REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

REVA Academy for Corporate Excellence (RACE)

CLTV using Machine Learning – A Comparative Study

Name: Anand KN

SRN: R19MBA52

Date: Aug 2022

PGDM/MBA in Business Analytics

Capstone Project Presentation
Year: II

race.reva.edu.in

Agenda

01 Introduction
Back Ground | Current status | Why this study

02 Literature Review
Seminal works | Summary | Research Gap

03 Problem Statement
Business Problem | Analytics Solution

04 Project Objectives
Primary & Secondary Objectives | Expected Outcome

05 Project Methodology
Conceptual Framework | Research Design

06 Business Understanding
Business Context | Monetary Impact

07 Data Understanding
Data Collection | Variables

08 Data Preparation
Pre-processing | Process | Techniques

09 Descriptive Analytics
Univariate | Bivariate | Hypothesis

10 Modeling
Machine Learning | Model Evaluation | Insights

11 Model Deployment
Applications | Demo

12 Suggestions and Conclusions
Insights | Next Step | Future Scope

13 Annexure
References | Publications | Plagiarism Score

Introduction

The importance of Customer Lifetime Value has been understated for very long. Researcher and Analysts have been talking about the importance of CLTV is for years, but it is still being ignored or underutilized. Traditionally most successful businesses measure their growth in two ways: the first is how it acquires new customers and second is how well the existing customer are retained and their worth or Customer Lifetime Value (CLTV) increased. But studies show that it is more profitable to retain existing customer than acquiring new customers.

Several factors account for the growing interest in the concept of CLTV. Marketing departments are coming under pressure and are held. IT today has capacity to process huge amounts of data which makes sampling redundant as the entire customer base can be made available for analysis. Also, analysts can convert this data into meaningful insights very easily with the help of many sophisticated in modelling techniques.

Different industry segments adopt different ways of predicting CLTV. Analysis of transactional data in calculating CLTV along with predictive modeling is a very extensive task. As part of this activity, both Linear and Non-Linear modeling techniques have been used to model the retail apparel firm’s data to predict CLTV. CLTV thus predicted using various modeling techniques is further analyzed using regression to derive the R-squared value which provides insights into which modeling technique will make a best fit for the data to predict CLTV.

Literature Review

CLTV stands for Customer Lifetime Value and measures all the potential profits a particular customer can bring to the organization. For instance, you have an online shop selling bicycles and all the additional products, and a new customer has just bought one. In the future, they may buy a helmet, new tires, a basket, etc. At some point, they may come for another bike. All these potential purchases and revenues are CLTV.

Lifetime Value is generally defined as the present value of all future profits obtained from a customer over his or her life of relationship with a firm. CLV is similar to the discounted cash flow approach used in finance. However, there are two key differences.

First, CLTV is typically defined and estimated at an individual customer or segment level. This allows us to differentiate between customers who are more profitable than others rather than simply examining average profitability. Second, unlike finance, CLV explicitly incorporates the possibility that a customer may defect to competitors in the future.

By Understanding their customers, a company can make strategies by which they can retain their customers and increase overall profitability. Calculating CLTV can help companies to investigate the parameters that companies generally ignore. At the beginning of a relationship, customers are more valuable due to the future potentials that they offer.Many studies use different methods, including generalized regression, logistic regression, quantile regression, latent class regression, CART, Markov chain modelling, neural network to create past customer behaviour models.

Problem Statement

The key problem areas that are addressed as part of this project work is

1. Determine the best model fit to predict CLTV by exploring various linear and non-linear machine learning algorithms

2. Perform a comparison of the key metrics of these techniques to determine the best fit model to recommend for prediction of CLTV

3. Calculate the monthly average value of customers for business to drive their marketing strategies

Project Objectives

Below are the objectives for predicting CLTV:

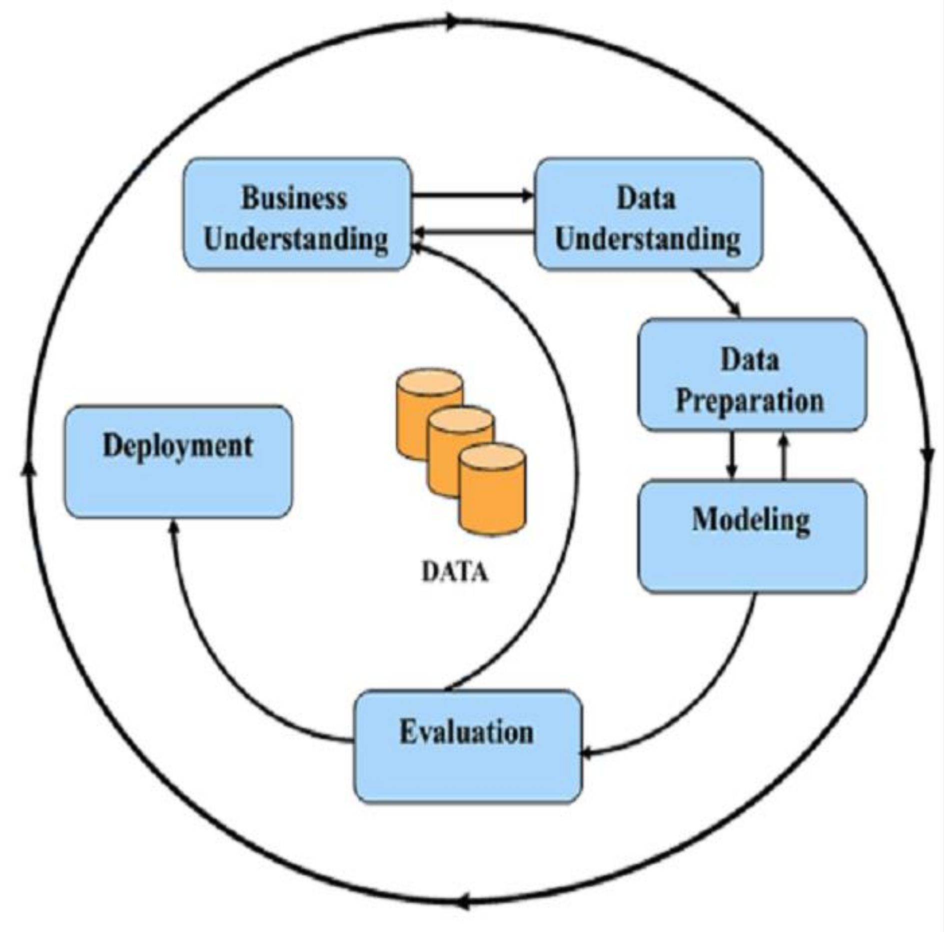
1. Compute CLTV for the customers using Linear and Non-Linear machine learning algorithms

2. Compare the metrics like R-squared values for the linear techniques and the Mean Absolute Error for the non-linear techniques to identify the best modelling technique

3. Deploy the best modelling technique for prediction of CLTV based on the comparison outcomes

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

Project Methodology



Project methodology that will be used as part of this project is being discussed in this chapter with details of the approach.

The process that will be followed as a project methodology is CRISP-DM framework that starts with understanding the business and then narrowing down into specific areas of interest like understanding the data, preparing the data for more insights, building the model, evaluating the model and deployment.

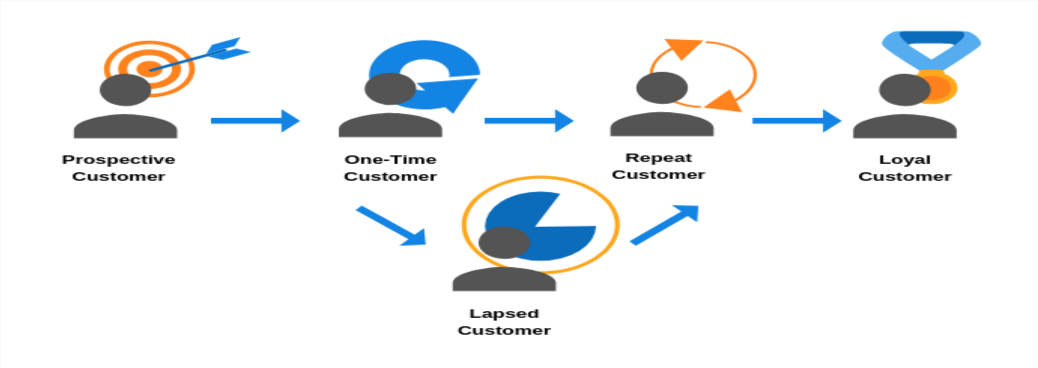
CLTV for the Apparel_POS data: Data used for this project consists of all the purchases from 2018 to 2020 for a period of 3 years. The company is an Apparels Retailer that sells apparels across India.

The main target variable that needs to be predicted here by looking at the data is if a customer continues to purchase or not with the retailer. If the customer continues his purchases, then the CLTV value would be on the higher side. As part of predicting the CLTV, we are more interested in identifying the repeat customers who give more business.

As part of this project, CLTV is predicted using Linear Modeling Techniques like Logistic Regression along with Variable Selection methods under Logistic Regression and Non-Linear modelling techniques like Decision Tree Methods – CART, CHAID, Exhaustive CHAID, and Quest. Neural Networks – Multi Layer Perceptron and Radial Basis Function and K Nearest Neighbors. Then the R-squared is derived from the predicted CLTV values using regression.

Business Understanding

Retail business is more challenging than ever. Competition is exponential along with Amazon being a major player in this domain which drives the acquisition costs to shoot up while customer bringing down the profitability. The days of increasing the size of the businesses by acquiring new customers are gone, it is the value that these customers can bring is what matters.



To increase revenue from the existing client base and to get a deeper understanding of the clients with high value is the only way today for the retailers to grow the business. Keeping the customers repeat their purchases is the key to get more value. To understand the client's preferences and behaviors, we need to make use of next level of analytics that is advanced in functionality and machine learning techniques. Client value is the best and candid way to run the business.

Data Understanding

Data used in this project is from retail Apparels business that sells apparels across India. This is set contains all the purchases occurring between 2018 and 2020 for a period of 3 years.

Here is the information on some of the key attributes:

- **Order No**: Order Number. Nominal, an integer number uniquely assigned to each purchase. The code starts with letter 'M'.
- **Quantity**: The number of each product (item) per purchase. Numeric.
- **Order Date**: Order Date and time. Numeric, the day and time when each item was ordered.
- **MRP**: MRP. Numeric, price of each item in Rupees.
- **Customer_ID**: Customer identification number. Nominal, an integer number uniquely assigned to each customer.

As per the formula for calculating CLTV below: (Hardie, 2006)

$$CLTV = \sum_{t=0}^{T} \frac{(p_t - c_t)\, r_t}{(1 + i)^t} - AC$$

Where

p_t = price of the item that a consumer pays at time t,
c_t = direct cost for servicing the customer at time t,
i = discount rate or cost of capital for the firm,
r_t = probability of customer repeat buying or being "alive" at time t,
AC = Acquisition cost, and
T = time horizon for estimating CLTV.

Data Preparation

As the data that we have for predicting CLTV in this project is a historic transactional data from a retail apparel firm, EDA is carried out on the data to prepare it for modeling. Here is snapshot of missing values check:

	Count	Proportion
Customer_ID	0	0.0
Order No	0	0.0
Order Date	0	0.0
Quantity	0	0.0
MRP	0	0.0
Total Sales	0	0.0

As part of feature engineering, new features are derived out of the existing columns to be able to use it for CLTV calculation based on the equation.

Below are the key features that are created to calculate:

t - this is value derived from the difference of recent order date and first order date
Sum of Product Discounted – This is based on the discount rate
Acquisition Cost = 0 (0 has been assumed as this is an online retailer)
Churn Propensity which is used to calculate Alive Propensity
Alive Propensity is 1 – Churn Propensity
Order Month – This is calculated based on the month of the order date

Descriptive Analytics

Descriptive analysis of the data to check on the key metrics has been carried out using python pandas library methods. Descriptive analytics summarizes the data by computing mean, median, mode, standard deviation likewise.

Below are the screenshots of descriptive analytics performed on the data along with further analysis on the data for details like the time range, total number of unique customers, total quantity sold, etc.

	Customer_ID	Quantity	MRP	TotalSales
count	17782.000000	17782.000000	17782.000000	17782.000000
mean	10263.435778	1.038916	1027.847824	1058.721179
std	6055.332535	0.284685	537.298888	568.953100
min	2.000000	1.000000	95.000000	95.000000
25%	4746.000000	1.000000	595.000000	595.000000
50%	10239.500000	1.000000	995.000000	995.000000
75%	15539.000000	1.000000	1295.000000	1295.000000
max	20610.000000	20.000000	3495.000000	6725.000000

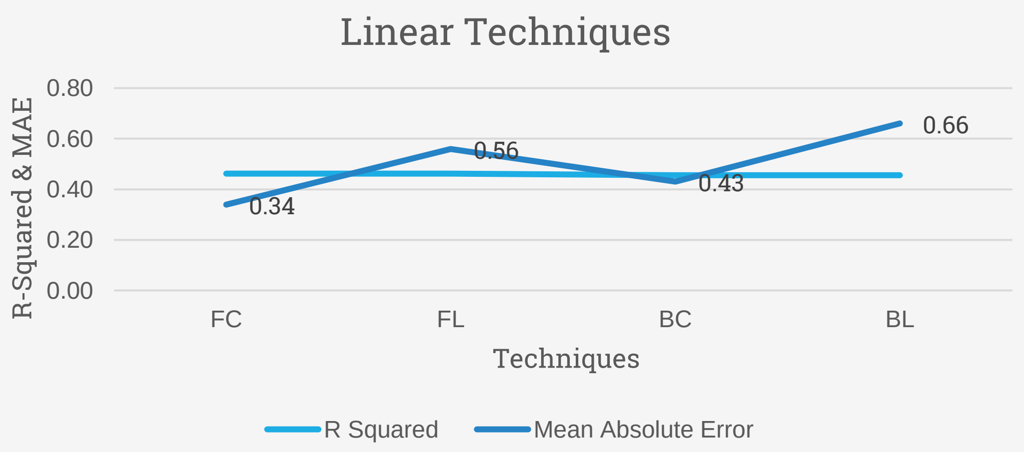
```
The Time range of transactions is: 2018-01-03 to 2020-12-02
Total number of unique customers: 8381
Total Quantity Sold: 18474.0
Total Sales for the period: 18826180.0
```

Modeling

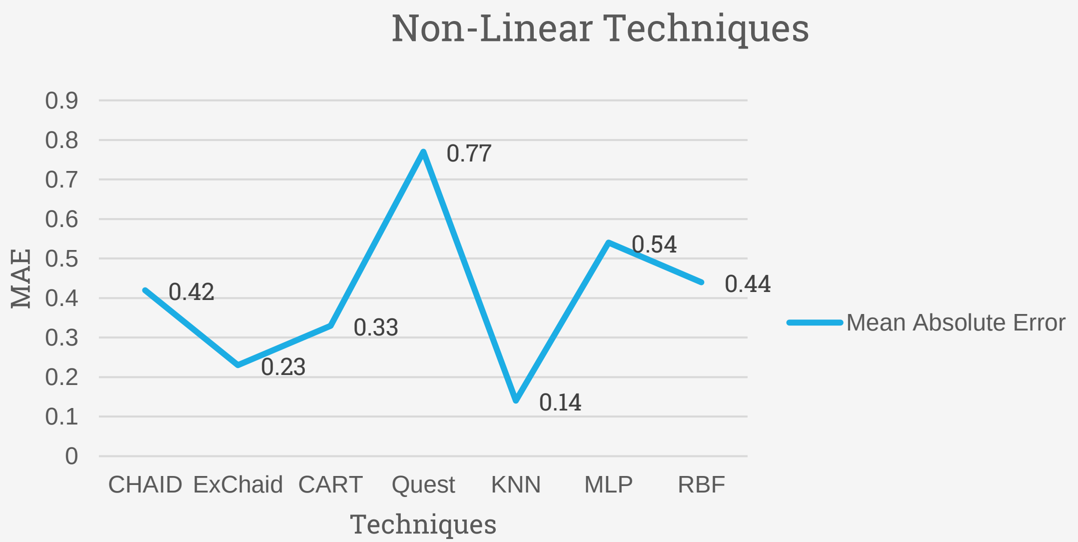
REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

To model the data in this project, various Linear and Non-Linear machine learning techniques have been used to predict the churn propensity which is a key driver in understanding if a customer will continue his business with firm or not.

As part of Linear machine learning technique, Logistic Regression and the variable selection methods like Forward Conditional, Forward Likelihood Ratio, Backward Conditional and Backward Likelihood Ratio have been used to model the data and calculate the R-squared value along with the Mean Absolute Error. Here is a snapshot of Linear Techniques:



As part of Non- Linear machine learning techniques, Decision Tree algorithms like CHAID, CART, QUEST. Neural Network algorithms like MLP & RBF along with KNN have been made use of.
Here is a snapshot of Non-Linear Techniques:

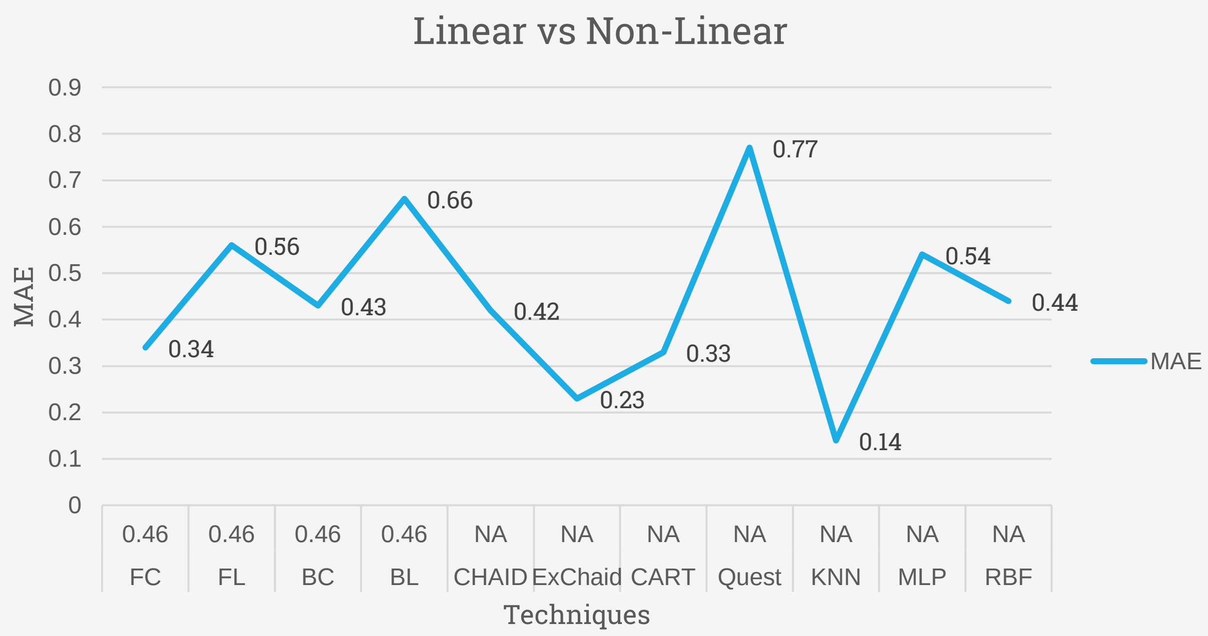


REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

Model Evaluation

CLTV is been predicted using various linear and non-linear modeling techniques. A deep analysis has been carried out to predict CLTV and then the related R-squared value along with the mean errors. Looking at the values of each of the techniques, it can be evaluated that the technique with a lower error value is the model that best fits the data for predicting CLTV. This can be used to make marketing decisions to target the high value customers who could increase the profit margins of the firm. Here are the outcomes of the model evaluation:

Comparison of the various modeling techniques based on the R-squared value and other key metrics.



Algorithm	R Squared	Mean Absolute Error	MAPE	Min. APE	Max. APE	RMSE
FC	0.46	0.34	19%	15%	25%	0.9
FL	0.46	0.56	27%	18%	30%	0.6
BC	0.45	0.43	30%	20%	35%	0.01
BL	0.45	0.66	45%	25%	50%	0.6
CHAID	NA	0.42	65%	33%	70%	0.3
ExChaid	NA	0.23	48%	30%	60%	0.2
CART	NA	0.33	36%	25%	45%	0.8
Quest	NA	0.77	38%	20%	50%	0.5
KNN	NA	0.14	8%	5%	13%	0.4
MLP	NA	0.54	33%	20%	40%	0.03
RBF	NA	0.44	32%	25%	45%	0.01

Model Deployment

As part of this project work, the modeling has been carried out using the data from csv file with python scripting on the Google Collab.

Modeling of the data has been carried out by making use of different Linear and Non-Linear machine learning techniques which are widely used across by different businesses.

A well planned deployment can be implemented once the business reviews and approves this study.

Results and Insights

The model with lower Mean Absolute Error is a good fit model for predicting the CLTV for the retail apparel firm. In this case, it is the KNN technique that has the lowest MAE compared to other linear and non-linear techniques.

The marketing team can now make use of the CLTV values to target high value customers and increase the sales.

Also, it is hard for the firms to target individual customers. We need to segment customers based on demographic data if it was available. Segmentation based on demographics could provide more insights into the customer profile to focus more on the customers.

Conclusion and Future Work

This project has been developed by using different linear and non-linear modeling techniques which provides an extensive insights into the CLTV of the customers for the data for a retail apparel firm that sells apparels in India.

The CLTV calculated or predicted helps the firm to take decision in terms of promotions and other offers that can be extended to their high value customers. The objective of predicting CLTV for retail firms is to ensure that the firms are in a state of mind to know who their high value customers are and can accordingly work on retaining them to drive higher profit margins.

The CLTV should start the shift from product-centric to customer-centric approach. With the increase in awareness of the CLTV metric and its benefits, the focus will be on its widespread adoption. Implementation of CLTV based strategies will give good insight to improve the customer experience

Some recommendations to increase customer Life Time Value are

1. Effective Communication
2. Loyalty Program
3. Retargeting

References

Burkard, K. (2020). *What is Customer Retention, and Why is it Important?* Smile Blog. https://cs-agents.com/blog/what-is-customer-activation-and-why-is-it-important/

CFI. (2019). *R-Squared - Definition, Interpretation, and How to Calculate*. https://corporatefinanceinstitute.com/resources/knowledge/other/r-squared/

Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. In *Family Medicine and Community Health* (Vol. 8, Issue 1). https://doi.org/10.1136/fmch-2019-000262

Customer Lifetime Value Prediction using Machine Learning | Addepto. (2019). https://addepto.com/customer-lifetime-value-prediction-machine-learning/

Hardie, B. (2006). Modeling Customer Lifetime Value. *Journal of Service Research*, *Volume 9*(Clv), 17. https://doi.org/10.1177/1094670506293810

How Retail Brands Can Predict Customer Lifetime Value. (2019). https://tinuiti.com/blog/ecommerce/customer-lifetime-value/

javatpoint. (2021). *Machine Learning Decision Tree Classification Algorithm - Javatpoint*. Javatpoint.Com. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

Manasson, A. (2019). *Why using CRISP-DM will make you a better Data Scientist | by Alexander Manasson | Towards Data Science*. https://towardsdatascience.com/why-using-crisp-dm-will-make-you-a-better-data-scientist-66efe5b72686

Pang, X., Zhou, Y., Wang, P., Lin, W., & Chang, V. (2020). An innovative neural network approach for stock market prediction. *Journal of Supercomputing*, *76*(3), 2098–2118. https://doi.org/10.1007/S11227-017-2228-Y

Predicting Customer Lifetime Value : A Definitive Guide. (2020). https://www.analyticsvidhya.com/blog/2020/10/a-definitive-guide-for-predicting-customer-lifetime-value-clv/

Sharapa, M. (2019).

“5 Simple Ways to Calculate Customer Lifetime Value | by Marie Sharapa | The Startup | Medium.” https://medium.com/swlh/5-simple-ways-to-calculate-customer-lifetime-value-5f49b1a12723 (August 6, 2022).

“Customer Lifetime Value Modelling.” https://dr.lib.iastate.edu/entities/publication/61e87bd9-a2ca-4fce-b7e5-5e196a332d48 (August 6, 2022).

“Customer Lifetime Value Prediction Using Machine Learning | Addepto.” https://addepto.com/customer-lifetime-value-prediction-machine-learning/ (March 23, 2022).

“Logistic Regression — Detailed Overview | by Saishruthi Swaminathan | Towards Data Science.” https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc (August 6, 2022).

“Predicting Customer Lifetime Value : A Definitive Guide.” https://www.analyticsvidhya.com/blog/2020/10/a-definitive-guide-for-predicting-customer-lifetime-value-clv/ (March 23, 2022).

Sharma, Shreya. 2021. “Customer Lifetime Value Modelling.”: 1–31.

Annexure

Additional Information | Plagiarism score

Customer Life Time Value with Machine Learning

ORIGINALITY REPORT

13%	12%	8%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	mafiadoc.com Internet Source	2%
2	ukcatalogue.oup.com Internet Source	2%
3	www.water.wa.gov.au Internet Source	2%
4	dr.lib.iastate.edu Internet Source	1%
5	Submitted to Mid-America Christian University Student Paper	1%
6	amfa.iau-arak.ac.ir Internet Source	1%
7	researcherslinks.com Internet Source	<1%
8	Submitted to University of Durham Student Paper	<1%
9	ieeexplore.ieee.org Internet Source	<1%

REVA UNIVERSITY

Bengaluru, India

Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

Thank you!