Using Machine Learning Technique for Analytical Customer Loyalty

Mohamed M. Abbassy [1†]
Moh.abbassy@gmail.com
Faculty of Computers and Artificial Intelligence, Beni-Suef University, Egypt

Summary
To enhance customer satisfaction for higher profits, an e-commerce sector can establish a continuous relationship and acquire new customers. Utilize machine-learning models to analyse their customer's behavioural evidence to produce their competitive advantage to the e-commerce platform by helping to improve overall satisfaction. These models will forecast customers who will churn and churn causes. Forecasts are used to build unique business strategies and services offers. This work is intended to develop a machine-learning model that can accurately forecast retainable customers of the entire e-commerce customer data. Developing predictive models classifying different imbalanced data effectively is a major challenge in collected data and machine learning algorithms. Build a machine learning model for solving class imbalance and forecast customers. The satisfaction accuracy is used for this research as evaluation metrics. This paper aims to enable to evaluate the use of different machine learning models utilized to forecast satisfaction. For this research paper are selected three analytical methods come from various classifications of learning. Classifier Selection, the efficiency of various classifiers like Random Forest, Logistic Regression, SVM, and Gradient Boosting Algorithm. Models have been used for a dataset of 8000 records of e-commerce websites and apps. Results indicate the best accuracy in determining satisfaction class with both gradient-boosting algorithm classifications. The results showed maximum accuracy compared to other algorithms, including Gradient Boosting Algorithm, Support Vector Machine Algorithm, Random Forest Algorithm, and logistic regression Algorithm. The best model developed for this paper to forecast satisfaction customers and accuracy achieve 88 %.
Keywords:
Satisfaction forecasting, Churn forecasting, Machine Learning, e-commerce.

1. Introduction

The dramatic increase and development of the internet have made fantastic success in online shopping [1]. Characterized by a high convenience and primary advantage, people prefer online shopping nowadays because of extensive comfort and timesaving [2]. Online shopping websites and mobile application platform provider for customers to purchase a lot of products online [3]. The most commonly known online shopping websites and mobile applications in the middle east where customers can also make purchases products online are Souq and Jumia [4]. Online shopping technology is increasingly increasing the customer base of its service providers with these developments in websites and smartphone apps. Customers look for better efficiency, value for money, higher quality of the product, in this competitive market, and furthermore customer acquisition costs have been increasing. The cost of retaining a current customer is reportedly cheaper than the risk of attracting new customers [9]. Thereby, from customer relationships to customer satisfaction, the focus of the business strategy changed.

The e-commerce websites and mobile application have hundreds of products which help the customers to purchase for competitive products. The aim of this analysis is to develop a model for machine learning that is more fulfilling. This paper attempts primarily mainly evaluate whether such an application of machine-learning to customer data can support practiced customer satisfaction, as compared to forecasts generated based on customer behaviour assumption [5]. This paper proposed employed a quantitative approach in this paper contains analysing customer data to develop machine learning models that can forecast the satisfaction of customers. This study framework for the execution of machine learning model. It consists of six phases: understanding the business environment, comprehension of data, pre-processing of data, model construction, evaluation.

Manuscript received August 5, 2023
Manuscript revised August 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.8.22

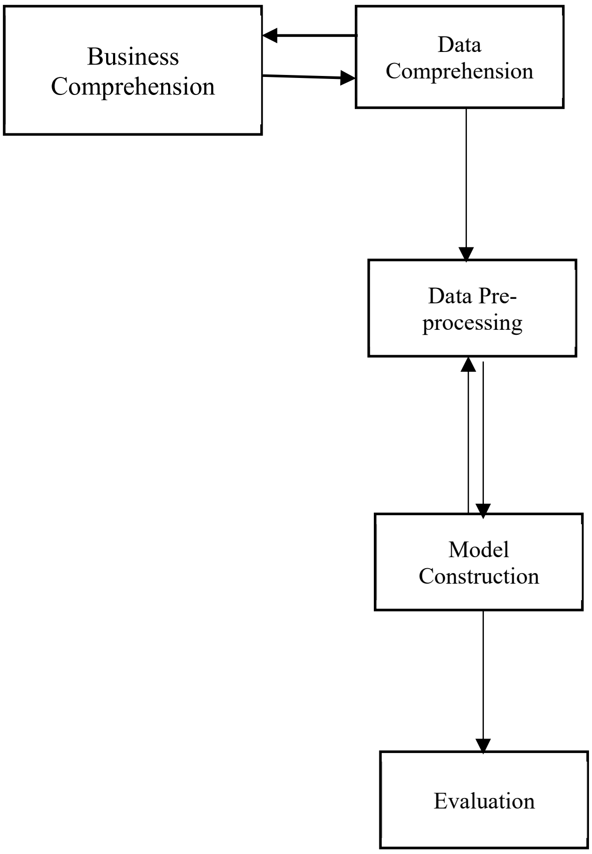

Figure.1 Framework for the execution of machine learning model

Customer behavior modelling and machine learning technology application can help to predict customer satisfaction and help this sector recognize the reasons why customers lose [6]. In addition, the correct detection of existing customers will help enhance the satisfaction of customers through optimizing rewards and deals [7]. throughout the effect saying, Data-driven marketing, can add value to the company compared to conventional marketing techniques by using predictive analyses of customer data [8]. This study aims to develop customer knowledge modelling to forecast satisfaction of customers on an e-commerce in Middle East, particularly Egypt.
This paper organized as in Section 2 related work was discussed, The detailed proposed model described in Section 3, The necessary result and discussion shown in section 4, The paper concludes with section 5.

2. Related Work:

Kolajo and Adeyemo have been focusing on data mining approaches in the telecommunications sector to combat churn behaviour. To do so, both descriptive and predictive data were used, and information about subscribers' calling behaviour was extracted to recognize the people that have a high chance of churns in the future. In this clustering phase, the clustering algorithms Simple K-Means and Expected Maximizing (EM) had been used, however, decision Stump algorithms, and ReTree Decision Tree in classification was applicable. This was used during the prediction process in which customers were defined, both during the clustering and the classification periods. The right algorithms were found. The study shows that the telecommunications service provider has yet to clearly differentiate the churner either from non-churner. Although their various option: either to provide motivation from all shoppers that are more costly than the effort to satisfy them, or to focus on the purchase program. Given that both alternatives have a negative effect on the company's finances, the best approach is to distinguish churner from non-churner [10].

To identify the probability of churn customers, machine learning models were commonly used [11]. Based on a churn prediction literature review, below present a brief introduction to the selected algorithms.

Logistic Regression Algorithm: primary focus of the Logistic Regression report is to analyse and forecast each the linkages between many features. Regression analysis uses multiple models to determine the interaction of the target/response predictor with different variables. The Regression analysis for the binary variable is a fitting analytical regression model. Regression analysis is a forecast analysis that explains how a binary issue is relevant to a set of separate variables. In order to evaluate the turnover rate, including a number of factors or attributes of the industry, Regression analysis was widely used for customer turnover [12].

Support Vector Machine (SVM): is learning how to analyse data for model determination. SVM reflects measurements in high dimensional space and seeks wants to clarify each optimally split hyperplane that separates between different classes provided with a list of training data labelled. New instances in a certain space have been viewed and categorized within a given character, depending on one's position relative to the space separation. SVM techniques have been extensively developed and studied for churn forecasts [13].

Random forest: It would be a cooperative learning technique that allows for classification and regression analysis. The framework designed for a single classification tree through the growth of several classification tree species is presented during the training phase. Each of the forests reacts to a specific instance, as well as the model seems to be the group mostly with the largest margin inside the entire forest. In contrast to traditional decision-making, the high-performance model also provides a major advantage to Random forest in the defence of overfitting [14].

Gradient Boosting Algorithm: It is focused mostly on ensemble methods that focus on either the Boosting. They are attempting to make a weak learner a better approach to learning. Each assumption would be that an ineffective classifier should be simpler than just another random estimation. Such algorithms that might accomplish at least somewhat smarter unlike random are, in fact, weak learners. The learning results are different in their inverse problem, producing weak learners. Adaboost separates observations by weighing up the poor or those that the weak learner does not properly predict. The main objective is to create a new weakness in effect making these attempts to make these incorrect observations. Due to their accuracy, The relatively high the accuracy, the stronger the final learner supports, the weaker learners will be connected to the best learner after their training. AdaBoost 's weak learners are divided decision-makers, and the combination of performance and accuracy of every weak learner is the focus of a single instance classification [15].

Conversely, gradient boosting is necessary for misclassified and dynamic situations that use most of the learner's errors. Errors are evaluated and the poor learner is adjusted for each replication. Rather, the weak contribution of the learner to strength is to a minimum the general error of the good learner [9]. AdaBoost [16] and gradient [17] are used for the churning model.

3. Proposed Model Method

The study's main objective is to check whether customer satisfaction is reliably predicted using machine learning as expected by the overall customer. The paper explains in greater depth how machine learning models are built utilizing techniques such as Gradient Boosting, Random Forest, SVM, and Logistic Regression are used for machine learning Algorithm. To evaluate the model's efficiency by using sampling approaches with a specific class difference ratio. Identify the optimal approach for consumers predicts specific. The retailer is an application-based e-commerce platform, with an increasing number of new customers on average annually created over the middle east. Machine learning is not used to measure the satisfaction of customers. Business resource able customers are categorized based on human experiences surrounding customer data [18].

One of the greatest concerns in many e-commerce applications is a customer. Studies have shown that keeping customer loyalty seems to be much lower in cost than attracting new customers [19]. Therefore, e-commerce platform focuses on designing valid and effective predictive models to recognize potential customers [20]. Satisfaction analysis covers several different steps as shown below:

- Market comprehension
- Selection, analysis, and processing of data
- Various classification algorithms to evaluate and select the best classifiers to forecast

3.1. Market Comprehension:

Throughout the market Comprehension process, the problem the need to tackle. The seller is an e-commerce platform for users from across the Middle East, with at least two thousand new customers on average. Machine learning in e-commerce applications is not actually used to estimate consumer satisfaction. The e-commerce sector is developing and is increasing in the Middle East. In every aspect of society, online shopping is continually more integrated, driving how we live and work, including trends in video streaming and social media advertising. This increase is accompanied by new competitors who are both opportunities and threats. E-commerce application would consider the actions of consumers who use Satisfaction programs and loyalty schemes. The purpose of this study is to construct machine-learning models using customer data from e-commerce applications that can predict customer satisfaction and loyalty with more than accuracy satisfaction.

3.2 Selection, analysis, and processing of data

Initially, data were obtained through customer reviews of the Souq e-commerce platform containing opinions of several online products, many negatives, and some positive. The preparation of data does not consist exclusively of transforming and cleaning of available data. It includes a good understanding of the features that have yet to be considered carefully and ensure that the data are appropriate first. Data preparation shortcuts will shorten your models. The preparation of analytical data begins with the exploration of data available. Analysts must then apply different data pre-processing techniques to clean up and reduce the data to manageable and relevant sizes. It consists of three phases: First Phase Tokenizing: It indicates that the customer review may be split into a variety of separate statements or split into a series of words. Second phase Cleaning: this requires eliminating special characteristics. Such as interrogative sign so this symbol is omitted. Third phase: Stop word Removal: all words that do not bring much meaning to interpretation are omitted in this method, such as is, is, was, he, she, etc.

Feature Extraction: This process minimizes the number of attributes into better subsets to improve accuracy, remove an overfitting problem, and minimize training time. To make it more accurate, remove the issue and time of training, this method reduces the number of attributes in suitable sub-sets [21]. Acc_score is a feature of “score” and “buy_attempt” variables. The metric of the buying performance of a customer can be considered better than the original 'Score' feature. A scoring feature is the number of shopping products in an e-commerce application, while the “Acc_score” the feature is an attribute that indicates the proportion of the customer's response to the product decided to purchase. The variable is extracted from the following:

Acc_$score$ = score/buy_ $attempt) * 100$ ____(1)

3.3. Classification Algorithms

The pre-processed data was used in this phase to develop machine learning models for the Satisfaction of customers. Since the data is available for training, models for machine learning are used in this research for modelling. The training standards for the forecasting of customer Satisfaction include random forest, logistic regression, and SVM algorithms. The best Satisfaction precision classifier is chosen for experiments along with the field basis of the recommendations of the classifiers. The pre-processed data was used in this phase to develop machine learning models for the Satisfaction of customers. Since the data is available for training, models for machine learning are used in this research for modelling. In this research, binary classification is used because the target variable is dichotomous. In the next series of experiment, sampled data are used to determine the output of models with a change in Imbalance. The e-commerce platform also applied the latest innovations to the optimized model to determine how it leads to Satisfaction accuracy. From the experiment, a model with the best Satisfaction accuracy is chosen and used for the evaluation.

3..1. Evaluate classifier:

The accuracy model is not treated as a test for determining the efficiency of the testing classifiers. Although the data set is highly imbalanced and most data are not retained customers, the model predictions are usually orientated to the non-retained class. As a result, even with low Satisfaction classification performance, the model can still be good overall correct. Thus, the generalized accuracy of the satisfaction class is used in this paper rather than the total sample accuracy to test the model results. These measurements can be calculated from the matrix of uncertainty. Matrix of Confusion: such That True Positive (TP): while both forecasted and actual values are true TP, True Negative (TN): while both forecasted and actual values are false TN, False Positive (FP): whenever the actual value is false and was forecasted to be true, as well as False Negative (FN): whenever the actual value is true but is forecasted to be false. The accuracy seems to be the calculation including its proper ratio including its retained customers listed below:

Satisfaction accuracy = TP/TP+FP_______(2)

Table1: Matrix of Confusion

	Actual Positive	Actual Negative
Forecasted Positive	TP	FP
Forecasted Negative	FN	TN

While its satisfaction accuracy is the metrics used for evaluating the study, such criteria as Satisfaction

Recall, Churn Precision, and Churn Recall have also been tested in this research to examine exactly how reliable the forecasts are and how accurately the majority groups are defined. Satisfaction Recall examines how many of the customers retained were accurately identified. This comes from

Satisfaction Recall = TP/ TP+FN

Although the priority is on the satisfaction class, the performance of churn class or majority class model learning models is also analyzed in order to fully understand how well the model also foresees the other classes. The Churn accuracy and Churn Recall of the model is:

Churn accuracy = TN/TN+FN __________(3)

Churn Recall = TN/TN+FP_____________(4)

4. Implementation:

Data set used in study is the customer data of an e-commerce application platform are the dataset used in this study. There are 8000 records, each with 10 features. Data from September 2019 to December 2019 are required for customer's activity. The target variable is transformed which indicates that a customer is retained or churned. The following table will describe all variables in the data. The data is common for this variable when customers who did not purchase products are excluded.

Table2: Description of variables

Attributes	Datatype	Description
Cust_at	a numerical variable	total number of vistas
buy_at	a numerical variable	number of product customer buy
seen_at	a numerical variable	number of products seen by customers
Score	a numerical variable	count number of purchases
A_day	a numerical variable	number of days
converted	target variable	represents satisfaction and churn

In following Table 3. show the statistical metrics such as count, mean, and standard deviation to show a better overview of data. The table 3 above shows that no data is incomplete. Most data features have differed greatly from the mean. The data before modelling needs to be standardized.

Data Pre-processing, data are analysed to suit pre-processing modelling given the results including its data comprehension phase. Data Pre-processing techniques include encoding, noise removal, standardization, and data splitting.

Stage	No records	Imbalance
Actual	8000	96.88
Noise Removal	6675	96.52

Table 4: Dataset After Noise Removal

Encoding means the transformation of one data form into another. Encoding is used to convert Boolean or Categorical variables designed to facilitate of the classifier. Standardization is a re-scaling process for data with a mean value of 0 and a standard deviation of 1. The performance of classifiers would be negatively affected by highly deviating data from the centre as the standard classifiers assume the data has a gaussian distribution. With the experience from the process of data comprehension, it was found that the data vary significantly from the mean. Therefore, before using machine learning algorithms, the data must be standardized. For each instance, the following formulation is used to generate a standardized score called xscore:

Table 3: Customer data description statistics

	Cust _at	buy_ at	seen_ at	Sco re	A_da ys	converted
Count	792 3	7923	7923	792 3	7923	7923
Mean	100. 23	5.61	1.87	33. 06	1.523	0.0125
SD	378	24.6 3	3.14	223 .3	3.03	0.11

Xscore = S-Nmean/sd_________________(5)

Where, S refer to Sample of data, Nmean refer to Sample mean, and Sd = Standard Deviation of the Sample.

Noise Removal, Noise refers to any unwanted data signal, that would have a significant impact on the performance of something like the classifiers. Data was analyzed from a commercial viewpoint in this segment to define the noise. In analyzing the "cust at" feature, 1325 customers did not try any kind of shopping (cust at = 0) could be deducted. In the commercial, the consumers have the register in an e-commerce platform but haven't used it. These customers are considered to be idle customers and do not contain data to assist with forecasts. These customers are therefore deleted from the initial dataset It will also contribute to enhancing the signal in the data but also reduces the imbalance in the data because most idle customers fall into the churn class.

Data Splitting, for the measurement of model performance, the research uses a 10-fold stratified k fold cross-validation method. Throughout the iteration, 9 folds of data are used and 1-fold is used for the training in each testing. For splitting the data into folds, a stratified approach is used to ensure that each fold has representative satisfaction and churn class. There are 6675 records for any plugin this cross-validation, with an Imbalance of 96.52.

Modelling of data, the pre-processed data is then used to build machine models to forecast customer satisfaction after the data understanding and data pre-processing stage. Many machine learning models are built with various algorithms, data-set sample methods, and other parameters in this step. A variety of studies in this process was carried out to build a reliable model that can accurately forecast the success of customers of the highest accuracy. The efficiency of machine learning is measured of satisfaction accuracy, the model that forecasts retainable customers from the overall customer market. Total classification accuracy is not used as a measure of the classification algorithm, as the forecast is biased in regard to the majority class given the high imbalance of the data. Throughout this scenario, an accuracy of Churn might be very high, leading to very high accuracy. For this purpose, Accuracy is the indicator for measuring the model efficiency of the class of minorities, i.e. the satisfaction class.

Classifier Selection, the efficiency of various classifiers like Random Forest, Logistic Regression, SVM, and Gradient Boosting Algorithm has been examined on the data set to determine a classifier algorithm for the study.

Random forest: It is a set of decision tree algorithms that perform for regression and classification problems.

Logistic Regression: The data in this study were represented by using a logistic regression classifier. The model is assessed using the 10-fold Stratified sampling method and parameters have been used by default.

SVM: Often used to develop machine learning models. The models have been developed and reliability is measured by different SVM kernels. For comparison with other algorithms, the kernel has the highest level of accuracy and retrieval.

Gradient Boosting Algorithm: An established decision trees coupled with a gradient that boosted the advantage. This model applies gradient boosting to model and evaluates customer data.

Table 5: Design for classifier Selection

	Data Size	Imbalance
Random forest	6000	75:25
Logistic Regression	6000	75:25
SVM	6000	75:25
Gradient Boosting	6000	75:25

5. Results:

The results obtained referred to in the implementation. Satisfaction accuracy is the measurement shown in the proposed model in paper to assess the effectiveness of the model. This is the total accuracy of the model due to issues of class variance is not taken into consideration for model performance evaluation. In this paper was performed to select a classifier algorithm for study purposes.

As shown in figure 2, It's the total of all website visits or customer requests. It ranges between 0 and 25788 with a standard deviation of 389.62. The mean value for the variable Cust_at is 101.78. This variable might be seen as extremely skewed to left. As shown in Figure 3, as well, Total customers view e-commerce platform with a mean value of 7.8 with a standard deviation of 30.64. These values range between 0 and 1873. As seen below, the data were extremely left-skewed.

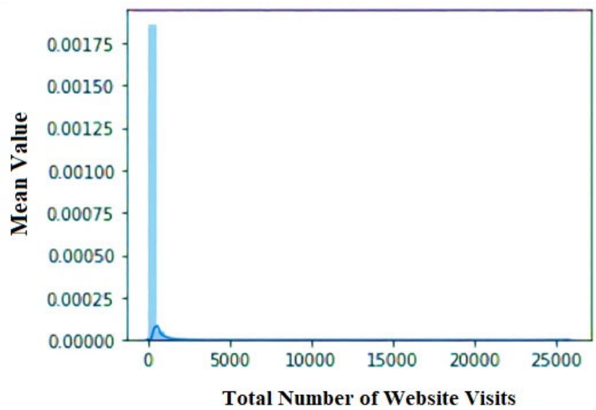

Figure. 2 Rate distribution of the total customer view

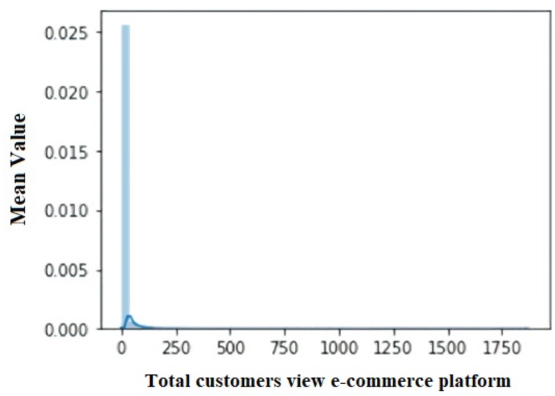

Figure. 3 Rate distribution of product purchase number

As shown in figures 2 and 3 it is worth noticing, with better accuracy and recall, that the gradient boosting algorithm exceeded all other learning models. For all classifiers, Churn Recall and Churn Precision are nearly constant. Either classifiers such as logistic and SVM are significantly different from the gradient booster classifier by 7% of satisfaction accuracy and 10 % of satisfaction recall. Furthermore, gradient boosting is selected as the best performing classifier algorithm in the study.

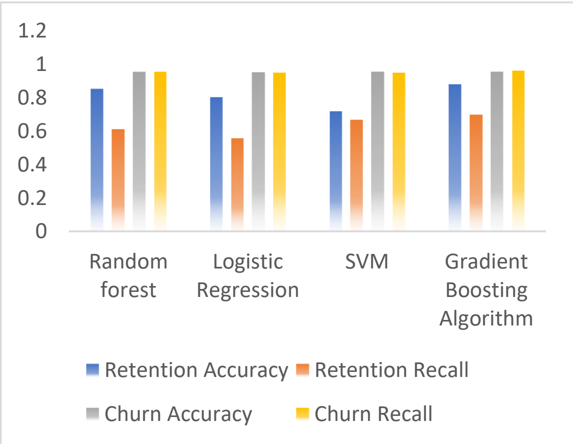

Figure4: Comparison of Classifiers

Table 6: Result for classifier

	Data Size	Satisfaction Accuracy	Satisfaction Recall	Churn Accuracy	Churn Recall
Random forest	6600	0.852	0.612	0.953	0.956
Logistic Regression	6600	0.802	0.557	.952	0.951
SVM	6600	0.721	0.669	.954	0.950
Gradient Boosting	6600	0.881	0.698	0.955	0.961

As seen in Figure 4, To recognize the best performing classifier, to assess the efficiency of the various classification algorithms with 6000 record data set list, representing 25% of customers proportion, and 75% of customers not retained. The results indicated that the Gradient Boosting Algorithm is highest accuracy, and recall algorithm, relative to other algorithms like the Logistic Regression, Random Forest, and SVM.

A customer data analysis was performed to determine the reasons for the satisfaction or churn. Some deductions on the behaviour of the retained/churned customers are subtracted from the data understanding phase. Figure 5,6,7,8 gives us a correlation of the retained customers and churned customers' patterns of activity. For the above graphs, a median measure is used to analyse the results. Particularly in comparison to churned customers, it could be noted that the retained customers were actively using the e-commerce websites or application. For the retained customer the median active days are 18 while for a churned customer it is 1. It indicates that, in the e-commerce websites or application, a customer with less probability of retaining will not be implicated.

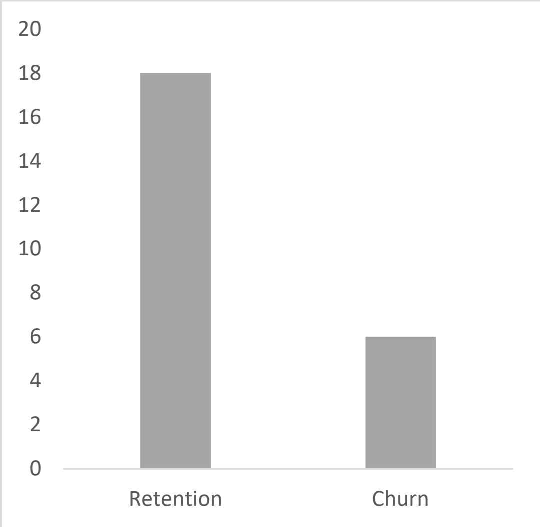

Figure. 5 Median active days

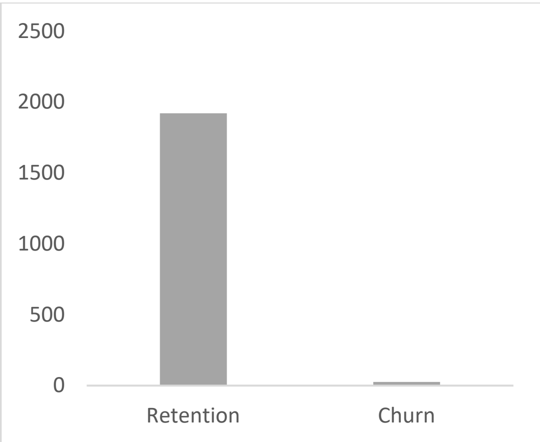

Figure. 6 Median of customer Vistas

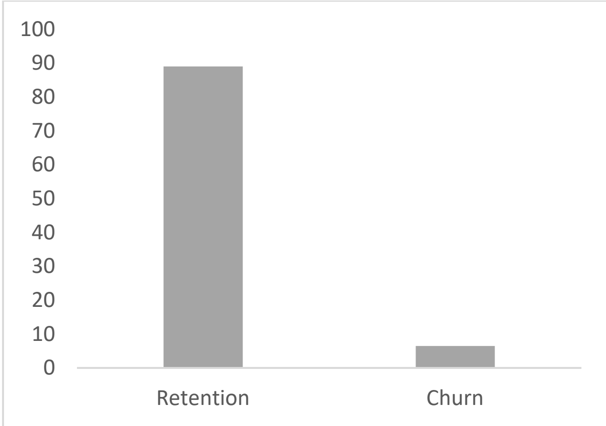

Figure. 7 Median of the Product seen with the customers

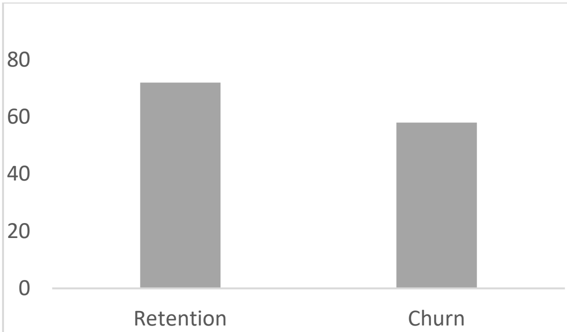

Figure 8 Median of customers' responses to the buying product

6. Conclusion:

This study aimed to provide a model for the satisfaction classification. The study is conducted to evaluate how machine learning models can work in forecasting customer satisfaction on e-commerce Websites and applications. This paper utilizes the machine learning model for forecasting customer satisfaction in e-commerce customer data. The machine-learning model has been utilized with data gathered by the e-commerce platform. Implemented through machine learning models, the proposed concept was evaluated by satisfaction accuracy measurement. The data was processed to avoid data inaccuracy, outliers, etc. during the preparation stage. The category variables have been encoded and the results are standardized in numerical values. In the data pre-processed stage for the modelling phase, to test the imbalanced data in a balanced data set, sampling methods were used, and their performance is evaluated. In the Evaluation stage of proposed model efficiency on target variable customer data set in customer satisfaction forecasting of machine learning models. To deal with this, customer data on e-commerce websites and applications for the satisfaction of customers were obtained and machine learning algorithms. The results achieved as maximum accuracy compared to other algorithms, including Gradient Boosting Algorithm, Support Vector Machine Algorithm, Random Forest Algorithm, and logistic regression Algorithm. The best model developed for this paper to forecast satisfaction customers and accuracy achieved up to 88 %.

References

[1] M. Vidhya and S. Aji, (2019) “A Mean Model Based Incremental Learning Technique for Extreme Learning Machine,” Procedia Computer Science, vol. 165, pp. 541–547.

[2] K. Lerman, S. N. Minton, and C. A. Knoblock, (2012) “Machine Learning Techniques for Wrapper Maintenance,” Virtual Enterprise Integration, pp. 334–350.

[3] C.-H. Yang, (2007) “Collaborative Mechanism of Manufacturing Enterprise Supply Chain Based on Multi-Agent,” 2007 International Conference on Machine Learning and Cybernetics, Aug. 2007.

[4] C. Ugwu, N. L. Onyejegbu, and I. C. Obagbuwa, (2012) “The Application of Machine Learning Technique for Malaria Diagnosis,” Machine Learning, pp. 2035–2043.

[5] Sharma, A., & Panigrahi, P. K. (2011). A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. International Journal of Computer Applications, 27(11), 26-31.

[6] Kolajo, T & A. B. Adeyemo (2012). Data Mining technique for predicting telecommunications industry customer churn using both descriptive and predictive algorithms. Computing Information Systems & Development Informatics Journal. 3(2). 27-34.

[7] Rahul J. Jadhav and Usharani T. Pawar. “Churn prediction in telecommunication using data mining technology” , (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.2, 2011.

[8] Andrew h. Karp, using logistic regression to predict customer retention,1998.

[9] De Caigny, A., Coussement, K., De Bock, K. W., & Lessmann, S. (2019). Incorporating textual information in customer churn prediction models based on a convolutional neural network. International Journal of Forecasting. doi:10.1016/j.ijforecast.2019.03.029

[10] Guangli Nieae, Wei Rowec Ling ling, Zhangab Yingjie, Tiana YongShi, “Credit card churn forecasting by logistic regression and decision tree”, Expert Systems with Applications Volume 38, Issue 12, November– December 2011

[11] Ali Dehghan, Theodore B. Trafalis, “Examining Churn and Loyalty Using Support Vector Machine”, Business and Management Research,Vol. 1, No. 4; 2012

[12] Ionut B. Brandusoiu Gavril Toderean, Predicting Churn In Mobile Telecommunications Industry", Acta Technica Napocensis Electronics and Telecommunications, Volume 54, Number 3, 2013

[13] Chen, Y., & Nevskaya, Y. (2019). “The Company We Keep:” Endogenous Network Formation and Peer Effects in Customer Churn. SSRN Electronic Journal. doi:10.2139/ssrn.3502849

[14] Schapire R.E. Explaining AdaBoost. In: Schölkopf B., Luo Z., Vovk V. (eds) Empirical Inference. Springer, Berlin, Heidelberg, 2013.

[15] Jerome Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", The Annals of Statistics, 29(5), 2001, 1189-1232. 13

[16] Adnan Idris, Asifullah Khan, yeon soo Lee, “Genetic programming and adaboosting based churn prediction for telecom. In the proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1328–1332, 2012.

[17] Manivannan, R., Saminathan, R., & Saravanan, S. (2019). An improved analytical approach for customer churn prediction using Grey Wolf Optimization approach based on stochastic customer profiling over a retail shopping analysis: CUPGO. Evolutionary Intelligence. doi:10.1007/s12065-019-00282-x

[18] Wu ; Sufang Meng, “E-commerce customer churn prediction based on improved SMOTE and AdaBoost”, In the proceedings of the 13th international conference on Service Systems and Service Management (ICSSSM), 2016.

[19] Lemmens, Aurelie and Gupta, Sunil, “Managing Churn to Maximize Profits” (May 8, 2017). http://dx.doi.org/10.2139/ssrn.2964906

[20] Gerald Fahner , Stochastic Gradient Boosting Approach to Daily Attrition Scoring Based on High-dimensional RFM Features”2015 Fair Isaac Corporation.

[21] C. Troussas, A. Krouska and M. Virvou, "Trends on Sentiment Analysis over Social Networks: Pre-processing Ramifications, Stand-Alone Classifiers and Ensemble Averaging", Machine Learning Paradigms, pp. 161-186, 2018