ISSN- 2394-5125 VOL 10, ISSUE 02, 2023

MAKING USE OF MACHINE LEARNING FOR CHURN PREDICTION IN RETAIL INDUSTRY

Dr. Munaga Ramakrishna Mohan Rao, MBA, Ph.D, PGDCA, LMISTE, IC-38
Professor & Director, P V Ram Reddy PG College, Hyderabad

Abstract- Every business has one overarching goal: to maximise sales and profits. When a company's regular clientele suddenly stops buying from it, it usually sees a precipitous decline in revenue. It has been established that keeping existing customers is less costly than finding new ones, making it a top priority in Customer Relationship Management, particularly in the retail sector. When a client quits patronising a store, they no longer provide an opportunity for more purchases or even cross-selling. As a result, businesses need to take preventative measures by identifying at-risk customers so they can be retained. This article demonstrates the utility of combining transaction data with machine learning for churn prediction in the retail sector. A total of 5,115,472 customer loyalty card records were pulled from a European retailer's data warehouse and used to train the machine learning models. According to the findings, machine learning models outperform their linear regression counterparts.

Keywords— Attrition Prediction, Churn rate, Retail Industry, Machine Learning

I. INTRODUCTION

All industries struggle with customer retention at some point. Keeping existing customers happy is cheaper than finding new ones [23], thus it's crucial for businesses to learn from the actions of their consumers who leave. There are a number of factors that might lead to a customer leaving. The rival may provide identical items at lower costs, provide superior Customer Service, or provide a more pleasant Online Shopping Experience [5]. According to studies, the cost of acquiring a new customer is higher than keeping an existing one [7]. Therefore, it is crucial to maintain the current clientele. Since customer turnover is often a slow and steady decline, rather than a rapid spike, it is possible that organisations might benefit by studying their customers' past purchase histories [6]. By using loyalty cards, businesses can keep track of countless customer details. Data like this contain valuable insight that is frequently lost in the sea of raw data. These datasets mostly consist of semi-structured [2] data like Excel, JSON, and CSV files, as well as structured [18] data that can be queried using SQL. These days, most people agree that Machine Learning effectively gleans hidden features from unprocessed data. In light of the widespread success of using Machine Learning techniques in a wide variety of contexts, this might be a viable strategy for mining unstructured datasets for insights. Colruyt France, a French grocery chain with 90 stores, mostly in the Franche-Comt'e area, generously provided an anonymized dataset for our analysis. With 105,488 consumers included, this dataset offers a representation of in-store purchases made by customers. There are now 5,115,472 rows of data from these clients. Using Machine Learning and Deep Learning on this kind of data is what makes this research

ISSN- 2394-5125 VOL 10, ISSUE 02, 2023

innovative. In addition to the data augmentation approach described in Section III-C, another significant contribution is the ability to link individual-level data (such as age, duration of client relationship, gender, and city population) to sales timeseries [9]. There is a best-case scenario where the model's accuracy is 75.60 percent. Here is how the rest of the paper is laid out. The background is provided in Section II, which discusses the nature of churn and the many challenges that may arise. Methodology, from data collection to analysis techniques, are detailed in Section III. Section IV describes the various models used, while Section VII summarises the findings and evaluates the various methods employed to identify at-risk customers. The paper's findings are summed up in Section VIII.

II. RELATEDWORKS

In business parlance, a "churned" client is one who has defected to a rival or who has ceased making purchases altogether. In this context, "churn" refers to the percentage of a company's current client base that is expected to cease doing business with the company within a certain time frame [7]. Another definition [19] is when a customer's typical purchase quantity goes below a certain level over a certain time frame. Equation 1 provides a formal definition of the typical shopping cart. Assume n = mod(P), where P is the total number of weeks in the period, and let PurchaseAmountC(i) represent the amount spent by customer C during week i.

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

It is difficult to predict exactly when a client would churn from a store. Not all of a store's customers will defect at once, but rather they will gradually quit making purchases there. In reality, they progressively move to a competitor [6]. For the sake of accuracy, let's use the following rule to define a churner in this setting. Let P1 be the time frame during which the typical spending habits of the consumer are studied. Future churn detection models will use it as input.

As soon as P1 concludes, we enter P2, the assessment phase, during which any shift in consumer purchasing behaviour becomes evident. The prediction model will be blind to P2. P2 is held constant at 12 throughout this research to accommodate the needs of the advertising service. That implies that there is always a 3-month window for feedback. This is the method that will be used for the labels. The consumer is considered at risk of churning if their average spending in P2 is less than 20% of their spending in P1, and is not considered at risk of churning if it is more than or equal to 20%. This factor, which is 20%, is the maximum allowable decrease. This indicates that the churn may be partial or complete. Let denote the discounting factor, and let C represent a customer, then you have the formal definition of churn. Assuming the following conditions, C is a churner.

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

ISSN- 2394-5125 VOL 10, ISSUE 02, 2023

Figure 1 shows some instances of churners, whereas Figure 2 shows some examples of non-churners.

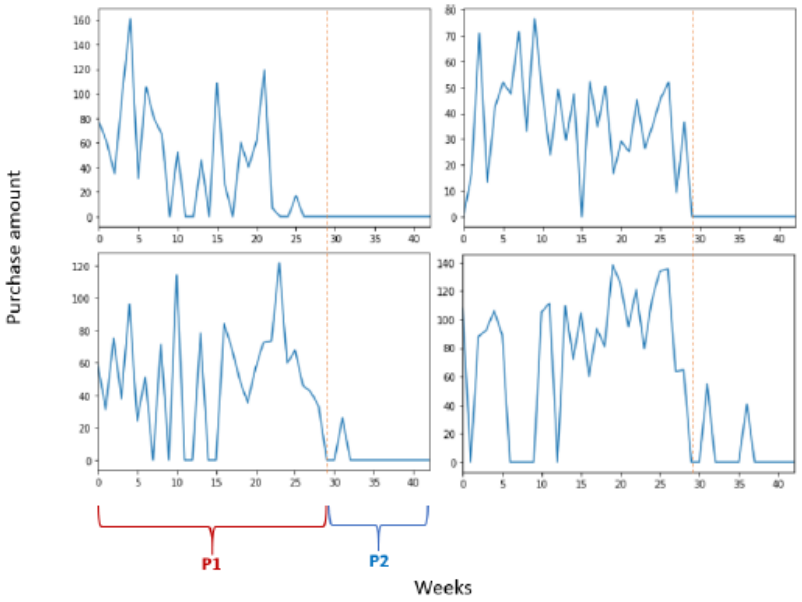

Figure 1: Four typical churners. They were regular purchasers during P1, but during P2 they have cut down to weekly or even less frequent purchases. We may infer three churn scenarios from these definitions and findings. The Most Common Types of Churn As can be seen in Figure 3, some instances include negative slopes [15] with smaller sounds. Assume a time series with a slope, denoted by

$$\hat{y}_i = \hat{S}x_i + c$$

The trend is not skewed by random fluctuations. A simple linear regression with a threshold may be used to identify these types of churners. Exceptions to churn if important factors are concealed: In these situations, the consumer's routine is influenced by factors outside their control, such as sales, soccer championships, weather, or any other external event. These confounding factors considerably affect the trend, making a linear regression model's categorization skewed. If there is a hidden variable in the data, the suggested model must be able to read that information from the input. This secret factor may also be represented as a time series. Case churning differs from case spotting in the following ways: Clients that only make infrequent purchases at our establishments make up anything from 20% to 30% of the total customers recorded in our sales databases. Periodic clients, on the other hand, are individuals who only visit every so often, such as once a week or once every two weeks. Customers who frequent our shops often perceive us to be their "go-to" spot, and as a result, they tend to make major purchases there. Even if they can make impulse buys at a competitor's shop that's closer to their home or more conveniently located. Churners are once-regular clients who suddenly become infrequent buyers. The occasional instances are not the main focus of our research.

ISSN- 2394-5125 VOL 10, ISSUE 02, 2023

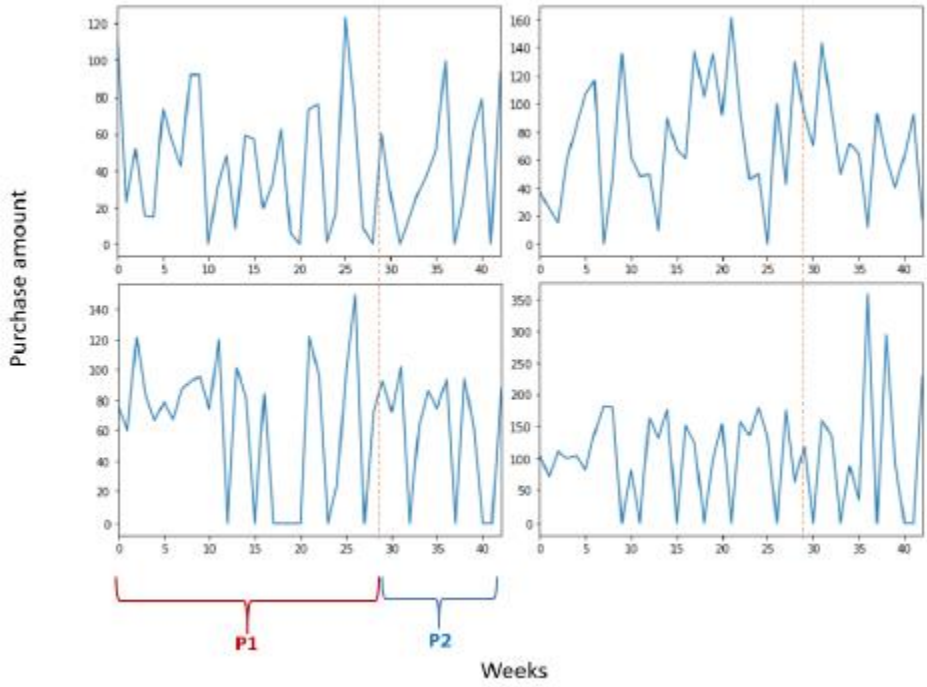

Figure 2 depicts four instances that do not churn. During P1 and P2, they used to shop weekly. Consumer behaviour did not significantly change.

There has been a lot of study on churn prediction during the last several years. According to the study's authors, keeping existing customers is a lot cheaper in the long run. Numerous studies have been conducted on churn, comparing machine learning and traditional methods on data from a variety of industries, including telecommunications [3], banking [20], and online subscriptions [17]. However, there is a dearth of research that focuses on the retail sector, which presents a unique set of challenges due to the nature of its contacts with and lifespans of its customers. Dingli [10] compared the RBM model to the CNN model in 2017. An F-measure (β=0.5) of 77% and a Precision of 74% were the best results that could have been obtained at that time using RBM. Convolutional networks and deep networks have both progressed since then. One strategy may be favoured over another depending on the circumstances and the kind of data being analysed. In the telecommunications industry [3], for instance, customers seldom sign up with several providers at once. When shopping, a customer may have a favourite store where they do most of their shopping and a few others where they do less of their shopping. The unique characteristics of the retail industry are the subject of this research. More importantly, no other studies have used the data augmentation method presented in this one.

ISSN- 2394-5125 VOL 10, ISSUE 02, 2023

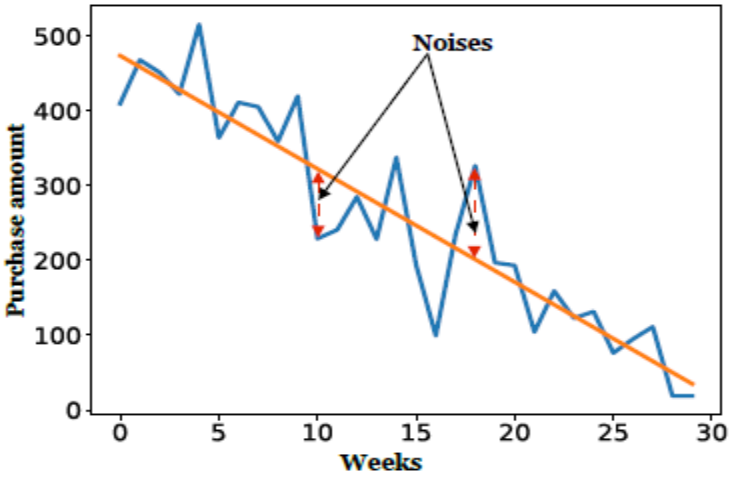

Figure 3 shows a churner with reduced noise, where the slope can be seen out.

III. PROPOSED SYSTEM ARCHITECTURE

A total of 5,115,472 rows of sales information from 105,488 clients at a French grocery chain were used in this study. All of the data used in this study has been pseudonymized, so it is possible to link individual characteristics (such as customer age, relationship duration, gender, and city population) to sales data [9]. Each set of rows represents a different client, and the whole thing is laid up as a 2D array. Customers with duplicate profiles with the same name and address, for example, were eliminated. Our company's Marketing department was consulted over the length of the P1 and P2 periods II-A, which was determined to be two and a half weeks respectively. A variety of lengths were tried out before arriving at the optimal one in terms of precision. The labelling method was used after (the reduction factor) was established at 0.2. A separate train and test set were created from the original dataset. Table I provides a summary of the information collected.

TABLE I. we can see a general breakdown of the information gathered. In This Study, We Didn't Even Try To Recognize The Importance Of Spot Customers.

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

Various forms of information provide unique details. Data mining techniques may struggle if the dataset exhibits any of these features. The class imbalance in the dataset employed in this research was a major limitation. Learning algorithms often only make predictions for the majority class when there is a disparity between them, since this helps to keep the prediction error to a minimum. For instance, out of 60,000 total consumers in the dataset, only 2,000 were

ISSN- 2394-5125 VOL 10, ISSUE 02, 2023

churners. There is a clear socioeconomic disparity between the two groups, with 96% of active customers compared to just 4% of churners. Re-sampling is a common method used to correct for this kind of class discrimination. There are two options for accomplishing this goal: over-sampling and under-sampling [12]. Through under-sampling, we only employ a small fraction of the whole majority class to educate our models. In this research, we balanced the amount of churners and non-churners by randomly excluding inputs from the set of current customers. This method involves adjusting the scale of an existing time series in order to generate new time series that include more samples of a certain class. A client who spends €100 per week may be used to create a fictitious one who spends €50 per week and still another who spends €200 per week. The second and third will operate in the same way as the first. The plan is to produce new clients who act just like the original ones, except on a larger scale. Figure 4 shows the original and the scaled-down version.

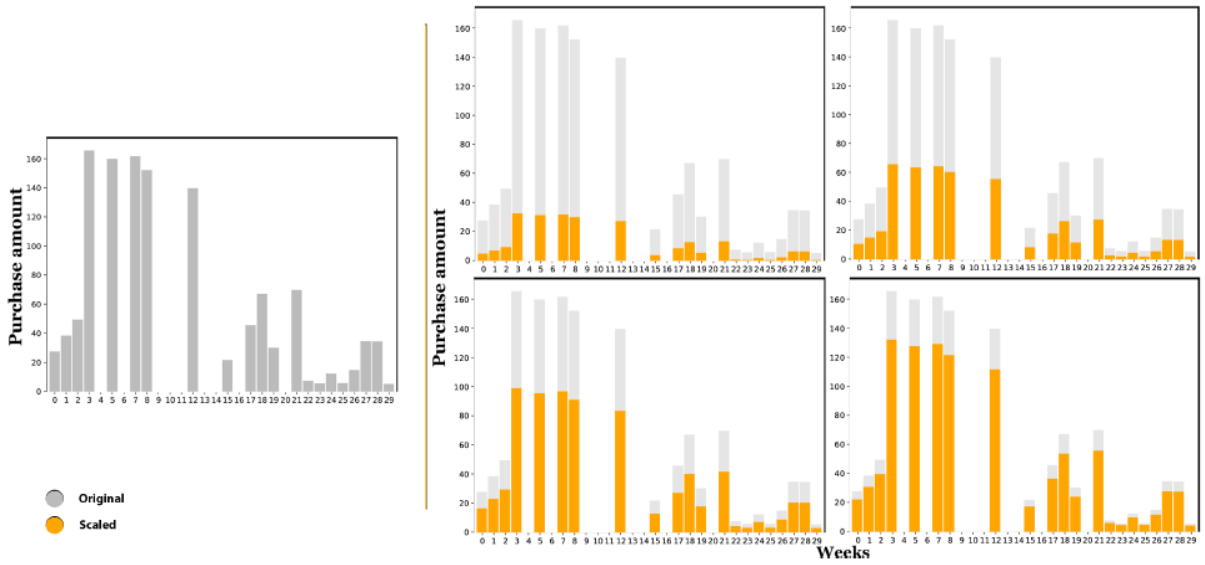

Fig.4 Scaled versions of the original data and data enhancements. Simplified versions are orange while the original series are shown in grey. In this case, for ease of reading, each scaled version is made to be less than one times the size of the original. Be aware that this may be multiplied by an additional 1.5 - 2 times.

IV. METHODOLOGY EMPLOYED

Below, we evaluate the effectiveness of a linear regression model against other machine learning methods that have been shown to be useful for churn prediction. During the comparing, both efficiency and dependability were taken into account. To describe the association between a continuous answer and one or more explanatory factors (also known as dependent and independent variables), statisticians use a technique called linear regression [1]. Linear regression is a simple statistical approach for learning about consumer behaviour, the nature of your firm, and the variables that affect your profitability [21]. Linear regression is only useful when the dependent variable is continuous, limiting its practicality in the business sector, but it is

ISSN- 2394-5125 VOL 10, ISSUE 02, 2023

nevertheless a well-known approach when it is applicable [11]. When compared to Rosenblatt's initial Perceptron model from 1950 [26], a Multi-Layer Perceptron (MLP) is a more advanced and flexible model. With the advent of Deep learning, MLP has been successfully used to several domains [14]. Between its input and output layers, it conceals one or more intermediate ones. As can be seen in Figure 5, the neurons have a layered organisation, with no interconnections between neurons in the same layer [24].

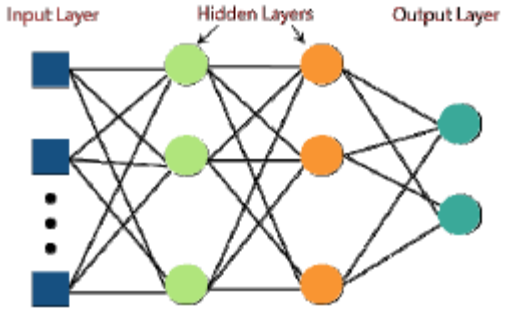

Fig.5 MLP

As a free and open-source package, Extreme Gradient Boosting (XGBoost) offers a powerful and effective implementation of the gradient boosting method [8]. In recent years, XGBoost has emerged as the dominant strategy for solving classification and regression issues in machine learning competitions, frequently serving as a decisive factor in the overall winning solution. Due to the categorization issue inherent in this research, it was judged instructive to put it through its paces so that its findings could be compared to those obtained by other methods.

When it comes to sequence prediction challenges, recurrent neural networks (RNNs) like Long Short-Term Memory (LSTM) networks may learn order dependencies [16]. However, although LSTM models were first developed in the late 1990s, they have only lately emerged as a viable and strong forecasting approach for timeseries. Short-term memory, or the inability to learn dependencies from extended sequences, is a major flaw of recurrent neural networks that may be fixed using an LSTM. An LSTM may remember, forget, or disregard data points depending on a probabilistic model by using a sequence of 'gates' [13] each with its own RNN. For an example of a long short-term memory (LSTM), see Figure 6.

ISSN- 2394-5125 VOL 10, ISSUE 02, 2023

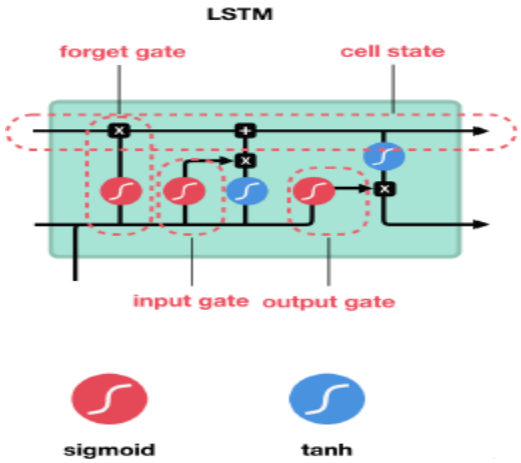

Fig.6 A pictorial illustration of long short-term memory cells.

V. PERFORMANCE METRICS

In this study, we evaluate the accuracy of several prediction models by using precision, recall, and the F-measure [22]. True positive (TP) and false positive (FP) samples are distinguished from True Negative (TN) and False Negative (FN) ones. How many customers who were assumed to be Churn were really Churn is represented by the Recall. Simply said, recall measures how well a classifier can track down all of the test samples it was trained on.

VI. RESULTS

If you want to test out a new regression or classification model, cross-validation [4] is a common tool to utilise. Some additional complexity is introduced when it is applied to timeseries or other naturally ordered data due to the sequential nature of the occurrences. It's possible to choose between two approaches. One method involves setting a time frame within which all customers' information has to be gathered using the k-fold method [25]. If you want to train and test your model, you'll need to split your dataset into k equal pieces, or folds, with k1 serving as your training set and k serving as your test or validation set. It should be repeated k times, with each subsequent fold serving as a new test set. The ultimate score is calculated by averaging the results of each round of validation. When trying out several values for k, it was found that k = 4 gave the greatest results. The second strategy is to divide the time frame into two halves before extracting the data. Then there will be four subsets total, two from each of the two initial sets. As a result, we have four distinct groups. The k-fold method may be used to these subsets. We'll just refer to it as "four-fold" from now on.

VII. DISCUSSION

All of the tested classifiers were put through their paces using a dataset consisting of time-series plus some extra information for each client. It has 62,852 samples, as was previously disclosed.

ISSN- 2394-5125 VOL 10, ISSUE 02, 2023

Linear Regression: Even if there are certain people for whom the churn is visible by simple linear regression, the results of the linear regression are inadequate, as shown in tables II and III.

MLP: Table IV shows that 200 neurons in the hidden layer were sufficient to collect the noise and slope information required for accurate categorization. If you train with more than 200 neurons, you won't see any substantial progress in your predictions. Thereafter, the model begins to overfit after reaching its best prediction after 10 epochs of training with an average accuracy of 73:30% and an average F-measure (=0.5) of 72:21%.

XGBoost: According to the data in Table V, the XGBoost model outperformed the MLP by a little margin when capturing slope information with noise using 50 estimators and a logistic regression for binary classification as objective function.

LSTM: Researchers found that Long Short-Term Memory (LSTM) models performed better than other methods of detection. Precisely because \sLSTM was built to be able to learn order dependence \sin sequence prediction tasks such as monitoring customer \ssales data over time to spot possible churners. In terms of prediction accuracy, three layers of LSTM proved optimal (Precision = 73:70%, F-measure = 75:60%). It was given in table VI.

TABLE II Predictions From The Linear Regression Model With P1 = 8 Weeks: Precision And F-Measure. (Greatest values are italicised, and the largest value pair (precision/F-measure) is highlighted)

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. The accuracy and F-measure of linear regression forecasts over a horizon of 30 weeks at P1. (Greatest values are italicised, and the largest value pair (precision/F-measure) is highlighted)

ISSN- 2394-5125 VOL 10, ISSUE 02, 2023

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

TABLE IV. F-Measure (Averages) And Precision (Mean) For Mlp (200, 200, 1) Estimates. (Top Values Italized 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. Extreme Gradient Boosting Prediction Accuracy And F-Measure (Averages) With 50 Estimators And A Maximum Depth Of 10. (Top Values Italized In Bold)

ISSN- 2394-5125 VOL 10, ISSUE 02, 2023

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. Prediction Accuracy And F-Measure (Average) Using Lstm (100, 50, 25). (Top Values Italized 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. FUTURE SCOPE AND CONCLUSION

The purpose of this study is to demonstrate the use of machine learning models on sales data in order to anticipate client attrition. In this investigation, we assess the prediction performance of four different statistical and machine-learning models. The hype around machine learning models suggests that they may be superior to more traditional methods. This study's recommended prediction model may be used by businesses to set clear objectives for future retention marketing initiatives. In the future, researchers may use other methods, such as the Transformers [28]. The use of external data like weather, neighbourhood characteristics, and average per capita income will be of great interest since newer machine learning models have proved to be better at handling additional data.

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. Ben´ıtez. 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. Migu´eis, 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.

ISSN- 2394-5125 VOL 10, ISSUE 02, 2023

[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 machinelearning 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.