Problem Identification for **Customer Lifetime Value**

P

Context

- Revenue growth is great to see in a young business
- sustaining that growth depends on how well a company retains customers as repeat buyers.
- Many company can spend money on marketing to convince people to try their product
- if those people don't like the product or the brand enough, it just requires more marketing money to acquire more customers in their place
- A valuable company will sell a product and brand that the begets customer loyalty and repeat business.

Specifics

- Customer and sales data orders from 2013 to September 2020 with 142,000 orders in over 200,000 rows and 72 columns
- Much of this data may not be useful features
- May have issues with the curse of dimensionality with so much data
- Stuck with data already acquired; cannot measure additional features actively (if we determine more would be useful)

Focus

- Customer and sales data orders from 2013 to September 2020 with 142,000 orders in over 200,000 rows and 72 columns
- Stuck with data already acquired; cannot measure additional features actively (if we determine more would be useful)

We turn to modeling for the solution….

Problem Statement for Customer Lifetime Value

P

How can we determine customer lifetime purchasing patterns at an online retail company: A. by exploring data for trends in customer retention, repeat rate, and churn B. by modeling Customer Lifetime Value (CLV)

Objectives

- Explore Customer Sales data from Shopify for an online retailer
- Model Customer Lifetime Value (CLV) based on available features
- Determine best models for CLV

Focus

- First purchase data by customers (later purchases maybe too late in the process - they already are repeat)
- Consider trends of when their first purchase is and the size of the order (money and number of items)

Constraints

- Customer data from Shopify in 5 CSV file with over 200,000 rows and 72 columns
- Must anonymize data so customer and company information are protected
- Cannot run A/B or other tests on customers
- Known that marketing surge occurred after August 2019; data should reflect that increase in customers after that date

Key Findings for Customer Lifetime Value

K

We found that Customer Lifetime Value, Customer Retention, and Repeat rate are affected by a number of factors: day of the week, month, first purchase value. We produced an accurate CLV model with Linear Regression and Random Forest Regressor.

Model Accuracy

- Root Mean Squared Error for our model (20% holdout test): 0.263
- Root Mean Squared Error for dummy model: 0.380
- This is on exponential scale (~$20 on avg or $1000 on big spenders)

Models used

- Final Model: Linear Regression with Random Forest on Residuals
- Also tried:
  - Gradient Boosting Regression
  - XG Boost Regression
  - Random Forest
  - Linear Regression

Other Findings

- Customer Lifetime Value and number of purchases are exponentially distributed
- Customers acquired on Monday had statistically significant higher CLV than customers acquired on Sunday
- Customers acquired in November had statistically significant higher CLV than customers acquired in September
- Wednesday and midweek days have higher repeat rates than weekends
- September had a much higher repeat rate than November

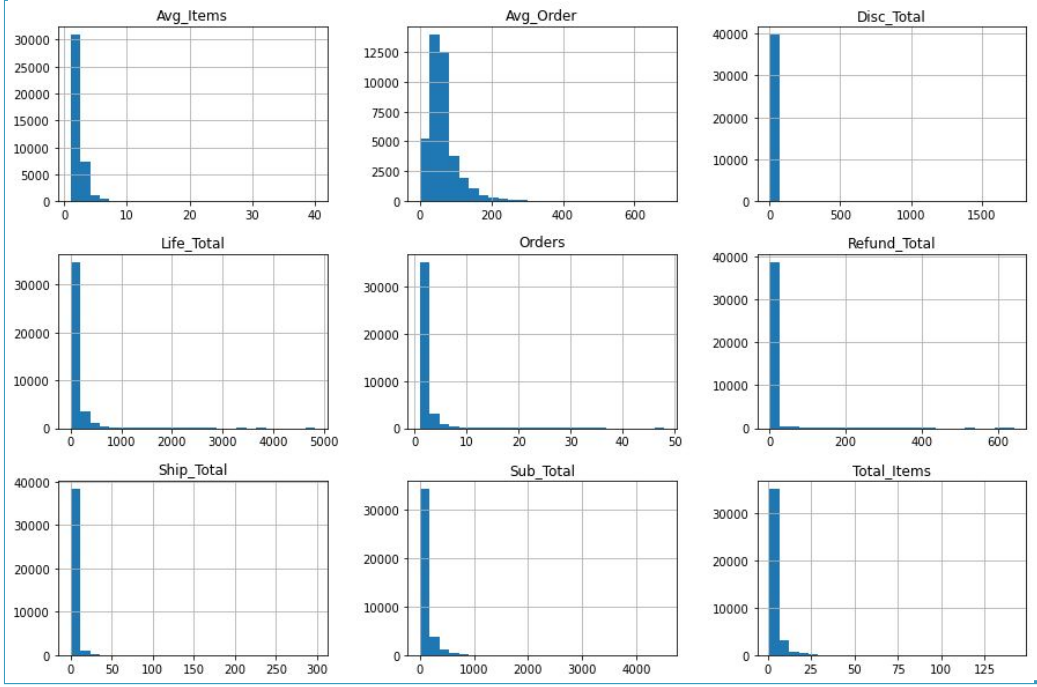
Exploratory Data Analysis for Customer Lifetime Value

K

Before we modeled the Customer Lifetime Value, we explored the data to find potential correlation and nuances

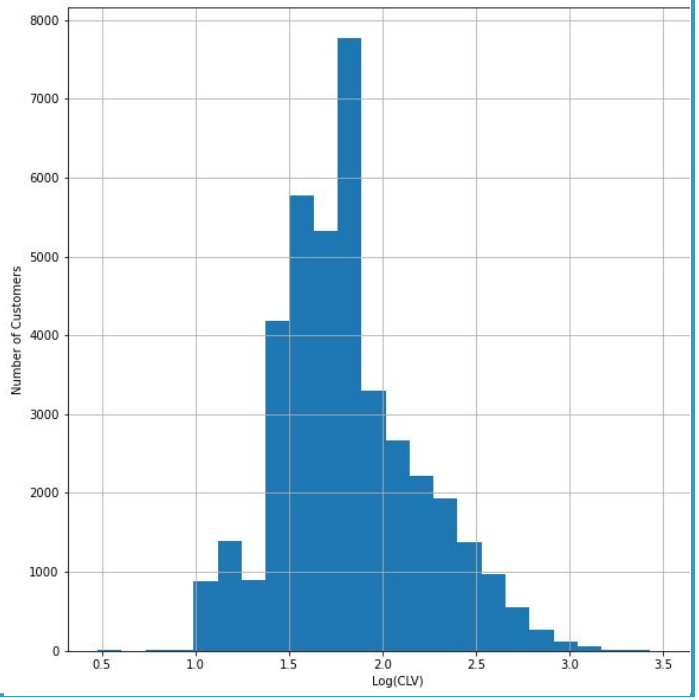
Exponential Distribution

- **We first looked at the distribution of many variables in the dataset**
- **Majority in early part of distribution with long tail = Exponential**



Log10 to Normalize

- **To convert this data to a better distribution to model, we converted by taking Log10**
- **Still skewed but close to Normal**



P K M R S