MATHEMATICS AND COMPUTATIONAL SCIENCES

Vol 5(3), 2024, pp:58-79
DOI:10.30511/mcs.2024.2032592.1189
ISSN:2717-2708
http://mcs.qut.ac.ir/

Qom University of Technology

Predicting customer churn in the fast-moving consumer goods segment of the retail industry using deep learning

Moein Mahdi[1], Mehdi jabbari *[2]

Abstract: The non-contractual environment, many brands, and substitute products make customer retention relatively tricky in the fast-moving consumer goods market .In addition, there is no such thing as a completely loyal customer, as most buyers purchase from several almost identical brands. If the customer leaves the transaction without notice, the company may need help responding and compensating. Companies should proactively identify potential customers before they leave the deal. Transactional data, readily available in point of sale (POS) systems, provides a wealth of information that can be harnessed to extract customer transactions and analyze their purchase patterns. This offers a robust foundation for predicting and preventing customer churn. This research shows how transactional data features are generated and are essential for predicting customer churn in the fast-moving consumer goods sector of the retail industry. This research presents data concerning the customers of a capillary sales and distribution company in the food industry. We have implemented standard machine learning methods with the available data in this research. However, we have also employed advanced deep-learning techniques to enhance our predictive capabilities. The results and accuracy of these methods, including Convolutional Neural Network (CNN) and Restricted Boltzmann Machine (RBM), have been thoroughly compared, providing a solid basis for our findings.

Keywords: Predicting customer churn, Customer retention, Retail industry (capillary sales), Customer loyalty, Fast-moving consumer goods.
2020 Mathematics Subject Classification: 68T07
Receive: 16 June 2024, **Accepted:** 21 September 2024

1 Introduction

In the fiercely competitive FMCG industry, companies increasingly realize the value of maintaining their existing customer base. This strategic shift helps reduce marketing costs and paves the way for increased customer loyalty and satisfaction, which are crucial for sustained business growth. Customer retention is a pressing issue across industries. Research consistently demonstrates that the cost of retaining a customer is significantly lower than attracting a new one [21]. This is primarily due to the high marketing costs associated with customer

[1]Department of computer science, Islamic Azad University, Naragh Branch, Iran
[2]*Corresponding author: Department of computer science, Qom University of Technology, Qom. Iran, Email: jabbari@qut.ac.ir

acquisition, making customer retention a more cost-effective strategy. For this reason, along with increasing competition, retaining existing customers is also very important. Customer loss usually happens gradually, not suddenly. By proactively analyzing customers' purchase history patterns, we can take an active approach to predicting churn. This empowers us to identify potential churners and gives us the control to take preventive measures before they leave, ensuring a more stable customer base. In today's complex FMCG distribution environment, several factors cause dissatisfaction and lead to customer churn [5]. For example, quality problems, inefficiency of distribution channels, offering new products, dynamic communication of customers with each other through social networks, aggressive competitive campaigns, laws and regulations, price changes, etc. [16] Accordingly, identifying churning customers who behave similarly and the root causes associated with their performance enables companies to improve customer retention. [7] Developing accurate and comprehensive churn prediction models is not just a step but a crucial and urgent task for FMCG companies. These models provide actionable business intelligence by identifying early signs of potential churn, enabling companies to take timely and effective measures to retain valuable customers and ensure more sustainable business growth. Therefore, they can proactively modify their marketing strategies to address the concerns of identified customers that are worth keeping, improving the bottom line. [17]

A slight improvement in the accuracy of predicting churn can have a significant positive impact on profitability. Attracting a new customer costs 5 to 6 times more than retaining an existing customer. [21] There is considerable evidence in relevant sources for the cost rate of new customer acquisition compared to customer retention, Such as 12 times [1], 20 times [16], and even 25 times [8]. Therefore, it is undeniable that acquiring a new customer is much more expensive than retaining an existing one. Furthermore, a 5% improvement in customer retention rates results in a 25-95% profit increase [8].

For a suitable strategy for customer retention, it is necessary to analyze the factors that encourage customers to leave the transaction (i.e., customer churn). Also, the profitable factors for maintaining and continuing the transaction should be identified [12]. Therefore, available, feasible, and accurate flow forecasting models must identify customers' intentions to leave the transaction and the reasons behind their behavior [5].

This research focuses on two aspects when predicting customer churn in the food retail industry. The first aspect is based on the features related to the model. Instead of using customer buying trends to cluster people, these values are created as attributes and transferred to the model. Therefore, different features are made for each customer to allow the model to learn and identify patterns for each individual. For this reason, two databases have been created to test and evaluate how to display data for predicting falls. The second aspect is the implementation of the algorithm.

This study is unique in that it uses deep learning to predict customer churn in the fast-moving consumer goods sector of the food industry. The strength of deep learning is that it can reveal hidden patterns in existing databases. The article is divided into five parts: the second part is the research literature, the third part is the research method, the fourth part is the evaluation results, and the fifth part is the conclusion.

2 Research literature

Miguis et al. (2013). [11] used logistic regression models and multivariable adaptive regression lines (MARS) to conclude that when variable selection methods are not used, MARS performs better than logistic regression. Jahrami et al.'s (2017) article is titled "Managing B2B customer churn, retention and profitability [19]."Decision trees, logistic regression, and adaptive boosting/adaptive boosting are among the models used in this research. Its private data set is the transaction records of 11,021 business customers in one year. In the article "Detection of dangerous exit events against the occurrence of accidents to improve the efficiency of incentives to reduce

customer attrition [5]", Calcio et al.'s (2015) article, from the random model (Pareto NBD) and the private data set of 199,352 customers that Randomly selected from the top four RFM sectors during 3.5 years are used. In this article, customer exit events (DOIs) are used to estimate the probability of being active, and a long period between DOIs is considered a potential customer churn signal. Solistiani and Tejahanto's (2017) [18] article, we will find that the feature selection methods can significantly affect the prediction accuracy of the random forest algorithm, and the Chi-square feature selection method can increase the accuracy of the random forest to 83.2 % improvement. Shoaib (2018), in an article titled "Prediction of customer churn in a retail store through machine learning algorithms [15]". In this research, the K-means clustering technique is used for segmentation. In the article "Management of classification imbalance in predicting customer churn [4]" written by Bores and Vandel-Pool (2009) of gradient boosting machine and weighted random forest models (with random sampling). and advanced), CUBE sampling technique, random forest, and logistic regression have been used. In the article "Partial Defection of Loyal Customers (Behaviorally) in a Non-Contractual FMCG Retail Environment [3]" by Buckinx et al. (2005), we will find that there is no significant difference in terms of performance between the methods in PCC and AUC. The alternative classification used is not observed. In the article "Using machine learning techniques to predict churn/churn of top customers" article by Buckinx et al. (2002) [2], which uses logistic regression models, linear discriminant analysis, quadratic discriminant analysis, C4.5, Neural and simple biz networks and data sets of personal records of 158,884 customers with loyalty cards (85% of all customers) have been used for ten months and with 32 features. Vu, V. H. (2024). In paper “Predict customer churn using combination deep learning networks model” , propose a combined deep learning network models to predict customers leaving or staying at the bank. The proposed model consists of two levels, Level 0 consists of three basic models using three Deep Learning Neural Networks, and Level 1 is a logistic regression model. The proposed model has obtained evaluation results with accuracy metrics of 96.60%, precision metrics of 90.26%, recall metrics of 91.91% and F1 score of 91.07% on the dataset “Bank Customer Churn Prediction” [22]. Saha, S., Saha, et al (2024). In paper “ChurnNet: Deep Learning Enhanced Customer Churn Prediction in Telecommunication Industry”, proposed a novel customer churn prediction architecture namely ChurnNet to predict customer churn in TCI [13]. Subramanian, R. S, et al (2024). In paper “Ensemble-based deep learning techniques for customer churn prediction model. Kybernetes” the data are collected from the WSDM-KKBox’s churn prediction challenge dataset. Here, the time-varying data and the static data are aggregated, and then the statistic features and deep features with the aid of statistical measures and “Visual Geometry Group 16 (VGG16)”, accordingly, and the features are considered as feature 1 [16].

2.1 Convolutional Neural Network (CNN)

A convolutional neural network is a well-known deep learning architecture inspired by living organisms' natural visual perception mechanism. In 1959, Hubelweil discovered that the cells of the visual cortex of animals are responsible for detecting light in receptive fields. Inspired by this discovery, Kunihiko Fukushima proposed Neon in 1980 (AD), which can be considered the father of a convolutional neural network. In 1990, Likan et al. published a seminal paper that established the modern CNN framework and later improved it. They developed a multi-layer artificial neural network called LeNet-5 to classify handwritten digits. LeNet-5 has several layers like other neural networks and can be trained with a backpropagation algorithm. This method can obtain effective representations of the original image that enable the detection of visual patterns directly from raw pixels with little to no pre-processing. In a parallel study, Zhang et al. used an invariant artificial neural network (SIN) to recognize characters in an image. However, due to the lack of extensive training data and computing

power at that time, their networks needed help handling more complex problems. For example, large-scale image and video classification worked well.
Figure 1-2 depicts the hierarchical classification of a convolutional neural network.

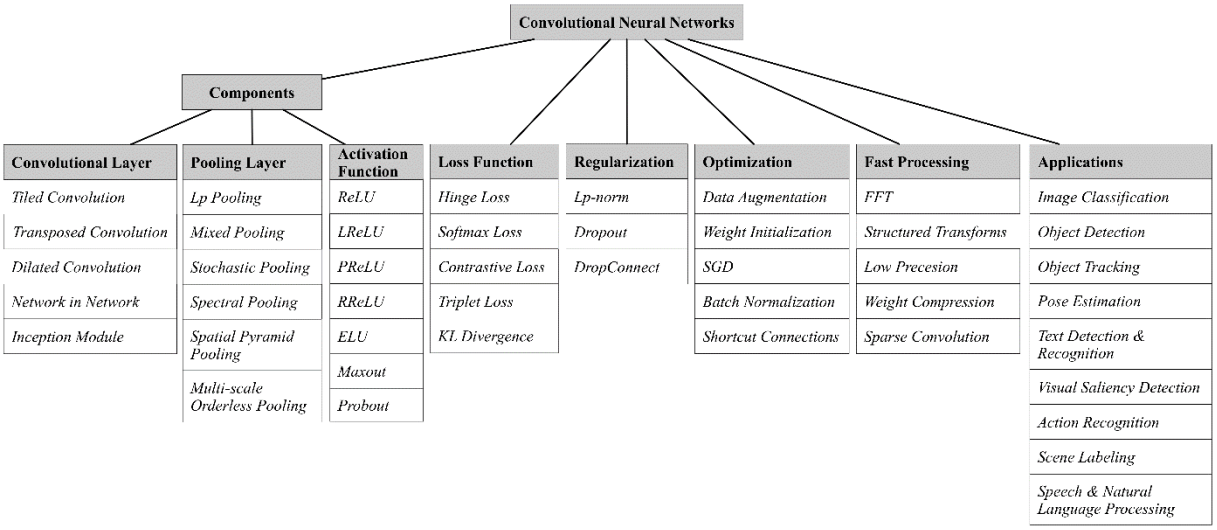
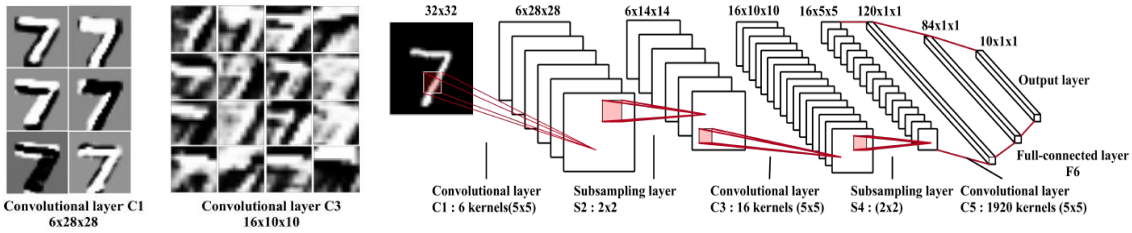


Figure 1-2 depicts the hierarchical classification of a convolutional neural network. [14]

Different convolutional neural network architecture types exist, but their primary components are similar. We consider the famous LeNet-5 example, consisting of three layers: convolution, integration, and fully connected. The purpose of the convolution layer is to learn to represent the features of the inputs. As shown in Figure 2-2 (right), the convolution layer consists of several kernels for calculating different feature maps. Specifically, each neuron has a feature map connected to a region of neighboring neurons in the previous layer. Such a neighborhood is called the neuron's receptive field in the last layer. The new feature map can be obtained by matching the input with a learned kernel and applying an element-wise nonlinear activation function to the



scrambled results. All input spatial locations share the kernel to generate each feature map. Complete feature maps are obtained using several different kernels.

Figure 2.2 Whole LeNet-5 network (right) – learned features (left) [23]

Figure 2-2 (right) LeNet-5 network architecture works well in digital classification. (Left) Visualization of LeNet-5 network features. Feature maps of each layer are displayed in a different block. Mathematically, the feature value at location (i, j) in the k feature map of the L layer of $z_{i,j,k}^{l}$ is calculated as follows:

$$z_{i,j,k}^{l} = w_{k}^{l} x_{i,j}^{l} + b_{k}^{l}$$

Where w_{k}^{l} and b_{k}^{l} are the weight vector and bias term of the k filter of layer l, respectively, and $x_{i,j}^{l}$ is the input patch at the center of location (i,j) of layer l. The kernel w_{k}^{l} l, which creates the feature map $z_{i,j,k}^{l}$, is shared. Such a weight-sharing mechanism has several advantages. Among other things, it can reduce the complexity of the model and make the training of the network easier; it introduces the activation function of nonlinearities to convolution, which is desirable for multi-layer networks to detect nonlinear features. [9]

2.2 Restricted Boltzmann Machine (RBM)

In 2006, Hinton et al. presented an efficient way to build deep networks called deep belief networks, which opened the research leap of deep learning. DBN first used the restricted Boltzmann machine to train the network weights layer by layer and then used the gradient descent method to fine-tune the weights. It is worth noting that the restricted Boltzmann machine and Auto-Encoder (AE) are both used as basic blocks in deep learning. However, unlike AE, RBM is an energy-based model, and RBM can also be viewed as a particular type of Markov Random Field (MRF). Generally, RBM provides a powerful tool for displaying the dependence structure between random variables. The restricted Boltzmann machine has attracted much attention in the artificial intelligence and machine learning community. RBM is mainly developed for classification and representation learning.

In RBM, we train products of experts (PoE) by maximizing the probability of recording information to learn weights. The gradient of the likelihood of entering the system in an RBM according to the connection weight can be expressed as the difference between data-dependent and model-dependent statistics. Sequential Gibbs sampling, fuzzy simulation, and continuous Markov chains are traditional methods of computing data- and model-dependent statistics, but they perform poorly in large datasets. Hinton developed adversarial divergence (CD) to train a PoE. Contrastive divergence also provides highly biased estimates of model-dependent statistics. RBM is an unsupervised learning method that can learn valuable data features.

The theories and applications of RBM have been widely studied during the last decade. For example, the original RBM was only suitable for addressing binary images in image processing. A series of RBM types is proposed to deal with authentic photos, such as Gaussian Binary Restricted Boltzmann Machine (GRBM), Covariance Restricted Boltzmann Machine (RBM), Mean and Covariance Restricted Boltzmann Machine (mcRBM), and Restricted Boltzmann Machine. Spiked and slab (ssRBM). Many types of RBM have been proposed to meet specific application requirements.

The restricted Boltzmann machine is a generative random network and includes a layer of observable units $\boldsymbol{v} = \{v_i\}_{i=1}^{D}$, and a layer of hidden units $\boldsymbol{h} = \{h_j\}_{j=1}^{J}$ with The parameters θ= {W,b,c}. The energy function and probability function of RBM are expressed as follows:

$$E(\boldsymbol{v}, \boldsymbol{h}; \theta) = -\sum_{i=1}^{D}\sum_{j=1}^{J} v_i W_{ij} h_j - \sum_{j=1}^{J} b_j h_j - \sum_{i=1}^{D} c_i v_i \,,$$

$$P(\boldsymbol{v}; \theta) = \sum_{h} P(\boldsymbol{v}, \boldsymbol{h}; \theta) = \sum_{h}\left(\frac{1}{Z(\theta)} exp\big(-E(\boldsymbol{v}, \boldsymbol{h}; \theta)\big)\right) = \frac{1}{Z(\theta)}\sum_{h} exp\big(-E(\boldsymbol{v}, \boldsymbol{h}; \theta)\big)$$

Where $v_i \in \{0,1\}$, $h_j \in \{0,1\}$, $\boldsymbol{W} = (W_{ij}) \in R^{D\times J}$are weights that connect visible units and hidden units. do

$\boldsymbol{c} = \{c_i\}_{i=l}^{D}$is the visible layer bias condition, $\boldsymbol{b} = \{b_i\}_{i=l}^{J}$ is the hidden layer bias condition, $Z(\theta) = \sum_v \sum_h exp\big(-E(\boldsymbol{v},\boldsymbol{h};\theta)\big)$ partition function and function $\mathrm{P(v;\theta)}$ can also be called marginal distribution of function $\mathrm{P(v,h;\theta)}$. [24].

3 Solution Method

In this chapter, we will first analyze the data with different machine learning algorithms: logistic regression, k-fold logistic regression, augmented logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), decision tree, and random forest. Moreover, in XG Boost, we show the implementation and results. Then, in the next part, we implement and compare the proposed method of this thesis, which is deep learning algorithms of complex neural networks (CNN) and restricted Boltzmann machine (RBM) with the same data.

3.1 Implementation with Machine Learning Methods

3.1.1 Data Preparation and Analysis

Data set: An actual data set obtained from a sales and distribution company in the fast-moving consumer goods sector of the capillary sales industry in two consecutive years, 2019 (P1 period) and 2014 (P2 period), was used. This data set includes fourteen characteristics: customer retention time in the company, the valuable life of goods in the store, buyer rating, customer category type, shopping cart price, shopping cart items, settlement type, settlement duration, category purchase value, the product, the amount of the product category purchased, the number of times purchased, the last purchase, the number of similar brands consumed, and purchase promotions.

Sampling: Instead of focusing on the entire customer base, we chose the most profitable, valuable, and behaviorally loyal customers. To determine this section, we considered the first 2/3 of customers based on their purchase value in the P1 and P2 periods. Then, we identified familiar customers who were in the choices for two consecutive years and introduced them as a valuable customer base. As a result, the first 2/3 of customers in terms of purchase value are 8,414 out of 12,622 customers in the P1 period and 8,738 out of 13,105 customers in the P2 period, and the number of joint customers in both periods is 3,936 customers. In this way, we selected joint buyers with a high rank in the top 2/3 customers in terms of purchase value for the P1 and P2 periods to sample valuable customers.

Loyal customers: Loyalty is defined based on the ratio of the purchase amount of the product category to the purchased amount. Valuable customers whose loyalty ratio is higher than 0.33 in the P1 period are considered loyal.

With a meticulous approach, we identified our loyal customers by removing 616 customers from a pool of 3,936 valuable customers based on their loyalty ratio.

Customer attrition: In a non-contractual environment in the FMCG market, customers can easily switch from one brand to another or divide their purchases between selected products. For this reason, customer churn should be carefully defined according to the desire to buy.

In this context, we divided loyal customers into two groups: 1. Customers whose loyalty ratio is between 0.33 and 0.5 in the P1 period; If their loyalty ratio is zero, i.e., no purchase in the P2 period, they are considered dropouts. 2. Customers whose loyalty ratio is higher than 0.5 in the P1 period: If their loyalty ratio in the P2 period is at least 0.5 and less than their P1 loyalty ratio, they are considered dropouts.
For example, a loyalty ratio of 0.2 or less in the P2 period is considered a drop. In contrast, a loyalty ratio of more than 0.2 in the P2 period and a loyalty ratio of 0.7 in the P1 period is regarded as no drop.

3.1.2 Data Analysis

The RFM model was used to analyze the data. RFM analysis is a market research model based on customer databases and direct marketing, mainly for retail products and professional services. RFM criteria are essential indicators for analyzing customer behavior because the number of customer purchases (Recency) and purchase value (Monetary value) indicate customer satisfaction, and their lifetime value and repeat purchases (Frequency) are signs of retention and customer loyalty. As part of the analysis, customers were grouped using a cluster dendrogram after rescaling the relevant data. This process identified eight distinct customer groups, as shown in the output (Figure 1-3).

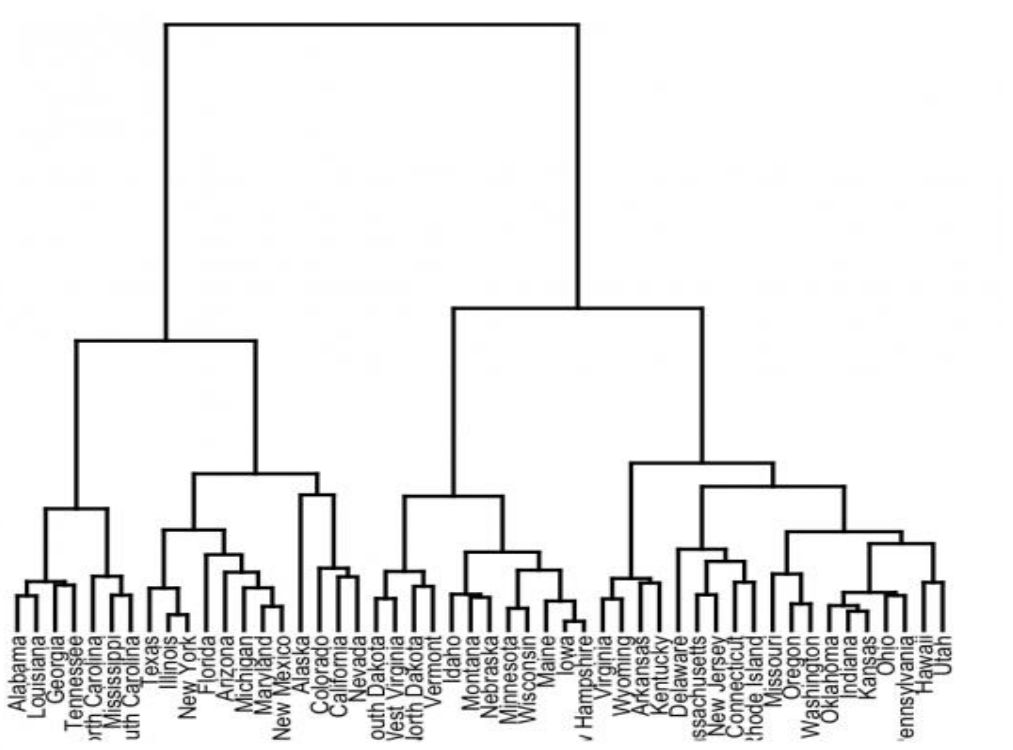


Figure 3-1 Cluster dendrogram

These groups are shown in the following three-dimensional RFM scatterplot, where X represents repetition.

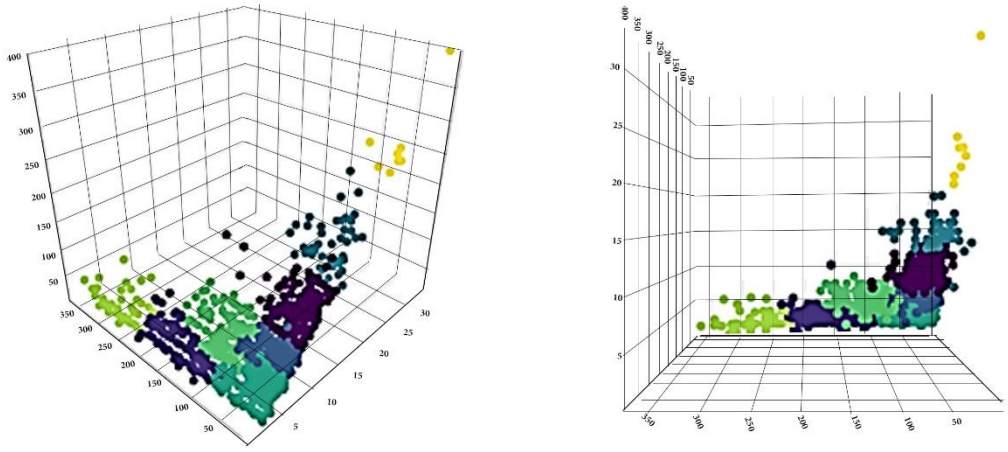


Figure 3-2 3D RFM

Y represents the number of customer purchases, and Z represents the value (Figure 2-3).

3.1.3 Segmentation by Clustering Method

Three different approaches are used to determine the optimal number of clusters. According to the diagram of the sum of squares within the cluster, the slope flattened after eight groups. The gap statistics method predicts that at least 6 clusters are enough. Our conclusive findings, as evidenced by the average silhouette chart (Figure 3-3), indicate that 8 to 9 clusters are the most suitable for our customer segmentation.

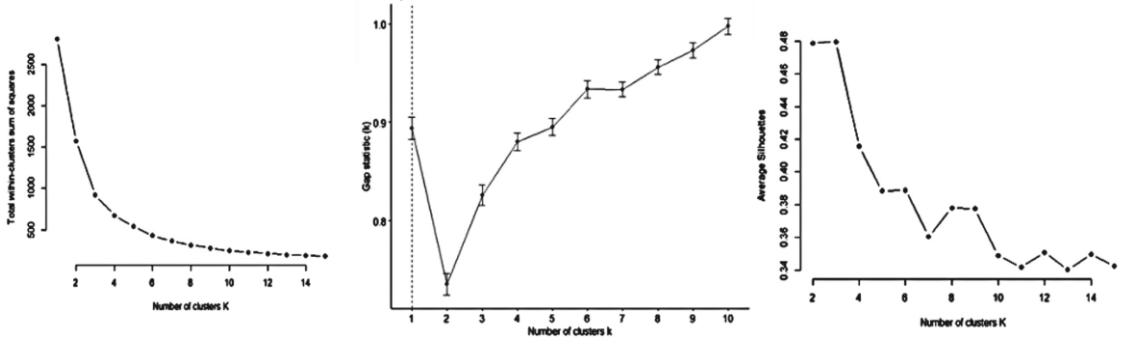


Figure 3-3 Diagram of intra-cluster sum of squares

(left), graph of gap statistics (middle) and graph of average silhouettes (right)

3.1.4 Exploratory Data Analysis, Visualization, and Feature Engineering

After testing different combinations of features, the set of features with the best performance was defined according to Table 1-3.

Table 3-1 Features used

LABEL	VARIABLES FORMULA	CATEGORY
SES_GROUP	SES_GROUP	FACTOR
PAX	PAX	FACTOR
KENT_KIR	KENT_KIR	FACTOR
CUSTOMER_CAT	CUSTOMER_CAT	FACTOR
SHELF_LIFE	SHELF_LIFE	FACTOR
GRADE	GRADE	FACTOR
REGION	REGION	FACTOR
DOIL	[OIL(P2)-OIL(P1)*1.2]	NUMERIC
DFMCG	[FMCG_TOTAL(P2)-FMCG_TOTAL(P1)*1.2]/[FMCG_TOTAL(P1)*1.2]	NUMERIC
DSPECIALSPM	[SPECIAL_SPM(P2)-SPECIAL_SPM(P1)*1.2]	NUMERIC
DSPECIALGRC	[SPECIAL_GRC(P2)- SPECIAL_GRC(P1)*1.2]	NUMERIC
PIRCE	[SPECIALPRICE(P2)- SPECIALPRICE(P1)*1.2]	NUMERIC
DCOUNTDIFCAT	[DCOUNTDIFCAT(P2)- DCOUNTDIFCAT(P1)]	NUMERIC

1. The 'factor ()' function was crucial in our data analysis process. It converted all character columns into factor columns, significantly enhancing our data interpretation.

2. A pivotal step in our data cleaning process was using the 'sum (is. na ())' function. This function was instrumental in identifying the total missing values in the data set, a crucial aspect of data quality assurance.

3. The 'Summary ()' function provided a comprehensive and detailed overview of the data set, including information on factor and numerical variables. For instance, it revealed 202 falling labels and 735 non-falling labels in the data set, painting a complete picture of the data.

4. Unselect function removes all columns; Include factors to show the correlations and density of all numerical variables with the ggpairs () function (Figure 5-3).

- Distribution of the DSPECIALGRC feature is not extended; zero values dominate the whole column, so these columns look like this. The skewness of the distribution of these features is positive and higher than the skewness of the normal distribution.
- The distance of the correlations shown above, from 1 and -1, indicates that there is no correlation between the features. The correlation value between RPRICE and DCOUNTDIFCAT is very close to zero. This finding is particularly intriguing as they are expected to be parallel, sparking further curiosity and interest.
- Scatters between each pair indicate the distribution of observations between them, and the line on them suggests the direction of the observations.

5. The graph in Figure 4-3 is made with the corrplot () function, which shows the correlation between the features after rescaling. The Ggpairs diagram is also shown in Figure 5-3.

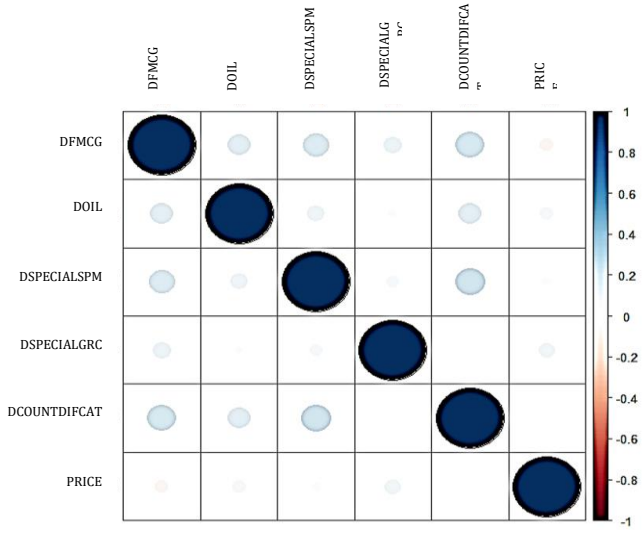


Figure 3-4Correlation diagram

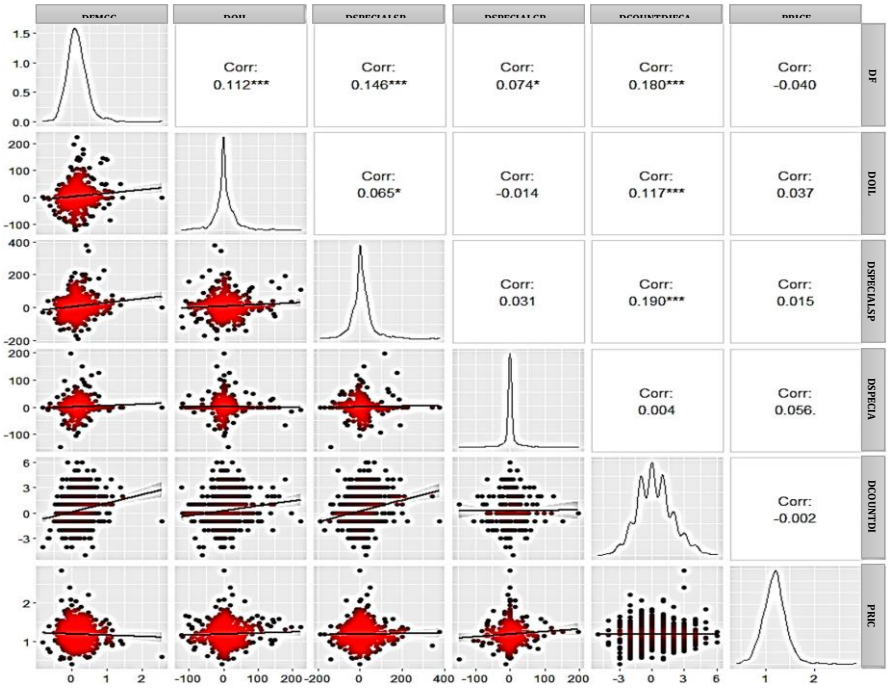


Figure 3-5 Ggpairs diagram

3.2 Suggested Method: Using Deep Learning Algorithms

3.2.1 Data Preparation and Parameter Selection

Data collection is essential because accurate data is needed for the results to be actual. As mentioned in the previous sections, this thesis uses an actual data set obtained from a sales and distribution company in the field of fast-moving consumer goods in the capillary sales industry.

Using the Kimball method, the available data is extracted, refined, and loaded (ETL).

- **Extraction**

Once the existing data has been cleared of any anomalies, the existing data may be extracted from the database and placed in the data warehouse. This is done by running scripts to extract data from the central database. The designed scripts are for truth tables and dimensions. The truth table contains a denormalized form of data that stores transaction data along with customer ID and inventory ID. Dimensions separate scripts are available for extracting descriptive status.

- **Refinement**

In this process, rules and aggregation are created. The number of repeat customers and receipts are pre-calculated to be pre-aggregated when reading the data count. In order to keep the data safe, the processing of customers' names, inventory brands, categories, and descriptions of the branch or each department will be removed from the data set.

- **Upload**

After the data is refined, it is loaded into a data warehouse that contains three dimensions and a truth table. This includes customer and inventory dimensions with primary keys, and indexes set up. The truth table contains inventory, customer and time identifiers, sales quantity, and value.

After a comprehensive analysis of the data mentioned in the previous section, the parameters are meticulously formed and transferred to the selected algorithms in the next step. Considering the purpose of the thesis, the parameters created revolve around customers' buying process. To study this, we have created two datasets based on different aspects of customer purchases. The first dataset is based on the number of customer purchases or recent customer purchases, purchase value, and repeat purchases (RFM). The recent purchases indicate the last time the customer did business with the company. The number of times the customer has gone to the supermarket is determined as the repeat purchase value, and the financial value spent by the customer is also determined as the purchase value. The second dataset includes 18 months of RFM data, providing a more comprehensive view of the customers' buying behavior over time. Customer attrition, also known as customer churn, refers to losing customers over time. It is a key concern for businesses, as it can significantly impact their revenue and profitability. In this study, we are particularly interested in understanding the factors contributing to customer attrition in the sales and distribution industry and how deep learning algorithms can help us predict and prevent it. Therefore, the data set is divided into two time frames or time windows to identify drops. The first window is the predictive window that identifies active customers. Activity is defined by customers who have transactions during this period. Active customers in the first window are labeled "non-churning," while

the remaining customers are labeled "churning". The second case is removed from the analysis because it was previously considered as a spill and was removed. The following window, the churn assessment window, classifies the remaining customers as churn or non-churn. If the customers make a transaction or trade during this period, they are non-falling; if no transaction is observed, they fall.

3.2.2 Implementation of Deep Learning Algorithms

Two deep learning algorithms are meticulously implemented on the discussed dataset. The dataset is carefully balanced to ensure the model doesn't overfit or create a majority class, maintaining an equal number of samples for both classes. Furthermore, outliers are meticulously removed before the data set is divided into training and testing. The division for this work is done through a random division of the data set in the ratio of 75 to 25, ensuring a fair representation of the data.

- **Convolutional Neural Network (CNN)**

A prerequisite for CNN inputs is that the features are placed in a matrix. Therefore, the class label and the features of the training and testing datasets are placed in two separate matrices. Before placing them in the matrix, the features are normalized and transformed.

The first dataset, with its 16 features, is meticulously designed into a four-by-four input matrix, ensuring the most precise representation of the data.

This network features a neural network with a three-by-three kernel. The choice of activation function, in this case, the sigmoid function, is significant as it ensures the outputs are binary. The subsequent layer is the sum integration layer, where the sum of the inputs is taken. This step is crucial as it helps control the fit by reducing the spatial size of the input.

A two-by-two kernel with a one-by-one step specifies the shift of the handle to the left, right, top, or bottom of the matrix. Then, in order to ensure that there is no misfit, random elimination of 0.1 is used. In the following, two fully connected networks are used, whereby the first network has five hidden layers, a sigmoid activation function, and random removal of 0.1. It is followed by the second fully connected network with three hidden layers.

The parameters used in the CNN, detailed in Table 2-3, play a crucial role in the model's training and learning process. The number of repetitions for training the model is set to 30, ensuring the model stays within the dataset. The batch size, set to 100, indicates the 'movement' or 'jumping' between parameters. The learning rate, set to 0.00625, determines the speed at which the model learns, while the momentum, set to 0.9, influences the convergence rate of the model.

These values are used because having a high learning rate may cause the system to drift and get stuck in local minima or maxima. Similarly, momentum is used as a convergence rate in deep networks. The penalty shown as Wd is set to 0.0003, which is used to increase the regularization. The second data set is sent to the designed CNN. The difference is in the number of inputs used. For this model, 25 features have been implemented along with a label. The features are converted into a 5x5 matrix. The same transformation is applied to the class label and test data.

Table 3-2 Complex neural network parameters

PARAMETER	VALUE
X	Train.Array
Y	Train.Y
NUM.ROUND	30
ARRAY.BATCH.SIZE	100
LEARNING.RATE	0.00625
MOMENTUM	0.9
WD	0.0003
EVAL.METRIC	Mx.Metric.Accuracy
EPOCH.END.CALLBACK	100

- **Restricted Boltzmann Machine (RBM)**

The data set is converted into binary for RBM, and a binary value separates each feature. This is a complex process because a binary identifier value is required for each feature defined in the dataset. Therefore, each customer has a vector of binary values representing their characteristics.

For example, Table 3-3 shows that the repeat purchase measure is grouped into nine categories. Therefore, each customer will have nine repeat purchase attributes for a particular month. If customers buy between 7 and 12 times, the binary value for this attribute is set to 1. For all other categories, the binary value is set to 0. This process is repeated for all features. The final number of features is 178 entries and one tag.

Table 3-3 An example of converting the repeat purchase feature to binary

FEATURE NAME	DESCRIPTION
FREQ1	Freq < 6
FREQ2	7 < Freq < 12
FREQ3	13 < Freq < 18
FREQ4	19 < Freq < 24
FREQ5	24 < Freq < 30
FREQ6	Freq > 30

The RBM training function requires the data to be in matrix format. Therefore, the data is converted and transferred to a matrix according to the function's needs. RBM architecture includes a visible layer and a hidden layer. One hundred seventy-eight nodes are used as input, visible in the visible layer, while 15 are available in the hidden layer. Every node in the visible layer is connected to every node in the hidden layer. A limitation of RBM is that nodes within a layer are not connected. This restriction turns the graph into a bipartite graph in which the nodes of the first layer are only connected to the nodes of the second layer.

Since the data is represented in binary values, the sigmoid activation function is implemented. This function limits the values to 1 and 0. One of the disadvantages of the sigmoid activation function is that if the neuron's weight is too high, the neurons may not learn. Therefore, to overcome this problem, an initial impulse of 0.004

during pre-training and a final impulse of 0.008 are determined. In addition, the learning rate is set to 0.05, which is the multiplier value for each period. Stochastic gradient descent is used to minimize the mean probability of negative registration. To do this, a partial derivative is defined by comparing the positive and negative phases. A contrasting divergence is used, and each variable is sampled according to the use of other variables. It is an iterative process in which a variable is randomly selected.

4 Computational Research

In this chapter, the effectiveness of the presented solution method is investigated. First, we implement the considered data set with machine learning methods and check its results. Then, we implement the proposed solution method and deep learning algorithms and compare their results. The results indicate the superiority of the proposed method.

4.1 Using Machine Learning Methods

The first method used before pruning is the decision tree. The training set's accuracy, sensitivity, and specificity were 81, 21, and 98 percent, respectively. After pruning, the accuracy of the test set is 77%, the sensitivity is 10%, and the specificity is 96%. The final decision tree produced after pruning is as follows (Figure 1-4):

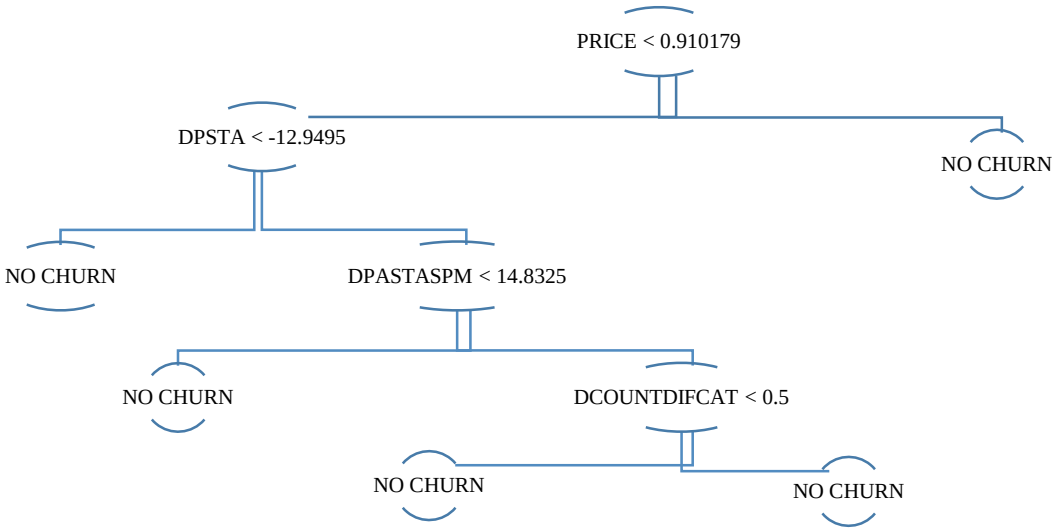


Figure 4-1 Decision tree

The most dominant variable was RPRICE and its final value was calculated as 0.91. After pruning, the accuracy of the test set was 77%, sensitivity 13%, and specificity 95%; Table 2-3 shows the clutter matrix of the test set of the decision tree after pruning.

Table 4-1 Clutter matrix of test set of decision tree after pruning

	CHURN	NO CHURN
CHURN	8	11
NO CHURN	53	210

Applying logistic regression by showing the participation of variables in the prediction of shedding made a better understanding of shedding behavior possible (Table 3-3).

Table 4-2 Variable share of logistic regression

Coefficients:	Estimate	Std. Error	z value	Pr(>\z\)	
(Intercept)	-0.474	0.620	-0.765	0.444	
CUSTOMER_CATA	0.547	0.342	1.598	0.110	
CUSTOMER_CATB	1.770	0.624	2.834	0.005	**
REGIONDowntown	0.766	0.512	1.495	0.135	
REGIONAroundtown	0.081	0.341	0.239	0.811	
REGIONSundry	0.099	0.343	0.287	0.774	
REGIONCentral	0.374	0.295	1.266	0.206	
REGIONUptown	0.739	0.349	2.114	0.034	*
DFMCG	0.365	0.373	0.979	0.328	
DOIL	-0.009	0.003	-2.926	0.003	**
DSPECIALSPM	0.005	0.002	2.223	0.026	*
DSPECIALGRC	0.015	0.006	2.522	0.012	*
PRICE	0.753	0.414	1.819	0.069	

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The most important variables were CUSTOMER_CAT, REGION, DFMCG, DOIL, DSPECIALSPM, DSPECIALGRC and RPRICE. The variables CUSTOMER_CAT, REGIONUptown, DFMCG, DSPECIALSPM, DSPECIALGRC and PRICECHANGE have positive coefficients.

RPRICE has a positive coefficient indicating that all things are equal.

When the average price of a certain product category increases, customers are more likely to churn. DOIL has a negative coefficient indicating: all things being equal; Customers who bought the oil at an inflated price are less likely to drop.

Among these variables, the most important factor is the increase in the average price of the product category and customer area; Because their coefficients are bigger. The accuracy of the training set was 78%, the sensitivity was 0.05%, and the specificity was 98%, and the accuracy of the test set was 77%, the sensitivity was 0.05%, and the specificity was 78% (Table 3-4).

Table 4-3 Confusion matrix of logistic regression test set

	CHURN	NO CHURN
CHURN	0	61
NO CHURN	2	219

The ROC curve also shows the algorithm's predictive power (Figure 6-3).

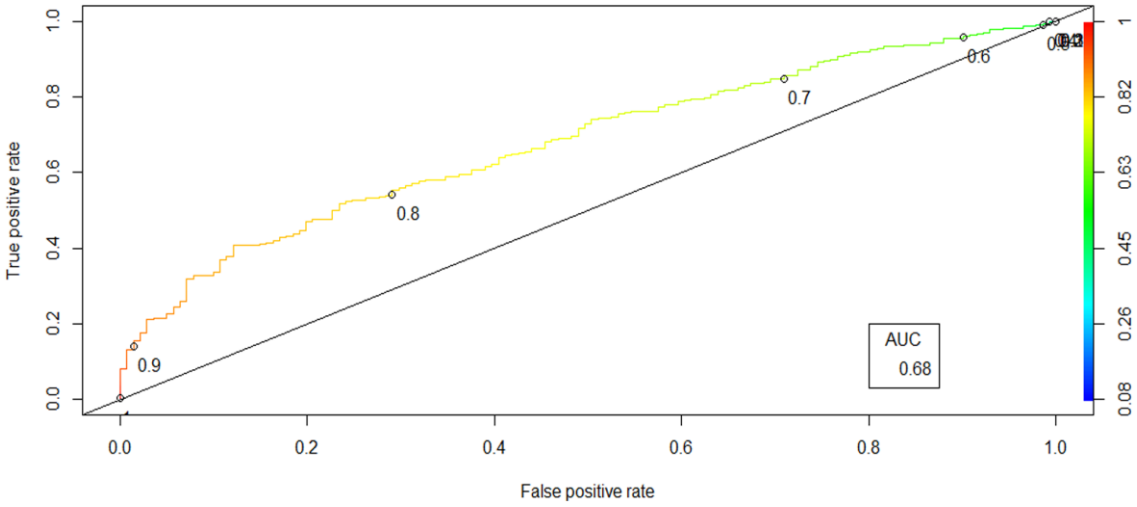


Figure 4-2 ROC curve

In order to apply the random forest technique, the data set was divided into experimental and training parts like the decision tree method. However, the SmoteClassif () function was used in the random forest algorithm to generate synthetic observations that depend on the "CHURN" label. When the model was run, k-fold helped build it with a value 5. Another critical concern in applying this method was determining the optimal n trees and mtry values (Figure 3-4).

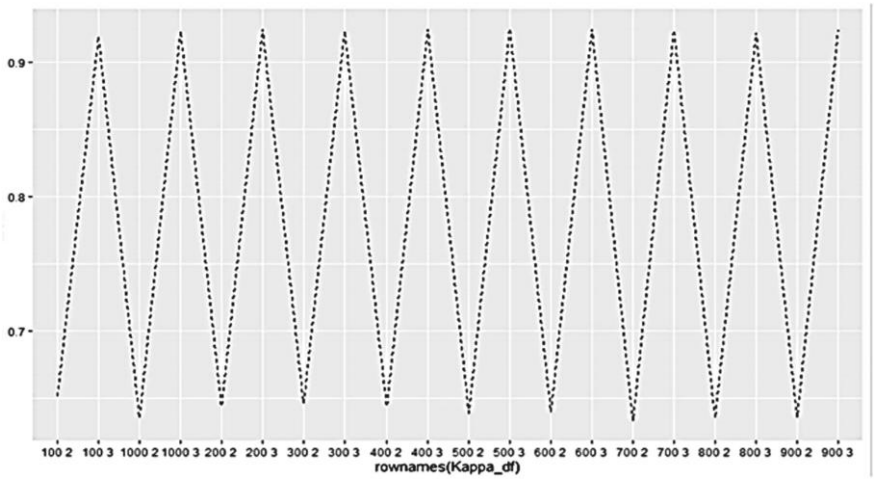


Figure 4-3 Random Forest classification diagram

The "mtry" setting parameter was kept constant at three due to the highest level of accuracy. The ntree parameter is also selected as 1000. Accuracy, sensitivity, and specificity were calculated as 74, 15, and 91% respectively (Table 4-4).

Table 4-4 Confusion matrix of random classifier

	CHURN	NO CHURN
CHURN	9	20
NO CHURN	52	201

XGBoost algorithm was another applied method, and the accuracy, sensitivity, and specificity were calculated as 87, 17, and 93%, respectively (Table 5-4).

Table 4-5 XGBoost clutter matrix

	CHURN	NO CHURN
CHURN	4	21
NO CHURN	19	273

Boost provides insight into drop behavior because it sorts variables based on their contribution (Figure 4-4).

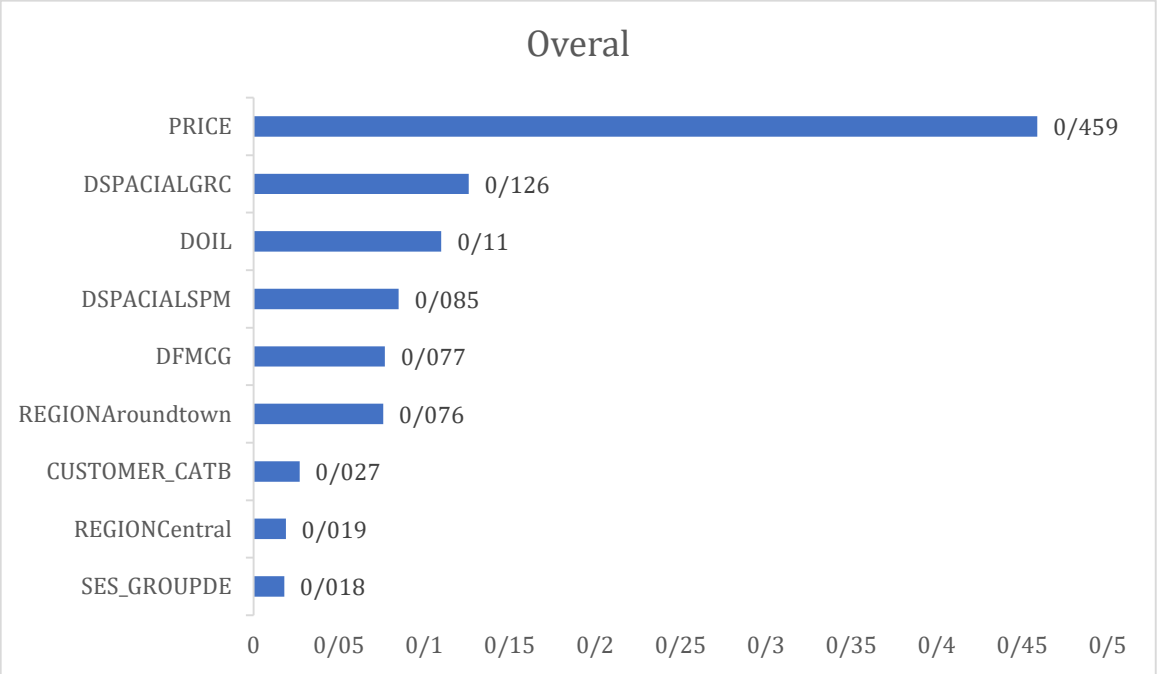


Figure 4-4 Importance of the XGBoost variable

PRICE (average price change of product category), DSPECIALGRC (price change of special product category), and DOIL (corn oil price change) were the most important factors. For Boosted logic; The accuracy of the test set was 87%, the sensitivity was 39%, and the specificity was 91% (Table 6-4).

Table 4-6 Boost logic scrambling matrix

	CHURN	NO CHURN
CHURN	9	27
NO CHURN	14	267

The QDA model calculated the test set's accuracy, sensitivity, and specificity to be 70, 48, and 72%, respectively (Table 7-4).

Table 4-7 QDA clutter matrix

	CHURN	NO CHURN
CHURN	11	82
NO CHURN	12	212

The LDA model is the next applied model, and the test set's accuracy, sensitivity, and specificity are calculated as 87, 0, and 94%, respectively (Table 8-4).

Table 4-8 LDA clutter matrix

	CHURN	NO CHURN
CHURN	0	17
NO CHURN	23	277

Finally, logistic regression with k times was used, and the test set accuracy, sensitivity, and specificity were calculated as 77, zero, and 98%, respectively (Table 9-4).

Table 4-9 Confusion matrix of k-fold logistic regression

	CHURN	NO CHURN
CHURN	0	11
NO CHURN	141	503

A summary table of the models used to categorize specific products is shown in the table below (Table 10-4).

Table 4-10 Comparison of machine learning models for specific product categories

Model	Accuracy %	Sensitivity %	Specificity %	Error Rate %	Precision %	Negative Predictive Value %
Decision Tree	77.305	13.115	95.023	22.695	42.105	79.848
Logistic Regression with k fold	76.794	0.000	97.860	23.206	0.000	78.106
Logistic Regression	77.660	0.000	78.214	22.340	0.000	99.095
Random Forest	74.468	14.754	90.950	25.532	31.034	79.447
XG Boost	87.382	17.391	92.857	12.618	16.000	93.493
Boosted Logic	87.066	39.130	90.816	12.934	25.000	95.018
LDA	87.382	0.000	94.218	12.618	0.000	92.333
QDA	70.347	47.826	72.109	29.653	11.828	95.848

4.2 Evaluating the Results of Machine Learning Models

The accuracy of the models used with base rates ("NOCHURN ratio") for all product categories is shown in the table below (Table 11-4).

Table 4-11 Comparison of accuracy of machine learning models

Model	Spaghetti Accuracy %	Special Accuracy %	Shaped Accuracy %	Oil Accuracy %
Decision Tree	68.280	77.305	96.369	68.794
Logistic Regression with k fold	62.628	76.794	96.491	56.308
Logistic Regression	63.441	77.660	96.067	65.957
Random Forest	62.903	74.468	93.041	64.539
XG Boost	75.000	87.382	98.951	45.098
Boosted Logic	55.435	87.066	96.970	49.020
LDA	66.304	87.382	93.939	45.098
QDA	52.174	75.393	98.718	61.765
Base	0.646	0.784	0.965	0.631

4.3 Evaluating the Results of Deep Learning Methods

To evaluate the performance of the algorithms in each data set, the label of the unseen data set is removed and placed in a separate vector. Then, the features are transferred to the trained models. After that, the predicted result is compared with the actual label to calculate the algorithm's accuracy.

- **Convolutional Neural Network (CNN)**

Comparing the two models, as shown in Table 12-4, the second model achieved a high accuracy of 74%, while the first model reached 68%. The sensitivity results for model 1 and model 2 are 59% and 66%, respectively. This shows that the percentage of non-drops is correctly classified for this algorithm. The classification of falls is shown in the obtained results for the degree of specificity. The models obtained 87% and 82% for model 1 and model 2. By analyzing the positive predicted values (Pos et al.), the model correctly classified 93% of non-falls in model 1 and 86% in model 2. By evaluating the negative predicted value (Neg Pred Value), which indicates the classification of customers who have dropped, the obtained algorithm is 41% and 60% for model 1 and model 2. The accuracy measure defining the classifiers' accuracy was 59% for model 1 and 66% for model 2. The recall criterion was used to understand the completeness, and 93% and 86% were obtained for model 1 and model 2, respectively. F1 measurement is calculated by considering the customers who are wrongly classified as churn and non-churn. Model 1 resulted in 73%, while Model 2 reached 74%. The F1 measurement is an efficient criterion for evaluating the accuracy of an experiment. This measure considers Precision and Recall together. The F1 measurement is one in the best situation and zero at worst.

Table 12-4 Evaluation Criteria for CNN

EVALUATION METRIC	MODEL 1	MODEL 2
SENSITIVITY	0.59	0.66
SPECIFICITY	0.87	0.82
ACCURACY	0.68	0.74
POS PRED VALUE	0.93	0.86
NEG PRED VALUE	0.41	0.60
PRECISION	0.59	0.66
RECALL	0.93	0.86
F1	0.73	0.74

- **Restricted Boltzmann Machine (RBM)**

By evaluating the criteria in Table 13-4, model 2 obtained the best results. The first criterion, Measurable sensitivity, calculates the ability to identify customers who will not drop correctly. The results obtained for model 1 and model 2 are 62% and 74%, respectively.

Similarly, the feature criterion tests the model's ability to classify customers who will churn correctly. This criterion is 87% for model 1 and 92% for model 2. Pos Pred Value reflects the probability of classifying real non-falls as non-falls. This criterion was obtained for model 1 and model 2, 83% and 92%, respectively. To analyze the correctness of the models in the definition of spillers, the results are 74% for model 1 and 77% for model 2. The accuracy measure defines the accuracy of the classifiers, so this criterion calculates the number of predicted non-precipitations divided by the total number of actual non-precipitations. This criterion was obtained for model 1, 67%, while for model 2 it reached 74%. To understand completeness, Recall is used, and the results obtained for model 1 and model 2 are 83% and 92%, respectively. The F1 measure is calculated by considering the customers who are wrongly classified as churn and non-churn. Model 1 resulted in 74%, while Model 2 reached 82%.

Table 4-13 Evaluation Criteria for RBM

EVALUATION METRIC	MODEL 1	MODEL 2
SENSITIVITY	0.67	0.74
SPECIFICITY	0.87	0.92
ACCURACY	0.77	0.83
POS PRED VALUE	0.83	0.92
NEG PRED VALUE	0.74	0.77
PRECISION	0.67	0.74
RECALL	0.83	0.92
F1	0.74	0.82

5 Conclusion

Customer satisfaction plays a significant role in the retail industry. Due to the increasing competition in this industry, companies must ensure that customers are satisfied with the services and quality of the products available. Predicting customer churn gives the company a competitive advantage in acting proactively and retaining customers with a high tendency to churn. This research discusses the importance of maintaining data available in food distribution companies or factories and the best parameters to predict spillage.

The first method used before pruning is the decision tree. The training set's accuracy, sensitivity, and specificity were 81, 21, and 98 percent, respectively. After pruning, the accuracy of the test set is 77%, the sensitivity is 10%, and the specificity is 96%.

The most dominant variable was RPRICE and its final value was calculated as 0.91. After pruning, the accuracy of the test set was 77%, sensitivity 13%, and specificity 95%. The most important variables were CUSTOMER_CAT, REGION, DFMCG, DOIL, DSPECIALSPM, DSPECIALGRC and RPRICE. The variables CUSTOMER_CAT, REGIONUptown, DFMCG, DSPECIALSPM, DSPECIALGRC and PRICECHANGE have positive coefficients.Among these variables, the most important factor is the increase in the average price of the product category and customer area; Because their coefficients are bigger. The accuracy of the training set was 78%, the sensitivity was 0.05%, and the specificity was 98%, and the accuracy of the test set was 77%, the sensitivity was 0.05%, and the specificity was 78%.

XGBoost algorithm was another applied method, and the accuracy, sensitivity, and specificity were calculated as 87, 17, and 93%, respectively. PRICE (average price change of product category), DSPECIALGRC (price change of special product category), and DOIL (corn oil price change) were the most important factors. For Boosted logic; The accuracy of the test set was 87%, the sensitivity was 39%, and the specificity was 91%.

To evaluate the performance of the algorithms in each data set, the label of the unseen data set is removed and placed in a separate vector. Comparing the two models, the second model achieved a high accuracy of 74%, while the first model reached 68%. The sensitivity results for model 1 and model 2 are 59% and 66%, respectively.

The first criterion, Measurable sensitivity, calculates the ability to identify customers who will not drop correctly. The results obtained for model 1 and model 2 are 62% and 74%, respectively. Similarly, the feature criterion tests the model's ability to classify customers who will churn correctly. This criterion is 87% for model 1 and 92% for model 2.

Machine learning and deep learning algorithms were used for implementation and compared with each other. From the implemented techniques, it is evident that analyzing customer purchase history behavioral patterns is essential to identify churn. This is because the number of customers in the retail industry is gradually falling. The results obtained from the proposed algorithms are satisfactory, with 83% in RBM and 74% in the implementation of CNN.

In future research, the analysis of customers' shopping baskets, that is, products that are purchased together regularly, can be investigated. This work provides insight and perspective for distribution companies on the best way to promote their products. In addition, for each product purchased for each customer, a pattern is defined to observe the purchasing habits of each product. Recurrent Neural Networks (RNN) will be applied to a sequence of customer actions. This algorithm will be used for sequence analysis due to its good performance. Based on purchase patterns, the warehouse manager can ensure that the defined items in the warehouse are available and available for customers to purchase.

Reference

[1] Abbasimehr, H., Setak, M., Tarokh, M. J. (2011). A neuro-fuzzy classifier for customer churn prediction. International Journal of Computer Applications, 19(8), 35-41.

[2] Buckinx, W., Baesens, B., Van den Poel, D., Van Kenhove, P., Vanthienen, J. (2002). Using machine learning techniques to predict defection of top clients. WIT Transactions on Information and Communication Technologies, 28.

[3] Buckinx, W., Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. European journal of operational research, 164(1), 252-268.

[4] Burez, J., Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.

[5] Calciu, M., Crie, D., Micheaux, A. (2015). Recognising dangerous drop out incidents as opposed to accidents to improve the efficiency of triggers reducing customer churn. Application to RFM customer segments of a fast moving customer goods retail chain. In Proceedings International Marketing Trends Conference.

[6] Cao, J., Yu, X., Zhang, Z. (2015). Integrating OWA and data mining for analyzing customers churn in E-commerce. Journal of Systems Science and Complexity, 28(2), 381-392.

[7] Figalist, I., Elsner, C., Bosch, J., Olsson, H. H. (2019). Customer churn prediction in B2B contexts. In Software Business: 10th International Conference, ICSOB 2019, Jyväskylä, Finland, November 18–20, 2019, Proceedings 10 (pp. 378-386). Springer International Publishing.

[8] Gallo, A. (2014). The value of keeping the right customers. Harvard Business Review.

[9] GU, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Chen, T. (2018). Recent advances in convolutional neural networks. Pattern recognition, 77, 354-377.

[10] Lipowski, M. M. (2018). Customer churn as a purchasing journey stage.

[11] Miguéis, V. L., Camanho, A., e Cunha, J. F. (2013). Customer attrition in retailing: an application of multivariate adaptive regression splines. Expert Systems with Applications, 40(16), 6225-6232.

[12] Murphy, J. A. (2001). The lifebelt: the definitive guide to managing customer retention. John Wiley Sons.

[13] Saha, S., Saha, C., Haque, M. M., Alam, M. G. R., Talukder, A. (2024). ChurnNet: Deep Learning Enhanced Customer Churn Prediction in Telecommunication Industry. IEEE Access.

[14] Sharkas, M., Attallah, O. (2024). Color-CADx: a deep learning approach for colorectal cancer classification through triple convolutional neural networks and discrete cosine transform. Scientific Reports, 14(1), 6914.

[15] Shoaib, T.: Customers Churn Prediction in Retail Store (2018). https://doi.org/10.13140/RG. 2.2.30545.38242.

[16] Subramanian, R. S., Yamini, B., Sudha, K., Sivakumar, S. (2024). Ensemble-based deep learning techniques for customer churn prediction model. Kybernetes.

[17] Sulistiani, H., Tjahyanto, A. (2017). Comparative analysis of feature selection method to predict customer loyalty. IPTEK the Journal of Engineering, 3(1), 1-5.

[18] Sulistiani, H., Tjahyanto, A. (2017). Comparative analysis of feature selection method to predict customer loyalty. IPTEK the Journal of Engineering, 3(1), 1-5.

[19] Tamaddoni Jahromi, A., Stakhovych, S., Ewing, M. (2017). The impact of personalised incentives on the profitability of customer retention campaigns.

[20] Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1-9.

[21] Verbeke, W., Martens, D., Mues, C., Basins, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert systems with applications, 38(3), 2354-2364.

[22] Vu, V. H. (2024). Predict customer churn using combination deep learning networks model. Neural Computing and Applications, 36(9), 4867-4883.

[23] Yadav, B., Indian, A., Meena, G. (2024). Recognizing Off-line Devanagari Handwritten Characters Using Modified Lenet-5 Deep Neural Network. Procedia Computer Science, 235, 799-809.

[24] Zhang, N., Ding, S., Zhang, J., Xue, Y. (2018). An overview on restricted Boltzmann machines. Neurocomputing, 275, 1186-1199.