DOI: 10.53469/jgebf.2025.07(12).01

Data-Driven Decision-Making in Retail

Manojdeep Singh Jasrotia, Nithin Narayan Koranchirath

Abstract: ***This study does a deep dive into potential use cases that a Retail company can take advantage of. Here, we try to explain how data driven decision making can boost Retail business. By presenting practical predictive analysis, statistical modeling, deep learning, a retail worker working in the B2B space can transform their business with their transactional logs, customer reviews, etc.Furthermore, the study tries to explain models that can be used in each of the use cases and the working algorithm behind it. It also covers how to assess those model performance and do a necessary model performance check.***

Keywords: Customer Segmentation, Customer Propensity, Demand Forecasting, Machine Learning, Artificial Intelligence, B2B, Retail

1. Introduction

In this competitive business age, every organization wants to be a step ahead than others. To do so, there is no better tool than data. Predictive analysis or inference drawn from statistical modeling is something that can show an organization they way ahead or open a new horizon of business.

Machine learning, which is a branch of Artificial Intelligence is used to identify patterns in data on which it is being trained and make predictions based on parameters trained while learning that pattern. There are several algorithms that can be used to do predictive analysis using machine learning and some important ones are being discussed in this study in the retail B2B space.

Customer segmentation is a technique that uses customer information in terms of demographic, physiological, behavioral, geographic, membership tags, etc. to create customer groups that can be used to do further analysis within these groups or create separate marketing strategies.

Customer propensity modeling is used to predict whether a given customer would buy in a specified future timeframe. This is done using historical customer buying behavior, factors like RFM (recency, frequency, monetary), seasonal effects, etc.

Demand forecasting helps an organization predict future demand for a product or service. There can be multiple factors on which it might depend. Some of those factors are customer historical buying patterns, current and future price points, advertising spend, competitor factors, etc.

2. Literature Review

2.1 Customer Segmentation

Authors of article [12] stated that Customer segmentation as a whole falls into one of two categories, either the business-to-consumer (B2C) or business-to-business (B2B) setting. The major difference between these two categories is the type of customer a company conducts business with. Products of B2B companies are purchased by distributors that act as intermediaries between the manufacturer and the end consumer. This is in contrast to B2C companies that sell their products directly to the end consumer [1]. Another difference of B2B customers is their demographics, which are instead referred to as firmographics. Some examples of firmographics are the customer type (ex: wholesale, retail, etc.), the customer's years in business, and the location of the customer with respect to the company.

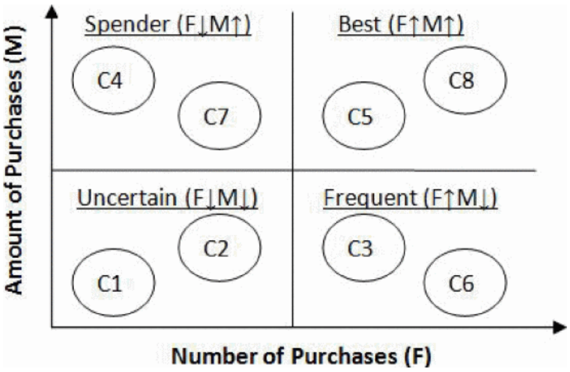

Figure 1: Customer value matrix (marcus [2])

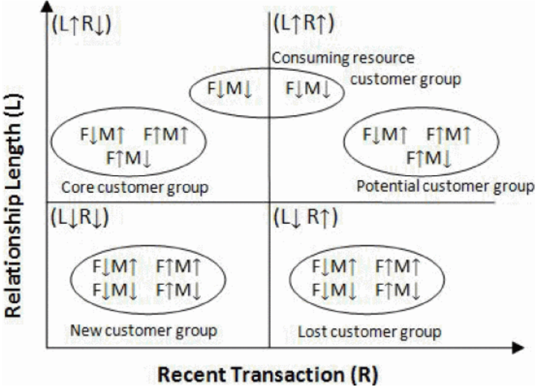

Figure 2: Customer loyalty matrix (chang and tsay [3])

Segmentation of customers in the B2B context is considered of equal importance as the B2C one [4]. With the nature of large transactions in B2B companies, it is imperative to fully identify and understand the characteristics of different groups business customers fall into. This is done with the aim of continuously increasing the value of each customer to the company, by means of discovering the needs of customers and addressing them accordingly [5]. Moreover, according to [1], research in the B2B customer segmentation area is considered under-researched in comparison to studies in the B2C customer segmentation field. In this paper B2B customer segmentation is applied.

2.2 Customer Propensity

A well-known method for assessing customers' purchasing behavior is the Recency, Frequency, Monetary (RFM) analysis developed by Arthur Hughes [6]. It is also claimed to be the basis for an average of 80% of all customer's behavioral segmentation techniques by researchers [7]. An RFM analysis is conducted in order to group customers of similar spending patterns together. The main goal is to handle each resulting segment in a fashion tailored specifically for it. This is achieved by basing the company's marketing strategies on the discovered information about the different types of customers the company has [8].

The RFM analysis consists of three predefined variables extracted from customers' transactional data, and are defined as follows [6] [8] [10]:

Recency: The time between the latest transaction a customer made with the company and the present (measured in days, months, or years).

Frequency: The total number of transactions a customer has made in a predefined time period.

Monetary: The total monetary amount of purchases a customer has spent during the allotted time period.

It is also important to note that taking the average frequency and monetary values per unit time is a more well rounded approach than taking the values' totals [8]. Moreover, a fourth variable was added to the RFM analysis by Chang and Tsay [3], namely the length (L) attribute, which is the length of the relationship between the customer and the company measured in unit time. According to [9], it is found that the LRFM method is a better alternative than the RFM when clustering data.

2.3 Demand Forecasting

As mentioned by Henri & Johanna [11] by utilizing a highly accurate demand forecast, retailers can predict the demand for goods at each store location and channel. This ensures high availability for customers while minimizing stock risk. A reliable forecast can also aid in capacity management, staff allocation in stores and distribution centers, and assist buyers in managing long lead-time purchasing.

Generating an accurate forecast is relatively simple under stable conditions. However, the retail industry is dynamic, with numerous factors impacting demand. Retail demand planners must take into account a multitude of variables, including:

- Variations in baseline demand that recur, such as those related to weekdays and seasons.
- Internal business decisions aimed at capturing consumer attention and gaining a competitive edge, such as promotions, price adjustments, or changes to in-store displays.
- External factors, such as local events, changes in a store’s neighborhood or competitive situation, or even the weather.

With so much data to consider, it is impossible for a human planner to account for all potential factors. Machine learning makes it possible to consider the impact of these factors at a detailed level, by individual store or fulfillment channel. As a result, many retailers are transitioning to machine learning-based demand forecasting.

Machine learning makes it possible to incorporate the wide range of factors and relationships that impact demand on a daily basis into your retail forecasts. This is enormously valuable, as just weather data alone can consist of hundreds of different factors that can potentially impact demand. Machine learning algorithms automatically generate continuously improving models using only the data you provide them, whether from your business or from external data streams. The primary benefit is that such a system can process retail-scale data sets from a variety of sources.

Machine learning is an extremely powerful tool in the data-rich retail environment. It should be leveraged in any context where data can be used to anticipate or explain changes in demand. In some instances, it can even fill in the gaps where the data is lacking.

1) **Recurring Demand Patterns:** These patterns can be influenced by a variety of factors, including seasonal variations, holidays, and other events that impact consumer behavior. For example, demand for certain products may increase during the holiday season or decrease during the summer months. Retailers can use demand forecasting techniques to anticipate these changes in demand and adjust their inventory levels accordingly to ensure they have the right amount of stock on hand to meet the demand.
2) **Internal Business Decisions:** Internal business decisions, such as promotions, price adjustments, and changes to in-store displays, can significantly impact retail B2B operations by affecting demand, inventory levels, and customer satisfaction.
3) **External Factors:** External factors such as economic conditions, the political/legal environment, competition, and the social environment can all influence retail B2B operations. These factors can impact demand, inventory levels, and customer satisfaction, and carefully consider them when making business decisions.
4) **Unknown Factors:** Changes in demand for which the impacting factor has not been recorded, such as competing companies or other natural calamities that impacts demand.

3. Methodology

3.1 Modeling

3.1.1 Customer Segmentation

To do customer segmentation one of the most common techniques is clustering which is an unsupervised form of machine learning. The segments or groups that come out of this exercise can benefit the business to form advertising and marketing strategies. There are several clustering techniques that can be used.

K-means clustering:

Algorithm steps [13]:

- Pick up number of clusters(k)
- Random k points are being made cluster centroids in the coordinate space. K-means++ initializes points out of the data points available.

- Categorize each data points to each of the cluster centroids and recompute cluster centroid
- This recomputation goes on till the cluster centroid does not converge

Objective function:

$$\sum_{j=1}^{k} \sum_{i=1}^{n} | x_i^{(j)} - c_j |$$

where, k is the number of clusters and x denotes a data point.

DBSCAN: is a density-based clustering algorithm that groups data points based on their proximity and density.
Steps of the algorithm follows:

- For each data point, find the number of points within a given radius (eps).
- If the number of points is greater than or equal to a minimum threshold (minPts), mark the point as a core point. Otherwise, mark it as a noise point.
- For each core point, find all the directly reachable points (points within eps radius) and assign them to the same cluster as the core point. If a directly reachable point is also a core point, find its directly reachable points and add them to the cluster as well. This process is repeated until no more points can be added to the cluster.
- Mark any point that is not reachable from any other point as an outlier.

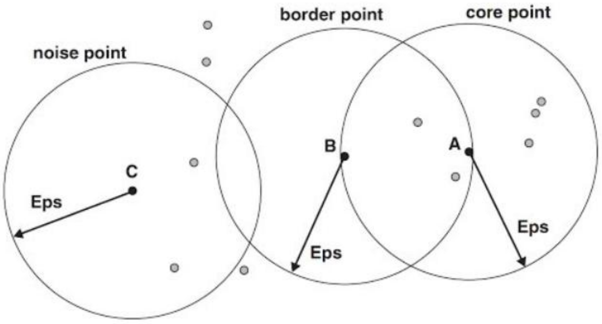

Figure 3: Classification of data points in DBSCAN (Dhanya Thailappan [16])

Hierarchical Clustering:
Hierarchical clustering is a technique used in cluster analysis to identify patterns in data. It involves building a hierarchy of clusters. There are two main types of hierarchical clustering: Agglomerative and Divisive.

Agglomerative clustering is a “bottom-up” approach, where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Divisive clustering, on the other hand, is a “top-down” approach, where all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. The choice of metric and linkage can have a significant impact on the result of the clustering. The metric determines which objects are most similar, whereas the linkage criterion influences the shape of the clusters.

The dendrogram is a tree-like diagram that displays all the merges and splits performed during hierarchical clustering. The optimal number of clusters can be determined by selecting a cut-off point on the dendrogram where the vertical distance between two branches is greatest [19].

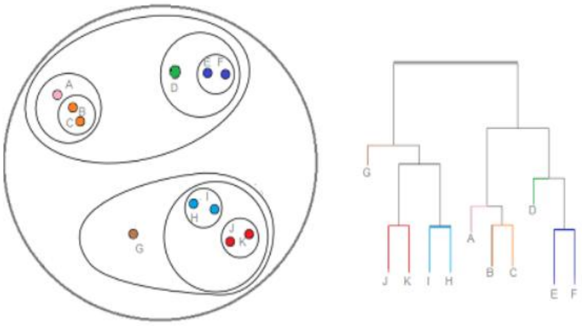

Figure 4: Dendogram [18]

Metrics to evaluate the models:

Silhouette score:
The silhouette score is a measure of how similar an object is to its own cluster compared to other clusters. It is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample. It ranges from -1 to 1, where -1 means clusters have been wrongly assigned and 1 means the ideal state of clustering is attained.

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

where,
C(i): The cluster assigned to the ith data point
N[C(i)]: Number of data points in cluster i
a(i) is a measure of how well assigned the ith data point is to it’s cluster:

$$a(i) = \frac{1}{N[C(i)] - 1} \sum_{C(i), i \neq j} distance(i, j)$$

b(i) is the average dissimilarity to the closest cluster which is not it’s cluster:

$$b(i) = min\ i \neq j(\frac{1}{N[C(i)]} \sum_{j \epsilon C(j)} distance(i, j))$$

Rand Index(RI) and Adjusted Rand Index(ARI) are measures for similarity check between two clustering methods [15].

$$R = (a + b) \div C_2^n$$

a: The number of times a pair of elements belongs to the same cluster across two clustering methods.
b: The number of times a pair of elements belong to different clusters across two clustering methods.
C_2^n: The number of unordered pairs in a set of n elements.

The Rand index value ranges between 0 and 1, when

0 it indicates that two clustering methods do not agree on the clustering of any pair of elements. Whereas, 1 indicates that two clustering methods perfectly agree on the clustering of every pair of elements

Adjusted Rand Index:
Suppose we have a dataset of six elements: {A, B, C, D, E, F}. We use two clustering methods that place each element in the following clusters:

Method 1 Clusters: {1, 1, 2, 2, 3, 3}
Method 2 Clusters: {2, 2, 1, 1, 3, 3}

To calculate the ARI between these clustering methods, we first need to create a contingency table that shows the number of pairs of elements that are assigned to each combination of clusters in both methods:

	Method 2 Cluster 1	Method 2 Cluster 2	Method 2 Cluster 3
Method 1 Cluster 1	0	2	0
Method 1 Cluster 2	2	0	0
Method 1 Cluster 3	0	0	2

Next, we can calculate the marginal totals for each row and column:

	Method 2 Cluster 1	Method 2 Cluster 2	Method 2 Cluster 3	Row Total
Method 1 Cluster 1	0	2	0	2
Method 1 Cluster 2	2	0	0	2
Method 1 Cluster 3	0	0	2	2
Column Total	2	2	2	

We can also calculate the total number of pairs:

Total Pairs = (n choose k) = (6 choose 2) = (6 * (6 -1)) / (2 * (1)) = 15

Next, we can calculate the expected index for each row and column using the formula:

Expected Index = (Row Total * Column Total) / Total Pairs

For example, the expected index for Row 1 and Column 2 is:

Expected Index = (Row Total * Column Total) / Total Pairs

Expected Index = (2 * 2) /15

Expected Index = 0.27

We can then calculate the ARI using the formula:

ARI = (Index - Expected Index) / (max(Index) - Expected Index)

where Index is the observed index calculated from the contingency table. For example, the observed index for Row 1 and Column 2 is:

Index = Number of pairs in Row i and Column j

Index = 2

Using this formula and the values from our contingency table and marginal totals, we can calculate the ARI as follows:

ARI = ((4 - (0.27 * (4 +4 +4))) / ((4 +4 +4) / (2))) - ((0.27 - ((4 *4) / (15))) / ((4 +4 +4) / (2)))

ARI = -0.14

3.1.1 Customer Propensity

Customer propensity model is used to predict the next purchase of a customer. To do so, the most common method is to use any simple classification machine learning model like Logistic Regression, Decision Trees, etc.

Decision trees are a type of supervised learning algorithm used for classification and regression tasks. The algorithm works by recursively splitting the data into subsets based on the values of the input features until a stopping criterion is met. The result is a tree-like model where each internal node represents a test on an input feature, each branch represents the outcome of the test, and each leaf node represents a class label or a numerical value [24].

Here are the steps involved in building a decision tree:

1) Select the best feature to split the data based on some criterion such as information gain or Gini impurity.
2) Split the data into subsets based on the values of the selected feature.
3) Recursively apply steps 1-2 to each subset until a stopping criterion is met, such as reaching a maximum depth or having all instances in a subset belong to the same class.

Bagging, Boosting, and Stacking are three popular ensemble learning techniques used in machine learning to improve the predictive performance of models by combining multiple models into one .

Bagging (Bootstrap Aggregating) is an ensemble learning technique that involves training multiple models on different subsets of the data and then combining their predictions. Bagging is used to reduce variance and overfitting by creating multiple models using bootstrapped samples . It is mostly applied to tree-based machine learning models such as decision trees and random forests .

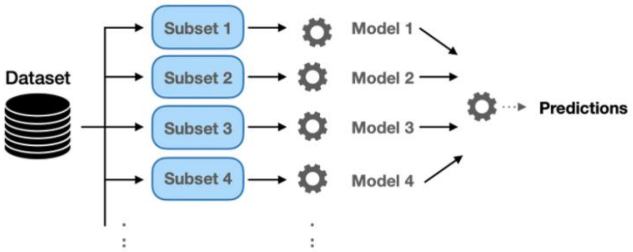

Figure 5: Bagging [21]

Boosting is another ensemble learning technique that involves training a sequence of models. Each model is trained on a weighted training set, with weights assigned based on the errors of the previous models in the sequence. The main idea behind sequential training is to have each model correct the errors of its predecessor. Boosting is used to reduce bias.

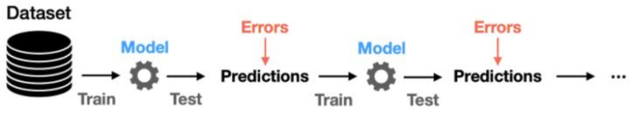

Figure 6: Boosting [21]

Stacking is an ensemble learning technique that involves training multiple models and then using another model to combine their predictions. Stacking combines the predictions of several base models to enhance the overall prediction accuracy. It aims to improve prediction accuracy by introducing a meta-level and using another model/ approach to estimate the input together with outputs of every model to estimate the weights or, in other words, to determine what models perform well and what badly given these input data.

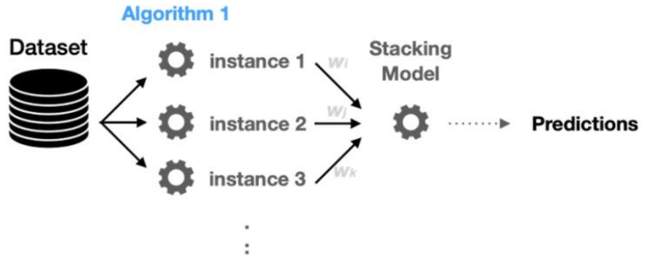

Figure 7: Stacking [21]

In summary, bagging reduces variance, boosting reduces bias, and stacking aims to improve prediction accuracy by combining the predictions of several base models using another model [22].

Random Forest is an ensemble learning method that uses bagging to improve the performance of decision trees . Here are the steps involved in the Random Forest algorithm [23]:

1) Randomly select k features from the dataset.
2) Construct n decision trees using the selected features and bootstrapped samples of the data.
3) For each decision tree, at each node, randomly select the best split from a random subset of m features.
4) Repeat steps 1-3 for n times to create n decision trees.
5) To make a prediction, pass the input through all the decision trees and output the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Here, k and m are hyperparameters that can be tuned to improve the performance of the algorithm.

Evaluating Classification models:

The task of evaluating classification models is to measure the degree to which the classification suggested using the model corresponding to the actual classification of the case. Depending on the method of observing, there are different measures for evaluating the performance of the model. Selection of the most appropriate measures shall be done depending on the characteristics of the problem and ways of its implementation.

In the evaluation of classification models the basic concept is the notion of fault. If the application of the classification models in selected case leading to the prediction of a class that is different from the actual class examples then there is an error in classification. If any mistake is equally important, then the total number of errors in the observed set can be an indicator of the work of a classifier. This approach is based on accuracy as a measure for evaluating the quality of the classification model. This measure can be defined as the ratio of the number of correctly classified examples according to the total number of classified examples.

$$Accuracy = \frac{number\ of\ correctly\ classified\ examples}{total\ number\ of\ cases}$$

The main disadvantages of accuracy as a measure for evaluation are as follows: (1) neglects the differences between the types of errors; (2) dependent on the distribution of class in the dataset. It is often important in practical problem solving to distinguish certain types of errors. In retail B2B, a system might be used to classify suppliers as reliable or unreliable based on their past performance. If the system incorrectly classifies an unreliable supplier as reliable, the error is more important because the retailer may continue to do business with that supplier and suffer losses. On the other hand, if the system classifies a reliable supplier as unreliable, the error has less importance because further investigation and due diligence will likely reveal that the supplier is, in fact, reliable.

The largest number of measures for evaluation of classification models related to classification problems with two classes. This is not a particular limitation for the use of these measures, given that problems with larger numbers of classes can be displayed as a series of problems with two classes. Each of these measures in particular stands out one of the classes as a target class, with the data set is divided into positive and negative examples of the target class. The negative examples include examples of all other classes. That is why below we consider a classification problem with two classes.

Confusion matrix in classification problems with two classes is shown in figure 8. It can be concluded from the figure that there are possible four different results forecasts. Really positive and really negative outcomes are correct classification, while the false positive and false negative outcomes are two possible types of errors.

False positive example is a negative example class that is wrongly classified as positive and false negative is a positive example of the class who is wrongly classified as negative:

- a is the number of correct predictions that instances are negative,
- b is the number of incorrect predictions that instances are positive
- c is the number of incorrect predictions that instances are negative.
- d is the number of correct predictions that instances are positive.

		Predicted Class	
		Negatives	Positives
Actual Class	Negatives	a	b
	Positives	c	d

Figure 8: Confusion matrix in classification problem with two classes.

A few standard terms are defined in a matrix with two classes: accuracy, true positive rate, false positive rate, true negative rate, false negative rate and precision. The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined. Accuracy may be determined using the equation:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{a+b}}{\boldsymbol{a+b+c+d}}$$

True positive rate is the proportion of positive cases that are properly identified and can be calculated using equation:

$$\boldsymbol{True\ Positive\ rate} = \frac{\boldsymbol{d}}{\boldsymbol{c+d}}$$

The false positive rate is the proportion of negative cases that were incorrectly classified as positive, and calculated with equation:

$$\boldsymbol{False\ Positive\ rate} = \frac{\boldsymbol{b}}{\boldsymbol{a+b}}$$

The true negative rate was defined as the proportion of negatives cases which are classified correctly, and is calculated using the equation:

$$\boldsymbol{True\ Negative\ rate} = \frac{\boldsymbol{a}}{\boldsymbol{a+b}}$$

The false negative rate is the proportion of positive cases that were incorrectly classified as negative, and are calculated using equation:

$$\boldsymbol{False\ Negative\ rate} = \frac{\boldsymbol{c}}{\boldsymbol{c+d}}$$

Finally, precision or positive predictive value presents the fraction of predictive positive cases that are accurate, and is calculated using the equation:

$$\boldsymbol{Precision} = \frac{\boldsymbol{d}}{\boldsymbol{b+d}}$$

There are cases when accuracy is not adequate measures. The accuracy determined by the Accuracy equation can’t be an adequate measure of performance when the number of negative cases is much higher than the number of positive cases. If there are two classes and one is significantly smaller than the other, it is possible to obtain high accuracy if all instances are classified in a larger class.

For example, if the number of reliable suppliers is much higher than the number of unreliable suppliers, a system that classifies all suppliers as reliable would have high accuracy, but would be unusable because it would miss all the unreliable suppliers. In such cases, the sensitivity of the classifier is an important measure and its ability to correctly identify unreliable suppliers is crucial.

In machine learning, most classifiers assume equal importance of classes in terms of the number of instances and the level of importance, which means that all classes have the same significance. Standard techniques in machine learning are not successful when predicting a minority class in an unbalanced data set or when the false negatives are considered more important than false positives. In practical terms, unequal costs of inaccurate classifications are common, so that the asymmetric misclassification costs must be taken into account as an important factor.

Cost-sensitive classifiers adapting models to costs of misclassification in the learning phase, with the objectives to reduce the costs of misclassification rather than to maximize the accuracy of classification. Because many practical problems of classifications have different costs associated with different types of errors, various algorithms for the evaluation of the sensitivity of classification are used.

Complementarity is one of the important characteristics of the evaluation of classification models. Using the pairs measures can be displayed specific accuracy of classification models with somewhat opposed positions. For example, by varying the parameter selected modeling techniques can be at the expense of one of the specific measures to increase the accuracy of the model shown in another measure. This is an optimization problem in which the selection with the appropriate settings based on one measure, maximizes other measures. In some cases, the quality of the classifier needs to be expressed by a number, not a pair of dependent measures, which is achieved by using pairs measures. Using the pair's value of measures, one measure is fixed and is observed only in the second measure. Thus, for example, can be considered measures of accuracy with fixed value of the response to 20% and in this case the derived measure is called the precision of 20%.

Besides derived measure, there are measures that are not based on fixing one component of a pair of measure, for example F-measure, which is defined as follows:

$$F - Measure = \frac{2 * response * accuracy}{response + accuracy}$$

Another way to test the performance of the classifier is the ROC graph. ROC graph is the two-dimensional representation which on the X axis represents the false positive rate and the Y axis represents the true positive rate. Item (0,1) is the perfect classifier: classifies all positive and all negative cases correctly. This is (0, 1), because the false positive rate is 0 (zero), a positive real rate is 1 (all). Point (0, 0) is a classifier that predicts all cases to be negative, while point (1, 1) corresponds to the classifier which ensures that every case is positive. Point (1, 0) is a classifier that is incorrect for all classifications. In many cases, the classifier has a parameter which can be adjusted to increase the real positive rates at the cost of increasing false positive rates or reducing the false positive rate based on the dropping value of real positive rates.

Each setting parameter gives par value for a false positive rate and positive real rates and the number of such pairs can be used to represent the ROC curves. Nonparametric classifier is presented ROC to one point, which corresponds to the par value of the false positive rate and positive real rate. Below figure shows an example of a ROC graph with two ROC curves and two ROC points marked P1 and P2.

Nonparametric algorithms produce a single ROC point for a particular data set. Characteristics of ROC graph are:

- ROC curve or point is independent of the distribution of the class or the cost of errors.
- ROC graph contains all the information contained in the matrix of errors.
- ROC curve provides a visual tool for testing the ability of the classifier to correctly identify positive cases and negative cases that were incorrectly classified.

The area under the one ROC curve can be used as a measure of accuracy in many applications, and it is called the measurement accuracy based on the surface.

The use of the classification accuracy of the classifier comparison is not an adequate measure unless the cost classification and distribution of class is unknown, but one classifier must be chosen for each situation. They propose a method of assessing the classifier using the ROC graph, imprecise costs and distribution of class. Another way of comparing ROC points is the equation that balances accuracy with Euclidean distances from the perfect classifier, i.e. from the point (0, 1) on the graph. In this way we include weighting factors that allow us to define the relative cost of improper classification, if such data are available.

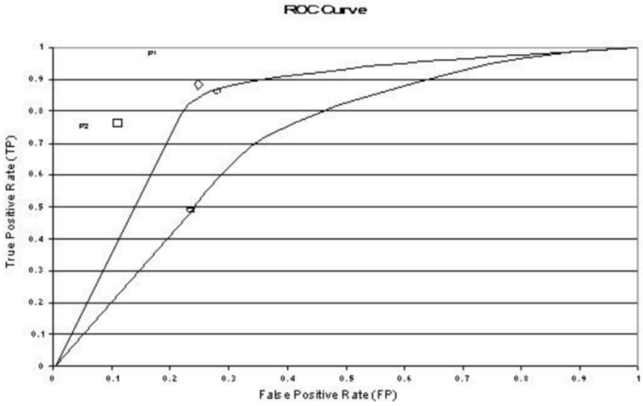

Figure 9:ROC Graph

3.1.2 Demand Forecasting

Time series models are fitted on historical data and are used to predict volume (i.e. sales) over a period of time.We are examining in this study the simple moving average, simple exponential smoothing, Holt's model, winter's model[24] & ARIMA model.

Simple moving average: To determine the next forecast, this method requires historical data. This method will be useful if we assume that demand will stay steady over time or then the demand has no trend or seasonality. The advantages of this method are the simplicity of us, easy to understand and implement [25].

The formula of moving average is expressed as:

The level in period t is intimated as the average demand over the recent N, the equation is shown below:

$L_t = (D_t + D_{t-1} + \dots + D_{t-N+1})/N$

The current forecast for all future periods is the same and is based on the current estimates of level. The forecast is stated as:

$$F_{t+1} = L_t \ and \ F_{t+n} = L_t$$

Where:

Dt = Observed demand in time period t

Ft+1 = Forecasted demand for t+1 made in time period t

N = Number of time periods

Simple exponential smoothing: This method is used if there is no trend or seasonality in the demand. The Simple Exponential Smoothing technique takes into account the smoothing factor which helps in reacting more strongly to recent changes in demand [25]. The formulas are expressed below:

The Lo is consider to be the average from all historical data, given demand data from period 1 until n:

$$L_0 = \frac{1}{n}\int_{i=1}^{n} D_i$$

For all future periods is the same as the current level and is given as:

$$F_{t+1} = L_t and \ F_{t+n} = L_t$$

Where:

α = Smoothing constant for level (0<α< 1)

Dt = Actual demand in time period t

Ft = Forecast made in period t

Holt's model: This method uses if there is a level and a trend in the demand but no seasonality thus, uses linear regression between the demand and the time to determine the initial level

The future forecast period is expressed as:

$$F_{t+1} = L_t + T_t \ and \ F_{t+n} = L_t + nT_t$$

Then, the revised estimate is the weighted average that has been observed and also from the old estimates. How to revise the approximate levels are as follows:

$$L_{t+1} = \alpha D_{t+1} + (1-\alpha)(L_t + T_t)$$
$$T_t = \beta(L_{t+1} - L_t) + (1-\beta)T_t$$

Where:

α = Smoothing constant for level (0<α<1)

β = Smoothing constant for trend (0<β<1)

n = Number of periods ahead to be forecast

F_{t+n} = Holt's forecast for period t+n

Winter's model: This model will be appropriate if the demands have a level, trend and also a seasonal factor. The systematic component of demand is level plus trend and times seasonal factor. The level component would represent the average demand, the trend component would represent the increase or decrease in demand over time, and the seasonal component would represent the cyclical pattern of demand throughout the year. For instance, demand might be higher during certain times of the year due to various factors.

To start the forecast, we must find out the level, trend, and seasonal factors first using a static model [24].

To find out the future forecast is expressed as:

$$F_{t+1} = (L_t + T_t)S_{t+1} \text{ and } F_{t+1} = (L_t + lT_t)S_{t+1}$$

Then, revise the estimate level, trend, and seasonal factors as follows:

$$L_{t+1} = \alpha(D_{t+1}/S_{t+1}) + (1-\alpha)(L_t + T_t)$$
$$T_{t+1} = \beta(L_{t+1} - L_t) + (1-\beta)T_t$$
$$S_{t+p+1} = \gamma(D_{t+1}/S_{t+1}) + (1-\gamma)S_{t+1}$$

Where:
α = Smoothing constant for level (0<1)
β = Smoothing constant for trend (0<β<1)
γ = Smoothing constant for seasonality factor
n = Number of periods in forecast lead period
p = Number of periods in seasonal cycle
F_{t+n} = Winter's forecast for period t+n

And to measure demand forecast error, we can use mean Squared error (MSE) , mean absolute deviation (MAD) and mean absolute error (MAPE) [26].

Mean Squared Error measures the average of the squared differences between the forecasted and observed values. This method is appropriate when forecast error has a distribution that is symmetric about zero and if the cost of large error is much larger than the gains from very accurate forecast. The formula is:

$$MSE = \frac{\sum \quad (Forecast\ Errors)^2}{n}$$

Where:
Fe= Forecast error of demand value at time t
n = Number of time period

Mean Absolute Deviation is done by using the absolute value of the estimated error divided by the number of periods. This method is appropriate if the forecast error does not have a symmetric distribution and if the cost of a forecast error is proportional to the size of the error. It is expressed as shown as below:

$$MAD = \frac{\sum \quad |Actual - Forecast|}{n}$$

Where:
Ft = Forecast demand value at time t
Dt = Actual demand at time t
At = Absolute value of forecast error at time t
n = Number of time period
Mean Absolute Percent Error is useful to avoid the large number of MSE and MAD when the forecast item is measured in thousands. This method is appropriate when the forecast has significant seasonality and demand varies considerably from one period to the next. This MAPE is calculated as:

$$MAPE = \frac{\int_{i=1}^{n} \quad 100|Actual - Forecast|/\ Actual}{n}$$

Where:
Ft = Forecast demand value at time t
Dt = Actual demand at time t
At = Absolute value of forecast error at time t
n = Number of time period

Autoregressive integrated moving average (ARIMA) model:
The ARIMA model has been extensively studied and applied in studies of forecast due to their attractive theoretical properties and because of the various empirical supporting evidences. In addition, ARIMA model has equivalence with most models of exponential smoothing, except for the multiplicative form of Holt-Winters[28].
Any forecasting method involves two steps: (i) the analysis of time series and (ii) the selection of the forecasting model that best fits to the data set. Likewise, for ARIMA, is used a similar sequence of analysis and selection by decomposition methods and regression. In this sense, this section is divided in two main parts: first, the basic concepts of autoregressive moving average models that support the ARIMA model are described, and then, the application of this model in time series forecasting.

The regression model takes the form:

$$Y_t = b_0 + b_1X_1 + b_2X_2 + \ldots\ldots. + b_pX_p + e_t$$

Where is the predicted variable, to are explanatory variables, to are linear regression coefficients and represents the error. If, however, these variables are defined as $X_1 = Y_{t-1}$, $X_2 = Y_{t-2}$, $X_3 = Y_{t-3}$ …$X_p = Y_{t-p}$, now becomes

$$Y_t = b_0 + b_1Y_{t-1} + b_2Y_{t-2} + \ldots\ldots. + b_pY_{t-p} + e_t$$

This equation still represents a regression equation, but differs from above since it has different explanatory variables that are, in fact, previous values of the predictor variable , called autoregressive (AR). Just as it is possible to regress past values of a series again, there is a time series model that uses past errors as explanatory variables:

$$Y_t = b_0 + b_1e_{t-1} + b_2e_{t-2} + \ldots\ldots. + b_pe_{t-q} + e_t$$

In this equation, a dependency relationship is established between successive errors and the equation is called a moving average model (MA). Many stationary random processes cannot be modeled purely as moving or as auto-regressive averages because they have qualities of both types of processes [4]. In this situation, the autoregressive (AR) can be effectively connected to the moving average model to form a common and general class of time series models called autoregressive moving average models (ARMA).

The ARMA model can only be used on stationary data. In practice, many of the time series are non-stationary, so that the characteristic of the underlying stochastic process changes over time. To extend the use of the ARMA model for nonstationary series is necessary to differentiate the data set. In this situation, the model is now called the autoregressive integrated moving average (ARIMA).

Any homogeneous non-stationary time series can be modeled as an ARIMA process of order (p,d,q). The practical problem is to choose the most appropriate values for p, d e q, i.e. specify the ARIMA model. This problem is solved in part by examining the autocorrelation function and partial autocorrelation function for the time series of interest. The first step is to determine the degree of homogeneity d, that is, the number of times that the series needs to be differentiated to produce a stationary series. Then it examines the correlation and partial autocorrelation function to determine possible specifications of p and q.

p: Lag order - the number of lag observations included in the model, also called the lag order

d: Degree of differencing - the number of times the raw observations are differenced

q: Order of moving average – the size of the moving average window

On a high level, we can abstract the prediction algorithm using ARIMA as following:

- Define the model ARIMA (p, d, q)
- Fit the defined model on the training dataset (smaller subset of the original time series)
- Run the predictions on the latter data using the fitted model
- If the outcome doesn't fall into the defined confidence level (the correctness of the prediction compared to the existing data), adapt the properties p, d and q of the model

The main benefit of introducing the ARIMA model into the sales predictions for the wholesale industry is the possibility of running the model automatically as the significant amount of data arrives into the data warehouse. Usually, the process of traditional planning and forecasting includes a vast amount of manual work using specialized tolls, while our implementation of the model on the top of the existing data warehouse can automatically adjust the parameter models through the auto_arima function even without any user intervention. The models can be fitted on a daily basis, providing the end users a possibility of faster determination of the strategic and tactical decisions without the delay which typical human intervention includes.

4. Conclusion

Contrary to the B2C setting, the B2B setup suffers from lack of enough actionable studies. Thus, while conducting our study in the B2B sector, we propose a methodology for retailer segmentation, customer propensity and demand forecasting based on the proposed models using the power of data to improve the decision making.

We have addressed a set of problems from a manufacturer-retailer-consumer chain. Because of their intermediary role in the relationship between a manufacturer and consumers, retailers’ role becomes significant for the survival of the manufacturer. We have provided a detailed algorithmic approach a B2B retail setup may use to enhance their business. Our approach mentions customer segmentation as a tool to understand customer behavior and derive actionable insights. We also provided how those segments can be fed into various other modeling setups like customer propensity and demand forecasting to obtain a more precise outcome. All the suggested techniques, if implemented with minimal bias and truthful data, will help the business obtain its desired goals and sustain in the industry.

The future scope of the study in the B2B sector is vast and promising. The proposed methods have the potential to revolutionize the retail setup. By exploring additional techniques and models to improve customer segmentation, customer propensity and demand forecasting, businesses can adapt to make more informed data driven decisions.

References

[1] A. Parvaneh, H. Abbasimehr and M. Tarokh, "Integrating ahp and data mining for effective retailer segmentation based on retailer lifetime value", Journal of Optimization in Industrial Engineering, vol. 5, no. 11, pp. 25-31, 2012

[2] C. Marcus, "A practical yet meaningful approach to customer segmentation", Journal of Consumer Marketing, vol. 15, no. 5, pp. 494-504, 1998.

[3] H. H. Chang and S. F. Tsay, "Integrating of som and k-means in data mining clustering: An empirical study of crm and profitability evaluation", Journal of Information Management, vol. 11, no. 4, pp. 161-203, 2004.

[4] C. Rygielski, J.-C. Wang and D. C. Yen, "Data mining techniques for customer relationship management", Technology in Society, vol. 24, no. 4, pp. 483-502, 2002.

[5] V. A. Stuntebeck, "B2b customer segmentation: Important considerations when segmenting business customers", October 2012.

[6] A. Hughes, Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable Customer-Based Marketing Program, McGraw-Hill Companies, Incorporated, 2006.

[7] R. Kohavi and R. Parekh, "Visualizing rfm segmentation", SDM, pp. 391-399, 2004.

[8] K. Tsiptsis and A. Chorianopoulos, Data Mining Techniques in CRM: Inside Customer Segmentation, Wiley, 2011.

[9] A. Mesforoush and M. Tarokh, "Customer profitability segmentation for smes case study:network equipment company", International Journal of Research in Industrial Engineering, vol. 2, no. 1, 2013.

[10] R. Blattberg, B. Kim and S. Neslin, Database Marketing: Analyzing and Managing Customers, Springer, 2008.

[11] Kaleva, Henri, and Johanna Småros “Machine Learning in Retail Demand Forecasting.” The

Complete Guide to Machine Learning in Retail Demand Forecasting, July 7,2023.[Online].Available: https://www.relexsolutions.com/resources/machine-learning-in-retail-demand-forecasting/

[12] D. A. Kandeil, A. A. Saad and S. M. Youssef, "A Two-Phase Clustering Analysis for B2B Customer Segmentation," 2014 International Conference on Intelligent Networking and Collaborative Systems, Salerno, 2014, pp. 221-228, doi: 10.1109/INCoS.2014.49.

[13] Prashant Sharma, 'Understanding K-means Clustering in Machine Learning(With Examples)' https://www.analyticsvidhya.com/blog/2021/11/understanding-k-means-clustering-in-machine-learningwith-examples/

[14] Nikita Sharma, 'Understanding the Mathematics behind K-Means Clustering' https://heartbeat.comet.ml/understanding-the-mathematics-behind-k-means-clustering-40e1d55e2f4c

[15] Zach, "What is the Rand Index? (Definition & Examples)" https://www.statology.org/rand-index/

[16] Dhanya Thailappan, 'Understand The DBSCAN Clustering Algorithm!' https://www.analyticsvidhya.com/blog/2021/06/understand-the-dbscan-clustering-algorithm/

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" dbscan.pdf (uh.edu)

[18] 'Hierarchical Clustering / Dendrogram: Simple Definition,Examples' https://www.statisticshowto.com/hierarchical-clustering/

[19] Sean Benhur,'Hierarchical Clustering: Agglomerative + Divisive Clustering' Hierarchical Clustering: Agglomerative + Divisive Explained | Built In

[20] M A Syakur et al 2018 IOP Conf. Ser.: Mater. Sci. Eng. 336 012017

[21] Mohammed Alhamid, 'Ensemble Models' https://towardsdatascience.com/ensemble-models-5a62d4f4cb0c

[22] Amy, 'Bagging vs Boosting vs Stacking in Machine Learning'. Bagging vs Boosting vs Stacking in Machine Learning | by Amy @GrabNGoInfo | GrabNGoInfo | Medium

[23] Breiman, L. (2001) Random Forests. Machine Learning, 45, 5-35. https://doi.org/10.1023/A:1010933404324

[24] J. R. Quinlan. Induction of decision trees. Mach. Learn., 1(1):81–106, March 1986

[25] Novaković, Jasmina Dj, AlempijeVeljović, Siniša Ilić S., Željko Papić, and Milica Tomović. 2017. Evaluation of classification models in machine learning. Theory and Applications of Mathematics & Computer Science 7,(1)(Spring):39-46, https://www.proquest.com/scholarly-journals/evaluation-classification-models-machine-learning/docview/1922445698/se-2 (accessed September 1, 2023).

[26] Chopra, S., and Meindl, P. (2016). Supply Chain Management; Strategy, Planning, and Operation. Edinburgh Gate: Pearson.

[27] S. Lakshmi Anusha, S. A. (2014). Demand Forecasting for the Indian Pharmaceutical Retail: A Case Study. Journal of Supply Chain Management Systems, 5.

[28] Heizer, J., and Render, B. (2014). Operations Managament; Sustainability and Supply Chain Management. Edinburgh Gate: Pearson.

[29] Yunishafira, Affiya. "Determining the Appropriate Demand Forecasting Using Time Series Method." KNE Publishing.Accessed September 2, 2023..[Online].Available: https://knepublishing.com/index.php/Kne-Social/article/view/3156/6693.

[30] Makridakis, S., Wheelwright, S. C., Hyndman, R. J.,Forecasting: methodsandapplications, 3a. Ed. EUA: John Wiley& Sons,1998, 642pp.

[31] Armstrong, J. S., Morwitz, V. G., Kumar, V.,Sales forecasts existing consumer products and services: do purchase intentions contribute to accuracy? International Journal of Forecasting, Vol. 16, No. 3, 2000, pp. 383-397

[32] T. Hlupić, D. Oreščanin and A. -M. Petric, "Time series model for sales predictions in the wholesale industry," 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 2020, pp. 1263-1267, doi: 10.23919/MIPRO48935.2020.9245255.