Customer Churn Prediction using Machine Learning

Conference Paper · December 2022

DOI: 10.1109/ICECA55336.2022.10009093

5 authors, including:

Peddarapu Rama Krishna
VNR Vignana Jyothi Institute of Engineering & Technology
16 PUBLICATIONS **25** CITATIONS

Customer Churn Prediction using Machine Learning

[1]Rama Krishna Peddarapu
Asst.Professor
Department of Computer Science and Engineering
VNR Vignana Jyothi Institute of Engg & Technology
Hyderabad, Telangana, India

[2]Sofia Ameena
Department of Computer Science and Engineering
VNR Vignana Jyothi Inst of Engg & Technology
Hyderabad, Telangana, India

[3]Surepally Yashaswini
Department of Computer Science and Engineering
VNR Vignana Jyothi Inst of Engg & Technology
Hyderabad, Telangana, India

[4]Nadipelli.Shreshta
Department of Computer Science and Engineering
VNR Vignana Jyothi Inst of Engg & Technology
Hyderabad, Telangana, India

[5]Muppidi.PurnaSahithi
Department of Computer Science and Engineering
VNR Vignana Jyothi Inst of Engg & Technology
Hyderabad, Telangana, India

2022 6th International Conference on Electronics, Communication and Aerospace Technology (ICECA) | 978-1-6654-8271-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECA55336.2022.10009093

Abstract-- The varying customer requirements and interests often result in subscription cancellation. Hence, running a subscription business necessitates an accurate churn forecasting model as even a minor change will result in a significant impact. If the seller is not informed that the customer is about to cancel the subscription, no action will be taken to retain them. As a result, this research study attempts to design and develop a churn prediction model to predict a subscription cancellation and provide incentives for that particular subscriber to stay back. This results in significant cost savings and generate an additional revenue source for any online business. The primary goal of this research work is to analyze different models for predicting the active churners with high accuracy. In existing systems, the service providers track down the clients before they leave in order to solve this problem. This study has compared the well-known machine learning techniques to solve the problem and also predict the results in a more accurate way.

Keywords– Customer churn, Customer churn prediction, Decision tree, SVM, Ensemble learning, Random forest, Logistic regression, XGBoost.

I. INTRODUCTION

1.1 CHURN PREDICTION

The pace at which consumers discontinue doing business with a firm is known as customer churn. Churn prediction is the process of identifying customers, who are most likely to stop using a service or cancel their membership. It is an important prediction used in many businesses since getting new customers can sometimes be more expensive than retaining existing ones. People frequently provide the example of cancelling their Netflix or Spotify subscriptions. A significant issue that is frequently related to the current business operation cycle is customer turnover. During the development stage of business lifecycle, the rate of increase of deals and churners is exponential and outweighs the count of churners. However, organizations in their later stages of development place a high priority on reducing the rate of client attrition. Unintentional and intentional factors make up the two different types of client churn.

Accidental churn happens when circumstances alter and stop customers from using the services later on, like when financial constraints become unaffordable for user. Intentional churn can be defined as something that happens when customers wish to choose another company that offers comparable services, like when competitors have better ideas, offer more advanced services, or charge a more affordable amount for a similar service. To address this issue, the telecom service providers must identify these customers before they depart. Machine learning algorithms like decision trees, logistic regression, KNN, Naive Bayes, etc. are used for attaining this goal. The best innovative qualities for predicting client turnover should be the main focus of this research work. For this, the data has been collected and analyzed, and on the basis of that analysis, four well-known machine learning methods have been utilized. Customers must cancel their service based on the subscription contract, tariff plan, contract term, number of services, average call duration in the previous month, and number of outgoing calls per month.

II. EXISTING SYSTEM

1. **Cohort Report:** A cohort report forecasts the number of clients and their rate of attrition through time. A cohort is a collection or group of customers, who made purchases from your business within a specific time period.

2. **Churn by behavior**: Churn may be predicted in addition to being examined by the cohort report by looking at customer behavior. This means that you must watch a certain customer behavior pattern when they use a particular feature or make a particular purchase activity and assess its effects on churn rate.

III. PROPOSED SYSTEM

Developing various systems to recognize data patterns by making use of machine learning algorithms and learning from it without explicit programming is the core characteristic of machine learning techniques. Here, different types of algorithms are used and the one which gives the most accurate result (Random Forest Classifier) is used for recognizing the data patterns.

These algorithms identify certain common consumer behavior patterns among people, who have previously left the business. In order to identify the prospective churners, ML algorithms compare the behavior of present consumers to these patterns.

It will make use of ensemble learning, which may build numerous models and then combine them to get better outcomes. Typically, ensemble approaches yield more precise results than a single model.

IV. LITERATURE SURVEY

In [1], Saran Kumar A,Chandrakala D demonstrated that over the past 20 years, technological improvements have caused the volume of data to expand rapidly. Numerous novel processes and approaches have been developed to process data and extract useful information hidden in raw data. The technique of obtaining important information from data is known as data mining. Numerous data mining techniques have been effectively used in numerous industries. Customers are the most precious resources in any firm because they are known to be the primary source of income. Companies these days are aware that they need to put up a lot of work to both retain their current customers and attract new ones. Churners are individuals who frequently change occupations for a number of causes. A business must be able to accurately predict consumer behavior, spot trends in customer attrition, and manage variables if it wants to lower customer turnover. A job called "churn prediction" employs binary classification to distinguish between those who churn and people who don't.

[2] is a novel method of features engineering and selection that Abdelrahim Kasem Ahmad, Assef Jafar, and Kadan Aljoumaa created employing ML methods on a huge platform of data. This procedure takes the longest to complete due to the vast number of columns and datasets. This kind of study is being done in the telecom sector to help firms make income. Churn prediction is a well-known source of significant revenue for telecom companies. The aim was to produce a system that could find customer attrition at the telecom company SyriaTel. These prediction models must reach greater AUC values. The data available for testing and training is split into 0.7 for training and 0.3 for testing. We opt to employ ten-folds for hyperparameter adjustment and cross-validation. To generate the features for machine learning algorithms, we made use of feature engineering, and selection approaches, effective feature transformation. Additionally, they encountered the problem of imbalanced data. Only 5% of the entries genuinely reflect consumer turnover.

Damandeep Singh, Vansh, and Dr. M. Kanchana, used clean data for the churn prediction model in [3], to prevent any irregularities. Only those features will be chosen such that they are taken into account that actually have an impact on customer attrition. The benefit data and the qualification rating filter are used to choose the features. CRM may boost efficiency, suggest appropriate advancement to a set of potential customers based on comparable designs, and significantly improve corporate marketing campaigns by knowing the essential churn numbers from customer data. The outcome of the model will deliver accurate data that is crucial for the industry. The model's predictions matched the real-world market quickly and accurately. According to Churn, it was quite effective and the customer's previous conduct can reveal a lot about how he will interact with the business going forward. This model analyzes consumer information pertaining to telecom businesses and forecasts whether churn will occur or not.

In [4], Rehan Ullah Khan, Mohamed Tahar Ben Othman, and Amir Ali Mustafa Qamar Due to the fierce rivalry in the tele-communications industry, prediction of client turnover for diverse service providers has become difficult. It is known that it is less expensive to retain an existing client than to acquire new ones. Long-term customers also contribute more revenue than new customers. As a result, it is critical for a business to plan for client loss. It aids in keeping the current clientele by actively lowering turnover. Due to its ability to identify patterns and subsequently learn these patterns for the prediction of new occurrences, machine learning has developed into a prominent paradigm for data mining.

In [5], To create scenarios that use artificial intelligence, Kolla Bhanu Prakash looked into the feasibility of picture identification by artificial intelligence using a machine learning approach called deep learning. This study used a neural network model to achieve deep learning. Neural networks have been used to assess the characteristics of the three-dimensional coordinates that lead to soil categorization and parameters. Soil classification can benefit from using Artificial Neural Networks (ANN).

Understanding the causes of client turnover is crucial for any business, as Kriti says in [6] presentation. The price a customer is offered, the benefits they receive, the length of time the customer has a relationship with the business, etc., are all common elements across several market sectors. Regression modeling has been utilized in earlier works to understand consumer behavior. Clustering has been utilized in certain research articles to divide up client groups with comparable habits. One of those research forecasts client attrition using data mining techniques: A Review on Customer Churn Prediction in Telecommunication Using Data Mining Techniques (S. Babu1, Dr. N. R. Ananthanarayanan). In this research study, the telecommunications business is specifically discussed and the elements that affect it are categorized.

Awais Adnan, Adnan Amin, Feras Al-Obeidat, Babar Shah, Jonathan Loo, and Sajid Anwar proposed a mixed neural network approach for CCP which was used on a CRM dataset from the American Telecommunications Corporation in [7].To construct the CCP model, they used a

technique that combined an artificial neural network with a self-organized map. The ANN was used to remove unrepresentative data from the training set, hence reducing the amount of data required. The first step's output is next fed into the SOM, which creates a prediction model.. The findings demonstrate that a self-organized map and artificial neural network combination works more accurately than a single neural network. As observed, samples from the training set are lost when the first method of data reduction and filtering is used.

In [8], Pronay Ghosh recommended The issue is focused on the banking industry, where a bank wishes to forecast a client's churn based on past data from that customer. By churn, we mean that the bank is attempting to forecast whether a client would default in the upcoming quarter based on their past credit behavior. The customer retention for a bank's product, savings accounts, is something the bank wishes to handle. The bank wants you to find clients who frequently have negative balances. You have information about the customers, including their demographics, age, and bank transactions. From a bank's perspective, it's crucial to keep up with business and client connections. Additionally, if someone could be identified as a defaulter, simple precautions may be taken to prevent similar violations from occurring.

In [9], with the use of the EfficientNetB6 Algorithm, Peddarapu Rama Krishna and Pothuraju Rajarajeswari developed a deep learning approach to solving the issue. The primary justification for using EfficientNEtB6 is that, when compared to other lasting CNNs at ImageNet and other Efficient Net models, the Efficient Net models achieved higher precision and more productivity. In light of this work, they therefore suggested the Deep Learning approach for non-invasive computer vision-assisted clinical diagnosis in medicine.

Total customer turnover, according to Nurulhuda Mustafa1, Lew Sook Ling, and Siti Fatimah Abdul Razak in [10], is the number of customers who switch providers. The churn rate is a metric used in the telecommunications sector to assess market competitiveness. Telecommunications includes phone, internet, and mobile services. For one Internet Service Provider (ISP), losing one of its 20 consumers will result in a 5% decrease in annual income. Daily, the telecom industry loses twenty percent to forty percent of its clients. If pricing and subscription alternatives weren't offered, 83% of Malaysians would transfer cellular providers. However, 66% of people had no trouble ending their subscription to their present service provider.

According to [11] K. Sandhya Rani et al., the data used in this covers entire client data over a length of time. In order to estimate churn in the telecom industry, this research primarily focuses on machine learning methods and approaches that leverage regression and trees. For a good customer churn prediction model, it largely focused on the decision tree algorithm and logistic regression. Many businesses develop their own models for predicting customer turnover since it costs six times more to acquire a new client than it does to retain an existing subscriber. ROC Curves for Voting Classifier show the level of accuracy determined by AUC, which are used to present the results. Despite the stacked model's huge increase in precision, the blended model has a higher F1 and was considered as the best model. The ROC Curves of it are plotted. Then, one approach is to display the false positive rate vs the true positive rate. This method displays the data rapidly as bar plots and pie charts and makes it simpler to understand why customers desire to churn.

In [12], Srinivas Kolli, Peddarapu Rama Krishna and Parvathala Balakesava Reddy, suggested a unique idea for extracting information from web news pages. The suggested approach combines the linked information available from many online sources, clarifies the information, and looks for semantic connections between the page content and the client's research inquiry to provide only the most important information. Future work will make use of placement computations like learn to Rank in order to raise the caliber and importance of the indexed listings. The results indicate that review is high for the classification of the summed-up inquiry while accuracy is lower. This is as a result of the words being channeled using the word recurrence coordinate as a rule. The exactness increases along with the number of words that coordinate precisely. The article's usage of pronouns to refer back to formal people, places, or things when managing them may lead these to be skipped during word recurrence coordination, which results in less accuracy and review esteem for the questions involving formal people, places, or things.

Since the focus of modeling in [13], Ming Zhao et al., is on prediction issues, Such logistic regression must be performed three times, with independent variables drawn from the current period and dependent variables drawn from the lag period. In other words, In order to forecast the dependent variables of the following month, the independent variables of the present month are used. Each month's data are neutralized using the R tool to make the regression coefficients roughly equal, and then divided at random into training and testing sets. The preparation of the training dataset is easily modifiable. According to Kim and Kwon's findings, there is a link between network scale and the loss of Korean telecom subscribers.

In [14], Benjamin Ghaffari and Yasin Osman provided and assessed a variety of methods and tactics on how machine learning may be utilized to forecast attrition, creating value for companies providing subscription-based services.The goal of this study is to increase our understanding of churn prediction. For the proposed model, they compared one base learner, Naive Bayes, with two ensemble learners, XGBoost and Random Forest. The study creates and assesses the developed model iteratively utilizing metrics like accuracy, precision, recall, and F1-score in accordance with the design science process. The findings suggest that machine learning can be used to predict client turnover. It is expected that a balanced training dataset will increase classifier performance, and the findings reveal that ensemble learners outperform single base learners. The goals of this study are to create a binary churn prediction model, evaluate the effects of balancing the training dataset for the model under discussion, and compare various types of algorithms for the model.

To predict client turnover, the author of [15] of Praveen Lalwani used tree-based techniques such as

decision trees, random forests, the GBM tree method, and XGBoost. Comparative analysis showed that, in terms of AUC accuracy, XGBoost beat rivals. In their comparative analysis of the models for forecasting customer attrition, A variety of machine learning models, such as logistic regression, decision trees, support vector machines, and naive bayes, were used by Praveen et al. They then looked at how boosting approaches affected categorization accuracy. The SVM-POLY P. Lalwani et al. surpassed the competition in the findings using AdaBoost. The accuracy of classification can be improved even more by incorporating feature selection strategies like univariate selection and others. Horia Beleiu et al. employed three machine learning techniques to predict customer attrition: support vector machines, neural networks, and Bayesian networks. Principal component analysis is used during the feature selection process to reduce the dimensionality of the data.

III. DESIGN AND METHODOLOGY

In the model building several machine learning techniques are used and the data used for training the models is 'CHURN MODELING '

A. *Data set Description*:

The churn modeling detail dataset (CSV format) has a total of 14 attributes and 1000 records. Dummy values are dropped from the dataset; duplicate records are not considered and are removed before hand passing to the model and NULL values are replaced with the aggregate functionality (Average of the column attribute). Dataset for statistical measurements is collected which contains the churn reports of various customers of a Bank. Fig 1 shows the Count Plots representing the relation with categorical variables.

RowNumber	10000
CustomerId	10000
Surname	2932
CreditScore	460
Geography	3
Gender	2
Age	70
Tenure	11
Balance	6382
NumOfProducts	4
HasCrCard	2
IsActiveMember	2
EstimatedSalary	9999
Exited	2
dtype: int	64

Output Column: Predicts Customer churn rate

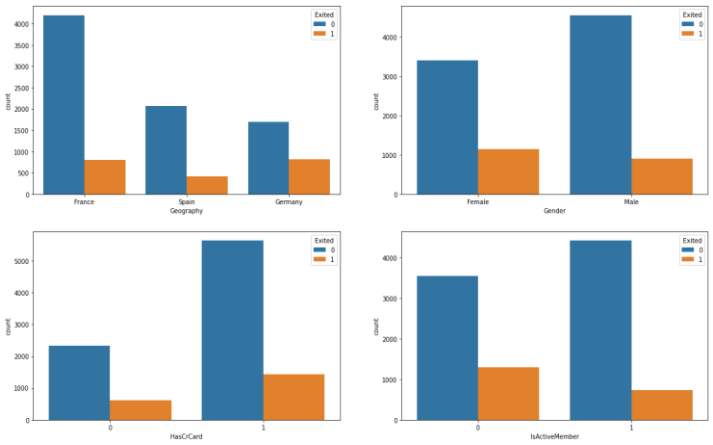

Fig 1. - Count Plots representing the relation with categorical variables

Feature Selection:

Since we prepared several new qualities using feature tools, we will face the issue of choosing features to take into account. Tree-based classification models work well because this is a classification issue with many categorical inputs. The characteristic that divides a large amount of data between estimators will be extracted, starting with a conventional decision tree or random forest. We can choose features by identifying the main factors that account for the majority of variations. Another choice is to use strategies like PCA to reduce dimensionality. Changing PCA features is challenging, therefore even if talking about influencing factors is a key component of our goal, we will choose the first approach.

We experimented with a number of methods, including XGBoost, SVM, Logistic Regression and Random Forest to perform feature selection. The best model for predicting the most important characteristics to classify clients was Random Forest. As a result, this model has been utilized to determine the traits that are most effective for classifying clients. The model has been subsequently assessed by utilizing the crucial elements to comprehend the AUROC curve's volatility. This model has worked well when only the top features were used.

B. *Techniques Used*:

1) Random Forest Algorithm:

With the Random Forest algorithm, various decision trees are constructed using various subsets of data, and the output of each decision tree is then mixed. The bootstrap method is used to create several subsets from the dataset before applying the bagging principle to the random forest. Now that each model has completed its work on these subsets and produced results, the process is known as aggregation. Random forest then combined the results of the various models to produce results based on voting. Fig 2 shows the n-D-Trees-based classification

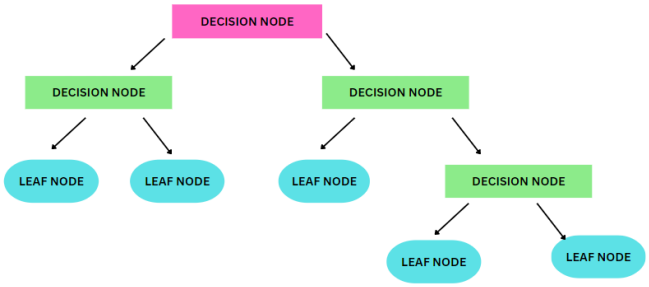

Fig 2. -n-D-Trees-based classification

2) Support-Vector-Machine (SVM):

For classification and regression problems, supervised machine learning procedural methodology called SV-Machine is employed. On the other hand, this algorithm is specifically used for classification issues in real-world problems and scenarios. Each data item is represented as a point in n-D space (n is the number of features) with a value corresponding to a specific position and a unique feature in this approach. To identify the hyper-plane that separates the classes in the supplied data set, the classification technique is then called. In nD-space, it (referring to H-P) can refer to a variety of lines or boundaries, but we need to identify the appropriate boundary line that aids in classifying the dataset's data points. Following classification, the method's optimal boundary is taken into consideration as the S-VM hyperplane. Fig 3 shows the hyperplane of support vector machine.

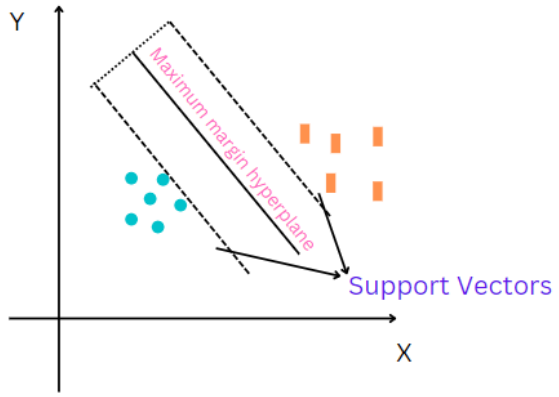

Fig 3. -SV-M[Best Hyper Plane to classify the data]

3) Logistic-[LR]-Regression:

A classification procedure known as logistic regression works by determining probability and then doing classification using the data. The likelihood that a row belongs to a class is determined by the sigmoid function, which takes as input a parameter vector and theta transpose product and outputs a probability between 0 and 1. A row is assigned a class if its probability is less than the threshold, and a different class if its probability is greater than the threshold. Threshold is the value used to determine classification.

4) XGBoost Algorithm

A distributed gradient boosting library called XGBoost was created to be exceptionally efficient, flexible, and portable. It implements machine learning methods using the Gradient Boosting framework. It provides a parallel tree boosting to effectively handle a range of data science tasks.

IV. RESULTS

By the conclusion of the project, we had developed a machine learning model that could analyze customer data to identify whether a client had canceled their subscription or not. This model clearly outperforms more conventional techniques for detecting whether a consumer will leave or stay. Scikit Learn was mainly used for implementing the inbuilt classification algorithms. The main reason for selecting the hybrid model is accuracy of individual models is comparatively less but hybrid model accuracy is higher and hence we consider the output of 4 different classification algorithms to determine the most accurate prediction, with this process of bagging multiple models we got an accuracy of 86.300%. Fig 4 shows the Random Forest scores. The comparison of all the algorithms was done using an ROC curve as shown in Fig 5. The highest accuracy was given by the Random Forest classifier as mentioned in the ROC curve in Fig 6.

	precision	recall	f1-score	support
0	0.87	0.98	0.92	1607
1	0.80	0.40	0.53	389
accuracy			0.86	1996
macro avg	0.83	0.69	0.72	1996
weighted avg	0.86	0.86	0.84	1996

Fig 4. -Random Forest scores

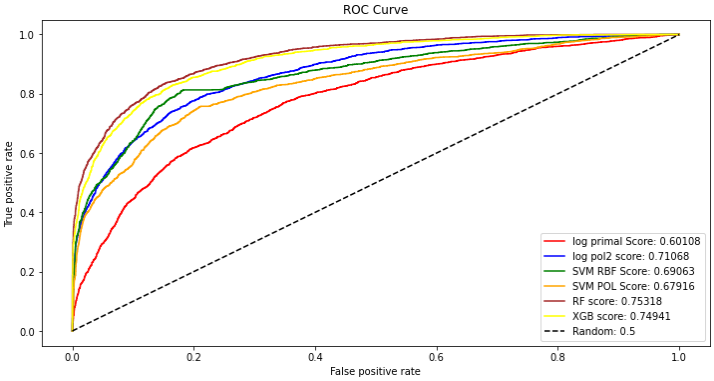

Fig. 5 -Receiver Operator Characteristic (ROC) curve

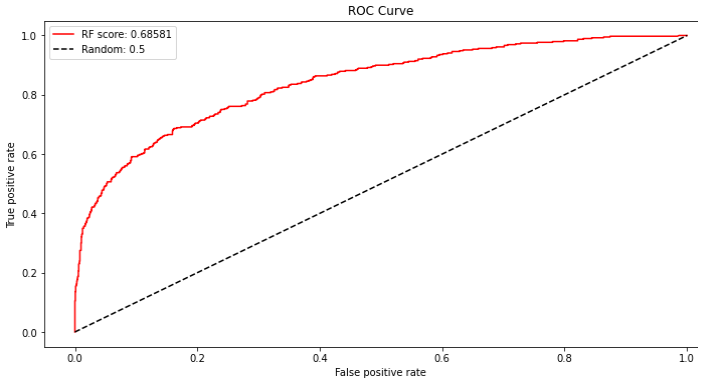

Fig. 6 -Receiver Operator Characteristic (ROC) curve

V. CONCLUSION

In this study, we explore how machine learning can be used to forecast client attrition in a B2B setting. One of the two components of customer churn management is churn prediction, as was already indicated. It would be fascinating to look into which features to employ and how they affect churn prediction in the B2B space. Depending on the variables, it might also be worthwhile to consider prospective retention tactics and the proactive steps that businesses could take to retain clients. Finding out what other techniques may be utilized for feature selection and sampling and how they would affect the results is another interesting research area. The basic scope of developing a machine learning-powered application to forecast client turnover is guided by a predetermined ML project architecture. The models of ensemble learning are providing better accuracy scores but time constraints can be improved in future.

REFERENCES

[1] Saran Kumar A,Chandrakala D "A Survey on Customer Churn Prediction using Machine Learning Techniques" November 2016,International Journal of ComputerApplications,154(10):13-16,DOI:10.5120/ijca2016912237 volume Article number: 28 (2019)

[2] Abdelrahim Kasem Ahmad, Assef Jafar & Kadan Aljoumaa "Customer churn prediction in telecom using machine learning in big data platform" Journal of Big Data, Article number: 28 (2019)

[3] Damandeep Singh , Vansh ,Dr. M. Kanchana, Associate Professor, "Survey Paper on Churn Prediction on Telecom", SRM, India

[4] Nasebah Almufadi , Ali Mustafa Qamar, Rehan Ullah Khan, Mohamed Tahar Ben Othman, "Deep Learning-based Churn Prediction of Telecom Subscribers", International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 12 (2019), pp. 2743-2748

[5] N Lakshmi Kalyani and Kolla Bhanu Prakash, "Soil Color as a Measurement for Estimation of Fertility using Deep Learning Techniques" International Journal of Advanced Computer Science and Applications(IJACSA), 13(5), 2022.

[6] Kriti, "Customer Churn: A Study of Factors Affecting Customer Churn using Machine Learning", Iowa State University Ames, Iowa 2019

[7] Adnan Amin , Feras Al-Obeidat , Babar Awais Adnan , Jonathan Loo , Sajid Anwar, "Customer churn prediction in telecommunication industry under uncertain situation", Center for Excellence in Information Technology, Institute of Management Sciences, Peshawar, 25000 Pakistan. College of Technological Innovation, Zayed University,144534 Abu Dhabi, United Arab Emirates. Computing and Communication Engineering, University West London.

[8] Pronay Ghosh, "Project report on customer churn prediction using supervised machine learning"

[9] Peddarapu Rama Krishna, Pothuraju Rajarajeswari, "Early Detection Of Melanoma Skin Cancer Using Efficient Netb6", International Conference on Advanced Computing and Communication Systems (ICACCS), Vol 1 & pg.01-05,07-Jun-2022

[10] Nurulhuda Mustafa, Siti Fatimah Abdul Razak, "Customer churn prediction for telecommunication industry": A Malaysian Case Study

[11] K.Sandhya Rani,Shaik Thaslima ,N.G.L Prasanna R.Vindhya ,P. Srilakshmi ."Analysis of Customer Churn Prediction in Telecom Industry Using Logistic Regression International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ,Volume-9,July 2021, https://doi.org/10.21276/ijircst.2021.9.4.6

[12] Srinivas Kolli, Peddarapu Rama Krishna and Parvathala Balakesava Reddy, A Novel NLP And Machine Learning Based Text Extraction Approach From Online News Feed, ARPN Journal of Engineering and Applied Sciences, Vol. 16, No. 6, pg no's:679-685,Mar-2021

[13] Ming Zhao ,Qingjun Zeng , Ming Chang , Qian Tong , and Jiafu Su "A Prediction Model of Customer Churn considering Customer Value: An Empirical Research of Telecom Industry in China", Research Center for Economy of Upper Reaches of the Yangtze River, Chongqing Technology and Business University, Chongqing 400067, China

[14] Benjamin Ghaffari & Yasin Osman, "Customer churn prediction using machine learning Benjamin Ghaffari & Yasin Osman, A study in the B2B subscription based service context"

[15] Praveen Lalwani, Manas Kumar Mishra, Jasroop Singh Chadha, Pratyush Sethi, "Customer churn prediction system: a machine learning approach"