Applying Machine Learning Models to Improve Customer Lifetime Value
Final Project Report For BUS228, Spring 2022

Vahe Baghdasaryan
American University of Armenia
vahe_bagdasaryan@edu.aua.am

Mary Margaryan
American University of Armenia
mary_margaryan@edu.aua.am

1 Introduction

This paper explores the ways Machine Learning Models can be utilized in order to improve the Customer Lifetime Values for retail firms. As defined by Caldwell (2021), "Customer lifetime value (CLV) is a measure of the average customer's revenue generated over their entire relationship with a company." Venkatesan and Kumar (2004) believe that CLV helps managers select customers and allocate their resources by developing ways to improve their relationships with their customers.

For this project, we have defined the following problem to analyze and solve:

"The lack of CLV forecast or customer segmentation based on CLV increases companies' revenue churn and customer acquisition costs." Hence, our solution to that problem is to use Machine Learning to "Predict and classify CLV of the e-commerce store customers and recommend strategies to increase the CLV of their customers." By doing so, we hope to solve the problems mentioned above.

To understand why is CLV so important, we need to discuss the fact that those customers who are targeted based on their lifetime value prove to be more profitable for the company over the future periods, and other metrics that marketing managers use might be underperforming compared to CLV; thus the best suggestion for them is to focus primarily on maximizing the CLV of their customers (Venkatesan et al., 2004).

2 Previous Work

Throughout the review of other research papers in this discourse, we synthesized various approaches to analyzing the data and finding ways to predict it and utilize it for further marketing purposes.

Chen and Fan (2013) did a study on lifetime value prediction by utilizing longitudinal data, which will account for the dynamic nature of the variable. For predicting the CLV, they used an improved multiple kernel support vector regression, which helps them select the best marketing strategy for a given customer.

Alternatively, Tsai, Hu, Hung, and Hsu (2013) focused on researching the different hybrid models for predicting the customer value and found that classification+classification hybrid model outperforms clustering+classification hybrid model, with a higher accuracy rate of a two-stage decision tree.

3 CLV Calculations

To be able to do the predictive and classifying analysis, first of all, we spent a lot of time manipulating the data and calculating the CLV or customer lifetime value. We used the panda's function drop_duplicates() to get a flawless dataset, which drops the duplicates or repeated records from the dataset. We have filtered the necessary columns for calculating CLV and only used the following columns: CustomerID, InvoiceDate, InvoiceNo, Quantity, UnitPrice, Visits, Avg Time Spent, Age, and Salary.

After the filtration, we have performed a number of calculations, including calculating the number of days between the present date and the date of last purchase from each customer, calculating the number of orders for each customer, and calculating the sum of the purchase price for each customer, calculate the average order value for each customer. After the calculations mentioned above, we have done retention metric

calculations, including churn, retention rate, and purchase frequency. To calculate the purchase frequency, we divided our total number of orders by the number of unique customers for the same time frame. To calculate the retention rate, we divided the number of return customers by the total number of customers. The churn rate was calculated by subtracting the retention rate from 1. The profit margin was 5%. Then we calculated the CLV by following the granular CLV formula: CLTV = ((Average Order Value x Purchase Frequency)/Churn Rate) x Profit margin.

4 Data

The data that we used for the analysis (attached to this paper) is E-commerce retail store data. The variables in the dataset have records of the transactions with spent, Quantity, and other values. However, in order to improve the data and be able to perform more analyses, we added some other metrics to it. The variables that we added were the number of website visits, average time spent (seconds) on the website, age, and salary of our customers.

5 Methods

We've used linear regression (OLS), Decision Trees (Random Forest), Logistic Regression, and Lasso & Ridge Regression to complete the analysis. The analysis has been done to predict and classify the customer lifetime value for the particular e-commerce shop as well as to come up with recommendations to improve the customer lifetime value (CLV). For this research purposes the Average Quantity, Visits, Average Time Spent, Age & Salary served as feature variables and CLV served as a label variable. You can access all codes and data here.

5.1 Linear Regression

The linear regression was calculated to predict the CLV on the dependent variables mentioned earlier in the paper. A significant regression equation has found F-statistics 2074, with an R^2 of 0.799 or 79.9%. Additionally, we see that all feature variables are predictive and statistically significant (being p-value less than alpha). So, based on the linear regression and OLS output, we can conclude that the model explains 79.9% of the variance in our dependent variable. Based on our model, we see that as Average Quantity increases by 1, CLV will increase by 2.942e-05.

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.820
Model:	OLS	Adj. R-squared (uncentered):	0.819
Method:	Least Squares	F-statistic:	2074.
Date:	Sun, 22 May 2022	Prob (F-statistic):	0.00
Time:	21:56:20	Log-Likelihood:	-1262.0
No. Observations:	2744	AIC:	2536.
Df Residuals:	2738	BIC:	2571.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Quantity_average	2.942e-05	5.36e-06	5.490	0.000	1.89e-05	3.99e-05
Quantity	-3.435e-05	1.95e-06	-17.589	0.000	-3.82e-05	-3.05e-05
Avg Time Spent	0.0002	2.32e-05	7.040	0.000	0.000	0.000
Visits	0.0011	0.000	4.096	0.000	0.001	0.002
Age	0.0076	0.001	11.585	0.000	0.006	0.009
Salary	3.225e-06	3.33e-07	9.677	0.000	2.57e-06	3.88e-06

Omnibus:	565.048	Durbin-Watson:	1.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	972.125
Skew:	-1.363	Prob(JB):	8.05e-212
Kurtosis:	4.035	Cond. No.	8.02e+03

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 8.02e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 1. OLS Regression Results

5.2 Decision Trees & Random Forest

Decision Trees have been used to classify customers as healthy and unhealthy based on the number of their customer lifetime value. We have calculated the average number of customer lifetime values and denoted all values above that (1) being healthy and (0) being unhealthy. We received a 0.84 accuracy score during the analysis, which is the same as almost 84%. Then we used the Random Forest library to calculate the MSE (mean squared error), which was almost 0.13. Again Random Forest has been used to find out the importance of feature variables in our decision tree model. Based on the analysis, we find out that Quantity Average has the highest impact and is the most important feature variable.

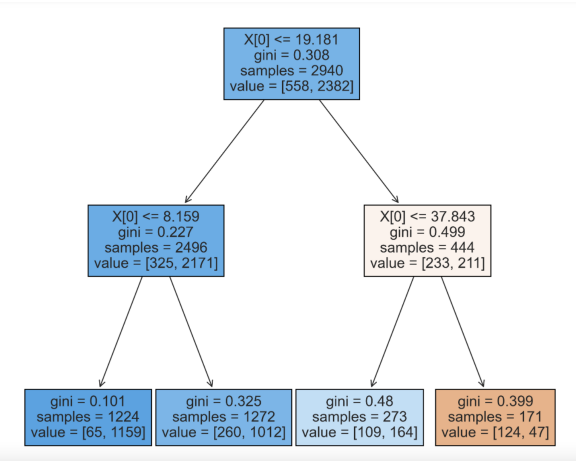

Figure 2. Decision Tree Visualization

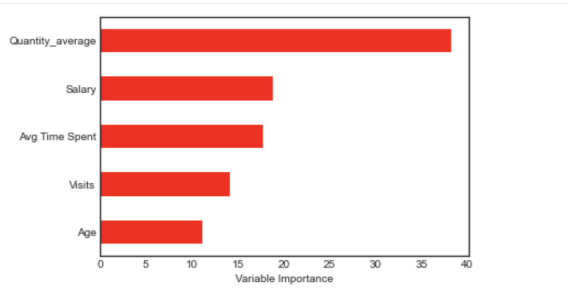

Figure 3. Variable Importance

5.3 Logistic Regression

The logistic regression has been used to do classification and come up with recommendations for the e-commerce shop. The logistic regression has been done by following the same principles mentioned for Decision Trees. During the regression, we received the accurrence score being 0.81, which is the same as almost 81%. Thus, the Logistic Regression model predicted that 81% of the time, the CLV is dependent on our feature variables. The precision score has been found as 0.814, which is the same as being almost 81.4%%. This means that our model can identify almost all the users whose CLV is dependent on our feature variables. The recall score has been found as 0.99, which means that the model we've built can recall almost all cases where the CLV depends on our feature variables.
We printed the confusion matrix during the logistic regression studies, where most of the cases were on the true negative side. This means that our machine learning program has been set on test data where there is an outcome termed negative that the machine has successfully predicted.
Besides this, we have also used the AUC and ROC curves to support our other findings. "The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve" (Reddy, 2020). On the other hand, "An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds" (Guanga, 2018). Our model predicted the AUC score to be 0.559. Typically, the AUC score of 1 represents a perfect classifier, and 0.5 represents a weak classifier. In our case, we almost got a perfect score. "The dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible" (Kharwal, 2021). In our case, we got a good classifier.

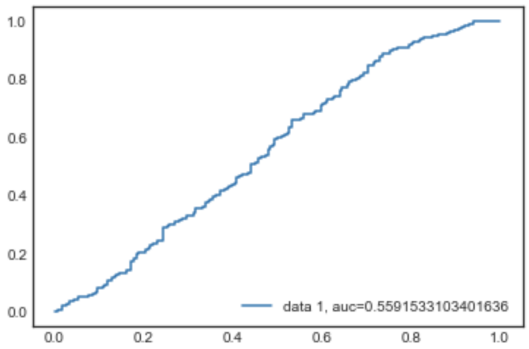

Figure 4. Logistic regression AUC and ROC curve

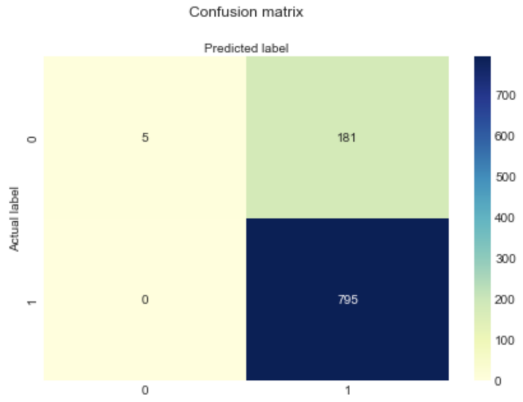

Figure 4. Logistic regression confusion matrix

5.4 Lasso Regression

Typically, lasso regression aims to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. Lasso does this by imposing constraints on model parameters. This shrinks the regression coefficients of some variables towards zero. (Mlwhiz, 2017). The mathematical formula for lasso regression is the following: Residual Sum of Squares + λ * (Sum of the absolute value of

the magnitude of coefficients). We've done lasso regression to support our findings and later on recommendations to improve the customer lifetime value for the e-commerce shop.

5.6 Ridge Regression

Ridge regression has been performed to support our findings and come up with solid recommendations for the e-commerce shop. The Ridge regression is the method used to analyze multicollinearity in multiple regression data. Ridge, as well as the lasso regressions, allowed us to regularize ("shrink") coefficients. This means that the estimated coefficients were pushed towards 0 to make them work better on new data sets ("optimized for prediction"). This allowed us to use complex models and avoid over-fitting simultaneously.

6 Results

Based on our research findings, we conclude that classification analysis (Decision Trees and Random Forest) worked the best than the regression models (OLS, Ridge, and Lasso). Classification analysis helped us better understand customers' health and develop strategies to improve the profitability of the e-commerce shop and the customer lifetime value of its customers.

7 Recommendations

After finalizing our regression and classification models, we have incorporated our findings to come up with strong and data-backed recommendations for the shop to improve the customer lifetime value of its customers.

7.1 Cross-Selling

For the users, those predicted CLV is lower than the average; we recommend the e-commerce shop increase the Quantity they purchase and dollar per order. When the user is on the checkout page, the shop can show them related items and ask them to add those items to their shopping card as Amazon does.

7.2 Personalized CRM

The company can integrate the machine learning predicting algorithm into its CRM system. For the users, those predicted CLV is lower than the average; we recommend the e-commerce shop increase the purchasing frequency + Quantity of the purchases. Through marketing automation, send out emails and personalized offers to those whose CLV is predicted to be lower than the average based on the regression tree.

7.3 Personalized Discounts

The company can integrate the machine learning predicting algorithm into its CRM system. For the users, those predicted CLV is lower than the average, we recommend the e-commerce shop increase the purchasing frequency and Quantity of the purchases. Through marketing automation, send out personalized discount offers to those whose CLV is predicted to be lower than the average based on the regression tree.

8 Conclusion and next steps

In conclusion, our proposed concept can help the e-commerce shop increase the customer lifetime value of its clients. Although the regression models couldn't predict the future customer lifetime value with high accuracy, the classification models have provided a good understanding of how to improve the customer lifetime value. From the above results, we can hypothesize that if the e-commerce shop incorporates recommendations, there is a high possibility that the customer lifetime value will be improved for those in the unhealthy stage. On the other side, we got highly high accuracy scores in Logistic regression and Decision Trees. We might try to rescale the data and bring the accuracy score to more realistic stages. We also have to think of ways to improve our regression models (linear, OLS), increase the significance of our feature variables, and make more accurate predictions about the forecasting CLV of the clients.

References

Caldwell, A. (2021, April 13). What Customer Lifetime Value (CLV) Is & How to Calculate It. Oracle NetSuite. https://www.netsuite.com/portal/resource/articles/ecommerce/customer-lifetime-value-clv.shtml?mc24943=v2

Chen, Z. Y., & Fan, Z. P. (2013). Dynamic customer lifetime value prediction using longitudinal data: An improved multiple kernel SVR approach. Knowledge-Based Systems, 43, 123-134.

Guanga, A. (2018, October 19). *Understand Classification Performance Metrics - Becoming Human: Artificial Intelligence Magazine*. Medium. https://becominghuman.ai/understand-classification-performance-metrics-cad56f2da3aa

Jain, D., & Singh, S. S. (2002). Customer lifetime value research in marketing: A review and future directions. Journal of interactive marketing, 16(2), 34-46.

Kharwal, A. (2021, June 22). *ROC Curve in Machine Learning*. THECLEVERPROGRAMMER. https://thecleverprogrammer.com/2020/07/26/roc-curve-in-machine-learning/

Mlwhiz. (2017, November 21). *LASSO Regression to predict the "School Connectedness" of a student*. MACHINE LEARNING HACKER. https://mlhackerblog.wordpress.com/2017/11/18/lasso-regression-to-predict-the-school-connectedness-of-a-student/

Reddy, S. K. R. (2020, November 1). *AUC ROC score and curve in multiclass classification problems :: InBlog*. Inblog. https://blog.ineuron.ai/AUC-ROC-score-and-curve-in-multiclass-classification-problems-2ja4jOHb2X

Tsai, C. F., Hu, Y. H., Hung, C. S., & Hsu, Y. F. (2013). A comparative study of hybrid machine learning techniques for customer lifetime value prediction. Kybernetes.

Venkatesan, R., & Kumar, V. (2004). A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy. Journal of Marketing, 68(4), 106–125. http://www.jstor.org/stable/30162020