National College of Ireland

A Machine and Deep Learning Framework to Retain Customers based on their Lifetime Value

MSc Research Project
Data Analytics

Kannan Kumaran
Student ID: x20195061

School of Computing
National College of Ireland

Supervisor: Dr.Pramod Pathak & Dr.Paul Stynes

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Kannan Kumaran
Student ID:	x20195061
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr.Pramod Pathak & Dr.Paul Stynes
Submission Due Date:	31/01/2022
Project Title:	A Machine and Deep Learning Framework to Retain Customers based on their Lifetime Value
Word Count:	5928
Page Count:	16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	30th January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Machine and Deep Learning Framework to Retain Customers based on their Lifetime Value

Kannan Kumaran
x20195061

Abstract

Customer Lifetime Value (CLV) measures the average revenue generated by a customer over the course of their association with the firm. CLV is measured by RFM-Recency, Frequency, and Monetary factors using their previous purchasing history. This research proposes a Machine and Deep Learning Framework to predict the Customer Lifetime Value in order to retain customers through targeted product promotions. The proposed framework combines clustering and regression models to analyse the significant variables for predicting the value of customers. Customers are grouped based on that value into levels such as high medium and low profitable customers. To identify the optimum model, this research compares Deep Neural Network and Machine Learning to probabilistic models Gamma-Gamma and Beta-geometric/negative binomial in order to predict the level of profitable customer class of following years by segmentation with help of K-means and Hierarchical ML clustering algorithms. Results of the five models are presented in this paper based on accuracy(R^2), Mean Squared Error and Mean Absolute Error. This research shows promise for Deep Neural Network(R^2-71%) in projecting the CLV. Considering the predicted CLV, the e-commerce decides on which customer group to invest for achieving a long-term Customer Relationship Management strategy.

1 Introduction

Customer Lifetime Value is a concept in customer relationship management (CRM) that is defined as the present customer value based on the predicted future revenue contributed to the firm or products over lifetime of a customer. Examining the potential customers and by offering discounts and promotions to retain them helps to establish a long-term relationship between the customers and firm. Online retailers should focus on customer retention because acquiring a new customer is more expensive than to keep an existing one (Wu et al.; 2005). The following aspects that support e-commerce in marketing decision and activities based on lifetime value are 1) Customer profitability analysis: Data mining may be used to track changes in customer profitability by analyzing customer transactions, and it can also assist businesses in identifying the most valuable customers. As a result, marketing expenses may be reduced, and offer customers with targeted products and services having high profitability to achieve customer retention and value. 2) Cross-selling: It is a marketing strategy for providing exiting customers with new products and services which helps maintain the customer relationship. 3) One to one marketing: Provide personalized services according to the different requirements of customer (Zhang and Zhang; 2013).

Lifetime value determination is one of the methods to classify customer groups based on the RFM model. Using recency, frequency and monetary attributes, the potential customer value could be found. Lifetime Value or the monetary value of a customer is predicted before churning the firm. CLV is one metric that may be used to manage customers, and it must be precise enough to be used effectively. It is crucial to maintain customer satisfaction over the span of service and hence initially it is important for identifying what a consumer requires. Since the e-commerce faces several purchase entries that will be difficult to track and cannot be done manually using standard methods. To solve this problem, customer data examination has to be done to analyze the purchase behavior and characteristics. This could be accomplished by creating a machine & deep learning model which is capable of recognizing patterns and accurately predict the outcomes for the years ahead employing huge amounts of past purchase data.

Customer segmentation is another essential factor in the process of e-commerce data analysis that helps in classification of customer groups based on similarities. Due to numerous transactions, customer group classification is hard to attain using traditional practices. Because of this customer segmentation utilizing ML clustering is used in this project to categorize customers ranging high to low value. This divides the customers into many clusters based on their shared qualities. Also, customer retention is maintained by the firm using marketing techniques that would bring in more profit thereby minimizing the investment risk (Koul and Philip; 2021).

The aim of this research is to investigate to what extent a Machine and Deep Learning Framework can retain Customers based on their Lifetime Value. The major contribution of this research is a novel integrated analysis of lifetime value and customer clustering implementing unsupervised ML algorithms, Deep Neural Network and Boosting techniques. A minor contribution of this research is to compare with probabilistic methods gamma-gamma and Beta Geometric/Negative Binomial Distribution. The future number of transactions and spending by each client can be obtained by performing this research. The categorization enables easier management, and any firm may decide which kind of customers they desire to consider and target with division as: high, medium and low-level customers. By carrying out evaluation measures such as accuracy and root mean square error, the model performance is compared, and best fit is discovered. Additionally, clustering algorithms is used to segment or classify consumers and find the most valuable customer groups, allowing the firm to conduct marketing campaigns accordingly. In the future, a firm can expand its sales into several nations and employ various insights of predicted outcomes to acquire consumers, and there is a good possibility that every customer will recognize the firm.

This paper discusses machine & deep learning models used for predicting Customer lifetime value and application of clustering methodology for Customer relationship management in section 2 related work. The research methodology is explained in section 3. Section 4 discusses the design components for the CLV machine and deep learning framework. The implementation of this research is discussed in section 5. Section 6 presents and discusses the evaluation results. The research is concluded, and future work is discussed in section 7.

2 Related Work

CLV aids during tough competition among companies which drives heavy investments on marketing along with acquiring new customers. It is essential not only to attract new consumers but also to ensure that existing customers stay with the firm for as long as feasible in order to be truly profitable. Venkatakrishna et al. (2021) discusses on CLTV that provides recommendations on where to invest, which may be useful to a telecom company when developing a marketing plan. The aim of this project is to examine the company's customer sales data and forecast the customer lifetime value. Also, Customer segmentation is done to develop focus on groups based on the predicted CLTV. Based on machine learning models that predict CLV and segmentation, this paper guides in planning and decision-making marketing strategies for future. Gradient Boosting Regressor model outperforms other ML model producing 84% and through segmentation it is inferred that nearly half of the value is contributed by 17 percent of customers. This analysis helps companies to categorize the most important customers by identifying customer segments based on their value in order to determine their loyalty as well as revenue. Right products and effective strategies could be applied on the relevant customer class.

Customer segmentation is the process of categorizing customers based on shared parameters such as age, region, and purchasing habits. Hossain (2017) represented centroid-based and density-based techniques for data clustering incorporating the k-means and DBSCAN algorithms. The results of implementing these two algorithms show that they can both be used for customer segmentation; however, unlike k-means, DBSCAN provides an additional option for finding unusual customers with different spending habits thereby ensuring customer satisfaction and optimal profit. Furthermore, the results obtained through the use of a density-based clustering method appear to be relevant to the dataset in consideration. As inference from this paper, density-based clustering techniques should be considered for use in order to achieve adequate consumer segmentation with experimenting Neural Network for cluster analysis, other types of clustering algorithms will be used to different datasets and their performances will be assessed as well to procure satisfying consumer segmentation in the future.

In customer relationship management, the RFM model is an essential quantitative analytical model. Research on improved RFM customer segmentation model based on k-means algorithm differentiated customers depending on various cluster groups to produce an advanced RFM model. Huang et al. (2020) used acquired data to create parameters such as R, F, M, and C. C is a newly added parameter that identifies consumers who placed orders at the same time and belong to the same cluster group. The data was standardized after the RFMC values were calculated. The elbow technique was used to find the best cluster and was found to be 5. Both the traditional RFM and the enhanced RFMC model have been clustered. Finally, both approaches yielded similar results, with high recency values. The author performed well on the advanced RFMC model but a snake plot to highlight the variations between the models could have been made.

An essential first step is to examine the historical data and identify the highly related features. Relevant resources may be targeted towards profitable clients based on particular clusters with usage of ML. Segmenting Bank Customers via RFM Model and Unsupervised Machine Learning was used by Aliyev et al. (2020) to calculate the value supplied by customers to financial institutions, and three models were created using three different types of clustering algorithms. RFM values of customers in the data were used to create customer segments. The first model used K-method Mean's twice to divide con-

sumers into 5 interesting groups based on their RFM parameters. The second model was built by integrating noise and the K-Means method with density-based spatial(DBSCAN) was deployed to identify dataset outliers & noise, and K-Means to categorize these outliers/noise with 2 groups based on their recency. As a result of the approach, the financial institution's most valued customers emerged. The third model used the agglomerative clustering approach and exposed the dataset to 4 clusters, which produced alike clusters to those produced by K-Means with four clusters. The last model was omitted for previous reason and its tremendous computational complexity. In the future, the technique presented in this study might be improved by looking at additional forms of bank customer behavior loans, investments, deposits, and others, in addition to the ones retrieved from past record of transactions.

Customer segmentation using machine learning was carried out analyzing data from local retail shops to divide clients into several groups, ranging from highly valued to lowly valuable by Kansal et al. (2018). As part of the data pre-processing, data scaling was performed. The cluster was then formed using three clustering approaches, including k-means, agglomerative, and mean shift. The elbow technique was applied in the k-means clustering method, and the best cluster was found to be 5. The silhouette score assessment technique was used to evaluate the outcomes of clustering techniques, and it was discovered that clustering methods(Agglomerative & K-means) perform better (0.56) than mean shift method (0.52). The author did an excellent job of research by incorporating additional models and evaluation methods to obtain better clustering which provides reason to opt and perform comparisons in the research.

Customer Lifetime Value Model Framework Using Gradient Boost Trees with RANSAC Response Regularization was performed by Singh et al. (2018) to offer a mathematical multi-layer model framework for calculating customer lifetime value based on a rigorous theoretical taxonomy as well as assumptions grounded on client characteristics. Rather than using a gradient boosting technique to directly boost a base learner, this work uses RANSAC regularization to boost weak learners. The results are compared to a gradient boost with lasso regularization fitted by the complete training set and tested on a genuine customer base. In severely skewed customer data, experimental assessment reveals that the suggested framework delivers an accuracy of 80 percent with 0.12 MSE-Mean Square Error when compared to another evaluated method. From a technical and business standpoint, there is still a lot of improvisations required.

CLV modelling assists to determine a customer's expected business value and enables merchants to allocate resources efficiently in their businesses. Win and Bo (2020) Predicted Customer Class using Customer Lifetime Value with Random Forest Algorithm for the following year based on their CLV, which will assist the online retailer in determining which clients should be engaged in for long-term CRM. Model is trained using the Random Forest (RF) method, and Random Search tuning is used to get the greatest prediction accuracy. On the same dataset, an experimental study is done to compare with the AdaBoost method. Models using the ideal hyperparameter value from Random Search beat AdaBoost models in terms of accuracy. The model's accuracy with default hyperparameters and all features is 81.46 percent with improvement to (82.26%) when only the selected characteristics from feature selection were used, which is an excellent model. The probability of a Random Forest model with the best hyperparameters adjusted by Random Search rose by 2%. Product interest will be advised based on the client class and preferences, which will lead to an soar in sales and assist to establish a stronger relationship with potential consumers, also motivate Low class customers to

enhance retention strategy.
Summary: In summary, state of the art implies that several models such as Probabilistic, Machine learning and Boosting frameworks have been improved, and there is a requirement to determine a further efficient CLV model for customer retention. Deep Neural Network could be experimented according to current research. Also, state of the art with regards to customer segmentation indicates usage of RFM factors for classification of customer profitable class. This research proposes a Deep Learning integrated CLV along with ML clustering to guide effective marketing decisions. DNN will be identified as optimal model compared to BG/NBD, Gamma-Gamma, Linear Regression, Random Forest and Gradient Boosting. K-means clustering will be identified as better classifier compared to Hierarchical and manual RFM segmentation.

3 Methodology

The research methodology consists of five steps namely data collection, data pre-processing, data transformation, data modelling and conversion, evaluation and results as shown in Fig.1

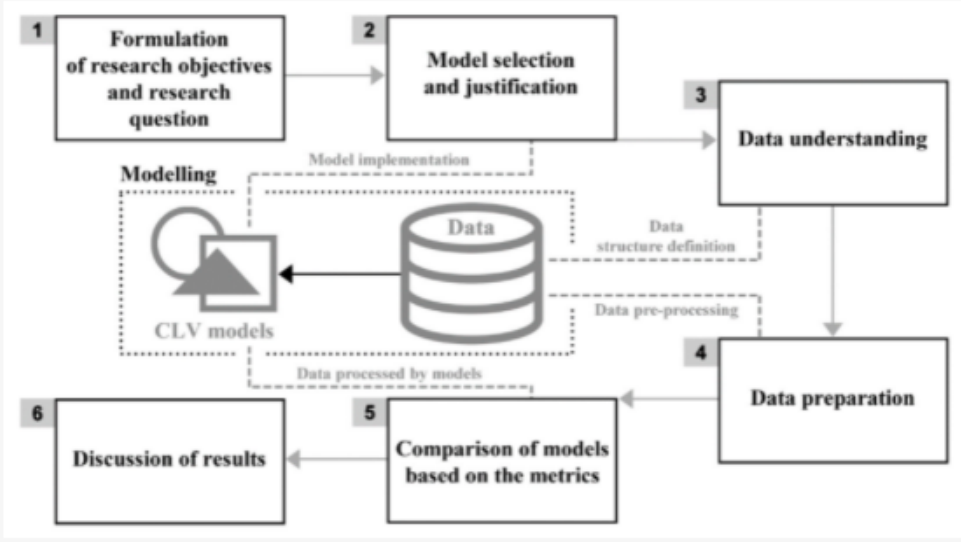


Figure 1: Methodology

The first step, Data Collection involves fetching of data from UCI machine learning repository to perform the research. A UK based online retail data consisting of transactions between 01/12/2009 and 09/12/2011 was collected. There are 541910 on year(2010-2011) records of customer transactions having 8 features namely customer ID, Invoice, stock code, description, invoice, date, price, quantity and country. The retrieved data has no ethical concerns and is used for academic research. Data source: `https://archive.ics.uci.edu/ml/datasets/Online+Retail+II`

The second step, Data Pre-processing involves selection of variables of interest from the given dataset such as Invoice, StockCode, Quantity, Price, InvoiceDate and Customer ID. Creation of an aggregated field named Total Price defining total price spent per product in each transaction obtained by multiplying Quantity with Price. Division of the variable InvoiceDate into two variables InvoiceDate and InvoiceDay that helps

distinguish different transactions by same customer at different times on the same day. The transactions having missing values are filtered out with regards to Customer ID. To avoid incorrect CLV and customer classification predictions data cleaning steps such as outlier identification and removal, duplicate elimination and value scaling was done.

The third step, Data Transformation/ Feature engineering involves generation of recency, frequency, and monetary parameters for each customer ID since our research is to predict CLV based on RFM model. To generate the RFM, customer id was grouped using the group-by function and aggregated with invoice date, invoice number, and total sales amount . The scores will be provided as input for customer lifetime value prediction and clustering model. Recency: To construct the recency values in invoice dates, a lambda function was built using the differentiation between last purchase and recent date of the customer. Frequency: The count of invoice numbers for each customer was acquired using the count() method to frame the frequency values. Monetary: To construct the monetary values, the sum of the sales amounts for each customer was acquired using the sum() function. Thus, classification of high, medium and low customers was possible with RFM calculation. The data set was split into a ratio of (80:20) for training and validation.

The fourth step, Data Modelling and Conversion involves model training, model conversion, and model implementation. The data has been modelled for CLTV prediction and customer segmentation. For this goal, both statistical and machine learning approaches are employed. Statistical methods such as beta geometric and gamma-gamma were used. Gradient Boosting and Deep Neural Network models were trained to learn and forecast continuous values. The predictor variable is the calculated CLV, and the model is trained on the training set and then validated on the test set in order to understand the model accuracy. Customer Segmentation is achieved by ML clustering algorithms such as K-means and Hierarchical clustering to group customers into different clusters based on their shared purchasing behaviour.

The fifth step, Evaluation and Results involves evaluating the performance of each of the deep & machine learning models using Accuracy(R^2), Mean Squared Error, Mean Absolute Error and Snake Plot. Snake plot was used in this study to compare RFM groups generated using ML clustering to the actual RFM groups through visualizations for easy identification. The metrics explained above was utilized to evaluate the models of CLV prediction . The model with higher accuracy and low error score will be illustrated as the best fit model.

4 Design Specification

The customer retention deep and machine learning framework architecture combines a deep neural network and ML models with segmentation using ML clustering as shown in Fig.2. The components of the framework include RFM, probabilistic models (BG/NBD and Gamma-Gamma), Clustering models (K-means and Hierarchical) and Regression Models(DNN, Linear Regression, Random Forest and Gradient Boosting).

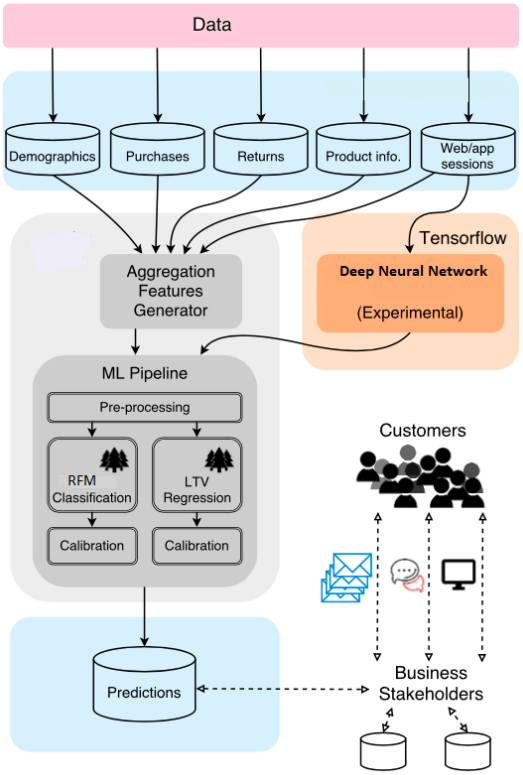


Figure 2: Design of CLV prediction and segmentation

4.1 Customer Segmentation models

1) RFM Model:
The customer groups ranging from profitable to non-profitable can be formed by the execution of RFM parameters based on the previous transactions of customer.
Recency: The recency parameter indicates how recently a customer made a purchase. The invoice and current date will be used to measure this.
Frequency: Frequency refers to the number of times a customer makes a purchase in a certain time period. The frequency will be determined using the customer ID.
Monetary: The entire amount spent by the customer in all transactions is referred to as monetary. Total Price and customer ID is used to measure this factor.
2) K-Means clustering:
K-means clustering is used to segment customers with similar characteristics into various clusters. This approach will use every data point and store it in a cluster having identical properties. This will be repeated until every data point has been allocated to a cluster. Every generated group will be distinct from the others with varying mean values. This method was chosen because the number of clusters may be fine-tuned before the model is implemented, and the clusters can be limited as desired. Moreover, several related works suggest that the k-means technique outperforms alternative clustering models. To find out the optimum number of clusters, Elbow method and silhouette score is used.
3) Hierarchical clustering:
Clusters with a specified order from top to bottom are generated using hierarchical clustering. Dendrogram is used to determine the number of clusters for hierarchical clustering. Two clusters are joined in this dendrogram once they are merged, and the height of the

join will equal the distance between these locations(Tripathi et al.; 2018).

4.2 Customer Lifetime Value Prediction Models

1) Beta-geometric/ Negative binomial distribution (BG/NBD) model:
The BG/NBD method is used to determine whether or not the customer is alive/active. It is a prediction system that uses previous customer transactions to identify the number of purchases made by each consumer thereby developing a customer retention strategy. This model performs the prediction of customer purchases that are a portion of customer lifetime value using the recency and frequency scores from the RFM. BetaGeoFitter() function was used and the probability of customers being active is specifically discovered with support of this model.
2) Gamma-Gamma model:
The Gamma-Gamma method is to predict each customer's monetary value. It calculates the amount spent by each customer using the recency and monetary factors. Lifetime Value (LTV) = future number of transactions * revenue per transaction *margin The output of the BG/NBD model will also be utilized to predict revenue using the R and M parameters in GammaGammaFitter() function.
3) Linear Regression:
To report the relationship between two variables, this technique fits a linear equation to the observed dataset. The main idea is to create a line that relates the data the best. The line with the minimal overall prediction error is the best line and the error is measured as difference in distance between each data point and regression line.
4) Random Forest:
Random forest is used as it does average to forecast data by fitting a number of classification decision trees on the data sub samples. Due to usage of numerous data instances, overfitting is minimized, and accuracy improved with changes in parameters.
5) Gradient Boosting:
It is employed along with 3 elements namely a weak learner for making prediction, a loss function which is optimized and additive model to insert the weak learners.
6) Deep Neural Network:
DNN are supervised procedures and include three layers namely input, hidden and output layer. Each node is linked to others containing weight and threshold. Node is activated in case the output of node exceeds certain threshold and hence data is passed to next layer. Certain parameters such as epochs, layers, batch size, optimizer, loss, activation function and metrics play a major role (Yi et al.; 2016).

5 Implementation

The Machine and Deep learning framework was implemented using google Colab and python language was used to carry out this research. The data source was loaded into dataframe using pandas library. The detailed implementation of the code can be viewed in the GitHub repository:`https://github.com/kannan-kumaran/Research-Project`. As per the Fig.3 , Customer ID and Description columns had missing values with percentage of 24.9 and 0.3 respectively and so those records were removed.

Duplicates entries using drop_duplicates() and transactions having cancelled orders and product quantity less than 1 were removed. To handle the outlier, two custom functions outlier_thresholds() and replace_with_thresholds() were defined. The shape of

	Missing Values	% of Total Values
Customer ID	135080	24.9
Description	1454	0.3

Figure 3: Missing value fields

dataframe was (392733, 10) after data preparation. In order to understand the processed data, several visualizations were plotted, and exploratory data analysis was performed. The topmost purchased product was "White Hanging T-light Holder" as shown in Fig.4 .This ensures that the demand is fulfilled by e-commerce through stock replenishment. Fig.4 demonstrates that UK stands first in number of products sold since the company is based in United Kingdom. Production of items that meet the needs of customers from various nations would help expand the business.

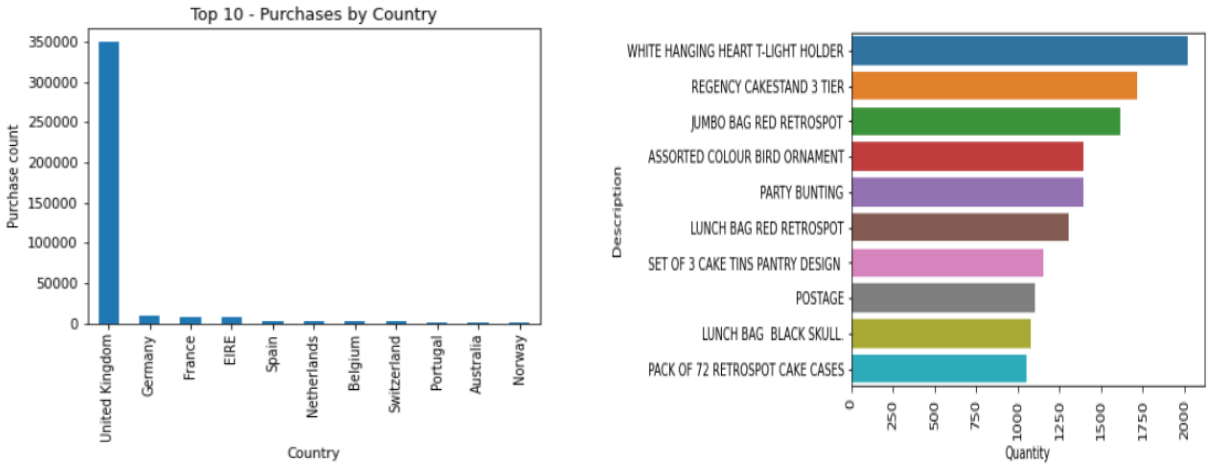


Figure 4: Top 10 Countries and Products in Sales

The bar plot in Fig.5 illustrates the month with the most sales is November 2011 and days from Monday to Thursday experience a steady increase in orders which later decreases . Also, the month with the lowest sales is unknown as the data set only includes transactions up through December 2011.

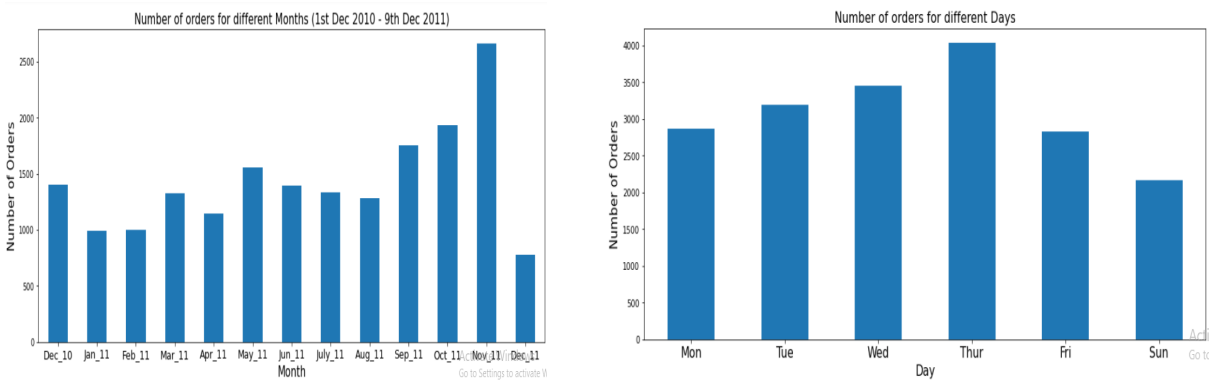


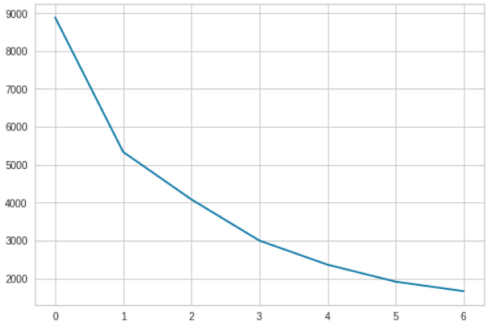
Figure 5: Months and days with top sales

1) Experiment on customer segmentation:
RFM parameters were generated after comprehension of the data using the qcut() function with value as 3. R, F, and M fields were formed with respective quartile values which were added to create RFM score. Robust RFM level was measured into 3 categories of customer group such as low, medium and high with consideration of RFM score. Fig.6 depicts the customer segmentation produced using the RFM score.

Customer ID	Recency	Frequency	Monetary	R	F	M	RFM_Score	Robust RFM Level
12346.0	325	1	310.44	1	1	1	3	Low
12347.0	2	182	4310.00	3	3	3	9	High
12348.0	75	31	1770.78	2	2	3	7	Medium
12349.0	18	73	1491.72	3	3	3	9	High
12350.0	310	17	331.46	1	1	1	3	Low

Figure 6: RFM level segmented dataframe

For the purpose of comparing the actual RFM levels obtained as discussed above with the ML predicted outcome, clustering techniques were used. K-means identified the optimum number of clusters using elbow method and silhouette score whereas Hierarchical clustering using dendrograms. Initially, K-means was opted and imported from the library of sklearn.cluster. The method KMeans() was then given the k variable to determine the cluster number. Using RFM data, a for loop was created to produce clusters depending on the sum of squared distance between each data point. An Elbow plot is displayed in Fig.7(a) which was graphed using pyplot library. Its is inferred that the curve flattens after three clusters with minor changes between each cluster. As a result, the customers are segmented in accordance to cluster number (k =3). Hierarchical Clustering was also analysed with different methods namely single, complete and average linkage. Using the Euclidean metric for the linkages, a dendrogram was plotted and based on inference clusters were segregated into 3. In the end, the outcomes of two clustering were added as columns H_Cluster (Hirarchical) and Kmeans_Cluster (Kmeans) to RFM dataframe for comparisons. Hierarchical clustering did not show promising results in segmentation.



(a) Elbow method

Kmeans_Cluster	Recency Mean	Frequency Mean	Monetary Mean	Monetary Count
0	46.125967	57.844936	960.054504	2715
1	248.696106	25.265907	416.192000	1053
2	23.579798	279.729293	4954.500808	495

(b) Cluster Mean Values

Figure 7: Cluster analysis

Thus K-means algorithm was executed with three clusters and mean of each cluster group was measured. The RFM mean values for each cluster is depicted in Fig.7(b)

Insights from mean values of recency, frequency and monetary for each cluster appear to be distinct from one another, indicating that customers have been segmented well. Clusters 1 and 2 appear to have higher mean values compared to Cluster 0 which is indication of valuable customers. Customers in Cluster 2 are more valuable than those in Clusters 0 and 1. Similar;y, Hierarchical clustering was performed with 3 clusters based on the interpretation of dendograms.The final RFM customer segmentation utilizing RFM values and the k-means clustering algorithm is shown in Fig.8 . Cluster 0 was assigned as low, cluster 1 was assigned medium, and cluster 2 was high profitable customer class based on the mean weightage .

CLV model implementation:

Customer lifetime value is a forecast of future buying and spending of each customer. For CLV prediction, Lifetime package has been installed in python and features invoice, customer id, invoice date, quantity, and total price were selected for prediction.

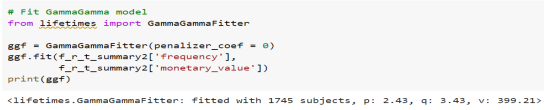
Train and test data split: The split date has been set to "2011-08-01 " corresponding to the end of the training data. The data was divided into train and test utilizing a data parser library. Separation of 8 months period train and 2-month test data was done.

2) Experiment on the probabilistic models:

BG/NBD model: Using summary_data_from_transaction_data() from lifetime library, RFM attributes have been generated with the relevant variables. The model is fitted with BetaGeoFitter and value of 0.1 is assigned to penalizer coefficient parameter in order to manage huge data fields for a smaller sample size. The parameter t is included for determining the first purchase and last period differences. Future buying pattern were predicted by the addition of t, frequency and recency attributes into the model. The alpha, beta, and r parameters, which were assigned to F- frequency, R- recency, and t respectively are shown in Fig.8(a) .In the end, a data frame was developed using merge function from Panda's library was used to join true and predicted purchases.

Gamma-Gamma model: The real amount spent by each customer was derived by aggregating the sum of the total price attribute. The train dataset was used to hold monetary values greater than zero in order to fit the model utilizing GammaGammaFitter lifetime and find projected spending. The penalizer coefficient has been set to 0 in this case with the frequency and monetary value into the model which is depicted in Fig.8(b)

(a) BG/NBD



(b) Gamma-Gamma

Figure 8: Probabilistic fitted models

3) Experiment on Deep Neural Network:

As discussed above for statistical models, data division into train and test is not required. This is because probabilistic models employ the features period to estimate latent variables. The feature periods are X and Y in the model and is unsupervised modelling. However, in case of machine learning algorithms values have to be trained for prediction as well as need a Y subset that is new to the model for performance evaluation. Some features must be produced before the machine learning models can be implemented. The transactional data features were defined using get_features() function and data were split

into train and test corresponding to dates. The target variable is chosen as the next purchase period and other features as independent variable to perform prediction. Deep neural Network was applied with build_model() where various parameters are defined . Relu activation function was used for the sequential layers. The DNN model was compiled with Adam optimizer and Mean Squared Error as metrics and loss function. Patience parameter is used for improvement by epoch number check in the EarlyStopping() of callback library. Finally, the model is fitted with X_train, y_train data with batch size of 32 and five epochs. Using the model, prediction results were acquired for the next 89 days as portrayed in Fig.9

```
#Deep Neural Network
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers
from keras.callbacks import ModelCheckpoint, EarlyStopping

def build_model():
    model = keras.Sequential([
    layers.Dense(256, activation='relu', input_shape=[len(X_train.columns), ]),
    layers.Dropout(0.3),
    layers.Dense(64, activation='relu'),
    layers.Dropout(0.3),
    layers.Dense(32, activation='relu'),
    layers.Dense(1)
    ])

    optimizer = tf.keras.optimizers.Adam(0.001)

    model.compile(loss='mse',
            optimizer=optimizer,
            metrics=['mae', 'mse'])

    return model

# The patience parameter is the amount of epochs to check for improvement
early_stop = keras.callbacks.EarlyStopping(monitor='mse', patience=30)

model = build_model()
#Should take 10 sec
early_history = model.fit(X_train, y_train,
                    epochs=5, batch_size=32, verbose=0,
                    callbacks=[early_stop])
```

Figure 9: Deep Neural Network Model

For the purpose of identifying the best model, DNN was compared to ML algorithms such as Linear Regression, Random Forest and Gradient Boosting. To forecast future purchases, the target field was chosen as the next purchase period while other features were used as the independent fields. The LinearRegression, Random Forest regressor and GradientBoostingRegressor techniques were imported from sklearn metrics and fitted with training attributes. The actual test variable was compared to predicted test variable by model and evaluated results were stored in dataframe.

6 Evaluation

6.1 Evaluation of segmentation

The aim of this experiment is to compare customer segmentation undertaken using RFM analysis and clustering methods (K-Means & Hierarchical). To guide retail industry in determination of optimum segmentation, the models had to be compared (Ahalya and Pandey; 2015). This was achieved using heatmap & snake plot for evaluation purpose. The snake plot of K- means clustering for customer classes(high, medium, low) are shown as (K_High, K_Medium, K_low) respectively in Fig.10. This plot helps in the classification customers group based on profitability and determines how well are segments

created. The peak reached on the graph are the highly profitable customers with value of nearly 3000 implying quantile discretization value of clusters. On inference, it indicates that segmentation done by clustering algorithms have more diverse clusters with greater weightage than by manual RFM analysis.

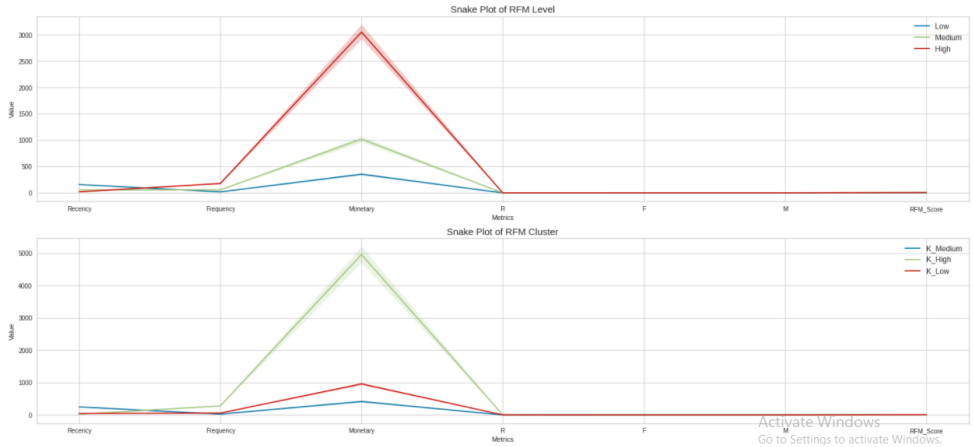


Figure 10: Snake Plot for comparison of RFM cluster and manual segmentation

Fig.11 demonstrates heat map for RFM segmentation attained with k-means clustering and also with RFM levels. Heat Map is used to find the relative importance and if the segmented groups are widely classified. On comparison of mean values, it was discovered that K-means segmented RFM outperforms manual RFM segmentation level. When comparing high profitable level customers from both heat maps, clustering based segmentation has higher mean values of 2.70 and 2.84 implying that customers are valuable and that segmentation is more effective.

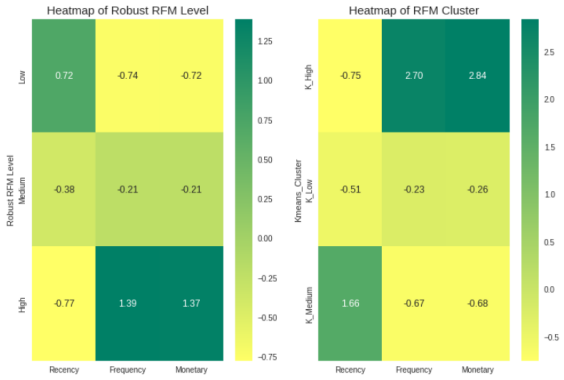


Figure 11: Heat Map for RFM comparison

6.2 Evaluation of Customer Lifetime Value Models

The aim of this experiment is evaluation of individual model performance accomplished using evaluation metrics such as R^2 score, mean absolute error, and mean squared error after the implementation of BG/NBD and Gamma-Gamma probabilistic models along with Deep Learning and ML approaches such as DNN, Linear Regression, Random Forest

and Gradient Boosting Regression. Future customer spends and purchases were obtained with help of these models. The final results of models implemented to predict CLV are concatenated into a data frame and displayed in Model Results Table.

Model Results			
Algorithm	Mean Squared Error(in million)	Mean Absolute Error	R^2 score (Accuracy)
Gamma-Gamma	1.8	300	57%
BG/NBD	1.2	0.66	52%
DeepNeuralNetwork	5.9	734	71%
Linear Regression	6.4	760	68%
Random Forest	7.7	811	63%
Gradient Boosting	7.6	840	63%

With 51 percent accuracy, the BG/NBD model predicted future customer purchases. The model accuracy is referred to by the R2 score. When the actual and projected values were compared, the the mean absolute error (MAE) was 0.66.

Then gamma-gamma model was used and evaluated for customer spend forecast that produces 57 percent accuracy. The error values are MAE -300. This model did not perform well in forecasting future spending.

After application of probabilistic models, ML algorithms were implemented. Linear Regression resulted with R2 score-(68%), mean absolute error-760. Random Forest resulted with R2 score-(63%), mean absolute error-811. Gradient Boosting resulted with R2 score-(63.4%), mean absolute error-840. In order to identify whether it is feasible to obtain an improved model accuracy by Deep Learning, DNN was implemented and outperformed rest of the models by generating results with highest accuracy of (71%) and lower mean absolute error value-734.

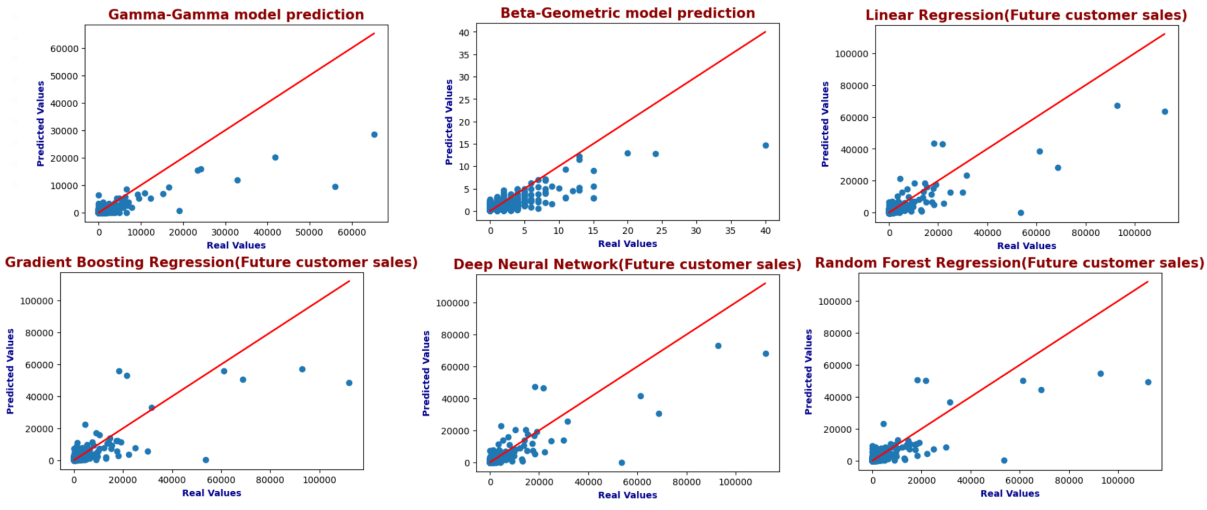


Figure 12: Actual vs Predict CLV model results

The comparison of actual and forecasted values for all five models are shown in Fig.12 and notice that few models have forecasted values close to actual values. Thus, guides in better model identification through visualizations. However, taking into account both model results table and plots, shows promise for DNN with best accuracy and contributes

a higher performance than beta-geometric, gamma-gamma model, Linear Regression, Random Forest and Gradient Boosting Regression.

6.3 Discussion

The proposed Machine and Deep Learning framework for customer retention was developed with focus on CLV adoption. Therefore, sales of the retail company were analysed using the previous transactions and predicted CLV with various models to ensure high accuracy, higher performance and minimum loss. This research is novel from previous related works by the usage of a Deep Learning along with ML and probabilistic models. Comparison of DNN results with the state of art models strengthened the fact of building an improved performance model for accurate predictions. Also, integration of customer segmentation concept by ML clustering is an added advantage to strategical marketing based on classes (high, medium and low) profitable customers.

• The inferences from exploratory data analysis are customers that has placed the most orders are from the United Kingdom and that spends the most price on purchases is from the Netherlands. Product in demand was "White Hanging T-light Holder". The month of November in year 2011 had the highest sales. According to days of week, Monday to Thursday faced a steady increase and later decreases.

• Deep Neural Network shows promise if the motivation is for accuracy as well as loss.

• RFM values obtained by clustering are promising with more diverse clusters and weightage than manual RFM calculation.

• Improved performance and accuracy is possible with the availability of additional data points. A minimum of 3 years of transaction history could achieve more data which is thereby a limitation in this research.

7 Conclusion and Future Work

The aim of this research was to ensure customer retention through analysis of customer sales data of online retail with customer segmentation and predict the CLV. This research proposes a Machine and Deep Learning framework to perform predictions and the outcomes guide the retail industry for effective planning and decision making of marketing strategies. On conduction of RFM analysis, three customer groups were obtained based on sum of recency, frequency and monetary scores. This was compared to clustering (K-means & Hierarchical) techniques and noticed diverse segments having larger mean values to achieve a promising customer segmentation.

Following this, the CLV was forecasted using probabilistic models such as BG/NBD to find future purchases and Gamma-Gamma to find the future spending. This was compared to CLV predicted using Deep Neural Network, Linear Regression, Random Forest and Gradient Boosting Regressor. In order to determine the best performing model, evaluation was done using R2 score, Mean Absolute Error and Mean Squared error. Results demonstrate that Deep Neural Network shows promise in terms of accuracy (71%) and loss (MAE - 731).

The restriction of data transactions is a limitation to truly enhance the prediction performance. This work can be improved by data gathering of at least 3 years of retail transactions. In addition, customer segmentation should be focused on generating cluster groups based on the products in the retail dataset. The future work could be to experiment new features and data samples for feasible model training. To improve the

accuracy of the used models, a larger sample size and a longer time span can be utilized as input. Also, hyperparameter tuning on the parameters and integration of algorithms can be carried out as part of future study.

References

Ahalya, G. and Pandey, H. M. (2015). Data clustering approaches survey and analysis, *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, pp. 532–537.

Aliyev, M., Ahmadov, E., Gadirli, H., Mammadova, A. and Alasgarov, E. (2020). Segmenting bank customers via RFM model and unsupervised machine learning.
URL: *https://arxiv.org/abs/2008.08662*

Hossain, A. S. (2017). Customer segmentation using centroid based and density based clustering algorithms, *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, pp. 1–6.

Huang, Y., Zhang, M. and He, Y. (2020). Research on improved rfm customer segmentation model based on k-means algorithm, *2020 5th International Conference on Computational Intelligence and Applications (ICCIA)*, pp. 24–27.

Kansal, T., Bahuguna, S., Singh, V. and Choudhury, T. (2018). Customer segmentation using k-means clustering, pp. 135–139.

Koul, S. and Philip, T. M. (2021). Customer segmentation techniques on e-commerce, *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 135–138.

Singh, L., Kaur, N. and Chetty, G. (2018). Customer life time value model framework using gradient boost trees with ransac response regularization, *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Tripathi, S., Bhardwaj, A. and Eswaran, P. (2018). Approaches to clustering in customer segmentation, *International Journal of Engineering Technology* **7**: 802.

Venkatakrishna, M. R., Mishra, M. P. and Tiwari, M. S. P. (2021). Customer lifetime value prediction and segmentation using machine learning.

Win, T. T. and Bo, K. S. (2020). Predicting customer class using customer lifetime value with random forest algorithm, *2020 International Conference on Advanced Information Technologies (ICAIT)*, pp. 236–241.

Wu, L., Liu, L. and Li, J. (2005). Evaluating customer lifetime value for customer recommendation, *Proceedings of ICSSSM '05. 2005 International Conference on Services Systems and Services Management, 2005.*, Vol. 1, pp. 138–143 Vol. 1.

Yi, H., Shiyu, S., Xiusheng, D. and Zhigang, C. (2016). A study on deep neural networks framework, *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, pp. 1519–1522.

Zhang, X. and Zhang, J. (2013). Crm applications in e-commerce strategy, *2013 International Conference on Computational and Information Sciences*, pp. 605–608.