International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 9 Issue 8 ǁ 2021 ǁ PP. 36-48

Customer Lifetime Value Prediction and Segmentation using Machine Learning

Ms. Ramamani Venkatakrishna
REVA Academy of Corporate Excellence, Reva University, *Bengaluru, India*
Mr. Pradeepta Mishra
Director of AI, Lymbyc, LTI
Ms. Sneha P Tiwari
REVA Academy of Corporate Excellence, Reva University, *Bengaluru, India*

Abstract - *There is a fierce competition in the telecom sector that is prompting the companies to invest heavily on marketing including acquiring new customers. But to be truly profitable, it is crucial not only to attract new customers, but to make sure old customers are retained with the company for as long time as possible. This turns the focus on customer lifetime value (CLTV). Knowing what drives CLTV gives ideas of what is best to invest in, and this information can be very valuable for the telecom company in designing their marketing strategy [1]. Many companies are now considering changing their marketing approach from Product centric to Customer-centric. For this approach to work, it is essential to understand each customer's worth or value, which then helps focus the resources on targeted marketing. The purpose of this project is to Analyze the Customer sales data of the company and predict the Customer lifetime value. Based on the predicted CLTV, customer segmentation is done to determine focus groups. The goal of this project is to provide a guide for marketing decision making and planning marketing strategies and plans for future, using a machine learning models to predict customer lifetime values and segmentation.*

Keywords: *CLTV, Predictive modelling, Regression, Probabilistic models, Customer-centric, Product-centric, RFM, Recency, Frequency, Monetary*

Date of Submission: 14-08-2021 Date of acceptance: 29-08-2021

I. 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 gowth 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.

It is a well-known fact that acquisition of new customers is expensive and difficult. Focusing on existing customers and nurturing them is also not very simple. The health of n organization is often measure by the growth in CLTV. In short, an awareness of CLTV will help marketers to reach the most valued customers consistently. It will also help focus on enhancing the customer relationship either through better targeting, improvements to customer experience, or value-added services. Though CLTV is focus of most marketing circles, its application is not so prevalent and is major challenge. But with the evolution in technology and better access to insightful data, measuring the value of the customer is becoming more and more relevant and is increasing in priority.

Several factors account for the growing interest in the concept of CLTV. Marketing departments are coming under pressure and are held. Some of the more traditional marketing metrics like brand awareness and value, market share, sales growth does not often justify marketing investment. Second, financial metrics like stock price and total profit also do not measure the success of the business. Though these measures are useful, they have a limited diagnostic capability. Third, advances in IT have made it easy for organizations to collect huge number of transactional data, especially in the area of sales. This data can be analyzed and used to understand revealed preferences of customers rather than intentions. Furthermore, IT 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. These insights can be leveraged to customize marketing programs for focused group of customers or even individual customers again with the help of technology.

II. LITERATURE REVIEW

Marketing strategy is an ever evolving and constantly changing phenomena in any industry. Marketing planning and decision making in current digitized world is becoming more and more dependent on data driven. The focus of this project is CLTV and customer segmentation which are some of the key concepts that are now largely driven by analytics. The literature survey done to understand the past work in this area puts them into 5 distinct areas:

1.1 Change in market strategy

Both academic and marketing practitioner now accept the concept of regarding customers as assets and whose value should be measured and managed. Large organizaitons were traditionally built upon the product-centric approaches. With the change in market scenario, customers have evolved and become more aware. This is prompting companies to slowly shift from a product-centric towards a customer-centric approach. Despite this shift, product centric vs customer centric still remain the biggest dilemma that any big company has to deal with. But with the changes in competition and customer behaviour, customer-centric approach seems to be more effective in driving business growth. Mass-market techniques become less and less effective with the increase in competition and resulting increase in product availability and its variety. This makes firms to pay more attention to the markets rather than the products. With an increasing emphasis on customers, their segmentation a logical step towards effective marketing strategy. For customer-centric marketing to be successful, firms should define themselves as "customer specialists".[4]

The customer is at the centre of all strategies in a customer-centric company. They understand the segments in which their products are sold. Customer-centric companies not only focus on the solution they offer to the customer, they also articulate the whole customer journey to enhance customer experience. Their emphasis will be on building customer loyalty which will in turn result in the growth of their business. They will focus on maximizing the CLTV by turning every customer profitable. Other important aspect of Customer lifetime value is the role it plays in the valuation of the company thus influencing investor decisions.

1.2 Customer Lifetime Value (CLTV)

CLTV is a mainstay concept in Marketing Management for the past few decades. However, most of the literature on the topic is dedicated to extolling the use of CLTV as a decision-making criterion or considered it in the context of business profitability. It is also discussed for the important role it plays in customer acquisition/retention trade-offs and customer acquisition decisions.[5] It is important to measure CLTV as this can be used as a metric in evaluating marketing decisions. It is important for a firm to have an estimate of the customer’s lifetime value when the customer first starts doing business with them, and at each of their subsequent purchases.[6]

1.3 Customer Lifetime Value measurement

Customer lifetime value (CLTV) is increasing being accepted as a metric that will help understand the how effectively the company can acquire, grow, and retain their most profitable customers. In view of this, a study of literature on some of the best practices in this area of customer value management along with how it can be adopted in different business conditions was conducted. Adoption of this concept in real life has always been a challenge. So, it is important to understand what organizational and implementation challenges that an firm faces during the implementation of customer value management by the firm. The major difficulty for a firm is in coming up with an optimal blend of the differential levels of treatments of different customers. This will help the firm in maximising the profits earned from each customer over that customers’ lifetime.[7]

1.4 Data Analytics to predict Customer Lifetime Value

Nicolas Glady, Bart Baesens, Christophe Croux in their paper “Modeling churn using customer lifetime value” explore Customer lifetime value as an important concept churn prediction in a non-contractual setting. [8]. CLTV is computed based on the predicted future number of transactions of a customer and the profit from these transactions. Probabilistic models like Pereto/NBD and BG/NBD are powerful techniques for predicting the future activity of a customer.[9] There are many studies on using regression techniques for prediction of CLTV using real customer transaction datasets. Data sets based on RFM segmentation are used most commonly for building CLTV prediction models. In these models’ value of customers are measured based on companies’ historical sales records.[10]

1.5 CLTV and Customer Segmentation

Different industry segments adopt different ways of calculating customer lifetime value. There have been some work on the way CLTV is calculated and used the telecommunications industry[11]. Usage of transactional data in calculating CLTV and predictive modelling, performing customer segmentation with the result of customer value derived has been an area of extensive study. This helps in developing marketing strategies based upon the result of customer analysis.

III. OBJECTIVE

CLTV is a key matrix for any customer centric business model. It tells us how valuable a customer is to the business and will help us determine the total worth of the customer throughout the course of their relationship. The purpose of this project is to Analyze the Sales of the company and predict the Customer lifetime value. Customer segmentation based on the predicted CLTV will help focus marketing resources effectively. By identifying customer segments based on their value, company can use data insights to group the most valuable customers. This will help in assessing the customers loyalty, their projecting revenue. Effective strategies can be devised to nurture the customer who are most profitable to the company. Right products and services can be pushed based on the context and relevance to the customer.

IV. PROJECT METHODOLOGY

CRISP-DM methodology was used to execute this project structurally, where all the project activities were grouped into the following pre-define steps.

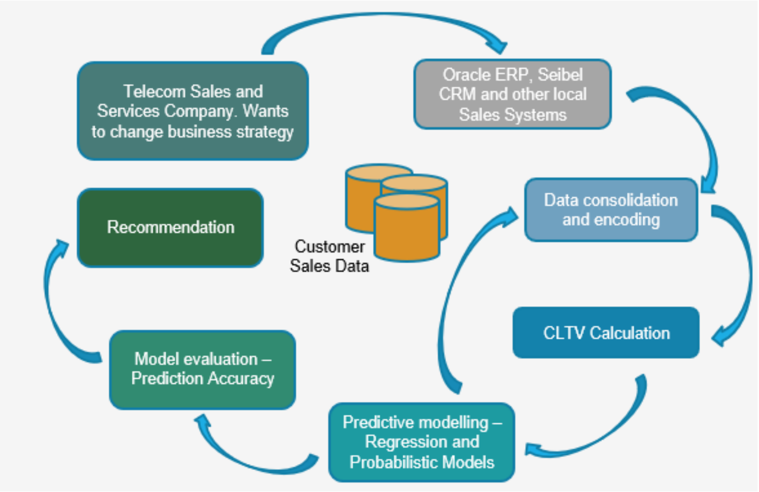

Telecom Sales and Services Company. Wants to change business strategy
Oracle ERP, Seibel CRM and other local Sales Systems
Data consolidation and encoding
CLTV Calculation
Predictive modelling – Regression and Probabilistic Models
Model evaluation – Prediction Accuracy
Recommendation
Customer Sales Data

Figure 1: Project Methodology

V. BUSINESS UNDERSTANDING

The telecommunications industry is critical in Alaska. It provides an essential link between people in all regions, no matter how remote, which is important in helping communities and businesses move forward together. None of this infrastructure would exist without quality equipment and support. According to the Alaska Telecom Association, which was first founded in 1949, Alaska's economic future is intrinsically linked to the availability of communications services. There are a number of telecommunications companies that work hard to keep Alaska's networks functioning reliably and the company chosen for this project is one such. The company is based in Anchorage and along with its partners and subsidiaries distributes a range of telecommunications equipment from mobile phones wi-fi solutions and extenders to internet and TV equipment to phone accessories. It also provided telecom services like mobile, internet and TV network and has a wide reach in the remote areas of Alaska. It is currently enjoying about 30% market share in Alaska region.

The company has an annual revenue of $9.66 million in sales (USD). It has a customer base of over 4000, both within and outside Alaska. Though bulk of its revenue is from Alaskan region, it has seen an impressive growth outside Alaska in the past 5-6 years. Customer base is broadly classified into Big Corporates, Small businesses and Retail outlets depending on the market segment they cater to. For the past year the segment catering to the retail showrooms has seen tremendous amount of competition and there is a slight downward trend in their sales revenues. The company is currently planning on changing its marketing strategy to address this problem.

VI. DATA UNDERSTANDING

Data is the most important component that will help understand customer behavior and gain insight. But sourcing the right information for the correct timeframe and applying it in the right manner remains a challenge. Collection of proper data (made even more complicated because of decentralized data storage) and the application of this information is the key to the success of the project.

Following data is analyzed and consolidated into sales data for a one-year time-period:

- Customer base
- Product listing
- Customer Contracts and Orders
- Invoicing data
- Customer Payments

The data set contains all the sales transactions for 1 year between 01/01/2020 to 31/12/2020. The period of 1 year was chosen to make sure we account for all the seasonal changes.

Attribute Information:

- Customer_ID: A unique identification number for each customer. Nominal, 5 digit integer.
- Invoice_Number: Invoice number. Nominal, a 6-digit integer that uniquely identifies a transaction.
- Item_number: Product or Item code. Nominal, a 5/6 character alpha-numeric uniquely identifies the product.
- Quantity: The quantity of each in the transaction. Numeric.
- Invoice_Dt: Invoice Date. Date, Transaction date.
- Unit_Price: Unit price. Numeric, Product selling price per unit in Dollars.
- Cust_Cat: Customer Category. Each Customer is identified as Wholesale Dealers, Oversees Customers and Retail Showrooms

Analysis was done to understand data. Category wise sales trend as shown in the figure 2 below indicates that sales are decreasing for category 3 customers. This customer segments needs extra focus in marketing.

.

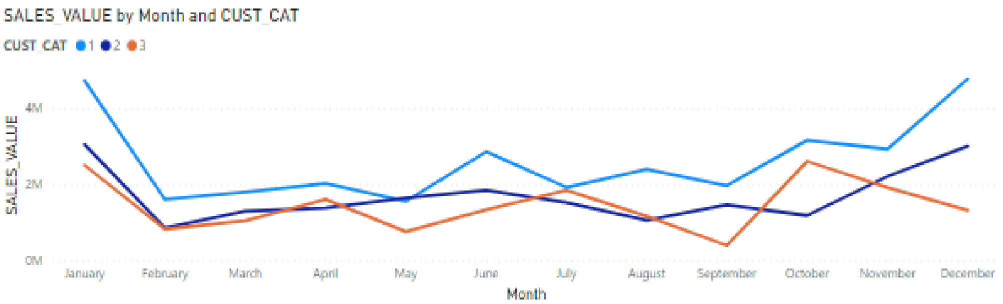


Figure 2 Category wise Sales Trend

VII. Data Preparation

7.1 Missing value handling

Missing values were seen in Customer_ID, Invoice_Number, Item_number and Quantity. Missing values for Item_number were imputed with Mode and Quantity with Mean, Invoice_number with a random 6-digit unique number. Records with missing Customer_ID had to be discarded (~2%)

7.2 Outlier handling

Unit_Price had few Outliers in terms of 0s and were removed. Outliers on total sales value were removed so that the data is not skewed.

7.3 CLTV Calculation

Customer Lifetime Value is generally measured by computing the total worth or value of a customer to the business during the whole period of their relationship. It considers the total revenue a customer brings in along with the company's expected customer lifespan. There are many ways of calculating CLTV based on the type of business, the industry segment, etc., and is decided based on the business processes of the company and the different factors that influence the company's revenue. One of the most popular and simple way of calculating CLTV involves the following matrices:

- Average Purchase Value = Total Sales Revenue/Total number of Orders
- Purchase Frequency = Total number of Orders/Number of Customers
- Average Customer Life Span = Average number of years a customer continues with the company.
- This is the CLTV Calculation formula used
- CLTV = Average Purchase Value * Purchase Frequency * Average Customer Life Span

CLTV was calculated for all customers based on the above formula. This will be used as Target variable for Prediction.

7.4 Feature Engineering

With only the sales data available, the challenge was to Identify the features that should be used to predict future and create them. RFM scores for each customer ID along with the customer category was used as features to predict CLTV. RFM stands for Recency - Frequency - Monetary Value. RFM model allows the business to segment the customers based on the above three criteria calculated using existing customer's transaction history. Using RFM we will have segments of Low, Mid and High Value based on the Recency, Frequency and Revenue scores generated. Once RFM Clustering is done, these segments were used as features.

7.4.1 Recency

Recency means when was the last time the customer purchased a product/service from the company. To calculate recency, based on the most recent purchase date of each customer the number of inactive days was calculated. Then, K-means clustering was applied to assign customers a recency score. Figure 8.1 below show the distribution of recency across customer. Based on the inertia value, number of clusters considered for Recency was 3. Figure 8.2 below shows the Recency segmentation.

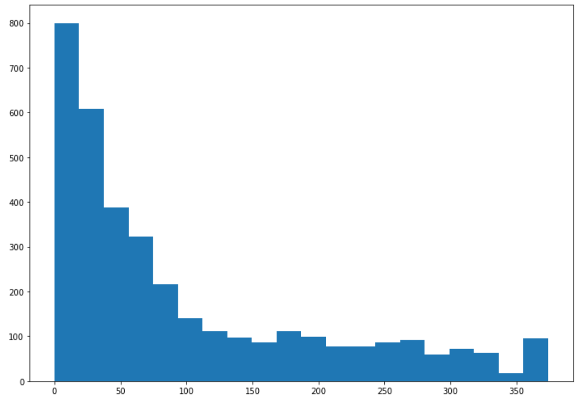
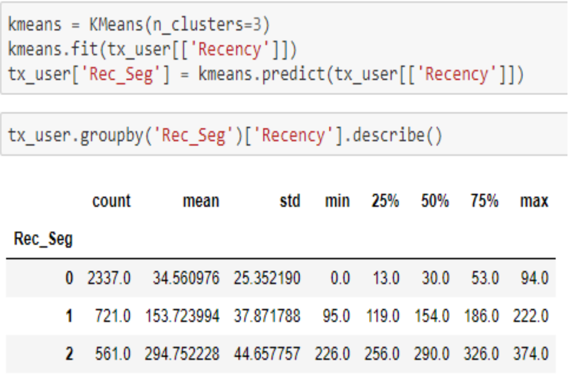


Figure 3 Recency histogram



```
kmeans = KMeans(n_clusters=3)
kmeans.fit(tx_user[['Recency']])
tx_user['Rec_Seg'] = kmeans.predict(tx_user[['Recency']])
```

```
tx_user.groupby('Rec_Seg')['Recency'].describe()
```

Rec_Seg	count	mean	std	min	25%	50%	75%	max
0	2337.0	34.560976	25.352190	0.0	13.0	30.0	53.0	94.0
1	721.0	153.723994	37.871788	95.0	119.0	154.0	186.0	222.0
2	561.0	294.752228	44.657757	226.0	256.0	290.0	326.0	374.0

Figure 4 Recency Clusters

7.4.2 Frequency

Frequency means how often the customer purchases in a year or any fixed time period. This is calculated based on the total number of orders for the customer. Figure 8.3 below shows the distribution of frequency. Based on the inertia value, number of clusters considered for Frequency was 3. Figure 8.4 below shows the Frequency segmentation.

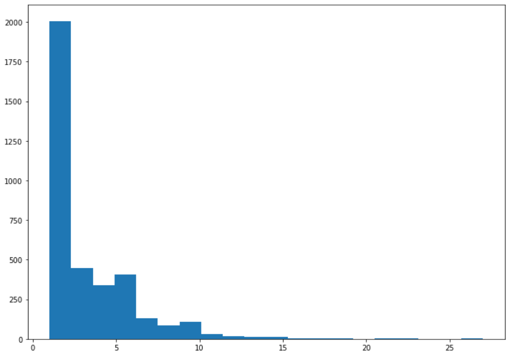
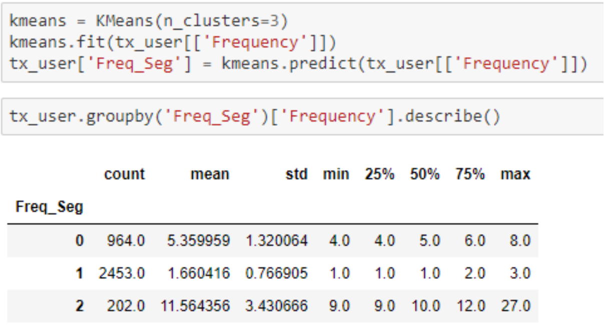


Figure 5 Frequency histogram



```
kmeans = KMeans(n_clusters=3)
kmeans.fit(tx_user[['Frequency']])
tx_user['Freq_Seg'] = kmeans.predict(tx_user[['Frequency']])
```

```
tx_user.groupby('Freq_Seg')['Frequency'].describe()
```

	count	mean	std	min	25%	50%	75%	max
Freq_Seg								
0	964.0	5.359959	1.320064	4.0	4.0	5.0	6.0	8.0
1	2453.0	1.660416	0.766905	1.0	1.0	1.0	2.0	3.0
2	202.0	11.564356	3.430666	9.0	9.0	10.0	12.0	27.0

Figure 6 Frequency Clusters

7.4.3 Revenue

Revenue or the monetary value means the amount of money the customer has spent on the company's products and services. This is calculated based on the total sales value of the customer. Figure 8.5 below show the distribution of customer sales value. Based on the inertia value, number of clusters considered for Frequency was 3. Figure 8.6 below shows the Revenue segmentation.

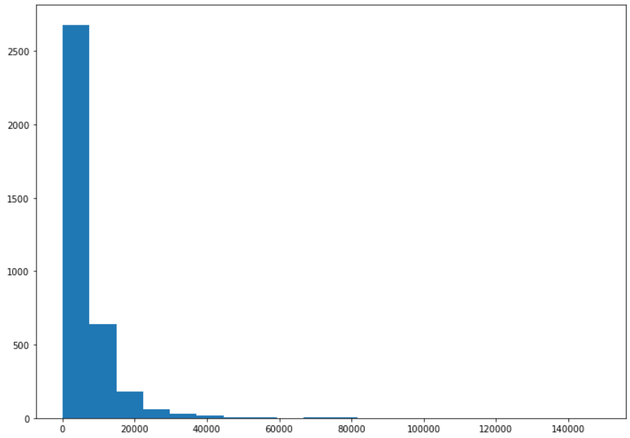
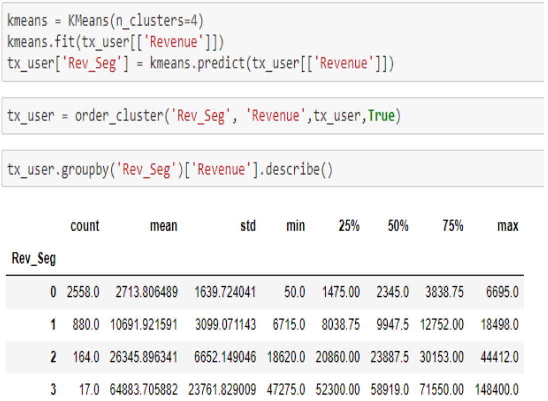


Figure 7 Revenue histogram



```
kmeans = KMeans(n_clusters=4)
kmeans.fit(tx_user[['Revenue']])
tx_user['Rev_Seg'] = kmeans.predict(tx_user[['Revenue']])
```

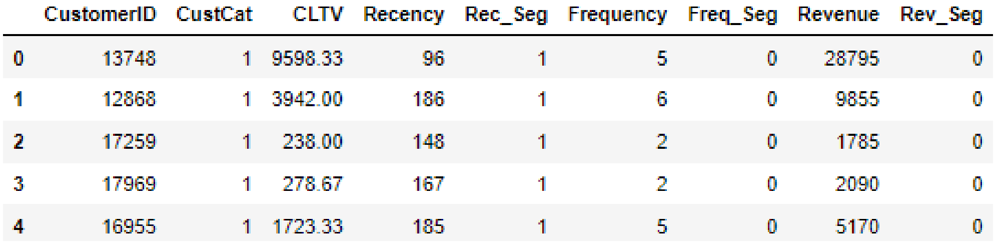
```
tx_user = order_cluster('Rev_Seg', 'Revenue',tx_user,True)
```

```
tx_user.groupby('Rev_Seg')['Revenue'].describe()
```

	count	mean	std	min	25%	50%	75%	max
Rev_Seg								
0	2558.0	2713.806489	1639.724041	50.0	1475.00	2345.0	3838.75	6695.0
1	880.0	10691.921591	3099.071143	6715.0	8038.75	9947.5	12752.00	18498.0
2	164.0	26345.896341	6652.149046	18620.0	20860.00	23887.5	30153.00	44412.0
3	17.0	64883.705882	23761.829009	47275.0	52300.00	58919.0	71550.00	148400.0

Figure 8 Revenue Clusters

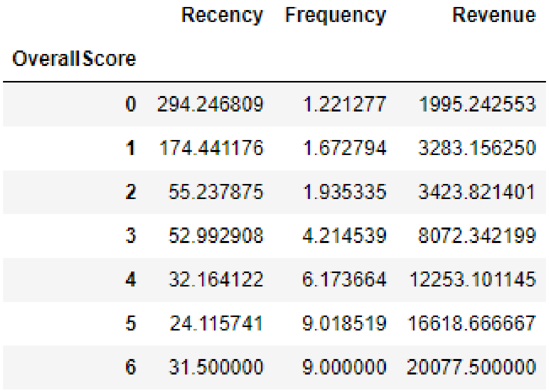
Figure 9 below shows the RFM segmentation. Overall RFM scores for every customer is calculated based on these segments



	CustomerID	CustCat	CLTV	Recency	Rec_Seg	Frequency	Freq_Seg	Revenue	Rev_Seg
0	13748	1	9598.33	96	1	5	0	28795	0
1	12868	1	3942.00	186	1	6	0	9855	0
2	17259	1	238.00	148	1	2	0	1785	0
3	17969	1	278.67	167	1	2	0	2090	0
4	16955	1	1723.33	185	1	5	0	5170	0

Figure 9 Data snippet after RFM Scoring

Overall Score is calculated by averaging the recency, frequency, and monetary scores together, then sort customers from highest to lowest to find the most valuable customers. Figure 8.8 below shows the overall scoring for the customers.



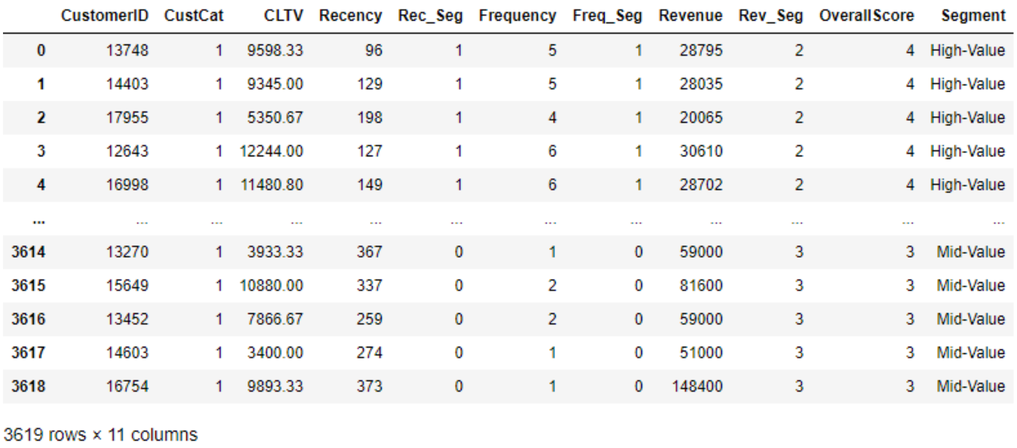
	Recency	Frequency	Revenue
OverallScore			
0	294.246809	1.221277	1995.242553
1	174.441176	1.672794	3283.156250
2	55.237875	1.935335	3423.821401
3	52.992908	4.214539	8072.342199
4	32.164122	6.173664	12253.101145
5	24.115741	9.018519	16618.666667
6	31.500000	9.000000	20077.500000

Figure 10 Overall Score

Based on the scoring as shown above in figure 8.8, we can conclude that the customers with score 6 are our best customers and 0 are the worst. These are binned into 3 segments

- 0 to 1: Low Value
- to 3: Mid Value
- 4+: High Value

Once the RFM segmentation is completed, these along with the Customer category were considered as features. Figure 8.9 shows the prepared data with all the features. This dataset is used for CLTV prediction and segmentation based on CLTV.



	CustomerID	CustCat	CLTV	Recency	Rec_Seg	Frequency	Freq_Seg	Revenue	Rev_Seg	OverallScore	Segment
0	13748	1	9598.33	96	1	5	1	28795	2	4	High-Value
1	14403	1	9345.00	129	1	5	1	28035	2	4	High-Value
2	17955	1	5350.67	198	1	4	1	20065	2	4	High-Value
3	12643	1	12244.00	127	1	6	1	30610	2	4	High-Value
4	16998	1	11480.80	149	1	6	1	28702	2	4	High-Value
...	...	...	...	...	...	...	...	...	...	...	...
3614	13270	1	3933.33	367	0	1	0	59000	3	3	Mid-Value
3615	15649	1	10880.00	337	0	2	0	81600	3	3	Mid-Value
3616	13452	1	7866.67	259	0	2	0	59000	3	3	Mid-Value
3617	14603	1	3400.00	274	0	1	0	51000	3	3	Mid-Value
3618	16754	1	9893.33	373	0	1	0	148400	3	3	Mid-Value

3619 rows × 11 columns

Figure 11 Snippet of prepared data

VIII. DATA MODELING

8.1 CLTV Prediction

With businesses increasingly adopting CLTV for their marketing planning, both industry and research efforts have increased to help shape may different methods for CLTV estimations. Both statistical and machine learning techniques are being is used extensively for the purpose. Some of the methods using machine learning models are:

Clustering Models: CLTV estimation can be thought of as a clustering problem where the model simply learns how to distinguish between different groups of users, depending on the features that represent them. Here features generally go beyond purchase history and will customers demographic and purchasing characteristics are considered. But predicting exact CLTV value for each customer cannot be done using clustering method. K-means is a popular clustering technique used for CLTV clustering.

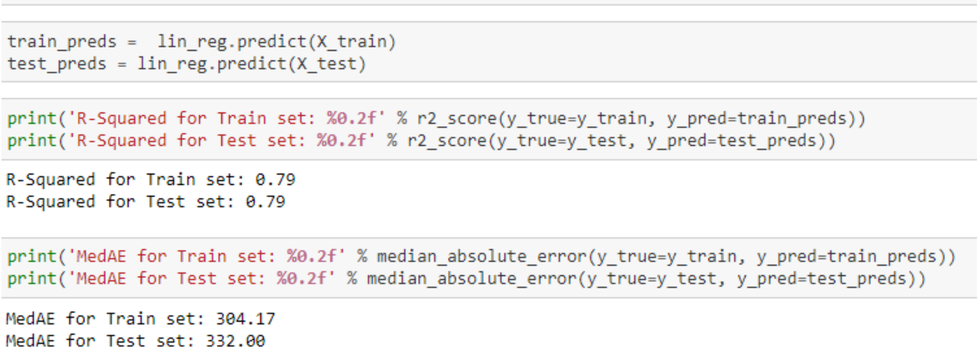
Multi-class Classification: CLTV estimation can also be tackled as a multi-class classification problem when there is a labelled dataset. Random forest is a very powerful approach used for multi-class classification.

Regression Models: Formulating the CLTV estimation as a clustering or a multi-class classification problem may not be the most appropriate way since the problem is naturally a regression problem in which the goal is to estimate a continuous value. In regression models the models are trained to learn and predict continuous values.

Calculated CLTV is used as the target variable and the data set is divided into train and test sets. Training set is used to train the model and the model is validated on the test set. This helps in understanding the model accuracy.

8.1.1 Linear Regression

One of the most well-known and powerful algorithms in machine learning is Linear regression. This algorithm fits a linear equation on the observed dataset to model the relationship between two variables. The central idea is to generate a line that best fits the data. The line which has the least total prediction error is the best fit line. The distance between each data point to the regression line is used to calculate the error. Figure 10.1 below shows the accuracy of Linear regression model on the dataset.



```
train_preds =  lin_reg.predict(X_train)
test_preds = lin_reg.predict(X_test)
```

```
print('R-Squared for Train set: %0.2f' % r2_score(y_true=y_train, y_pred=train_preds))
print('R-Squared for Test set: %0.2f' % r2_score(y_true=y_test, y_pred=test_preds))
```

```
R-Squared for Train set: 0.79
R-Squared for Test set: 0.79
```

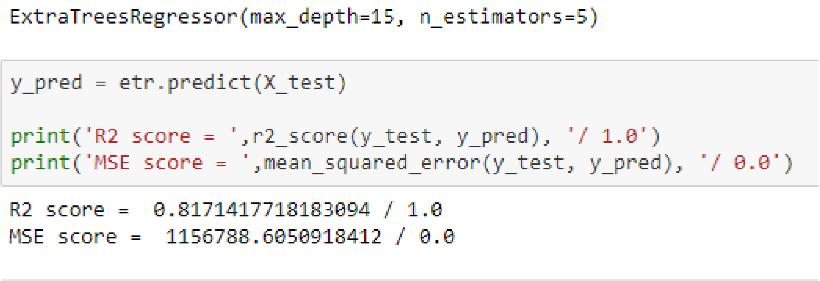
```
print('MedAE for Train set: %0.2f' % median_absolute_error(y_true=y_train, y_pred=train_preds))
print('MedAE for Test set: %0.2f' % median_absolute_error(y_true=y_test, y_pred=test_preds))
```

```
MedAE for Train set: 304.17
MedAE for Test set: 332.00
```

Figure 12 Accuracy of Linear Regression model

8.1.2 Extra Trees Regressor

Extremely Randomized Trees (or Extra-Trees) is an ensemble learning method. This method creates extra trees randomly in sub-samples of datasets to improve the predictivity. By this approach, the method reduces the variance. The method averages the outputs from the decision trees and hence overfitting is reduced. Figure 10.2 below shows the accuracy of Extra Tress Regressor model on the dataset.



```
ExtraTreesRegressor(max_depth=15, n_estimators=5)
```

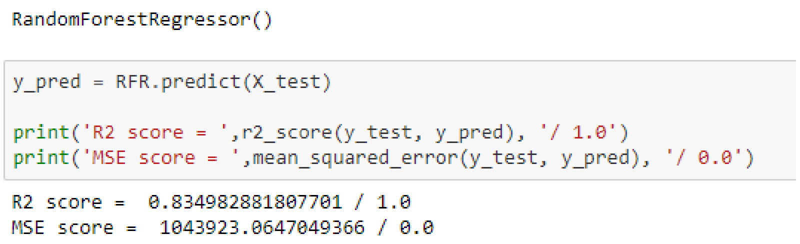
```
y_pred = etr.predict(X_test)

print('R2 score = ',r2_score(y_test, y_pred), '/ 1.0')
print('MSE score = ',mean_squared_error(y_test, y_pred), '/ 0.0')
```

```
R2 score =  0.8171417718183094 / 1.0
MSE score =  1156788.6050918412 / 0.0
```

Figure 13 Accuracy of ExtraTree model

8.1.3 Random Forest Regressor

A random forest fits a number of classifying decision trees on sub-samples of the dataset and uses averaging for prediction. Since it used multiple instances of the data, predictive accuracy is improved, and over-fitting avoided. Accuracy of the model can be improved by changing the parameters. Figure 10.3 below shows the accuracy of Random Forest Regressor model on the dataset.



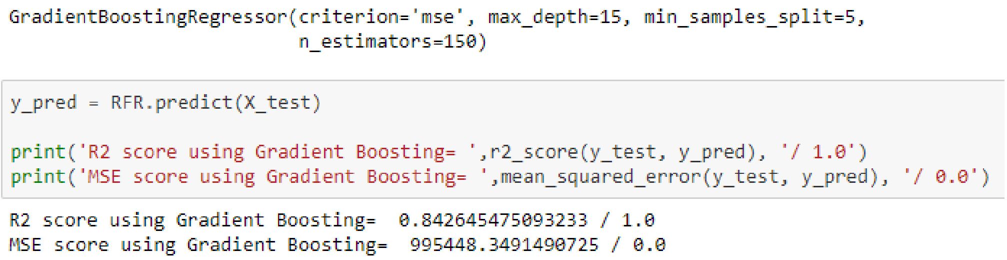
```
RandomForestRegressor()

y_pred = RFR.predict(X_test)

print('R2 score = ',r2_score(y_test, y_pred), '/ 1.0')
print('MSE score = ',mean_squared_error(y_test, y_pred), '/ 0.0')

R2 score =  0.834982881807701 / 1.0
MSE score =  1043923.0647049366 / 0.0
```

Figure 14 Accuracy of Random Forest model

8.1.4 Gradient Boosting Regressor

Gradient boosting is a very powerful technique for building predictive models. It involves three elements:

- An optimized loss function.
- A weak learner to make predictions.
- An additive model to add weak learners to minimize the loss function.
- Figure 10.4 below shows the accuracy of Gradient Boosting Regressor model on the dataset.
-

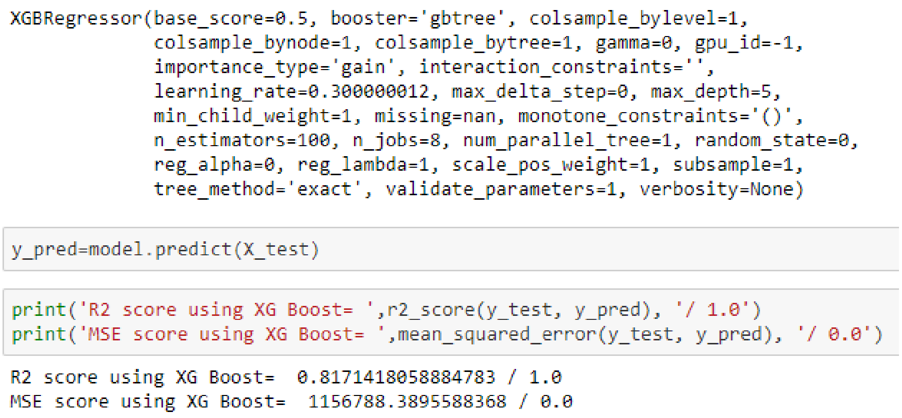


```
GradientBoostingRegressor(criterion='mse', max_depth=15, min_samples_split=5,
                          n_estimators=150)

y_pred = RFR.predict(X_test)

print('R2 score using Gradient Boosting= ',r2_score(y_test, y_pred), '/ 1.0')
print('MSE score using Gradient Boosting= ',mean_squared_error(y_test, y_pred), '/ 0.0')

R2 score using Gradient Boosting=  0.842645475093233 / 1.0
MSE score using Gradient Boosting=  995448.3491490725 / 0.0
```

Figure 15 Figure 15 Accuracy of Gradient Boosting model

8.1.5 XGBOOST

XGBoost is a powerful approach for building supervised regression models. It is an ensemble learning method and provides an efficient and effective implementation of the gradient boosting algorithm. Advantages of XGBoost model are its ease of use, Model Accuracy and Feasibility, and computational efficiency. Figure 10.5 below shows the accuracy of XGBoost Regressor model on the dataset.



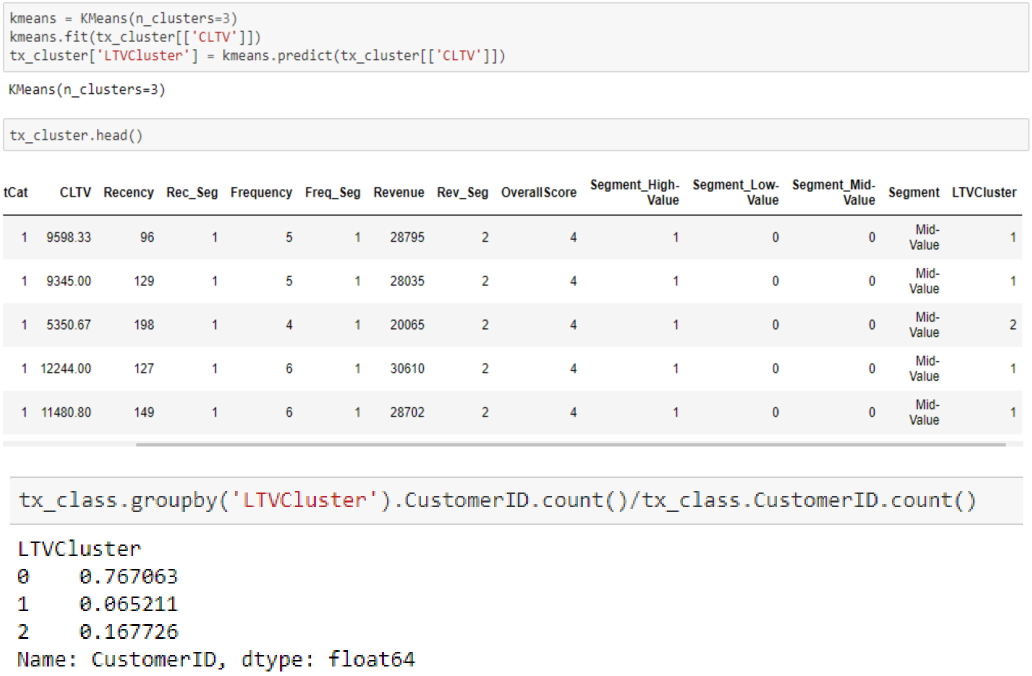
```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
             importance_type='gain', interaction_constraints='',
             learning_rate=0.300000012, max_delta_step=0, max_depth=5,
             min_child_weight=1, missing=nan, monotone_constraints='()',
             n_estimators=100, n_jobs=8, num_parallel_tree=1, random_state=0,
             reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
             tree_method='exact', validate_parameters=1, verbosity=None)

y_pred=model.predict(X_test)

print('R2 score using XG Boost= ',r2_score(y_test, y_pred), '/ 1.0')
print('MSE score using XG Boost= ',mean_squared_error(y_test, y_pred), '/ 0.0')

R2 score using XG Boost=  0.8171418058884783 / 1.0
MSE score using XG Boost=  1156788.3895588368 / 0.0
```

Figure 16 Accuracy of XGBoost model

Based on the above prediction methods accuracy of Gradient boosting model was selected for CLTV prediction.

8.2 Customer Segmentation

CLTV prediction gives total customer values for each customer. This will be useful to the marketing team as it gives them specific customers to focus on. But to strategize and build marketing plan, it is necessary to understand the characteristics of different segments of customers, not just a single customer. Hence it becomes important to build clusters or segments which makes it more actionable. Based on the predicted CLTV, groups were formed using K-means clustering technique.



```
kmeans = KMeans(n_clusters=3)
kmeans.fit(tx_cluster[['CLTV']])
tx_cluster['LTVCluster'] = kmeans.predict(tx_cluster[['CLTV']])
```

```
KMeans(n_clusters=3)
```

```
tx_cluster.head()
```

tCat	CLTV	Recency	Rec_Seg	Frequency	Freq_Seg	Revenue	Rev_Seg	OverallScore	Segment_High-Value	Segment_Low-Value	Segment_Mid-Value	Segment	LTVCluster
1	9598.33	96	1	5	1	28795	2	4	1	0	0	Mid-Value	1
1	9345.00	129	1	5	1	28035	2	4	1	0	0	Mid-Value	1
1	5350.67	198	1	4	1	20065	2	4	1	0	0	Mid-Value	2
1	12244.00	127	1	6	1	30610	2	4	1	0	0	Mid-Value	1
1	11480.80	149	1	6	1	28702	2	4	1	0	0	Mid-Value	1

```
tx_class.groupby('LTVCluster').CustomerID.count()/tx_class.CustomerID.count()
```

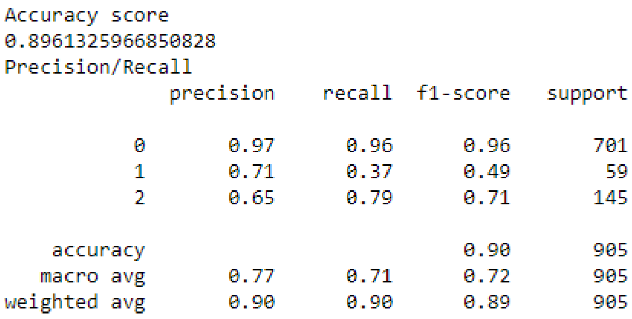
```
LTVCluster
0    0.767063
1    0.065211
2    0.167726
Name: CustomerID, dtype: float64
```

Figure no. 17 Clustering with K-means

The figure 10.7 shows the customer segmentation using K-means. Segment 0 with 76.7% customers are the low value customers and segment 2 with close to 16% customers are the high value ones. This segment value is used as the target variable and train and test data sets are created. Using classification algorithms, CLTV based customer segmentation is achieved.

8.2.1 Logistic Regression

Logistic Regression classification determines the probability that the target variable belongs to a certain class as a function of a linear summation of all the features. A logistic function is used which will model the probability of possible outcomes in this algorithm. Logistic regression is mostly used to analyse the how much the multiple dependent variables influence a single outcome variable and is designed for classification. For this algorithm to work the predicted variable should be binary and the assumption is that all predictors are independent of each other. Also, this model work best if the data is free of missing values. 89.6% accuracy was achieved using logistic regression model on the dataset. This is not very high though segment 0 positive precision is quite high at 97%.

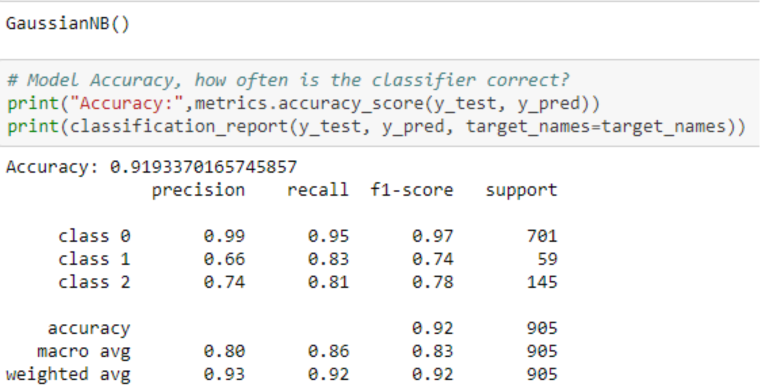


```
Accuracy score
0.8961325966850828
Precision/Recall
              precision    recall  f1-score   support

           0       0.97      0.96      0.96       701
           1       0.71      0.37      0.49        59
           2       0.65      0.79      0.71       145

    accuracy                           0.90       905
   macro avg       0.77      0.71      0.72       905
weighted avg       0.90      0.90      0.89       905
```

Figure 18 Model Accuracy for Logistic Regression

8.2.2 Naive Bayes

Naive Bayes algorithm assumes that every pair of features are of independent if each other. This model works well in many of the real-world situations like spam filtering and document classification. A very small amount of training data is sufficient to estimate the necessary parameters in this technique. Compared to other sophisticated methods, Naive Bayes classifiers are very fast, but results may not always be accurate. Figure 10.9 below shows the accuracy (92%) of Naive Bayes model on our data. Segment 0 precision at 99% is very good and this model seems to be best for segmentation.



```
GaussianNB()

# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
print(classification_report(y_test, y_pred, target_names=target_names))

Accuracy: 0.9193370165745857
              precision    recall  f1-score   support

     class 0       0.99      0.95      0.97       701
     class 1       0.66      0.83      0.74        59
     class 2       0.74      0.81      0.78       145

    accuracy                           0.92       905
   macro avg       0.80      0.86      0.83       905
weighted avg       0.93      0.92      0.92       905
```

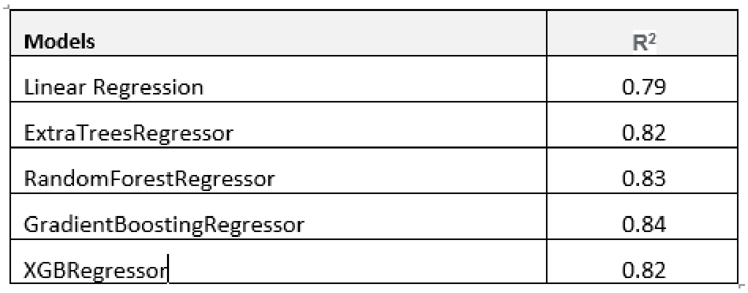
Figure 19 Model Accuracy for Naive Bayes

IX. MODEL EVALUATION

Accuracy of a machine learning model is measured by checking how good a fit they can achieve with the given data. There are various matrices based on which we can determine the model that best fits the data.

9.1 CLTV Prediction

In a regression model, once the regression like fit, the distance of the data point to this line is measured, based on which R-squared is calculated. The closer the data points are to the fitted regression line, better the model. This statistic gives a numeric measure of the goodness-of-fit of a model. This matrix is considered to determine the best fit model for CLTV predictions. Figure 20 below gives comparative scores for each of the models.



Models	R^2
Linear Regression	0.79
ExtraTreesRegressor	0.82
RandomForestRegressor	0.83
GradientBoostingRegressor	0.84
XGBRegressor	0.82

Figure 20 Regression Model Evaluation Matrix

Above table clearly shows Gradient Boosting Regressor is the best among all the Machine learning techniques for predicting CLTV.

9.2 Customer Segmentation

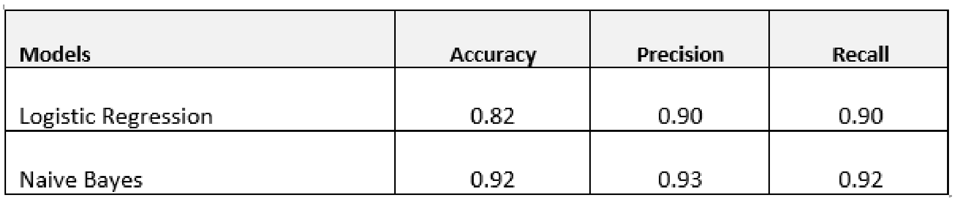
For classification models, apart from accuracy, precision and recall are also important matrices that should be considered before selecting the best fit model.

Accuracy: The ratio of the number of correct predictions to the total number of predictions give accuracy of the model.

Precision: Precision is the ratio of True Positives to Total Positives.

Recall: Recall is the measure that shows if the model is correctly identifying True Positives. Recall also gives a measure of how good our model is in identifying the relevant data.

Figure 21 below shows these matrices for both the models used for Customer classification. Based on these scores, Naïve Bayes model is found better suited for segmentation.



Models	Accuracy	Precision	Recall
Logistic Regression	0.82	0.90	0.90
Naive Bayes	0.92	0.93	0.92

Table 21 Classification Model Evaluation Matrix

X. CONCLUSION

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. For the purpose of this project, Gradient Boosting Regressor model is considered for CLTV prediction based on the prediction accuracy.

Some recommendations to increase customer Life Time Value are

- Effective Communication
- Loyalty Program
- Retargeting

It is seen from customer segmentation based on predicted CLTV, that about 17% customers contribute to almost 50% of the Value. This is the segment that should be targeted by the marketing team. These customers should be nurtured so that they continue with the company and efforts should be made to increase their CLTV.

XI. REFERENCES

[1]. Flordahl, P., & Friberg, J. (2013). Modeling Customer Lifetime Value in the Telecom Industry.

[2]. Chamberlain, B. P., Cardoso, A., Liu, C. B., Pagliari, R., & Deisenroth, M. P. (2017, August). Customer lifetime value prediction using embeddings. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1753-1762)

[3]. Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., ... & Sriram, S. (2006). Modeling customer lifetime value. Journal of service research, 9(2), 139-155.

[4]. Sheth, J. N., Sisodia, R. S., & Sharma, A. (2000). The antecedents and consequences of customer-centric marketing. Journal of the Academy of marketing Science, 28(1), 55-66.

[5]. Berger, P. D., & Nasr, N. I. (1998). Customer lifetime value: Marketing models and applications. Journal of interactive marketing, 12(1), 17-30.

[6]. Borle, S., Singh, S. S., & Jain, D. C. (2008). Customer lifetime value measurement. Management science, 54(1), 100-112.

[7]. Kumar, V., Ramani, G., & Bohling, T. (2004). Customer lifetime value approaches and best practice applications. Journal of interactive Marketing, 18(3), 60-72.

[8]. Glady, N., Baesens, B., & Croux, C. (2009). Modeling churn using customer lifetime value. European Journal of Operational Research, 197(1), 402-411.

[9]. Glady, N., Baesens, B., & Croux, C. (2009). A modified Pareto/NBD approach for predicting customer lifetime value. Expert Systems with Applications, 36(2), 2062-2071.

[10]. 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.

[11]. Lu, J., & Park, O. (2003). Modeling customer lifetime value using survival analysis—an application in the telecommunications industry. Data Mining Techniques, 120-128.

[12]. Mauricio, A. P., Payawal, J. M. M., Cueva, M. A. D., & Quevedo, V. C. (2016, May). Predicting customer lifetime value through data mining technique in a direct selling company. In 2016 International Conference on Industrial Engineering, Management Science and Application (ICIMSA) (pp. 1-5). IEEE.

[13]. Honga, L. I. U., & Weib, X. I. E. (2008). Research on Churner Based on Customer Lifetime Value. Journal of Hefei University of Technology (Social Sciences), 06.

[14]. Kim, S. Y., Jung, T. S., Suh, E. H., & Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. Expert systems with applications, 31(1), 101-107.

[15]. Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert systems with applications, 26(2), 181-188.

[16]. Lemon, K. N., & Mark, T. (2006). Customer lifetime value as the basis of customer segmentation: Issues and challenges. Journal of Relationship Marketing, 5(2-3), 55-69.

[17]. Kahreh, M. S., Tive, M., Babania, A., & Hesan, M. (2014). Analyzing the applications of customer lifetime value (CLV) based on benefit segmentation for the banking sector. Procedia-Social and Behavioral Sciences, 109, 590-594.

[18]. Pramono, P. P., Surjandari, I., & Laoh, E. (2019, July). Estimating customer segmentation based on customer lifetime value using two-stage clustering method. In 2019 16th International Conference on Service Systems and Service Management (ICSSSM) (pp. 1-5). IEEE.

[19]. Cuadros, A. J., & Domínguez, V. E. (2014). Customer segmentation model based on value generation for marketing strategies formulation. Estudios Gerenciales, 30(130), 25-30.

[20]. Khajvand, M., & Tarokh, M. J. (2011). Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. Procedia Computer Science, 3, 1327-1332.

[21]. Cheng, C. J., Chiu, S. W., Cheng, C. B., & Wu, J. Y. (2012). Customer lifetime value prediction by a Markov chain based data mining model: Application to an auto repair and maintenance company in Taiwan. Scientia Iranica, 19(3), 849-855.

[22]. Kumar, V., & Rajan, B. (2009). Profitable customer management: Measuring and maximizing customer lifetime value. Management accounting quarterly, 10(3), 1.

[23]. Hosseni, M. B., & Tarokh, M. J. (2011). Customer segmentation using CLV elements. Journal of Service Science and Management, 4(03), 284.

[24]. Hiziroglu, A., & Sengul, S. (2012). Investigating two customer lifetime value models from segmentation perspective. Procedia-Social and Behavioral Sciences, 62, 766-774.

[25]. Khajvand, M., & Tarokh, M. J. (2011). Analyzing customer segmentation based on customer value components (case study: a private bank). Advances in Industrial Engineering, 45(Special Issue), 79-93.