Machine Learning Implementation in Electronic Commerce for Churn Prediction of End User

Neha Sharma, Aayush Raj, Vivek Kesireddy, Preetham Akunuri

Abstract*: *Client conduct can be addressed from numerous points of view. The client's conduct is distinctive in various circumstances will give his concept of client conduct. From an overall viewpoint, the conduct of the client, or rather any individual around there, is taken to be irregular. When noticed distinctly, it is regularly seen that the future conduct of an individual can rely upon different variables of the current circumstance just as the conduct in past circumstances. This examination establishes the forecast of client beat, for example regardless of whether the client will end buying from the purchaser or not, which relies upon different components. We have chipped away at two sorts of client information. To start with, that is reliant upon the current elements which don't influence the past or future buys. Second, a period arrangement information which gives us a thought of how the future buys can be identified with the buys before. Logistic Regression, Random Forest Classifier, Artificial neural organization, and Recurrent Neural Network has been carried out to find the connections of the agitate with different factors and order the client beat productively. The correlation of calculations demonstrates that the aftereffects of Logistic Regression were somewhat better for the principal Dataset. The Recurrent Neural Network model, which was applied to the time-arrangement dataset, additionally gave better outcomes.

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

I. INTRODUCTION

The part of AI is an immense branch and has spread quickly to every one of the fields lately. Utilizing constant information to get close to precise outcomes and thus, arranging answers for a given issue is far reaching and is being advanced by each other individual. Client Behavior Modeling is characterized as discovering connections of the conduct with different variables that address the acquisition of the client and finding on the basic focuses on which the organization needs to attempt to fulfill its clients [1]. Client conduct, albeit thought about irregular, can be speculated or anticipated, when distinctly noticed. On the off chance that there is a path for a dealer to know which of its clients are troubled or are needing to leave, they can be given uncommon consideration, and maintenance arrangements can be planned. In spite of the fact that there are approaches to apply different models to discover client practices, it is as yet a difficult errand on the grounds that the conduct is addressed contrastingly in various settings. Additionally, each organization has its approach to manage the clients and would need to observe its essential principles to fulfill the clients. In this way, ensure that the model meets the organization's necessities can be an exceptionally infuriating undertaking. In the space of client investigation, it is smarter to do prescient demonstrating as opposed to the uninvolved speculating of client conduct. The Prediction Results permits advertisers and maintenance specialists to deal with future conduct as opposed to the detached instructed conjectures of the past. The upside of this is that advertisers can zero in on explicit clients with explicit methodologies. On the off chance that we take particulars, if an organization can foresee which of its clients are well on the way to agitate, at that point they can move toward those clients in an unexpected manner in comparison to they were arranging prior and ideally foster procedures to hold them. This, along these lines, fills in as possible income to the organization. Along these lines, foreseeing client stir can be considered as an arrangement issue, and we can apply different characterization calculations to anticipate the agitate [2]. Forecast can be made by considering, the clients present buys like what thing has the client been generally keen on the past and what administrations would the client wish to proceed later. Likewise, a record of how reliable the client has been in buying the administrations of the organization can help in prediction. The part of AI is an immense branch and has spread quickly to every one of the fields lately. Utilizing continuous information to get close to precise outcomes and thus, arranging answers for a given issue is boundless and is being advanced by every other individual. Client Behavior Modeling is characterized as discovering relationships of the conduct with different variables that address the acquisition of the client and finding on the basic focuses on which the organization needs to attempt to fulfill its clients [1]. Client conduct, albeit thought about arbitrary, can be speculated or anticipated, when acutely noticed.

Manuscript received on May 19, 2021.
Revised Manuscript received on May 25, 2021.
Manuscript published on May 30, 2021.
* Correspondence Author

Neha Sharma*, Assistant Professor, Department of Information Technology, Manipal University Jaipur (Rajasthan), India. nehav.sharma@jaipur.manipal.edu

Aayush Raj, Department of Information Technology, Manipal University Jaipur (Rajasthan), India. Email: aayush.189302085@muj.manipal.edu

Vivek Kesireddy, Department of Information Technology, Manipal University Jaipur (Rajasthan), India. Email: vivek.189302171@muj.manipal.edu

Preetham Akunuri, Department of Information Technology, Manipal University Jaipur (Rajasthan), India. E-mail: akunuri.189302176@muj.manipal.edu

Retrieval Number: 100.1/ijsce.F35020510521
DOI: 10.35940/ijsce.F3502.0510521
Journal Website: www.ijsce.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication
© Copyright: All rights reserved.

If there is a path for a vender to know which of its clients are troubled or are needing to leave, they can be given uncommon consideration, and maintenance arrangements can be planned.

Even though there are approaches to apply different models to discover client practices, it is as yet a difficult undertaking in light of the fact that the conduct is addressed contrastingly in various settings.

Likewise, each organization has its approach to manage the clients and would need to adhere to its essential principles to fulfill the clients. Along these lines, ensure that the model meets the organization's necessities can be an annoying errand. In the space of client investigation, it is smarter to do prescient demonstrating instead of the aloof speculating of client conduct. The Prediction Results permits advertisers and maintenance specialists to chip away at future conduct instead of the aloof instructed theories of the past. The benefit of this is that advertisers can zero in on explicit clients with explicit systems. If we take points of interest, if an organization can foresee which of its clients are destined to stir, at that point they can move toward those clients in a fairly unexpected manner in comparison to they were arranging prior and ideally foster systems to hold them. This, along these lines, fills in as likely income to the organization. Thus, anticipating client stir can be considered as a grouping issue, and we can apply different order calculations to foresee the agitate [2]. Forecast can be made by considering, the clients present buys like what thing has the client been generally intrigued by the past and what administrations would the client wish to proceed later on. Likewise, a record of how steady the client has been in buying the administrations of the organization can help in forecast.

II. RELATED WORK

The Customer beat forecast is extremely questionable to foresee; it has the tension on business fulfillment market. Zuhao et. Al. [3], executed Support Vector Machine on the Dataset of remote beat industry to anticipate the improved exactness in contrast with conventional techniques. The exploration utilized improved one-class support vector machine and looked at the exhibitions of the distinctive part. The outcomes introduced the 87.15% exactness, which was contrasted with ANN, Decision tree, and Naïve Bayes. Backing Vector Machine outflanks the customary methodologies since it requires less measure of time for preparing and testing. The expectation by Support Vector Machine is better in contrast with ANN, Decision tree, Logistic Regression, and guileless Bayes [4]. Xia and Jin carried out SVM on telecom information and contrasted and the few methodologies. The examination showed that the agitate rate is more among the clients who have solid charge eagerness, long versatile help, impressive client care, and standard call and message use. The Dataset of media transmission and banking industry has various similitudes which show that SVM can be applied to the financial area moreover. The Data hotspots for forecast of clients stir incorporates the historical backdrop of the uses, bills, and client administrations, however Lian Yan et. Al. [5], utilized Call Detail Record. The examination utilizes the postponed Call Detail Record as essential to appraise client conduct. The Call Detail record information was utilized to extricate the calling joins and distinguishes a few distance measures. The Data recovered from the calling joins was given as contribution to the neural organization, and the worthy exactness is accomplished. The Calling connections can be utilized for promoting and offer in a particular local area.

More highlights can be processed from the current highlights to improve the expectation of agitate clients. Huang et al. [6], registered new Dataset from the current information, which incorporates call subtleties, record and bill data, line data, installment data, grumble and administration data. The aggregate Dataset 827,124 client's information was chosen from this present reality data set of Ireland. The testing and preparing dataset has 13,562 churners and 400,000 non-churners information where 738 highlights address every client. Seven techniques were applied to anticipate the agitate. The examination gave correlation among the procedures, and correlation between separated list of capabilities and existing list of capabilities. The new highlights dataset introduced preferable outcomes over existing Dataset.

Vafeiadis et al. [7] directed the investigation utilizing five techniques for AI to foresee client stir. In the main section of the exploration, the models were executed on the Dataset gathered from the UCI Machine Learning vault, public space dataset, and further, the models were assessed utilizing cross-approval techniques. In the second portion of examination, the model exhibition was improved utilizing boosting calculations. To recognize the most effective component blend, Monte Carlo Simulation was performed on each strategy. The outcomes showed that the supported model outcomes outflanked the straightforward (non-helped) model outcomes. SVM-POLY utilizing AdaBoost performed better among every one of the models. Ahmad et al. [8] zeroed in on stir in the media transmission area, where agitate influences the income of the organization. The examination expected to anticipate the beat of clients utilizing AI methods on huge information. The creators utilized the Area Under Curve standard measure to quantify the presentation of the model, the worth produced by the Area Under Curve is 93.3%. The model was created and tried on huge Dataset given by SyriaTel Telecom Company, which contained the data of clients more than nine months. XGBOOST calculation performed better among four AI strategies. Routh et al. [9] proposed a model to conquer the vulnerabilities, for example, unstable conduct of the client and expanding culmination hazard. The model figures the conceivable danger and recognizes the connection among hazard and client conduct. The model utilized information from the accommodation business, and models give 20% improved precision in contrast with the current traditional models. The consequences of the examination assist merchant with understanding the justification stirring and produce new designs to manage the conceivable impending agitates.

III. METHODS AND MATERIALS

The examination was led on two distinctive datasets. The main Dataset is the record of administrations bought by clients of an anecdotal telecom organization and the second Dataset is the record of procurement history of the clients of an online business site.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

The Dataset is gathered from kaggle.com [10] containing subtleties of a telecom organization. The sections in the Dataset incorporate vital qualifications of the clients just as the administrations bought by every client alongside the agitate variable (0 or 1), for example if the client will stir.

The second Dataset contains the online retail of a web based business site from the UCI Machine Learning Repository [11].

It has records of the different buys made by the clients for one year from 1-12-10 to 9-12-11. The information is not quite the same as the principal Dataset as it is the time-arrangement information.

IV. METHODOLOGY

The examination proposed models to foresee if a specific client will beat. We have utilized different AI ideas and calculations for forecast. The calculations have additionally been utilized to test if the exactness of characterizing the beat clients is likewise kept up on the information on which the model isn't prepared.

A. Logistic Regression

Logistic Regression utilizes a capacity called strategic capacity at the center of the strategy. The strategic Regression utilizes the sigmoid capacity. It applies the capacity on the condition of straight Regression which gives us the outcomes as the likelihood of an event. The state of the bend is S-like which can fit any genuine number worth to a worth somewhere in the range of 0 and 1. It computes the worth by the accompanying equation.

1 / (1 + e^-value)

Where e is the base of the natural logarithms,

Moreover, value is the actual numerical value that is to be transformed.

B. Random Forest Classifier

Random Forest calculation is a managed order AI calculation. It haphazardly makes the woods with a few choice trees. Each tree makes various choices and predicts values. In Random Forest classifier, a more critical number of trees will in general give more exact outcomes. Random Forest calculation deals with the issue of missing qualities and doesn't overfit the model when we have a greater number of trees present. There are two phases in the irregular forests calculation, one is arbitrary timberland creation, and the other is to anticipate the irregular woods classifier made in the main stage. In Random Forest classifier, the worth of the yield variable is the lion's share esteem anticipated by every one of the trees.

C. Artificial Neural Networks

Artificial neural network (ANN) is an AI calculation which chips away at the possibility of a human cerebrum. By and large, an ANN comprises of a progression of interconnected hubs that change, and the data is gone through a progression of layers. An ANN has three essential layers, an info layer, a secret layer and a yield layer. Even though as per the prerequisites, different altered renditions of the models comprising of an enormous number of covered up layers are being utilized nowadays.

D. Recurrent Neural Networks

These are the neural organizations which are applied for time arrangement examination. At the point when we have information that relies upon past conduct, RNNs are utilized. They are not quite the same as the straightforward fake neural organizations such that they permit the progression of data starting with one cell then onto the next. This data is the portrayal of the information data of the past layer.

The information is preprocessed to eliminate unessential data from the data set. The all out information is changed over into the numeric information. The invalid qualities in the Dataset are dealt with, and the information is then scaled to make the units of different segments immaterial. The information and the objective yield is indicated, and information is spat into the preparation set and the test set. Figure 1 shows the exploration interaction.

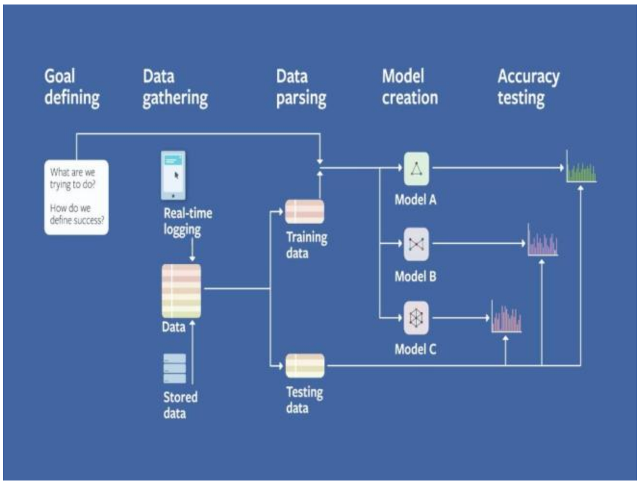

Figure 1: Flow chart indicating the research process

For the online retail/E-business dataset, the agitate is determined utilizing the mean standard deviation of the client and the individual deviation from the mean of a solitary buy. On the off chance that the individual deviation is more prominent than the mean standard deviation, the client is said to have stirred in the middle of the two buys and subsequently a column is affixed in the middle of making the worth of the agitate trait 'valid' and making the buy things and unit cost as 0. The info information is then made by adding every one of the past acquisitions of the client comparing to its most recent buy in the objective variable, and afterward the information is parted into the preparation set and the test set. Different diagrams depicting the relationships in the information are built to give us a superior comprehension of what we are managing.

V. RESULT

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

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

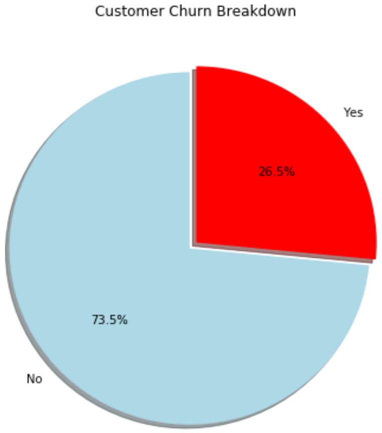

Figure 2: Churn Breakdown of Dataset1

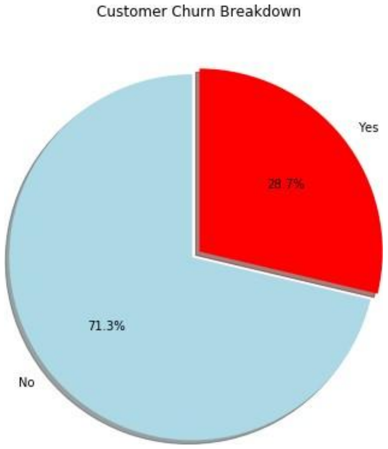

Figure 3: Churn Breakdown of Dataset2

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

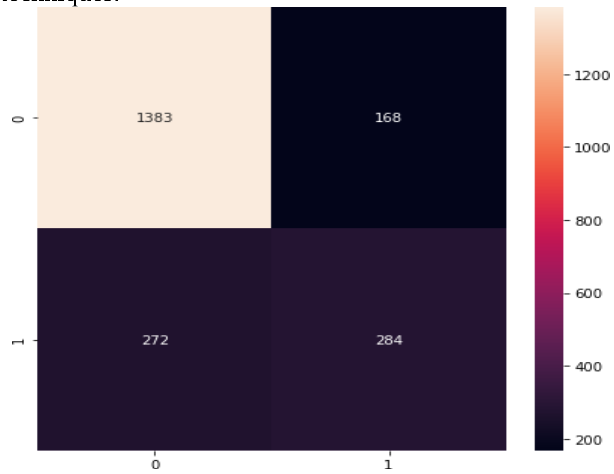

Figure 4: Confusion Metrics of Logistic Regression

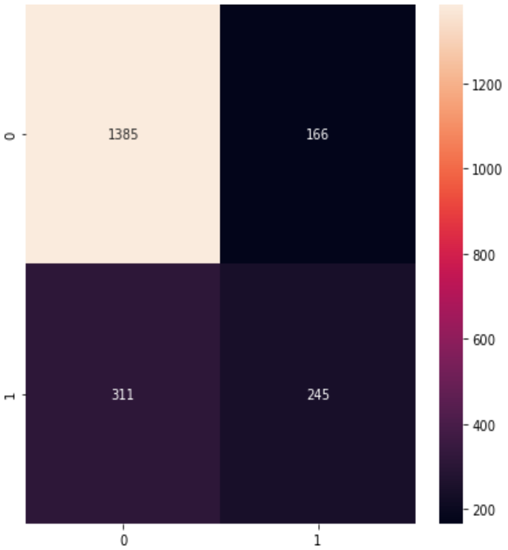

Figure 5: Confusion Metrics Random Forest

Table 1: Confusing Metrics of Logistic Regression, Random Forest and artificial Neural Network.

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

For the E-Commerce dataset, Recurrent Neural Network model is applied usingdifferent activation functions under the concept of Long Short Term Memory. After applying 10-fold cross-validation and grid search for specific parameters, we calculate the results for the maximum accuracy achieved. Figure 6 represents the correlation between customer ID and invoice No. Figure 8 and Table 4 present the confusion matrix of the recurrent neural network.

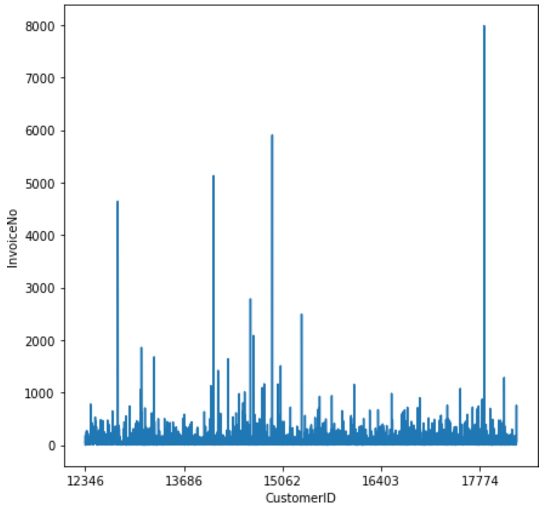

Figure 6: Correlation between Invoice and Customer ID

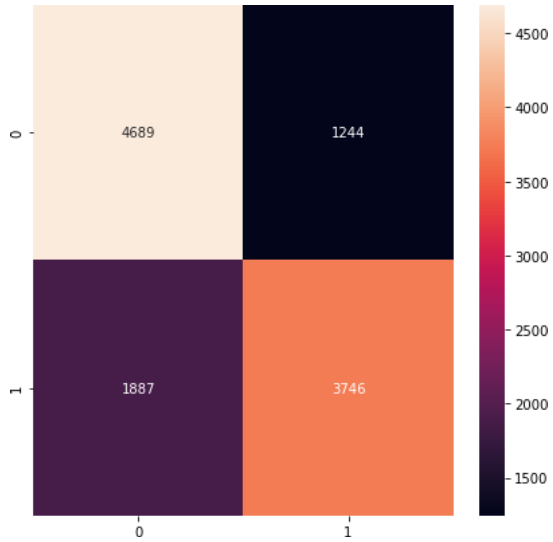

Figure 7: Confusion Metrics of Recurrent Neural network

Table 2: Confusion Metrics of Recurrent neural Network

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

VI. CONCLUSION

The examination acquaints models with foresee if a specific client will agitate. The main Dataset is the record of administrations bought by clients of an anecdotal telecom organization and the second Dataset is the record of procurement history of the clients of an internet business site. Different Machine Learning AI ideas and calculations are applied for forecast and have additionally been utilized to test if the exactness of characterizing the beat clients is likewise kept up on the information on which the model isn't prepared.

On looking at the outcomes and investigating, it is seen that the consequences of Logistic Regression are somewhat better compared to other characterization calculations in this situation. The outcomes demonstrate that the clients who bought costly administrations and additionally were senior residents are the ones who are well on the way to agitate. The outcomes additionally portray that the future acquisition of the client firmly relies upon the past buys, for clients buying more things in a single buy made their next buy after a moderately prolonged stretch of time. Additionally, the clients who make some reliable memories distinction in two back to back buys are well on the way to remain.

In view of these outcomes we can infer that various organizations can apply this model to their client information to sort out who is destined to stir and work on planning the maintenance arrangements so as not to endure misfortunes.

Different models can be fabricated and attempted, particularly the profound learning models with an alternate mix of layers and initiation capacities. Figuring the beat rate is useful, however on the off chance that conceivable, the following most conceivable acquisition of a client can be anticipated so the organization can recognize the client's premium rapidly and pitch those items to the client, where he/she is exceptionally intrigued.

REFERENCES

1. S. B. Borah, S. Prakhya, and A. Sharma, “Leveraging service recovery strategies to reduce customer churn in an emerging market,” J. of the Acad. Mark. Sci., vol. 48, no. 5, pp. 848–868, Sep. 2020, doi: 10.1007/s11747- 019-00634-0.
2. M. Panjasuchat and Y. Limpiyakorn, “Applying Reinforcement Learning for Customer Churn Prediction,” J. Phys.: Conf. Ser., vol. 1619, p. 012016, Aug. 2020, doi: 10.1088/1742-6596/1619/1/012016.
3. Y. Zhao, B. Li, X. Li, W. Liu, and S. Ren, “Customer Churn Prediction Using Improved One-Class Support Vector Machine,” in Advanced Data Mining and Applications, vol. 3584, X. Li, S. Wang, and Z. Y. Dong, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 300–306.
4. G. Xia and W. Jin, “Model of Customer Churn Prediction on Support Vector Machine,” Systems Engineering - Theory & Practice, vol. 28, no. 1, pp. 71–77, Jan. 2008, doi: 10.1016/S1874-8651(09)60003-X.
5. Lian Yan, M. Fassino, and P. Baldasare, “Predicting customer behavior via calling links,” in Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., Montreal, Que., Canada, 2005, vol. 4, pp. 2555–2560, doi: 10.1109/IJCNN.2005.1556305.
6. B. Huang, M. T. Kechadi, and B. Buckley, “Customer churn prediction in telecommunications,” Expert Systems with Applications, vol. 39, no. 1, pp. 1414–1425, Jan. 2012, doi: 10.1016/j.eswa.2011.08.024.
7. [7] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. Ch. Chatzisavvas, “A comparison of machine learning techniques for customer churn prediction,” Simulation Modelling Practice and Theory, vol. 55, pp. 1–9, Jun. 2015, doi: 10.1016/j.simpat.2015.03.003.
8. A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer churn prediction in telecom using machine learning in big data platform,” J Big Data, vol. 6, no. 1, p. 28, Dec. 2019, doi: 10.1186/s40537-019-0191-6.
9. P. Routh, A. Roy, and J. Meyer, “Estimating customer churn under competing risks,” Journal of the Operational Research Society, vol. 0, no. 0, pp. 1–18, Aug. 2020, doi: 10.1080/01605682.2020.1776166.
10. “Telco Customer Churn | Kaggle.” https://www.kaggle.com/blastchar/telco-customer-churn (accessed Sep. 15, 2020).
11. “UCI Machine Learning Repository: Data Sets.” https://archive.ics.uci.edu/ml/datasets.php (accessed Sep. 15, 2020).

AUTHORS PROFILE

Neha Sharma, is an Assistant Professor in department of Information Technology, Manipal University Jaipur, India. She is currently pursuing her PhD. in Network Security from Manipal University Jaipur, India. She has an overall experience in industry and academics of more than 10 years. She has many National and International publications to her credit.

Aayush Raj, is currently pursuing his bachelor’s in Information Technology at Manipal University Jaipur, Jaipur, India. He has also worked on various projects based on Machine Learning, Web Development, and App Development. His areas of interest include Data Analytics, Data Science, Machine Learning. He was also Chair of one of the most renowned technical club IEEE SB MUJ he was very disciplined and handled all his responsibilities with utmost diligence during his tenure.

Vivek Kesireddy, is currently pursuing his bachelor’s in Information Technology at Manipal University Jaipur, Jaipur, India. He has also worked on various projects based on Machine Learning, IoT, Artificial Intelligence, and Augmented Reality. His areas of interest include Data Analytics, Data Science, Machine Learning, and IoT-based projects.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

Preetham Akunuri, Being a student in Manipal University Jaipur, he is currently pursuing his Bachelor's in Information technology. He has designed model and software for an app and worked on projects using Convolutional Neural Networks. Computer graphics, Data science, Machine learning and Programming languages are his areas of interest.