Customer Lifetime Value (CLV) Prediction

Authors: Zachary John Bowers, Abhishek K. Gawande, Ajay Krishna Manoj, Dianjin Xu

Course: ISYE 6414 Statistical Modeling and Regression Analysis

School: Georgia Institute of Technology

Level: Masters

Contact: zbowers3@gatech.edu; agawande7@gatech.edu; amanoj31@gatech.edu; dxu327@gatech.edu

Summary:

Customer Lifetime Value (CLV) is an important metric for businesses as it measures the total worth/profit from a particular customer to a business over a period of time. This metric informs business leaders on decisions regarding strategic focus on existing customer retention or new customer acquisition; marketing strategies targeted at specific segments of the populace; and recommended advertisement budget projections. This project aims to predict the CLV of customers based on 1-year transaction data (2012). Our team created 18 customer level quantitative and qualitative attributes based on invoice data to predict the future CLV, as measured by Revenue. We prepared the data, performed variable selection, checked the validity of the model assumptions, and checked for the errors on test data. Based on the performance, variable selection, and goodness of fit metrics identified, three models namely: Full-Transformed, Step-Wise-Transformed, and Lasso Transformed were found to be good candidates for the purpose. The net total revenue for the first 6 months of 2013, with 3546 customers is predicted to be $ 28.03 M with 64.8 % of the customers belong to the $0-$1000 segment.

Table of Contents

1. Introduction

Businesses are reliant upon customers for generating revenue and profits, but different customers segments often contribute to these profits in unequal proportions. Quantifying this measure can facilitate business strategies focused on acquiring and retaining customers that contribute the most to a business' profits. Many customer metrics, such as Net Promoter Score and Customer Satisfaction Scores, facilitate the evaluation of the intangible value of a business: including customer loyalty, brand engagement, and customer satisfaction.[i] These measures fail to directly tie these intangible valuations to profits. Alternatively, Customer Lifetime Value (CLV) is "the amount of profit your company can expect to generate from a customer, for the time the person (or company) remains a customer."[ii] There are numerous methods of deriving a customer's CLV, but all provide a metric that allows companies to determine the relative value of a customer to a company in relation to a company's profits. CLV provides value in improving customer retention, drives repeat sales, encourage higher value sales, and ultimately increase profitability.[iii] While sometimes challenging to measure, CLV has become of immense importance to business and industry leaders.

At its most basic level, CLV is determined by (1) calculating a customer's average order value; (2) determining the average number of orders/transactions per time period; (3) measuring a customer's lifetime in terms of the time periods; and (4) producing a CLV as the product of these three metrics.[iv] The variations of approaching these four basic steps have resulted in numerous different CLV modeling techniques, including generalized regression, quantile regression, classification and regression trees, Markov Chain models, and neural networks.[v] Understanding a business' desired use of the derived CLV can help with determining the appropriateness of the model, but it is important to understand the underlying CLV model to avoid misinterpretation of the calculated CLV.

This is of particular importance when considering the desire for appropriate customer segmentation as well as ensuring an accurate determination of a customer's contribution to a business' profits. For example, some models include customer acquisition cost while others do not.[vi] Identifying relevant business decisions [i.e. customer loyalty programs directed at large businesses versus small businesses; promotions on high profit products; marketing strategies targeting online versus in-person customers] will impact how the desired CLV model should be developed to facilitate effective customer segmentation. These concerns of both producing accurate CLV predictions as well as effectively segmenting customers to facilitate decision-making must be addressed by any CLV model.

2. Background/Purpose

To address the accurate CLV prediction, this paper will explore and apply relevant linear regression methods and techniques to develop a Recency, Frequency, and Monetary (RFM) model to assist a small online retail company in improving its profitability. The online retailer is a UK-based business established in 1981 with approximately 80 staff personnel selling unique all-occasion gifts. Two years prior to the data acquisition, the company shifted completely from direct mailing catalogues to online sales while maintaining its customer base from all parts of the UK and Europe.[vii] Our group used a Recency, Frequency, Monetary Value (RFM) model to assist in CLV prediction. An RFM score helps with customer segmentation by deriving a standardized score from the product of these three values. Recency (R) is the time duration since a customer's last purchase. Frequency (F) is how many purchases a customer has made over a certain time period. Monetary Value (M) is the total revenue a customer has generated over a certain time period.[viii] These values can then be standardized, ranked, and weighted before deriving an RFM score per customer. These scores in addition to other features can help with customer segmentation when building a model to predict CLV.

As such, our group hypothesized that customers who have purchased items more recently, purchased items more frequently, and purchased the highest grossing items would have a higher CLV score. Additionally, our group expected that customer behavior concerning each individual purchase transaction – total unique items bought per order and total items bought per order – would be positively correlated with a customer's CLV value. Alternatively, our group anticipated that customer's who return items more frequently or in higher proportions would be negatively correlated with CLV. Finally, our group was interested to see if a relationship existed between a customer's country of residence as well as type of item purchased when predicting CLV.

3. Methodology

3.1 Dataset and Preparation

Dataset Overview

In order to develop a predictive CLV model, our group utilized the online company's transaction data for a 12 month period – totaling 541,909 customer transactions from December 1st, 2010 until December 10th, 2011.[ix] Each transaction included the 8 attributes listed in Table 1.[x]

Attribute Name	Description	Example
InvoiceNo	Invoice Number: a unique six-digit number assigned to each transaction. Cancellation Invoice Numbers start with letter "C". A single invoice/basket often contains multiple transactions, i.e. multiple items.	536365 C536379
StockCode	Product/Item Number: a unique five-digit number assigned to each product or item to differentiate products from each other. On occasion a character is included at the end of the five-digit number.	22568 84030E
Description	Product Description: name of product or item	English Rose Hot Water Bottle
Quantity	Product Quantity: the number of each product/item per transaction. Negative values indicate cancellations	1 -5
InvoiceDate	Invoice Date: date and time of each generated transaction	12/1/2010 8:26:00 AM
UnitPrice	Unit Price: numerical value of price per unit in sterling (British Pounds)	£10.55
CustomerID	Customer Identification Number: a unique 5-digit number assigned to each customer	17850
Country	Customer's Country of Residence	United Kingdom France

Table 1: Dataset Transaction Attributes

First, our group cleaned the data by removing 5269 duplicated transactions. Next, we identified 135,037 transactions which were missing one of the eight attributes listed above – primarily CustomerID. As customer segmentation and customer specific CLV are the primary purposes of this study, our group removed these rows resulting in 401,603 remaining transactions to be utilized for analysis. These transactions equated to 22,190 unique invoices from 4372 customers originating from 37 countries.

As our model objective is to predict future CLV scores (and hence customer contributions to business revenue), our group then split the data displayed in Figure 1 into two different time frames: 147,327 transactions from the first six months from December 1st, 2010 to May 31st, 2011 and 254,276 transactions from the second six months from June 1st, 2011 to December 11th, 2011. While there is an increase in total transactions over time, the data appears relatively stable with few observable abnormalities. The first six months of data was utilized to analyze customer attributes – including RFM scores, transaction history, purchasing behavior, and country of residence. The analysis of this data would provide the model inputs for predicting future CLV. In contrast, the second six months was used to determine the customer's actual CLV score / contribution to the company's revenue. The analysis of this data would provide the model response of the actual future CLV.

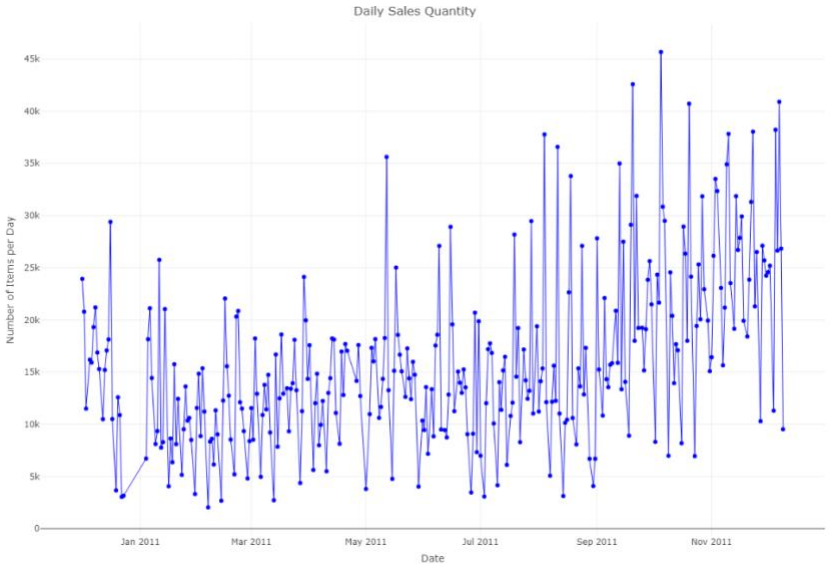

Figure 1: Daily Sales

Data Exploration

When conducting the initial data exploration, different categorical and numeric variables were analyzed for the first six months of transactional data to determine the potential for these variables to contribute either explanatory value to CLV calculations or assist in customer segmentation. We first analyzed the country of residence and number of order cancellations/returns to determine if these variables were frequent enough to provide effective segmentation. Figure 2 revealed that over 85% of the transactions originated from the United Kingdom in comparison to the 34 other countries. As such, we determined that an effective customer segmentation would be customers residing in the United Kingdom – home country of the business – and customers residing outside of the UK. While we only found 3574 of the 147,327 transactions were classified as order cancellations, the potential impact of such an event on CLV and revenue was deemed significant enough to remain in consideration as a potential explanatory variable.

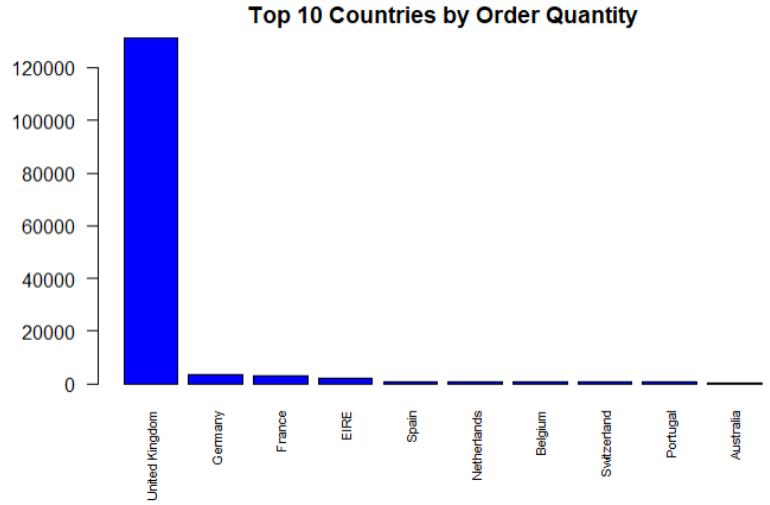

Figure 2: Transactions per Country

Figure 3 revealed that approximately 87% (2418 of 2767) of customers made 100 transactions or less over the first 6 months. This suggests that the Frequency variable of the RFM model will potentially provide effective customer segmentation in helping determine CLV. Due to the original dataset only including 8 transaction level attributes, our group determined that additional customer level features needed to be created from the transactional data to facilitate the development of an effective predictive CLV model.

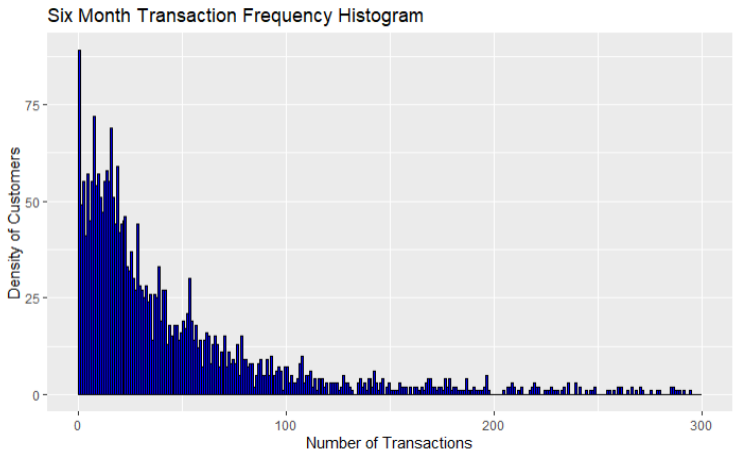

Figure 3: Number of Transactions per Customer Histogram

Customer Feature and Response Variable Creation

From the original transaction level data set, our group created customer specific attributes by grouping data per CustomerID in order to assist in analyzing customer attributes and behavior to assist in customer segmentation when deriving CLV models. These attributes ranged from Recency, Frequency, Monetary Value measures to invoice and transaction level behavioral data. Table 2 lists the 18 customer level attributes which would be used in subsequent variable selection procedures.

Column Name	Description	Value Range
Orders_Unique	Total number of sales purchase invoices per customer over six month period.	0 – 86 Sales Invoices
Returns_Unique	Total number of return/cancellation invoices per customer over six month period.	0 – 19 Return/Cancel. Invoices
Total_Items_Purchased	Total quantity of items purchased per customer over six month period	0 – 78,758 Total Items Purchased
Quantity_Basket	Average quantity of items purchased per customer per invoice over six month period.	0 – 4583 Items/Invoice
Total_Items_Returned	Total quantity of items returned/cancelled per customer over six month period	0 – 9360 Total Items Returned
Types_Items_Purchased	Total number of categories of items purchased per customer over six month period	0 – 881 Types of Items Purchased
Unique_Item_Per_Basket	Average number of categories of items purchased per customer per invoice over six month period.	0 – 173 Types of Items Purchased/Invoice
Types_Items_Returned	Total number of categories of items returned/cancelled per customer over six month period	0 – 52 Types of Items Returned
Unique_Item_Per_Return	Average number of categories of items returned/cancelled per customer per invoice over six month period.	0 – 45 Types of Items Returned/Invoice
Sales_Revenue	Total sales revenue generated per customer over six month period.	0 – 110,713 Pounds in Revenue
Return_Refund	Total amount refunded per customer over six month period	0 – 15,878 Pounds in Returned Revenue
Monetary_Value	Total net revenue per customer (sales revenue – return refund)	-72 – 110,668 Pounds in Net Revenue

Average_Unit_Price_Purchase	Average amount spent per item purchased per customer	0 – 295 Pounds per Item Purchased
Average_Unit_Refund_Return	Average amount refunded per item returned per customer	0 – 1716 Pounds per Item Returned
Country	Country of Residence	1793 Customers in UK 183 Customers Outside UK
Is_Buying_Most_Popular	Binary Variable. Determines whether customer buys one of top three most popular items (Item 23166 – Ceramic Jar; Item 22423 – Regency Cake Stand; Item 85123A – T Light Holder)	764 Customers Bought One of Three Most Popular Items 1212 Customers Did Not Buy Three Most Popular Items
Recency	Number of Days since the Customer's last purchase until the end of the reporting period (May 31st 2011)	0 – 181 Days Since Last Purchase
RFMSeg	Recency/Frequency/Monetary (RFM) Segmentation of Customers into three groups. 0th Group represents customers contributing to the lower third of revenue values. 1st Group represents customers contributing to middle third of revenue values. 2nd Group represents customers contributing to higher third of revenue values.	Group 0: 50 Customers Group 1: 1536 Customers Group 2: 390 Customers

Table 2: Final Dataset Customer Level Features

After creating the customer level features describing customer attributes and behaviors from the first six months of transactional data, the customer's future revenue contribution to the company was calculated from the next six months of data. This response value per customer equates to the customer's future CLV which is what our model is to predict. It should be noted that during the first six months of data there were 2767 customers whereas in the second six months of data there were 3577 customers. As the purposes of our model was only to consider characteristics of customers and their behavior to help predict future CLV, we only included the data for 1976 customers that were present in both six month segments. Future research focused on marketing and customer retention would be useful in identifying factors related to customer acquisition, retention, and departure and how these impact CLV calculations.

3.2 Modeling

Once all potential predicting variables and the response variable are created, a secondary data exploration procedure was conducted. When considering the RFM segmentation of customers in Figure 4, there does appear to be some explanatory value from this segmentation in predicting future revenue particularly with regard to customers in the highest RFM segment. When comparing the individual Frequency, Recency, and Monetary values of customers in the first six months compared to future net revenue in the second six months (Figures 5, 6, and 7, respectively), some relationships became apparent. As expected, past monetary value attributes appeared to have a strong positive relationship with future sales revenue expenditures by customer. This same positive relationship was also generally noted with regard to past order frequency and future revenue, albeit to a significantly lesser degree. Other than several outliers, a relationship between recency and future revenue was not apparent. This same type of exploration was conducted on all 18 potential predicting variables. Orders_Unique and Total Items Purchased also appeared to have a strong positive relationship to future sales revenue while the predictors related to returns/cancellations appeared to have a negative relationship with future sales revenue. In order to avoid unnecessarily excluding potentially important predicting variables in the final model, all variables would initially be included until appropriate variable selection techniques could be conducted later in the study.

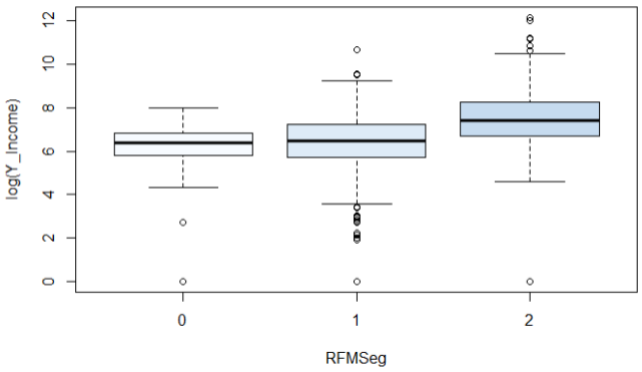

Figure 4: RFM Segment Boxplot

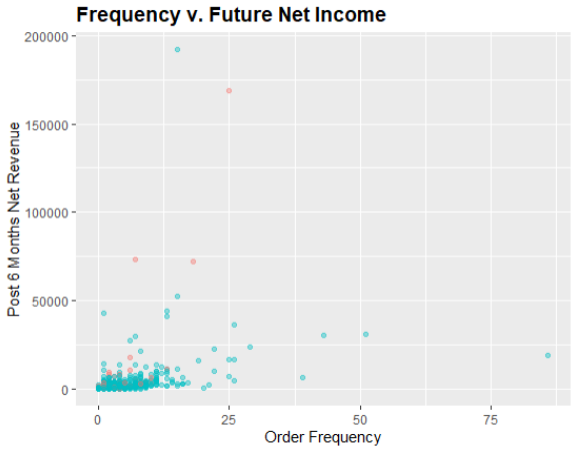

Figure 5: Frequency v. Future Revenue Scatterplot

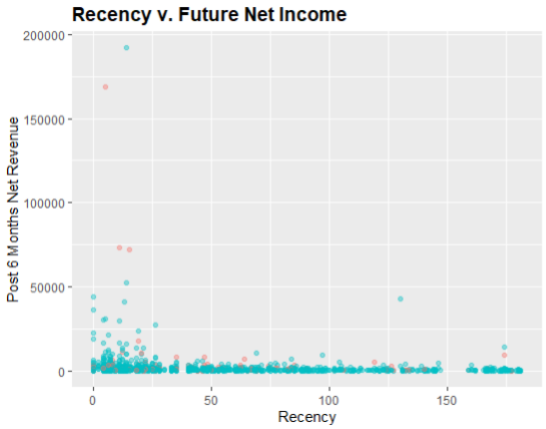

Figure 6: Recency v. Future Revenue Scatterplot

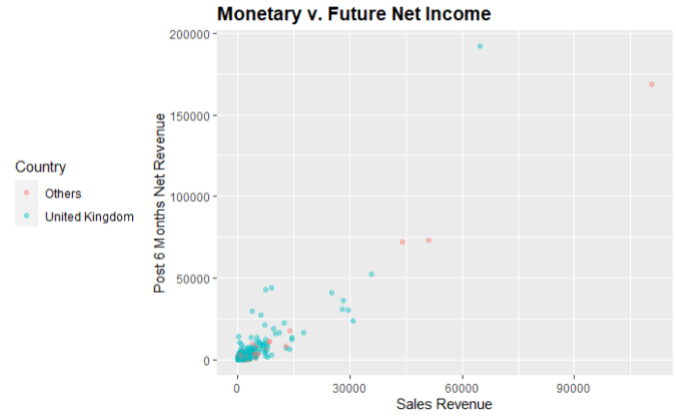

Figure 7: Monetary Value v. Future Revenue Scatterplot

In addition to looking at the graphical relationships between the predictors and response variables, we also assessed the dependent relationships between the predictor variables through correlation analysis in Figure 8. The most apparent strong correlation existed between Sales_Revenue and Y_income while there also existed lesser positive and negative correlations between several variables. Based on this initial correlation analysis, our group realized that multicollinearity might affect the predicted model and should be assessed during subsequent analysis.

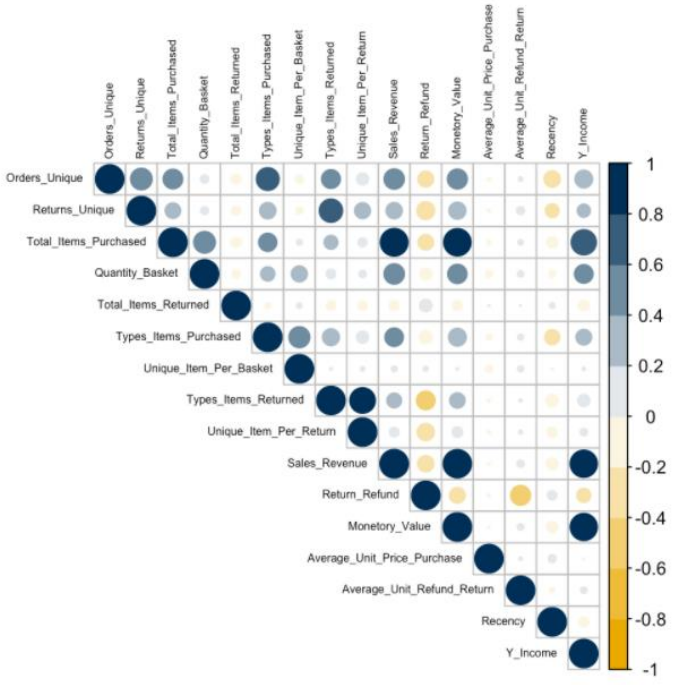

Figure 8: Predicting Variable Correlation Chart

After all of the data was properly cleaned, organized, and explored, our group planned to approach the development of a predictive CLV model in the following manner. First, our group performed a 70-30 split of this data set into a train dataset for fitting the model and a test dataset for assessing predictive performance. Next, to avoid unnecessarily excluding important predictive variables, we developed a multiple linear regression model encompassing all 18 predictive variables. Through checking for outliers, multicollinearity analysis, and transformations to improve on modeling assumptions, our group found a full model that would be used as the baseline for comparison of all subsequent models. After arriving at a full multiple linear regression model, our group utilized a variety of variable selection techniques – including Forward/Backward Regression, Lasso Regression, and Elastic Regression – to determine reduced models with potentially more optimal prediction risk values that would take into account the bias/variance tradeoff of more complex models. After determining several candidate models through each of these techniques, our group calculated prediction performance metrics for each of the models. These metrics provided the foundation of our recommended prioritization of each model for the online companies' leadership depending on the companies goals and risk tolerance.

Uniqueness of Modeling Approach

As explained in the introduction, CLV modeling has recently become an area of research interest to include the utilization of this particular dataset. Our group sought to expand on these previous research efforts in a variety of ways. Whereas other researchers removed cancellations from the dataset prior to feature creation, we instead used this data to create six additional customer level features which had the potential to provide additional explanatory power from a more holistic approach of customer behavior rather than simply looking at purchases. Additionally, we hypothesized that the particular items being purchased by customers may have also had a relationship with CLV calculations which led to our creation of the "Most Bought Item" feature. These additional features represented over a 50% increase in potential features/predictive variables in comparison to other approaches. Our group also developed a more holistic interaction variable called "RFM Score" to explore the impact that a potential singular metric of overall customer behavior would have on predicted CLV values. While other researchers have performed similar studies, these efforts often focused on simply the modeling of RFM scores as a response variable through various methods rather than using the derived score as a predicting variable for CLV values. In summary, our project was designed to expand on previous research efforts by looking at the inclusion of a wider range of customer features and behaviors through the use of existing datasets thereby avoiding the additional data collection costs normally associated with the inclusion of additional predicting variables.

4. Results

Fitted Full Multiple Linear Regression Model:

After performing the full multiple linear regression model for future revenue as a function of all predicting variables, we arrived at the model summarized in Figure 9.

Individual Predictors Statistical Significance:

At the 99% Confidence Level, we found that 9 of the 18 predicting variables were statistically significant given all other variables in the model. Specifically, "Returns_Unique", "Total_Items_Purchased", Quantity_Basket", "Total_Items_Returned", "Types_Items_Purchased", "Sales_Revenue", "Return_Refund, Country", and "Is_Buying_Most_Popular1"

Full Model Overall Significance:

After completing a partial F-test at the 99% Confidence Level, we determined a p-value to be ~ 0. As such, we rejected the null hypothesis that all predicting variable coefficients are equal to zero and concluded that the overall model is statistically significant and provides explanatory power to future CLV predictions.

Model Coefficient of Determination:

With an R-square value of 0.9052, the model explains almost 90.5% of the variability in the training data. While this value cannot be used in isolation to interpret the utility of the model, this score is a positive indication that the model may have explanatory value.

Interpretation of regression coefficients:

Regression coefficient for sales revenue is 2.728, which means that for every unit increase in sales revenue, expected revenue increases by $2.728 keeping all other variables constant.

Analysis of Unique Interaction Predicting Variables:

As discussed in our group's unique modeling approach, we created several variables to take business perspective into account and have a more detailed explanatory analysis in comparison to previous CLV studies. Examples include "Unique_Item_Per_Basket" and "Types_Items_Returned" through which we are attempting to capture the specific effect of the interactions of individual variables in the original dataset.

```
## lm(formula = Y_Income ~ ., data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32124   -583    -79    522  31922
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -1.013e+03  4.963e+02  -2.041 0.041489 *
## Orders_Unique                 -3.383e+01  3.109e+01  -1.088 0.276779
## Returns_Unique                -3.359e+02  9.077e+01  -3.701 0.000223 ***
## Total_Items_Purchased         -2.033e+00  7.571e-02 -26.854  < 2e-16 ***
## Quantity_Basket                1.060e+00  3.242e-01   3.270 0.001101 **
## Total_Items_Returned          -1.964e+00  2.700e-01  -7.274 5.88e-13 ***
## Types_Items_Purchased         -7.281e+00  2.698e+00  -2.699 0.007047 **
## Unique_Item_Per_Basket        -1.323e+00  6.113e+00  -0.216 0.828661
## Types_Items_Returned           1.729e+01  4.983e+01   0.347 0.728583
## Unique_Item_Per_Return        -5.829e+01  6.313e+01  -0.923 0.355988
## Sales_Revenue                  2.728e+00  1.124e-01  24.270  < 2e-16 ***
## Return_Refund                  1.888e+00  5.118e-01   3.689 0.000234 ***
## Monetory_Value                 2.273e-01  1.004e-01   2.264 0.023720 *
## Average_Unit_Price_Purchase -7.624e+00  7.566e+00  -1.008 0.313844
## Average_Unit_Refund_Return     1.205e+00  1.522e+00   0.792 0.428321
## CountryUnited Kingdom          8.221e+02  2.488e+02   3.304 0.000978 ***
## Is_Buying_Most_Popular1       -4.102e+02  1.541e+02  -2.662 0.007858 **
## Recency                        2.266e+00  1.550e+00   1.462 0.144057
## RFMSeg1                        5.139e+02  4.321e+02   1.189 0.234494
## RFMSeg2                        2.255e+02  4.884e+02   0.462 0.644316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2545 on 1363 degrees of freedom
## Multiple R-squared:  0.9052, Adjusted R-squared:  0.9038
## F-statistic: 684.7 on 19 and 1363 DF,  p-value: < 2.2e-16
```

Figure 9: Full CLV Model Summary

- *Null hypothesis*: All regression coefficients related to unique interaction variables are 0 and do not add any significant explanatory power to the models.
- *Alternative Hypothesis*: At least 1 unique interaction variable has non-zero coefficient and does add significantly to the explanatory power to the models.
- *Anova F-Partial Test Results*: P-value for this test was 1.37e-07. Consequently, we reject the null hypothesis that the coefficients for the interaction variables terms are all 0 and conclude that at least one of these variables is statistically significant and adds to the explanatory power of the model.

Outliers:

After performing checks for overall and predicting variable coefficient significance, our group sought to identify any outliers in the data with a Cook's Distance Plot. As depicted in Figure 10, we found that several data points were clear outliers in the model. To measure the impact of these data points, our group removed the outliers and retrained the model. As this did reduce the R-square value to 0.8956, these datapoints did have some influence on the overall model. After further evaluation of each of these individual observations, these observations were deemed to be anomalies that (from a business perspective) were unlikely to be encountered in the future and were removed for the duration of our group's study.

Multicollinearity Assessment:

After removing outliers, our group assessed the level of correlation between predictors that could negatively affect the model's performance. After calculating each of the variables' Variance Inflation Factor (VIF) to check for this multicollinearity, three important impacts to the study were observed:

1. VIF values were high for the predicting variables "Total_Items_Purchased", "Sales_Revenue", and "Monetory_Value".
2. As "Total_Items_Purchased" is regarded as important separate metric from a business perspective to the model, we chose to include this in the model.

3. Alternatively, as the correlation between "Sales_Revenue" and "Monetory_Value" was extremely high – correlation coefficient of 0.955 – our group removed Monetory_Value as it was a redundant predictor for "Sales_Revenue".

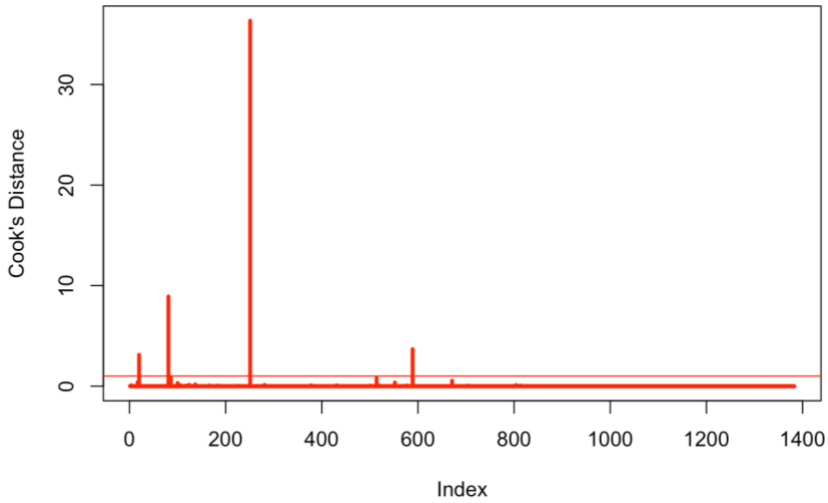

Figure 10: Full Model Outlier Analysis

Model 2: Removing Outliers and Redundant "Monetory_Value" Predictor:

After creating a second model with Y_Income as the response variable and all predicting variables excluding "Monetory_Value", we evaluated the overall and individual variable significance again and performed residual analysis to test model assumptions.

Model 2 Significance:

- Overall Significance: Significant at 99% Confidence level with a p-value of (<2.2e-16).
- Individual Predictor Significance: A major observation is that number of significant variables(given every other variable in the model) has been reduced to 8 to 4, thereby confirming the impact of outliers and multicollinearity previously identified.
- Interaction Term Explanatory Power Significant: ANOVA F-Partial Test resulted in a p-value of 0.006394. We can conclude that at least one of the interaction variables is statistically significant and adds to the explanatory power of the model.

Goodness of Fit:

Our group then performed Residual Analysis to assess how well the model assumptions of linearity, uncorrelated error, constant variance, and normality hold.

1. Linearity Assumption: After plotting each predicting variable versus the standardized residuals, the assumption of linearity appeared to hold as the observations are evenly distributed across the mean 0 line, as displayed in Figure 11.

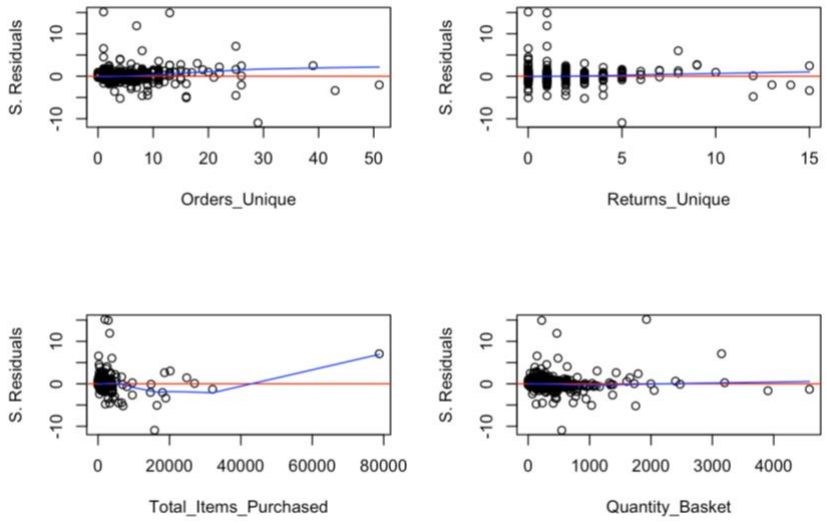

Figure 11: Linearity Assumption Check (Standardized Residuals v. Predictors)

2. Constant Variance and Independence Assumptions: After plotting the fitted values versus the standardized residuals as shown in Figure 12, the constant variance assumption did not appear to hold as the variance increases as the fitted values increase. When assessing uncorrelated errors, we observed no apparent clusters and conclude that this model assumption is valid.

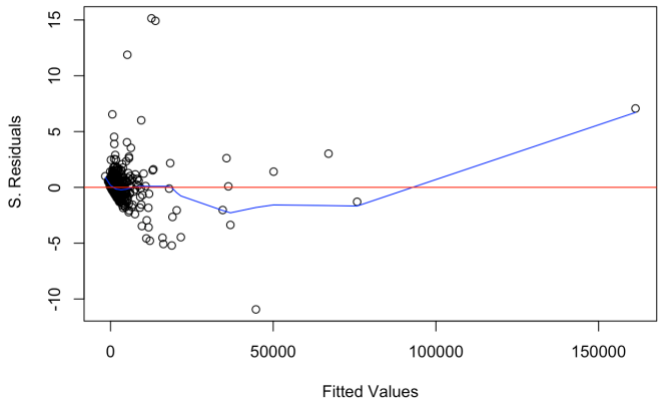

Figure 12: Constant Variance & Uncorrelated Error Checks

3. Normality Assumption: After reviewing the histogram and qq-plots in Figure 13, the normality assumption appears to be violated as the tails deviate from the expected normal distribution.

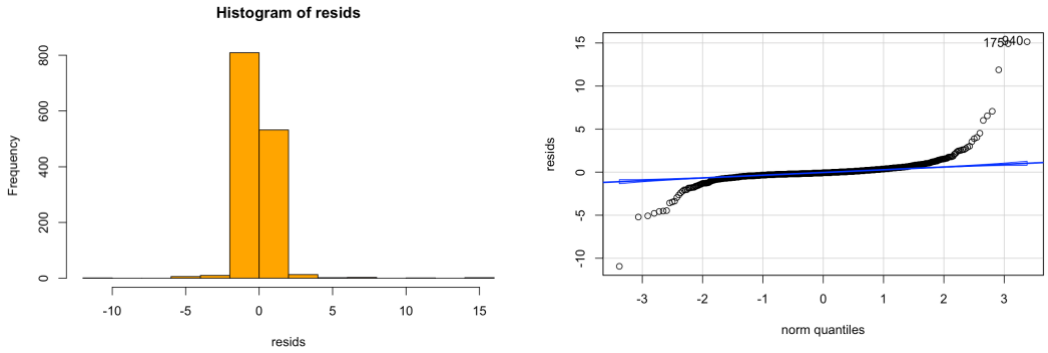

Figure 13: Normality Check – QQ-Plot and Residuals Histogram

Full Model Transformation:

In order to improve the deviations observed in the goodness of fit analysis, we decided to explore transformation techniques. Using the box-cox transformation method, we determined a square root transformation to be the most appropriate transformation. After performing this transformation on the response variable, we produced another model with the following observations:

- R-square value decreases from 0.8955 to 0.6997. While not as high as before, this value still represents a very high proportion of the variability in the data.
- Subsequent residual analysis reveals that linearity and uncorrelated assumption remain valid, and as depicted in Figure 14, normality and constant variance assumptions have significantly improved. We can see that the transformed model is a better fit although with a lesser R-square value.

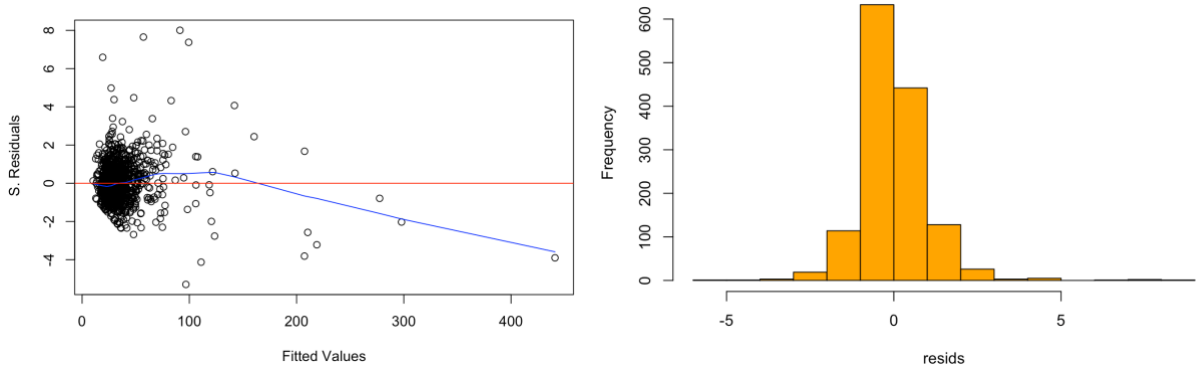

Figure 14: Constant Variance and Normality Checks

Variable Selection:

After developing a model incorporating all predicting variables – excluding redundant variables – we performed variable selection to determine if reduced models could be even more effective at predicting future CLV without the artificially high performance metrics inherent in more complex models. The three variable selection techniques utilized were Stepwise Regression, Lasso Regression, and Elastic-Net Regression approaches.

Stepwise Regression

First, our group performed backward-forward stepwise selection with AIC. As we did not force any variables to be selected, our minimum model only included an intercept and the upper model was our full model.

This variable selection model technique identified 10 variables to be excluded from the full model:
"Returns_Unique", "Quantity_Basket", "Total_Items_Returned", "Unique_Item_Per_Basket", "Unique_Item_Per_Return", "Return_Refund", "Average_Unit_Price_Purchase", "Is_Buying_Most_Popular1", "Recency", "RFMSeg1", "RFMSeg2"

LASSO Regression:

Next, our group performed lasso regression on the training set by using the R function cv.glmnet() to determine the lambda value that minimized the cross-validation error using 10 fold CV. This optimal lambda value was found to be 0.006786743 with the Regression Path Chart displayed in Figure 15.

This variable selection model technique identified 8 variables to be excluded from the full model:
"Total_Items_Purchased", "Total_Items_Returned", "Unique_Item_Per_Basket", "Types_Items_Returned", "Return_Refund", "Average_Unit_Price_Purchase", "Is_Buying_Most_Popular", "RFMSeg1", "RFMSeg2".

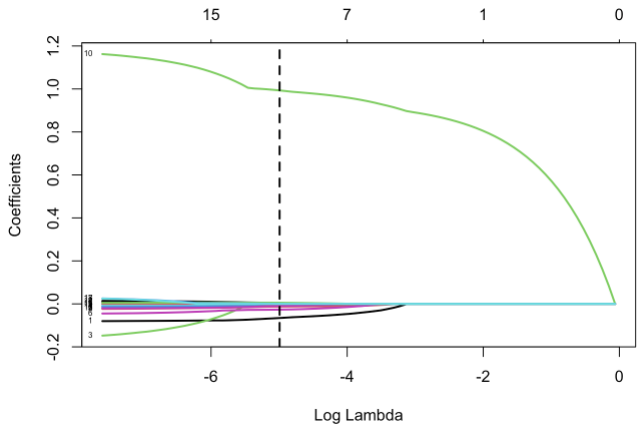

Figure 15: Regression Path Chart

For both the Stepwise Regression and LASSO models, variable transformations were conducted in order to improve on the deviations from the model assumptions when conducting Goodness of Fit analyses.

Elastic-Net Model:

Finally, our group performed elastic net regression on the training set and used the R function cv.glmnet() to find the lambda value that minimizes the cross-validation error using 10 fold CV while giving equal weight to both penalties. The optimum lambda value was found to be 0.008524545.

This variable selection model technique identified 7 variables to be excluded from the full model: "Total_Items_Purchased", "Total_Items_Returned", "Unique_Item_Per_Basket", "Types_Items_Returned", "Average_Unit_Price_Purchase", "Is_Buying_Most_Popular1", "RFMSeg2"

Model Comparison:

In order to help business leaders assess which models had the best predictive power, our group performed predictions utilizing the test data. Four different prediction metrics as well as the R Squared and adjusted R Squared values are displayed in Table 3.

	MSPE	MAE	MAPE	PM	R.Squared	Adj.R.Squared
Full	13025690.442	1187.425	48.053	0.242	0.895	0.894
Full-Transformed	4621711.506	939.890	38.181	0.086	0.700	0.696
Step-Wise	12953893.831	1170.067	48.534	0.241	0.895	0.894
Step-Wise-Transformed	6125621.072	990.763	24.206	0.114	0.659	0.657
Lasso	13762029.042	1165.035	45.121	0.256	0.894	0.893
Lasso-Transformed	12525764.223	1022.407	31.182	0.233	0.675	0.673
Elastic Net	14255321.677	1191.445	45.077	0.265	0.894	0.893

Table 3: Model Prediction Comparisons

Variable Selection Summary:

	Number_of_coefficients	Number_of_significant_coefficients
Full	19	4

Full-Transformed	19	11
Step-Wise	8	3
Step-Wise-Transformed	8	6
Lasso	11	4
Lasso-Transformed	11	6
Elastic Net	12	6

Table 4: Variable Selection Comparison

Goodness of Fit Comparison

	OveralL_GOF	Linearity	Constant Variance	Independence	Normality
Full	Inadequate	Valid	Violated	Valid	Violated
Full-Transformed	Average	Valid	Valid	Valid	Improved
Step-Wise	Average	Valid	Inconclusive	Valid	Violated
Step-Wise-Transformed	Average	Valid	Valid	Valid	Improved
Lasso	Average	Valid	Violated	Valid	Violated
Lasso-Transformed	Average	Valid	Valid	Valid	Improved
Elastic Net	Average	Valid	Violated	Valid	Violated

Table 5: Goodness of Fit Comparison

Three Candidate Models

Based on the performance, variable selection, and goodness of fit metrics identified in Tables 3, 4, and 5, our group selected three candidate models for assisting in business decisions centered on CLV: Full-Transformed, Step-Wise-Transformed, and Lasso Transformed, listed below.

Full-Transformed:

$$Revenue\_in\_Next\_6\_Months = 14.2106116 + 1.5244188 * Orders\_Unique - 0.0058896 * Total\_Items\_Purchased + 0.0191830 * Quantity\_Basket - 0.0362494 * Total\_Items\_Returned + 1.0104432 * Types\_Items\_Returned - 1.6161617 * Unique\_Item\_Per\_Return + 0.0070713 * Sales\_Revenue - 0.0282769 * Recency + 5.8480868 * RFMSeg1 + 9.7567188 * RFMSeg2$$

Step-Wise-Transformed:

$$Revenue\_in\_Next\_6\_Months = 23.7851972 + 0.0060357 * Sales\_Revenue + 1.3435685 * Orders\_Unique - 0.0029727 * Total\_Items\_Purchased + 0.0312137 * Types\_Items\_Purchased + 0.3591615 * Types\_Items\_Returned$$

Lasso-Transformed:

$$Revenue\_in\_Next\_6\_Months = 23.474078 + 1.561800 * Orders\_Unique + 1.100590 * Returns\_Unique + 0.015342 * Quantity\_Basket + 0.018765 * Types\_Items\_Purchased + 0.003314 * Sales\_Revenue - 0.021147 * Recency - 2.209740 * RFMSeg1$$

From Table 5, all three moderately satisfy the multiple linear regression modeling assumptions after variable transformation. From Table 4, the Full-Transformed model incorporated nineteen coefficients with ~60% (11/19) being statistically significant. The Step-Wise-Transformation model involved eight coefficients with ~75% (6/8) being statistically significant. Lastly, The Lasso-Transformation model included eleven coefficients with ~35% (4/11) being statistically significant. These coefficients are listed in Table 6.

X* => *Selected Features;* ***X **=> *Selected Feature is Significant***	**Full-Transformed**	**Step-Wise-Transformed**	**Lasso-Transformed**
Orders_Unique	**X**	**X**	**X**
Sales_Revenue	**X**	**X**	**X**
Types_Items_Purchased	**X**	**X**	**X**
Quantity_Basket	**X**	**X**	
Recency	**X**	**X**	
RFMSeg	**X**	**X**	
Unique_Item_Per_Return	**X**	**X**	
Returns_Unique	**X**	**X**	
Types_Items_Returned	**X**	**X**	**X**
Total_Items_Returned	**X**		
Country	**X**	**X**	**X**
Average_Unit_Refund_Return	**X**	**X**	**X**
Is_Buying_Most_Popular	**X**		
Unique_Item_Per_Basket	**X**		
Return_Refund	**X**		
Average_Unit_Price_Purchase	**X**		

Table 6: Variables Selected

Six key features – "Order_Unique", "Sales_Revenue", "Types_Items_Purchased", "Type_Items_Returned", "Country", and "Average_Unit_Refund_Return" – were included in all models.

- While Country and Average_Unit_Refund_Return are not statistically significant, they should be included as controlling variables.
- "Order_Unique" is the total number of sales purchase invoices per customer over six months and has a positive relationship with predicted revenue. Holding all other predicting variables constant, encouraging a customer to become a more frequent purchaser should encourage a higher CLV.
- "Sales_Revenue" is the total sales revenue generated per customer over six month period and has an unsurprising positive relationship with predicted future revenue.
- "Types_Items_Purchased" varies between a negative and positive coefficient depending on the model chosen. Further investigation into the relationship of this predictor with future CLV is recommended.
- "Types_Returned" is the total type of items returned/canceled per customer over six month period. Both the Full_Transformed model and the Step-Wise-Transformed model are indicating that as the types of items return/cancel increase, the more revenue this customer will generate in the future while holding other predicting variables unchanged.

Based on the results in Table 3, the full-transformed model has the lowest MSPE and MAE among all models. The R^2 of this model is 0.7 which means that 70% of the variance can be explained by using this model. Although, this model performed quite well in the forth mentioned categories. However, this model included many more predicting variables than others. More predicting variables usually means more data mining and storage efforts. The Step-Wise-Transformed model also returned a smaller MSPE and MAE among all other models. However, the R^2 is around 0.659 which is the lowest among all models. Due the smaller size of the predicting variable categories, it is a reasonable trade-off. Lastly, the Lasso-Transformed model returned a MSPE = 12525764.223 and a MAE = 1022.407. The R^2 of Lasso-Transformed model is approximately 0.673. Thus, the Step-Wise-Transformed model and Lasso-Transformed model both gave us smaller model with a reasonable explanatory power.

5. Conclusions

As Customer Lifetime Value directly ties customer level data to business revenue, this metric has been of increasing importance in comparison to other traditional customer level data assessing loyalty, satisfaction, and more intangible valuations. Whereas invoice and sales data traditionally has been used for evaluating overall business performance – Sales Growth, Gross Profit Margin, Cash Flow Margins – utilizing this data for customer level analysis can provide significant insight into a tangible evaluation of customer contribution to these overall metrics. This study

took this idea and expounded on existing research to create 18 customer level quantitative and qualitative attributes based on invoice data to predict the future CLV, as measured by Revenue, individual customers contribute to a business. By utilizing readily understandable multiple linear regression techniques and verifying that the model did not violate the theoretical assumptions for its appropriate utilization, our group developed a predictive CLV model incorporating these predictive variables to assist business leaders in predicting future CLV based on characteristics of its existing customer base. Our group then performed three variable selection processes to create reduced models with fewer predictive variables to both avoid potentially inflated model performance measures inherent in complex models as well as help business leaders prioritize which customer level characteristics should be focused on for increasing a customer's CLV. Based on the performance, variable selection, and goodness of fit metrics identified, three models namely: Full-Transformed, Step-Wise-Transformed, and Lasso Transformed were found to be good candidates for the purpose.

Business Interpretation and Considerations:

Our dataset consists of 2012, the first 6 months of which were used to train the models to predict the revenue for the last 6 months. Based on our models, we can predict the revenue for a customer given their past 6-month behavior. We used the last 6 months of 2012 to predict the expected revenue outside our given dataset timeline for early 2013 as this is the aim of our project. The model used is full-transformed. We have also taken the assumption of no churn.

Our prediction results are as follows:

1. The net total revenue for 3546 customers is $ 28.03 M.
2. The distribution of these customers as net expected revenue is as below:

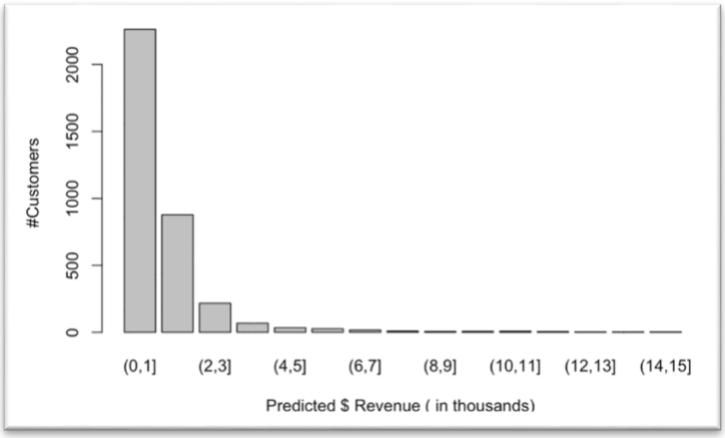

Figure 16: Scored Dataset for 2013

*64.8 % of the customers belong to the $0-$1000 segment

*24.8 % of the customers belong to the $1000-$2000 segment

This CLV information from Figure 16. can be used for target marketing, determining customer acquisition costs, customer service settings as well as contract management for long term customers

Future Scope:

- Another way to calculate the revenue/CLV is by predicting it separately for different segments of customers rather than individually customer prediction. K-Means clustering of customers can be used to identify the segments.
- One of the assumptions in our CLV approach is regarding customer churn. However, we can predict the churn of a customer through classification. This can be multiplied with revenue to get a relatively lower estimate of CLV.
- In our approach, we were restricted with features from Transactional Data. However, we can find and utilize others like Clickstream Website Behavior features, Demographics and Promotional and Campaign information.
- We have used Multiple Linear Regression for our modelling; however, we can try out different models like Random Forest, XGBoost, etc. to have a wider base to select models.

References

[i] Qualtrics XM. (2021) How to Measure Customer Loyalty. Accessed 1 December 2021. <https://www.qualtrics.com/uk/experience-management/customer/measure-customer-loyalty/>.; Qualtrics XM. (2021) What is CSAT (Customer Satisfaction Score)?. Accessed 1 December 2021. <https://www.qualtrics.com/uk/experience-management/customer/what-is-csat/>.

[ii] Gallo, Amy. Harvard Business Review (15 July 2014). "How Valuable are Your Customers?" Accessed 01 December 2014. <https://hbr.org/2014/07/how-valuable-are-your-customers>.

[iii] Caldwell, Austin. Oracle Netsuite (13 April 2021). "What is a Customer Lifetime Value (CLV) & How to Calculate?" Accessed on 01 December 2021. <https://www.netsuite.com/portal/resource/articles/ecommerce/customer-lifetime-value-clv.shtml>.

[iv] Ibid.

[v] Sharma, Shreya. Capstones, Theses, and Dissertations; Ivy College of Business; Iowa State University; Ames, Iowa (Spring 2021). "Customer Lifetime Value Modeling." <https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=1882&context=creativecomponents>, p.9.

[vi] Ibid.

[vii] Chen, Daqing; Sain, Sai Laing; Guo, Ken; Database Marketing & Customer Strategy Management (Vol 19, 3, 197-208). "Data Mining for the Online Retail Industry: A Case Study of RFM model-based customer segmentation using Data Mining." <https://link.springer.com/article/10.1057/dbm.2012.17>, p. 199.

[viii] GL, Manoj. Analytics Vidhya (27 April 2021). "Customer Lifetime Value using RFM Analysis". Accessed 2 December 2021. <https://www.analyticsvidhya.com/blog/2021/04/customer-lifetime-value-using-rfm-analysis/>.

[ix] Dr Daqing Chen, Director: Public Analytics group. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK. <https://archive.ics.uci.edu/ml/datasets/online+retail>.

[x] Sharma, 2021, 12.