International Journal of Research in Engineering, Science and Management
Volume-3, Issue-6, June-2020
www.ijresm.com | ISSN (Online): 2581-5792

Predicting the Propensity of Order Cancellation in the Ecommerce Domain

I. Chidroop[1*], Minal Moharir[2]
[1]*Student, Department of Computer Science and Engineering, Rashtreeya Vidyalaya College of Engineering, Bangalore, India*
[2]*Associate Professor, Department of Computer Science and Engineering, Rashtreeya Vidyalaya College of Engineering, Bangalore, India*
*Corresponding author: chidroop.i@gmail.com

***Abstract*: Ecommerce has been a revolution in the retail domain. Shoppers from around the world can shop in a click. This ease of access creates a great change to the business, but it also increases the risk of exposure to bad actors with the same ease. Understanding the behavioral patterns that customers exhibit is key to providing an edge amongst the competition and also security to business goals. This study tries to predict the propensity to cancel an order. This can influence the order journey in triggering payment options or requiring additional authentication or even help manage inventory blocking. This study aims to understand the mechanics of human behavior and parameters that affect orders. The problem is tackled with two sets of data, the current order data and the history of the customer placing the order. As this is a two class classification problem this study tries to apply the two class boosted decision tree, two class decision forests, Neural networks and Locally Deep Support vector machine algorithms to datasets with and without customer history data. The study subjects the models to a dataset without customer history as this would be the scenario for new customers (Cold Start). This study compares the different models with multiple metrics such as accuracy recall precision and f1 score. Boosted decision trees with the customer history data fared the best with metrics of 93.1%,91.6%,92.6% and 92.1%, Accuracy precision recall and F1 score respectively.**

***Keywords*: Classification in Ecommerce Cancellation, Neural Networks, Boosted Decision tree, Support Vector Machine, Decision Forests.**

1. Introduction

In recent years ecommerce has become a pivotal concept in the retail domain. Not to say brick and mortar stores have lost their purpose but they have their pros and cons. They can offer an in-person experience, buyers love trying the product out before they pull the trigger but there have been two vital changes in the consumer space. First, the widespread spawn of boutique and small-scale productions exponentially multiplying the options in a particular product category. Finite retail shelf space often cannot cater to offer all options. Second, the novelty of trying the product does not extend to repeatedly purchased items. On the other side of the field, ecommerce platforms have a wider audience, reach, real-time inventories and a massive catalogue to choose from. All this just at your fingertips. Major corporations have started moving to multichannel experience to make a whole some customer journey.

With new systems and innovation comes new problems. Access to the ever so powerful internet is a double-edged sword. Yes, more customers can be reached than before, but the bad actors can reach the business just as easily. So, it is very essential to understand the demographic the platform serves. There is an eminent need to mine customer data to understand their preferences, habits and shopping patterns. Recommendation of items also forms a key challenge in these platforms. [1] It is not only a security concern but is also a key business advantage to make customer journeys wholesome. To keep a major ecommerce platform afloat customer understanding becomes primal.

When the shopping paradigm was completely changed it was bound to cause a new niche of issues. Issues like fake orders, fake cancellations, Inventory hijacking, price gouging have become commonplace. This paper tackles the problem of fake orders and cancellations. To identify if an order is fake or is it is going to be cancelled are two different problems sharing some common ground, but from a business perspective it is a loss of dollars. So diving in to take a look at how Machine learning and classification algorithms can serve this purpose.

2. Related Work

There has been a significant amount of research in the domain of customer segmentation. Customer segmentation is the process of grouping like customers together by finding patterns or common points. Common strategies to achieve this in the past was by using algorithms like K-means clustering [2]. K-means clustering provided a starting point for unsupervised scenario. The main draw back with this technique is that the number of clusters is pre-determined. It becomes a user decision to set the number of clusters. In a real-time scenario data flows in continually and clustering the entire set repeatedly is highly inefficient. Stream clustering is the approach that takes a input stream as a data source and there are global and local clusters created [3].

3. Methodology

The aim is to create a model that can predict the propensity of cancelling an order placed in the ecommerce realm. The challenging aspect of this problem is the need to make a prediction only based on the information it has at the time of purchase. This data can be broadly classified into two categories: order related data and customer related data.

A. Order related data

Order data is all the metrics revolving around the particular purchase being made these are (non exhaustively):

1. List of items purchased
2. Current value of purchase
3. payment method chosen
4. Platform from which order was placed (Mobile / Web etc.)
5. Time of Order
6. Delivery address

7 Coupons/Promotion Applied

Order related data tends to throw light on current patterns or a local finding. There are many latent relationships that can be derived from order data. The idea is to pick up patterns from the type of products being ordered. Order data provides a local understanding of the input package.

B. Customer history data:

Customer history data is all the metrics revolving around the customer who placed the order. This study looks at these parameters for the lifetime of the customer:

1. Number of Orders placed
2. Number of Orders Accepted/payed for
3. Number of Cancelled orders
4. Number of Returned orders
5. Average number of days before Order was returned
6. Average Value of the order
7. Average Frequency of Orders placed

Having a global understanding of the input package is also key to having a high accuracy. It provides a global perspective to the model. Customer data tends to throw light on how this particular customer has interacted with the platform and might indicate strongly how she/he will interact in the future as well.

These two parts form the input package for the model. The aim is to explore the latent relationships between the Customer and order history as well. In the current data set there is a clear balance between the classes but if there is a stark delta between classes that is if one class is <5% and the other >95% then there is a need to use some technique to normalize this difference, else the model may fit to side with the bigger class. This study recommends a technique like SMOTE (Synthetic Minority Oversampling Technique) to balance the dataset.

This paper aims to compare the performance of the following algorithms:

1. Two Class Boosted Decision tree
2. Two Class Decision Forests
3. Two Class Neural Networks
4. Two Class Locally Deep Support Vector Machine.

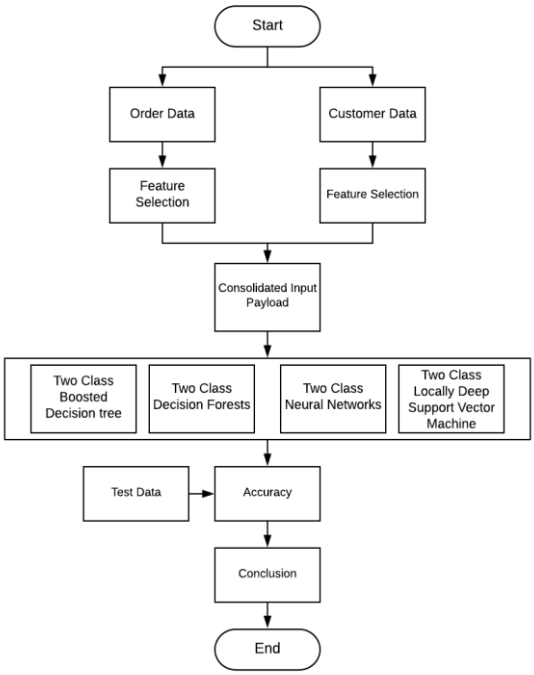

Fig. 1. Research Steps

C. Decision trees

A decision tree is a structure of nodes where internal nodes represent a comparison or test given an attribute, the branches of the internal nodes will be the outcomes. The leaf nodes represent the class of outputs. Given an input tuple the respective attributes at each node is compared and advanced down till you reach a terminal node of the tree which becomes the output for a given input. Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is a statistical method which uses a supervised learning to create a predictive model. [4]

D. Boosted Decision Trees

Simple decision trees are notoriously known to over fit the model. To reduce this risk the study uses techniques to combine multiple trees. Boosting is the process of combining these trees serially to avoid overfitting. Weak learners are combined together to get a better model.

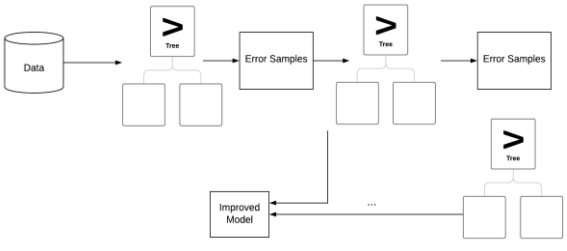

Fig. 2. Boosted Decision Trees generation

Let each function denote a tree and y is the output of the tree. Then using the gradient descent algorithm we can derive the method to boost trees. log-loss function is used to calculate the loss as it is a classification model. An ensemble of weak learners make up the fit model.

- $f_1(x) \approx y$
- The residual is $y - f_1(x)$
- $f_2(x) \approx y - f_1(x)$
- The residual is $y - f_1(x) - f_2(x)$
- $f_3(x) \approx y - f_1(x) - f_2(x)$

Fig. 3. Boosted decision tree algorithm

E. Two class decision forests

Simple decision trees easily over fit so decision forest is another ensemble learning technique in classification problems. Ensemble models provide better coverage and accuracy than single decision trees. The algorithm is implemented by construction of multiple decision trees and then a majority vote for the output class decides the final output. Voting is one of the better-known methods for generating results in an ensemble model.

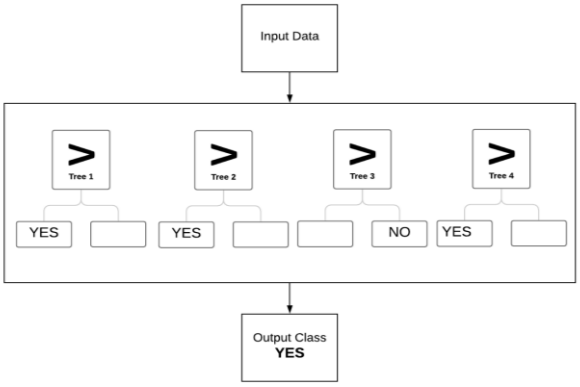

Fig. 4. Classification in a Decision Forest by Voting

F. Two class neural networks

In an attempt to replicate the processing of a human brain. Basic element of the Neural Network is modeled after the human neuron. The Neuron takes a finite number of inputs and with a mathematical function that outputs one output and the network is built up of many such elements. These elements are arranged in layers namely input, output and hidden layers. Once the output layer produces a result it is compared with the actuality and the error is back propagated to correct the model. When repeatedly performed with an appropriate network structure and sufficient time great results can be obtained.

The number of nodes in the input layer will be equal to the number of features. There is one hidden layer with 100 nodes and 2 nodes in the output layer one for each class.

Table 1
Confusion Matrix for Boosted Decision Tree (BDT)

True Positive	False Negative
28182	8866
False Positive	True Negative
7682	41602

Table 2
Performance Metrics for Boosted Decision Tree

True Positive	False Negative
28182	8866
False Positive	True Negative
7682	41602

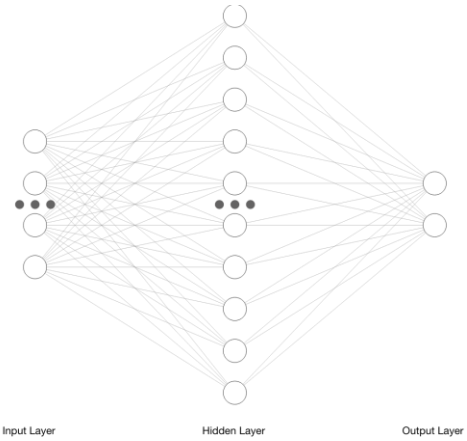

Fig. 5. Architecture of Neural Network

G. Two class locally deep support vector machine

Support vector machines have recently gained popularity in classification problems. LD-SVM is a nonlinear classifier that performs very well in binary classification tasks. It can be used when we can't use linear classifiers or when nonlinear classifiers give bad results. LD-SVMs primarily train much quicker than other algorithms, sometimes at the cost of accuracy. LD-SVM models are a good alternative when the data is complex enough to perform poorly on linear models (such as logistic regression). Also, LD-SVM models are small enough to be used in mobile devices or other situations where complex models (such as neural networks) are too large for efficient use. Conversely, this model should not be used if you don't care about the size of the sample, or if you need a linear sample for simplicity or speed prediction. There is also no point in changing to LD-SVM if there are already good results provided by linear classifiers, or if you can get high classification accuracy by adding small amounts of non-linearity. [5]

4. Results

A. Prelude

This study consists of two data segments i.e. the order related data and customer history. It considers 88612 order entities and the history of the customers related to these entities. On analysis the data set is well balanced between the prediction classes so we do not need any transformation technique like SMOTE (Synthetic Minority Oversampling Technique). In the customer journey paradigm, this model assumes that we have customer

history but the problem of cold start is to be considered so we compare all models with and without customer history to make the research wholesome. This study compares the performance metrics across models and the corresponding data sets. This study presents the following results for each model and two corresponding data sets for the model:

1. Confusion Matrix
2. Performance Metrics
3. ROC Curve for the model.

B. Performance Metrics

In the classification paradigm just comparing the accuracy gives us a incomplete picture of the model performance. We use the following metrics to understand the characteristics of the model better.

1) Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Fig. 6. Calculation of accuracy in classification models

Figure 6 shows the calculation of accuracy in classification models where TP = True Positive, TN = True Negatives, FP = False Positives and FN = False Negatives.

2) Precision and Recall

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

Fig. 7. Calculation of Precision and Recall in classification models

3) F1 Score

The F1 Score is used as a metric when we want a balance between precision and recall. It is useful when there is an uneven class distribution. Figure 8 shows the calculation of F1 score in the classification models.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Fig. 8. Calculation of F1 Score in classification models

C. Two Class Boosted Decision Tree

The data set contains a class ratio that is well divided so we can directly proceed with the model. Table 1 shows the confusion matrix for cases of cold start without customer history data.

When we subject the model to customer data as well Table 3 shows the confusion matrix of the model. Table 4 shows the performance metrics of the same. We can clearly see a huge performance boost in all aspects of the model. This ensemble approach has absorbed the new data fairly well.

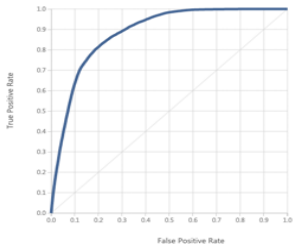

Fig. 9. ROC Curve for Two Class Boosted decision trees (BDT)

Table 3
Confusion Matrix for Boosted Decision Tree with Customer Data (BDT+CD)

True Positive	False Negative
35293	2826
False Positive	True Negative
3254	47239

Table 4
Performance Metrics for Boosted Decision Tree with Customer Data (BDT+CD)

Accuracy	Precision
0.931	0.916
Recall	F1 Score
0.926	0.921

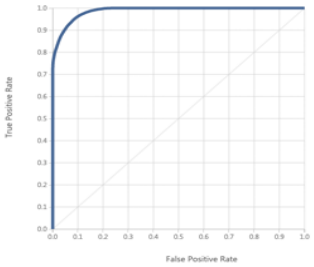

Fig. 10. ROC Curve for two Class Boosted decision trees with customer data (BDT+CD)

D. Two Class Neural Network

Table 5 depicts the confusion matrix results and Table 6 are the performance metrics. The training time taken for a neural network is significantly more.

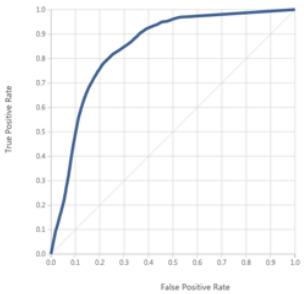

Fig. 11. ROC Curve for two Class Neural Networks trees (NN)

When we subject the model to customer data as well with Neural Networks.

Table 5
Confusion Matrix for Neural Network (NN)

True Positive	False Negative
31389	5659
False Positive	True Negative
14717	34567

Table 6
Performance Metrics for Neural Network

Accuracy	Precision
0.764	0.681
Recall	F1 Score
0.847	0.755

Table 7
Confusion Matrix for Neural Network with Customer Data (NN+CD)

True Positive	False Negative
30075	8044
False Positive	True Negative
8939	41554

Table 8
Performance Metrics for Neural Network with Customer Data (NN+CD)

Accuracy	Precision
0.808	0.771
Recall	F1 Score
0.789	0.78

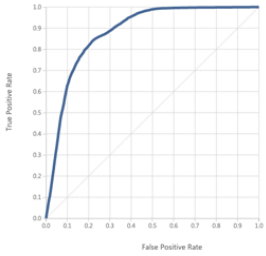

Fig. 12. ROC Curve for Two Class Neural Networks with customer data (NN +CD)

E. *Two Class Decision Forest*

Table 9
Confusion Matrix for Decision Forests (DF)

True Positive	False Negative
24818	12230
False Positive	True Negative
7029	42255

Table 10
Performance Metrics for Decision Forests (DF)

0.777	0.779
Recall	F1 Score
0.67	0.72
0.777	0.779

The Second ensemble method using decision trees with a voting technique.

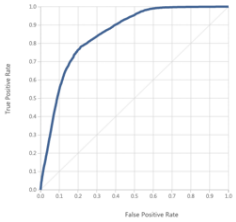

Fig. 13. ROC Curve for two Class Decision Forests (DF)

When we subject the model to customer data as well with Decision Forest.

Table 11
Confusion Matrix for Decision Forests with Customer Data (DF+CD)

True Positive	False Negative
3215	34904
False Positive	True Negative
149	50344

Table 12
Performance Metrics for Decision Forests with Customer Data (DF+CD)

Accuracy	Precision
0.604	0.956
Recall	F1 Score
0.084	0.155

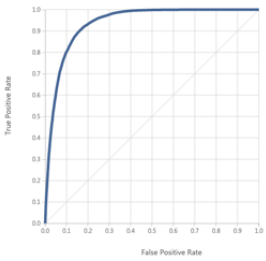

Fig. 14. ROC Curve for Two Class Decision Forests with customer data (DF +CD)

F. *Two class locally deep support vector machine*

LD-SVM offers a great nonlinear model with very fast training times.

Table 13
Confusion Matrix for Locally Deep Support Vector Machine (LD-SVM)

True Positive	False Negative
29797	7251
False Positive	True Negative
9750	39534

Table 14
Performance Metrics for Locally Deep Support Vector Machine (LD-SVM)

Accuracy	Precision
0.803	0.753
Recall	F1 Score
0.804	0.778

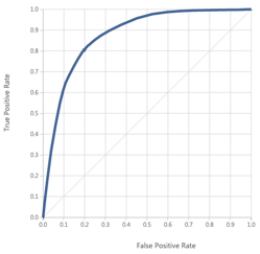

Fig. 15. ROC Curve for two Class Locally Deep Support Vector Machine (LD-SVM)

When we subject the model to customer data as well with Decision Forest.

Table 15
Confusion Matrix for Locally Deep Support Vector Machine with customer data (LD-SVM+CD)

True Positive	False Negative
31980	6139
False Positive	True Negative
10395	40098

Table 16
Performance Metrics for Locally Deep Support Vector Machine with customer data (LD-SVM+CD)

Accuracy	Precision
0.813	0.755
Recall	F1 Score
0.839	0.795

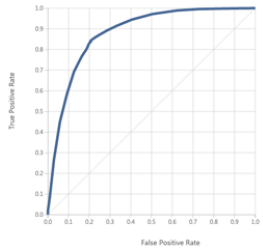

Fig. 16. ROC Curve for two Class Locally Deep support vector machine with customer data (LD-SVM +CD)

G. *Comparative analysis*

The following tables compare all 8 scenarios metric wise. It depicts how each model fairs with each other and with and without customer data. The model without customer data are represented by a "blue bar" and "yellow bars" represent datasets with customer data. As predicted the figures 17 through 20 show that Adding customer data significantly improved the model in most cases. The largest impact was seen by the Boosted decision tree technique with a 13 % boost approximately. Decision forest sees a reduction in accuracy but precision sees a healthy boost so we can predict that the model fits more towards the positive class that is "Delivered". So DF can be helpful if we are concerned about the false positives as much. BDT sees a healthy boost in all metrics and provides a more balanced model. So taking a deep dive into the model, looking at Table 17 it shows how the model scored across the data sets and presents a distribution on the classification. The table is structured to divide the classification into 10% bin based on the Scored probability of the class. In this study the proximity to 1 is identified as "Delivered" and 0 as "Classified" the two prediction classes. The models Threshold is 0.5 and predictions above 0.5 are classified as the positive class that is Delivered and vice versa. From Table 17 we understand that the concentrations of positive examples at 1 and Negative at 0 indicates that most of the times the model is confident with its prediction and there are clear differences in the classes that the model has picked up. It also lists the Performance metrics at a score bin level.

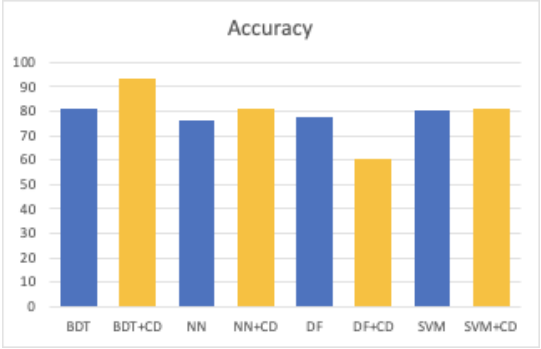

Fig. 17. Performance Metric Comparison: Accuracy

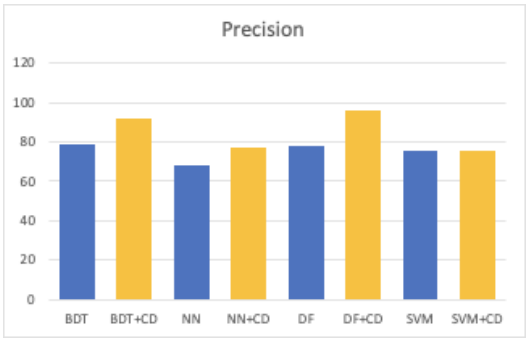

Fig. 18. Performance Metric Comparison: Precision

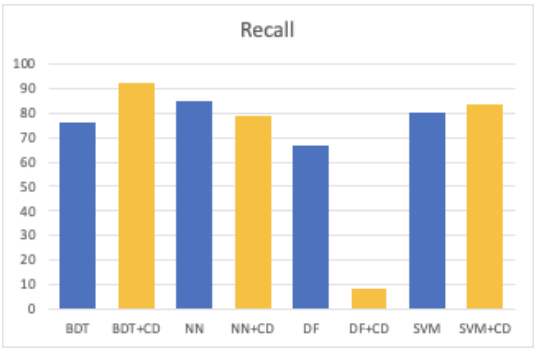

Fig. 19. Performance Metric Comparison: Recall

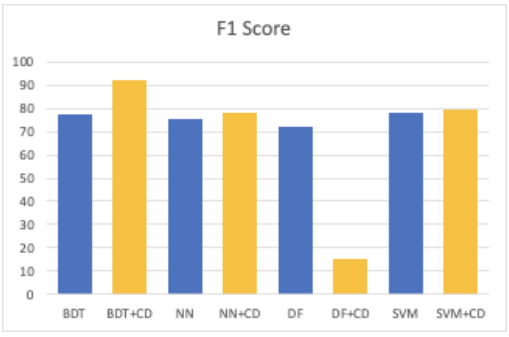

Fig. 20. Performance Metric Comparison: F1 Score

5. Conclusion

As the dataset was subject to four models and in two paradigms i.e. is with and without customer history data. The model with the most promising result was achieved with the boosted decision tree subjected to both order and customer data.

With an accuracy of 93.1% and a precision and recall of 91.6% and 92.6% respectively. LD_SVM also provided a great model in the case of no customer data. LD_SVM's also have extremely fast training times. With this instance of the data, decision trees provide a very viable solution to a fast and accurate model to solve this ecommerce conundrum.

The models also consistently showed that the customer history data helped improve the model performance metrics. This indicates that customer's behavioral patterns tend to repeat and they can have a significant impact on the decisions they take on the platform. The e commerce systems will need to start understanding these patterns. AI and ML play a key role in impacting business decisions and the way we tackle these problems as well.

6. Future Work

The current models can be improved to add more features to the dataset to provide a more insightful take on the demographic. The hyper parameters to each of the techniques can be modified to derive better accuracy and metrics. The neural network can benefit from additional layers and some more training time.

The current school of thought can be extended to other problems in the ecommerce domain as well like predicting inventory shortages, identifying fraud orders or fraud sellers etc. This could generate great impact to the customer and business and can be the edge or other competing platforms.

References

[1] L. Narke, and A. Nasreen, "A Comprehensive Review of Approaches and Challenges of a Recommendation System", International Journal of Research in Engineering, Science and Management, vol. 3, no. 4, pp. 381-384, April 2020.

[2] Chinedu Pascal Ezenkwu; Simeon Ozuomba; Constance Kalu, Application of K-Means Algorithm for Efficient Customer Segmentation; A Strategy for Targeted Customer Services, International Journal of Advanced Research in Artificial Intelligence, 2015.

[3] Tushar Kansal; Suraj Bahuguna; Vishal Singh; Tanupriya Choudhury, Customer Segmentation using K-means Clustering, CTEMS, 2018.

[4] Himani Sharma and Sunil Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining", International Journal of Science and Research, 2016

[5] Jose, C., Goyal, P., Aggrwal, P. & Varma, M., "Local Deep Kernel Learning for Efficient Non-linear SVM Prediction", Proceedings of the 30th International Conference on Machine Learning, in PMLR, vol. 28, no. 3, pp. 486-494, 2013.

[6] Beatriz Nery, Rodrigues Chagas, Julio Augusto and Nogueira Viana, "Current Applications of Machine Learning Techniques in CRM: A Literature Review and Practical Implications", International Conference on Web Intelligence (WI), 2018.

[7] Emad Elaziz Dawood, Essamedean, Improve Profiling Bank Customer's Behavior Using Machine Learning, IEEE Access, Volume, 2019.

[8] B. N. Krishna Sai and T. Sasikala, Predictive Analysis and Modeling of Customer Churn in Telecom using Machine Learning Technique, ICOEI, 2019.

[9] Andrew Aziz, Customer Segmentation based on Behavioral Data in E-marketplace, Teknisk- naturvetenskaplig fakultet UTH-enheten, 2017.

[10] Tushar Kansal; Suraj Bahuguna; Vishal Singh; Tanupriya Choudhury, Customer Segmentation using K-means Clustering, CTEMS, 2018.

[11] Carnein Matthias; Trautmann Heike, Customer Segmentation Based on Transactional Data Using Stream Clustering, Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2019.

[12] Carnein; Matthias; Assenmacher; Dennis; Trautmann; Heike, An Empirical Comparison of Stream Clustering Algorithms, The Computing Frontiers Conference, 2017.

[13] Asosiasi Penyelenggara Jasa Internet Indonesia, " Magazine APJI (Asosiasi Penyelenggara Jasa Internet Indonesia)" (2019): 23 April 2018.

[14] Asosiasi Penyelenggara Jasa Internet Indonesia, "Mengawali integritas era digital 2019 - Magazine APJI (Asosiasi Penyelenggara Jasa Internet Indonesia)" (2019).

[15] Laudon, Kenneth C., and Carol Guercio Traver. E-commerce: business, technology, society. 2016.

[16] statista.com. retail e-commerce revenue forecast from 2017 to 2023 (in billion U.S. dollars). (2018). Retrieved April 2018, from Indonesia, https://www.statista.com/statistics/280925/e-commerce-revenueforecast-in-indonesia/.

[17] Renjith, S. Detection of Fraudulent Sellers in Online Marketplaces using Support Vector Machine Approach. International Journal of Engineering Trends and Technology (2018).

[18] Roy, Abhimanyu, et al. "Deep learning detecting fraud in credit card transactions." 2018 Systems and Information Engineering Design Symposium (SIEDS). IEEE, 2018.

[19] Zhao, Jie, et al. "Extracting and reasoning about implicit behavioral evidences for detecting fraudulent online transactions in e-Commerce." Decision support systems 86 (2016): 109-121.