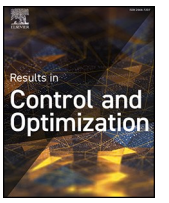
Results in Control and Optimization 14 (2024) 100342

Contents lists available at ScienceDirect

Results in Control and Optimization

journal homepage: www.elsevier.com/locate/rico



Customer churn prediction in telecom sector using machine learning techniques

Sharmila K. Wagh [a,*], Aishwarya A. Andhale [b], Kishor S. Wagh [c], Jayshree R. Pansare [a], Sarita P. Ambadekar [d], S.H. Gawande [e]

[a] Department of Computer Engineering, M.E.S. College of Engineering, S.P. Pune University, Pune, Maharashtra 411001, India
[b] Department of Information Technology, MKSSS's Cummins College of Engineering for Women, Pune 411052, India
[c] Department of Computer Engineering, AISSMS Institute of Information Technology, S.P. Pune University, Pune, Maharashtra 411001, India
[d] Department of Computer Engineering, K. J. Somaiya Institute of Technology, Sion, Mumbai, India
[e] Industrial Tribology Laboratory, Department of Mechanical Engineering, M.E.S. College of Engineering, S.P. Pune University, Pune, Maharashtra, 411001, India

ARTICLE INFO

Keywords:
Churners
Customer churn prediction
Up-sampling
Classifiers
Survival analysis

ABSTRACT

In the telecom industry, large-scale of data is generated on daily basis by an enormous amount of customer base. Here, getting a new customer base is costlier than holding the current customers where churn is the process of customers switching from one firm to another in a given stipulated time. Telecom management and analysts are finding the explanations behind customers leaving subscriptions and behavior activities of the holding churn customers' data. This system uses classification techniques to find out the leave subscriptions and collects the reasons behind the leave subscription of customers in the telecom industry. The major goal of this system is to analyze the diversified machine learning algorithms which are required to develop customer churn prediction models and identify churn reasons in order to give them with retention strategies and plans. In this system, leave subscriptions collects customers' data by applying classification algorithms such as Random Forest (RF), machine learning techniques such as KNN and decision tree Classifier. It offers an efficient business model that analyzes customer churn data and gives accurate predictions of churn customers so that business management may take action within the churn period to stop churn as well as loss in profit. System achieves an accuracy of 99 % using the random forest classifier for churn predicts, the classifier matrix has achieved a precision of 99 % with a recall factor of 99 % alongwith received overall accuracy of 99.09 %. Likewise, our research work improves churn prediction, scope other business fields, and provide prediction models to hold their existing customers customer service, and avoid churn effectively.

1. Introduction

Role of predictive model is to bring the churned customers to light. The proposed model's purpose is to bring churned customers to light. In a targeted approach industry try to identify which customers are likely to churn. The industry then targets those customers or clients and provides them with special incentives, offerings, and plans except normal customers. This approach can bring a huge loss to industry, if churned measures are inaccurate because the industries are wasting a lot of money to the customers who would have stayed

* Corresponding author.
E-mail address: skwagh135@gmail.com (S.K. Wagh).

https://doi.org/10.1016/j.rico.2023.100342
Received 30 March 2023; Received in revised form 24 October 2023; Accepted 11 November 2023
Available online 13 November 2023

anyways, irrespective of short or long distance. Communication has become an important part of today's life. It's being used in every field [1–3]. The telecom industry needs to build the best predictive model for churning customers.

Churn Customer refers to the number of existing customers who may leave the service provider over a given period. These customers can be called as churners. The main aim of churn is to predict the churnable customers at the earliest, to identify the reason for churning. The primary goal of churn analysis is to identify and anticipate churnable consumers as soon as possible. This will help us to rectify the issues of the customer. This will be helpful to satisfy the customer needs and will continue to use that service. This will help to meet the needs of the customers, and they will continue to utilize the service. There are promotional costs known as acquisition costs and retention costs in a telco company. The acquisition cost is the price a company pays to gain new consumers. Retention costs, on the other hand, are the costs of keeping existing clients. It is very difficult to predict which customers would churn and which customers will be maintained due to human limitations. As a result, the allocation of money may be incorrect, resulting in a higher amount of cash being issued.

Furthermore, according to some reports, the acquisition cost is 5 times that of the retention cost. If it is incorrect in projecting a client who will churn, but it turns out that we are correct in anticipating a customer who will be kept, we will have to spend more than we should. This paper attempts to develop a Machine Learning model that can predict customer churn prediction and retention. Retaining cost of the company for the existing customers is far economical as compared to addition of new customers within the network. The customer churn is the direct loss in terms of revenue to the company. If the information of the development of churning of customers is known well in advance, then appropriate steps can be taken, and a better service can be provided to such customers. It is observed that the long-term customers add more revenue to the company as they are not much responsive to slight changes. Today's most challenging and critical problems faced by the telecommunication industry is the management of churning customers. Recent studies show that the main objective is to identify the valuable churn customers using a huge amount of data received from the telecommunication industry. Practically there are so many restrictions in using current models, which faces lots of difficulties & hurdles towards the problem of churning in today's present environment. While modeling development, a lot of information-rich features are neglected in In Feature selection Process. Mostly statistical methods are being used in a diverse domain, which tends to give undesirable results of the present predictive models. The feature selection is another huge problem with the existing models. Every customer may be an individual or a group and have different churning reasons. Classification of a churn customer can be as a churner, irrespective observing reasons and factors for churn. There are different patterns of behavior during the churn process and should not treat all of them in the same manner. Some customers may not churn easily than others. Today's need is for a more realistic prediction model which can predict churn customers in advance. This will be of great help to provide strong retention strategies for the different groups of churners may it be different promotions depending upon the churn factors for the different group of churners. Encouraged by the above-mentioned facts & observations, in this study, a model for prediction of churners with the help of different machine learning algorithms is proposed.

Ullah, et al. [1] provided a customer attrition model for data analytics that is validated using common assessment measures. The results demonstrated that utilizing machine learning techniques improved the performance of their proposed churn model. The F-measure result from Random Forest and J48 was 88 % better. The authors used the dataset to identify the primary churn variables and did cluster profiling based on their churn risk. Finally, the authors presented recommendations for telecom decision-makers on client retention. Ammar, et al. [2] presented a hybridized algorithm technique to predict churners that was quite efficient. When compared between the proposed algorithm and normal firefly algorithm, the accuracy obtained by both is found to be similar. Yet the hybrid firefly algorithm performed better than the normal firefly algorithm in terms of time latency which is very low. The study was carried out of these algorithms with respect to F - Measure, Accuracy, Time, PR and ROC. The prediction model in the telecommunication companies should be as efficient as accurate [3]. The efficiency of the model can be gained by making the best training model with reduced dimension and size. The extraction and feature selection techniques used, helps to get the efficient features along with predicting accurately. Thus, this paper shows that being smaller in size i.e., 232 KB this prediction model performs the similar tasks even with more accurate results i.e., 92 % than the initial model of 303 KB size. In many researches, was represented in different ways and data mining is most usable in churn prediction in telecommunication industry [4]. The concept of multilayer perceptron neural network was used to proposed churn classification model and showed that the logistic boost and logical regression both are useful to build a churn prediction model. Black box models are complicated, but actual work tell only logistic regression and using logit boost training example for weak prediction that play an important role in churn prediction. Effective classifiers like KNN, Decision tree classifier, random forest are studied [5]. Classifiers consist of four stages: data access, data wrangling, training, and acquiring insights that all aspects play an important role to analyze churn prediction data. Here, data access is gives input to the proposed model. Data wrangling is the process of collecting distributed data [6]. The training process is constructed by model and test the data set for processing and acquiring insights is useful to predict final output and analyze the final result of churn prediction.

Machine learning techniques used [7,8] in developing this model for this work. The telecom company can predict churn customers with the help of electronic learning technology with developed model. Industries can provide best services so as to reduce the churn level. Such types of models help telecom services for making them profitable. Random Forest and Decision Tree are used for this model. In [9] the PMM (Predictive Mean Matching) algorithm helps to manage lost values, rather than feature removal or recognition of missing data. The combination of two Ensemble classifiers is embedded within the customer speculation model to manage large databases, time-dependent feature labels, and the distribution of inequitable data in the Telecommunication industry. To evaluate the effectiveness of customer churn uplift models, the author presented a novel profit-driven assessment method dubbed the maximum profit uplift measure in [10]. The proposed MPU measure extends the maximum profit measure for customer churn prediction models, allowing for evaluation of a customer churn uplift model's performance in terms of profit per customer in the customer base earned when a retention campaign targets the optimal proportion of customers with the highest uplift scores. Maximizing the profit generated

by the retention campaign determines the best proportion of customers to target, which is proven in this article to be strongly tied to the uplift model's capacity to identify the so-called persuadable i.e., Customers who are about to churn but will be kept if the campaign is targeted. Saran Kumar et al. [11] gives a thorough examination of the strategies used to anticipate client churn. Each of their proposed churn prediction models has a low forecast accuracy. To avoid the problem of client turnover, a solid prediction model is essential. Kraljević and Gotovac [12] given a logical foundation for developing data mining applications by defining an enhanced technique for designing applications based on Data Mining technologies. The methodology proposed has been deployed and proven in the telecom industry using data mining applications to predict prepaid subscriber attrition.

Real-life examples of customers are used who decided to go and learn the qualities and behaviors that precede customer profits [13]. There are results that show a clear height for advanced versions of models compared to plain (non-expanded) versions. The best separator was SVM-POLY using AdaBoost with approximately 97 % accuracy and F-rating over 84 %. Alzubaidi, et al. [14], used CDR attributes to determine the strength of social ties between users for each identified community, and then used a model to propagate churn influences on the call graph to determine the net aggregated influences from the churner node. This effect was employed in the logistic regression method to calculate the churn tendency for individual users. When building the Telecom social network, examine the relationship between users to improve the performance of the churn prediction model. CDR attributes were employed to characterize the relationship strength of social communication between edges in the graph. CDR attributes contain information about phone calls and SMS messages. In this work the quality of prediction measurements for various prediction models is compared. Khalid, et al. [15] looked at different prediction models and compared the quality of prediction models such classification algorithms and decision trees. They discovered that the decision tree's accuracy is higher than the other techniques (3 % higher than the second result and 6 % higher than the lowest-achieving algorithm), showing that the decision tree is an excellent churn prediction technique.

The primary goal of study presented in [16] is to develop a method to predict high-value customer turnover based on existing research and customer attribute characteristics in the telecom industry. This study accomplishes customer churn prediction based on the telecom business based on the analysis of big data in the telecom industry and historical information estimation of customers, combined with logistic regression method. It identifies possible churned customers in the customer library by studying the features of customer churn behavior in the telecom industry, and it assists organizations in taking targeted win back efforts based on the characteristics of the potential churned A review of customer churn prediction in the telecommunications industry is presented in [17]. The study demonstrates a huge number of attributes that are put into practice by a big number of paper reviewers to construct a customer churn prediction model. Lalwani et al. [18] presented a comparative study of customer churn prediction in the telecommunication industry using well-known machine learning techniques like Logistic Regression, Nave Bayes, Support Vector Machines, Decision Trees, Random Forest, XGBoost Classifier, CatBoost Classifier, AdaBoost Classifier, and Extra Tree Classifier in this research paper. The experimental results demonstrates that two ensemble learning algorithms, Adaboost classifier and XGBoost classifier, have the highest accuracy when compared to other models, with an AUC score of 84 percent for the churn prediction problem. They outscored other algorithms across the board in terms of accuracy, precision, F-measure, recall, and AUC score. In [19] an improved churn prediction for credit card system using supervised learning and rough clustering technique is presented. Whereas, Amin et al. [20] implemented churn prediction using Naïve Bayes for customers of telecommunication industry with focus on the challenges to get new customers in telecommunication industry instead of handling old customers. Model to analyze customer behavior was developed along with improvement in accuracy of prediction. In line with this work again Amin et al. [21], highlighted telecommunication sector's churn prediction using Just-in-Time approach. It handled various problems of customers intention that may switch from one service provider to another. In this context, different churn prediction methods are applied by practitioners to resolve issues related to preserving customer retention.

In other work, Amin et al. [22] presented churn prediction for customers of cross-company by using transformation of data techniques. It emphasizes target company with lacking data which may be used to source company to predict customer churn effectively. Even, it is also including impact of these techniques on performance using various classifiers needed in telecommunication sector. Amin et al. [23] used rough set method to predict customer churn of telecommunication sector. Customer behavior was used to differentiate the churn from non-churn customers. The major objective was to desperate the necessity of businesses in retention of existing customers. Moreover, this work uses intelligent rule-based techniques for decision making along with rough set to detect customer churn from non-churn customers. Ahmad et al. [25] focused on multi dataset approach for churn prediction which is applicable for telecommunication industry. In this work, major contribution related to feature engineering phase for creation of custom features with the help of machine learning is presented. Moreover, the big data approach along with comparative results of multiple algorithms of machine learning has been used.

In the telecom sector, a number of customer retention tactics are being used to lower churn and boost loyalty. Customer loyalty in the telecom industry is far from guaranteed. Many companies combined its cellular and wire line customer loyalty programmers as part of a reorganization that aimed to give customers more rewards options. Customers can now take advantage of a variety of benefits for their loyalty, including: hardware enhancements include fiber optic service and internet content. To lower their churn rates, telecoms are investing more and more in cutting-edge technologies. Tools for customer analytics and insight can be used to forecast customer behavior, allowing service providers to decide which strategies will increase retention rates the most effectively. Additionally, cutting-edge technology uses artificial intelligence and machine learning techniques to pinpoint the clients most likely to churn. From this discussion it is observed telecom industry playing a very major role in our daily life and generating an enormous amount of customer base. New customer base generation is costlier than retaining the existing customers. Churning is the process of switching of customers in a given time. Telecom management and analysts find the explanations behind customers leaving subscriptions and behavior activities of the holding churn customers' data. The model proposed uses classification techniques to find the leave subscriptions and collects the reasons behind it in the telecom industry. The aim of this model is to analyze the various machine learning algorithms

required to develop customer churn prediction models and identify churn reasons in order to give them with retention strategies and plans. The main motivation behind churn prediction in the telecom sector is to reduce churns and retain the existing customers. Section 2 emphasizes system architecture of proposed system and proposed system model is presented in detail. Sequentially, Section 3 focuses on detailed dataset description based on single dataset including data pre-processing, and feature selection and also highlighted multi dataset approach concisely. In Section 4, experimental analysis using various ways such as decision tree classifier, Random Forest algorithm, survival analysis, Cox proportional hazard model, and retention strategy is presented. Finally, concluding remarks are summarized in Section 5.

2. System architecture

In this section, system architecture and proposed system model are discussed subsequently.

2.1. System architecture of the proposed system

The implementation for churn will require the latest version of Anaconda with built in features that consist of Jupiter notebook for training and testing data. The latest version of Anaconda with built-in functionality, like Jupiter notebook for training and testing data, will be required for the churn implementation. Churn predictions for the telecom industry have been carried out using literature with various methods that includes machine learning algorithms, data mining techniques and retention strategies. These techniques effectively support many companies for predicting, identifying, and retaining churners which help in CRM (Customer relationship management) and decision making. CRM deals with the data to identify a loyal customer for industry. High revenue generating customers (loyal customers) for a company have no impact on the competitor companies. Such loyal customers help to grow profitability of a company by referring to the other people such as their family members, colleagues, and friends. Hence, the role played by CRM is very important in churn prediction and it also helps to retain the churning customers. Fig. 1 depicts a cycle through which churn prediction can be made. For prediction there are many algorithms such as Support vector machine, K-nearest neighbor, J48, naive Bayes, logistic regression, LWL, Random Forest, Decision tree classifier which are used to resolve classification problems. Random forest and Decision tree classifier are considered relevant with better accuracy and performance.

2.2. Proposed system model

Let S is the proposed system.

$$S = \{I,\ O,\ DP,\ FS,\ EF\}$$

Where

1) I (Input): Dataset
2) (Output): CM (Confusion matrix) ={R->Actual False, Actual True} {C->Predicted False, Predicted true}
3) DP (Data Processing)
4) FS (Feature Selection) is measured by Pearson correlation formula (Cr) where set of numerical attributes is taken 'X' ranging from X_1, X_2, X_3……., Xn consider a set of attributes X, form a subset of two attributes each, X= {{X_1, X_1,}, {X_1, X_2}, {X_1, X_3} .... {Xi, X_j}… {X_n, X_n}}

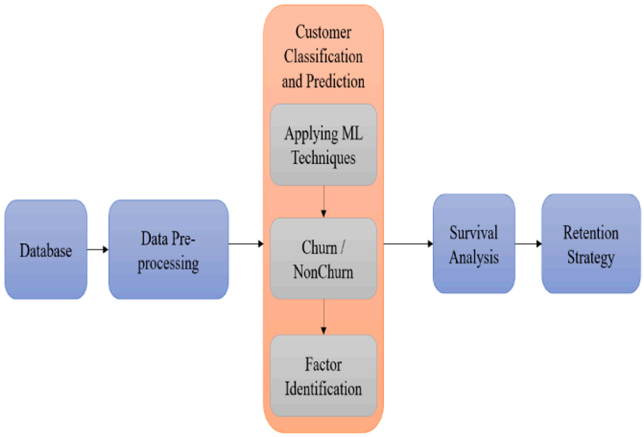


Fig. 1. Proposed system.

If each subset = X then evaluates $\sum X_i$, $\sum X_j^2$ & $\sum (X_i * X_j)$
Cr= $\sqrt{\sum} (X_i * X_j) \div \sum X_i^2 * \sum X_j^2$
Where Cr = 1 (+ve correlation), Cr = 0 (no correlation), Cr = −1 (-ve correlation)
Convert string values into numerical (S→N)
Then apply ML algorithm on new Dataset (consider as C)

5) EF – Efficiency of Proposed Model

The proposed system for churn prediction is derived using accuracy, precision, recall, f-measure. Accuracy calculates the accuracy metric.

TP=True positive, TN=True negative, FP=False positive, FN=False negative values, AP=Actual positive, AN=Actual negative.

a) Accuracy= (TP + TN) ÷ (TP + TN + FP + FN)
b) True positive and true negative are values of the confusion matrix after applying classification algorithms. True positive rate is the value that shows us which part of data is classified as correct and false positive classifies incorrect values.
c) (TP) rate = TN ÷ TN +FP
d) (FP) rate= FP ÷ FP+TN
e) Precision = TP ÷ (TP + FP)
f) Recall = TP ÷ (TP + FN)
g) F-Measure = (2 × Precision × Recall) / (Precision + Recall)

Constraints: The time is the key for predicted churners. If C (customer) churns after one month, but sometimes dataset shows that it would churn after one week that will lead to incorrect churn. Hence, C depends upon T. Incorrect dataset (ID) with irrelevant features will be other constraints to predict churn model.

Failure of S: If customer (C) gets churned out for one month time (T) but then customer rejoins again would lead to loss of company and system (S).

3. Dataset description

Here, specifically focus is on Single Dataset presented in sub Section 3.1. Sequentially, Section 3.2 emphasizes multiple datasets to validate proposed system.

3.1. Single dataset

The dataset used for experiments in this paper, contains results of Telco-Customer-Churn dataset obtained from Kaggle website (it is also known as IBM Watson dataset which was released in 2015).Each row represents a customer, each column contains attribute described on the column Metadata. It consists of 7043 customer information. Every customer has 21 features and the "Churn" it contains 11 missing values in the Total Charges column. The last attribute contains labelled data with two classes where 26.53 % of total customers are labelled as "T'' indicating true customers i.e., categorized as churning customers and the remaining 73.46 % customers are labelled as "F'' indicating false customers i.e., categorized as non-churning customers. The attribute selection depends on the results of techniques of feature selection that find useful, the most similar and effective attributes to predict the churning customers. A total of 5174 are non-churners and 1869 are churners. The dataset contains 16 categorical columns and 5 numeric columns. The dataset helps to figure out customer prophecy and build retention possibilities. Figs. 2 and 3 shows details of database.

3.1.1. Data pre-processing

This step is required to remove all the irrelevant and dirty data of real world. As the data is congregated from many resources it is important to overcome this issue. Without execution of this step decision makers cannot predict outcomes even if they did it won't be

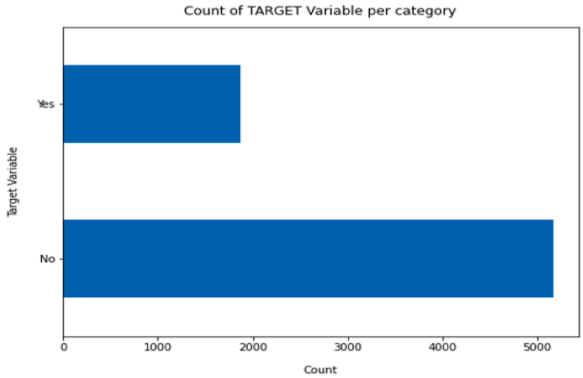


Fig. 2. Count of target Variable per category.

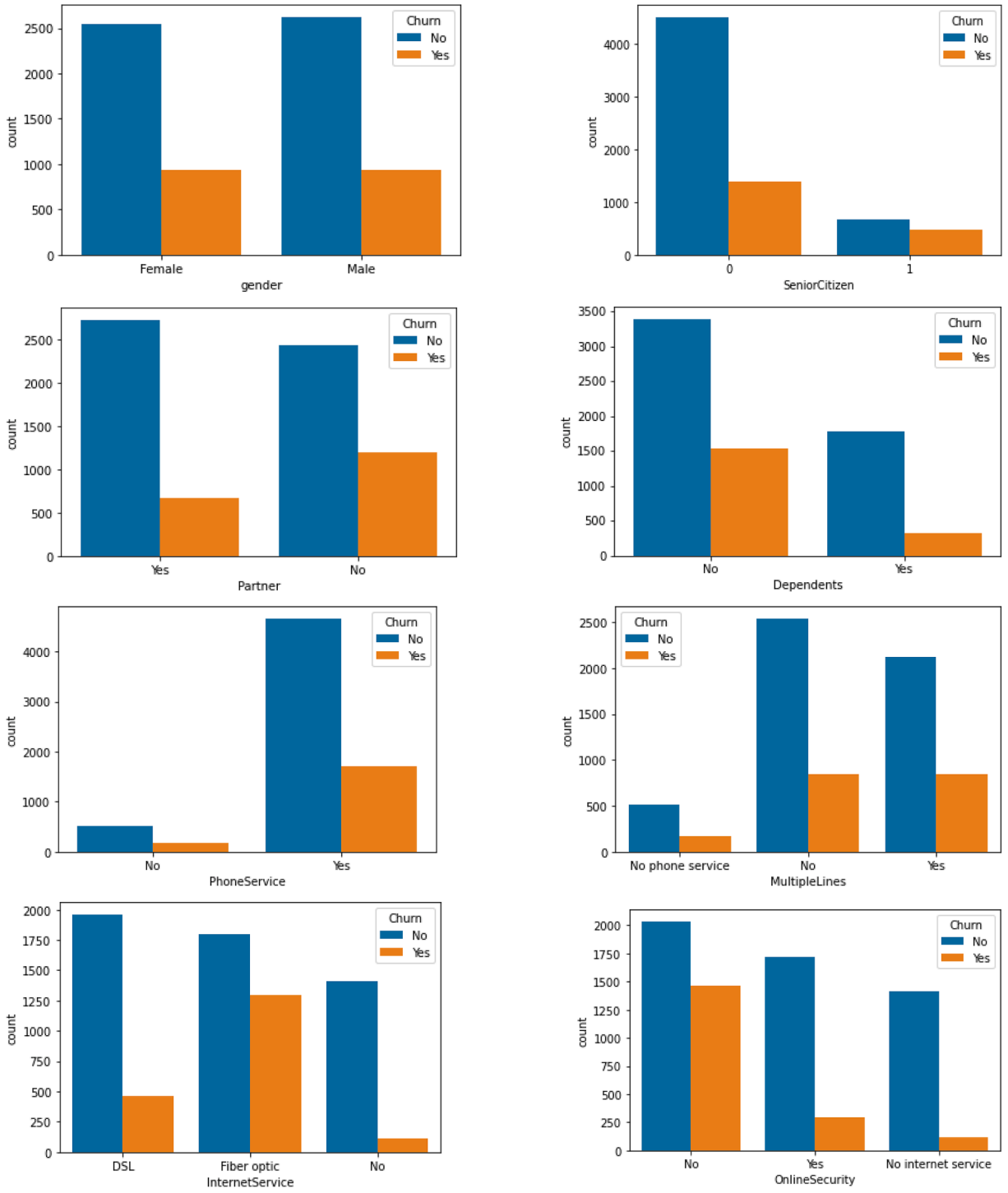


Fig. 3. Database attribute details.

correct that leads to no quality data and no quality mining. To solve these issues following methods are considered for cleaning the data. Likewise, there are many such features taken as input to predict model. Classification of data and performing pre-processing cleans the data and makes it easy to use.

Data classification and pre-processing clean the data and make it easier to use. There are three steps in data preparation.

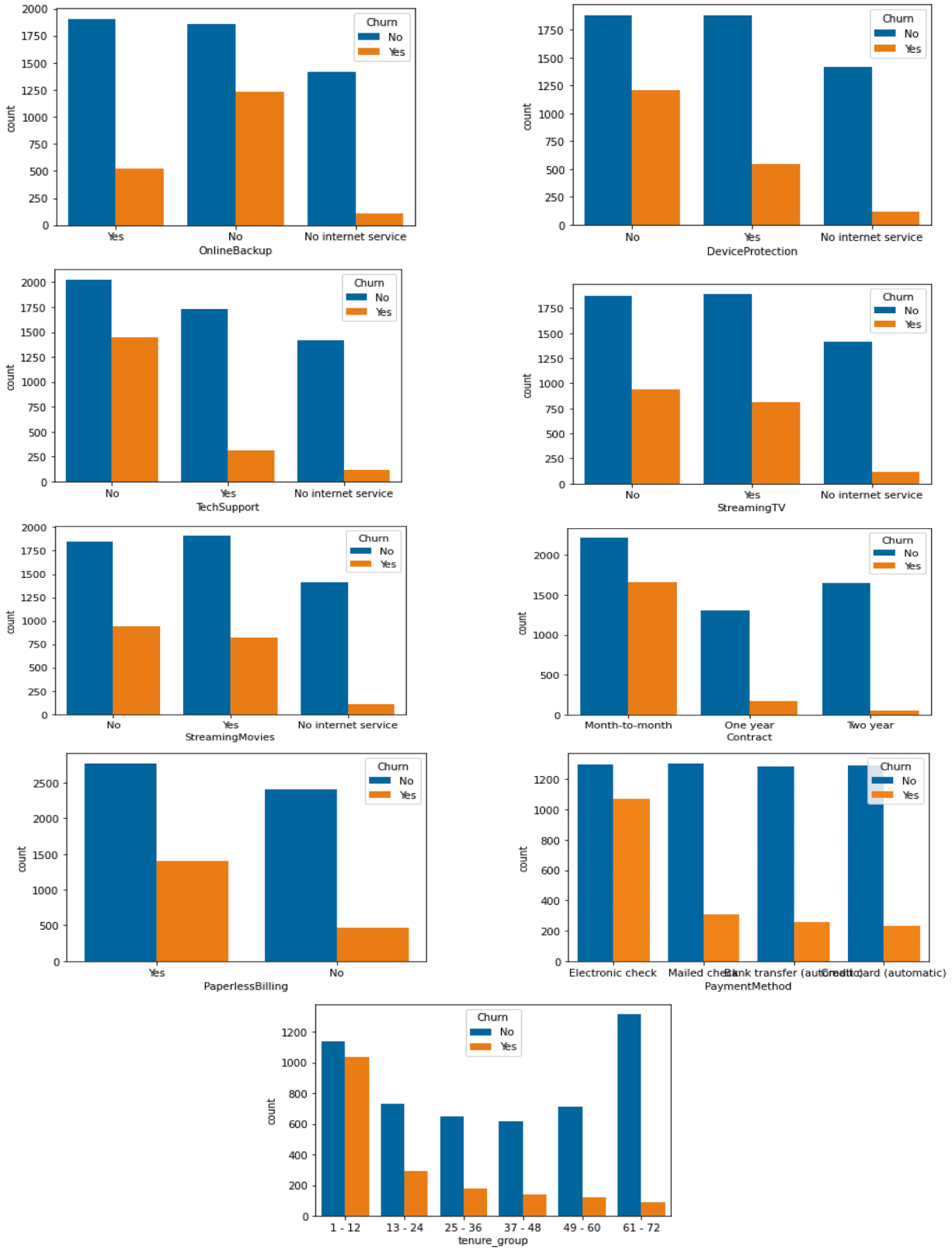


Fig. 3. (*continued*).

- A data comparison is carried out in the first stage to identify redundant data. The properties that are repeated are immediately removed.
- Each instance searches for the missing value in step two. If missing values are detected, replacement procedures are used; otherwise, missing values are replaced with the most suited value. When there is no way to replace the instances, they are discarded.
- The final stage entails defining the type of each value in order to eliminate extraneous data. If the information is no longer useful, it is discarded. In other words, noisy data is deleted from the data that has already been pre-processed. Check the data types of each column first.

Because the total charges column's data type is object type, we convert it to numeric and make a duplicate of the underlying data to manipulate and process. This conversion is performed using panda's library imported in the program and using to-numeric function of library. Next, check for duplicate values in dataset, but there are no duplicate values present in dataset. Further continue checking for missing values. So according to general thumb rule the feature with less missing values there to fill means values or simply use regression to predict the missing values based on the particular feature. Similarly, in case of feature with high number of missing values, it would better to drop those columns due to less analysis and insights. Also, the columns with more than 30–40 % are deleted. Now by using is null and sum function there are 11 missing values present in the Total Charges column so drop this missing value. The tenure consists of maximum value of 72 and also dividing the tenure into 6 categories of 1–12, 13–24, 25–36, 37–48, 49–60, 61–72 make easy in visualization of column data of tenure period. Dropping the columns that are not required for processing. It was found the customer ID is not useful and contains unique value which won't affect the prediction results, so drop customer ID and similarly drop the tenure column which is not necessary to evaluate the results.

1–12	2175
61–72	1407
13–24	1024
25–36	832
49–60	832
37–48	762

Name: tenure_group, dtype: int64

Univariate analysis - In data exploration, the distribution of individual predictions by churn will be determined, and the number of churners representing non-seniors will be determined. Here, 1 represents senior citizens, and the plot demonstrates that if a client is a senior citizen, they are more likely to churn. Similarly, when a candidate is single and does not have a partner, the partner churn ratio is high. People with phone service are more likely to churn, similarly payment method, if there is an electronic check appears higher than the case of credit card appears lowest churner because of this they could be having auto debit features, which is one of the important features for churn, similarly remaining people with phone service are more likely to churn, similarly remaining people with phone service are more likely to churn, similarly remaining people with phone service are more likely to churn, remaining figures also shows the churn count respectively above selected features.

Bivariate analysis - It is used to find a value prediction for a single variable. Correlations between variables are simple to find. A relationship between two variables is defined as bivariate. There are numerous features in our dataset, and we presented the results. Two variables were examined, and two new data frames for churners and non-churners were generated. A function is created that maintains a data frame that is passed with column, title, and hue information for each feature, similar to how bar graphs for different features may be shown in diagrams. Gender characteristics considers 2500 female and male participants, with a churner/non-churner ratio of around 50 % for each gender. According to gender feature analysis, females are more likely to churn if they have a relationship, but males are more likely to churn if they do not have a partner. The groups are classified into churners and non-churners.

3.1.2. Feature selection

This is an important step in achieving our model's goal. Unnecessary data is discovered in datasets while training the model, resulting in a reduction in model accuracy. As a result, feature selection on a dataset is used to solve these issues.

The following are the advantages of feature selection:

- Reduced over fitting means less chance of making conclusions based on noise.
- Accuracy is improved because there is fewer misleading data.
- Training time is reduced providing lesser algorithm complexity with algorithms that train faster.
- Applying Machine Learning Algorithms

[1] *Decision tree*: Decision tree are used to solve both classification and regression problem in the form of trees that can be incrementally updated by splitting the dataset into smaller dataset, where the result are represented by the leaf node. Each branch represents the possible decision outcome or reaction. It is like a flowchart diagram that shows the various outcomes from a series of decisions. It can be used as a decision-making tool, Decision tree has some series of same craft questions regarding attributes of test data record and it's use to solve classification-based problems, every time it gets solution from it and follows until the final conclusion of class label record. Then visit several decision trees for achieving target value. It can be either true or false. Now pick a majority vote of trees or count the target values provided, then based upon decision trees predict

if customer churn is true or false for research analysis, or for planning strategy. A primary advantage for using a decision tree is that it is easy to follow and understand. Decision tree classifier is simple and adaptive classification technique, this is basically implying a straightforward process to analyze and solve the problem.

[2] *Random Forest tree*: The random forest is a classification algorithm consisting of many decision trees. More numbers of trees in the forest led to more robustness in prediction with higher accuracy. It uses bagging and feature randomness, when building each individual tree to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. In this algorithm, each tree will have its own output from a dataset provided, such that output generated will be considered from the majority of trees. Decision trees made in this algorithm are of numeric type in which the tree picks any random attribute in the dataset. Advantages of Random Forest that it helps to solve both regression and classification problems.

[3] *Important Features or factors responsible for churning according to Decision Tree model*: Finding the important factors that are responsible for churning makes it possible to find the service that is required to customer to prevent from attrition. This can be done using feature importance. Here the feature is ranked according to their importance. The most important feature at the top of list, while least important are at end of list.

In Fig. 4 it is seen that Contract-Month-To-month ranked first in the list with its importance as 0.517, Total Charges with importance as 0.104, No Internet Service is 0.093, DSL Internet Service as 0.0795, Monthly Charges as 0.0517, Contract of Two years as 0.0464. Contract of one year as 0.0419, tenure group of 1–12 as 0.0397, No Streaming Movies as 0.00696, Fiber Optic Internet Service as0.00643, and last in list is with No Paperless Billing as 0.004727.

[4] *Important Features or factors responsible for churning according to Random Forest Tree model*: In this model from Fig. 5 the features importance ranked from 1st to last as,

- Contract month to month = 0.1245
- Tenure group = 0.10518
- Internet Service Fiber Optic = 0.07693
- Total Charges = 0.07000
- Contract Two year = 0.06123
- Tenure group 61 -72 = 0.04302
- Online Security yes = 0.03977
- No tech support = 0.03943
- Online Backup No Internet Service = 0.0360
- Streaming Movies No Internet Service = 0.03479
- Internet Service DSL = 0.0338
- Monthly Charges = 0.0309
- Online Security No Internet Service = 0.0294
- Contract One year = 0.02819
- No Online Security = 0.02620
- Tech Support Yes = 0.02491
- Tech Support No Internet Service = 0.024313
- Partner Yes = 0.02422
- No Multiple Lines = 0.001289

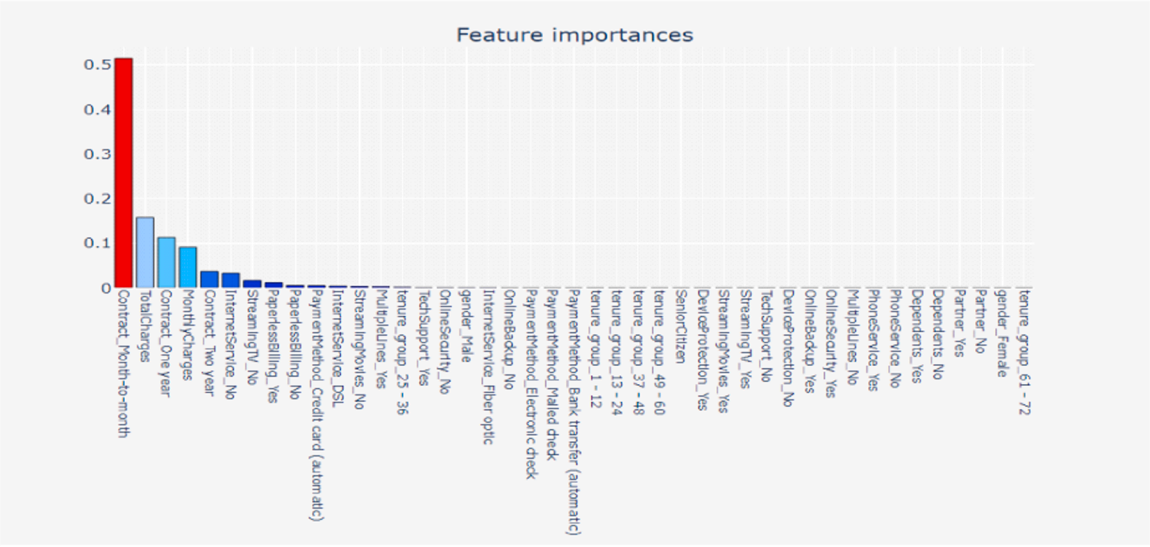


Fig. 4. Feature importance for decision tree.

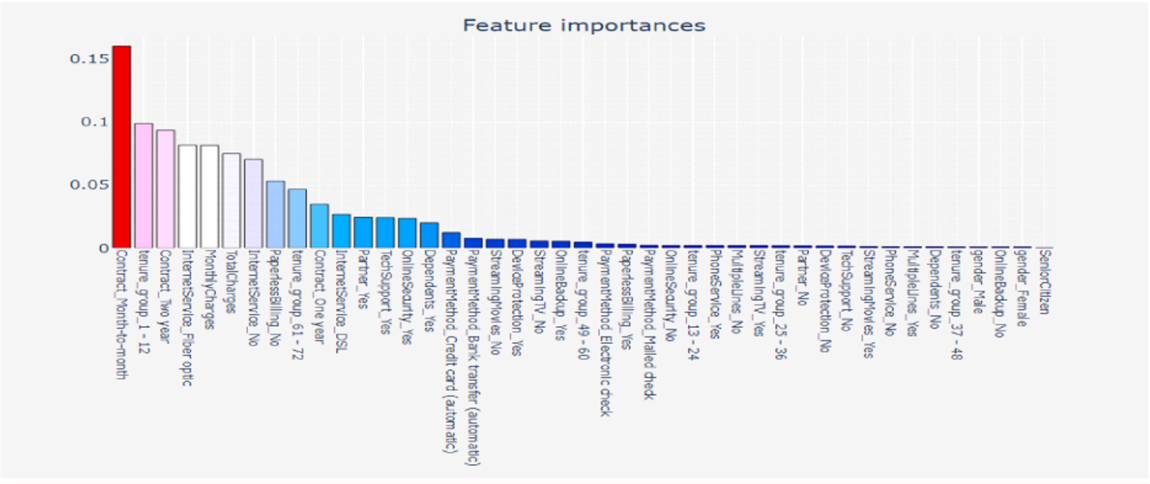


Fig. 5. Feature importance for Random Forest classifier.

- Gender Male = 0.0011822
- Multiple Lines No phone service = 0.0011689
- No phone Service = 0.001107
- Gender Female = 0.001101

3.2. Multiple dataset

There exist multiple types of data in SyriaTel [25] as classified below which may be applied to construct our churn prediction model used in telecom sector. It is observed to present dataset structure using spark engine, it is required to include phase-based exploration with suitable pre-preparation for algorithms based on classification.

3.2.1. Towers and complaints database

In this database, the detailed data of actual location is shown as in the form of digits. The serial digits are mapped with database of towers which offers the actual location of the transaction and providing state, city, area, sub-area, latitude and longitude. Database including various complaints provides all submitted complaints along with inquiries statistics based on coverage, any problem related to the telecom business and issues related to packages and offers.

3.2.2. Customer data

Customer data comprises the data-based contact information and services of customers. Moreover, various packages, services and offers taken by customer. Also, it contains CRM system including information generated from all customer GSMs such as gender, birth date, the location and type of subscription etc.

3.2.3. Network logs data

Network logs data includes sessions related to calls, SMS and internet for each transaction used by telecom operator such as required time to initialize a session for call ending and to check internet status. Moreover, it represents whether session is expired or not due to error occurred in the internal network.

3.2.4. Mobile IMEI information

Mobile IMEI information comprises the model, type, model of the mobile phone and whether mono or dual SIM device. Data may have large size which may require information in detailed. It requires a lot of time for understanding. It also needs to know the original sources along with format for storage. Moreover, related to these records, data must be linked to each other logically using relational databases that actually represent customers detailed information.

3.2.5. Call details records

Call Details Records (CDRs) includes modifiable data related to MMS, calls, SMS etc. Also, transaction made by customers using internet which is ultimately generated in the form of text files.

4. Experimental analysis

After preprocessing on dataset following are the observations:

- Strong correlation exists between tenure and total charges, means as tenure increases so does total charges.
- Strong correlation exists between monthly charges and total charges as well.
- Tenure and Contract duration seems to be strong factors in determining churn.
- Among service types, phone service seems to be most popular.
- CSP should investigate if customers receiving digital invoice have any concern with understanding the bill details.
- Also, they should encourage customers to move to automated payment modes to improve customer experience.
- Gender does not play an important role. However, CSPs should take care of the experience of senior citizens.

From Figs. 6 and 7 it is seen that churn is higher when monthly rates are high. Even with modest overall charges, there is large churns, as shown in the Fig. 7. When all three parameters (Total Charges, Monthly Cost, and Tenure) are combined, higher monthly charges at low tenure result in lower total charges, implying that all of these characteristics are associated to higher churn. Fig. 8 shows used all features.

4.1. Experiment analysis using decision tree classifier

So, in our proposed model using a Decision Tree classifier, the obtained accuracy of the model is 78 %, which is very low, and printing the classification report led to the dataset unbalance, which results in less accuracy. Decision Tree classifier took Number of Leaves - 331, Size of the tree - 552, Time taken to build model-1.06 s for execution.

As a result, the accuracy of the Decision tree classifier model before up-sampling & ENN should not be used as a meaningful measure because it leads to unbalanced datasets. As a result, when checking recall, precision, and F1 scores for the minority class, it's clear that the precision, recall, and F1 ratings for Class 1, i.e. churned consumers, are far too low.

For up-sampling training data into a decision tree classifier that differentiates into x train and y train and creates a prediction variable and calls the classification input to process input to produce output with accuracy, and using SMOTE (synthetic minority oversampling technique) by performing oversampling and cleaning using ENN (Edited Nearest Neighbors), the dataset is balanced with dataset values of 493, 40, and 599, 34, and it provides a solution. Tables 1 and 2 shows details results of Decision tree classifier model before and after up-sampling & ENN. The precision of the classifier matrix is 93 %, the recall factor is 93 %, and the F1 score is 93 % and provides 93.85 % accuracy.

4.2. Experimental analysis using random forest algorithms

Missing values are handled with carefully, and accuracy is maintained. It's even capable of handling big, multi-dimensional datasets. So, the accuracy of our model using the Random Forest Tree classifier is 98.91 %. Table 3 shows results of Random Forest classifier model before Up-sampling and ENN. There are several techniques used in random forests. Generally bagging technique is used known as ensemble classifier.

For the minority class of churned consumers, employ a classification matrix to increase the model's performance. The imbalance database and its merely oversampled minority class can be addressed by using SMOTE (synthetic minority oversampling technique) and ENN (edited nearest neighbors). In the required minority class, it would include duplicate examples.

The results of the Random Forest classifier model after up-sampling and ENN are shown in Table 4. As a result, the random forest classifier predicts churn with an overall accuracy of 99 %. The classifier matrix has a precision of 99 %, a recall factor of 99 %, and an accuracy of 99.09025616471152 %.

4.3. Survival analysis

The survival analysis technique is a valuable statistical technique for predicting how long a client would keep a subscription when

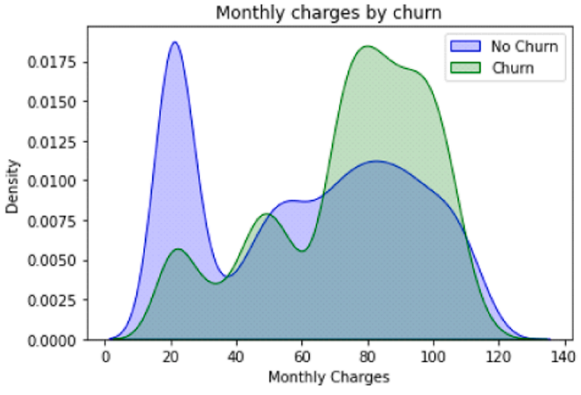


Fig. 6. Monthly charges by Churn.

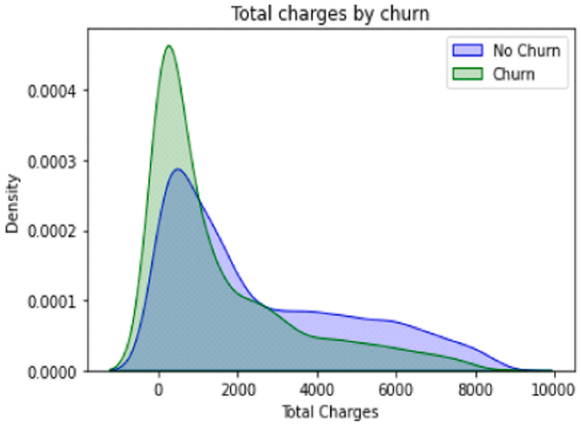


Fig. 7. Total charges by Churn.

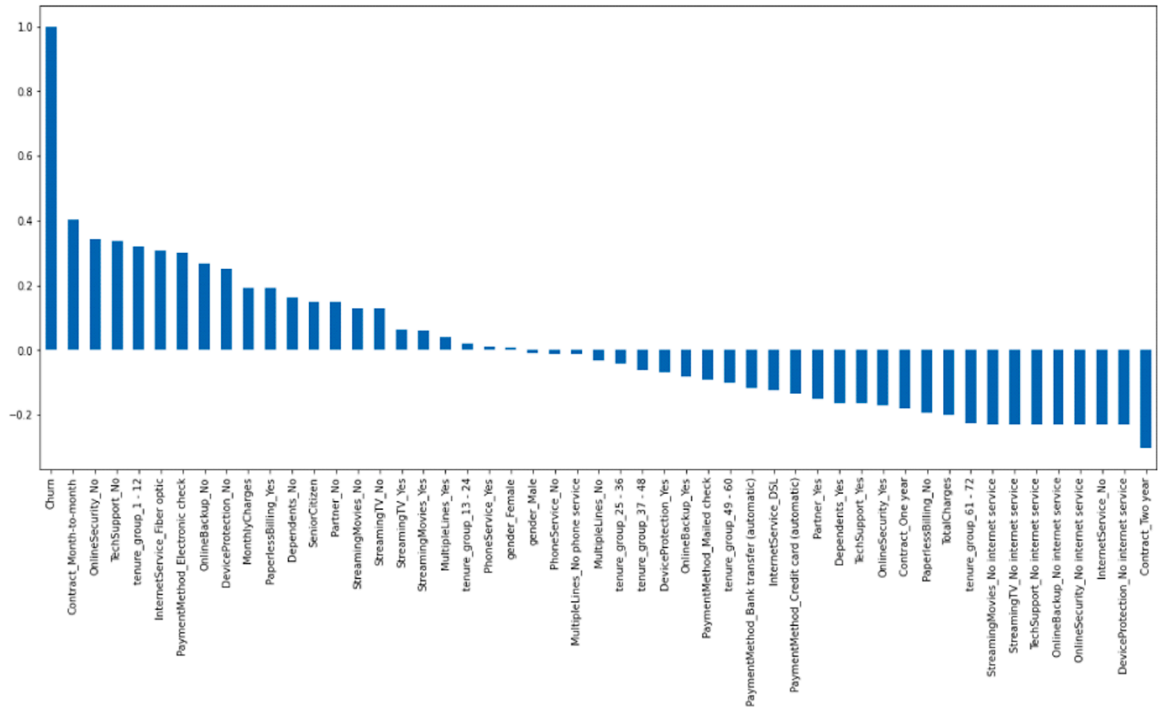


Fig. 8. All features in bar graph.

Table 1
Results of decision tree classifier model before up-sampling & ENN.

	Precision	Recall	F1-score	Support
Class 0 (No)	0.83	0.89	0.86	1046
Class 1(Yes)	0.61	0.49	0.54	361
Weighted avg	0.77	0.78	0.77	1407

	No. of Instances	Percentage
Correctly Classified Instances	5493	77.9923%
Incorrectly Classified Instances	1550	22.0077%

Table 2
Results of decision tree classifier model after up- sampling & ENN.

	Precision	Recall	F1-score
Class 0 (No)	0.93	0.92	0.93
Class 1(Yes)	0.94	0.94	0.94
Weighted avg	0.93	0.93	0.93
	No. of Instances	Percentage	
Correctly Classified Instances	6615	93.85%	
Incorrectly Classified Instances	427	6.15 %	

Table 3
Results of Random Forest tree classifier model before Up-sampling & ENN.

	Precision	Recall	F1-score
Class 0 (No)	0.99	1.00	0.99
Class 1(Yes)	0.99	0.96	0.97
Weighted avg	0.99	0.99	0.99
	No. of Instances	Percentage	
Correctly Classified Instances	6971	98.91%	
Incorrectly Classified Instances	77	1.09%	

Table 4
Results of Random Forest classifier model after Up-sampling and ENN.

	Precision	Recall	F1-score
Class 0 (No)	0.98	1.00	0.99
Class 1(Yes)	1.00	0.98	0.99
Weighted avg	0.99	0.99	0.99
	No. of Instances	Percentage	
Correctly Classified Instances	7010	99.09%	
Incorrectly Classified Instances	38	0.91 %	

they churn. "Time to event analysis" is another name for survival analysis. Customer retention is heavily influenced by survival analysis. To avoid churn, we concentrate on a large number of consumers with a short survival span. This analysis determines the value of a customer's life time. The event is defined as the precise time when a customer cancels or leaves a subscription, and the time is specified as the time when the consumer joins the service.

Survival function: -

$$\mathrm{S(t)} = P\ \mathrm{r}\ (T > t) = 1 - \mathrm{F(t)} = dx$$

Here T = event time, f(t)= density function

4.4. Cox proportional hazard model

'Time-to-event' data is analyzed using the Kaplan-Meier (KM) approach. All-cause mortality is a common outcome in KM analyses. However other outcomes such as the occurrence of a cardiovascular event could also be included. The Cox Proportional Hazard model is useful to predict better survival probability of individuals. In this model, some characteristics include partner, monthly costs, phone service, gender, and remaining variables are covariates, which impute on the survival probability, taking into account each customer's tenure at the time they churned. All variables and survival functions are likewise included in this model.

The log-hazard of an individual is a linear function of their variables and a population-level baseline hazard that changes with time, according to Cox's proportional hazard model. Mathematically:

There are a few things to notice about this model: the baseline hazard, b0, has the only temporal component (t). The partial hazard is a time-invariant scalar element in the preceding equation that solely raises or decreases the baseline hazard. As a result, changes in variables will only affect the baseline hazard.

Fig. 9 show the coefficient in another way. For instance, the coefficient for PhoneService Yes (having a phone service) is around 0.69. In the Cox proportional hazard model, a one-unit increase in PhoneService Yes increases the baseline hazard by a factor of exp (0.69) = 2.00, or nearly 20 %. A greater hazard indicates that the event is more likely to occur. The hazard ratio is defined as exp (0.69) divided by 1.

The key thing to notice here is that although though the (coef) values for covariates MonthlyCharges and gender Male are close to

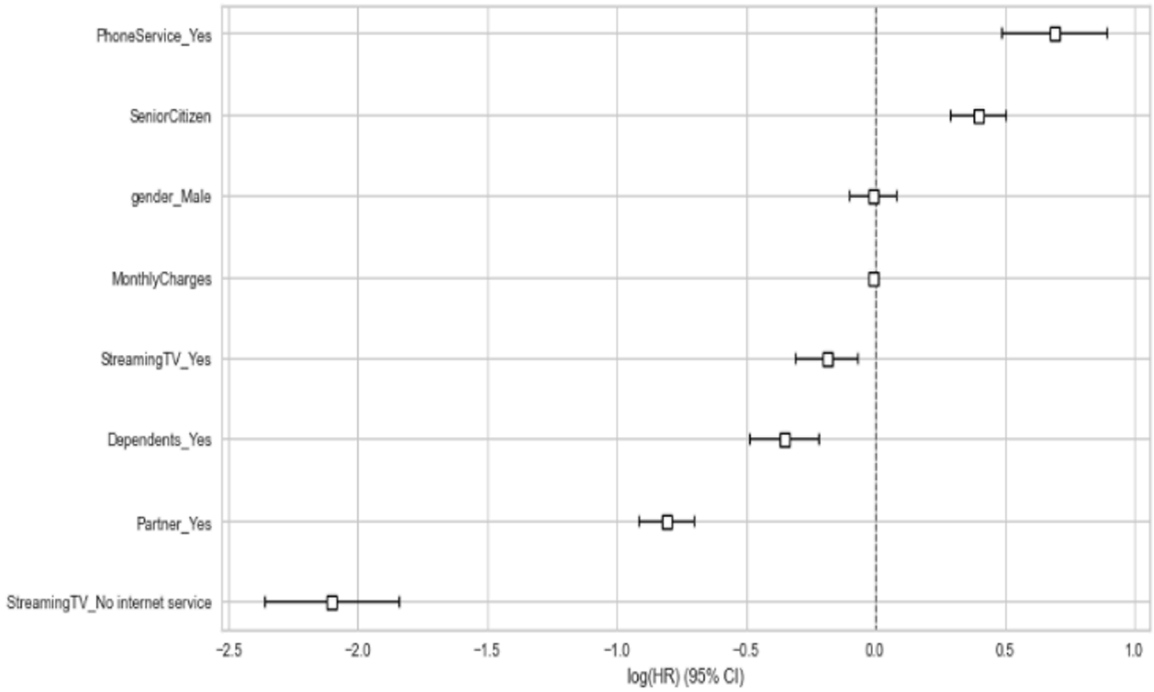


Fig. 9. Significance of covariate in predicting churn risk.

zero (−0.01), the former still has a substantial impact in forecasting churn, whereas the latter is inconsequential. The reason for this is because MonthlyCharges is a set of continuous values that can change from one month to the next.

Fig. 10 shows survival curve for the selected customers. So, based on the Fig. 10, it is concluded that customer 2 has the highest chance of churning. Creating survival curves at the customer level allows us to develop a proactive plan for high-value clients for various survival risk segments along the timeline.

4.5. *Retention strategy*

Some high-effort interactions to support customer retention in the telecom industry include Customers Must Be Educated, attractive

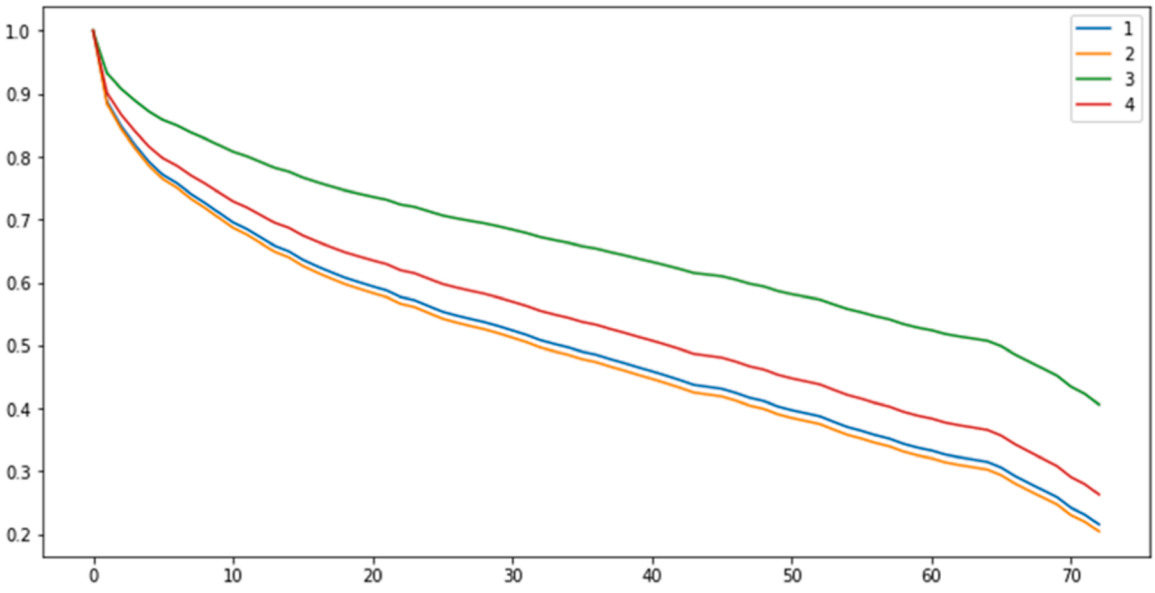


Fig. 10. Survival curve for the selected customers (Customer 1, 2, 3 and 4).

Offering Plans, repeat contacts, an emphasis on wasting customers' time, shoddy self-service, avoiding unneeded robotic service and giving complicated instructions.

5. Conclusion

Our system works effectively for achieving the major objective to analyze the various machine learning algorithms required to develop customer churn prediction models and in identification of churn reasons in order to give them with retention strategies and plans. Our system focuses on resolving an unavoidable issue Customer churn arise in the telecommunications sector for a variety of reasons. The most perplexing aspect about client turnover is that churn is impossible to manage. Moreover, customer turnover has several causes, some of which are visible and others which are not. On the other hand, operators in the telecommunications industry should be aware that client loss will occur sooner or later and they must be prepared to respond. In this context, our system plays important role on solving the problem of customer churn has become critical for telecom companies' survival.

A customer churn system is based on machine learning methods, a decision tree classifier, and a random forest algorithm. Here, prior to and after up sampling and ENN, both techniques are applied to the system. Decision tree classifier models initially produced poor results on an unbalanced dataset that did not take ultimate accuracy into account when matrices were used to evaluate the model. In comparison to a decision tree classifier, a random forest classifier produces better results. With an overall accuracy of 99 %, the random forest classifier predicts churn. The classifier matrix has a precision of 99 %, a recall factor of 99 %, and an accuracy of 99.09 %. System comprises churn prevention that is based on survival analysis using Cox Proportional Hazard model and retention plans. Survival curve is used for the selected customers plays an important role in customer churn prediction. In the telecom sector, it appears obvious that lowering customer effort is a method to boost customer retention. This churn prevention system may be made more complex and sophisticated in the future to provide better and more precise recommendations. More advanced algorithms such as deep learning recurrent neural networks that help to identify nonlinear complex relationships between data variables which may be applicable in future study to estimate survival likelihood. Moreover, customer churn prediction in the telecommunication sector is possible a rough set approach and data certainty [24] in future.

Financial and ethical disclosures

This work is not supported fully or partially by any funding organization or agency.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] Ullah I, Raza B, Malik AK, Imran M, Islam SU, Kim SW. A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. IEEE Access 2019;7:60134–49. https://doi.org/10.1109/ACCESS.2019.2914999.
[2] Ahmed AAQ, Maheswari D. Churn prediction on huge telecom data using hybrid firefly based classification. Egypt Inform J 2017;18(3):215–20. https://doi.org/10.1016/j.eij.2017.02.002.
[3] E V, Ravikumar P, S C, M SK. An efficient technique for feature selection to predict customer churn in telecom industry. In: Proceedings of the 1st International Conference on Advances in Information Technology (ICAIT); 2019. p. 174–9. https://doi.org/10.1109/ICAIT47043.2019.8987317.
[4] Jain H, Khunteta A, Srivastava S. Churn prediction in telecommunication using logistic regression and logit boost. Procedia Comput Sci 2020;167:101–12. https://doi.org/10.1016/j.procs.2020.03.187.
[5] Pamina J, Beschi R, SathyaBama S, Soundarya S, Sruthi MS, Kiruthika S, Aiswaryadevi VJ, Priyanka G. An effective classifier for predicting churn in telecommunication. J Adv Res Dyn Control Syst 2019;11(01–Special Issue). https://ssrn.com/abstract=3399937.
[6] Berger P, Kompan M. User modeling for churn prediction in e-commerce. IEEE Intell Syst 2019;34(2):44–52. https://doi.org/10.1109/MIS.2019.2895788. March-April.
[7] Gaur A, Dubey R. Predicting customer churn prediction in telecom sector using various machine learning techniques. In: Proceedings of the international conference on advanced computation and telecommunication (ICACAT); 2018. p. 1–5. https://doi.org/10.1109/ICACAT.2018.8933783.
[8] Borja B, Bernardino C, Alex C, Ricard G, David MM. The architecture of a churn prediction system based on stream mining. Front Artif Intell Appl 2013:157–66. https://doi.org/10.3233/978-1-61499-320-9-157.
[9] Yildiz M, Varlı S. Customer churn prediction in telecommunication. In: Proceedings of the 23rd signal processing and communications applications conference, SIU 2015 - proceedings; 2015. p. 256–9. https://doi.org/10.1109/SIU.2015.7129808.
[10] Devriendt F, Berrevoets J, Verbeke W. Why you should stop predicting customer churn and start using uplift models. Inf Sci 2021;548:497–515. https://doi.org/10.1016/j.ins.2019.12.075 (Ny).
[11] Saran Kumar A, Chandrakala D. A survey on customer churn prediction using machine learning techniques. Int J Comput Appl 2016;154:13–6. https://doi.org/10.5120/ijca2016912237.
[12] Kraljević G, Gotovac S. Modeling data mining applications for prediction of prepaid churn in telecommunication services. Automatika 2010;51(3):275–83. https://doi.org/10.1080/00051144.2010.11828381.
[13] Hung SY, Yen DC, Wang HY. Applying data mining to telecom churn management. Expert Syst Appl 2006;31(3):515–24. https://doi.org/10.1016/j.eswa.2005.09.080.

[14] Alzubaidi AMN, Al-Shamery E. Predicting customer churn in telecom sector based on penalization techniques and ensemble machine learning. Int J Eng Technol 2018;7:657–64. https://doi.org/10.14419/ijet.v7i4.19.27977.
[15] Khalid LF, Mohsin Abdulazeez A, Zeebaree DQ, Ahmed FYH, Zebari DA. Customer churn prediction in telecommunications industry based on data mining. In: Proceedings of the IEEE symposium on industrial electronics & applications (ISIEA); 2021. p. 1–6. https://doi.org/10.1109/ISIEA51897.2021.9509988.
[16] Zhao M, Zeng Q, Chang M, Tong Q, Su J. A prediction model of customer churn considering customer value: an empirical research of telecom industry in China. Discrete Dyn Nat Soc 2021;2021:12. https://doi.org/10.1155/2021/7160527. Article ID 7160527.
[17] Nadeem AN, Umar S, Shahzad MS. A review on customer churn prediction data mining modeling techniques. Indian J Sci Technol 2018;11(27):1–7. https://doi.org/10.17485/ijst/2018/v11i27/121478.
[18] Lalwani P, Mishra MK, Chadha JS. Customer churn prediction system: a machine learning approach. Computing 2022;104:271–94. https://doi.org/10.1007/s00607-021-00908-y.
[19] Rajamohamed R, Manokaran J. Improved credit card churn prediction based on rough clustering and supervised learning techniques. Cluster Comput 2018;21(1):65–77. https://doi.org/10.1007/s10586-017-0933-1.
[20] Amin A, Adnan A, Anwar S. An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes. Appl Soft Comput 2023;137. https://doi.org/10.1016/j.asoc.2023.110103.
[21] Amin A, Al-Obeidat F, Shah B, Al Tae M, Khan C, Ur Rehman Durrani H, Anwar S. Just-in-time customer churn prediction in the telecommunication sector. J Supercomput 2020;76:3924–48. https://doi.org/10.1007/s11227-017-2149-9.
[22] Amin A, Shah B, Khattak AM, Lopes Moreiraet FJ, Ali G, Rocha A, Anwar S. Cross-company customer churn prediction in telecommunication: a comparison of data transformation methods. Int J Inf Manag 2019;46:304–19. https://doi.org/10.1016/j.ijinfomgt.2018.08.015.
[23] Amin A, Anwar S, Adnan A, Nawaz M, Alawfi K, Hussain A, Haung K. Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing 2017;237:242–54. https://doi.org/10.1016/j.neucom.2016.12.009.
[24] Amin A, Shah B, Khattak AM, Lopes Moreiraet FJ, Ali G, Rocha A, Loo J, Anwar S. Customer churn prediction in telecommunication industry using data certainty. J Bus Res 2019;94:290–301. https://doi.org/10.1016/j.jbusres.2018.03.003.
[25] Ahmad A, Jafar A, Aljoumaa K. Customer churn prediction in telecommunication using machine learning in big data platform. J Big Data 2019;6:1–24. https://doi.org/10.1186/s40537-019-0191-6.