**Customer Churn Predictive Analysis in the Telecom Industry Using Random Forest Classifier**

**Anjali A Sunil**  
*Department of Advanced Computing St.Joseph’s University*Bangalore, India  
[anjaliasunil@gmail.com](mailto:anjaliasunil@gmail.com)

**Jesse Tabitha E**  
*Department of Advanced Computing*  
*St. Joseph’s University*Bangalore, India  
[jes.tabi117@gmail.com](mailto:jes.tabi117@gmail.com)

**Dr. Jayati Bhadra***Department of Advanced Computing St.Joseph’s University*Bangalore, India  
[jayatibhadra@sju.edu.in](mailto:jayatibhadra@sju.edu.in)

**Je**  
*Department of Advanced Computing*  
*St. Joseph’s University*Bangalore, India [jes.tabi117@gmail.com](mailto:jes.tabi117@gmail.com)

*Abstract*— The telecom sector, which is distinguished by its vast data generation, is becoming more and more important for contemporary company operations. Identification of client loyalty and churn rates is necessary because managing and keeping consumers is a top priority for telecom companies. Predictive analytics are being used more and more in the telecom sector to manage and keep customers. In terms of predicting customer turnover, models like Random Forest, decision tree, SVM, and Logistic Regression have demonstrated encouraging accuracy rates. With the use of these techniques, telecom firms may increase retention tactics and boost overall business performance while actively managing customer attrition. Using data from an Iranian telecom company, the Random Forest classifier was used to predict customer attrition in the telecom sector. The findings demonstrated that the Random Forest classifier performed better in predicting customer attrition than other methods. The Random Forest classifier is a valuable tool in identifying and retaining loyal customers, contributing to ongoing efforts to enhance customer relationship management and market strategies in the telecom sector.

Keywords— Customer Churn, Random Forest Classifier, Telecom Companies, Customer Retention, Machine Learning Models, Hybrid Models

# Introduction

The telecommunications industry generates vast volumes of data annually due to its continuous expansion. It's imperative to analyze this data to effectively retain and manage customers. Assessing customer loyalty and predicting churn rates is crucial for businesses. These predictions help in identifying customers inclined towards loyalty and those considering switching to competitors. This is particularly significant for contemporary telecom companies aiming to enhance profits through the analysis of customer relationships and the implementation of strategic marketing approaches.

The telecom sector offers industries that provide a range of services such as internet, telephone, and television services. Due to the rising challenges telecom industries face, such as tough competition from OTT platforms such as Netflix and Prime, and the need to invest in the latest infrastructure to support 5G services and other new technologies, customer retention has become a struggle. Customers often end their contract with a telecom industry if they are unhappy with the services and other aspects of the product used. They need to develop products to satisfy customers by keeping track of their purchasing behaviour and offering products to suit the needs of the customers. In this way, they can retain customers and reduce the cost of acquiring new customers.

Telecom companies scrutinize customer attrition rates, a pivotal business metric, as the expenses involved in acquiring new customers far exceed those of retaining existing ones. Therefore, companies often offer customer service to win back defecting clients as it is worth recovering long-term customers rather than recruiting new clients. If a company has a constant high customer churn rate, it means that the business is not sustainable. It is not enough to just get new customers, the customers must stay in your company for a long period of time.

This paper will deal with customer churn or customer turnover, customer retention or customer defection is the loss of customers or clients. Churn prediction models will be used to do predictive analytics for predicting customer churn by assessing their potential to churn.

# Literature review

Throughout the years, numerous researchers have conducted precise and dependable churn prediction studies, which businesses can utilize to formulate suitable customer retention tactics. These researchers employ a range of methods, encompassing machine learning, data mining, and hybrid techniques, to empower telecom companies in recognizing, forecasting, and retaining customers at risk of churning.

Adnan Amin et al[1] proposed an adaptive learning approach using evolutionary computation and a Naïve Bayes classifier, coupled with a Genetic Algorithm(subclass of the evolutionary algorithm). The effectiveness of this suggested method on openly accessible datasets like BigML Telco churn, IBM Telco, and Cell2Cell has demonstrated remarkable success, achieving an impressive accuracy rate of 95% with an F-score of 89%.

R. Sudharsan et al[2] proposed an improvisation of RNN, the Swish Recurrent Neural Network (S-RNN) was introduced for customer churn prediction. This deep learning model incorporated a novel feature selection strategy, resulting in an impressive 95.99% accuracy rate.

Shrestha et al[3] used S-RNN which outperformed traditional models, including RNN, DNN, CNN, and ANN. This innovation holds significant promise for enhancing customer retention strategies in various industries. A study on customer churn prediction in Nepal's telecommunication industry uses the XGBoost technique, achieving an impressive 96.25% accuracy rate. This model is promising for telecommunication companies in Nepal, enabling proactive customer churn management, enhancing retention strategies, and improving business performance.

Another study utilized machine learning techniques and Information Gain Filter Feature Selection Algorithms. Saheed et al[4] used SVM, MLP, RF, and NB classifiers. The results showed a 95.02% accuracy with feature selection, indicating its effectiveness in improving predictive performance compared to 92.92% without feature selection. In the study titled Churn Prediction in Telecommunication using Logistic Regression and Logit Boost, two prominent machine learning techniques were employed by Hemalata et al [5] to tackle the critical issue of customer churn prediction. Their research yielded promising results, with both Logistic Regression and Logit Boost achieving high accuracy rates of 85.1785% and 85.2385%, respectively.

Alboukaey et al [6] employed two models based on different feature extraction methods from multivariate time series data: a statistics-based model and an RFM-based model. Additionally, they explored deep learning techniques with LSTM and CNN-based models. Their results indicated a significant superiority of the LSTM-based model over the CNN-based model, while both LSTM and CNN models performed equally as well as the RFM-based model. All three of these models outperformed the statistics-based model.

Another study by Wassouf et al [7] involved applying various classification algorithms, including random forest classifier, decision tree classifier, gradient-boosted tree classifier, and multi-later perceptron (MLPC). Among these, the random forest classifier demonstrated the highest accuracy at 95.50%.

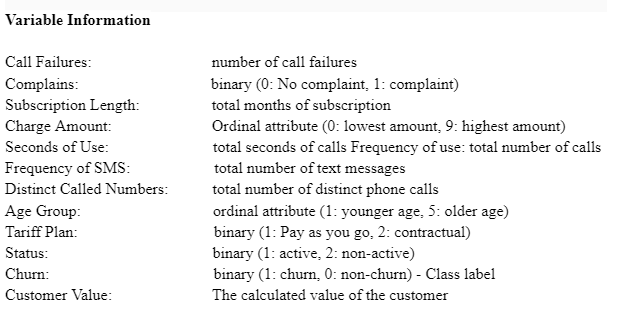
Ullah et al[8] conducted a study focusing on churn prediction models in the telecom sector, utilizing Random Forest and k-means clustering. Their model achieved an accuracy of 0.883, effectively predicting customer churn. This research offers valuable insights into the effectiveness of Random Forest and factor identification techniques in enhancing churn prediction models within the telecommunications industry.

# Metehodology used

This section elaborates on the comprehensive procedure employed in our research, which is centred around predicting customer churn in the telecommunications industry. The primary model used in this context is the Random Forest model.

## Data description

We utilized a dataset sourced from an Iranian telecom company, spanning a 12-month period in the year 2020. This dataset comprises 3,150 records, each representing a customer and featuring information distributed across 13 columns. The dataset encompasses various attributes, including call failures, frequency of SMS, number of complaints, number of distinct calls, subscription length, age group, charge amount, service type, usage in seconds, status, frequency of use, and Customer Value. It's important to note that all attributes, except for Churn, are derived from data collected over the initial 9 months. The Churn labels indicate the customer's status with the company at the end of the full 12-month period. This designated 3-month interval is referred to as the planning gap. The dataset was sourced from the UCI Machine Learning Repository, and the details can be found at this DOI link: <https://doi.org/10.24432/C5JW3Z>.

****

.

## Data pre-processing

## The dataset must be made useful and free from noise, incorrect data and invalid features, before applying the model. This telecom dataset was explored to understand the attributes and their plausible contribution to the model. Null and duplicate values were checked for and the attribute labels were renamed to make the dataset more legible. The unique values and value counts were determined to explore the dataset and the distribution of the categorical features in the dataset. The outliers were removed by filling them up with Nan values and then replacing them with the mean of the attribute values. Now, the dataset is ready for the model to be applied.

## Feature importance

A Feature Importance chart offers valuable information to telecom companies about the important features that affect customer churn. This information can be used to develop strategies to promote customer retention.

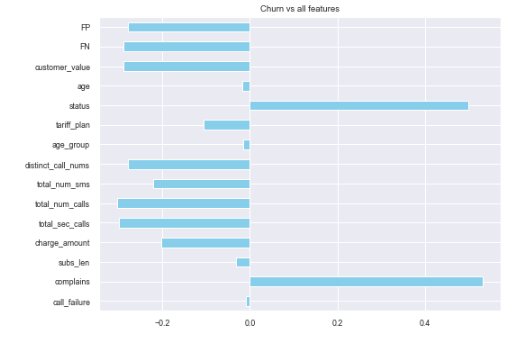


Fig 2. shows the most important features predicting customer churn is customer value, status, complains, total\_num\_calls, total\_sec\_calls and distinct\_call\_nums. This is not surprising as the customers who hold more value are less likely to churn and more likely to be retained. The chart also knows that customers who call and message more are more likely to churn. We see that status and complains are strongly negatively correlated to churn, which means that the more complain and unactive a customer is, the more likely he/she is to churn.

## Proposed methodology

A Random forest classifier is the machine learning technique applied to the churn dataset. This powerful machine-learning technique is generally used for classification, regression and anomaly detection. It can handle non-linear data efficiently and performs better when correlation exists in the dataset. Another benefit of this classifier model is that it can handle a large dataset and it can handle any missing values during model training.

Random Forest is a machine learning technique that leverages an ensemble of decision trees to make predictions. It operates by training multiple decision trees on different subsets of the dataset, and each tree produces its own prediction. These individual predictions are then aggregated to generate a final prediction.

In the context of the Random Forest Classifier, this approach follows a divide-and-conquer strategy. It creates numerous decision trees, each of which selects a random subset of attributes from the telecom churn dataset. These trees grow to a certain depth or level, depending on the features included in their subset. Ultimately, the combination of all these individual decision trees forms the final model used to predict outcomes for the test dataset. This ensemble method often results in improved predictive accuracy and generalization.

Before splitting the data into train and test sets, the categorical features are encoded into numerical values. This is essential as sometimes models do not handle categorical features well. Also, the minority class is sampled to balance the dataset to ensure that the model learns from all classes equally.

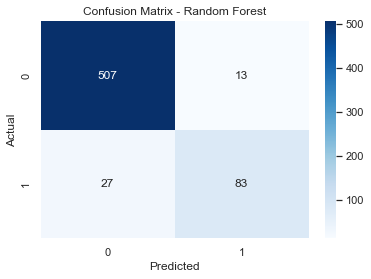
Random Forest proved to be highly effective in our analysis due to its robust handling of our dataset, resulting in superior performance compared to other techniques.

# Results and discussion

The disparate performance metrics from the proposed work with prevailing techniques were analyzed and compared in detail. In order to prove the effectiveness of the work, a performance analysis along with a comparative analysis was performed. The inferences drawn from the proposed model and comparative study are described in this section.

## Performance Analysis of Proposed Classification Technique

## On performing the random forest model for customer churn prediction on our Iranian telecom company dataset, an excellent accuracy of 98.31% was achieved. These are the disparate results of the performance metrics achieved.



*Fig 2. Confusion Matrix of Random Forest Model*

The confusion matrix shown in Fig 2. shows the classification measures. It is seen that the TP and TN values are greater than the FP and FN values. This means that the Random Forest Classifier makes the correct prediction.

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Accuracy | 98.31% |
| Precision | 0.99 |
| Recall | 0.97 |
| F1 score | 0.98 |

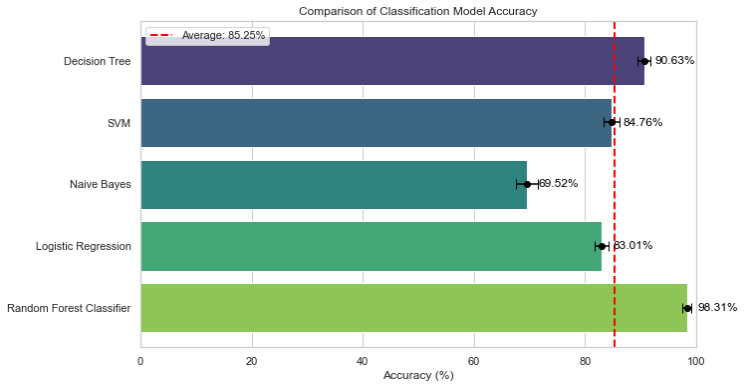
*Table 1. Performance Metrics of Random Forest Classifier*

The high accuracy score suggests that the model is capable of learning complex patterns in the data that can be helpful in making accurate predictions about which customers are most likely to churn from the telecom company. The Precision and Recall values are high, almost close to 1, which signifies the correctness of the prediction model

The model's impressive F1 score indicates its ability to effectively adapt to the dataset and accurately identify customers at risk of churning while minimizing the occurrence of false positives. Avoiding false positives is critical as it prevents businesses from allocating resources to customers who are unlikely to churn. In summary, this study underscores the effectiveness of the Random Forest model as a powerful machine-learning technique for predicting customer churn. It suggests that Random Forest models can be a valuable asset for businesses striving to reduce customer churn rates.

## Comparative Analysis of Machine Learning and Deep Learning Algorithms

In order to perform a comparative study of the telecom dataset using various models, a number of deep learning and machine learning models were applied to understand how each of the algorithms performs with the dataset.

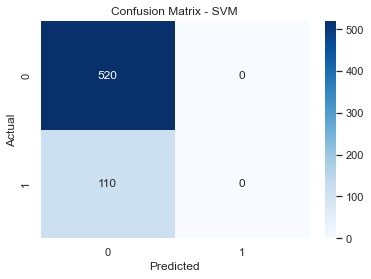


*Fig 3. Graphical representation of Comparison of Machine Learning Models and their Accuracy*

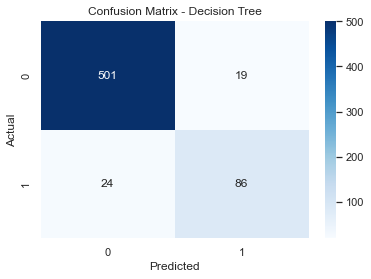
Table 2 provides a comprehensive overview of the performance of various machine learning models, where the Random Forest Classifier stands out with the highest accuracy score of 98.31%. On the other end of the spectrum, the Naive Bayes model yields the lowest accuracy score at 69.52%. This diversity in accuracy scores underscores the varying effectiveness of different models in handling the dataset.

In Figure 3, the graphical representation illustrates the performance of SVM and Logistic Regression models. Notably, these models exhibit accuracy levels that closely align with the average accuracy line observed across all machine learning models. This suggests that SVM and Logistic Regression, while not reaching the extreme highs of the Random Forest Classifier, still demonstrate competitive and consistent performance within the context of the dataset.

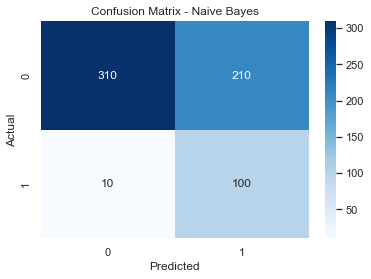
The results from both the table and the graphical representation emphasize the Random Forest Classifier's standout accuracy and highlight the relative performances of other models, such as Naive Bayes, SVM, and Logistic Regression, in comparison to the average accuracy levels observed across the machine learning models.



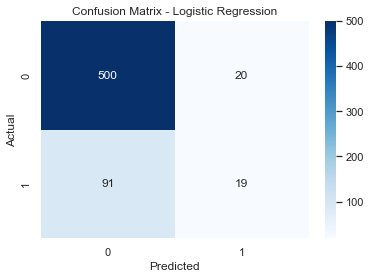
(a)



*(b)*



*(c)*



*(d)*

*Fig 4. Comparison of Confusion Matrix of Machine Learning Models*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ML Models** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| Decision Tree | 90.63% | 0.95 | 0.96 | 0.96 |
| SVM | 84.76% | 0.83 | 1 | 0.9 |
| Naive Bayes | 69.52% | 0.97 | 0.6 | 0.74 |
| Logistic Regression | 83.01% | 0.85 | 0.96 | 0.9 |
| Random Forest | 98.31% | 0.99 | 0.97 | 0.98 |

*Table 2. Result Comparison of Machine Learning Models*

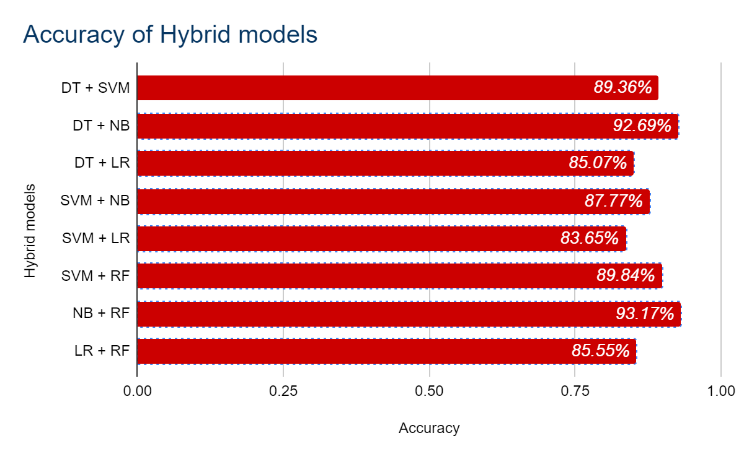
Table 2 summarizes performance metrics derived from confusion matrices in Figure 4. When comparing F1 scores, a balance of Precision and Recall, Random Forest ranks highest with a score of 0.98, closely followed by Decision Tree. SVM and Logistic Regression share the same F1 score, while Naive Bayes trails behind with the lowest F1 score. This highlights Random Forest's superior predictive ability and offers insights into the relative performance of other models based on their F1 scores.

| **DL Models** | **Accuracy** |
| --- | --- |
| SwishRNN | 47.50% |
| SimpleRNN | 52.00% |
| DNN | 55.50% |
| ANN | 93.49% |
| Feedforward Neural Network | 94.54% |
| One-hot Encoding | 67.23% |
| TabNet Accuracy: | 87.94% |
| Ensemble Accuracy: | 88.41% |
| TabNet model with hyperparameter tuning | 91.59% |
| GBTs | 93.02% |
| GRU model | 90.63% |
| 4 layer MLP | 94.41% |

*Table 3. Comparison of Deep Learning Models and their Accuracy*

Based on the data presented in both Table 2 and 3, it is evident that the deep learning models have surpassed the machine learning models in terms of performance. Among the various deep learning models, the Feedforward Neural Network, Four-layer MLP, ANN, and GBT models achieved the highest accuracy scores, reaching 94.54%, 94.41%, 93.49%, and 93.02%, respectively. In contrast, the Simple RNN, Swish RNN, and DNN models exhibited the lowest accuracy rates, with scores of 47.50%, 52.00%, and 55.50%, respectively. This comparison underscores the superior performance of deep learning models, particularly the Feedforward Neural Network and Four-layer MLP, in this context.

## Comparative Analysis of Hybrid Models



*Fig 5. Graphical Representation of Comparison of Hybrid Machine Learning Models and their Accuracy*

So far we have seen how five different machine learning classifiers showed different demeanours. Since our churn dataset is smaller than usual to be used as a train-set system. The idea of a hybrid methodology was proposed to get a more robust prediction. Out of the several combinations of ML models performed, Fig 5 shows eight hybrid models of varying accuracy scores. The hybrid model of Naive Bayes and Random Forest Classifiers produced the highest accuracy of 93.17%. We see that the average accuracy of the hybrid models is greater than that of the ML models. This proves that hybrid models give improved accuracy and an overall better performance.

# CONCLUSION

In the highly competitive telecom sector, predicting customer churn is of utmost importance to retain valuable customers. Effective strategies involve segmenting customers based on their interests and offering tailored, competitive services and offers to each group. Researchers continuously strive to provide data-driven solutions for customer retention, which is more cost-effective than acquiring new customers. In this study, using a dataset from an Iranian telecommunications company, we successfully predicted customer churn and validated our results through standard metrics.

The findings clearly indicate that deep learning models outperformed other techniques. However, the machine learning approach, specifically Random Forest, achieved an exceptional F-measure result of 98%. Notably, the hybrid model combining Random Forest and Naive Bayes achieved the highest accuracy score.

For future research, the study could be extended to include customer profiling methods for retaining customers based on specific attributes. Additionally, conducting time series analysis could help uncover customer churn trends over time, providing valuable insights for predicting future churn patterns.

##### References

[1]Adnan Amin, Awais Adnan, Sajid Anwar, An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes, Applied Soft Computing, Volume 137, 2023, 110103, ISSN 1568-4946, <https://doi.org/10.1016/j.asoc.2023.110103>.

[2]R. Sudharsan & E. N. Ganesh (2023) A Swish RNN-based customer churn prediction for the telecom industry with a novel feature selection strategy, Connection Science, 34:1, 1855-1876, DOI: [10.1080/09540091.2022.2083584](https://doi.org/10.1080/09540091.2022.2083584)

[3]Sagar Maan Shrestha, Aman Shakya, A Customer Churn Prediction Model using XGBoost for the Telecommunication Industry in Nepal, Procedia Computer Science, Volume 215, 2022, Pages 652-661, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2022.12.067>.

[4]Y. K. Saheed and M. A. Hambali, Customer Churn Prediction in Telecom Sector with Machine Learning and Information Gain Filter Feature Selection Algorithms, 2021 International Conference on Data Analytics for Business and Industry (ICDABI), Sakheer, Bahrain, 2021, pp. 208-213, doi: 10.1109/ICDABI53623.2021.9655792.

[5]Hemlata Jain, Ajay Khunteta, Sumit Srivastava, Churn Prediction in Telecommunication using Logistic Regression and Logit Boost, Procedia Computer Science, Volume 167, 2020, Pages 101-112, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2020.03.187>.

[6]Nadia Alboukaey, Ammar Joukhadar, Nada Ghneim, Dynamic behavior-based churn prediction in mobile telecom, Expert Systems with Applications, Volume 162, 2020, 113779, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2020.113779>.

[7] Wassouf, W.N., Alkhatib, R., Salloum, K. *et al.* Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. *J Big Data* 7, 29 (2020). <https://doi.org/10.1186/s40537-020-00290-0>

[8]Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, A Churn

Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector, in IEEE Access, vol. 7, pp. 60134-60149, 2019, doi: 10.1109/ACCESS.2019.2914999.

1. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, Electron spectroscopy studies on magneto-optical media and plastic substrate interface, IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
2. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.