A Machine Learning Model for Personalized Tariff Plan based on Customer’s Behavior in the Telecom Industry

Lewlisa Saha[1], Hrudaya Kumar Tripathy[2], Fatma Masmoudi[3], Tarek Gaber[4]
School of Computer Engineering, KIIT, Deemed to be University, Bhubaneswar, India[1, 2]
College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Alkharj, 11942, Saudi Arabia[3]
School of Science, Engineering, and Environment, University of Salford, Manchester, UK[4]

Abstract—**In the telecommunication industry, being able to predict customers’ behavioral pattern to successfully design and recommend a suitable tariff plan is the ultimate target. The behavioral pattern has a vital connection with the customers’ demographic background. Different researches have been done based on hypothesis testing, regression analysis, and conjoint analysis to determine the interdependencies among them and the effects on the customers’ behavioral needs. This has presented us with ample scope for research using numerous classification-based techniques. This work proposes a model to predict customer’s behavioral pattern by using their demographic data. This model was built after investigating various types of classification-based machine learning techniques including the traditional ones like decision tree, k-nearest neighbor, logistic regression, and artificial neural networks along with some ensemble techniques such as random forest, adaboost, gradient boosting machine, extreme gradient boosting, bagging, and stacking. They are applied to a dataset collected using a questionnaire in India. Among the traditional classifiers, decision tree gave the best result of 81% accuracy and random forest showed the best result among the ensemble learning techniques with an accuracy of 83%. The proposed model has shown a very positive outcome in predicting the customers’ behavioral pattern.**

Keywords—Customer behavior; data analytics; ensemble learning; machine learning; telecommunication industry

I. INTRODUCTION

Telecommunication industry is one of the world’s largest services providing industry and is known for having the highest number of customers. Customers are the most vital entity of any business platform, and the telecom sector is no exception to that. Bringing in new customers, understanding their needs and requirements, dividing the customers into proper groups to be able to offer the right product or service, gain their loyalty and retain them in the long term are the targets of any customer-centric business. As for the telecommunication sector, retaining customers has become a necessity to survive in the competitive market.

In [1], the authors have discussed the framework containing the development over the recent years in the Indian mobile communication sector. Over 23 years, the Indian telecom sector has seen a fundamental shift in communication modes from fixed line to mobile, making India the world's second-largest telephone market. The exponential rise of smart phones and the Internet in 2013 marked the beginning of the era of convergence of two technologies: mobile and the Internet. In India, mobile penetration (density) has increased from 4% in 1995, when mobile telephony was first introduced, to 88.50% in 2019 (with 1165.46 million subscribers). The low density of mobile telephony in rural India remains an issue. This high mobile density in urban India is owing to higher per capita income, education, and network externality (as compared to rural India), resulting in a single user owning numerous subscriber identity modules (SIM). The Indian population's total teledensity has surged up to 90.1% as of March 2019 [2]. In this paper, authors have given a deep insight into the telecommunication industry and their relationship with their customers and how it has developed over the years. The necessity to maintain existing consumers has grown over time as the number of new entering customers decrease. The urban India had a teledensity of 159.66% in March 2019, while rural India had a teledensity of 57.5%. This demonstrates the significant differences in people's socioeconomic position based on where they live. For example, metro cities like Delhi had a teledensity of 238.57%, Kolkata had a teledensity of 165.51%, and Mumbai had a teledensity of 165.62 %, whereas the states of Bihar had a teledensity of 59.95 %, Assam had a teledensity of 68.81 %, and so on. As a result, the location is crucial in determining the demands and requirements of customers. This latter compels us to go on to the next level in customer relationship management: customer segmentation.

In the telecommunication sector, geographic location, socio-economic background, age group, gender identity, and so on have a profound impact on the selection of tariff plans by their respective customers. A significant variation in the selection of tariff plans can be observed between customers belonging to a metropolitan city to that of a rural area. Hence, this invites scope for better customer segmentation and customization of services. To address this existing generalized tariff system in the Indian telecommunication industry, our work focuses on predicting customer behavioral pattern that would lead to better-personalized tariff plans by designing a flexible tariff plan for the customers in accordance with their needs.

Keeping those aspects in the hindsight, we propose to analyze the customers’ demographic data using different classification-based machine learning techniques such as decision tree, logistic regression, k-nearest neighbor, and artificial neural network among traditional classifiers and random forest, Adaboost, extreme gradient boosting, gradient boosting machine and stacking among ensemble learning techniques. This study is used to train the model into predicting the customers' behavioral pattern in the form of call duration and data consumption. The result of the proposed model is at par with the desired outcome. With a larger amount of data the performance of the model could be improved and could be enhanced significantly.

The rest of the paper is organized as follows: Section II shows the related works in the research area. Section III discusses the machine learning techniques used in the proposed research model. Section III also provides a detailed idea about the dataset used and the methodology followed in the research. The results that are obtained by the research are detailed in Section IV, Section V highlights the discussion of the present work and the whole work has been concluded in Section VI.

II. Related Work

The Telecom sector offers services that broadcast voice, data, sound, text, video, and other signal forms. It also consists of wired and wireless activities. Nowadays, when the Internet is the dominating factor, the telecommunication industry has grown tremendously. In order to meet the increasing demand for data consumption, the telecommunication service industry needs to rise so that it can keep up with the high usage of social media platforms. Over the years, authors across the globe have worked with different datasets belonging to the telecommunication sector spanning various countries. Hence, we try in our work to analyze and assimilate a few of the previous works based on tariff plan design in the telecommunication sector. Authors in (Nkordeh et al., 2017) [3] discussed the growth of GSM in the Nigerian telecommunication industry over the period time of 2001 to 2016. They performed a detailed market analysis of the annual growth based on subscribers, revenue growth, and market penetration and how it caused the loss of subscribers. In (Shao et al., 2017) [4], the authors have performed a study which adopts MTFPI (Malmquist total factor productivity index) and data envelopment analysis to conclude that because of their high acceptance of technology developments, the telecommunication service industries had a higher rate of productivity growth. In [5] the author intended to offer a CRM optimization scheme for the telecommunications industry utilizing business intelligence (BI). In the first phase of their work, they conducted a literature survey for all available assignments in the Scopus and web of science. New concepts were introduced to the current work in its second phase and the third phase, the new approach was implemented by a Brazilian telecommunications operator. Authors in [6] used hypothesis testing to study the impact of business intelligence on CRM from the perspective of the employees of telecommunication companies in Oman. Tong et al. discussed a way of increasing customer loyalty through the use of net promoter data mining (NPS) which depended on information gained on decision tree model and k-means clustering [7]. They attempted to identify the reasons for customer loyalty and conclude the connection between the NPS and the financial performance of the customer groups. Jose et al. sought to determine the important parameters affecting the performance of the telecommunication contact center which was vital to customer satisfaction [8]. A case study on Big Data Analytics in the incoming and external contact center was undertaken by the authors by using principal component analysis (PCA), factor analysis, regression analysis, and cross-tabulation. Bahri-Ammari and Bilgihan proposed a framework that uses hypotheses testing for examining the relationship between customer satisfaction and loyalty and customer retention [9]. In reconciling customer satisfaction among mobile telecommunication providers, Newton and Ragel reviewed the potential relational linkages concerning customer loyalty with the help of correlation and regression analysis [10]. The authors intended to detect the degree, relationships, and impact of customer satisfaction and reliability. In their article, Belwal and Amireh have been focused on the Omani telecommunications market and have evaluated the service quality of two main Omani telecom companies: Omantel and Ooredoo [11]. This was tested on customer attitudinal loyalty to see how the five SERVQUAL factors influenced it, by using a machine learning technique called structural equation modeling.

In a variety of industries, including telecom, understanding how consumers make tariff decisions is a key issue. Most telecommunication service providers assign customers with a variety of mobile plans from which they can choose one that best suits their monthly needs. As mentioned in [12] due to many service providers, the telecom business has become extremely competitive. Their study proposed an association-based rule mining technique to help telecom operators select optimal recharge combo deals that are appropriate for their clients. In another work, Gerpott and Meinert used hypotheses testing on data obtained from the German subsidiary of a large multinational MNO to analyze the degree of tariff plan misfit depending on the socio-demographic features [13].

Consumer overspending in mobile plan selections may result in higher revenues for telecom service providers. However, it could ultimately lead to less satisfied customers and are thus more prone to churn, which is a major problem for telecom service providers. As a result, enhancing customer happiness by encouraging them to make better choices is a superior long-term approach. Author in [14] studied the impact of launching of JIO on the Indian telecommunication market which led to a one-dimensional design of the tariff plans in the long term and resulted in the need for getting a better understanding of the customers' requirements. Bibim and Ramanathan tried to use conjoint analysis to determine the best combination of data, voice, and SMS for postgraduate and undergraduate students, as well as to investigate how they use it [15]. The suitable effects on customer loyalty and customer relationship management in the Indian telecommunications sector had been shown by [16] with the help of Exploratory Factor Analysis and Regression Analysis.

The most challenging issue is the difficulty to comprehend consumers' tariff options. Although customers aspire to save money by selecting a cost-cutting strategy, they are only partially focused on resolving the issue. As a result, rather than making tariff decisions based on predicted demand, they tend to simplify the process by employing cognitive shortcuts, such as heuristic thinking, which leads to systemic tariff biases. Ignoring how consumers utilize heuristic thinking can lead to failures in consumer protection and market control policies. The authors in [17] have investigated the quality of decision making in a lab setting where consumers were presented with a limited number of mobile phone plan options. However, they have not dealt with varying degrees of tariff complexity as well as unknown usage. According to Jin et al., the goal of their study has been to check whether consumers employ a specific type of heuristic reasoning, known as salience-based decision making when choosing mobile plans and they preferred the approach of hypotheses testing [18]. Based on market segmentation theory, Shuochen et al. have introduced a matching model that links tariff packages and consumers' usage behavior (e.g., total minutes utilized, data consumption, etc.) [19]. Gerpott and Meinert have looked at the relationship between outgoing mobile voice minutes and monthly mobile internet data traffic in a sample of 11,614 residential postpaid members over 25 months, from October 2011 to October 2013 by using regression analysis [20]. Díaz examined the factors determining customer satisfaction and customer loyalty in the Peruvian mobile market [21]. Multinomial logit and the GSEM estimates have showed, based on a survey of 1259 customers, how customer satisfaction determinants can be assessed when satisfaction has been measured using ordered categorical data.

Ascarza et al. have examined the efficacy of retention campaigns with the help of a large-scale field test where recommendations for plans have been provided to some customers and not others [22]. Lee et al. have used conjoint analysis in their research to study the effects of B2B service introduction in the Korean telecommunication industry [23]. In their research, Garcia-Marioso and Suárez have examined the reasons that motivate consumers to transfer mobile operators by using logit model over data from a longitudinal survey of 4110 Spanish mobile users conducted over the period of 2015 and 2016 [24]. Xu et al. presented a system for predicting customer turnover that employed an ensemble-learning technique, including stacking models and soft voting [25]. Bachan and Gaber performed churn prediction by using decision tree, logistic regression, and support vector machine [26]. Capponi et al. used a formal economic model to investigate the incentives for businesses and to offer tailored pricing plans when customers were on the verge of quitting and their service utilization was heterogeneous [27]. Kim et al. addressed a few of the research gaps such as lack of customer support and subscription work that had been based on the benefits and rewards that result in customer subscription, subscription switching costs, and retention of customer support [28].

After analyzing various previous works, we notice that even though various machine learning techniques, alongside many non-IT-based techniques, have been used time and again by researchers to get a better understanding of the customers. These works have primarily revolved around regression analysis and hypothesis testing. To provide personalized tariff plans, many works have tried to study the customers' behavioral pattern. The authors have mostly concentrated on finding the interrelationship among the constraints by using the correlation coefficient and likelihood ratio to get a better understanding of the customers' behavior. Hence, the objective of our work is to take a different approach where irrespective of the interdependency among the constraints. It is not the focal point of the work. A step-by-step learning approach has been considered where the pattern of different constraints and how it can impact the main governing aim has been comprehensively analyzed.

III. Materials and Methods

Classification falls under the supervised learning category, where the targets are also given access to the input data. Supervised learning is a type of machine learning where the output is predicted by the machines using well-labeled training data that has been used to train the machines. The term "labelled data" refers to input data that has already been assigned the appropriate output. This section accounts for a detailed description about different traditional as well as ensemble learning based classifiers used in the proposed model.

A. Traditional Classifiers

1) Decision tree: A decision tree is considered one of the few most generally used grading systems in classifying data or understanding the hidden pattern of a particular data collection. It consists of a group of nodes among which the first node is designated as the root node while the rests are known as internal and leaf nodes. The last layer of the decision tree consists of leaf nodes which usually have a predefined class target value. The primary backbone for building a decision tree is repeatedly dividing the nodes on each level based on the criteria for splitting [29]. This splitting and expanding phases last until a stopping criterion are encountered. The various criteria can be represented as follows numerically,

$$InformationGain(b_i, R) = Entropy(z, R) - \sum_{u_{i,j} \in dom(b_i)} \frac{\left|\sigma_{b_i = u_{i.j}} R\right|}{|R|} \cdot Entropy\left(z, \sigma_{b_i = u_{i.j}} R\right) \tag{1}$$

Where,

$$Entropy(z, R) = \sum_{d_j \in dom(z)} - \frac{\left|\sigma_{z = d_j} R\right|}{|R|} \cdot \log_2 \frac{\left|\sigma_{z = d_j} R\right|}{|R|} \tag{2}$$

$$Gini(z, R) = 1 - \sum_{d_j \in dom(z)} \left(\frac{\left|\sigma_{b = u_{i.j}} R\right|}{|R|}\right)^2 \tag{3}$$

Where, R= a training set; b_i= a discrete attribute; z= target attribute; $u_{i,j}$= values.

2) K-Nearest Neighbor: The KNN classifier assesses the similarity between the new system and the training process instances to classify a new process into either normal or intrusive classes. It also uses the class-label of the nearest k classes to predict the new process class. The assumption behind the procedure is that the processes of the same class are grouped into the vector space. The computation of the neighbor elements depends on the value of k which is the number of neighbors to describe the class of the data. The election of the immediate neighbors is done by majority voting. This is a straightforward solution but depends on the value of k. Euclidean distance metric can be used to calculate the distance and the mathematical representation has been shown in equation 4 [30, 31].

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2} \quad (4)$$

3) Logistic regression: Logistic regression is a sort of statistical probabilistic model of categorization. It is also used to forecast a category variable that depends on the predictor variables of one or more of them (such as client characteristics). In our problem, it is used after extensive data preprocessing on the original dataset [25]. The logistic regression uses sigmoid function and it will output the probability which will then be mapped to the two or more classes.

$$s(z) = \frac{1}{1+e^{-z}} \quad (5)$$

where, s(z) = output between 0 and 1 (probability estimate); z = input to the function; e = base of natural log.

For multiclass logistic regression, the binary classification is run multiple times, once for each class. Prediction = max (probability of the classes).

4) Artificial neural network: The artificial neural network (ANN) is a mathematical representation that is strongly based on the functional and structural aspects of the human nervous system. Each input of a neural network is multiplied by a weight which is then added to the sum of these weighted inputs, and the bias in the layers following the input layer. The network's last layer contains the transfer function, also known as an activation function, which helps in achieving the desired scalar output. The following is the mathematical representation of an artificial neuron:

$$O(t) = f\left(\sum_{i=1}^{n} v_i(t) \cdot I_i(t) + c\right) \quad (6)$$

where; $O(t)$ is the output which is found from the neural network at a given time, f represents the activation or transfer function, c is the bias, $I_i(t)$ and $v_i(t)$ are the inputs and their weights respectively [32].

B. Ensemble Learning

Ensemble learning is a collection of distinct learning machines or classifiers that work together to produce more accurate and reliable results by combining all methods. An inducer, also known as a strong learner in ensemble methods, is a learner that is randomly well correlated with the actual classification of the labeled training set. While a weak learner is related to the true classification to some extent. Boosting, bagging, and random forests are some of the new techniques that have emerged because of the ensemble learning framework. In the training of the base model, parameter values vary as well, allowing different ensemble components to offer their uniqueness to the framework. As a result, ensemble systems are incredibly versatile and efficient in real-world applications, giving them the ability to solve and approach a wide range of problems. In ensemble learning, there are various aggregation strategies; bagging is one of them, and it is employed in prediction models to reduce variance. Other strategies include: stacking which tries to reduce prediction bias, and boosting which aids in the conversion of several weak learners into an aggregated strong model [29,25]. Fig. 1 shows a generalized tree representation of the work behind ensemble learning.

1) Random forest: Random forest is an ensemble study that uses more than one decision-making tree for classification and regression. There is some randomness while selecting the subset and characteristics for the nodes of each tree in the random forest classifier. The Gini index is one of the criteria to divide the data into random forest. The random forest is also important for variables, which does not only contribute to the development of a precise model but also contributes to prediction [29,30]. Fig. 2 shows the tree representation of the working behind a random forest ensemble model.

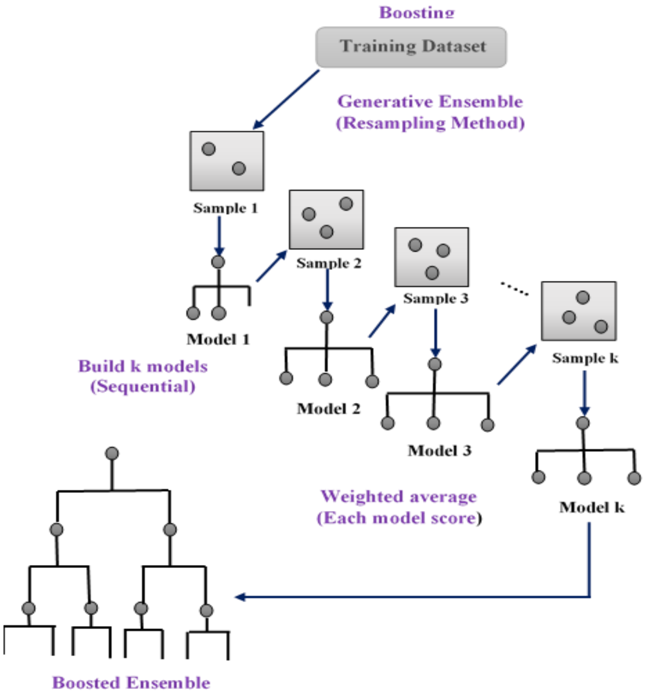

Fig. 1. Tree Representation of Ensemble Learning.

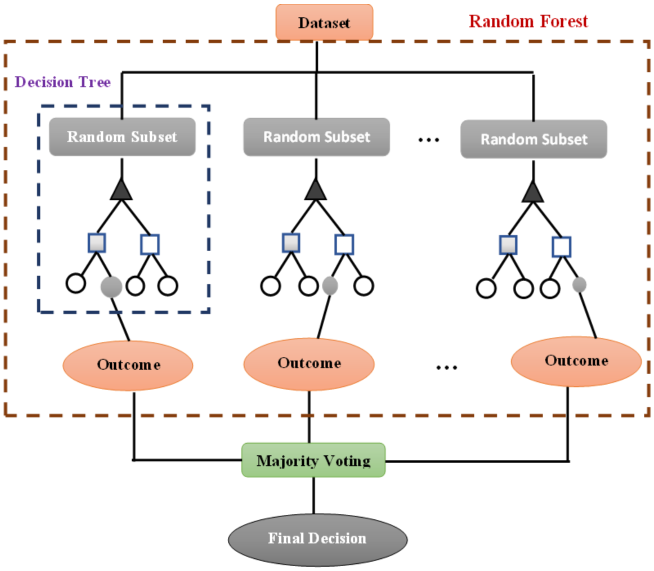

Fig. 2. Tree Representation of Random Forest.

2) Gradient Boosting Machine (GBM): GBM is a frequently used machine learning approach for both regression and classification issues. It is highly effective in solving a wide variety of practical applications. The GBM approach is considered an approach for numerical optimization, which seeks to identify an additive model to minimize loss. The GBM approach builds a new decision tree (i.e., "weak learners") at each step, which will lower loss function most effectively [33, 34, 35].

$$F_M(x) = F_{M-1}(x) + \gamma_M h_M \qquad (7)$$

where, M= number of iterations; γ_M = multiplier; $h_M(x)$ = a base learner.

3) Extreme Gradient Boosting (XGB): XGBoost is a regression tree with the same Decision Tree principles. It promotes regression and classification. This approach is an effective and scalable gradient booster (GBM) variation which is frequently used for computer vision, data mining, and other applications. The training is carried out via an "additive strategy": a tree ensemble model utilizes *t* additive functions to predict the output, given a molecule *i* using a descriptor x_i vector [36,37]. The mathematical representation of a simplified objective of XGBoost has been shown in equation 7.

$$L^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \qquad (8)$$

where, $L^{(t)}$is the objective function at iteration *t* and g_i and h_i are the first and second order gradient statistics of the loss function.

4) AdaBoost: AdaBoost is the acronym for adaptive boosting. It is a form of classification algorithm model of dichotomy which trains and combines a range of weak classifiers to fulfill the classification criteria of datasets. AdaBoost is adapted to increase the weight of a sample misclassified by the previous weak classifier and to reduce the weight of the correctly classified sample to the following weak learner. AdaBoost's instance categorization can be expressed mathematically as:

$$H(x) = sign(\textstyle\sum_{t=1}^{T} \alpha_t \cdot M_t(x)) \qquad (9)$$

where M_t represents the classifiers while α_t is the individual weight of each classifier [38, 39].

5) Bagging: Bagging is an ensemble learning approach for improving the accuracy of other learning algorithms' predictions. An ensemble method of learning is a strategy that integrates numerous machine algorithms for the prediction that can be done with better accuracy and stability than with only a single algorithm. Bagging is very effective for decision tree, even though it may be used for various types of classification techniques. Bagging is a technique for combining numerous complex models and then averaging (in regression) or majority voting (in classification) the outputs of the different models of the same kind to create a more powerful prediction model [40, 30].

```
Input: Data set $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$;
Base learning algorithm $L$; Number of base learners $T$.

Process:
    1. for $t = 1, \ldots, T$ :
    2.     $h_t = L(D, D_{bs})$ //$D_{bs}$ is the bootstrap distribution
    3. end
Output: $H(x) = \arg max_{y \in Y} \sum_{t=1}^{T} (h_t(x) = y)$        (10)
```

IV. Proposed Methodology

A. Dataset Used

A survey in the form of a questionnaire was used to collect the data from different mobile phone users. It has been prepared after consulting with an Indian telecommunication service provider. It was designed by using google form and has been shared among different Indian telecommunication subscribers through different social media platforms, emails, and messages. The data was collected from the legal users of different Indian service providers in the form of a categorical form dataset. The questionnaire contains multiple-choice questions. The unclean dataset contains 476 user information and 88 attributes. Then, we have proceeded to the preprocessing steps which led to the formation of the final refined dataset containing 476 user information and 16 attributes. Table I shows the attribute list of the cleaned dataset and Table II shows the lists of attributes used to calculate the three target attributes obtained after the preprocessing.

TABLE I. Final Attribute List and their Description

Attribute Name	Labels Assigned	Label Distribution	Description
Type of service used	• Cellular Prepaid • Cellular Post-paid	418 58	Connection type
Age group	• Under 18 • 19 – 25 Years • 26 – 40 Years • 41- 55 Years • 56 and above	4 347 77 19 29	Customer age group
Gender	• Male • Female • Non-Binary • Others	309 166 0 1	Customer gender identity
Occupational status	• Employed • House wife • Business • Student • Retired	71 17 8 359 21	Customer occupational background
Income group (income per month)	• Below Rs. 20,000 • Rs.20,000 – 50,000	47 64 28 29	Customer monthly income

	• Rs. 50,000 – 1 Lakh • Above Rs. 1 Lakh • Dependent	308	
Education	• Non-graduate • Graduate • Post-graduation and above	300 115 61	Customer educational background
You have been using cellular services for	• Less than 1 Year • 1 to 3 years • More than 3 Years	21 117 338	Duration for which the customer has been using the telecom service
Average expense on cellular services per month will be about	• Less than Rs. 500 • Rs. 500 – 1000 • Rs. 1001 – 3000 • More than Rs. 3000	308 141 23 4	Approximate monthly expenditure on monthly recharge
Do you use internet	• Yes • No	471 5	If the customer uses internet
Preferable medium for messaging:	• SMS/ Offline messaging Platform • Online Messaging Platform	13 463	Customers' preferred medium of messaging
How frequently do you travel?	• Once a year • Twice a year • More than twice	177 126 173	An approximate number of times the customer travels
Where do you travel the most?	• Within the country • Outside the country	444 32	Travelling preference
Do you use the same number while travelling?	• Yes • No	450 26	Preferred number used during travelling
Total time spent on call during weekdays	• 1 to 3 hours • 3 hours to 5 hours • 5 hours to 10 hours • More than 10 hours	0 204 144 128	Number of hours spent on call during the weekdays and over the weekend. **(Target Data)**
Total time spent on call during weekends	• 1 to 3 hours • 3 hours to 5 hours • 5 hours to 10 hours • More than 10 hours	0 172 154 150	
Total data consumptions	• Less than 500 MB • 500 MB to 1 GB • 1 GB to 3 GB • More than 3 GB	74 66 211 125	Amount of mobile data consumed. **(Target Data)**

TABLE II. TARGET ATTRIBUTES AND THEIR DETAILS

Target Attributes	Attributes used to obtain the Target Attributes	Description
Total time spent on call during weekdays	Preferable time to connect during the weekdays (Monday- Friday): [Family]	Preferred time of the day to connect over calls during the weekdays. • Morning (6am – 12 pm) • Afternoon (12pm – 5pm) • Evening (5pm – 10pm) • Night (10pm – 6am)
	Preferable time to connect during the weekdays (Monday- Friday): [Relatives]	
	Preferable time to connect during the weekdays (Monday- Friday): [Friends]	
	Preferable time to connect during the weekdays (Monday- Friday): [Colleague]	
	Preferable time to connect during the weekdays (Monday- Friday): [Others]	
	Time spent for total calls per day during weekdays (Monday- Friday): [Family]	Time duration spent on calls during the weekdays at different time of the day. • Less than 1 hour • 1 hour to 3 hours • 3 hours to 5 hours • More than 5 hours
	Time spent for total calls per day during weekdays (Monday- Friday): [Relatives]	
	Time spent for total calls per day during weekdays (Monday- Friday): [Friends]	
	Time spent for total calls per day during weekdays (Monday- Friday): [Colleague]	
	Time spent for total calls per day during weekdays (Monday- Friday): [Others]	
Total time spent on call during weekends	Preferable time to connect during the weekends (Saturday- Sunday): [Family]	Preferred time of the day to connect over calls during the weekend. • Morning (6am – 12 pm) • Afternoon (12pm – 5pm) • Evening (5pm – 10pm) • Night (10pm – 6am)
	Preferable time to connect during the weekends (Saturday- Sunday): [Relatives]	
	Preferable time to connect during the weekends (Saturday- Sunday): [Friends]	
	Preferable time to connect during the weekends (Saturday- Sunday): [Colleague]	
	Preferable time to connect during the weekends (Saturday- Sunday): [Others]	
	Time spent for total calls per day during weekends (Saturday- Sunday): [Family]	Time duration spent on calls during the weekdays at different time of the day. • Less than 1 hour • 1 hour to 3 hours • 3 hours to 5 hours • More than 5 hours
	Time spent for total calls per day during weekends (Saturday- Sunday): [Relatives]	
	Time spent for total calls per day during weekends (Saturday- Sunday): [Friends]	
	Time spent for total calls per day during weekends (Saturday- Sunday): [Colleague]	

	Time spent for total calls per day during weekends (Saturday- Sunday): [Others]	
Total data consumptions	Usage of internet in hours: [Morning (6am-12pm)]	Number of hours internet is being used. • Less than 1 hour • 1 hour to 2 hours • 2 hours to 3 hours • More than 3 hours
	Usage of internet in hours: [Afternoon (12pm- 5pm)]	
	Usage of internet in hours: [Evening (5pm- 10pm)]	
	Usage of internet in hours: [Night (10pm- 6am)]	
	What are you surfing? [Morning (6am-12pm)]	Type of content being used on the internet. • Videos • Social Media • Websites • Chats
	What are you surfing? [Afternoon (12pm- 5pm)]	
	What are you surfing? [Evening (5pm- 10pm)]	
	What are you surfing? [Night (10pm- 6am)]	
	What type of websites are you surfing?	• Educational • Business • Entertainment
	What do you download from Internet?	• Videos • Images • Documents • Mp3 files • Software
	What type of calls are you making through Internet?	• Video Calls • Voice Calls • None
	Do you use paid music streaming platform?	• Yes • No

From Fig. 3, it can be seen that a greater number of weekday calls ranging from three to five hours, and during the weekends a larger number of calls exceeds 10 hours. Our dataset lacks any record which ranges between one to three hours. Fig. 4 shows the count of customers based on their data consumption. As it can be seen in both graphical representations, data imbalance is an issue. However, balancing out the data doesn't have any positive effect on the model outputs. Hence, the step of data balancing was not further considered during the execution of the models.

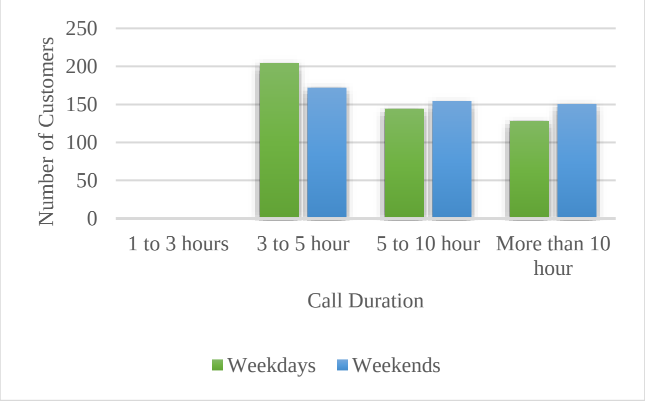

Fig. 3. Count of Customers for Each Call Duration Range during Weekdays and Weekends.

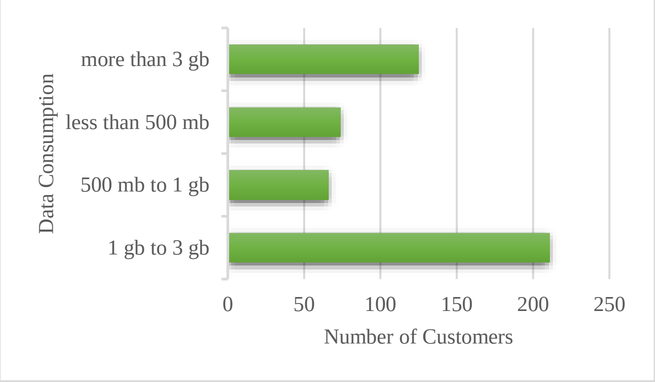

Fig. 4. Count of Customers for Each Data Consumption Range.

B. *Proposed Model*

The work proposes a research model which predicts customers' behavioral data like call duration and data consumption, by analyzing their demographic data such as age, gender identity, occupation, education and so on. Different constraints have been taken into consideration to design a generalized model that would provide a better prediction. Even though the machine learning techniques that are used here are already existing and well-known, the proposed approach has not been used in the Indian telecommunication sector. Fig. 5 shows the diagrammatic approach toward the research objective in a step-by-step style.

Step 1: Data collection. As discussed in Section 4.1, the data has been collected using a questionnaire through a survey among regular telecom subscribers of the Indian telecom industry. After the data has been collected, it has been divided into two parts: the customer demographic data and customer behavioral data.

Step 2: Data Preprocessing. Since the main target is to predict the customers' call duration and data consumption, the behavioral data obtained are used to calculate target attributes as mentioned in Table II. For calculating the data consumption, depending on the browsing history using Internet, the option is replaced by an Internet data unit such as Streaming Video: 353 Mb/Hour; social media: 20 Mb/Hour; Websites: 25Mb/Hour; Chats (if both Video and Voice calls or just Video call): 100 Mb/Hour; Chat (if just voice call): 60 Mb/Hour. For situations where multiple options are selected by the users, the option having the maximum value is selected. Once the target is calculated, the attributes that are used for the purpose and all the other attributes irrelevant to the study are dropped. On the other hand, the demographic data are only cleaned by removing redundant data and null values. After completing all the preprocessing steps, the data has been encoded using one-hot encoding technique.

Step 3: Training, Testing, and Validation. After the preprocessing has been completed, the data have been encoded by using the one hot encoding technique and partitioned for the training and testing purpose. The split for the training and testing has been 80% and 20% respectively. As discussed in Sections 3.1 and 3.2, multiple ensemble models and traditional classifiers have been used for the prediction purpose. Other than all the techniques discussed previously, a stacking ensemble model has been used as one of the ensemble techniques for the predictive system which has been shown in Fig. 6. The stacking model consists of two levels. Level 0 comprises four base classifiers: Decision tree (DT), Random Forest (RF), Logistic Regression (LR), k-nearest neighbor (KNN); and level one consists of LR as the meta classifier. A soft voting technique has been used to predict the final output of the stacking ensemble model. Majority voting, plurality voting, and weighted voting can be used for individual classifiers that produce clear class labels, although soft voting is typically preferred for individual classifiers that produce class probability outputs. When all the individual classifiers are treated equally, the simple soft voting approach simply averages all the individual outputs to produce the combined output, but if we combine the individual outputs by using various weights, a weight specific to each classifier is generated, and the total output for class cj is,

$$H^j(x) = \sum_{i=1}^{T} w_i h_i^j(x) \tag{11}$$

where w_i is the weight assigned to the classifier h_i.

Other used ensemble models: Random Forest, AdaBoost, XGB, GBM, and Bagging. Among these models, Random Forest has given the best accuracy of 83%. The traditional classifiers used have been Decision Tree, Logistic regression, k-Nearest Neighbor, and ANN. Table III presents the detailed tuning parameter specifications of different classifiers that are used in the research during the implementation by using Python.

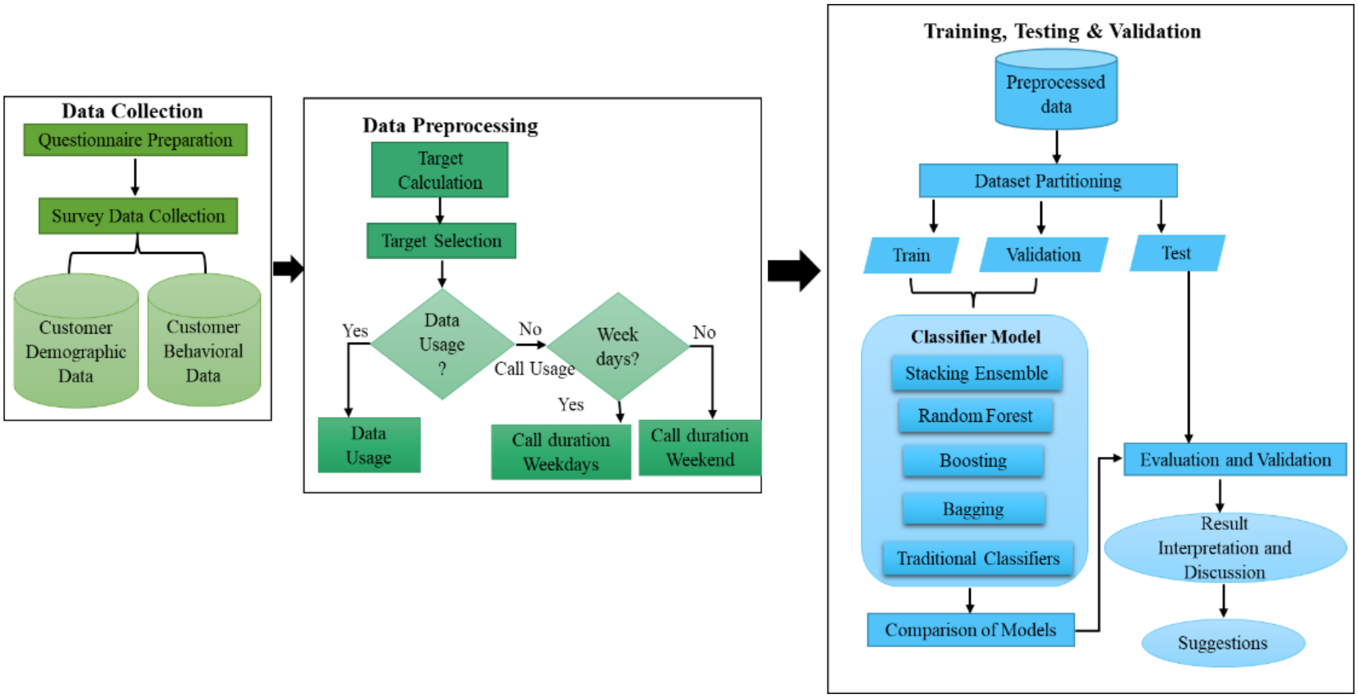

Fig. 5. The Proposed Research Model.

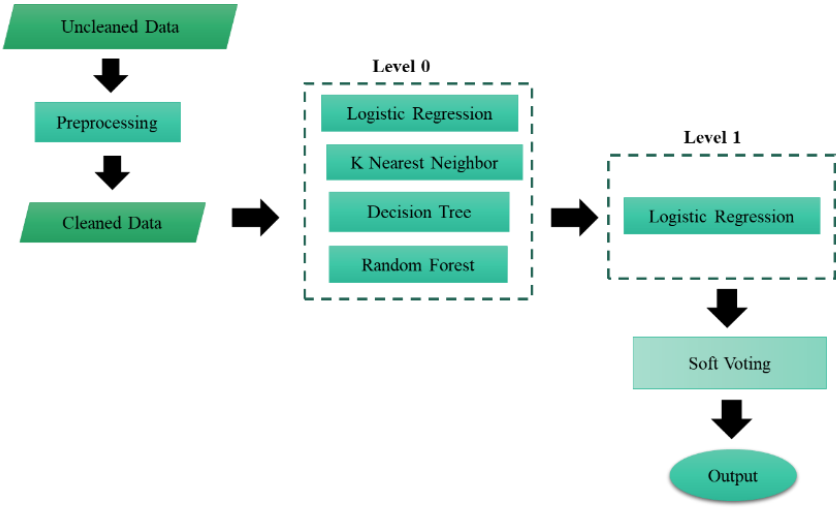

Fig. 6. Schematic Diagram of the Stacking Model.

TABLE III. PARAMETER SPECIFICATIONS

Technique used	Specifications
Random Forest	Criterion =Gini, Max Depth=3, Min Samples Split=5, Estimators=10
AdaBoost	Learning Rate=0.09, Estimators=10
XGM	Booster= Gbtree, Learning Rate=0.1, Max Depth=2, Estimators=20
GBM	Criterion=Friedman Mse, Learning Rate=0.01, Max Depth=3, Min Samples Split=2, Estimators=60
Bagging	Estimators=300, Random State=10
Stacking	Cv=10, Estimators= [Decision Tree Classifier, Random Forest Classifier, Logistic Regression, K Neighbors Classifier], Final Estimator= Logistic Regression
Decision Tree	Max Depth=2, Criterion = Gini, Min Samples Split=2
Logistic Regression	Solver = Liblinear, Max Iter=10000
k-Nearest Neighbor	Metric= Minkowski, Neighbors=38, P=2

Artificial Neural Network	Dense = 32, Activation=Relu
	Dense = 16, Activation=Relu
	Dense = 3, Activation=Softmax
	Optimizers = Adam, Learning Rate=0.001
	Loss = Categorical Crossentropy

C. Performance Metrics

In this work, the proposed predictive model for customer behavior has been evaluated by using accuracy, precision, sensitivity, specificity, F1 score, and kappa analysis. The fraction of total samples properly classified by the classifier is called accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

where TP denotes true positive, FP denotes false positive, TN denotes true negative, and FN denotes false negative. Precision is what percentage of positive forecasts were truly positive.

$$Precision = \frac{TP}{TP+FP} \tag{13}$$

The sensitivity is the fraction of all positive samples that the classifier accurately identified as positive.

$$Sensitivity = \frac{TP}{TP+FN} \tag{14}$$

Specificity is the fraction of all negative samples that have been accurately identified as negative.

$$Specificity = \frac{TN}{TN+FP} \tag{15}$$

F1 score is the combination of precision and sensitivity.

$$F1\ score = 2 \times \frac{Precision \times Sensitivity}{Precision+Sensivity} \tag{16}$$

For qualitative items, kappa analysis is a statistical measure of inter-rater reliability.

$$k = \frac{p_o - p_e}{1-p_e} = 1 - \frac{1-p_o}{1-p_e} \tag{17}$$

The kappa coefficient varies from 0 to 1 and following statements can be made,

< 0 = agreement equivalent to a chance

0.1– 0.2 = slight agreement

0.21 – 0.40 = fair agreement

0.41 – 0.60 = moderate agreement

0.61 – 0.80 = substantial agreement

0.81 – 0.99 = almost perfect agreement.

V. RESULTS

In this work, different classification techniques have been applied to the customers' demographic data to predict their behavioral pattern. As discussed in the previous section, the data collected has been preprocessed to calculate the target attributes: total call duration during weekdays, total call duration during the weekend, and total data consumption. After preprocessing, the desired target has been selected for the classification. In this work, few ensemble learning techniques and some traditional classifiers have been applied. K-fold cross validation has been used, as well, to validate the models, where k is been maintained at 10 irrespective of models. As shown in Tables IV and V, random forest has given the best accuracy of 83% among all the techniques applied. Apart from the performance metrics, the reliability of the model has been tested by using kappa analysis. Random Forest has shown the best result among all the classifiers that are applied with a kappa value of 0.74, which comes under the categorization of substantial agreement. When kappa analysis was performed on the ensemble models, except for the bagging technique, all models have given values within the range of 0.61 to 0.88 which signifies substantial agreement. Models that recorded accuracy of over 80% were AdaBoost, XGB, and the stacking ensemble model whose kappa values have been 0.71, 0.71, and 0.69, respectively. Among traditional classifiers, decision tree has given the best result with an accuracy of 81% and 0.71 kappa value. The other traditional classifying techniques with kappa value of substantial agreement are logistic regression and k-nearest neighbor.

Fig. 7- 9 show the bar graph plot of the accuracy, specificity, and precision measures of all the classifiers that are used in this work. Fig. 10 shows the obtained F1-scores and sensitivity from all the classifiers. Concerning the specificity, it can be inferred that there is a small number of false negative samples. As kappa values have been calculated for substantiating the reliability of the proposed approach, Fig. 11 presents the bar plot of the accuracy of each classifier, as well as their corresponding kappa values and Fig. 12 shows a bar plot of the kappa values in decreasing order.

TABLE IV. PERFORMANCE OF ENSEMBLE MODELS

Classifiers	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**	**F1-score**	**Kappa**
Random Forest	0.83	0.83	0.83	0.91	0.83	0.74
AdaBoost	0.81	0.82	0.81	0.90	0.81	0.71
XGB	0.81	0.82	0.81	0.90	0.81	0.71
GBM	0.78	0.77	0.78	0.89	0.78	0.66
Bagging	0.73	0.72	0.73	0.88	0.72	0.58
Stacking	0.80	0.80	0.80	0.93	0.79	0.69

TABLE V. PERFORMANCE OF TRADITIONAL CLASSIFIERS

Classifiers	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**	**F1-score**	**Kappa**
Decision Tree	0.81	0.82	0.81	0.90	0.81	0.71
Logistic Regression	0.74	0.74	0.74	0.87	0.74	0.61
k-Nearest Neighbor	0.76	0.76	0.76	0.89	0.76	0.63
Artificial Neural Network	0.72	0.72	0.72	0.86	0.72	0.57

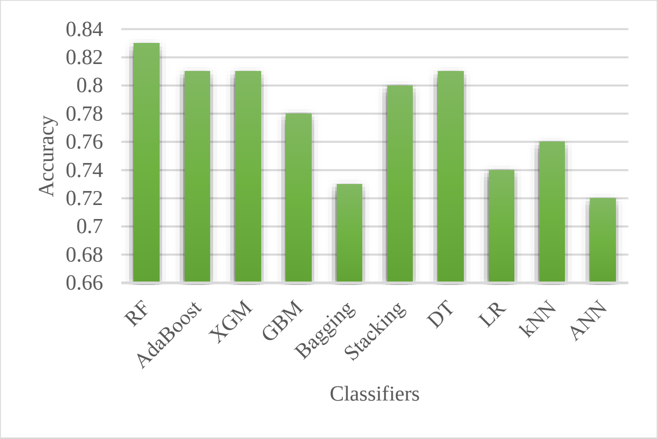

Fig. 7. Accuracy Plot of different Classifiers.

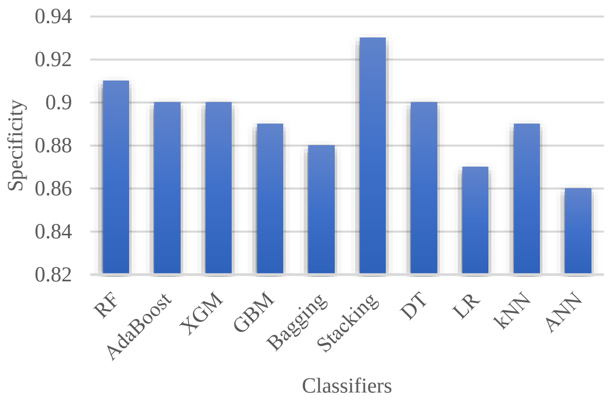

Fig. 8. Specificity Plot of different Classifiers.

Table VI shows some instances of test cases that have been used to test the classification models. It can be seen that test cases 1, 2, and 3 are predicted correctly by using all the classifiers except KNN, which has the least average success rate among all classifiers. Test case 4 is incorrectly predicted by most of the classifiers other than bagging, stacking, and ANN. Even though the performance accuracies obtained during training of the bagging and ANN models are lower than the other classifiers, yet when applied to certain test cases they have performed better.

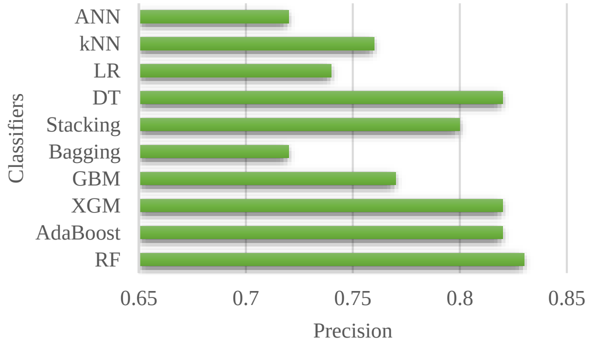

Fig. 9. Precision Plot for different Classifiers.

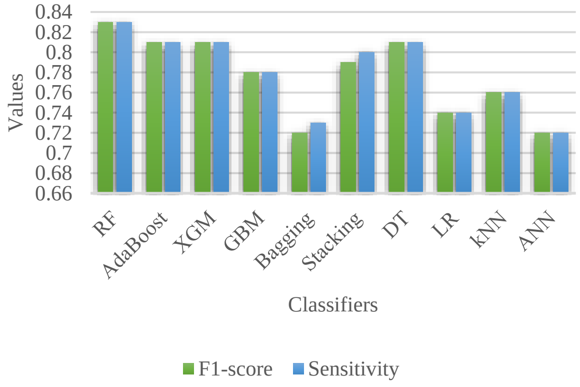

Fig. 10. F1-score and Sensitivity Plot of the Classifiers.

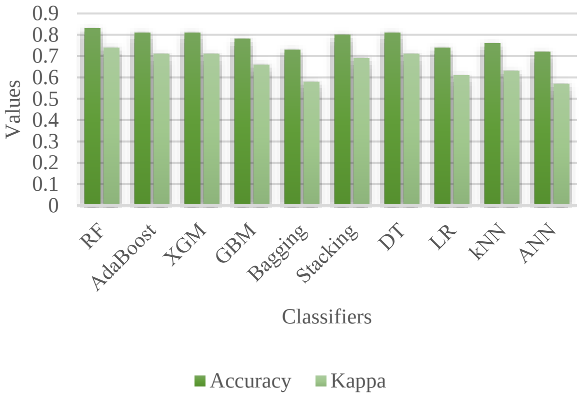

Fig. 11. Accuracy and Kappa Value Plot of the Classifiers.

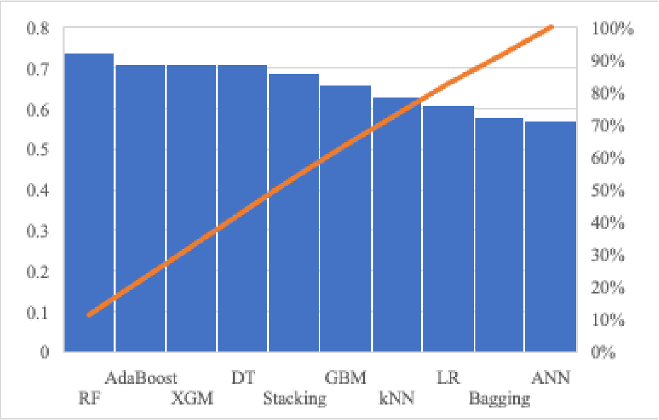

Fig. 12. Kappa Value Plot for Different Classifiers in Decreasing Order.

TABLE VI. DIFFERENT TEST CASE SCENARIOS

Test Case	Actual Class	Predicted Class/ Output									
		RF	AdaBoost	XGM	GBM	Bagging	Stacking	DT	LR	kNN	ANN
1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	1	2
3	0	0	0	0	0	0	0	0	0	0	0
4	1	2	2	0	2	1	1	0	1	2	1
5	2	2	2	2	2	2	2	2	1	2	2
Average success rate (%)		80	80	80	80	100	100	80	80	60	100

VI. Discussion

Personalized tariff plans can solve a lot of problems for the Indian telecom industry. The current generalized status of the tariff, plan design has led to a lot of dissatisfaction among the customers. This can be solved by better analyzing the customers' behavioral pattern and how it varies depending on their socio-economic background. To offer a better recommendation to customers, flexible tariff plans need to be designed and hence the requirement for an effective predictive system.

In this work, a predictive system has been designed to be able to predict customers' behaviors: call duration and data consumption. Different classification models have been tested on the data obtained from the customers to get a better prediction of the customers' behavioral pattern. This can help with understanding what kind of tariff plan should be recommended and for which segment of the customers. Every customer's needs vary based on their lifestyle and background. An ideal scenario for the proposed work would be the current situation where students, teachers, and IT sector personnel has been working online which have led to a drastic increase in data consumption among these people. Therefore, the tariff plans that are recommended for these people should include higher data allocation than those who are working offline. Again, the total call duration of females who are housewives and in the age group over 56 is different from those who are working and of the age group of 26 to 40. Thus, a desperate need for flexible tariff plans. As shown previously, this work is based on customers' age group, gender identity, educational background, occupational status, monthly income, expenditure on mobile plans, and so on to predict their behavioral pattern.

Comparing to different works of literature on customers' behavioral pattern, this work focuses on determining the feasibility and efficiency of using classification techniques on customers' demographic data to detect patterns. According to the results obtained through this work, ensemble learning techniques and ANN can be used to predict the customers' call duration and data consumption with high efficacy. Service providers can use this proposed model to predict the call duration and data consumption which can be used to design more flexible tariff plans catering to the customers' needs in a much more personalized form. This research model can also be used for customer segmentation depending on their socio-economic background, which will make it easier for recommending the right tariff plan. This can also help with customer retention since retaining the right customer is one of the major goals of any telecom service provider. Investing in the right customer can lead to higher profit for the business. As the dataset varies in all the previous works, hence a usual comparative study was not feasible for the proposed work. Therefore, Table VII only highlights the different inferences that are reached for the general goal of tariff plan design.

TABLE VII. HIGHLIGHTS OF DIFFERENT WORKS BASED ON TARIFF PLAN DESIGN

Reference	**Techniques Used (IT based)**	**Dataset used**	**Performance Metric/ Inference Drawn**
Bahri-Ammari and Bilgihan, 2017	Hypothesis Testing	Customer survey data from mobile telecom industry in Tunisia.	Path Coefficients and t-test: 0.815 (5.888)
Newton and Ragel, 2017	Correlation and Regression Analysis.	Customer data in mobile telecommunication industry in Batticalao.	Adjusted R square of Model A= 0.786. Adjusted R square of Model B= 0.804.
Belwal and Amireh, 2018	Partial Least Square based Structural Equation Modelling	Survey data from Omani Telecom industry.	Cronbach's alpha score greater than 0.7 and composite reliability scores 0.70–0.90
Tong et al., 2017	Information Gain on Decision tree Model and k mean Clustering.	Survey data from customers of telecom industry.	Accuracy of 71.24%
Jose et al., 2017	Principle Component Analysis (PCA), Factor Analysis, Regression Analysis and Cross Tabulation.	Data from Telecom contact center.	Sampling adequacy is validated with KMO test score of 0.739.
Gerpott and Meinert, 2017	Hypothesis testing	Data obtained from the German subsidiary of a large multinational MNO.	McFadden's pseudo R2 of 0.085 and coefficient of determination of 0.318.
Bibin and Ramanathan, 2018	Conjoint analysis	Survey data of students.	Result showed that better network and low cost are the main reasons for their choice of service provider.
Dubey and Srivastava, 2016	Exploratory Factor Analysis, Regression Analysis.	Customer Survey data from Indian telecommunication industry.	KMO test score of 0.833.
Jin et al., 2021	Hypothesis testing	Major telecommunication operator in China.	Probability of switching up = 0.002 Probability of switching down = 0.010
Shouchen et al., 2017	Distance method by norm method	Dataset from China telecommunication industry.	The matching result indicates that 302 676 users' current tariff package are not optimized selection and should shift their consumption suit.
Gerpott and Meinert, 2016	Regression analysis	A dataset of residential users of mobile communication services.	63.2% of sampled subscribers share a positive/ complementary relationship between the two services whereas a considerable proportion of 36.8% of subscribers substitutes call minutes by data consumption.
Diaz, 2017	Generalized Structural Equation Modelling (GSEM), Multinomial Logit	Real-life survey data from Peru.	Wald Tests shows that the estimation is statistically significant at 1%.
Lee et al., 2018	Conjoint Analysis	Real life survey data from South Korean Telecom	Conjoint analysis show that free data service provides significantly greater benefits on average than smartphone interphone and Enterprise messenger services do.
García-Mariñoso and Suárez, 2019	Logit model	Real-life survey data collected by Spanish Markets and Competition Authority.	estimated correlation coefficient 0.043, p value= 0.669
Kim et al., 2019	Statistical Analysis	Real life data from Korean company.	In the model 1, amount of the variance (R2=0.210, p < 0.001). In the model 2, amount of the variance (the increase of R2=5.6%, p < 0.001)
This Work	Random Forest	Survey data collected through questionnaire	Accuracy= 83%, Kappa Value= 0.74.

Most of the existing works implementing machine learning are applied on real-life customer data and the inference drawn from those works also varies based on the techniques and the used dataset. (Tong et al., 2017) have used machine learning technique of decision tree and have achieved an accuracy of 71.24%. Compared to our work, this work has produced higher accuracy. Whereas works like (Bahri-Ammari and Bilgihan, 2017; Newton and Ragel, 2017; Gerpott and Meinert, 2016) have used correlation coefficients to draw the inference for their work. In another study (Maji et al., 2017) have implemented pattern recognition by using rule-mining based apriori algorithm to study their customers. Considering the varying factors in every research study on personalizing tariff plans drawing a comprehensive comparative study seems to be not feasible. As governing hypotheses of each problem varies greatly among each other, hence drawing a common point of inference is not reasonable enough in this kind of situation.

The major challenge that are faced throughout our work is the data collection. In the telecommunication industry, customer data involves the intricate policies of privacy and security preservation. Therefore, collecting enough data has been a huge challenge. For any data related research, the quality of the data is a major concern. In this work, the data collected from customers directly has been assumed to be accurate and thorough. But the concern regarding the accuracy of the information provided by the customers remains.

Highlights of the contributions of the present work:

- Predicts customers' behavioral pattern by analyzing their demographic data by using classification-based learning methods.
- Helps with customer segmentation which in turn will help recommending the right service to the customers.
- Contributes to designing a personalized tariff plan which will cater to the customers' requirements at a much more reasonable price.

VII. Conclusion

Because of the high efficiency of classification-based learning models, the proposed model produced a higher success rate in predicting the customers' behavior. Among traditional classifiers, the decision tree has given the best result with an accuracy of 81% among other techniques. Also, the reliability of the model has been tested using kappa analysis and showed to be higher than all other investigated classifiers with a value of 0.71. While the ensemble classifier, random forest, has given the best accuracy of 83% among other techniques with kappa value of 0.74. Although the proposed work has provided satisfactory results, it has certain limitations. The major shortcoming of the work can be identified as the size of the dataset which has impacted the performance of certain classifiers to a great extent. Therefore, in future, the work can be extended by using a much larger dataset such that inherent patterns can be identified with more precision. Also, a more robust and generalized model could be designed that would work irrespective of the type of dataset.

References

[1] Gupta, R., & Jain, K. (2020). What drives Indian mobile service market: Policies or users?. Telematics and Informatics, 50, 101383.

[2] Saha, L., Tripathy, H. K., Nayak, S. R., Bhoi, A. K., & Barsocchi, P. (2021). Amalgamation of Customer Relationship Management and Data Analytics in Different Business Sectors—A Systematic Literature Review. Sustainability, 13(9), 5279.

[3] Nkordeh, N., Bob-Manuel, I., & Olowononi, F. (2017). The Nigerian telecommunication industry: Analysis of the first fifteen years of the growths and challenges in the GSM market (2001–2016).

[4] Shao, B. B., Lin, W. T., & Tsai, J. Y. (2017). An empirical study of the telecommunications service industries using productivity decomposition. IEEE Transactions on Engineering Management, 64(4), 437-449.

[5] Valentim, L. C., Quelhas, O. L. G., & Ludolf, N. V. E. (2019). Proposição de sistemática para implantação de Customer Relationship Management apoiado por Business Intelligence a organizações do setor de telecomunicação. Sistemas & Gestão, 14(3), 232-244.

[6] Al-Zadjali, M., & Al-Busaidi, K. A. (2018). Empowering CRM through business intelligence applications: A study in the telecommunications sector. International Journal of Knowledge Management (IJKM), 14(4), 68-87.

[7] Tong, L., Wang, Y., Wen, F., & Li, X. (2017). The research of customer loyalty improvement in telecom industry based on NPS data mining. China Communications, 14(11), 260-268.

[8] Jose, B., Ramanan, T. R., & Kumar, S. M. (2017, November). Big data provenance and analytics in telecom contact centers. In TENCON 2017-2017 IEEE Region 10 Conference (pp. 1573-1578). IEEE.

[9] Bahri-Ammari, N., & Bilgihan, A. (2017). The effects of distributive, procedural, and interactional justice on customer retention: An empirical investigation in the mobile telecom industry in Tunisia. Journal of Retailing and Consumer Services, 37, 89-100.

[10] Newton, S., & Ragel, V. R. (2017). The effectiveness of relational bonds on customer loyalty mediated with customer satisfaction: telecommunication industry, Batticaloa. Asian Journal of Economics, Business and Accounting, 1-11.

[11] Belwal, R., & Amireh, M. (2018). Service quality and attitudinal loyalty: Consumers' perception of two major telecommunication companies in Oman. Arab economic and business journal, 13(2), 197-208.

[12] Maji, G., Mandal, S., Bhattacharya, S., & Sen, S. (2017, March). Designing combo recharge plans for telecom subscribers using itemset mining technique. In 2017 IEEE International Conference on Industrial Technology (ICIT) (pp. 1232-1237). IEEE.

[13] Gerpott, T. J., & Meinert, P. (2017). Choosing a wrong mobile communication price plan: An empirical analysis of predictors of the degree of tariff misfit among flat rate subscribers in Germany. Telematics and Informatics, 34(4), 303-313.

[14] Haq, N. (2017). Impact of Reliance JIO on the Indian telecom industry. International Journal of Engineering and Management Research (IJEMR), 7(3), 259-263.

[15] Bibin, P. B., & Ramanathan, H. N. (2018). Identifying the Best Mobile Combo Tariff Plan for Professional Students: An Application of Conjoint Analysis. International Journal of Business Analytics and Intelligence, 6(2), 36.

[16] Dubey, A., & Srivastava, A. K. (2016). Impact of service quality on customer loyalty-A study on telecom sector in India. IOSR Journal of Business and Management (IOSR-JBM), 18(2), 45-55.

[17] Friesen, L., & Earl, P. E. (2015). Multipart tariffs and bounded rationality: An experimental analysis of mobile phone plan choices. Journal of Economic Behavior & Organization, 116, 239-253.

[18] Jin, H., Lu, Z., Huang, L., & Dou, J. (2021). Not too much nor too little: Salience bias in mobile plan choices. Telecommunications Policy, 45(4), 102071.

[19] Shuochen, X., Lianju, N., & Wenying, Z. (2017). Study of matching model between tariff package and user behavior. The Journal of China Universities of Posts and Telecommunications, 24(3), 91-96.

[20] Gerpott, T. J., Meinert, P. (2016). The impact of mobile Internet usage on mobile voice calling behavior: A two-level analysis of residential mobile communications customers in Germany. Telecommunications Policy, 40(1), 62-76.

[21] Díaz, G. R. (2017). The influence of satisfaction on customer retention in mobile phone market. Journal of Retailing and Consumer Services, 36, 75-85.

[22] Ascarza, E., Iyengar, R., & Schleicher, M. (2016). The perils of proactive churn prevention using plan recommendations: Evidence from a field experiment. Journal of Marketing Research, 53(1), 46-60.

[23] Lee, H., Choi, H., & Koo, Y. (2018). Lowering customer's switching cost using B2B services for telecommunication companies. Telematics and Informatics, 35(7), 2054-2066.

[24] García-Mariñoso, B., & Suárez, D. (2019). Switching mobile operators: Evidence about consumers' behavior from a longitudinal survey. Telecommunications Policy, 43(5), 426-433.

[25] Xu, T., Ma, Y., & Kim, K. (2021). Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping. Applied Sciences, 11(11), 4742.

[26] Bachan, L., & Gaber, T. (2021, March). Predicting Customer Churn in the Internet Service Provider Industry of Developing Nations: A Single, Explanatory Case Study of Trinidad and Tobago. In International Conference on Advanced Machine Learning Technologies and Applications (pp. 835-844). Springer, Cham.

[27] Capponi, G., Corrocher, N., & Zirulia, L. (2021). Personalized pricing for customer retention: Theory and evidence from mobile communication. Telecommunications Policy, 45(1), 102069.

[28] Kim, M. K., Park, M. C., Lee, D. H., & Park, J. H. (2019). Determinants of subscriptions to communications service bundles and their effects on customer retention in Korea. Telecommunications Policy, 43(9), 101792.

[29] Chakrabarti, S., Swetapadma, A., & Pattnaik, P. K. (2021). A channel independent generalized seizure detection method for pediatric epileptic seizures. Computer Methods and Programs in Biomedicine, 209, 106335.

[30] Alsouda, Y., Pllana, S., & Kurti, A. (2019, May). Iot-based urban noise identification using machine learning: performance of SVM, KNN, bagging, and random forest. In Proceedings of the international conference on omni-layer intelligent systems (pp. 62-67).

[31] Liao, Y., & Vemuri, V. R. (2002). Use of k-nearest neighbor classifier for intrusion detection. Computers & security, 21(5), 439-448.

[32] Chakrabarti, S., Swetapadma, A., Ranjan, A., & Pattnaik, P. K. (2020). Time domain implementation of pediatric epileptic seizure detection system for enhancing the performance of detection and easy monitoring of pediatric patients. Biomedical Signal Processing and Control, 59, 101930.

[33] Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy and Buildings, 158, 1533-1543.

[34] Drucker, H., Schapire, R., & Simard, P. (1993). Improving performance in neural networks using a boosting algorithm. In Advances in neural information processing systems (pp. 42-49).

[35] Beygelzimer, A., Hazan, E., Kale, S., & Luo, H. (2015). Online gradient boosting. arXiv preprint arXiv:1506.04820.

[36] Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., & Song, F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Computers in biology and medicine, 121, 103761.

[37] Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J., & Gifford, E. M. (2016). Extreme gradient boosting as a method for quantitative structure–activity relationships. Journal of chemical information and modeling, 56(12), 2353-2360.

[38] Wang, F., Jiang, D., Wen, H., & Song, H. (2019). Adaboost-based security level classification of mobile intelligent terminals. The Journal of Supercomputing, 75(11), 7460-7478.

[39] Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In icml (Vol. 96, pp. 148-156).

[40] Sreng, S., Maneerat, N., Hamamoto, K., & Panjaphongse, R. (2018). Automated diabetic retinopathy screening system using hybrid simulated annealing and ensemble bagging classifier. Applied Sciences, 8(7), 1198.