PalArch's Journal of Archaeology of Egypt / Egyptology

CHURN PREDICTION OF CUSTOMER IN TELECOMMUNICATION AND E-COMMERCE INDUSTRY USING MACHINE LEARNING

Ms Kavita[1], Neha Sharma[2], Gaurav Aggarwal[3]
[1]Manipal University Jaipur
[2]Manipal University Jaipur
[3]Manipal University Jaipur
Email: [1]Kavita.chaudhary@outlook.com, [2]Nehav.sharma@jaipur.manipal.edu,
[3]Gaurav.aggarwal@jaipur.manipal.edu

Ms Kavita[1], Neha Sharma[2], Gaurav Aggarwal[3], Churn Prediction Of Customer In Telecommunication And E-Commerce Industry Using Machine Learning– Palarch's Journal of Archaeology of Egypt/Egyptology 17(9) (2020), ISSN 1567-214X.

Keywords: Churn Prediction, Machine Learning, Telecommunication Churn, E-commerce Churn, Logistic Regression, Random Forest Classifier, Artificial neural network, Recurrent Neural Network.

Abstract:
Customer behaviour can be represented in many ways. The customer's behaviour is different in different situations will give his idea of customer behaviour. From a general perspective, the behaviour of the customer, or rather any person in any area, is taken to be random. When observed keenly, it is often seen that the future behaviour of a person can depend on various factors of the present situation as well as the behaviour in past situations. This research constitutes the prediction of customer churn, i.e. whether the customer will terminate purchasing from the buyer or not, which depends on various factors. We have worked on two types of customer data. First, that is dependent on the present factors which do not affect the past or future purchases. Second, a time series data which gives us an idea of how the future purchases can be related to the purchases in the past. Logistic Regression, Random Forest Classifier, Artificial neural network, and Recurrent Neural Network has been implemented to discover the correlations of the churn with various factors and classify the customer churn efficiently. The comparison of algorithms indicates that the results of Logistic Regression were slightly better for the first Dataset. The Recurrent Neural Network model, which was applied to the time-series dataset, also gave better results.

Introduction:

The branch of machine learning is a vast branch and has spread rapidly to all the fields in recent years. The idea of using real-time data to get near accurate results and hence, planning solutions for a given problem is widespread and is being promoted by every other person. Customer Behavior Modeling is defined as finding out correlations of the behaviour with various factors that represent the purchases of the customer and finding on the critical points on which the company has to work to satisfy its customers [1]. Customer behaviour, although

considered random, can be guessed or predicted, when keenly observed. If there is a way for a seller to know which of its customers are unhappy or are wanting to leave, they can be given special attention, and retention solutions can be designed. Though there are ways to apply various models to find out customer behaviours, it is still a challenging task because the behaviour is represented very differently in different settings. Also, each company has its way to deal with the customers and would want to follow its basic rules to satisfy the customers. So, make sure that the model meets the company's requirements can be a very irksome task. In the area of customer analytics, it is better to do predictive modelling rather than the passive guessing of customer behaviour. The Prediction Results allows marketers and retention experts to work on future behaviour rather than the passive educated guesses of the past. The advantage of this is that marketers can focus on specific customers with specific strategies. If we take specifics, if a company is able to predict which of its customers are most likely to churn, then they can approach those customers in a somewhat different way than they were planning earlier and hopefully develop strategies to retain them. This, thus, serves as potential revenue to the company. So, predicting customer churn can be considered as a classification problem, and we can apply various classification algorithms to predict the churn [2]. Prediction can be made by taking into account, the customers present purchases like what item has the customer been most interested in the past and what services would the customer wish to continue in the future. Also, a record of how consistent the customer has been in purchasing the services of the company can help in prediction.

Related Work:

The Customer churn prediction is very uncertain to predict; it has the pressure on commercial completion market. Zuhao et. Al. [3], implemented Support Vector Machine on the Dataset of wireless churn industry to predict the improved accuracy in comparison to traditional methods. The research used improved one-class support vector machine and compared the performances of the different kernel. The results presented the 87.15% accuracy, which was compared to ANN, Decision tree, and Naïve Bayes. Support Vector Machine outperforms the traditional approaches because it requires less amount of time for training and testing.

The prediction by Support Vector Machine is better in comparison to ANN, Decision tree, Logistic Regression, and naïve Bayes [4]. Xia and Jin implemented SVM on telecommunication data and compared with the several approaches. The study showed that the churn rate is more among the customers who have strong charge willingness, long mobile service, considerable customer care, and regular call and message usage. The Dataset of telecommunication and banking industry has numerous similarities which indicate that SVM can be applied to the banking sector also.

The Data sources for prediction of customers churn includes the history of the usages, bills, and customer services, but Lian Yan et. Al. [5], used Call Detail Record. The study uses the delayed Call Detail Record as primary to estimate customer behaviour. The Call Detail record data was used to extract the calling links and identifies several distance measures. The Data retrieved from the calling links was given as input to the neural network, and the acceptable accuracy is achieved. The Calling links can be used for marketing and offer in a specific community.

More features can be computed from the existing features to improve the prediction of churn customers. Huang et al. [6], computed new Dataset from the existing data, which includes call details, account and bill information, line information, payment information, complain and service information. The total

Dataset 827,124 customer's data was selected from the real-world database of Ireland. The testing and training dataset has 13,562 churners and 400,000 non-churners data where 738 features represent each customer. Seven methodologies were applied to predict the churn. The research provided comparison among the techniques, and comparison between extracted feature set and existing feature set. The new features dataset presented better results than existing Dataset.

Vafeiadis et al. [7] conducted the experiment using five methods of machine learning to predict customer churn. In the first segment of the research, the models were implemented on the Dataset collected from the UCI Machine Learning repository, public domain dataset, and further, the models were evaluated using cross-validation methods. In the second segment of research, the model performance was enhanced using boosting algorithms. To identify the most efficient feature combination, Monte Carlo Simulation was performed on every method. The results demonstrated that the boosted model results outperformed the simple (non-boosted) model results. SVM-POLY using AdaBoost performed better among all the models. Ahmad et al. [8] focused on churn in the telecommunication sector, where churn affects the revenue of the company. The study aimed to predict the churn of customers using machine learning techniques on big data. The authors used the Area Under Curve standard measure to measure the performance of the model, the value generated by the Area Under Curve is 93.3%. The model was developed and tested on large Dataset provided by SyriaTel Telecom Company, which contained the information of customers over nine months. XGBOOST algorithm performed better among four machine learning techniques. Routh et al. [9] proposed a model to overcome the uncertainties such as volatile behaviour of the customer and increasing completion risk. The model computes the possible risk and identifies the relationship between risk and customer behaviour. The model used data from the hospitality industry, and models give 20% improved accuracy in comparison to the existing conventional models. The results of the research help vendors to understand the reason for churning and generate new plans to deal with the possible upcoming churns.

Methods and materials:

The research was conducted on two different datasets. The first Dataset is the record of services purchased by customers of a fictional telecom company and the second Dataset is the record of purchase history of the customers of an e-commerce website. The Dataset is collected from kaggle.com [10] containing details of a telecommunication company. The columns in the Dataset include necessary credentials of the customers as well as the services purchased by each customer along with the churn variable (0 or 1), i.e. whether the customer will churn or not.

The second Dataset contains the online retail of an e-commerce website from the UCI Machine Learning Repository [11]. It has records of the various purchases made by the customers for one year from 1-12-10 to 9-12-11. The data is different from the first Dataset as it is the time-series data.

Methodology:

The research proposed models to predict whether a particular customer will churn or not. We have used various machine learning concepts and algorithms for prediction. The algorithms have also been used to test if the accuracy of classifying the churn customers is also maintained on the data on which the model is not trained.

LOGISTIC REGRESSION

Logistic Regression uses a function called logistic function at the core of the method. The logistic Regression uses the sigmoid function. It applies the function on the equation of linear Regression which gives us the results as the probability of an occurrence. The shape of the curve is S-like which can fit any real number value to a value between 0 and 1. It calculates the value by the following formula.

$$1 / (1 + e^{-value})$$

Where e is the base of the natural logarithms,
Moreover, value is the actual numerical value that is to be transformed.

RANDOM FOREST CLASSIFIER

Random forest algorithm is a supervised classification machine learning algorithm. It randomly creates the forest with several decision trees. Each tree creates different decisions and predicts values. In random forest classifier, a more significant number of trees tend to give more accurate results. Random forest algorithm manages the problem of missing values and does not overfit the model when we have more number of trees present. There are two stages in the random forest algorithm, one is random forest creation, and the other is to predict the random forest classifier created in the first stage. In random forest classifier, the value of the output variable is the majority value predicted by all the trees.

ARTIFICIAL NEURAL NETWORKS

Artificial neural network(ANN) is a machine learning algorithm which works on the idea of a human brain. In general, an ANN consists of a series of interconnected nodes that transform and the information is passed through a series of layers. An ANN has three primary layers, an input layer, a hidden layer and an output layer. Though according to the requirements, various modified versions of the models consisting of a large number of hidden layers are being used these days.

RECURRENT NEURAL NETWORKS

These are the neural networks which are applied for time series analysis. When we have data that depends on past behaviour, RNNs are used. They are different from the simple artificial neural networks in a way that they allow the flow of information from one cell to the next. This information is the representation of the input information of the previous layer.

The data is preprocessed in order to remove irrelevant information from the database. The categorical data is converted into the numeric data. The null values in the Dataset are taken care of, and the data is then scaled to make the units of various columns irrelevant. The input and the target output is specified, and data is spat into the training set and the test set. Figure 1 demonstrates the research process.

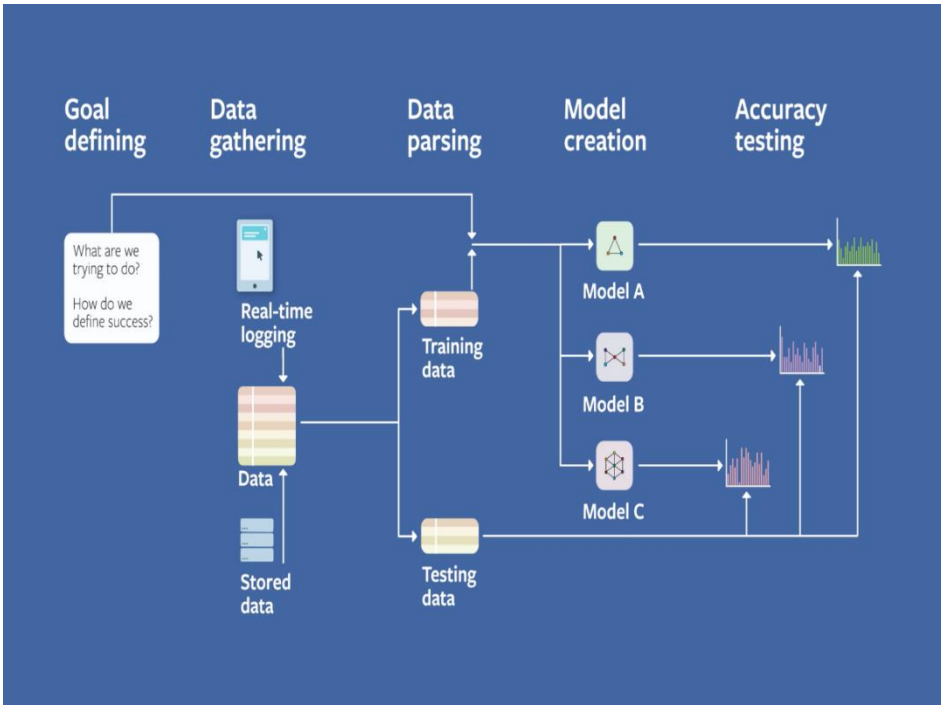

Figure 1: Flow chart indicating the research process

For the online retail/E-commerce dataset, the churn is calculated using the mean standard deviation of the customer and the individual deviation from the mean of a single purchase. If the individual deviation is greater than the mean standard deviation, the customer is said to have churned in between the two purchases and hence a row is appended in between making the value of the churn attribute 'true' and making the purchase items and unit price as 0. The input data is then made by adding all the previous purchases of the customer corresponding to its latest purchase in the target variable, and then the data is split into the training set and the test set. Various graphs describing the correlations in the data are constructed to give us a better understanding of what we are dealing with.

Result:
For the Telecommunications dataset, Logistic Regression, Random Forest and Artificial Neural Network models are applied, and an attempt to modify the basic versions of models is made to achieve the maximum accuracy possible. Figure 2 represents the breakdown of the churn of telecommunication Dataset.

For the E-Commerce dataset, Recurrent Neural Network model is applied using different activation functions under the concept of Long Short Term Memory. After applying 10-fold cross-validation and grid search for specific parameters, we calculate the results for the maximum accuracy achieved. Figure 3 represents the breakdown of the churn of E-commerce dataset.

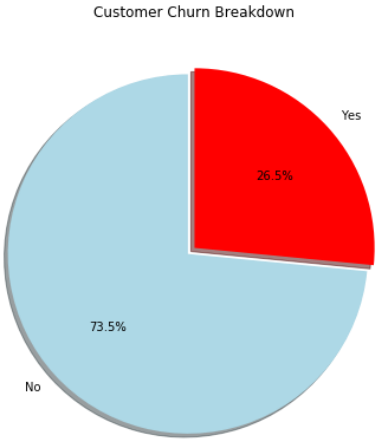

Figure 2: Churn Breakdown of Dataset1

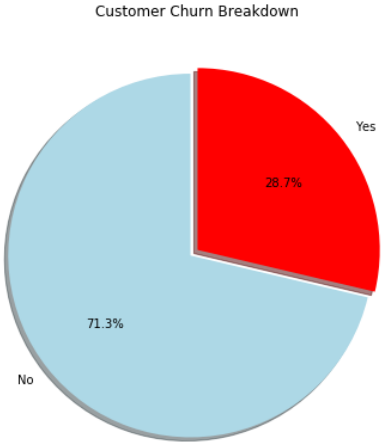

Figure 3: Churn Breakdown of Dataset2

The Confusion matrix has been generated to represent the result. Figure 4, figure 5, and figure 6 demonstrate the confusion metrics of Logistic Regression, Random Forest and Artificial neural network, respectively. Table 3 contains the combined confusion metrics data for all the machine learning techniques.

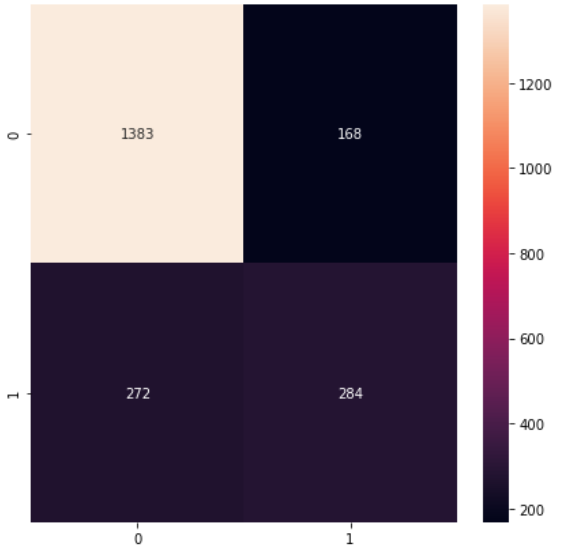

Figure 4: Confusion Metrics of Logistic Regression

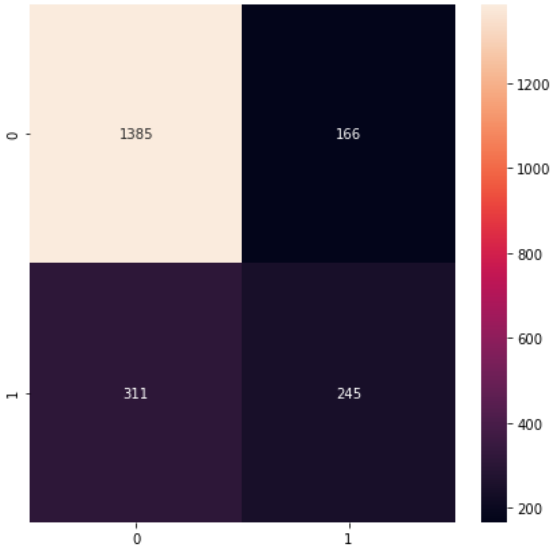

Figure 5: Confusion Metrics Random Forest

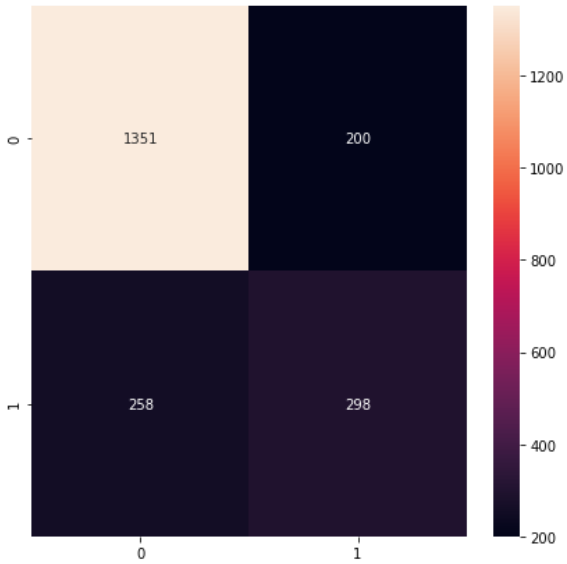

Figure 6: confusion Metrics of Recurrent Neural Network

Table 3: Confusing Metrics of Logistic Regression, Random Forest and artificial Neural Network.

		Predicted Class		
		0	**1**	**Methodology**
Actual Class	**0**	**1383**	**168**	**Logistic Regression**
	1	**272**	**284**	
	0	**1385**	**166**	**Random Forest**
	1	**311**	**245**	
	0	**1351**	**200**	**Artificial Neural Network**
	1	**258**	**298**	

For the E-Commerce dataset, Recurrent Neural Network model is applied using different activation functions under the concept of Long Short Term Memory. After applying 10-fold cross-validation and grid search for specific parameters, we calculate the results for the maximum accuracy achieved. Figure 7

represents the correlation between customer ID and invoiceNo. Figure 8 and Table 4 present the confusion matrix of the recurrent neural network.

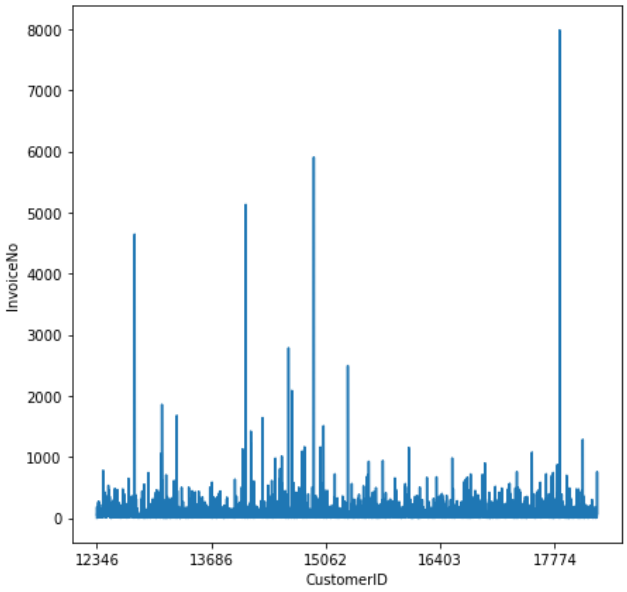

Figure 7: Correlation between Invoice and Customer ID

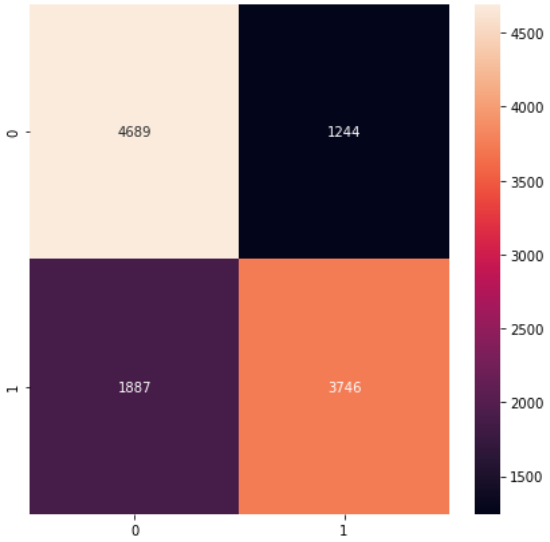

Figure 8: Confusion Metrics of Recurrent Neural network

Table 4: Confusion Metrics of Recurrent neural Network

RNN	0(Predicted)	1(Predicted)
0(Actual)	4689	1244
1(Actual)	1887	3746

Conclusion:

The research introduces models to predict whether a particular customer will churn or not. The first Dataset is the record of services purchased by customers of a fictional telecom company and the second Dataset is the record of purchase

history of the customers of an e-commerce website. Various Machine Learning machine learning concepts and algorithms are applied for prediction and have also been used to test if the accuracy of classifying the churn customers is also maintained on the data on which the model is not trained.

On comparing the results and analyzing, it is observed that the results of Logistic Regression are slightly better than other classification algorithms in this scenario. The results indicate that the customers who purchased expensive services and/or were senior citizens are the ones who are most likely to churn. The results also depict that the future purchase of the customer strongly depends on the past purchases, for customers purchasing more items in one purchase made their next purchase after a relatively long time. Also, the customers who have a consistent time difference in two consecutive purchases are most likely to stay.

Based on these results we can conclude that different companies can apply this model to their customer data to figure out who is most likely to churn and work on designing the retention solutions so as not to suffer losses.

Various other models can be built and tried, especially the deep learning models with a different combination of layers and activation functions. Calculating the churn rate is helpful, but if possible, the next most possible purchase of a customer can be predicted so that the company can identify the customer's interest quickly and pitch those products to the customer, in which he/she is highly interested.

References:

[1] S. B. Borah, S. Prakhya, and A. Sharma, "Leveraging service recovery strategies to reduce customer churn in an emerging market," *J. of the Acad. Mark. Sci.*, vol. 48, no. 5, pp. 848–868, Sep. 2020, doi: 10.1007/s11747-019-00634-0.

[2] M. Panjasuchat and Y. Limpiyakorn, "Applying Reinforcement Learning for Customer Churn Prediction," *J. Phys.: Conf. Ser.*, vol. 1619, p. 012016, Aug. 2020, doi: 10.1088/1742-6596/1619/1/012016.

[3] Y. Zhao, B. Li, X. Li, W. Liu, and S. Ren, "Customer Churn Prediction Using Improved One-Class Support Vector Machine," in *Advanced Data Mining and Applications*, vol. 3584, X. Li, S. Wang, and Z. Y. Dong, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 300–306.

[4] G. Xia and W. Jin, "Model of Customer Churn Prediction on Support Vector Machine," *Systems Engineering - Theory & Practice*, vol. 28, no. 1, pp. 71–77, Jan. 2008, doi: 10.1016/S1874-8651(09)60003-X.

[5] Lian Yan, M. Fassino, and P. Baldasare, "Predicting customer behavior via calling links," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, Montreal, Que., Canada, 2005, vol. 4, pp. 2555–2560, doi: 10.1109/IJCNN.2005.1556305.

[6] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1414–1425, Jan. 2012, doi: 10.1016/j.eswa.2011.08.024.

[7] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. Ch. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, Jun. 2015, doi: 10.1016/j.simpat.2015.03.003.

[8] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J Big Data*, vol. 6, no. 1, p. 28, Dec. 2019, doi: 10.1186/s40537-019-0191-6.

[9] P. Routh, A. Roy, and J. Meyer, "Estimating customer churn under competing risks," *Journal of the Operational Research Society*, vol. 0, no. 0, pp. 1–18, Aug. 2020, doi: 10.1080/01605682.2020.1776166.

[10] “Telco Customer Churn | Kaggle.” https://www.kaggle.com/blastchar/telco-customer-churn (accessed Sep. 15, 2020).
[11] “UCI Machine Learning Repository: Data Sets.” https://archive.ics.uci.edu/ml/datasets.php (accessed Sep. 15, 2020).