Comparison of Deep Learning Algorithms to Predict Customer Churn within a Local Retail Industry

Alexiei Dingli, Vincent Marmara, and Nicole Sant Fournier

***Abstract*—A top priority in any business is a constant need to increase revenue and profitability. One of the causes for a decrease in profits is when current customers stop transacting. When a customer leaves or churns from a business, the opportunity for potential sales or cross selling is lost. If a customer leaves the business without any form of advice, the company may find it hard to respond and take corrective action. Ideally companies should adopt a proactive and identify potential churners prior to them leaving. Customer retention strategies have been noted to be less costly than attracting new customers. Through data available within the Point of Sales (POS) systems, customer transactions may be extracted and their buying patterns may be analysed. This paper demonstrates how through transactional data features are created and may be identified as significant to predict churn within the retail industry. The data provided within this paper pertains to a local supermarket. Therefore, the churners identified and results attained are based on real scenarios. The novelty of this paper is the concept of implementing deep learning algorithms. Convolution Neural Networks and Restricted Boltzmann Machine are the selected deep learning techniques. The Restricted Boltzmann Machine attained the best results that of 83% in predicting customer churn.**

***Index Terms*—Customer churn, deep learning, retail grocery industry.**

I. Introduction

Customer retention is a critical problem which is encountered across various industries. Common reasons as to why customers churn include: Competition are offering similar products at cheaper prices; Word of mouth or negative marketing through social media; Competition might have better customer service or customers might have relocated. Research [1] indicates that the cost of retaining a customer is less than attracting new ones. This is due to marketing costs required to appeal to new customers. For this reason, together with the increase of competition it have become pivotal that the current customers base is retained. Normally, customers churn gradually and not abruptly. This means that by analysing customers historic buying patterns one can adopt a proactive approach in predicting churn. Since all transactions are inserted through POS and stored in databases, understanding customers' needs and patters is possible as data is accessible. According to [2], executives are dedicating marketing budgets to focus on customer retention campaigns. Various models designed to predict churn focus on statistical and renowned machine learning algorithms including Random Forest and Logistic Regression. This paper focuses on two aspects when predicting churn within the grocery retail industry. The first is based on the features which will be passed on to the model. Instead of using customers buying trends to cluster the individuals, these values will be created as features and are passed to the model. Therefore, for each customer various features are created to allow the model to learn and identify patterns per individual. For this reason, two datasets are created to test and evaluate how data should be represented to predict churn. The second aspect is the implementation of the algorithms. The novelty of this study is the use of deep learning to predict churn within the grocery industry. To our knowledge, this is the first study which implements deep learning within this industry. The strength of using deep learning is that it can reveal hidden patterns within the available dataset.

Manuscript received July 24, 2017; revised October 3, 2017.

Alexiei Dingli and Nicole Sant Fournier are with the Department at University of Malta, Malta (email: alexiei.dingli@um.edu.mt, Nicole.sant-fournier.15@um.edu.mt).

Vincent Marmara is with the Faculty of Economics, Management and Accountancy; University of Malta, and also with Faculty of Business and IT; University of Malta, Malta (email: Vincent.marmara@um.edu.mt).

II. Background and Literature Review

Customers are being placed at the forefront of the business and have become the dominating factor. With this in mind, businesses need to provide customers with incentives so as to reduce the probability of them moving to competitors. A small negative interaction with a customer, might mean that the customer might churn [2]. An important aspect within the business is to have a good understanding of customers' needs, whereby holistic views of their patterns may be analysed. When customers are satisfied with the service or products, customer loyalty increases [3].

Authors [4], further discuss that revenues and margins will increase if customers are retained when compared to the cost of attracting new customers. Applying statistical techniques and machine learning algorithms on available data may guide companies in identifying hidden trends and customer behavioural patterns. Implementing data mining techniques to predict churn may give companies a competitive edge in improving the relationship with customers. Using customer churn models which correctly classify churn, companies have added value. These include increase in profitability and reduce churn [5].

A. Definition of Churn

Churn is a term used within the marketing field to indicate that a customer has moved to a competitor or has stopped transacting. Churn may be defined as customers who have a high probability to stop transacting with the company [1] or as described by [6]: churn may be identified when a customer's purchasing value falls beneath a threshold across a predefined

doi: 10.18178/ijmlc.2017.7.5.634

period of time. Within the Grocery Retail Industry, the identification of the exact moment a customer will churn is hard to define. Customers do not stop purchasing from the store abruptly, but rather they partially defect -in other words they gradually move to a competitor [7]. Based on this definition, [6] predicts churn for the grocery industry based on the first product category the customer has purchased. This provides a good indication on whether the customer is loyal towards the store or not.

B. Convolution Neural Network - CNN

CNN have shown excellent performance results in image classification and object recognition [8]. The architecture of a CNN is compiled of layers whereby the input of one layer is the output of the next layer. Multiple feature maps are required which will compute a smaller output than the input but will have the same depth. The input is passed to the convolution layer containing an activation function. Following this, the output of this layer is sent to a max pooling layer [9] discusses that this layer creates a smaller and concise feature map of the input. Next, is a fully connected layer whereby all nodes from the max pooling layer are connected to the neurons within this layer. The number of fully connected layers diff ers on the depth of the data. The output from this layer is the classification result for the image. Similar to neural networks, backward and forward passes are used to train the algorithm [10]. Following this, through back propagation, the loss function is computed as seen in (1).

$$E_{total} = \sum \tfrac{1}{2} (target - Output)^2 \quad (1)$$

A learning rate is identified whereby the weight which has created this loss is updated accordingly. Defining the learning rate is important as the higher the rate the quicker the model takes to converge. On the other hand, care needs to be taken as if the value is too high, the optimal point will not be reached.

C. Restricted Boltzmann Machine

Restricted Boltzmann Machine (RBM)are a stochastic neural network, whereby the algorithm is used to identify patterns within data [11]. A RBM consists of three layers, the visible layer, hidden layer and output layer. When seen as a graph, it is a depicted as a bipartite graph. Each node within the visible layer is connected to each node within the hidden layer [12]. The restriction of RBM is that the nodes within each layer are not connected [13]. Through the weights and biases passed through the layers, important features and patterns are defined. Restricted Boltzmann Machines is a novel technique for classification problems [14] designed a model using RBM as a supervised classification model named ClassRBM. Similar to RBM structure, this is a three-layer model. the first layer is the visible layer containing inputs; the second layer is a hidden layer with hidden units, and the third is the binary output. The advantage of this as discussed by [15] is that it can represent any distribution over binary vectors and the probability can be improved by increasing the number of hidden units. The approach adopted to train the model are two, the generative approach or the discriminative approach. The former focuses on improving the likelihood function for a joint distribution, whilst the second is a conditional distribution. The ClassRBM, may be used as a standalone classifier or in conjunction with other models [11].

III. Methodology

A. Extract Transform and Load

The data collection was a task in itself, as for the results to be realistic actual data was required. Various local supermarkets were contacted to examine the possibility to use their data. After various attempts, a supermarket responded and provided data for research purposes. Applying the Kimball methodology, the data available traverses through a process of Extract Transform and Load (ETL).

1) Extract

Once the available data has been cleaned from any anomalies, the data available may be extracted from the database and placed into the data warehouse. This is done by executing scripts which are used to extract the data from the source database. The scripts designed are for the fact table and dimensions. The fact table contains a denormalised form of data, whereby the transactional data together with the customer id and stock id are stored. As for the dimensions, separate scripts are available to extract the descriptive aspect.

2) Transform

Within this process, rules and aggregations are created. Counts for customers frequency and number of receipts are pre-calculated so that on reading data counts are pre-aggregated. So as to abide to data protection, the process eliminates customer names, stock brand, category and department descriptions from the dataset.

3) Load

Following the data transformation, data is loaded into a data warehouse which include three dimensions and one fact table. This include the Customer and Stock dimensions which have primary keys and indexes set up. The fact table contains the Ids for the stock, customer and time together with the sales quantity and sales value.

B. Parameter Selection

Having transformed and analysed the data as discussed in the previous section, the next exercise requires the formation of parameters which will be passed to the selected algorithms. Keeping in mind the objective of the project, the parameters created revolve around the customers buying trends. Two datasets are created based on the Recency, Frequency and Monetary (RFM) values. Recency represents the last time the customer has transacted with the supermarket, the number of times a customer has frequented the supermarket is defined as Frequency and the value spent at the supermarket is identified as Monetary. The first dataset is based on four months of historic RFM data whilst the second dataset includes 11 months of historic data.

C. Churn Defined for the Model

As defined by researchers [1] and [6], customer churn may be seen from two perspectives: the first is a slow churner, whereby the customer frequency and monetary values decreases over time. The second form is when the customer stops frequenting the supermarket. For the purpose of this

research, the later definition will be taken into consideration. So as to identify churners, the dataset is divided into two-time frames or time-windows. The first window is the predictive window which identifies active customers. Activity is defined by the customers having transactions within this period. Customers which have activity within the first window are tagged as non-churners, whilst the remaining customers are marked as churners. The latter are eliminated from the analysis as they have already churned. The next window, the churn assessment window, classifies the remaining customers as a churner or non-churner. If customers transact in this period they are non-churners, whilst if no transactions are seen then they are churners.

D. Implementation

Two deep learning algorithms are implemented on the discussed datasets to identify whether the overall accuracy is dependent on the volume of historic data. To abstain the model from over fitting or to avoid creating a bias to a majority class, the dataset is balanced having the same number of examples available for both classes. Furthermore, outliers are removed from the dataset prior to dividing the dataset into training and testing. The apportionment for this is by randomly dividing the dataset into 75:25 respectively.

a) Convolution neural networks

A pre-requisite to the CNN inputs is that features are placed into a matrix. Therefore, the class label and features of the training and testing dataset are placed into two separate matrices. Prior to placing them into the matrix, the features are normalised and transposed. The first dataset contains 16 features, therefore a 4 by 4 input matrix is designed.

The network consists of one convolution network having a 3 by 3 kernel. The sigmoid function is the selected activation function as the outputs are binary values. The next layer is the sum pooling layer, whereby the sum of inputs is taken. The objective of this level is to control over fitting by reducing the spatial size of the input.

A kernel of 2 by 2 with a stride of 1 by 1 which defines how the patch slides towards the left, right, up or downwards of the matrix. Following this, to ensure that over fitting does not occur, a drop out of 0.1 is used. Following this, two fully connected networks are used, whereby the first network has 5 hidden layers, sigmoid activation function and a drop out of 0.1. This is followed by the second fully connected network which has three hidden layers.

The parameters passed to the CNN are seen in Table I. The number of iterations to train the model is of 30, this number is selected so that the model will not over-train on the dataset. The batch size is set to 100 which indicates the 'wiggle' or 'jump' between parameters. The learning rate at which the model is trained is set to 0.00625 and momentum of 0.9. These values are used as having a high learning rate might result in the system moving around and getting stuck in the local minima or maxima. Similarly, momentum is used so as converge rates on deep networks. The penalty depicted as wd is set to 0.0003, which is used to increase regularisation. The second datasets are passed to the designed CNN. The difference is the number of inputs used, for this model 25 features together with one label are implemented. The features are transformed into a 5 by 5 matrix. The same transformation is applied to the class label and test data.

TABLE I: PARAMETERS FOR CONVOLUTION NEURAL NETWORK

Parameter	Value
X	Train.array
Y	Train.y
Num.round	30
Array.batch.size	100
Learning.rate	0.00625
Momentum	0.9
Wd	0.0003
Eval.metric	mx.metric.accuracy
Epoch.end.callback	100

b) Restricted boltzmann machine

For the RBM, the dataset is converted into binary as proposed by [15] whereby each feature is split into a binary value. This is a laborious process as for each of the features defined in the datasets, the representative value in binary is required. Therefore, each customer will consist of a vector of binary values which is a representation of their features. For instance, the measure frequency is grouped into nine buckets as seen in Table II. Therefore, each customer will have all nine features referring to the frequency for a specific month. If the customers frequents the supermarket between 7 and 12 times, the binary value will be set to 1 for this feature. For the rest of the buckets the binary value will be set to 0. This process is repeated for all features, the final number of features are 178 inputs and one label.

TABLE II: SAMPLE OF CONVERTING FEATURE FREQ INTO BINARY

Feature Name	Description
Freq1	Freq < 6
Freq2	7 < Freq <12
Freq3	13 Freq 18
Freq4	19 < Freq < 24
Freq5	24 < Freq 30
Freq6	. >30

The function to train the RBM requires data to be in a matrix format. Therefore, data is converted and transposed into a matrix as per requirements of the function. The architecture of the RBM is as designed by [16] which includes a visible layer and a hidden layer. 178 nodes are used as inputs which are seen within the visible layer, whilst 15 nodes are available within the hidden layer. Each node within the visible layer is connected to each node within the hidden layer. The restriction within the RBM is that nodes within the same layer are not connected. This makes the graph a bipartite graph whereby nodes from the first layer are only connected to nodes within the second layer.

Since data is represented in binary values, the sigmoid activation function is implemented. As discussed by [17], this function restricts values to 1 and 0. One of the disadvantages of the sigmoid activation function is that neurons might not learn if the weights of the neuron are too high. Therefore, to overcome this, an initial momentum of 0.004 during pre-training and a final momentum of 0.008 are set. Furthermore, the learning rate is set to 0.05 which is the values multiplied for each epoch.

To minimise the average negative log likelihood, the stochastic gradient descent is used. To do this, a partial derivative is complimented, whereby a comparison between the positive and negative phase is carried out. One constrastive divergence is used, whereby each variable is sampled given the other variables is used. This is an iterative process whereby a variable is selected randomly.

IV. Evaluation

To evaluate the algorithms performance per dataset, the label of the unseen dataset is removed and placed in a separate vector. The features are then passed on to the trained models. The predicted result is then compared to the actual label to calculate the accuracy of the algorithm.

TABLE III: Evaluation Metrics for CNN

Evaluation Metric	Model 1	Model 2
Sensitivity	0.59	0.66
Specificity	0.87	0.82
Accuracy	0.68	0.74
Pos Pred Value	0.93	0.86
Neg Pred Value	0.41	0.60
Precision	0.59	0.66
Recall	0.93	0.86
F1	0.73	0.74

A. Convolution Neural Network

Comparing the two models as depicted in Table III, the second model attained a higher overall accuracy of 74% whilst model 1 obtained 68%. The results for Sensitivity are of 59% and 66% for Model 1 and Model 2 respectively. This indicates the percentage of non-churners being correctly classified for this algorithm. The classification of churners is depicted in the results achieved for Specificity. The models obtained 87% and 82% for Model 1 and Model 2. Analysing the positive predicted values, the model correctly classified 93% of non-churners within Model 1 and 86% within Model 2. Evaluating the negative predicted value which represents the classification of customers who will churn, the algorithm achieved is of 41% and 60% for Model 1 and Model 2. Precision which defines the classifiers exactness obtained 59% for Model 1 and 66% for Model 2. To understand the completeness, Recall is used whereby 93% and 86% for Model 1 and Model 2 were obtained respectively. The F1 measure is computed by taking into consideration the customer misclassified as churners and non-churners. Model 1 resulted in a 73% whilst Model 2 attained 74%.

B. Restricted Boltzmann Machine

Evaluating the metrics in Table IV, Model 2 attained the best overall results. The first metric sensitivity, calculates the capability of correctly identifying customers who will not churn. The results obtained for Model 1 and Model 2 are 62% and 74% respectively. Similarly, specificity tests the ability of the model in correctly classifying customers who will churn acquired 87% and 92% for Model 1 and Model 2. The Pos Pred Value reflects the probability of actual non-churners being classified as non-churners, model 1 and model 2 attained 83% and 92% respectively. To analyse the models' correctness in defining churners, the results obtained are 74% for model 1 and 77% for model 2.

Precision defines the classifiers exactness, as this calculates the number of predicted non-churners divided by the total number of actual non-churners. Model 1 obtained 67% whilst Model 2 attained 74%. To understand the completeness, recall is used whereby the results achieved are 83% and 92% for Model 1 and Model 2 respectively. The F1 measure is computed by taking into consideration the customer misclassified as churners and non-churners. Model 1 resulted in a 74% whilst Model 2 attained 82%.

TABLE IV: Evaluation Metrics for RBM

Evaluation Metric	Model 1	Model 2
Sensitivity	0.67	0.74
Specificity	0.87	0.92
Accuracy	0.77	0.83
Pos Pred Value	0.83	0.92
Neg Pred Value	0.74	0.77
Precision	0.67	0.74
Recall	0.83	0.92
F1	0.74	0.82

C. Result Comparison

Evaluating the results attained per algorithm the second dataset (Model 2 as seen in Fig. 1 and Fig. 2) obtained superior results to the first. The increase in overall accuracy for CNN and RBM is of 6% from Model 1 to Model 2. The difference between the two datasets is the number of months taken into consideration to predict churn. This indicates that a relationship exists between historic data and churn prediction.

Including additional months to the dataset provides a holistic view of the customers buying patterns. From the data deep learning algorithms are used to identify patterns and insight on the customers buying trends across months. By having an increase in accuracy confirms that a customer churns gradually. This finding is instrumental to the datasets for churn models, as when additional data pertaining to RFM is included the overall accuracy results improve.

Reviewing the results attained by the algorithms, the RBM obtained the best results in classifying churners. With regard to CNN, considering the input to the model were features transposed into a matrix and not actual images the result is satisfactory

V. Conclusion and Future Works

Customer satisfaction plays a large role in the retail industry. Through the rise of competition in this industry, companies need to ensure that customers are satisfied with the service and quality of products found in the supermarket. Being in a position to predict customer churn gives the company the competitive advantage to be proactive and retain customer who have a high propensity to churn.

This paper discussed the importance of available data supermarkets retain together with the best parameters required to predict churn. From the techniques implemented it is evident that analysing customers historic behavioural

patterns is pivotal to identify churn. This is due to the fact that customers within the retail industry are gradual churners. The results attained from the algorithms are satisfactory attaining 83% with RBM and 74% when CNN was implemented.

Further work includes, analysing customers market basket analysis whereby through association rules products purchased together on a regular basis are defined. This will provide insight to supermarkets on how best to promote their products. Furthermore, for each product purchased per customer a pattern will be defined so as to observe the buying habits per product. Recurrent Neural Networks (RNN) will be applied to the sequence of consumer actions. This algorithm will be used as it performs well in cases of analysing sequences. Based on the purchasing patterns, warehouse manager can ensure that the items defined will be in stock and available for customers to purchase them.

REFERENCES

[1] C. Jie, Y. Xiaobing, and Z. Zhifei, "Integrating OWA and data mining for analyzing customers churn in e-commerce," The Editorial Office of JSSC and Springer- Verlag Berlin Heidelberg, vol. 28, pp. 381-391 2015.

[2] The science behind customer churn. [Online]. Available: http://financeinbusinesslife.info/the-sciencebehind-customer-churn/

[3] S. Neslin, S. Gupta, W. Kamakura, L. Junxiang, and C. H. Mason, "Defection detection: Measuring and understanding the predictive accuracy of customer churn models," *Journal of Marketing Research*, vol. 43, pp. 204-211, 2006.

[4] IBM. [Online]. Available: https://www.ibm.com/developerworks/library/badata-mining-techniques/, Developer works, Accessed: November 2016

[5] E. Siegel, *Predictive Analytics, The power to Predict Who Will Click, Buy, Lie or Die*, Wiley, 2013.

[6] V. Migueis, D. V. den Poel, A. Camanho, and J. Falcao, "Modelling partial customer churn: On the value of first product-category purchase sequences," *Expert Systems with Applications*, pp. 11250-11256, 2012.

[7] W. Buckinx and D. V. den Poel, "Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting," *European Journal of Operational Research*, pp. 252-268, 2005.

[8] Stanford University. CS231. [Online]. Available: http: //cs231n.github.io/convolutional−networks/

[9] M. Nielson, *Neural Networks and Deep Learning*, Online Book, 2016.

[10] Rumelhart, E David., G E. Hinton, and R. J. Williams. Learning representations by back propagating errors," *Cognitive Modelling*, 1988.

[11] D. C. Mocanu, E. Monacu, P. H. Nguyen, M. Gibescu, and A. Liotta, "A topological insight into restricted Boltzmann machines," Springer, vol. 104, pp. 243-270, 2016.

[12] L. Goodfellow, Y. Bengio, and A. Courville. (2016). Deep learning, Book in preparation for MIT Press. [Online]. Available: http://www.deeplearningbook.org

[13] A. Fisher and C. Igel, "An introduction to restricted boltzmann machines," Springer-Verlag Berlin Heidelberg, pp. 14-36, 2012.

[14] [14] H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio, "Learning algorithms for the classification restricted Boltzmann machine," *Journal of Machine Learning Research*, no. 13, pp. 643- 669, 2012.

[15] J. M. Tomczak and M. Zieba, "Classification restricted Boltzmann Machine for comprehensible credit scoring model," *Expert Systems with Applications*, no. 42, pp. 1789-1796, 2015.

[16] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proc. International Conference on Machine Learning*, 2009, pp. 791-798.

[17] B. Huang, M. T. Kechadi, B. Buckely" Customer churn prediction in telecommunications," *Expert System with Applications*, no 39, pp. 1414-1425, 2012.

Prof. Alexiei Dingli is a senior lecturer of Artificial Intelligence within the Faculty of ICT at the University of Malta. He is also one of the founder members of the ACM student chapter in Malta, founder member of the Web Science Research, founder member of the International Game Developers Association (IGDA) Malta and of the Gaming group at the same University. He pursued his PhD on the Semantic Web at the University of Sheffield in the UK under the supervision of Professor Yorick Wilks. While there, he worked on various large projects but his major contribution can be attributed to the Advanced Knowledge Technologies project, one of the largest Interdisciplinary Research Collaborations (IRC) funded by the Engineering and Physics Research Council (EPSRC). For this project, he created two systems which were rated World Class by a panel of international experts whose chair was Professor James Handler (one of the creators of the Semantic Web). These systems were later used as a core component of the application that won the first Semantic Web challenge (2003). His recent work in Mobile Technology and Smart Cities (2011) was also awarded a first price by the European Space Agency. He has published several posters, papers, book chapters and a book in the area. For four years, he also worked as a senior manager in a large government corporation where he got insight into the needs, potential and deficiencies of digital natives. During this time, he also pursued an MBA with the Grenoble Business School in France specialising on Technology Management.

Vincent Marmara is a statistician and researcher by profession. He obtained his first degree in Statistics, Operational Research and Mathematics at the University of Malta. He later advanced his studies by obtaining Masters of Science in Statistics at Sheffield University, UK. Furthermore, he has obtained his PhD in Mathematics (Statistics) from the University of Stirling, Scotland. He was entrusted with numerous research projects both at a national and international level. He led research groups and analyzed data to a high-level scientific extent.

Vincent has over 10 years of experience in the Remote Gaming Industry as a Business Intelligence Analyst and Consultant. He occupied several key important roles such as; member of the faculty board of science (University of Malta), President of the Science Student Society, Financial Officer of the National Youth Council, Deputy CEO and Chief Regulatory Officer within the Malta Lotteries and gaming authority Vincent is a lecturer at the department of Management (FEMA) at the University of Malta and a fellow of the Royal Statistical Society (UK). His main research interests are research and management in the gaming industry, epidemiology and health research, sampling surveys, regression models and Bayesian analysis.

Nicole Sant Fournier is a business intelligence consultant. Her first degree was in Information Systems and Management at London School of Economics. She now has started her Maters in Artificial Intelligence with University of Malta.

Throughout her career, she has had the opportunity to understand and meet key people within industries including the Financial, Retail and Distribution Sector as a business analyst and business intelligence Consultant. Having analysed and implemented solutions for various industries Nicole has an understanding of the business problems companies encounter. Therefore, her thesis focuses on solving the problem of churn prediction using data mining techniques.