Customer Churn Prediction in the Telecom Industry Using Machine Learning Algorithms

Mr. Abhinav Sudhir Thorat, Dr. Vijay Ramnath Sonawane
Department of Computer Science and Engineering,
Dr. A. P. J. Abdul Kalam University, Indore (M.P.) - 452016, India

Abstract:
Customer Churn Prediction in the Telecom Industry Using Machine Learning Algorithms Customer churn detection is one of the most important research topics in the telecommunications industry because the company must deal with retaining on-hand customers. Churn refers to the loss of customers as a result of competitors' exiting offers or network issues. In these cases, the customer may decide to cancel their service subscription. Churn rate has a significant impact on customer lifetime value because it affects the company's future revenue as well as the length of service. Companies are looking for a model that can predict customer churn because it has a direct impact on the industry's income. Machine learning techniques are used in the model developed in this work. We can predict which customers are likely to cancel their subscriptions by using machine learning algorithms. We can use this to provide them with better services and lower the churn rate. These models assist telecom services in becoming profitable. We used a Decision Tree, Random Forest, and XGBoost in this model.

Keywords: Telecom churn, Xgboost(Extreme Gradient Boosting) Classification algorithms, Decision Trees, Random Forest.
DOI: 10.24297/j.cims.2023.1

1. Introduction

The telecommunications industry has emerged as one of the most important in developed countries. Service businesses suffer, particularly from customer churn, or the loss of valuable customers due to competitors. The level of opposition increased as science advanced and the number of operators increased. Companies are working hard to survive in this competitive market, relying on complex strategies. Customer churn results in significant loss of telecom services and becomes a serious issue. Three major approaches have been introduced to increase profits by acquiring new customers, upselling current customers, and increasing customer retention. However, comparing these strategies using the value of return on investment (RoI) of each has revealed that the third approach is the most successful strategy, demonstrating that maintaining an existing customer costs much less than acquiring a new one, in addition to being held much easier than upselling tactics. To implement the third strategy, businesses must reduce the possibility of customer churn, also known as "customer movement from one provider to another." Customer churn is a major concern in service industries with highly competitive services. On the other hand, predicting which customers are likely to leave the company can serve as a potentially large-hearted extra revenue source if done early on. Several types of research have confirmed that machine learning technology is extremely effective at predicting this situation. This method is used to learn from previous data.

2. Existing System

Various techniques, such as data mining, machine learning, and hybrid technologies, have been used to predict customer churn. These techniques allow and assist businesses in identifying, forecasting, and retaining churn customers. They also assist businesses with CRM and decision making. Most of them used decision trees in common because it is a well-known method for determining customer churn, but it is not appropriate for complex problems [1]. However, the study found that reducing the data improves the decision tree's accuracy [2]. Data mining algorithms are sometimes used for customer prediction and historical analysis. Regression tree techniques were discussed in conjunction with other commonly used data mining methods such as decision trees, rule-based learning, and neural networks [3].

Proposed System

In this system, we use various algorithms such as Random Forest, XGBoost, and Logistic Regression to find accurate values and predict customer churn. Here, we implement the model by using a dataset that has been trained and tested, resulting in the greatest number of correct values. Figure 1 depicts and describes the proposed model for churn prediction. The first step is data preprocessing, which involves filtering data and converting it into a similar format before performing feature selection. To find accurate values and predict customer churn, we use various algorithms such as Random Forest, XGBoost, and Logistic Regression in this system. We use a trained and tested dataset to implement the model, resulting in the greatest number of correct values. The proposed model for churn prediction is depicted and described in Figure 1. The first step is data preprocessing, which entails filtering and converting data into a similar format prior to performing feature selection.

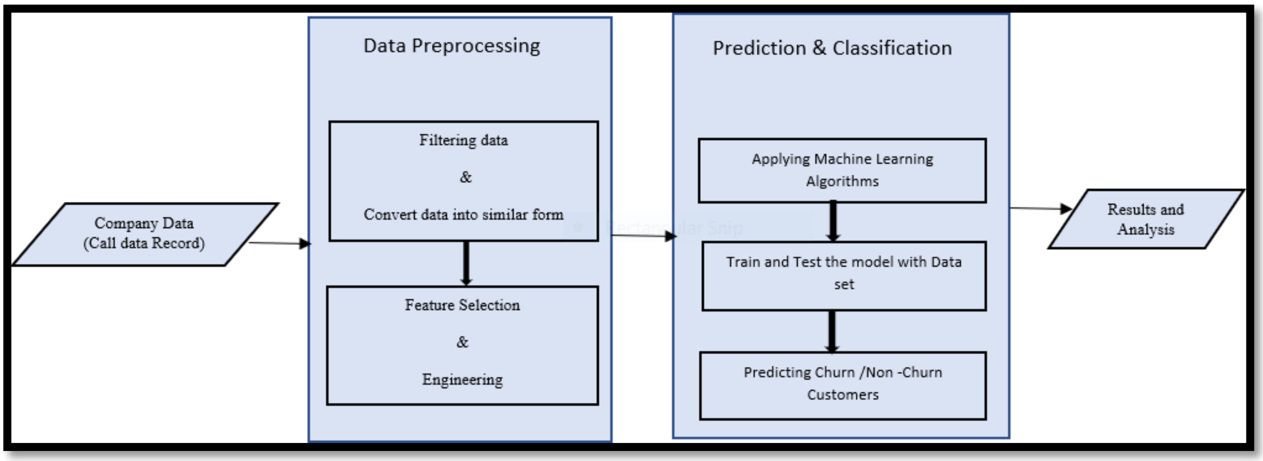

Company Data
(Call data Record)
Data Preprocessing
Filtering data
&
Convert data into similar form
Feature Selection
&
Engineering
Prediction & Classification
Applying Machine Learning Algorithms
Train and Test the model with Data set
Predicting Churn /Non -Churn Customers
Results and Analysis

Fig 1. Proposed Model for Customer Churn Prediction

Data Set

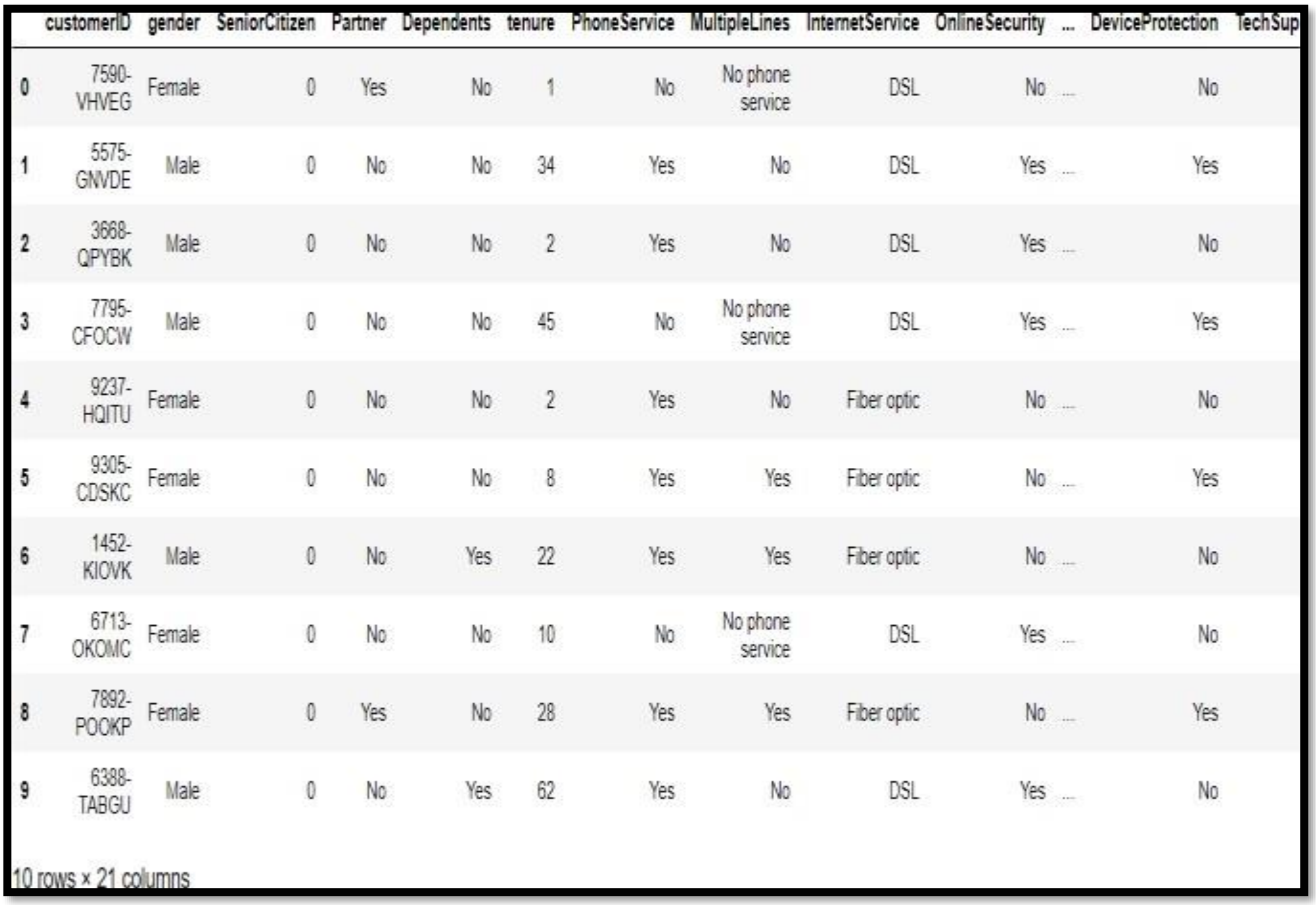
1) Customer demographic information: age, gender, occupation, education, location, marital status, etc.
2) Customer account information: account type, credit limit, balance, payment history, etc.
3) Customer behavior information: frequency of purchases, time spent on website, product usage, etc.
4) Customer feedback information: customer ratings, customer surveys, customer reviews, etc.
5) Marketing information: campaigns, offers, promotions, etc.
6) Transaction information: transaction history, purchase details, payment details, etc.

3. Methodology

Data Preprocessing

A data set is a collection of feathers with N rows. Many values are in various formats. They may be the starting point for everything in a dataset; it should have full-fledged data to help the machine learn about the problem. Scrap information on the internet can be used to generate or develop datasets. Some issues require us to create a dataset that makes sense and tells us how to respond based on real-time inputs for the problem datasets can be obtained from the internet on a daily basis. For each member of the data set, the data set lists the values of variables such as height and object weight. Every value is identified as a datum. As previously stated, the data set serves as the starting point for this process.

Table 1: Customer Data for preprocessing



	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSup
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	
5	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	...	Yes	
6	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	...	No	
7	6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	...	No	
8	7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	...	Yes	
9	6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	...	No	

10 rows × 21 columns

Data was collected from various sources, so each uses a different format to represent a single value, such as gender, which is represented by M/F or Male/Female. Because the machine can only understand 0 and 1, an image in 3-dimension data should be reduced to a 2-dimension format like data show to avoid noisy data, null values, and incorrect size. Panda's tabular data and OpenCV for images can be used to clean data.

Data Filtering and Noise Removal

It is critical to make the data useful because undesirable or null values can produce unsatisfactory results or lead to less accurate results. There are many incorrect and missing values in the data set. We examined the entire dataset and identified only the most useful features. The listing of features can result in greater accuracy and only includes useful features.

Feature selection & Engineering Feature selection is an important step in selecting the necessary elements from the data set based on knowledge.

The dataset used here has many features from which we selected the ones that will help us improve performance measurement and make decisions, while the rest will be of less importance. The classification performance improves when the dataset contains only valuable and highly predictable variables. Thus, having only significant features and reducing the number of irrelevant attributes improves classification performance.

Prediction & Classification

In the telecommunications industry, many techniques for predicting customer churn have been proposed. These three modelling techniques are used as churn predictors in this study. These methods are as follows:

Random Forest

We use Random Forest to forecast whether a customer will cancel his subscription. Random Forest employs Decision trees to predict whether a customer will cancel his subscription. The random forest is made up of many decision trees. A decision tree directs attention to a specific class. A class with the most votes will be the classifier for a specific customer. Decision trees are highly sensitive to the data on which they are trained. We use Bagging to avoid this. Bagging is a process in which we take a random sample from a dataset to train decision trees.

Logistic Regression

We can predict the likelihood of churn, or the likelihood of a customer cancelling their subscription, using logistic regression. Logistic regression is a classification algorithm that uses supervised learning. In logistic regression, we set a limit and only the classification is done using logistic regression. The threshold value is variable and is determined by the classification problem.

XGBoost

eXtreme Gradient Boosting is abbreviated as XGBoost. The primary reason for using XGBoost is because of its execution speed and model performance. XGBoost employs ensemble learning methods, which involve combining multiple algorithms to produce a single model. XGBoost supports parallel and distributed computing while utilising memory efficiently.

4. Proposed Work

Initially, we will obtain the dataset from Kaggle and remove all of the null values using data filtering. The data was then converted into a similar format that was easier to understand and analyse. We attempt to implement a predictor model for the Telecom company using Logistic regression and a novel approach. We start with a customer data set and divide it into training and testing by preprocessing and feature selection. We did some feature engineering for this algorithm in order to get more efficient and accurate results.

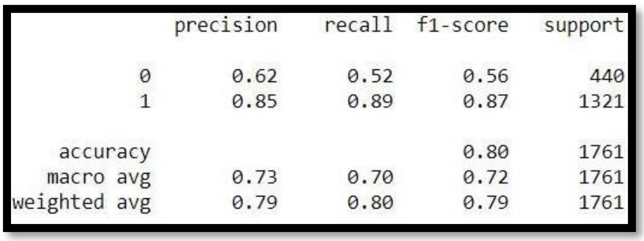
Logistic regression helps us to have a discriminative probabilistic classification and can estimate the probability of occurring event places. The dependent variable presents the event occurrence (e.g., it will be one if the event takes place, 0 otherwise). By training, the data to that model will get a result having their details, and then we will test the model with the remaining amount of data. As a result of the findings, we will have an accurate prediction of customer churn and a clear warning about the customer, which will allow the company to take some measures to avoid losing the existing customer from the service. (In this case, we divided the data into 80% training and 20% testing).

By obtaining both results, we will attempt to fix the y1 train data and y2 test data to the model fit in order for the model to learn from historical data. In this case, epochs are used to force the model to learn the same data multiple times. We visualised the data using the CNN model, and this allowed us to determine the model accuracy of the resulting data, as well as predict churn. (Fig.6,Fig.7)

Similarly, we used the other two techniques to determine which would yield more accurate results. We used the same dataset in the Random Forest, and by using the technique, we trained the model and tested it to get the results in the confusion matrix, which will show us the obtained output and the accuracy (fig.3). The XGBoost model result is shown in (fig.4), where we can see the accuracy obtained by using that technique.

5. Result And Analysis

On the dataset, we ran several experiments on the proposed churn model using machine learning algorithms. In Fig.2, we can see the results of the experiment performed with the Random Forest algorithm and check the accuracy. Random Forest (RF) is a useful classification algorithm that can handle nonlinear data very efficiently. When compared to the other techniques, RF produced better results, accuracy, and performance. We prefer to use the technique that results in better accuracy because we need better accuracy to predict customer churn. Similarly, the results obtained when using the Logistic regression technique (Fig.4) and XGBoost can be seen (Fig.5). Finally, we used the CNN model to visualise the data. The visualised data can be seen in Figs. 6 and 7.

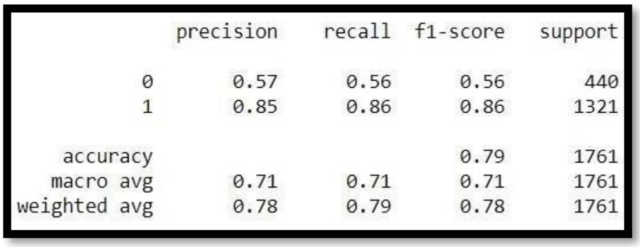


```
              precision    recall  f1-score   support

           0       0.62      0.52      0.56       440
           1       0.85      0.89      0.87      1321

    accuracy                           0.80      1761
   macro avg       0.73      0.70      0.72      1761
weighted avg       0.79      0.80      0.79      1761
```

Fig 3.Confusion matrix of Random Forest.

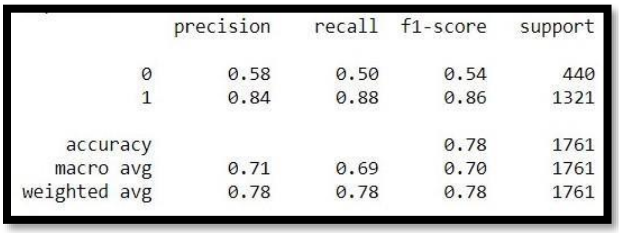


```
              precision    recall  f1-score   support

           0       0.57      0.56      0.56       440
           1       0.85      0.86      0.86      1321

    accuracy                           0.79      1761
   macro avg       0.71      0.71      0.71      1761
weighted avg       0.78      0.79      0.78      1761
```

Fig 4.Confusion matrix of Logistic regression.



```
              precision    recall  f1-score   support

           0       0.58      0.50      0.54       440
           1       0.84      0.88      0.86      1321

    accuracy                           0.78      1761
   macro avg       0.71      0.69      0.70      1761
weighted avg       0.78      0.78      0.78      1761
```

Fig 5. Confusion matrix of XGBoost.

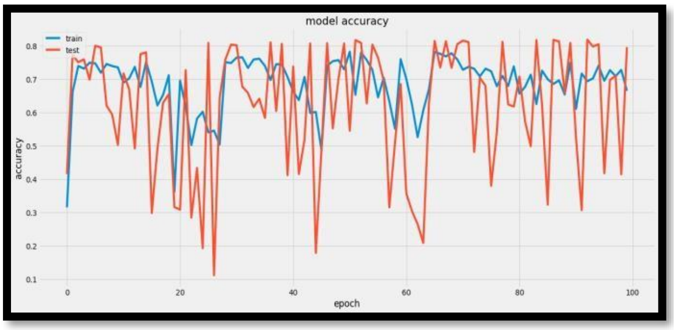


Fig.6.Visualizing CNN model val_acc and accuracy.

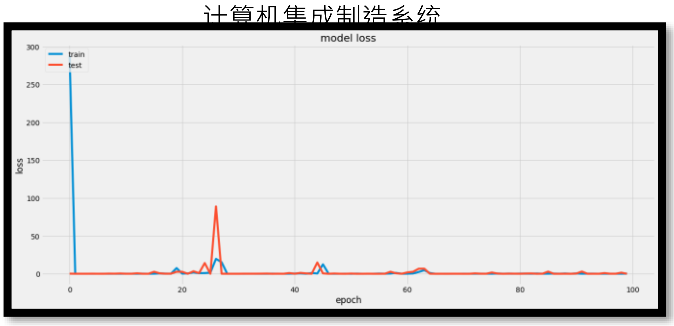


Fig.7.Visualizing CNN model val_loss and loss.

6. Conclusion

The importance of churn prediction will assist many companies, particularly those in the telecom industry, in achieving a profitable income and good revenue. Customer churn prediction is a major issue in the telecom industry, and as a result, companies prefer to keep existing customers rather than acquire new ones. Because of their applicability and diversity in this type of application, three tree-based algorithms were chosen. When compared to other algorithms, we will achieve greater accuracy by using Random Forest, XGBoost, and Logistic regression. We are using a dataset of some customers' service plans and checking the values to make a precise prediction, which will aid in identifying customers who are planning to migrate to other company services. This gives the telecom company a clear picture and allows them to make some exciting offers to stay in that service. The results show that using machine learning techniques, our proposed churn model produced better results and performed better. Among the various methods, Random Forest produced the highest accuracy. In the coming days, we will conduct additional research on lazy learning approaches to improve customer churn prediction. The study can be expanded to learn about changing customer behaviour by employing Artificial Intelligence techniques for trend analysis and customer prediction.

References

1. S. Babu, D. N. Ananthanarayanan, and V. Ramesh, ’ ‘A survey on factors impacting churn in telecommunication using datamininig techniques,’ ’ Int. J. Eng. Res. Technol., vol. 3, no. 3, pp. 1745–1748, Mar. 2014.
2. C. Geppert, ’ ‘Customer churn management: Retaining high-margin customers with customer relationship management techniques,’ ’ KPMG & Associates Yarhands Dissou Arthur/Kwaku Ahenkrah/David Asamoah, 2002.
3. W. Verbeke, D. Martens, C. Mues, and B. Baesens, ’ ‘Building comprehensible customer churn prediction models with advanced rule induction techniques,’ ’ Expert Syst. Appl., vol. 38, no. 3, pp. 2354–2364, Mar. 2011.

4. Y. Huang, B. Huang, and M.-T. Kechadi, ’ ‘A rule-based method for customer churn prediction in telecommunication services,’ ’ in Proc. Pacific– Asia Conf. Knowl. Discovery Data Mining. Berlin, Germany: Springer, 2011, pp. 411–422.
5. A. Idris and A. Khan, ’ ‘Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers,’ ’ in Proc. 15th Int. Multitopic Conf., Dec. 2012, pp. 23–27.
6. M. Kaur, K. Singh, and N. Sharma, ’ ‘Data mining as a tool to predict the churn behaviour among Indian bank customers,’ ’ Int. J. Recent Innov. Trends Comput. Commun., vol. 1, no. 9, pp. 720–725, Sep. 2013.
7. V. L. Miguéis, D. van den Poel, A. S. Camanho, and J. F. e Cunha, ’ ‘Modeling partial customer churn: On the value of first product-category purchase sequences,’ ’ Expert Syst. Appl., vol. 12, no. 12, pp. 11250–11256, Sep. 2012.
8. D. Manzano-Machob, ’ ‘The architecture of a churn prediction system based on stream mining,’ ’ in Proc. Artif. Intell. Res. Develop., 16th Int. Conf. Catalan Assoc. Artif. Intell., vol. 256, Oct. 2013, p. 157.
9. P. T. Kotler, Marketing Management: Analysis, Planning, Implementation and Control. London, U.K.: Prentice-Hall, 1994.
10. Mr. Umakant Dinkar Butkar, Manisha J Waghmare. (2023). Hybrid Serial-Parallel Linkage Based six degrees of freedom Advanced robotic manipulator. *Computer Integrated Manufacturing Systems*, *29*(2), 70–82.
11. J. Hadden, A. Tiwari, R. Roy, and D. Ruta, ’ ‘Computer assisted customer churn management: State-of-the-art and future trends,’ ’ Comput. Oper. Res., vol. 34, no. 10, pp. 2902–2917, Oct. 2007.
12. H.-S. Kim and C.-H. Yoon, ’ ‘Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market,’ ’ Telecommun. Policy, vol. 28, nos. 9–10, pp. 751–765, Nov. 2004.
13. Y. Huang and T. Kechadi, ’ ‘An effective hybrid learning system for telecommunication churn prediction,’ ’ Expert Syst. Appl., vol. 40, no. 14, pp. 5635–5647, Oct. 2013.
14. A. Sharma and P. K. Kumar. (Sep. 2013). ’ ‘A neural network based approach for predicting customer churn in cellular network services.’ ’ [Online]. Available: https://arxiv.org/abs/1309.3945
15. Ö. G. Ali and U. Aritürk, ’ ‘Dynamic churn prediction framework with more effective use of rare event data: The case of private banking,’ ’ Expert Syst. Appl., vol. 41, no. 17, pp. 7889–7903, Dec. 2014.

16. A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, ’ ‘Customer churn prediction in telecommunication industry using data certainty,’ ’ J. Bus. Res., vol. 94, pp. 290–301, Jan. 2019.
17. Mr. Umakant Dinkar Butkar, Manisha J Waghmare. (2023). An Intelligent System Design for Emotion Recognition and Rectification Using Machine Learning. *Computer Integrated Manufacturing Systems*, *29*(2), 32–42.
18. V. Lazarov and M. Capota, ’ ‘Churn prediction,’ ’ Bus. Anal. Course, TUM Comput. Sci, Technische Univ. München, Tech. Rep., 2007. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.7201&rep=rep1&type=pdf
19. R. Vadakattu, B. Panda, S. Narayan, and H. Godhia, ’ ‘Enterprise subscription churn prediction,’ ’ in Proc. IEEE Int. Conf. Big Data, Nov. 2015, pp. 1317–1321.
20. V. Umayaparvathi and K. Iyakutti, ’ ‘Applications of data mining techniques in telecom churn prediction,’ ’ Int. J. Comput. Appl., vol. 42, no. 20, pp. 5–9, Mar. 2012.
21. A. T. Jahromi, M. Moeini, I. Akbari, and A. Akbarzadeh, ’ ‘A dual-step multi-algorithm approach for churn prediction in pre-paid telecommunications service providers,’ ’ J. Innov. Sustainab., vol. 1, no. 2, pp. 2179–3565, 2010.
22. U D. Butkar, and et.al.. (2023). Novel Energy Storage Material and Topologies of Computerized Controller. *Computer Integrated Manufacturing Systems*, *29*(2), 83–95. [23] G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi, ’ ‘Credit card churn forecasting by logistic regression and decision tree,’ ’ Expert Syst. Appl., vol. 38, no. 12, pp. 15273–15285, Nov./Dec. 2011.
23. S. V. Nath and R. S. Behara, ’ ‘Customer churn analysis in the wireless industry: A data mining approach,’ ’ in Proc. Annu. Meeting Decis. Sci. Inst., vol. 561, Nov. 2003, pp. 505–510.
24. Y. Zhang, J. Qi, H. Shu, and J. Cao, ’ ‘A hybrid KNN-LR classifier and its application in customer churn prediction,’ ’ in Proc. IEEE Int. Conf. Syst., Man Cybern., Oct. 2007, pp. 3265–3269.
25. H. Yu et al., ’ ‘Feature engineering and classifier ensemble for KDD cup 2010,’ ’ Dept. Comput. Sci. Inf. Eng., National Taiwan Univ., Taipei, Taiwan, Tech. Rep., 2010, vol. 1, pp. 1–16.
26. Mr. Umakant Dinkar Butkar, Manisha J Waghmare. (2023). Novel Energy Storage Material and Topologies of Computerized Controller. *Computer Integrated Manufacturing Systems*, *29*(2), 83–95.

27. G. Holmes, A. Donkin, and I. H. Witten, ′ ‘WEKA: A machine learning workbench,’ ′ in Proc. Austral. New Zealnd Intell. Inf. Syst. Conf., Dec. 1994, pp. 357–361.
28. F. S. Gharehchopogh and S. R. Khaze. (2013). ′ ‘Data mining application for cyber space users tendency in blog writing: A case study.’ ′ [Online]. Available: https://arxiv.org/abs/1307.7432
29. U.D Butkar and et.al " Modelling and Simulation of symmetric planar manipulator Using Hybrid Integrated Manufacturing." *Computer Integrated Manufacturing Systems*, *29*(1), 464–476.
30. A. Amin et al., ′ ‘Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods,’ ′ Int. J. Inf. Manage., vol. 46, pp. 304–319, Jun. 2019.
31. A. Amin, B. Shah, A. M. Khattak, T. Baker, H. ur Rahman Durani, and S. Anwar, ′ ‘Just-in-time customer churn prediction: With and without data transformation,’ ′ in Proc. IEEE Congr. Evol. Comput., Jul. 2018, pp. 1–6.
32. A. Amin et al., ′ ‘Customer churn prediction in the telecommunication sector using a rough set approach,’ ′ Neurocomputing, vol. 237, pp. 242–254, May 2017.