A Machine Learning Approach to Identify Potential Customer Based on Purchase Behavior

Conference Paper · March 2019

DOI: 10.1109/ICREST.2019.8644458

Adil Mahmud Choudhury
American International University-Bangladesh
4 PUBLICATIONS **19** CITATIONS

Kamruddin Nur
American International University-Bangladesh
39 PUBLICATIONS **217** CITATIONS

A Machine Learning Approach to Identify Potential Customer Based on Purchase Behavior

Adil Mahmud Choudhury
Department of Computer Science
American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
adil@mahmud.ch

Kamruddin Nur
Department of Computer Science
American International University-Bangladesh (AIUB)
Dhaka, Bangladesh
kamruddin@aiub.edu

***Abstract*—To lift the revenue boundary and stay ahead of the competitors it is important to understand customer's purchase behavior. Different business industries proposed different policies to explore the potentiality of a customer based on statistical analysis. In this paper, we rather propose a machine learning approach to identify potential customers for a retail superstore. The paper proposed an engineered approach to classify potential customer, based on previously recorded purchase behavior. Using this classification as ground truth, we then apply machine learning algorithms to find a pattern to predict potential customers with an accuracy of 99.4%.**

***Keywords*—Machine Learning, Purchase Behavior Analysis, Customer Classification**

I. Introduction

The focus of modern grocery superstore business has been shifted to the customer-centric organization. Customers are the most important factor for a business. Some customers can help the business to generate more profit compared to the others. A loyalty-prone customer intends to stay with the supplier who can provide the quality products. On the other hand, a deal-prone customer will always look for a better offer from a competitor. Customers can be classified as profitable and unprofitable [1].

Bringing a new customer to the business is cost-intensive. Because it involves various marketing strategies without having prior knowledge about the customer. However, a business can easily apply specific marketing strategies based on previous customers information, when it approaches it's existing customers. According to the study of Reichheld and Teal [2], a business can increase its profit up to 95% by increasing five percent customer retention. So, businesses are more focused on building relationships with their customers. Companies must learn about their customers purchase preferences to build a long-term relationship with them.

Data accumulation is highly increased in recent years which leads to a great interest in the field of machine learning (ML). The computer has proven its usefulness in predicting and pattern findings by learning the data. The capabilities of machine learning (ML) were extensively explored and significant success were achieved in the field of language translation [3], [4], self-driving [5], text recognition [6], image recognition [7], [8] and voice recognition [9], [10]. Modern organizations are combining technology with business policies to gain a competitive advantage [11]. This has resulted in a competition of finding relevant talent to gain the advantage in technology. People expertise in machine learning and data science are in high demand with the technology itself evolving every day. In recent times, giant companies are spending a rigorous amount of money to gain the technological advantage over their competitors. Studies done in 2015 shows that companies like Google, Facebook and Amazon had spent over USD $8.5 billion to acquire companies working on Artificial Intelligence. This is an increase of 400% since 2010.

Grocery superstore business has been accumulating huge data from customers. Due to the digital revolution, most of the grocery superstores are using computer software to manage sales and customer data. Exploring and analyzing this data can provide new insights which can lead to profit acceleration. In this paper, we will analyze the purchase behavior of a customer using machine learning. Machine learning techniques can be divided into supervised and unsupervised learning. A supervised machine learning model is built based on previously known purchase behavior. Once the model is built, it can generate potentiality score for a new customer purchase pattern. A supervised model is built using labeled data. On the other hand, an unsupervised model does not have any labeled data, rather classifies customers into clusters based on similar purchase behavior.

According to The Pareto Principle [12], twenty percent of customers create eighty percent of the profit. These twenty percent customers are considered as the potential customer. The purpose of this research is to explore how machine learning technique can be used to identify potential customers in a business-to-customer (B2C) sales context, using customer sales data. There are two goals in this research. (i) Implementing machine learning models in a real-life scenario to find the limitations and possibilities. (ii) Application of machine learning in business data to predict potential sales.

The paper is organized as follows, section II discusses the literature review and related work. Research methodology is explained in section III. Section IV discusses the steps of the research. Experiments and result is briefly discussed in section V. Conclusion, limitation and future work is discussed in section VI.

II. Related Work

Studying customer behavior is age old problem which had initially attracted by the researchers of business industries. Every business wants to keep their customers on the long-term basis. For last few decades, business researchers were exploring the importance and strategies for building long-term relations with customers [13], [14]. In an early research by E. Gummesson [15] had discussed the importance of the long-term relationship between business and its customer. A close relation to customers will lower marketing cost per customer. This does not mean that a new customer would not be desirable, but it means the business should focus more on how it can build strong long-term relations with its customers. Relationship marketing concepts are introduced in 80s and few notable studies [13], [14], [16] discussed long-term relationship approaches. To build a strong relationship with your customer it is important to understand their purchase behaviors. Though there were many studies in 90s to understand customer purchase pattern, it was not sufficient. Approaches were way more expensive since lots of handwritten data analysis were involved. Due to the involvement of high cost, none of them were feasible to apply in a practical business scenario.

In 2000, the modern technology took over the business industries for data storing. Which brought a scope for the researchers to explore the data in minimalistic setup. To model a customer behavior, a range of approaches had been proposed. Marzia et al. [17] had discussed detailed literature reviews related to the application of predictive analytics in customer relationship management. Xu and Walton [18] had conducted a research on customer relationship management to understand customer demands. They had also proposed an analytical CRM system for customer knowledge accusation. Buckinx et al had proposed a prediction model for the customers future spending patterns [19]. Using a transactional database, they analyzed customer behavior. Vanderveld et al. [20] had used machine learning technique to describe customer lifetime value for a real-world business. Guimei et al. [21] had explored Alibaba sales data and proposed prediction requirements and most important features using feature engineering and machine learning.

As this is the first research on this topic of purchase behavior analysis using machine learning, all our approaches are unique. We have used the feature engineering technique to achieve more accurate results. Our study shows that machine learning and feature engineering can be a legitimate tool for potential customer prediction.

III. Methodology

Machine learning has many successful footprints in various domains. We used the machine learning approach to analyze a customers purchase behavior. A supervised machine learning technique is used in our study to classify potential customer. At first, we collect the data from sales software and pre-processed it. The quantity of item sold by the unit is considered as features. We used to engineer the dataset to label the data. Using this labeled data, we had trained our supervised model for classification. After assessing the purchase pattern of a customer, the system generates a potentiality score between 0 and 1, where 1 stands for higher potentiality and 0 for no potentiality. Classification related computational measures are learned from the methods. This system involves in a large amount of data pre-processing. Complete knowledge of the data required to determine the present features. To increase the accuracy, features were engineered. Here we have represented detailed explanations of data profile along with data pre-processing and labeling. We have also discussed the methods and tools used to apply machine learning algorithms in our problem domain to gain the insight knowledge and achieve the goal. Figure 1 presents the complete illustration of our proposed methods.

A. Dataset

Our dataset consists of 9259 customers sales data. Over the three months, 194439 invoice entries are given. On average a customer made 7 purchase each month. Each invoice contains minimum one and maximum 83 types of items. Items are grouped into 19 categories. We consider six glossary categories for this research. We extract all sales information related to our research from the dataset and applied machine learning approach to create the classification model.

Grocery superstore named ”Taradin Super Shop” provided the sales data for this research. Taradin is located in the Sylhet district of Bangladesh. Taradin has been doing grocery business since 2001. They have been using customized sales software for billing. Since the database of their software was designed to meet their custom requirement, we had to apply several techniques to tailor data and extract feature as per our research requirements.

B. Feature Selection

To identify significant features first we have to understand their database structure. From database engineering perspective, a single invoice data of a customer was saved into multiple tables. A database table named *product_sale_master (Figure 2)* stores invoice number, item barcode, the price of the item, sold quantity and measurement unit of that item.

Each item of an invoice was stored into this *product_sale_master* data table. For example, an invoice with five items will make five rows entries in *product_sale_master* table. This table did not provide any information on ”who bought the item?” or ”what was the item?”. For our research, it was important to find the relation between invoice, customer id, measurement unit and quantity of each sold items. *product_sale_master* table only gave us data on invoice, quantity and measurement unit, but it didn’t tell us item details and customer ID. So, we had kept looking and found another table named *billing_details (Figure 3)* contained invoice number, date of sale, time of sale, customer ID, amount of sale, vat and bill code.

From this *billing_details*, we had found the relation between an invoice number and customer ID but still, the product details were not known to us. We had to look into other

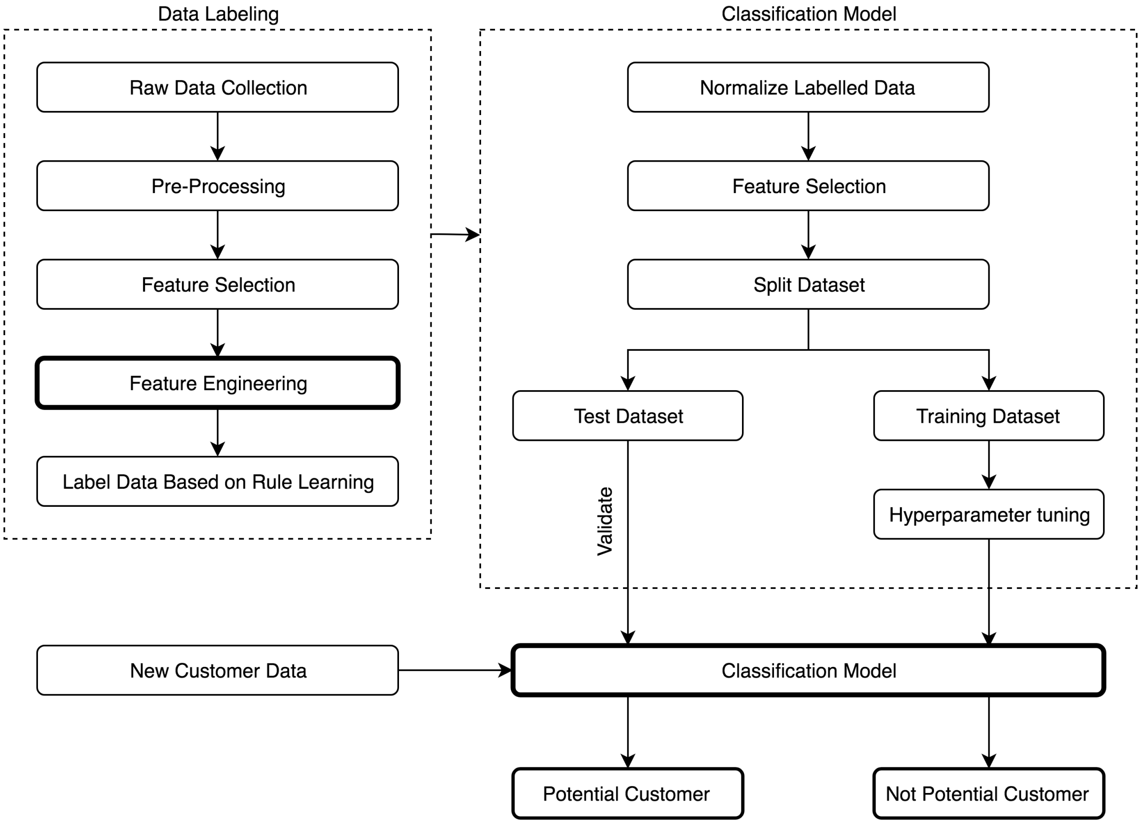
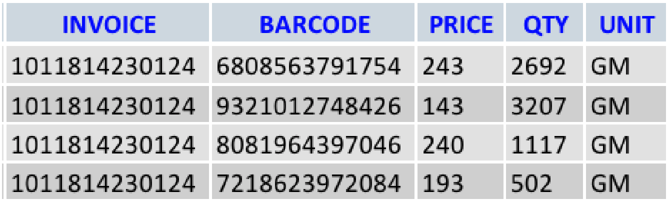
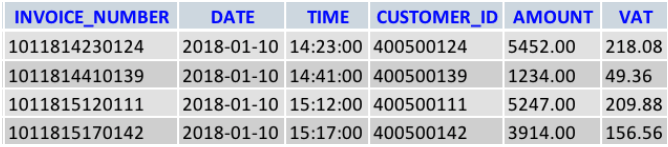


Fig. 1. The Complete Workflow of the Proposed Method



INVOICE	BARCODE	PRICE	QTY	UNIT
1011814230124	6808563791754	243	2692	GM
1011814230124	9321012748426	143	3207	GM
1011814230124	8081964397046	240	1117	GM
1011814230124	7218623972084	193	502	GM

Fig. 2. Partial view of product_sale_master table

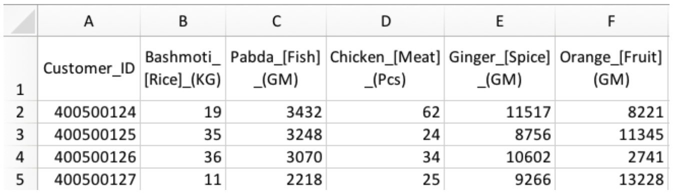


INVOICE_NUMBER	DATE	TIME	CUSTOMER_ID	AMOUNT	VAT
1011814230124	2018-01-10	14:23:00	400500124	5452.00	218.08
1011814410139	2018-01-10	14:41:00	400500139	1234.00	49.36
1011815120111	2018-01-10	15:12:00	400500111	5247.00	209.88
1011815170142	2018-01-10	15:17:00	400500142	3914.00	156.56

Fig. 3. Partial view of billing_details table

tables to get product details. We had mapped multiple complex database queries to extract data according to our requirements. We had collected 9259 customers data for three months for our research. A single quantity was calculated analyzing all invoices, for each item by summarizing the quantity of that particular item bought by a customer within our calculated time period. A partial view of our raw dataset can be seen in Figure 4.

From their billing software, we get explicit knowledge about the sale but none of these help us in any way to label the data. So from Taradin's accounting department, we had collected the data of profit with respect to each category. Sales on each category contribute a percentage in the total profit. Profit variation of each category was very different in each month, hence we consider the mean of profit contribution by each category for the given period. We can see the overall mean profit contribution by categories in Figure 5.



	A	B	C	D	E	F
1	Customer_ID	Bashmoti_ [Rice]_(KG)	Pabda_[Fish] _(GM)	Chicken_[Meat] _(Pcs)	Ginger_[Spice] _(GM)	Orange_[Fruit] (GM)
2	400500124	19	3432	62	11517	8221
3	400500125	35	3248	24	8756	11345
4	400500126	36	3070	34	10602	2741
5	400500127	11	2218	25	9266	13228

Fig. 4. Partial view of accumulate Raw Dataset

To analyze customer purchase behavior, it is important to explore the quantity of each item bought by a customer. So we collected the amount of all items bought by individual customers and consider each of those items as a feature. But those features did not give us good results. So we calculate the mean of each category and consider that as features. Those engineered features improve the overall prediction result.

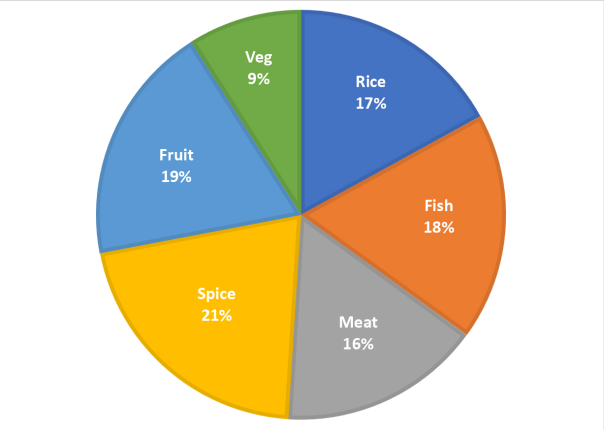


Fig. 5. Mean Profit contribution by categories

C. Data Pre-processing

Our tailored dataset of this research contained items with various unit scale. Using multiple unit measurement could bias the result. As the first step of our data preprocessing, we had to bring all our data into a common unit scale. We had ”Kilo Grams”, ”Grams” and ”Pieces” as measurement units in our dataset. We had converted all weight measurement from Kilo Grams to Grams. Generalizing measurement unit ”Pieces” with ”Grams” was challenging. We were informed that items sold in ”Pieces” were bought in ”Kilo Grams” from the wholesale market. From this knowledge, we had mapped items wholesale measurement unit with the retail measurement unit. Once all data had been formed in a single measurement unit, we had standardized and scaled before using in machine learning to ensure that the fields with larger numeric values that did not bias the results.

D. Customer Classification

Customers are classified into two classes. (a) High Potential Customer & (b) Low Potential Customer. For the purpose of machine understanding we present those two classes using 1 and 0. In the data set *High Potential Customer* is represented using *1* and *Low Potential Customer* is represented using *0*. Customer is classified using *Profit Threshold* λ.

Profit Threshold is calculated from the features of customer purchase historical data along with their respective profit percentage. We have customer purchase data for three months. Also, we have the profit percentage on each item sold along these three months. All item was categorized into $ic_1, ic_2, ic_3...ic_n$. Where n is the number of total categories. Each of those categories contribute to the total profit. If the profit percentage on each category is $p_1, p_2, p_3...p_n$, then total profit (P) is defined as,

$$P = \sum_{i=1}^{n} p_i \tag{1}$$

Now we calculate the mean σ from the quantity of items bought by a customer from each category. If a customer x buy $i_1, i_2, i_3...i_m$ items from a category ic_1, then the mean (σ) is calculated as,

$$\sigma = \frac{\sum_{i=1}^{m} i_i}{m} \tag{2}$$

For customer x, mean of each category is $x\sigma_1, x\sigma_2, x\sigma_3...x\sigma_n$. The total sold quantity of each category as $tsc_1, tsc_2, tsc_3...tsc_n$. Based on item purchased by a customer x, the profit contribution (γ) of customer x can be computed as,

$$\gamma_x = \sum_{i=1}^{n} (\frac{p_n}{tsc_n}) x\sigma_n \tag{3}$$

Using (2) and (3) calculate profit contribution for all customer. Calculate the mean of all customer’s profit contribution and that is our *Profit Threshold* λ. When $\gamma => \lambda$ classify that customer as *High Potential Customer* & for $\gamma < \lambda$, classify that customer as *Low Potential Customer*.

E. Definition

Engineered Features: To enhance performance classification matrix, features were extracted from native features based on comprehension of the data domain.

Profit threshold: A magnitude of profit calculated based on items profit distribution on a given period.

IV. Experiment Result and Discussion

We had started our experiment using our tailored raw dataset, which was extracted from the billing system of the Taradin. Our dataset was not labeled. Once the dataset was ready, we had started applying the machine learning algorithms to get our optimum model for classification. Table I shows the number of potential and non-potential customers distribution of our data.

TABLE I
Number of potential and non-potential customers

Customer Class	Distribution	Percentage
Potential	3973	43%
Not Potential	5286	57%

Our dataset is almost equally distributed into two classes. In machine learning, we must split the data for training and testing. As our dataset distribution is nearly 50% of each class, we will split the data into $40 : 60$ ratios. Forty percent data was used for training and sixty percent data was used for testing as a heuristically good ratio of data trains. Afterward, we had started to apply machine learning algorithms.

As our problem belongs to binary classification, at first we had applied Logistic Regression using the native features. Since the native features have different quantity values within a single item category, the result was very poor. We have achieved 56.78% accuracy using the native features. The recall (true positive rate) was only 1.89%. Which means our prediction model will only classify a customer potential when

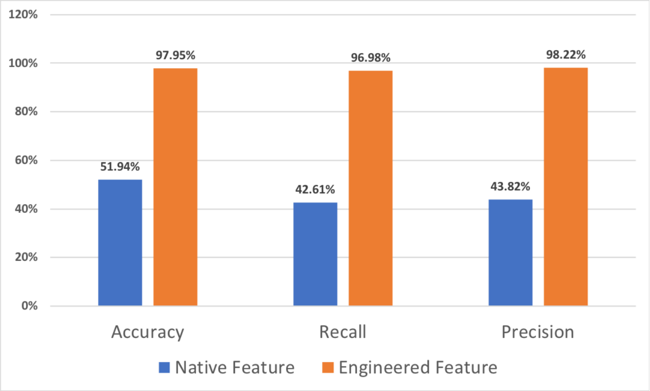


Fig. 6. Decision Tree Results

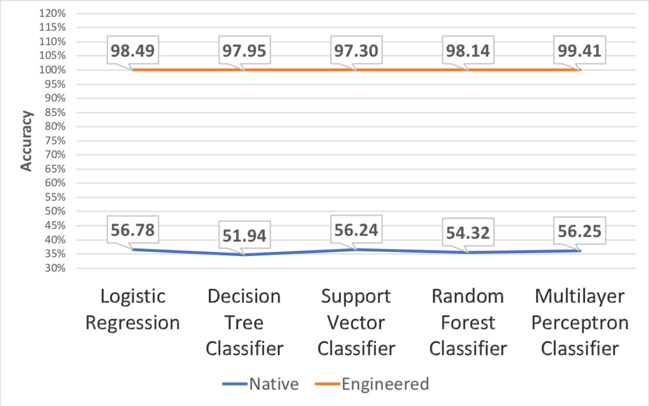


Fig. 7. Comparative Accuracy Result

its very much sure and that can cause misplacing a possible potential customer into a non-potential customer. But when we used engineered features for the same dataset, we had achieved a very good result of 98.49%. Also, the recall and precision were improved 99.56% and 97.00% respectively.

Next, we applied the Decision Tree algorithm. Decision Tree was chosen to apply because it has extreme transparency on rule-based decision making. This algorithm uses in-memory classification model, which makes it more memory efficient. When we had applied decision tree using the native features, the accuracy, recall, and precision was 51.94%, 42.61% and 43.82% respectively. Comparing this result with Logistic Regression, we can see that overall accuracy was decreased by 4.84% but the recall has gained 40.72%. This algorithm provides a fairer prediction compared to Logistic Regression. But using the Decision Tree on engineered feature performed a much better result. Figure 6 illustrates the comparative results of Decision Tree algorithm using the native and engineered feature.

We had also applied Support Vector Classifier and Random Forest at our dataset. For the native feature, both produced the accuracy and precision rate were similar to Logistic Regression and Decision Tree, but the recall percentage of SVC matched with Logistic Regression and the recall percentage of Random Forest was somewhere in the middle of Logistic Regression and Decision Tree. Results of engineered feature gave a similar result for all four machine learning approaches. Table III illustrates the results of the engineered feature.

After applying traditional machine learning, we had applied neural network machine learning to explore the problem. We have used Multilayer Perceptron Classifier. Though the neural network is more of black box approach, it often produces good results. When we had applied MPC on our problem for the native feature, the result was somewhat similar to the SVc, LR, RF, and DT. But there was an outstanding accuracy of 99.41%, which was never achieved before. Also the recall and precision were as good as 98.93% and 99.68% respectively. Now we have the results of multiple algorithms for both native and engineered features. We use Figure 7 to visually illustrate the accuracy differences of those algorithms.

A. *Performance Metrics*

Performance metrics are the established method to compare machine learning algorithms. Accuracy, precision, and recall are used as performance metrics to express the comparative results of machine learning algorithms. This research title belongs to classification problem domain. For classification, performance metrics are calculated using true positive (TP), true negative (TN), false positive (FP) and false negative (FN) of a classification model. True positive (TP) and true negative (TN) illustrate the number of correct classification. On the other hand, false positive (FP) and false negative (FN) express the number of incorrect classification. Equation (4), (5) and (6) define the performance metrics.

$$Accuracy = \frac{TN + TP}{TP + TN + FP + FN} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

In this study, true positive refers to the correct prediction of a potential customer; true negative refers to correct prediction of a non-potential customer. False positive refers to a customer who is not potential but the prediction model classifies it as a potential customer and false negative refers to a customer who is potential but the model prediction classifies it as non-potential.

The classification accuracy will be measured using the performance metrics. Classification Accuracy only fit well on those models where the training dataset distribution of the number of sample per class is similar. The unevenly distributed dataset will heavily bias the accuracy towards the heavy class. For our search training dataset is almost equally distributed, so we can rely on the accuracy of performance metrics. The Recall value of a performance metrics will describe the true positive rate of the model.

B. Native Features Performance

From the result, we can realize that it is very difficult to find a pattern from the native dataset. Logistic regression produced the highest accuracy among all five algorithms, but it is only 56.78%. The neural network algorithm MLP classification rate is 56.24%. Using the native feature both traditional machine learning and neural network produced a poor accuracy.

C. Engineered Feature Performance

Engineered features out performed the result of native features by a large margin. The MLP classifier produced the highest result with 99.41% accuracy. The approaches those produced accuracies more than 97% in finding potential customers are presented in Table II which are significantly improved results than the results of native features.

TABLE II
ENGINEERED FEATURE PERFORMANCE

Algorithm	Accuracy	Recall	Precision
Logistic Regression	98.49	99.56	97.00
Decision Tree Classifier	97.95	96.98	98.22
Support Vector Classifier	97.30	97.99	95.82
Random Forest Classifier	98.14	98.49	97.20
Multilayer Perceptron Classifier	99.41	98.93	99.68

V. CONCLUSION

Although there are many approaches proposed for identifying potential customers, a machine learning approach is rather rare. Our research is first of its kind where machine learning is used to study customer's purchase behavior for a retail superstore. The experiment of potential customer classification achieved prediction accuracy up to 99.4% with recall 98.9% and precision 99.7%. We engineered features to capture the relationship between categories, items, quantity, measurement unit, and sales. The difference in result between native and engineered feature is 42.6%. A business can be highly benefited by identifying their potential customer correctly. The potential customer can be approached with a customized marketing plan which can increase the sale of a business.

As the research is based on the data acquired from the grocery superstore Taradin as a case study, the area of the research only limits to the grocery superstore itself and the industry. However, this results possibly generalize enough to adhere to the problem definition for the whole grocery superstore system in Bangladesh. Taradins existing customers data is considered for this research. As a result, if there is an unexplored customer segment which is not in this dataset, will remain out of this research scope. In future, machine learning can be used to understand customers behavior, a product of interest, buying frequency which will help to plan more appropriate marketing plan and efficient supply chain management.

ACKNOWLEDGEMENT

We would like to thank Taradin Superstore for sharing their sales data for this research. Without their support and generosity, this study could not have been completed.

REFERENCES

[1] Berry, Leonard L. ”Relationship marketing of servicesgrowing interest, emerging perspectives.” Journal of the Academy of marketing science 23, no. 4, pp. 236-245, (1995).

[2] Reichheld, Frederick F., and Jr WE Sasser. ”Zero defections: Quality comes to services.” Harvard business review 68, no. 5, pp. 105-111, (1990).

[3] Collobert, Ronan, and Jason Weston. ”A unified architecture for natural language processing: Deep neural networks with multitask learning.” In Proceedings of the 25th international conference on Machine learning, pp. 160-167. ACM, 2008.

[4] Goldberg, Yoav. ”A primer on neural network models for natural language processing.” Journal of Artificial Intelligence Research 57, pp. 345-420, (2016).

[5] Ohn-Bar, Eshed, and Mohan Manubhai Trivedi. ”Looking at humans in the age of self-driving and highly automated vehicles.” IEEE Transactions on Intelligent Vehicles 1, no. 1, pp. 90-104, (2016).

[6] Zhang, Xiang, Junbo Zhao, and Yann LeCun. ”Character-level convolutional networks for text classification.” In Advances in neural information processing systems, pp. 649-657. 2015.

[7] Rosten, Edward, and Tom Drummond. ”Machine learning for high-speed corner detection.” In European conference on computer vision, pp. 430-443. Springer, Berlin, Heidelberg, 2006.

[8] Chapelle, Olivier, Patrick Haffner, and Vladimir N. Vapnik. ”Support vector machines for histogram-based image classification.” IEEE transactions on Neural Networks 10, no. 5, pp. 1055-1064, (1999).

[9] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. ”Speech recognition with deep recurrent neural networks.” In Acoustics, speech and signal processing (icassp), 2013 ieee international conference on, pp. 6645-6649. IEEE, 2013.

[10] Graves, Alex, and Navdeep Jaitly. ”Towards end-to-end speech recognition with recurrent neural networks.” In International Conference on Machine Learning, pp. 1764-1772. 2014.

[11] Ling, Christine Tan Nya. ”Knowledge management acceptance: Success factors amongst small and medium-size enterprises.” American Journal of Economics and Business Administration 3, no. 1, pp. 73-80, (2011).

[12] Pareto Principle In Wikipedia. Retrieved September 2, 2018, from https://en.wikipedia.org/wiki/Pareto_principle.

[13] Gronroos, Christian. ”Relationship approach to marketing in service contexts: The marketing and organizational behavior interface.” Journal of business research 20, no. 1, pp. 3-11, (1990).

[14] Rosenberg, Larry J., and John A. Czepiel. ”A marketing approach for customer retention.” Journal of consumer marketing 1, no. 2, pp. 45-51, (1984).

[15] Gummesson, Evert. ”The new marketingdeveloping long-term interactive relationships.” Long range planning 20, no. 4, pp. 10-20, (1987).

[16] Berry, Leonard L. ”Relationship marketing.” American Marketing Association, 1983.

[17] Mirzaei, Tala, and Lakshmi Iyer. ”Application of predictive analytics in customer relationship management: a literature review and classification.” In Proceedings of the Southern Association for Information Systems Conference, pp. 1-7. 2014.

[18] Xu, Mark, and John Walton. ”Gaining customer knowledge through analytical CRM.” Industrial management & data systems 105, no. 7, pp. 955-971, (2005).

[19] W. Buckinx, G. Verstraeften, D. Poel, ”Predicting Customer Loyalty Using the Internal Transactional Database”, IEEE, Expert Systems with Applications, 32(1), 2006.

[20] Vanderveld, Ali, Addhyan Pandey, Angela Han, and Rajesh Parekh. ”An engagement-based customer lifetime value system for e-commerce.” In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 293-302. ACM, 2016.

[21] Liu, Guimei, Tam T. Nguyen, Gang Zhao, Wei Zha, Jianbo Yang, Jianneng Cao, Min Wu, Peilin Zhao, and Wei Chen. ”Repeat buyer prediction for e-commerce.” In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 155-164. ACM, 2016.