CUSTOMER PREDICTIVE ANALYSIS

On Brazilian E-Commerce Dataset

By Nicholas Satyahadi

Presentation Overview

DATA EXPLORATION

MODEL DEVELOPMENT

PREDICTION ANALYSIS & DECISION MAKING

Introduction

- The Brazilian E-commerce Dataset will serve as an example on how one would create a machine learning model to conduct predictive analysis.
- **Focus of the model**: Predict whether if the customer will return after they made their purchase(s) and received it.

Data Exploration

- Data Cleaning:
  - The **product category name will be translated into English** due to language barrier.
  - **Products with NULL descriptions (e.g. name, length, weight) had been removed** in order to prevent errors.

- Data Standardization:
  - The order items dataset had been summarized to show how many items purchased per order instead of using order_item_id.

Data Exploration

Order Reviews

Not all order have reviews, and one order might have multiple reviews for multiple products. An order id is used as a perfect example of the conditions as shown below. Nevertheless, the review is still valid since not all customers will leave comments.

	review_id	order_id	review_score	review_comment_title	review_comment_message	review_creat
22585	2a74b0559eb58fc1ff842ecc999594cb	0035246a40f520710769010f752e7507	5	NaN	Estou acostumada a comprar produtos pelo barat...	2017-08-25
25802	89a02c45c340aeeb1354a24e7d4b2c1e	0035246a40f520710769010f752e7507	5	NaN	NaN	2017-08-29

Data Exploration

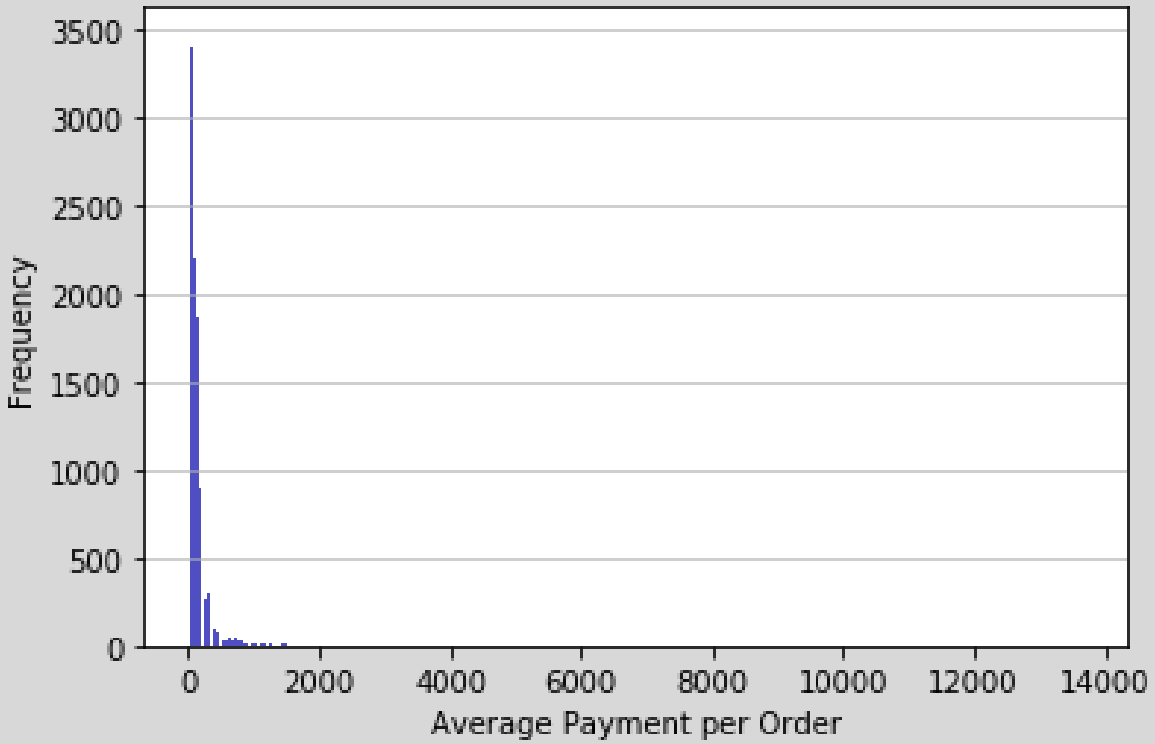
Order Status

- These are order statuses that the system will generate in order:
  1. **Created**
  2. Invoiced
  3. Approved
  4. Processing
  5. Shipped
  6. **Delivered**
- Since we want to understand the customers' purchasing behavior, it's essential for us to only observe valid delivered orders, meaning the orders have complete and correct (no NULLs) records of delivery.

Data Exploration

Payment Amounts

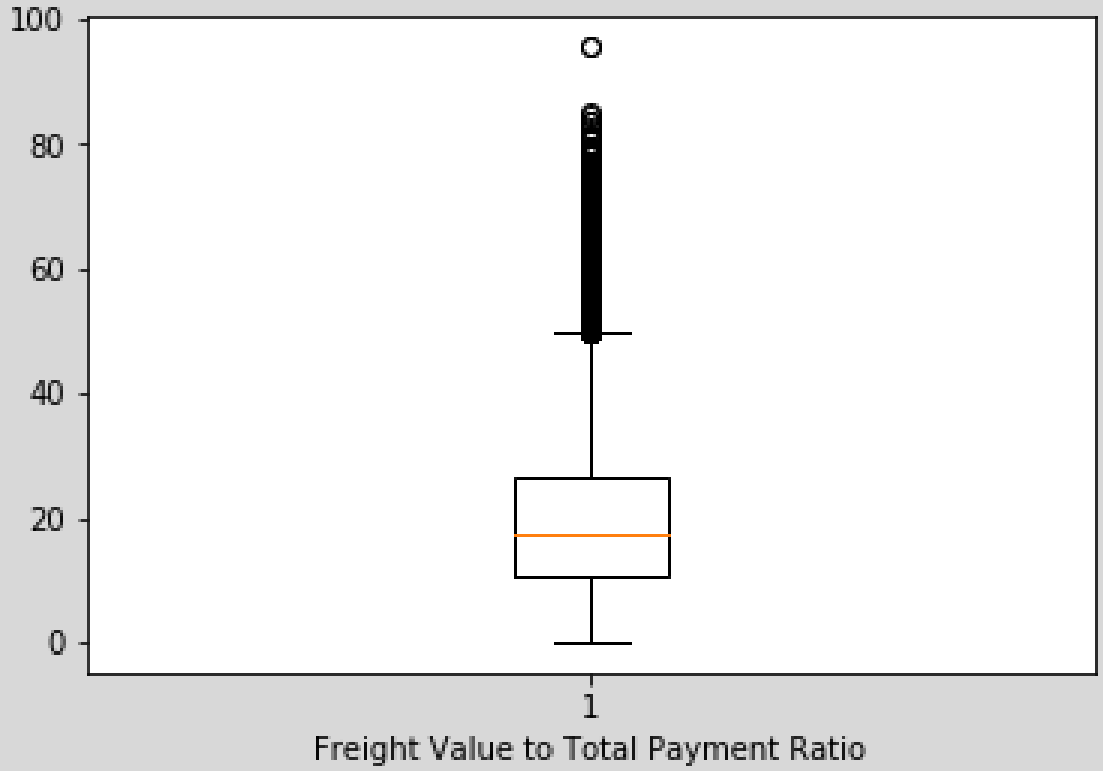
- Majority of the payment values below 2000 with the 3rd quantile around below 200 (exact number: 171.13).
- Less than 5% of orders have payment amount above 500.



Data Exploration

Freight Value to Payment Ratio

- Most of the orders have freight value to total payment ratio below 50%.
- It's appropriate for customers prefer to make a purchase when the freight value doesn't cover majority of the payment.
- Even so, it doesn't eliminate the possibility that there are some customers willing to pay such high freight value.



Model Development

The return of the customer relates on how "loyal" the customer is to the store. Hence, here are some variables that would affect the customer's loyalty:

Factors	Explanation
Order review	High reviews indicate trust from customer to the store
Average payment per order	High payment along with high review increases trust towards the store and indicates a high satisfactory purchase
Freight value to total payment ratio	Freight value relates to the distance between the seller and the customer, also the weight of the item. If the freight value covers most of the payment, it is less likely for the customer to return to make another purchase.

Model Development

- The dataset used in the model will be labeled into will return (1) or won't return (0) according to the variables or factors explained previously.
- Logistic Regression and Random Forest classification model will be used and compared to predict.

Model Development

Labeling process

- The dataset will be labeled into will return (1) and won't return (0) according to the records in each distinct customer id and order id. Then, the dataset will be divided into train and test dataset.
- There will be a condition switching for some records, picked randomly. This switching is conducted in order to create an unlikely condition, e.g. 5 Stars review but labeled as won't return.
- The labeling rules or value thresholds decided according to the findings explained in data exploration sections*.

*attached in the appendix of this slide.

Prediction Result

Returning Customers

- The prediction predicted 82.09% (19,488 customers) of customers from the test dataset will return for future purchase(s).
- The average freight ratio is 19.01%
- The table on the right summarizes the top 3 categories based on the number of items purchased.

Category	Number of Items Purchased	Average Payment per Order	Number of Vouchers Used
Bed, Bath, & Table	2,145	222,898	59
Health & Beauty	1,989	306,302	46
Sports & Leisure	1,739	222,390	31

Prediction Result

Non-returning Customers

- The prediction predicted 17.91% (4,251 customers) of customers from the test dataset will not return for future purchase(s).
- The average freight ratio is 23.03%
- Despite being predicted to not return, 11.88% (505 orders) of the orders have 5 star review.
- The table on the right summarizes the top 3 categories based on the number of items purchased with 5 star reviews.

Category	Number of Items Purchased	Average Payment per Order	Number of Vouchers Used
Electronics	62	2,202	0
Telephony	51	2,459	2
Health & Beauty	41	3,927	0

Decision Making

- **Promotions**:
  - **Cashbacks with minimum payment** would increase the basket size (number of items purchased) along with maintaining the total payment per order.
  - **Discounts** would reduce with the average payment per order but could effectively attract non-returning customers to return. The discounts could be in embedded in the product itself or using a certain payment method, preferably credit cards since most of the payments are through credit cards.
  - **Freight value discounts with minimum payment** would also attract non-returning customers to make a purchase since the average freight ratio from the non-returning customers are higher than the returning customers.
- **Additional inputs:**
  - Health & beauty products seems to be the main driver since it's included in both returning and non-returning customers top 3 categories. Maintaining the average payment per order would be recommended.
  - Raising public awareness about the abundant amount of vouchers that are able to be utilized is recommended to attract non-returning customers and new customers.

APPENDIX

Labeling Rules

Review Score	Payment	Freight Ratio	Labeled as
< 3	-	-	0
= 3	< 200	> 50%	0
= 3	< 200	< 50%	1
= 3	>= 200	-	0
> 3	-	> 50%	0
> 3	-	< 50%	1