Problem Identification for **Customer Lifetime Value**

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

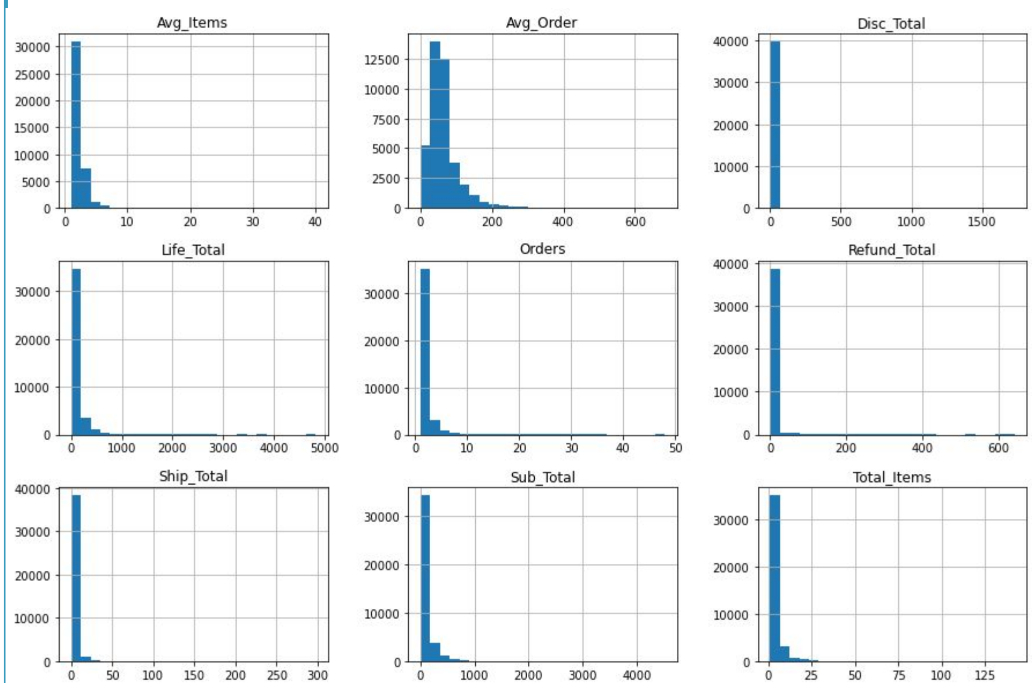
Before we modeled the Customer Lifetime Value, we explored the data to find potential correlation and nuances

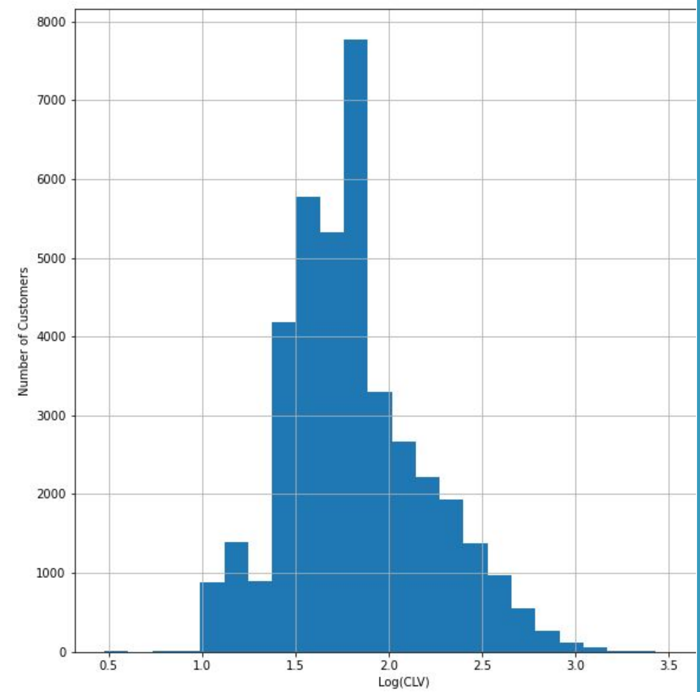
Exponential Distribution

- **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**





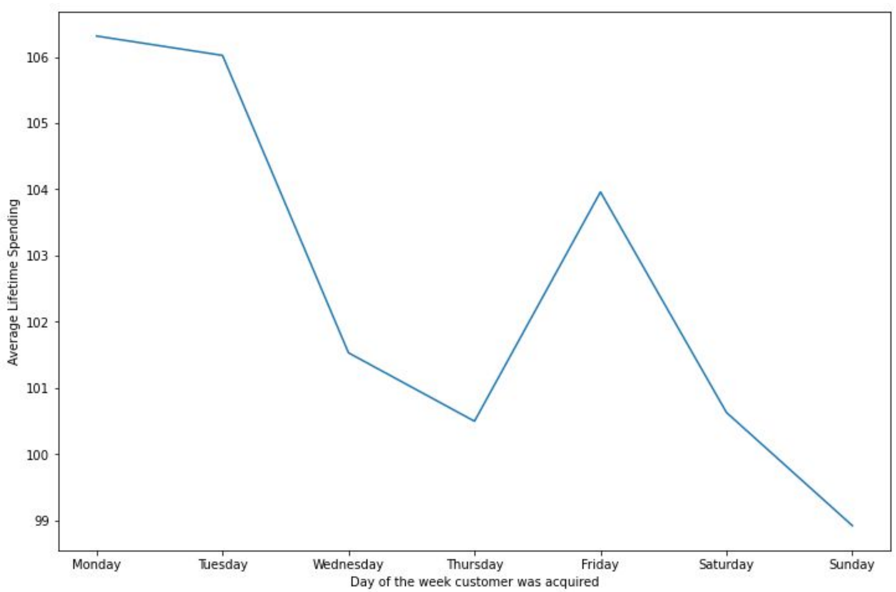
Exploratory Data Analysis for Acquiring Customers

K

In exploring Customer Lifetime Value with respect, customers acquired (first purchase) at some times had higher CLV

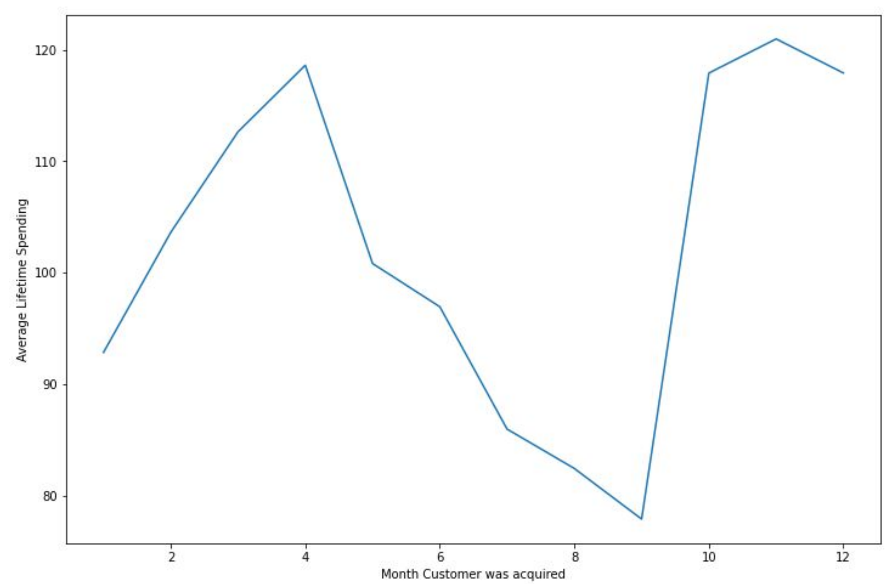
Mondays better than Sundays

- **Customers acquired on Monday (first purchase) had higher lifetime value (\$106 vs \$99)**
- **p-value = 0.007 that Monday average is higher than Sunday (99.3%)**



November better than September

- **More drastic than weekday difference (\$120 vs \$78)**
- **p-value = 6.05×10^{-12} that November average is higher than September**
- **Could have attracted low-quality customers in marketing boost**



P K M R S

Exploratory Data Analysis for Repeat Purchase

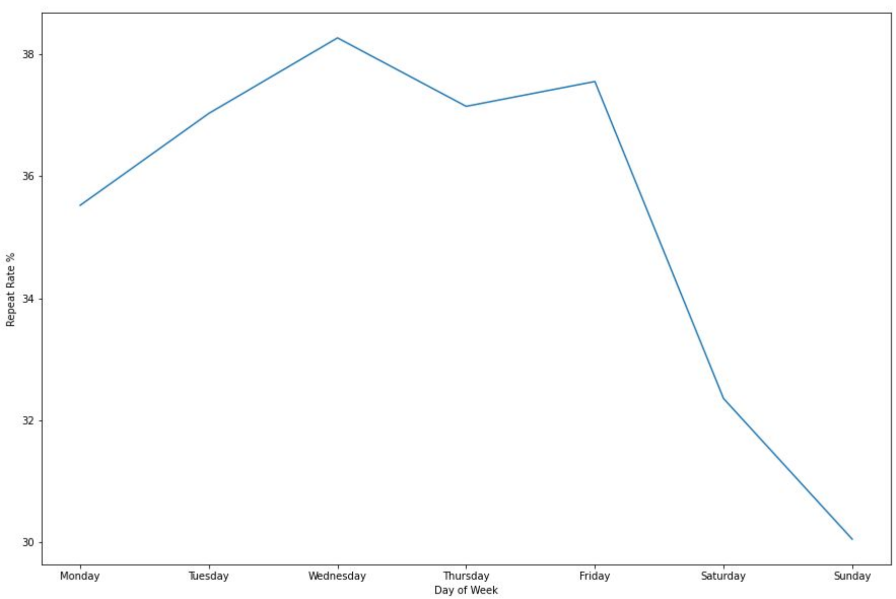
After customers are acquired, when do repeat purchases account for more of the sales?

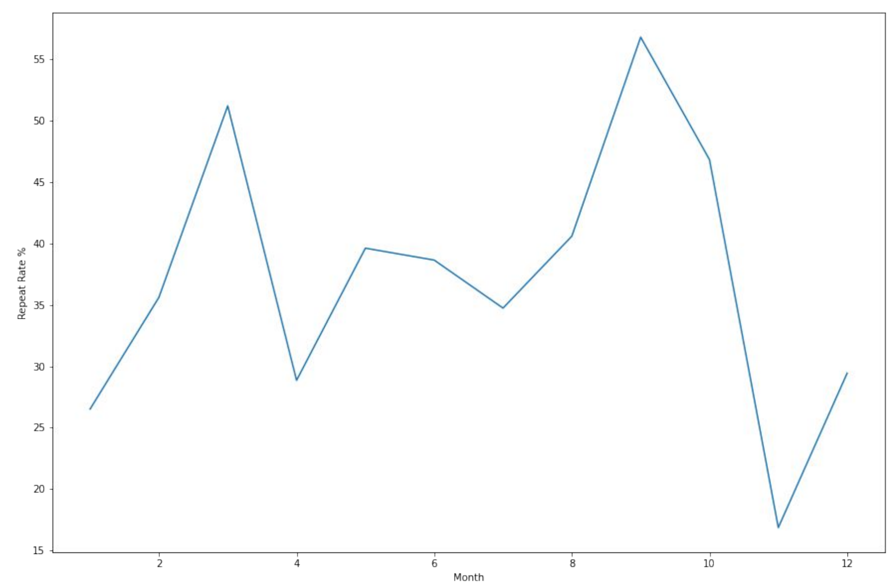
Wednesday better than Sunday

- **Higher repeat rate of purchase on Wednesday: 38% vs. 30%**
- **Tuesday, Thursday, and Friday are also strong repeat days**

September has more repeat business than November

- **More drastic difference (56% vs 16%)**
- **Could be slower new customer acquisition in September**
- **Could just be the marketing approach at these times**





Pre-processing for Customer Lifetime Value

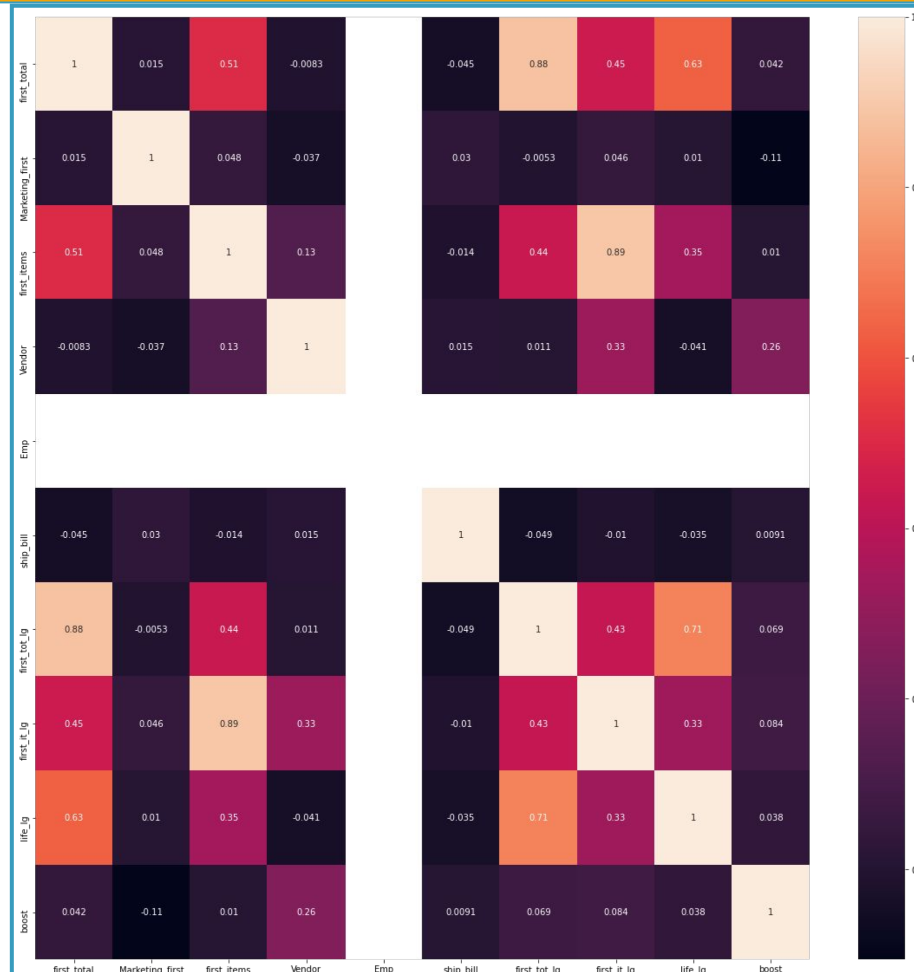
M

We had a lot of rows of data available to model customer lifetime value, so we had to drop a lot of irrelevant data

heatmap of correlation coefficients

Removed many variables

- **Removed variables that were not known at the time of the first purchase (total orders, total items, etc.)**
- **Removed variables that had little or no correlation to CLV**
- **Removed variables that were too closely correlated to each other**
- **Created dummy variables for categorical features: month, weekday, email domain, lead SKU**
- **Filled missing values with the most relevant data**



Modeling for Customer Lifetime Value

M

Built multiple models for customer lifetime values using Linear Regression, Random Forest Regression, and Gradient Boost Regression. Had to eliminate variable further for linear regression.

Initial model results

- **On the first round of modeling, linear regression was extremely off - likely from too many features.**
- **We dropped over 40 variables from consideration and linear regression went from horrible to the best model**
- **All of those top 5 models performed very well**
- **improvement over dummy is ~$20 near average CLV; ~$1000 near extreme end**

Customer Lifetime Value

Model	MAE	MSE	RMSE
Linear Regression 2	1.859726e-01	6.770457e-02	2.602010e-01
Random Forest 1	1.847552e-01	6.807910e-02	2.609197e-01
Gradient Boost 1	1.847552e-01	6.807910e-02	2.609197e-01
Random Forest 2	1.847164e-01	6.809799e-02	2.609559e-01
Gradient Boost 2	1.847164e-01	6.809799e-02	2.609559e-01
Dummy (Average)	2.906511e-01	1.445166e-01	3.801533e-01
XG Boost	4.628376e-01	3.046849e-01	5.519827e-01
Linear Regression 1	1.041380e+10	4.308612e+23	6.564002e+11

Models were built to predict log10 of CLV, so values and errors need to be used as exponent to calculate CLV: 10^model value

MAE = Mean Absolute Error; MSE = Mean Square Error

Model Tuning for Customer Lifetime Value

M

If a few models is good, more is better!
We tuned the hyperparameters on the models we used to squeeze out the best performance.

Final model results

- **We tuned the hyperparameters for Gradient Boost and Random Forest; both did not improve the models**
- **We built an Ensemble model using the top 2 models: Linear Regression and Random Forest on the residuals**
- **This stacked ensemble model performed the best.**
- **Resulted in RMSE 0.122 and MAE 0.108 better than dummy (average model)**

Model	MAE	MSE	RMSE
Ensemble: LR2 + RF	1.821710e-01	6.680951e-02	2.584753e-01
Linear Regression 2	1.859726e-01	6.770457e-02	2.602010e-01
RF Tune 1	1.839021e-01	6.798485e-02	2.607391e-01
Random Forest 1	1.847552e-01	6.807910e-02	2.609197e-01
Gradient Boost 1	1.847552e-01	6.807910e-02	2.609197e-01
Random Forest 2	1.847164e-01	6.809799e-02	2.609559e-01
Gradient Boost 2	1.847164e-01	6.809799e-02	2.609559e-01
RF Hypertune	1.847757e-01	6.767430e-02	2.609559e-01
GB Hypertune	1.832013e-01	6.889463e-02	2.624779e-01
Dummy (Average)	2.906511e-01	1.445166e-01	3.801533e-01
XG Boost	4.628376e-01	3.046849e-01	5.519827e-01
Linear Regression 1	1.041380e+10	4.308612e+23	6.564002e+11

Recommendations for Customer Lifetime Value

R

What should we do with these results?
What action can we take to improve the model?

Actions for Sales

- **Put more resources into acquiring customers on Monday and in November**
- **Market for repeat customers midweek (Tuesday - Friday) and in April and September**
- **Getting that second order is harder than getting order beyond that; retention goes up with order numbers**

Other potential investigations

It would be great to isolate time frames and test these values again (maybe eliminate Christmas season) and see if these conclusions still hold

Improved Modeling Approach

- **Include all SKU's from the first order**
- **Try Ridge or Lasso Regression to limit parameter values in model**
- **Model Future CLV (subtract out first order)**
- **Include time of day in models**
- **Try different ensemble models and Neural Net**
- **Focus model on identifying customers that we really want to retain**
- **Box-Cox Transformation on data**
- **Use log10 or Box-Cox transformation on more features**
- **Model for WHEN the second purchase occurs for repeating customers**

Summary and Conclusion for Customer Lifetime Value

S

- **Explored customer lifetime value (CLV), retention, and repeat**
- **Constructed a dozen models for CLV**
- **Models performed better than dummy (average)**
- **Ideas for future models and explorations**

Data used in Model Building

- **Acquiring customers that stay: Monday >> Sunday; Nov >> Sept**
- **Repeat purchases: Wednesday >> Sunday; September >> November**
- **Explored many other effects on retentions, repeat, and CLV**

Model Performance

- **Ensemble RMSE: 0.258**
- **Dummy RMSE: 0.380**
- **Model performed 32% better than dummy model (average CLV)**

Improved Approach Proposed

- **Any SKU from first order**
- **Model Future CLV**
- **Ridge or Lasso Regression**
- **Include time of day in models**
- **Use different ensemble models**
- **Focus model on customers we want to retain**
- **Model when second purchase occurs**
- **Log10 on more features**

Special Thanks to

- **Springboard TA’s**
- **Springboard Mentor, Raghu**