Churn detection using machine learning in the retail industry

K. Bernard Agbemadon
FEMTO-ST Institute, CNRS
Univ. Bourgogne Franche-Comte (UBFC)
Belfort, France
kodjo.agbemadon@univ-fcomte.fr

Raphaël Couturier
FEMTO-ST Institute, CNRS
Univ. Bourgogne Franche-Comte (UBFC)
Belfort, France
raphael.couturier@univ-fcomte.fr

David Laiymani
FEMTO-ST Institute, CNRS
Univ. Bourgogne Franche-Comte (UBFC)
Belfort, France
david.laiymani@univ-fcomte.fr

***Abstract*—The constant need to increase sales and profitability is a top priority in every company. Indeed, when the existing consumers stop purchasing from the company, its income tends to drop rapidly. For this reason, customer retention has been considered one of the most important issue in Customer Relationship Management, mostly in the retail industry, as it has been found to be less expensive than acquiring new consumers. The chance for future sales or even cross-selling is missed when a customer stops going to a particular shop. That is why companies have to be proactive and detect potential churners before they leave. This paper shows how transactional data and machine learning can be relevant for the retail industry to forecast churns. To train the machine learning models, a sample of 5,115,472 records of consumers with a loyalty card was obtained from the data warehouse of an European retail company. The results revealed that the machine learning models perform better than linear regression models.**

***Index Terms*—churn prediction, retail industry, machine learning, deep learning**

I. Introduction

Customer retention is a key challenge that is faced across different sectors. Since acquiring new customers is more expensive than retaining them [23], it has become essential for all companies to study the behavior of customers who churn, in order to avoid it in the future.

Customer could churn for several reasons. It can come from bad publicity on social medias or word-of-mouth, from similar products being sold at cheaper prices by the competition or the competition could also have better Customer Service or User Interface/Experience for online purchase [5]. Research reveals that attracting new clients is more expensive than customer retention [7] due to the marketing costs needed to bring new customers. Therefore the preservation of the existing client base has become essential. Customers usually churn gradually and not suddenly, that is why analyzing their historic purchasing patterns [6] could enable companies to detect a drop in their purchasing habits. When companies use loyalty cards, thousands of attribute values are stored for each buyer. Those data include useful knowledge which is often buried in the large array of raw data. It should be noticed that, these datasets contain mainly structured data that can be requested through SQL [18] and semi-structured [2] data such as Excel, JSON and CSV files.

Now, it is widely accepted that Machine Learning manages to perfectly extract hidden characteristics from raw data. Across many different fields, Machine Learning methods have been applied successfully, therefore it could be used to extract knowledge from those raw data.

This study uses a private dataset from Colruyt France, a retail company with 90 supermarkets (700 to 1200 m²), mainly located in the Franche-Comté region, France. This dataset represents customer purchases in the stores and contains 105,488 customers. These customers have so far produced a total of 5,115,472 rows of data.

The novelty of this study is the application of Machine Learning and Deep Learning Techniques on this type of data. Other main contributions are the possibility to bind personal data (age, length of the client relationship, gender, population of the city where the customer lives) to the sales time-series [9] and the data augmentation technique explained in section III-C. In the best scenario the model's precision is 75.60%.

The remainder of the paper is structured as follows. Section II presents the context, regarding the definition of churn and the different levels of difficulty that can be encountered. Section III explains the methodology, from data acquisition to evaluation methods. Section IV details the different models used and section VII presents the obtained results and compares the different techniques used to detect potential churners. Finally, Section VIII summarizes the conclusions drawn from the paper.

II. Context

A. *Definition of Churn*

Churn is a marketing term that refers to a customer who has switched to a competing company or stopped purchasing from your company. Churn can be defined as customers who are likely to stop transacting with the firm in a given period [7]. It can also be defined as [19]: when the average basket of a customer, namely the average amount spent over a period, falls below a threshold over a predetermined period of time. The average basket is described formally in Equation 1.

Let $PurchaseAmount_C(i)$ be the amount spent by a customer C during a week i; and $n = |P|$ the number of weeks of a period P;

$$AverageBasket_C(P) = \frac{\sum_{i=1}^{n} PurchaseAmount_C(i)}{n} \quad (1)$$

It is difficult to pinpoint the specific moment when a customer would churn in the supermarket retail industry. Customers do not suddenly stop buying from the store, rather they partly defect. In fact, they tend to switch to a rival gradually [6]. To be more factual, in this context, a churner is defined by the following rule. Let $P1$ be the period of observation, where the customer's usual purchase amounts are examined. That will be the input to the future churn detection model.

Immediately after $P1$, there is $P2$ which is the period of evaluation, where a change in customer buying habits can be seen. $P2$ will be unknown to the prediction model. Throughout this study, $P2$ is permanently set to 12 to match the requirement of the marketing service. Which means the evaluation period is always 3 months.

The labelling technique will be as follows. If the average purchase during $P2$ is $< 20\%$ of the average purchase during $P1$, then, that customer will be labelled as a churn, if not, he/she will be labelled as a non churner. This 20% is the allowed drop, and is called the reduction factor. It means that the churn could be partial or total. The formal definition of churn is:

Let α be the reduction factor, and C a customer. C is considered as churner if and only if :

$$AverageBasket_C(P1) < \alpha \times AverageBasket_C(P2) \quad (2)$$

Some examples of churners and non churners are shown respectively in figure 1 and figure 2.

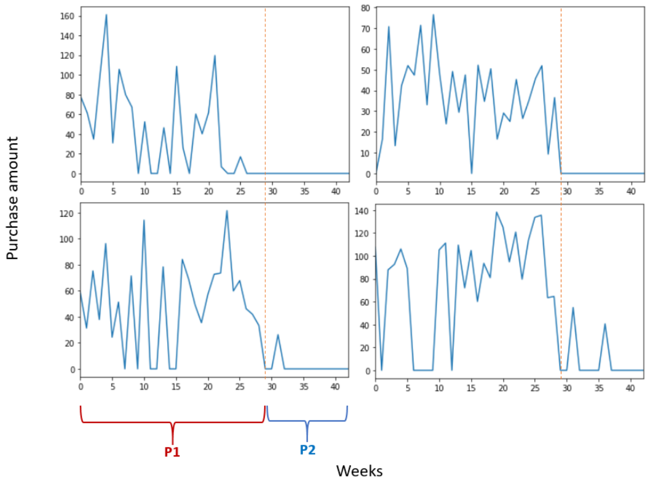

Fig. 1. Four examples of churners. During period $P1$ they used to buy each weeks, and they have totally or partially stopped buying during $P2$.

From these definitions and observations we can deduce three churn cases.

1) *Simple churn cases* These cases are negative slopes [15] with lower noises, as shown in figure 3. Let S the slope for the time-series values as $\hat{y}_i = Sx_i + c$. For each i the error $e_i = |y_i - \hat{y}_i|$ have to respect $e_i \leq \frac{y_i}{n}$, n being the number of periods or the length of the time-series. In other words, the noise does not distort the trend. This kind of churners are relatively easy to detect using a simple linear regression with a threshold.
2) *Churn cases with hidden variables* These cases are similar to the previous ones with the difference that they include hidden variables, namely promotions, soccer championships, weather, or any other external event that impacts the consumer's habit. These hidden variables significantly distort the trend, creating a bias in the classification by a linear regression model. The proposed model for these cases have to be able to read from the input, an extra information about a potential hidden variable. This hidden variable can be formatted as a

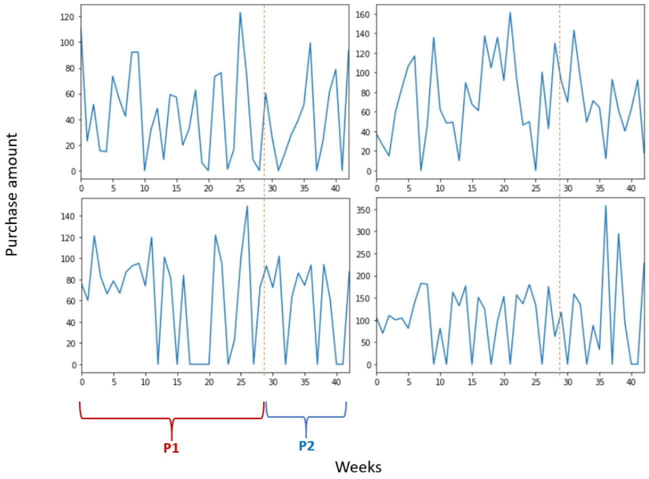

Fig. 2. Four examples of non churners. They used to buy each weeks, both during $P1$ and $P2$. There was no dramatic change in purchasing habits

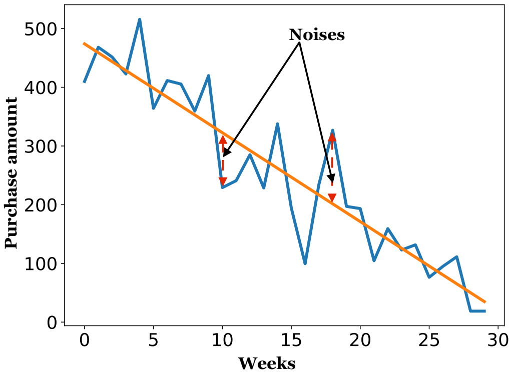

Fig. 3. Churner with lower noise, here the slope can be seen despite the noise.

time-series too.

3) *Difference between churn case and sporadic case* The sales records often include 20% to 30% of customers who come to buy in our stores occasionally. In contrast to regular customers, those who come periodically, i.e. once or several times a week or even every two weeks. The regular customers consider our stores as their main, they will make their biggest purchases there. Even if, when needed they can also make small purchases in a competitor's store closer to them or more easily accessible. Basically, churners are regular customers who are becoming sporadic. This study does not focus on the sporadic cases.

B. Related works

During the past few years, there have been a lot of research in the field of churn prediction. The fact that customer retention is much more economical than customer acquisition was explained in the work [23]. There has been a lot of studies on churn that compared machine learning approaches to classical approaches on data from different domains, namely, telecommunication [3], banking [20] and online [17] subscription. On the other hand, there are just a few studies that specifically target the retail industry, which has certain unique characteristics in terms of consumer interactions and life cycles. In 2017, Dingli [10] compared Restricted Boltzmann Machine (RBM) model to Convolutional Neural Network (CNN) model. At that time RBM achieved the best score with F-measure (β=0.5) of 77% and a Precision of 74%. Since then, convolutional networks have evolved, as have deep networks.

However, the context and the nature of the data can have an effect on the outcome, pushing to use one to privilege one approach over another. For example, in certain domains such as telecommunications [3], the customer rarely subscribes to more than one operator simultaneously. While in retail industry, the consumer may have a main store, where he/she buys periodically, and others where he/she buys sporadically. This study focuses on the retail sector in all its particularities. Also, the data augmentation technique proposed in this study has not been seen in any other similar study.

III. Methodology

A. Data acquisition

This research includes 5,115,472 rows of sales data, distributed in 105,488 customers from a supermarket chain company in France. A pseudonymisation was applied on the entire dataset, which allows the possibility to bind personal info (age, length of the client relationship, gender, population of the city where the customer lives) to the sales time-series [9] in this research. The rows were split into groups where each group identified a customer and is formatted as a 2 dimensions array. Outliers such as customers with 2 active profiles on the same name and address were removed. The length of $P1$ and $P2$ periods II-A in weeks was discussed with the Marketing service of our company. Multiple lengths were tested to finally converge to the ones which produce the highest accuracy. After setting α (the reduction factor) to 0.2, the labelling technique discussed in II-A was applied.

The dataset was then split into a train and test sets. See table I for an overview of the dataset.

TABLE I
An overview of the dataset. Sporadic customers were not considered during this study.

Labels	Values
Number of customers	105,488
Number of sporadic customers	42,636
Total number of rows	5,115,472
Number of customers labelled as Non churner	61,259
Number of customers labelled as Churner	1,593

B. Highly imbalanced dataset

Different characteristics can be found in various types of data. Any of these characteristics can make it difficult for data mining algorithms to extract useful patterns. One of the main issues with the dataset used for this study was the class imbalance. When there is an imbalance between classes, the learning algorithm basically tends to forecasts only the majority class to minimize the error. In the dataset for example, there were 60,000 active customers compared to just 2 thousands churners. In other words, 96% of active consumers compared to just 4% of churners, showing a classic case of class imbalance.

Re-sampling is a widely adopted technique for addressing the class imbalance issue. It can be done in two ways: either over-sampling or under-sampling can be used [12]. With under-sampling, just a subset of the majority class is used to train our models. In this study, a random sample of inputs was excluded from the set of active clients, so that the number of churners and the number of non churners becomes equivalent.

C. Data augmentation by scaling

It consists in generating other time-series to artificially increase the number of samples in one class by changing the scale of the original time-series. From a customer who always buys around 100 euros every week, we can generate a virtual one who will buy around 50 euros every week and another one who will buy around 200 euros every week. The latter two will have exactly the same dynamics as the first. The idea is to artificially create new customers with the same behavior as the original one, but at a different scale. A representation of original and scaled versions can be seen in figure 4.

IV. Models

In the following, a linear regression model was compared to some machine learning approaches that have been proven to work for churn prediction in general. Efficiency and reliability were considered during the comparison.

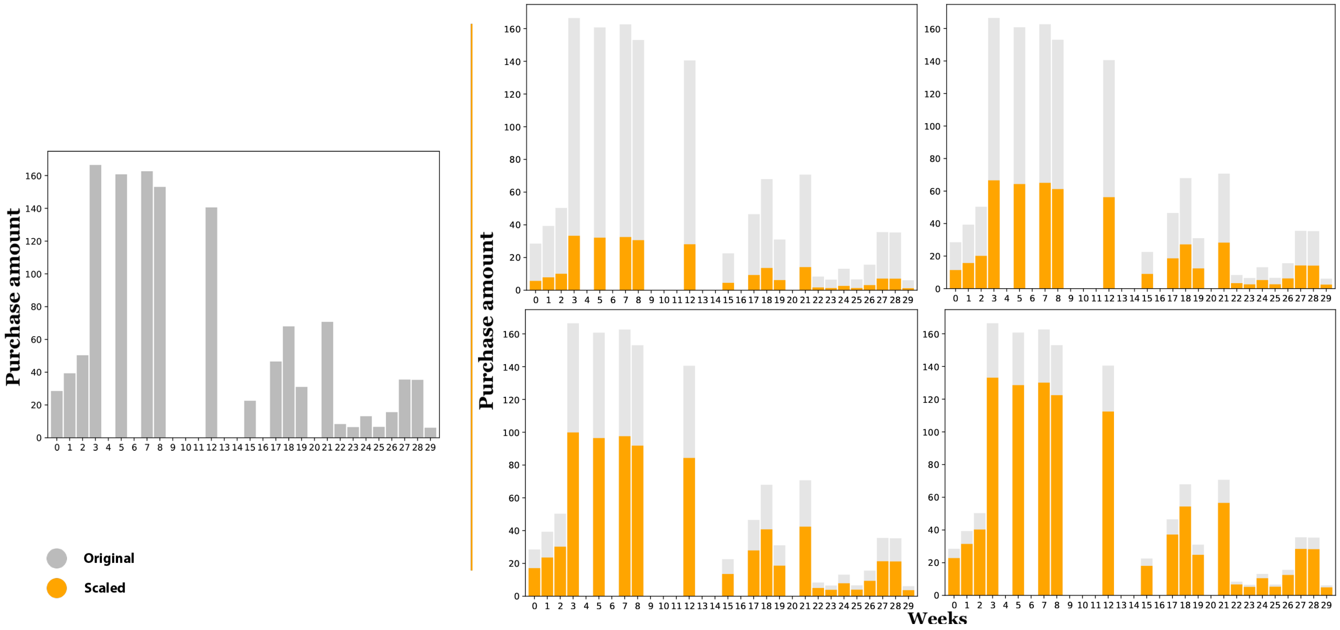

Fig. 4. Original data with its scaled versions with data augmentation. In grey, the original series, and in orange the scaled versions. For the sake of readability in this example, the scale chosen for each scaled versions is $<$ 1x the original. Note that this can be extended to 1.5 times, 2 times...

A. Linear Regression

Linear regression is known in statistics as a linear approach to modeling the relationship between a scalar response and one or more explanatory variables (also called dependent and independent variables) [1]. Linear regression is a lightweight statistical tool useful for gaining insights into customer behaviour, understanding the business and factors influencing profitability [21]. Although linear regression has limited applicability in the business world, as it can only work when the dependent variable is continuous in nature, it is still a well-known technique in situations where it can be used [11].

B. MLP

A Multi-Layer Perceptron (MLP) is a variant of the original Perceptron model proposed by Rosenblatt in 1950 [26]. MLP has proven itself in many fields with the emergence of Deep learning [14]. It has one or more hidden layers between its input and output layers. The neurons are structured in layers, connections are always directed from lower layers to upper layers, and neurons in the same layer are not interconnected [24] as shown in figure 5.

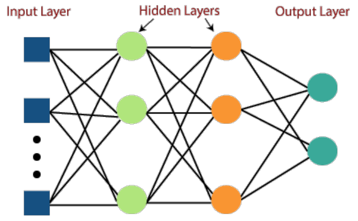

Fig. 5. A graphical representation of a Multilayer Perceptron.

C. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm [8]. XGBoost has become the go-to method and often the key component in winning solutions for classification and regression problems in machine learning contests. As this study is subject to a classification problem, it was deemed useful to test it in order to compare the results with other approaches.

D. LSTM

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) that can learn order dependences in sequence prediction problems [16]. While introduced in the late 90's, LSTM models have only recently become a viable and powerful forecasting technique for time-series. An LSTM rectifies a huge issue that recurrent neural networks suffer from: short term memory i.e. the inability to learn dependencies from long sequences . Using a series of 'gates' [13] each with its own RNN, an LSTM manages to keep, forget or ignore data points based on a probabilistic model. See figure 6 for an LSTM illustration.

V. Evaluation Metrics

Precision, recall, and F-measure are used as measurement tools in this paper to measure the reliability of the various prediction models [22]. TP and FP respectively denote True Positive and False Positive samples, while True Negative and False Negative samples are respectively denoted as TN and FN. The Recall is the proportion of Churn customers that were correctly identified. The recall is intuitively the ability of the classifier to find all the positive samples.

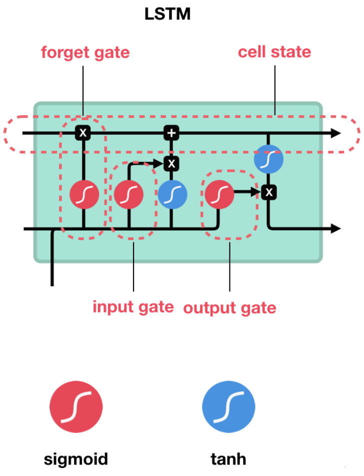

Fig. 6. A graphical representation of LSTM memory cells.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The Precision is the proportion of the predicted Churners that were correct. The precision is intuitively the ability of the classifier not to label as positive a negative sample.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

F-Measure / F-beta score weighs Recall more than Precision by a factor of β. When β equals 1.0, Recall and Precision are equally important.

$$F_\beta = (1 + \beta)\frac{\text{Precision} \cdot \text{Recall}}{\beta \cdot \text{Precision} + \text{Recall}} \tag{5}$$

In this context, the goal is to get as many TN as possible without spamming a lot of customers. The TP is necessary but not the priority. Which leads us to use the F-beta score metric, with β equal to 0.5.

VI. Hyperparameter

In this section, the hyperparameters [27] configurations used will be presented along with the reasons for these choices.

The MLP was $200 \times 200 \times 1$ in size, with a dropout of 0.2 separating each layer and a Dense layer of size 1 at the output. A range of (10^{-10} to 10^{-2}) was chosen to vary the learning rate, and the model learned better with a learning rate of 10^{-3}.

The XGBoost had 50 estimators, with a logistic regression for binary classification as objective function. A range of (5 to 50) was chosen to vary the maximum depth, and the model learned better with a maximum depth of 10.

For the LSTM model, there were 3 stacked layers of LSTM, with a dropout of 0.2 separating each layer and a Sigmoïd activator at the output. The model learned better with a learning rate of 10^{-7}.

VII. Results and discussion

A. Cross Validation

Cross-validation [4] is frequently used in the evaluation of regression and classification models. Applying it to the time-series or other naturally ordered data adds some complexity because of the chronology of events. Two techniques can be considered.

The first one is to define a time interval, where data should be retrieved for all customers, then apply the k-fold technique [25]. To do so, the dataset must be divided into k equal parts, called folds, where $k-1$ folds will be the training set and the remaining fold will be the test or validation set. Repeat it k times with different remaining fold as test set each time. The final score averages the validation results at the end of each iterations. In this context, several k were tested, the best result was obtained with k = 4.

The second technique consists in choosing 2 time intervals where to extract the data. There will be two subsets then, and each of them will be divided in two. Thus, four subsets are obtained. On these subsets the k-fold technique can be used. It will therefore be called 4-fold.

B. Discussion

The dataset which includes time-series with some additional information for each customer, was used to evaluate the performance of the tested classifiers. As previously exposed, it contains a total of 62,852 samples.

1) Linear Regression: It can be seen in tables II and III that the results of the linear regression are unsatisfactory, even if there are some individuals for whom the churn is detectable by simple linear regression as mentioned in section II-A-1,

it is unfortunately not the case for the majority. It is worth noting that the predictions are slightly better when the model is given a longer period at input. Table II shows results with $P1 = 8$ and Table III shows results with $P1 = 30$. With a longer period at input, the accuracy of the results suddenly becomes sensitive to the variation of the threshold. A period of 8 weeks was chosen at the beginning as input, to match the requirement of the marketing service. But, by varying the parameters in a naive way, and by looking for a better accuracy in the linear regression's predictions, parameters converged to a 30 weeks period and a threshold of 0.4. With these parameters we obtain a precision of 67%.

2) MLP: The results in table IV show that using 200 neurons in the hidden layer captured the slope information including noise, necessary for a good classification. Beyond 200 neurons, there is only a slowing down during training without a real improvement of the predictions. The model then reaches its best prediction for 10 epochs of training with an average precision of 73.30% and an average F-measure (β=0.5) of 72.21%, and starts to overfit beyond that.

3) XGBoost: The results in V show that using 50 estimators with a logistic regression for binary classification as objective function,

the XGBoost model was able to capture the slope information with the noise slightly better than the MLP. By varying

TABLE II
PRECISION AND F-MEASURE FOR THE LINEAR REGRESSION MODEL WITH $P1 = \mathbf{8}$ WEEKS PREDICTIONS. (HIGHEST VALUES IN BOLD, THE COUPLE (PRECISION/F-MEASURE) WITH THE LARGEST VALUES IS ALSO IN BOLD)

Threshold	Precision (%)	F-measure β=0.5 (%)
-1.0	56.09	55.27
-0.8	55.72	54.50
-0.6	55.43	53.85
-0.4	56.04	54.05
-0.2	55.97	53.47
-0.0	62.34	**57.52**
0.2	63.07	57.51
0.4	63.36	57.18
0.6	**63.78**	**57.05**
0.8	63.44	56.34
1.0	63.52	55.87

TABLE III
PRECISION AND F-MEASURE FOR THE LINEAR REGRESSION MODEL WITH $P1 = \mathbf{30}$ WEEKS PREDICTIONS. (HIGHEST VALUES IN BOLD, THE COUPLE (PRECISION/F-MEASURE) WITH THE LARGEST VALUES IS ALSO IN BOLD)

Threshold	Precision (%)	F-measure β=0.5 (%)
-1.0	64.19	**65.27**
-0.8	64.70	64.87
-0.6	65.94	64.61
-0.4	**67.19**	**64.13**
-0.2	67.52	62.40
-0.0	66.66	58.90
0.1	67.40	58.24
0.4	**67.84**	54.02
0.6	67.47	50.62
0.8	65.48	44.93
1.0	66.51	42.60

the maximum depth and the number of estimators, it was observed that beyond 70 estimators and a maximum depth of 15 there was only a decrease in the prediction accuracy. The best prediction (precision = 73.63%, F-measure = 74.45%) was achieved with 50 estimators and a maximum depth of 10.

4) LSTM: It was determined that the LSTMs outperformed the other detection approaches in this study. Precisely because LSTM was designed to be able to learn order dependence in sequence prediction problems such as analyzing customer sales data over time to detect potential churners. The best prediction (Precision = 73.70%, F-measure = 75.60%) was achieved with 3 stacked layers of LSTM.

TABLE IV
PRECISION AND F-MEASURE (AVERAGES) FOR MLP ($200 \times 200 \times 1$) PREDICTIONS. (HIGHEST VALUES IN BOLD)

Epochs	Resampling	P1 length	Precision (%)	F-measure β=0.5 (%)
10	Yes	8	70.01	70.60
10	Yes	30	**73.30**	**72.21**
10	No	8	50.04	55.60
10	No	30	52.40	56.51
50	Yes	8	68.80	**70.40**
50	Yes	30	**69.67**	69.38
50	No	8	50.02	55.60
50	No	30	52.20	57.70

TABLE V
PRECISION AND F-MEASURE (AVERAGES) FOR EXTREME GRADIENT BOOSTING (50 ESTIMATORS AND A MAXIMUM DEPTH OF 10) PREDICTIONS. (HIGHEST VALUES IN BOLD)

Epochs	Resampling	P1 length	Precision (%)	F-measure β=0.5 (%)
10	Yes	8	71.14	71.22
10	Yes	30	**73.63**	**74.45**
10	No	8	52.01	54.77
10	No	30	53.73	56.92
50	Yes	8	69.25	70.68
50	Yes	30	**71.89**	**70.71**
50	No	8	51.31	55.78
50	No	30	52.66	58.07

TABLE VI
PRECISION AND F-MEASURE (AVERAGES) FOR LSTM ($100 \times 50 \times 25$) PREDICTIONS. (HIGHEST VALUES IN BOLD)

Epochs	Resampling	P1 length	Precision (%)	F-measure β=0.5 (%)
10	Yes	8	70.86	70.92
10	Yes	30	**73.70**	**75.60**
10	No	8	51.87	55.98
10	No	30	52.78	57.32
50	Yes	8	69.17	70.42
50	Yes	30	**72.31**	**71.28**
50	No	8	52.14	55.91
50	No	30	53.41	58.53

VIII. CONCLUSION

This paper presents the application of machine learning models on customer sales data for churn prediction. A total of 4 statistical and machine-learning models have been compared in this study, focusing on predictive performance. With the reputation that precedes machine learning models, it is easy to guess that they might outperform conventional approaches, but this study explicitly shows the gap in section VII-B.

Companies can specify the goal of future retention marketing campaigns by using the prediction model proposed in this study.

Future studies may use other methodologies such as Transformers [28]. Since recent machine learning models have proven to be better at handling more data, the exploitation of external data such as weather, neighborhood details, average per capita income, will be of major interest.

ACKNOWLEDGEMENTS

This work was done as a part of a CIFRE (N 2019/1139) project with Colruyt France, funded by the Ministry of Higher Education and Research of France, managed by the Association Nationale de la Recherche et de la Technologie (ANRT) and was partially supported by the EIPHI Graduate School (contract ”ANR-17-EURE-0002”).

REFERENCES

[1] O. O. Aalen. A linear regression model for the analysis of life times. *Statistics in medicine*, 8(8):907–925, 1989.

[2] S. Abiteboul. Querying semi-structured data. In *International Conference on Database Theory*, pages 1–18. Springer, 1997.

[3] A. K. Ahmad, A. Jafar, and K. Aljoumaa. Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6(1):1–24, 2019.

[4] C. Bergmeir and J. M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

[5] M. Braun and D. A. Schweidel. Modeling customer lifetimes with multiple causes of churn. *Marketing Science*, 30(5):881–902, 2011.

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

[7] J. Cao, X. Yu, and Z. Zhang. Integrating owa and data mining for analyzing customers churn in e-commerce. *Journal of Systems Science and Complexity*, 28(2):381–392, 2015.

[8] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[9] P. S. Cowpertwait and A. V. Metcalfe. *Introductory time series with R*. Springer Science & Business Media, 2009.

[10] A. Dingli, V. Marmara, and N. S. Fournier. Comparison of deep learning algorithms to predict customer churn within a local retail industry. *International journal of machine learning and computing*, 7(5):128–132, 2017.

[11] W. W. Eckerson. Predictive analytics. *Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report*, 1:1–36, 2007.

[12] A. Estabrooks, T. Jo, and N. Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20(1):18–36, 2004.

[13] J. Gonzalez and W. Yu. Non-linear system modeling using lstm neural networks. *IFAC-PapersOnLine*, 51(13):485–489, 2018.

[14] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[15] R. A. Hoban. Introducing the slope concept. *International Journal of Mathematical Education in Science and Technology*, pages 1–17, 2020.

[16] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] S. M. Keaveney and M. Parthasarathy. Customer switching behavior in online services: An exploratory study of the role of selected attitudinal, behavioral, and demographic factors. *Journal of the academy of marketing science*, 29(4):374–390, 2001.

[18] J. Melton. Database language sql. In *Handbook on Architectures of Information Systems*, pages 105–132. Springer, 1998.

[19] V. L. Miguéis, D. Van den Poel, A. S. Camanho, and J. F. e Cunha. Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert systems with applications*, 39(12):11250–11256, 2012.

[20] T. Mutanen, S. Nousiainen, and J. Ahola. Customer churn prediction–a case study in retail banking. In *Data Mining for Business Applications*, pages 77–83. IOS Press, 2010.

[21] R. H. Myers, D. C. Montgomery, G. G. Vining, and T. J. Robinson. *Generalized linear models: with applications in engineering and the sciences*, volume 791. John Wiley & Sons, 2012.

[22] D. M. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2011.

[23] A. K. Rai and M. Srivastava. Customer loyalty attributes: A perspective. *NMIMS management review*, 22(2):49–76, 2012.

[24] H. Ramchoun, M. A. J. Idrissi, Y. Ghanou, and M. Ettaouil. Multilayer perceptron: Architecture optimization and training. *IJIMAI*, 4(1):26–30, 2016.

[25] J. D. Rodriguez, A. Perez, and J. A. Lozano. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):569–575, 2009.

[26] F. Rosenblatt. *The perceptron: a theory of statistical separability in cognitive systems (Project Para)*. Cornell Aeronautical Laboratory, 1958.

[27] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning. Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data. *arXiv preprint arXiv:1803.11266*, 2018.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.