International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN : 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET241143

To Study and Analyse the Customer Churn Prediction using Machine Learning Algorithm

Dr. Sonali Nemade, Dr. Sujata Patil, Mrs. Deepashree Mehendale, Mrs. Vidya Shinde, Mrs. Reshma Masurekar
Dr. D. Y. Patil Arts, Commerce and Science College Pimpri, Maharashtra, India

ARTICLEINFO

Article History:

Accepted: 25 June 2024
Published: 18 July 2024

Publication Issue :
Volume 11, Issue 4
July-August-2024

Page Number :
61-65

ABSTRACT

The customer churn prediction (CCP) is one of the challenging problems in the E-Commerce industry. With the advancement in the field of machine learning and artificial intelligence, the possibilities to predict customer churn has increased significantly. Our proposed methodology, consists of six phases. In the first two phases, data pre-processing and feature analysis is performed. In the third phase, feature selection is taken into consideration. Next, the data has been split into two parts train and test set in the ratio of 80% and 20% respectively. In the prediction process, most popular predictive models have been applied, namely, logistic regression, random forest classifier etc. on train set are applied to see the effect on accuracy of models.

In addition, K-fold cross validation has been used over train set for hyper parameter tuning and to prevent overfitting of models. Finally, the obtained results on test set have been evaluated using confusion matrix and AUC curve.

Keywords : Random Forest, Logistics Regression, Customer Churn, Machine Learning

I. INTRODUCTION

Digitalization and globalization have led to new ways of doing business, and organization round the world h ave had to adapt. Subscription based services are one r esult of the tremendous digitalization that has taken t he globe by storm. With this comes both possibilities and challenges that require new solutions. Digitalizati on has revolutionized how business is performed and i ncreased the supply of subscription-based services. Co mpanies may find it harder to keep customers as a res ult. This is vital for keeping competitive and gaining a n edge over other companies.

Since information technology is growing, the amount of data and information has expanded in recent years. This rapid rise has permitted the storage and processin g of large volumes of data and increased the need to au tomatically identify and create knowledge. By extracti ng valuable information from stored data, organizatio ns can grow. With this rise, data mining and machine

learning are used more often because they can handle and analyze large amounts of data.

The digitalization movement has also improved customer relationship management (CRM) data processing processes. Knowledge management and customer relationship management have gained increased attention in the subscription-based business model. The concepts focus on allocating resources to customer-centric activities to increase competitive advantage. Customer knowledge is information gleaned from customers. Customer relationship management systems collect, store, and analyze data to give organizations an overview of their consumers. Such systems have grown over the years, and by employing technology and data analysis tools, businesses can detect patterns in client behavior that would be hard to discover manually.

These patterns could represent a customer's buying habits or churning. In a subscription-based company model, success depends on minimizing client churn. A consumer churns when they leave their service provider. The word has gained popularity with the rise of online services. Previously said, studying customer behavior helps predict attrition, which is crucial for several reasons. For companies that rely on subscription-based income, it can affect whether they can sustain a stable income or need to adapt their offerings to keep clients. Compared to retaining consumers, acquiring new ones is more expensive, thus corporations save money by keeping existing ones.

II. Objectives

1. The purpose is to segment clients so that a marketing strategy may be developed.
2. As a result, the purpose of churn prediction is to spot clients who are going to abandon the service in advance, so that the service provider can utilize marketing to maintain them.

III. Proposed Methodology

The dataset is obtained directly from the Kaggle. The dataset consist of 7043 customers and each column consist of 21 features. The dataset explains the customer id, sign up details, customer account information, and demographic information about the customers. The dataset has to be preprocessed properly before applying the supervised classification techniques. The new features can be created from existing nature from the recurrent usage of peoples. These features are necessary to determine the usage of customer in advance and it is should be much needed information for the model. The dataset is obtained directly from the Kaggle. The dataset consist of 7043 customers and each column consist of 21 features. The dataset explains the customer id, sign up details, customer account information, and demographic information about the customers. The dataset has to be preprocessed properly before applying the supervised classification techniques. The new features can be created from existing nature from the recurrent usage of peoples. These features are necessary to determine the usage of customer in advance and it is should be much needed information for the model. The dataset is obtained directly from the Kaggle. The dataset consist of 7043 customers and each column consist of 21 features. The dataset explains the customer id, sign up details, customer account information, and demographic information about the customers. The dataset has to be preprocessed properly before applying the supervised classification techniques. The new features can be created from existing nature from the recurrent usage of peoples. These features are necessary to determine the usage of customer in advance and it is should be much needed information for the model. The dataset is obtained directly from the Kaggle. The dataset consist of 7043 customers and each column consist of 21 features. The dataset explains the customer id, sign up details, customer account information, and demographic information about the customers. The dataset has to be preprocessed properly before applying the supervised classification techniques. The new features can be cr

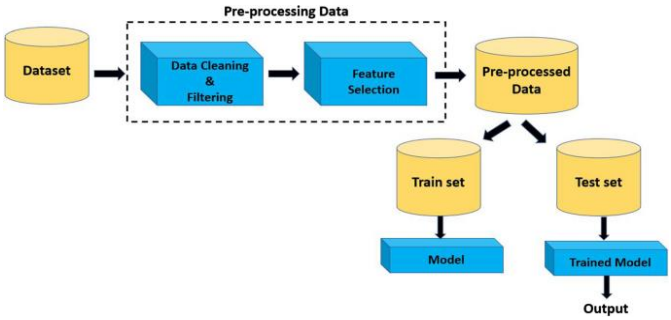
eated from existing nature from the recurrent usage of peoples. These features are necessary to determine th e usage of customer in advance and it is should be muc h needed information for the model. The dataset was acquired straight from Kaggle. There are 7043 consum ers in the dataset, and there are 21 features per colum n. The dataset needs to be appropriately preprocessed before using supervised classification methods. People 's frequent use of the already-existing nature can be us ed to build additional features. These features are esse ntial for anticipating consumer usage, and the model s hould require these information greatly.

We tested the impact of various predictor variables on customer churn. We applied machine learning model ling, which requires the following steps:

1. Pre-processing the variables present in the dataset so that they can be included in the model.
2. Defining the machine learning modelling methods to be used, in particular choice of the metric to be optimised and the type of model.
3. Training the model using various sets of variables, and the selection of independent variables which maximise the performance of the proposed model.

Running the predictions from the selected models.
The methodology used in this study can be divided into four broad categories:
Methods used in pre-processing applied to the variables present in the dataset.
Methods used for variable selection.

4. Machine learning modelling methods—choice of model, cross-validation, up-sampling, etc.



Regression analysis-logistic regression analysis

Regression is one of the statistical process for estimating how the variables are related to each other. It includes ample amount of techniques for establishing the model and analyzing several variables, when the epicenter of importance is on the bond which is shared between a dependent variable and one or many independent variables. In the light of customer churning, regression analysis is not broadly used because linear regression models are useful for predicting continuous values. But, Logistic Regression or Logit Regression analysis (LR) is a probabilistic statistical classification model. It is also used for binary classification or binary prediction of a categorical value (e.g., house rate prediction, customer churn) which depends upon one or more parameters (e.g., house features, customer features). In addressing the complex problem of customer churn prediction problem, data first has to be casted under proper data transformation from the initial data in order to achieve good performance and sometimes it performs as good as Decision Trees

Random forest classifier

It works on the divide and conquer approach. It is based on the random subspace method. In this method a number of trees are formed and each decision tree is trained by selecting any random sample of attributes from the predictor attributes set. Each tree matures up to maximum extent based on the attributes or parameters present. The final decision tree is formed

for the prediction mainly based on weighted averages. It has the ability to handle thousands of input parameters without deletion. It can also handle the missing values inside the data-set for training the predictive model.

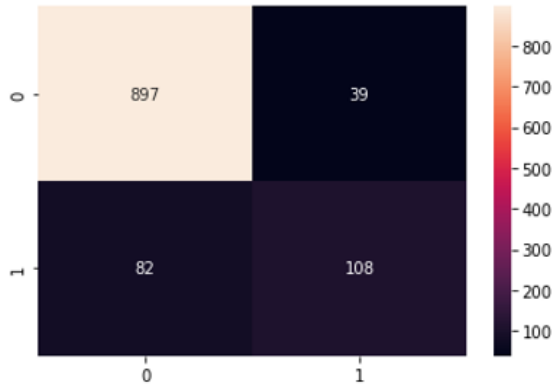
IV. Experimental Results

F-measure, precision, recall, accuracy, and throughput of Customer Churn Prediction have all been used as metrics to assess the effectiveness of applied models or throughput on the test set. It assesses how well the prediction algorithms can anticipate which consumers would eventually leave. The confusion matrix is used to calculate the aforementioned four metrics, which are displayed in Table. Table displays the confusion matrix representation. While false negative and true negative are represented as Fn and Tn, genuine positive and false positive are marked as Tp and Fp.

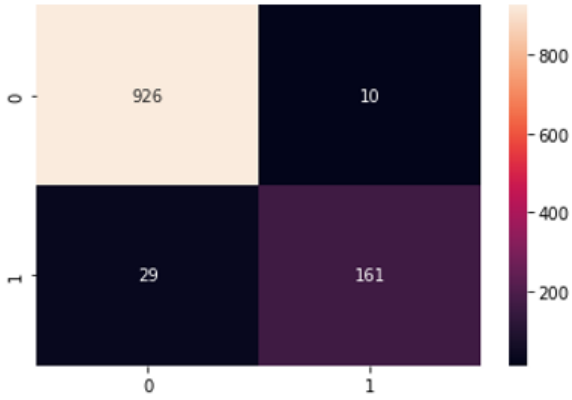
To comprehend the evaluation criterion, you should become familiar with the following four terms:

- **True Positive (Tp):** The quantity of clients that fall into the churner group and whose numbers the predictive model accurately forecasted.
- **True Negative (Tn):** The quantity of clients that the predictive model accurately forecasted and fall into the non-churner group.
- **False Positive (Fp):** The quantity of clients that the predictive algorithm has tagged or identified as churners even when they are not.
- **False Negative (Fn):** The quantity of clients who are churners but have been classified as non-churners by the predictive model.

Confusion matrix of Logistic Regression:



Confusion matrix of Random Forest:



The following is the test score and training score of the logistics regression and random forest:

	Test Score	Training Score	Accuracy Score
Logistics Regression	0.64094	0.64973	0.89253
Random Forest	0.891966	1.0	0.965364

V. CONCLUSION

Based on customer behaviour data from an actual e-commerce company, a churn prediction study was carried out. The study found that customer churn happens frequently due to the unique features of customers who use the rental business. The churn

prediction model was validated by using a machine learning algorithm to quantify churn risk information, and client contract information was monitored during the operations, making this work academically significant. Customers are specifically kept from leaving the field by defensive group actions, such enforcing a minimum number of months of use or using product groups that aren't tailored to each individual consumer.

This study aims to develop a machine learning-based churn defence tool that can accurately learn and forecast a customer's likelihood of churning. The tool can be used to undertake churn prevention marketing ahead of time for customers who have a high probability of churning.

In order to prevent customer churn, paid intervention alternatives are anticipated to include real customer marketing initiatives including qualitative care, CRM initiatives, and the application of varied prices for each customer. Future research will examine the efficacy verification outcomes, including projected revenues.

VI. REFERENCES

[1] Omar Adwan, Hossam Faris, Khalid Jaradat, Osama Harfoushi, Nazeeh Ghatasheh" Predicting customer churn in telecom industry using multilayer preceptron neural networks: modeling and analysis" Life Sci. J., 11 (3) (2014), pp. 75-81

[2] Mohammad Ridwan Ismail, Mohd Khalid Awang, M. Nordin A. Rahman, Mokhairi Makhtar"A multi-layer perceptron approach for customer churn prediction"International Journal of Multimedia and Ubiquitous Engineering, 10 (7) (2015), pp. 213-222

[3] Farquad, H. &Vadlamani, Ravi &Surampudi, Bapi. (2014). Churn Prediction using Comprehensible Support Vector Machine: an Analytical CRM Application. Applied Soft Computing. 19. 10.1016/j.asoc.2014.01.031

[4] Kumar, Dudyala& Ravi, Vadlamani. (2008). Predicting credit card customer churn in banks using data mining. International Journal of Data Analysis Techniques and Strategies. 1. 4-28. 10.1504/IJDATS.2008.020020.

[5] D. Sikka, Shivansh, R. D and P. M, "Prediction of Delamination Size in Composite Material Using Machine Learning," 2022 International Conference on Electronics and Renewable Systems (ICEARS), 2022, pp. 1228-1232, doi: 10.1109/ICEARS53579.2022.975212

[6] S. De, P. P and J. Paulose, "Effective ML Techniques to Predict Customer Churn," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 895-902, doi: 10.1109/ICIRCA51532.2021.9544785.

[7] Fatih Kayaalp"A review and analysis of churn prediction methods for customer retention in telecom industries" 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), IEEE (2017), pp. 1-7

[8] Davoud Gholamiangonabadi, Jamal Shahrabi, Seyed Mohamad Hosseinioun, Sanaz Nakhodchi, Soma Gholamveisy Customer churn prediction using a new criterion and data mining; A case study of Iranian banking industry Proceedings of the International Conference on Industrial Engineering and Operations Management (2019), pp. 5-7