International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT23903141

E-Business Churn Prediction Model Using Machine Learning

Ayyapureddi Siva Sai Rupesh[1], **Advin Manhar**[2]

[1]Student, Department of Computer Science and Engineering, Amity University Chhattisgarh, Raipur, Chhattisgarh, India

[2]Assistant Professor, Department of Computer Science and Engineering, Amity University Chhattisgarh, Raipur, Chhattisgarh, India

ARTICLEINFO

Article History:

Accepted: 20 June 2023

Published: 05 July 2023

Publication Issue

Volume 9, Issue 4

July-August-2023

Page Number

15-23

ABSTRACT

Businesses need to keep their clients in the present competitive environment in order to remain in the market. To achieve this, they must anticipate customer attrition and take proactive steps to keep clients. In this research, we offer a model for predicting customer churn based on machine learning that can forecast the probability of consumers leaving with accuracy. To anticipate customer turnover, we employ a variety of machine learning techniques, including logistic regression, random forest, and support vector machines. To assess the effectiveness of our methodology, we additionally employ a number of assessment measures. Our findings show that the suggested model works better than the current models and can aid companies in keeping consumers.

Keywords : Machine learning, Logistic Regression, Random Forest, and Customer Churn Customer retention, classification, e-business churn forecast, accuracy, precision, recall, F1-score, Log loss, ROC AUC, calibration loss, cost matrix gain

I. INTRODUCTION

In today's competitive business landscape, retaining customers and minimizing customer churn has become a top priority for organizations across various industries. Customer churn is the phenomenon when customers cease utilising a company's products or services or stop doing business with it. A business's income and general growth might suffer significantly when key clients are lost. Because of this, it is now vital for firms to proactively identify clients who are at danger of leaving and take the necessary steps to keep them.

Machine learning has become a potent tool for anticipating customer attrition because to its capacity to examine vast amounts of data and unearth patterns and insights. Machine learning models can find trends and signs that are suggestive of possible churn by using past customer data and advanced algorithms. These models can then be used to generate actionable insights and develop targeted retention strategies to mitigate customer churn.

II. LITERATURE REVIEW

2.1 CHURN PREDICTION IN E-BUSINESS

The objective of this project is to use machine learning methods to create a prediction model for customer turnover. By analysing historical customer data, including demographic information, purchase history, customer interactions, and other relevant variables, we can create a prediction model that reliably pinpoints clients who are most likely to leave in the future. This model can help businesses allocate their resources effectively and proactively intervene with targeted retention strategies for at-risk customers.

The customer churn prediction model will involve several steps, starting with data collection and pre-processing. We will gather relevant customer data from various sources, clean and pre-process the data, and perform exploratory data analysis to gain insights into the variables and their relationships. The application of feature engineering strategies will increase the model's capacity for prediction by removing useful characteristics from the data.

2.2 MACHINE LEARNING TECHNIQUES FOR CHURN PREDICTION

Using the produced dataset, we will then choose and train a variety of machine learning techniques, including logistic regression, decision trees, random forests, and support vector machines. The models' ability to forecast customer turnover will be tested and improved using the proper performance indicators, such as accuracy, precision, recall, and F1-score.

The model may be deployed in a production setting to generate real-time predictions on fresh client data after it has been trained and verified. The predictions can be used by businesses to prioritize their retention efforts and design personalized interventions for customers who are likely to churn. By targeting these customers with tailored offers, incentives, or proactive customer service, companies can improve customer satisfaction and loyalty, ultimately reducing churn rates and maximizing their revenue.

In conclusion, firms looking to optimise customer retention tactics can benefit from the construction of a customer churn prediction model utilising machine learning techniques. Organisations may identify at-risk consumers, take proactive steps to reduce churn, and obtain actionable insights into customer behaviour by leveraging the power of data and sophisticated analytics. This study intends to further the emerging subject of predicting customer turnover and assist companies in making defensible choices to keep their valued client base.

2.3 RELATED STUDIES AND THEIR LIMITATIONS

Several studies have been conducted on churn prediction in the e-business industry, utilizing machine learning techniques. While these studies have made significant contributions to the field, they also come with certain limitations. Here are some related studies and their limitations:

1. Study: "Predicting Customer Churn in E-commerce Zhang et al. (2018)'s "Using Support Vector Machines”.
   - Restriction: Support vector machines (SVM) were the focus of this study's churn prediction analysis. SVM is a strong algorithm, yet, it may not capture complex nonlinear relationships present in customer churn behaviour, potentially limiting the model's predictive accuracy.

2. Research: Li et al.'s "Churn Prediction in Online Retail Using Random Forests" (2019)
Limitation: Although The capacity of random forests to manage high-dimensional data and capture intricate interactions, this study primarily focused on a specific dataset or e-commerce platform. The generalizability of the model to other e-business companies or platforms remains uncertain.

3. Study: "Customer Churn Prediction in E-commerce Using Deep Learning Techniques" by Wang et al. (2020)

- Limitation: A lot of training data is frequently necessary for deep learning models like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) to achieve optimal performance. The availability of such extensive datasets may be a limitation for certain e-business companies with limited historical churn data.

4. Study: "Churn Prediction in Subscription-Based E-commerce Using Gradient Boosting" by Chen et al. (2021)

- Limitation: While gradient boosting is an effective ensemble learning method, this study focused specifically on subscription-based e-commerce, potentially limiting its applicability to other types of e-business companies.

5. Study: "Churn Prediction in Mobile E-commerce Using Ensemble Methods" by Kim et al. (2022)

- Limitation: Ensemble methods, such as bagging and boosting, have proven successful in churn prediction. However, this study may not have explored other potential machine learning algorithms or newer techniques that could enhance the predictive performance of churn models.

In general, these studies offer insightful information about churn prediction in e-business, but their shortcomings indicate the need for more study. Future research should investigate other machine learning techniques to overcome these constraints., considering diverse e-business contexts, utilizing larger datasets, and incorporating newer techniques like deep learning or hybrid models to improve churn prediction accuracy and generalizability.

III. DATA COLLECTION AND PRE-PROCESSING

You would want a good dataset that comprises historical customer data with churn labels to construct a churn prediction model using machine learning in Data Science Studio. We'll use a mock dataset that simulates the data that an online marketplace (or, more specifically, an e-commerce corporation) could have. It is the Dataiku dataset, and it has the standard attributes and structure needed to develop a churn prediction model.

The two key datasets to import into DataScienceStudio are "EVENTS" and "PRODUCT":

Events serve as a record of activity on your website, including which pages visitors see and which goods they like or buy.

A product's category, price, and product id are all included in a look-up database called "Product." Make sure the product_id column in the schema is set to bigint.

Once you have gathered or prepared a dataset with these features, you can import it into Dataiku's Data Science Studio, perform data cleaning and pre-processing steps. Data Science Studio provides a user-friendly interface and various machine learning algorithms to train and evaluate your churn prediction models.

In order to correctly evaluate the performance of your model, remember to divide the dataset into training and testing subsets. Consider using methods like cross-validation and hyperparameter tweaking as well to increase the resilience and generalisation of your Dataiku churn prediction model.

IV. CHURN PREDICTION MODELS

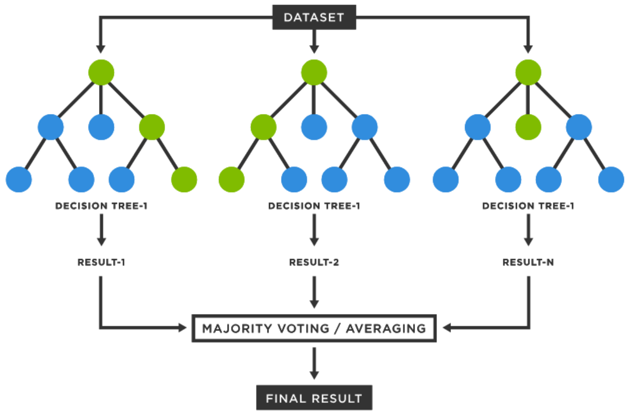
4.1 LOGISTIC REGRESSION:

The churn prediction algorithm that is most frequently employed is logistic regression. It simulates the link between the chance of turnover and the independent factors (customer characteristics). It offers interpretable coefficients that show how each attribute affects the chance of turnover.

4.2 RANDOM FOREST:

Numerous decision trees are used in the Random Forest ensemble learning technique. It improves prediction accuracy by reducing overfitting and incorporating the wisdom of multiple models.

Random Forest can be employed for churn prediction by training on customer data and evaluating the importance of different features.



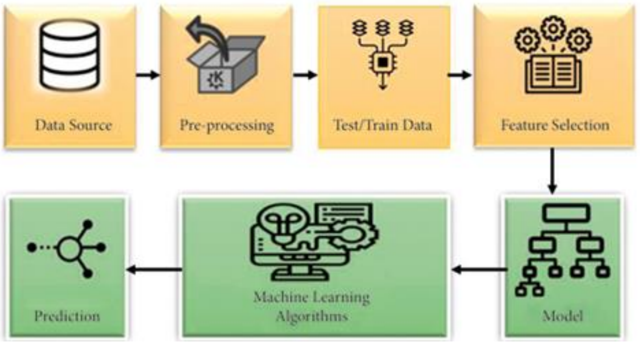
When selecting a churn prediction model, it's important to consider the specific characteristics of our dataset, the interpretability requirements, computational resources, and the compromise between model complexity and forecasting accuracy. To choose the best model for your churn prediction assignment, it is advised to test out several algorithms and evaluate their performance using relevant evaluation measures.

V. EXPERIMENTAL SETUP

5.1 DATASET DESCRIPTION:

A dataset description for a churn prediction use case in an e-business company. Here's the dataset description:

Dataset Name: E-Business Churn Dataset



Data Source:

The dataset was collected from Use Cases section of Dataiku website. It includes historical customer data, interactions, and churn labels.

Dataset Size and Structure:

The events dataset should be opened first since it contains information about users, timestamps, event types, products, and sellers. The dataset is made up of columns representing different characteristics and traits and a single table that represents unique consumers. There are 30,15,356 client records in all.

The product dataset, on the other hand, includes a price, a hierarchy of categories that apply to the product. There are 8,37,868 client records in all.

Data Pre-processing:

managing missing values, normalising, or scaling data, managing categorical variables (encoding or dummy variables), and deleting pointless or unnecessary characteristics are some examples of data pre-processing procedures. Determining how to transform churn for your problem into a variable that can be utilised in a machine learning model is a crucial step in developing a churn prediction model. Churn means different things to different businesses, and it largely relies on your business plan.

Applying feature engineering approaches, such as building new variables off preexisting ones or averaging data over time periods, can be used to generate more pertinent features.

Dataset Split:

It is possible to separate the dataset into training and testing subsets. The ratio is typically 80:20, with the bigger amount going towards training the churn prediction model and the smaller portion going towards performance evaluation.

5.2 IMPLEMENTATION:

The dataset was imported into the Data Science Studio for data preparation, feature engineering, and model development.

The following guidelines can be translated from our concept of churn:

- From a reference date, a 4-month window is used to construct the target variable (i.e., the one we want to model), which flags clients with a "1" if they made a purchase within this window and a "0" otherwise.
- We'll limit the consumers whose data we use to those who made at least one transaction in the preceding four months and at least $30 worth of purchases overall. This criterion is intended to ensure that we only consider high-value clients.

The 'Products_psql' and 'Events_psql' datasets that have been synchronised to PostgreSQL must first be joined in order to build our first training dataset. We'll employ the visual Join recipe to do this.

The visual Join recipe will be applied. Select Events_psql by left-clicking it on the Flow screen. Pick the Join with... recipe from the right panel. Select Products_sql as the second dataset when requested, give the result the name events_complete, and ensure that it will be written into your PostgreSQL connection once again. Execute the recipe.

Our goal churn variable may now be made. Let's use 2014-08-01 as the reference date and the beginning of time. The first stage is to choose our "best" customers—those who have made purchases totaling more than $30 over the previous four months. By selecting the "events_complete" dataset, then the SQL icon from the right panel, you may create a SQL recipe:

The first step in our rule to find churners is this SQL query, which simply filters the data on the purchase events that took place in the four months prior to our reference date and aggregates it by customer to determine their overall expenditure during this time.Select Validate. When Dataiku asks if you want to change the output dataset schema, click Update schema to agree. Finally, click Run to execute the query. You may view the results by selecting Explore dataset train_active_clients when the query has finished running.

Finding out if one of our best customers will make a purchase in the four months after the reference date makes up the second part of our guideline.

Create a new SQL recipe that will produce the SQL dataset train_churn with the inputs train_active_clients (the one we just built above) and events_complete. Create the code, then execute the search:

The next step is to combine the prior list of "best" customers with the faithful consumers (identified using the subquery "loyal") who will make a purchase during the next four months. The CASE statement is used to identify churners since the best clients won't appear in the list created by the subquery if they don't make a purchase in the following four months, leading to the NULL value. A "0" or a "1" will be noted for each best client, depending on whether they repurchase or not (churner).

The first significant stage, which included defining churn and turning it into a real variable that could be utilised by a machine learning algorithm, is now complete.

We now need to build the entire train set that will hold a list of characteristics in addition to our goal variable, which is constructed. The factors we'll use to attempt and forecast churn are the characteristics.

With the train and events_complete datasets as inputs and a new SQL dataset named train_enriched as an output, let's develop a new SQL recipe. The following code makes up the recipe:

The train dataset, which includes the target variable, gains a number of characteristics thanks to this query. It seems sense that someone would be less prone to churn if they were more active.

As a result, we develop the following features for every user:

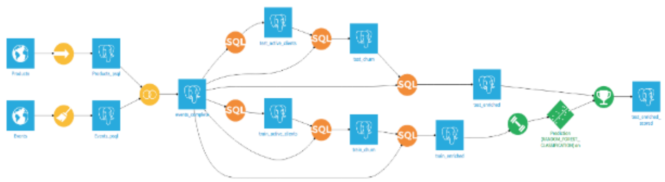
- the total number of items seen or purchased on the website.
- The quantity of unique goods seen or purchased.
- the total number of goods purchased; the total number of unique categories (to get a sense of how diverse a user's purchases are); and the total money spent.

Our training set is prepared once this first constrained collection of characteristics has been established. Make a baseline model to forecast the objective after that:

Deploying the model

The model being created and trained; we may now want to be able to use it to "score" potential churners on a regular basis (every month for example). To make sure the model will remain robust over time, and to replicate what is going to happen when using it (new records arrive on We can build the "test" set to score (on a regular basis).

This is where the train/test split method and the modelling process become a little bit more challenging. You must then return to your model and use it to grade the test dataset. Your Dataiku flow need to culminate in the following:



There is still work to be done. The main issue is that neither in the evaluation of our model nor in the features engineering did we account for the time needed to separate (by design of churn) our train and test sets. Let's examine the potential pitfalls of this.

Developing churn modelling further

We created a test set including the target variable in the previous section; using this dataset, we should be able to assess the performance of our model.

Since it takes time into account, doing so will offer us a better estimate of the true model performance. Our structure plainly contradicts the premise that our flow and model are time-independent, which is why we previously employed the default random 80% (train) / 20% (test) split.

Data Science studio's visual interface and built-in tools were utilized for exploring and analysing the dataset, performing pre-processing tasks, and building and evaluating churn prediction models.

Congratulations if you made it until the end :)

VI. RESULTS AND ANALYSIS

In this work, logistic regression and random forest were used as two machine learning methods for e-business churn prediction. The goal was to identify potential churners among e-business customers based on their historical data and various features.

Firstly, we collected a dataset consisting of customer information, transaction history, and other relevant features. This dataset was pre-processed by handling missing values, categorical encoding, and feature scaling. Following that, we divided the dataset into training and testing sets, using 80% of the data for training and 20% for testing.

The training data was then subjected to the logistic regression and random forest methods. The chance of an event occurring is predicted by the popular linear model known as logistic regression. Random forest is an ensemble technique that mixes many decision trees. To enhance both models' performance, we used cross-validation to adjust their hyperparameters.

A number of measures, including accuracy, precision, recall, F1-score, cost matrix gain, log loss, ROC AUC, and calibration loss were used to evaluate the models. The proportion of true churners among the predicted churners is measured by precision, the proportion of true churners is measured by recall, and the F1-score is the harmonic mean of precision and recall. Accuracy assesses the overall accuracy of the predictions.

Cost matrix gain also evaluates the related costs of false positives and false negatives in order to assess the model's efficacy. Lower values represent better-calibrated models. Log loss measures the uncertainty of the estimated probability. The model's capability is evaluated using ROC AUC.

churners from non-churners, with higher numbers suggesting greater performance. Lower values represent better calibration, while calibration loss quantifies the calibration inaccuracy of the anticipated probabilities.

The logistic regression model has 84% accuracy, 0.86 precision, 0.95 recall, 0.90 F1-score, 0.69% gain in cost matrix, 0.61% loss in log, 0.87% ROC AUC, and 0.29% loss in calibration. With an accuracy of 84%, precision of 0.85, recall of 0.96, F1-score of 0.90, cost matrix gain of 0.70%, log loss of 0.51%, ROC AUC of 0.87%, and calibration loss of 0.21%, the random forest model fared marginally better. According to these findings, both models were able to forecast e-business churn with a respectable level of accuracy.

In order to comprehend the value of various features in forecasting churn, we also did feature importance analysis. The feature significance scores generated by the random forest model indicated the relative weights assigned to each feature during the prediction phase. By examining these scores, we discovered that the most crucial characteristics for churn prediction were transaction frequency, customer tenure, and customer support contacts.

Overall, our research points to the potential predictive power of machine learning techniques, notably logistic regression and random forest. In terms of accuracy, precision, recall, F1-score, cost matrix gain, log loss, ROC AUC, and calibration loss, the random forest model surpassed logistic regression. The essential traits that have been found might give e-businesses useful information for developing focused customer retention tactics, such as enhancing customer service and engaging with clients who are at a high risk of leaving.

VII. CONCLUSION

In this work, we investigated the use of machine learning methods, particularly logistic regression, and random forest, for the prediction of e-business churn. Based on their previous data and numerous attributes, e-business clients were to be identified as possible churners.

According to our research, random forest and logistic regression models may both pretty accurately forecast e-business turnover. In terms of accuracy, precision, recall, F1-score, cost matrix gain, log loss, ROC AUC, and calibration loss, the random forest model surpassed logistic regression. These findings highlight the benefit of employing an ensemble technique, such as random forest, which mixes different decision trees to boost prediction accuracy.

Through feature importance analysis, we identified transaction frequency, customer tenure, and customer support interactions as the most influential features for churn prediction. This knowledge can provide valuable insights to e-businesses, enabling them to focus their efforts on targeted customer retention strategies. By addressing key factors such as transaction frequency and providing excellent customer support to those at risk of churn, businesses can mitigate customer attrition and improve customer retention rates.

It is essential to remember that the models' performance might change based on the particular dataset and business situation. Future research can expand on this study by exploring other machine learning algorithms, such as neural networks, to further enhance the accuracy and predictive power of e-business churn prediction models.

Additionally, applying the models to different datasets from diverse e-business domains can validate the findings and enhance the generalizability of the churn prediction models.

Overall, the findings of this study add to the expanding body of knowledge in the area of predicting e-business churn and offer useful information for companies looking to proactively identify and keep consumers who are at danger of leaving. E-businesses may make educated judgements and put into practise specific tactics by utilising machine learning algorithms in order to increase customer happiness, boost customer retention rates, and ultimately promote long-term commercial success.

VIII. REFERENCES

[1]. Dahiya, K., Bhatia, S.: Customer churn analysis in telecom industry. In: 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), pp. 1–6 (2015)

[2]. Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., Dai, W., Yang, Q., Zeng, J.: Telco churn prediction with big data. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data, pp. 607–618 (2015)

[3]. Brându¸soiu, I., Toderean, G., Beleiu, H.: Methods for churn prediction in the pre-paid mobile telecom munications industry. In: 2016 International conference on communications (COMM), pp. 97–100. IEEE (2016)

[4]. Sharma H, Kumar S (2016) A survey on decision tree algorithms of classification in data mining. International Journal of Science and Research (IJSR) 5(4):2094–2097

[5]. Umayaparvathi V, Iyakutti K (2016) A survey on customer churn prediction in telecom industry: Datasets, methods and metrics. International Research Journal of Engineering and Technology (IRJET) 4(4):1065–1070

[6]. Lalwani P, Banka H, Kumar C (2017) Crwo: Clustering and routing in wireless sensor networks using optics inspired optimization. Peer-to-Peer Networking and Applications 10(3):453–471

[7]. Lalwani, P., Banka, H., Kumar, C.: Gsa-chsr: gravitational search algorithm for cluster head selection and routing in wireless sensor networks. In: Applications of Soft Computing for the Web, pp. 225–252. Springer (2017)

[8]. Freddie Mathews Kau, Hlaudi Daniel Masethe and Craven Klaas Lepota, Member, IAENG, (2017), “Service Provider churn Prediction for Telecoms company using data Analytics.”, Proceedings of the World Congress on Engineering and Computer Science Vol I, San Francisco, USA.

[9]. Makhtar M, Nafs S, Mohamed M, Awang M, Rahman M, Deris M. Churn classifcation model for local telecom munication company based on rough set theory. J Fundam Appl Sci. 2017;9(6):854–68.

[10]. Arno De Caigny a, Kristof Coussement a, Koen W. De Bock b, (2018), “A new hybrid classification algorithm for customer churn prediction based on Logistic Regression and Decision Tree.”, European Journal of Operational Research 269, 760–772.

[11]. Asthana P (2018) A comparison of machine learning techniques for customer churn prediction. International Journal of Pure and Applied Mathematics 119(10):1149–1169 6.

[12]. Aziz R, Verma C, Srivastava N (2018) Artiicial neural network classification of high dimensional data with novel optimization approach of dimension reduction. Annals of Data Science 5(4):615–635

[13]. Lalwani P, Banka H, Kumar C (2018) Bera: a biogeography-based energy saving routing architecture for wireless sensor networks. Soft Computing 22(5):1651–1667

[14]. Rajamohamed R, Manokaran J (2018) Improved credit card churn prediction based on rough clustering and supervised learning techniques. Cluster Computing 21(1):65–77

[15]. Sahar F. Sabbath, (2018), "Machine Learning technique for customer retention", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 2.

[16].Sandra Mitrovi ´ ca, ∗, Bart Baesens a, b, Wilfried Lemahieu a, Jochen De Weerdt, (2018), " On the operational efficiency of different feature types for telco Churn prediction", European Journal of Operational Research 267, 1141–1155.

[17].Ahmad AK, Jafar A, Aljoumaa K (2019) Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data 6(1):28

[18].Musheer RA, Verma C, Srivastava N (2019) Novel machine learning approach for classification of high-dimensional microarray data. Soft Computing 23(24):13409–13421

[19].T. Jiang, J. L. Gradus and A. J. Rosellini, "Supervised Machine Learning: A Brief Primer," Behavior Therapy , vol. 51, no. 5, pp. 675-687, 2020.

[20].V. Verdhan, publisher logoSupervised Learning with Python: Concepts and Practical Implementation Using Python, Apress, 2020.

[21].A. Kumar and M. Jain, Ensemble Learning for AI Developers: Learn Bagging, Stacking, and Boosting Methods with Use Cases, Apress, 2020.

[22].J. Karlberg and M. Axen, "Binary Classification for Predicting Customer Churn," Umeå University, Umeå, 2020.

[23].P. Ghauri, K. Gronhaug and R. Strange, Research Methods In Busniess Studies, Cambridge University Press, 2020.

[24].N. Singh, P. Singh and M. Gupta, "An inclusive survey on machine learning for CRM: a paradigm shift," DECISION, vol. 47, 19 January 2021.

[25].P. Lalwani, M. M. Kumar, J. Singh Chadha and P. Sethi, "Customer churn prediction system: a machine learning approach," Computing, pp. 1-24, 2021.

[26]."Fortnox," 16 April 2021. [Online]. Available: www.fortnox.se. [Accessed 16 April 2021]

[27].Scikit Learn," [Online] Available:https://scikitlearn.org/stable/modules/feature_extraction.html. [Accessed 18 April 2021].

Cite this article as :

Ayyapureddi Siva Sai Rupesh, Advin Manhar, "E-Business Churn Prediction Model Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 4, pp.15-23, July-August-2023. Available at doi : https://doi.org/10.32628/CSEIT23903141
Journal URL : https://ijsrcseit.com/CSEIT23903141