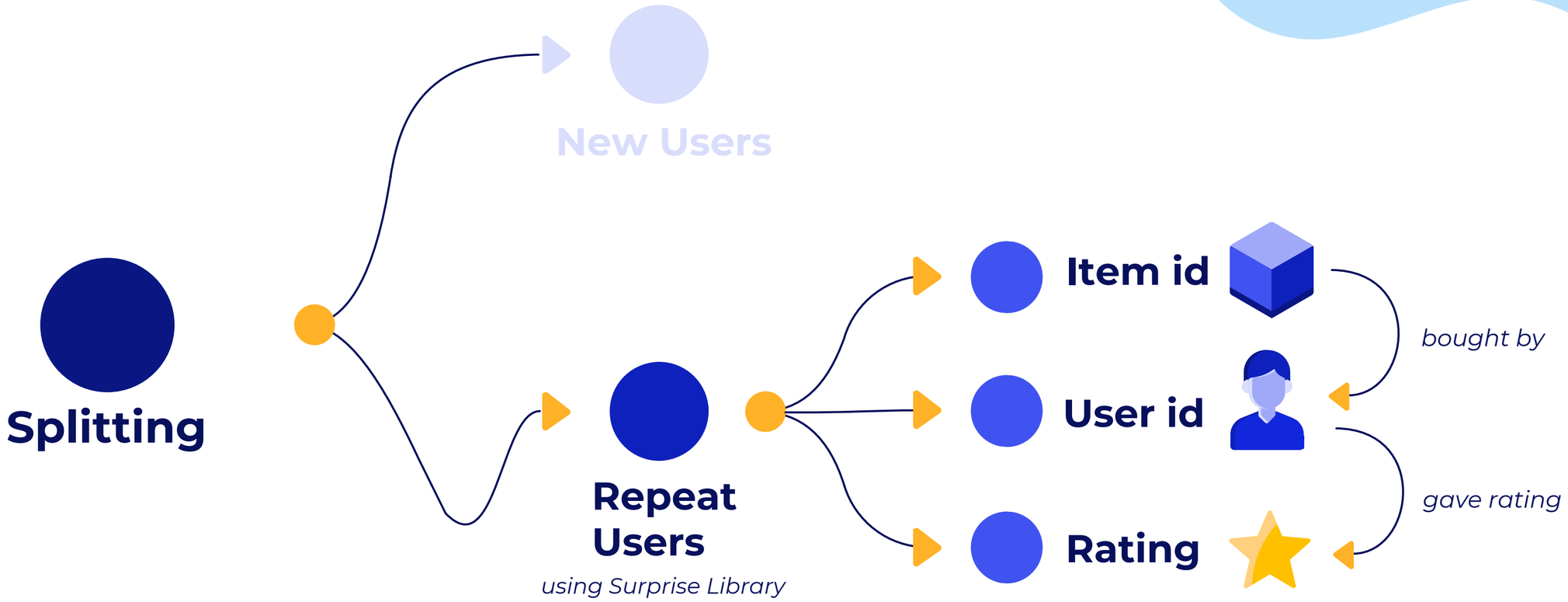
Transaction Value:
REPEAT VSFIRST-TIME CUSTOMER?

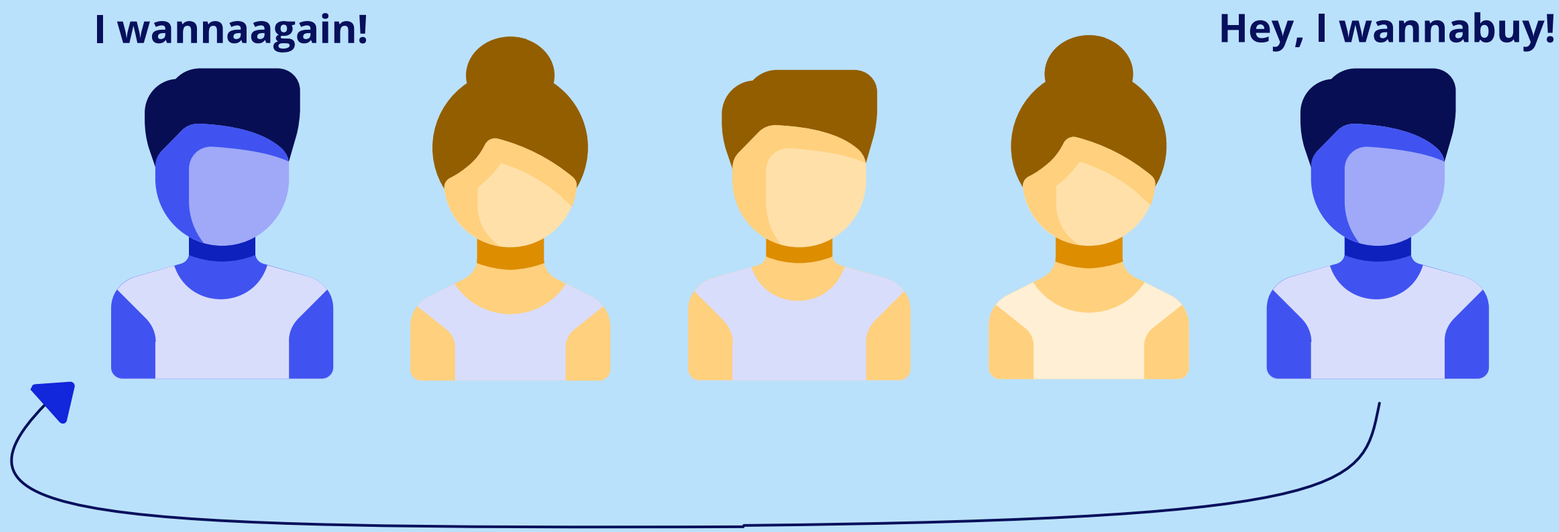
Repeat customers made repeat purchases of the product by seeing total transaction value is higher than their amount.



Hence there’re still lot of opportunities to gain more selling on **repeat customer** and **first-time customer**

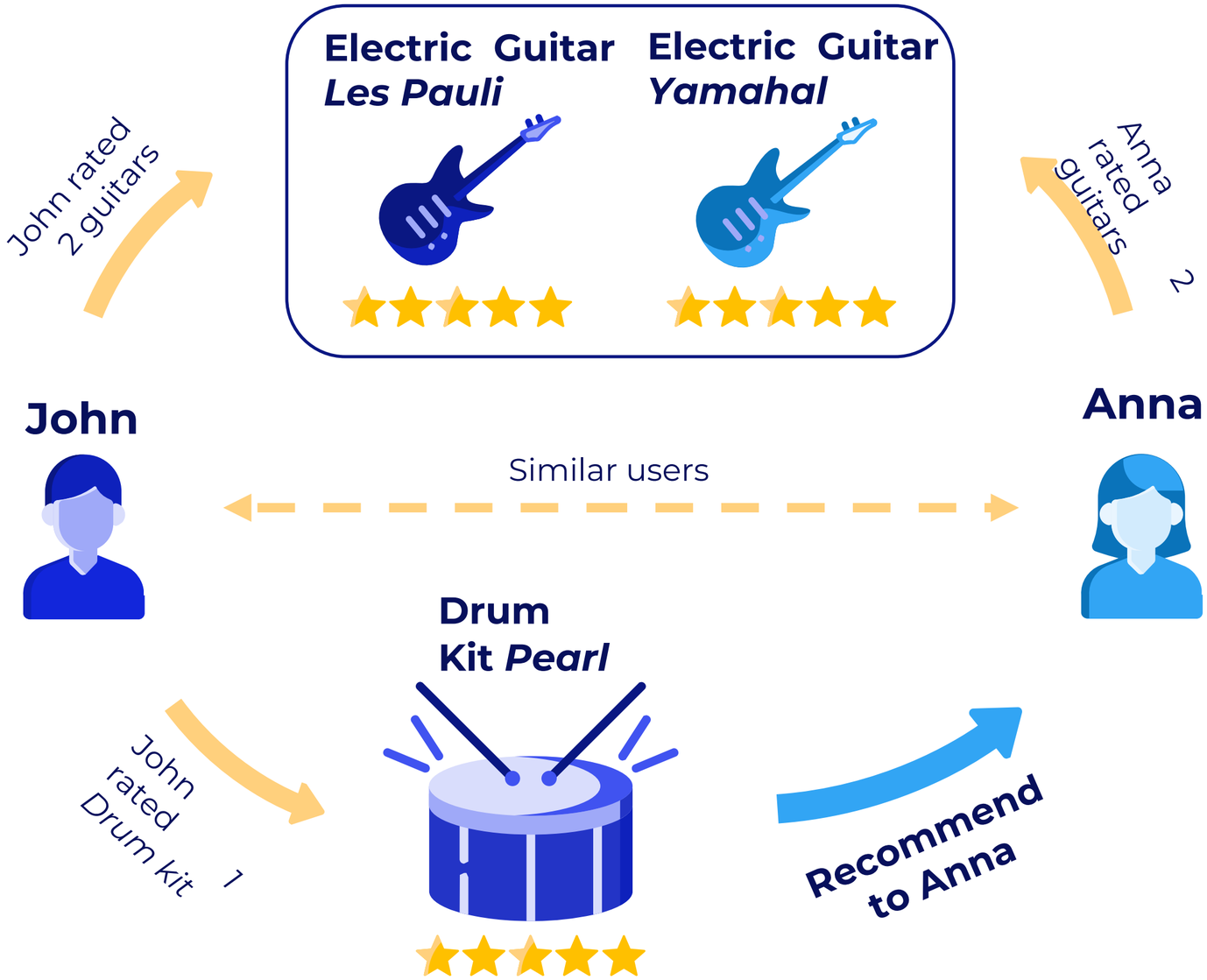
RECOMMENDER
FOR REPEAT USERS





Recommender for Repeat Customer

RECOMMENDATION SYSTEM



Collaborative Filtering models use the **collaborative power** of ratings provided by community ratings and user ratings

RECOMMENDATION SYSTEM

Collaborative Filtering -*User Based*

Jona	★★	★★★	
John	★★★★	★★★★★	★★
Anna	★★★★	?	★★

7 Algorithm Models

on Surprise Library

Normalpred

SVD

SVD++

NMF

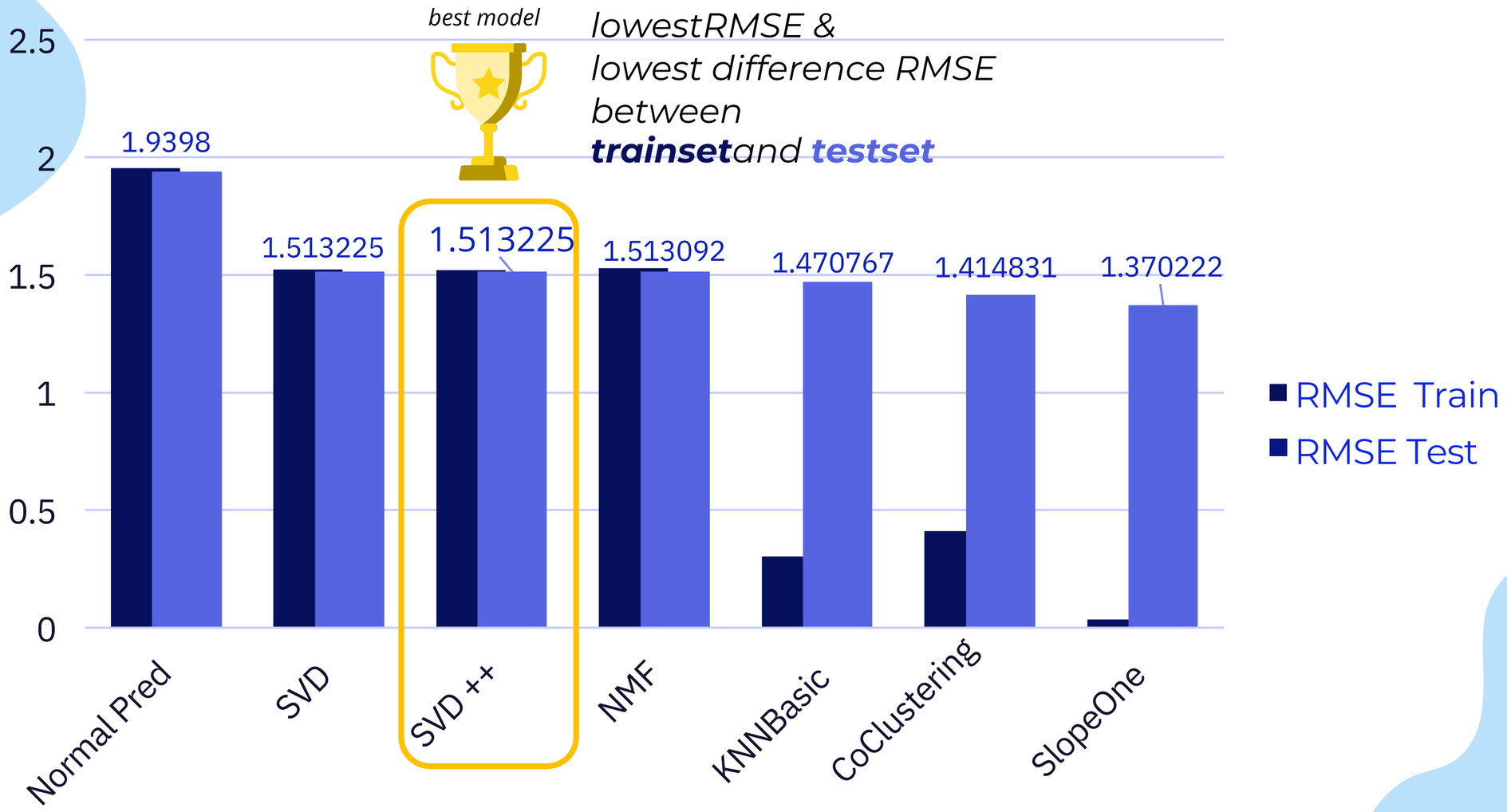
KNNBasic

Coclusterin

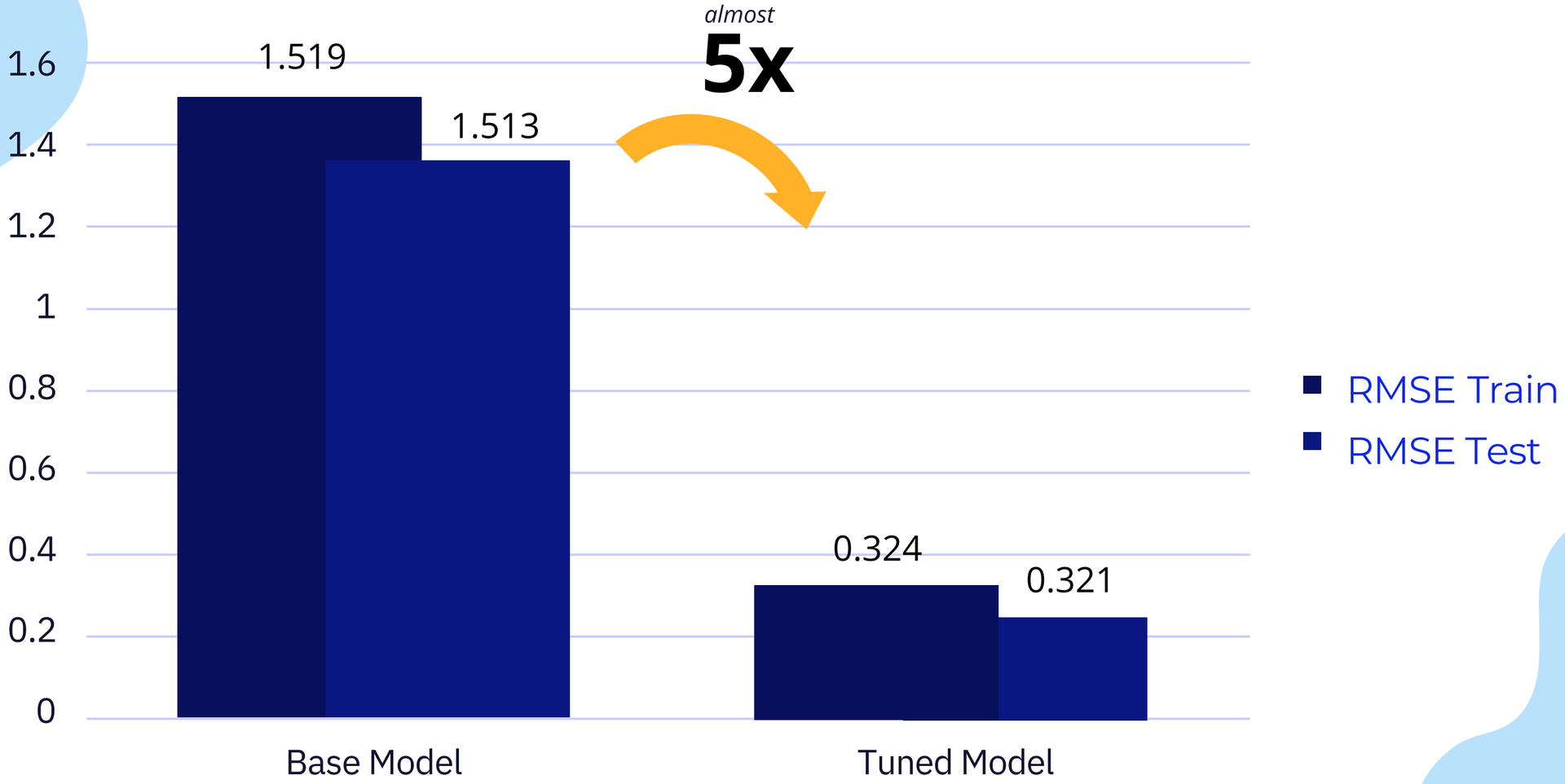
SlopeOne

FINDING
BEST MODEL

Modeling Result



Improved Model after using GridSearchCV



Testing The Model

Best Prediction

	uid	iid	rui	est	details	lu	Ui	err
3188	4f82360423c19895d2a8183e6d8ec701	bf66f5d110af02b50af37038afe90bd9	1.0	1.0	{'was_impossible': False}	2	2	0.0
100	a133f658a4d462264b8d4d8cbb282393	7c6fbb3a5346dfd607386155c6f628f8	5.0	5.0	{'was_impossible': False}	2	3	0.0
2123	b0d60f871dec79cae101d9e74c816407	65223c26538a2226610efc437e488b77	1.0	1.0	{'was_impossible': False}	2	2	0.0

User · Item · Predicted · Actual · Error

Joni bought LexungCar with **Actual Rating 1** Score otherwise the **Predicted Rating 1** Score too, Hence the **difference/error is 0**. Meanwhile this Car was rated by 2 person.

Worst Prediction

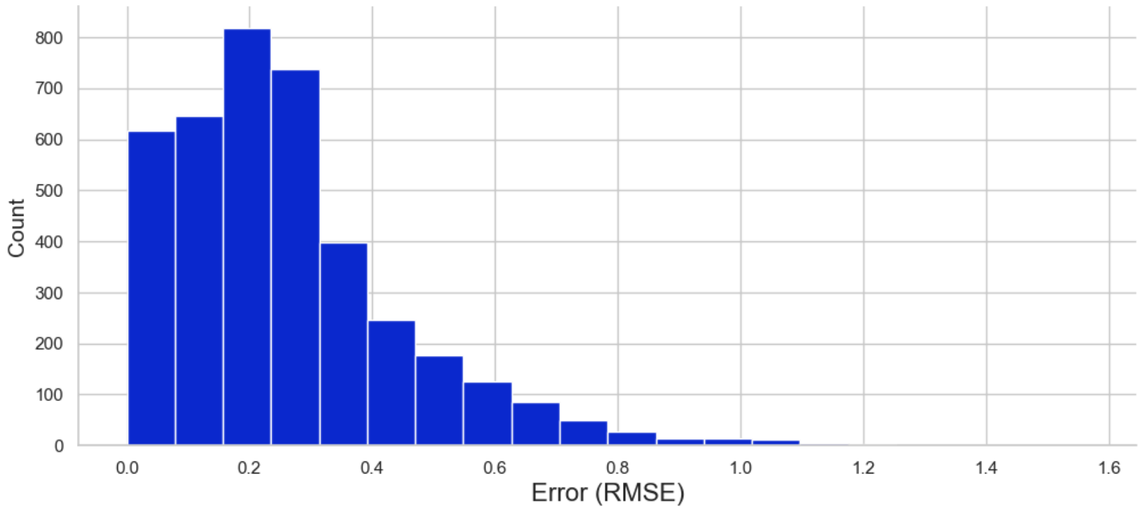
	uid	iid	rui	est	details	lu	Ui	err
2568	c49760f3652f57abae1908ae0a452350	87d780fa7d2cf3710aa02dc4ca8db985	1.0	2.076481	{'was_impossible': False}	1	11	1.076481
618	50ad7151ad494370e8dc57a57351d17e	f71973c922ccaab05514a36a8bc741b8	1.0	2.081859	{'was_impossible': False}	1	9	1.081859
260	4f88e6285f805c9a823ba23291a65ca3	f77dd338d9f75229a09cbb9a18fd0c9a	1.0	2.091208	{'was_impossible': False}	2	8	1.091208

User · Item · Actual · Predicted · Error

Kamila bought Cupid Doll with **Actual Rating 1** Score otherwise the **Predicted Rating 2.1** Score, Hence the **difference/error is 1.1**. Meanwhile this Doll was rated by 11 person.

Evaluating The Model

RMSE Distribution for All Rating



RMSE Distribution for Split Rating

The model had the **good performance** so far for **all rating score**

For all rating, the model had RMSE with skewed right distribution

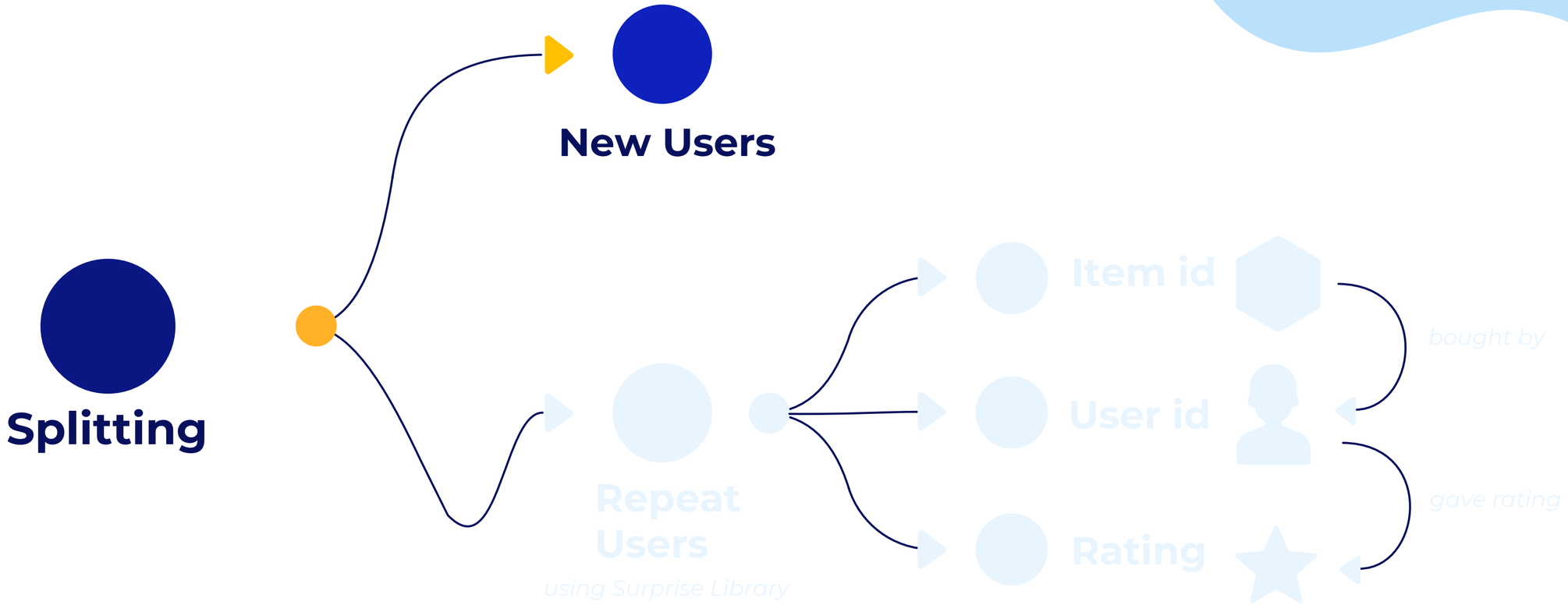
The model had the **good performance** so far for **above and below 3 rating score**

For above 3 rating score, the model had RMSE with skewed right distribution, meanwhile for below 3 rating score the model had same distribution but more long tail and slightly lower head

Hey, I wannabuy!

Recommender for First-Time Customer

RECOMMENDER
FOR NEW USERS

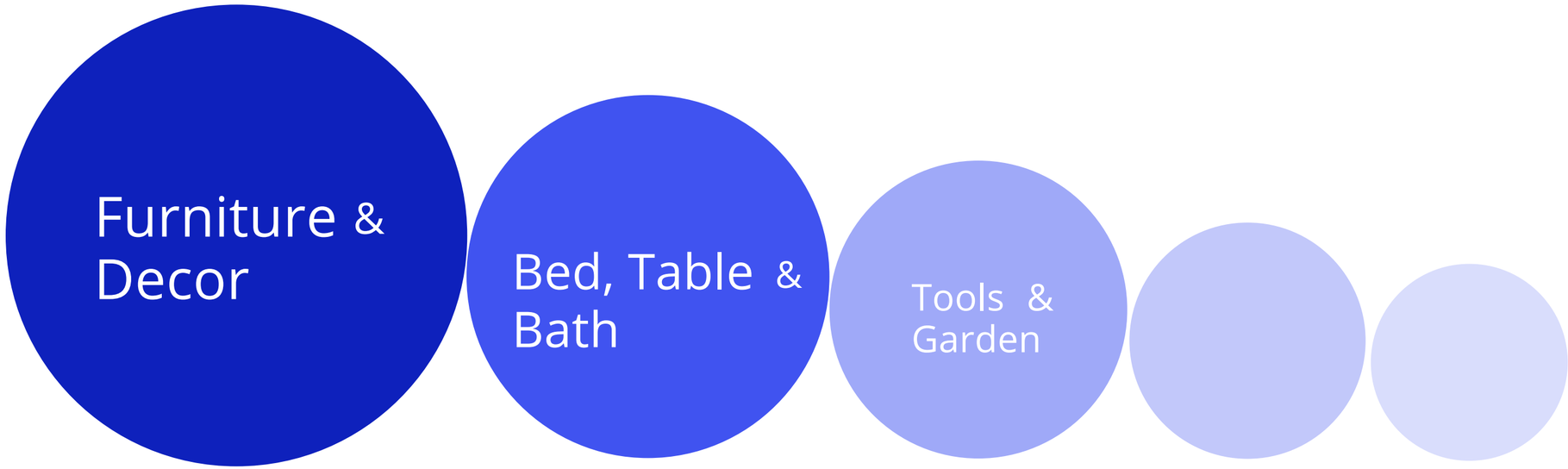


Cold-start Problem

from First-time Customer

Jona	★★	★★★	?
John	?	★★★★★	★★
New User	?	?	?

Top 3 Popular Product Category
for First Time-Customer



for all customers

Top 5 Popular Product Category
for First Time-Customer

Conclusion & Suggestions

Conclusion:
The analysis reveals three key problems. Firstly, most **of transaction value** is generated from **only 17 categories (23%),** indicating there's still **potentials** in the remaining **55 categories**. Secondly, **bad ratings still exist** in **top 17 categories**, which may impact to **customer satisfaction** and their **retention**. Thirdly, most of the clients are **first-time customer who bought 1 product** instead of repeat customer that bought **2-3 products.**

Recommendations:
1. Implement a system that leverages collaborative filtering, specifically the SVD++ algorithm to personalize product recommendations based on user-to-product data, improving the overall customer experience and driving sales.
2. Reduce bad ratings by identifying the underlying causes, from product quality, delivery, and customer support, to other factors affecting customer satisfaction.
3. Since there are still a lot more first-time customers than repeat customers, the platform must overcome the cold-start problem for first-time customers.

THANK YOU!