Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 2, pp. 613–630
DOI : 10.47974/JIOS-1599

Telecom customer churn prediction model : Analysis of machine learning techniques for churn prediction and factor identification in telecom sector

Anshul Pareek *
Poonam †
Shaifali Madan Arora §
Department of Electronics and Communication Engineering
Maharaja Surajmal Institute of Technology
New Delhi
India

Nidhi Gupta ‡
Department of Electrical and Electronics Engineering
Maharaja Surajmal Institute of Technology
New Delhi
India

Abstract

This research project leverages exploratory data analysis (EDA) on a telecom company's customer data to predict user churn. Utilizing Python and its libraries, including Pandas, NumPy, Matplotlib, and Scikit-Learn, it identifies key parameters crucial for accurate predictions. The results and predictions are visualized using the Flask framework and Power BI analytics tool. In the highly competitive telecommunications sector, customer churn poses a significant challenge, with an annual churn rate of 15-25%. This study addresses the escalating phenomenon of consumers freely switching between providers, leading to financial losses for businesses. By employing machine learning models, it aims to forecast potential churners, enabling companies to focus targeted retention efforts and mitigate losses. The tool's efficacy lies in its ability to analyze patterns among churned users, offering a solution to a pressing issue in the telecom industry. Beyond its immediate application, this

* *E-mail:* **er.anshulpareek@msit.in (Corresponding Author)**
† *E-mail:* **poonam.dahiya@msit.in**
§ *E-mail:* **shaifali04@msit.in**
‡ *E-mail:* **nidhi@msit.in**

tool can be extended to other industries, providing valuable insights for customer retention strategies.

Subject Classification: *94-XX.*

Keywords: *Churn prediction, Retention, Telecom, CRM, Machine learning, Imbalance data, Upsampling, Customer churn, Random forest model, Decision tree model, EDA.*

I. Introduction

A. *Background*

Customer churn refers to the discontinuation of business between customers or subscribers and a company or service [1]. In the telecom industry, where customers have a range of service providers, there's an annual churn rate of 10-30% due to high competition ([1-3]). While personalized customer retention is challenging for companies with a large customer base, predicting potential churn allows them to focus efforts on "high-risk" clients, aiming to expand coverage and boost loyalty. The telecommunications sector increasingly relies on customer churn prediction through data mining, statistical learning, and effective Customer Relationship Management (CRM) to enhance competitiveness [4]. Retaining existing customers is economically favorable, with studies showing that a 5% reduction in churn can lead to profit increases ranging from 25% to 85% [5]. Recognizing the significance of customer churn is crucial as it directly impacts EBITDA margin and poses a widespread industry challenge ([1,5]). Given the long-term nature of customer relationships in telecommunications, characterized by the use of a telephone code for several years, the emphasis on customer retention becomes paramount [4]. Research indicates that satisfied customers contribute significantly to a company's potential deals, whereas dissatisfied customers can negatively influence the purchasing decisions of a substantial number of individuals. Ignoring existing customers can result in losing up to half within five years ([3,5]).

To address the complex issue of customer churn, researchers globally have proposed various methods and strategies ([8-10]). These include employing feature selection and ensemble learning techniques, boosting models, cluster analysis ([6,8]), decision tree algorithms ([11-13]), and Bayesian network classifiers [14]. Despite these efforts, the scarcity of discussion on the data imbalance problem ([15-17]) in customer churn is evident. Telecom customer churn, being a rare yet immensely valuable event, poses a typical imbalanced dataset classification challenge in data mining [18].

A larger customer base reduces initiation costs and enhances profit potential. Hence, the key to success is minimizing client attrition and implementing a focused retention strategy for optimal outcomes ([19-20]).

B. *Objectives*

The main objective of this research paper is to propose methods of evaluating and predicting the likelihood of a customer churning from a telecom company. The formal objectives are:

- Determining the percentage of churned customers versus those actively using services.
- Examining the data to understand the features influencing customer churn.
- Assessing and comparing the performance scores of different models relevant to this study.
- Identifying the most suitable machine learning model for accurately classifying churn and non-churn customers.
- Predicting the likelihood of churn based on customer details.
- Developing an accessible and user-friendly application for a positive user experience and ease of use.

II. Related Works

Work has been done and telecom companies have collected data about their user base as well as acquisition and consumption patterns. This project utilizes a tangible dataset on Kaggle, which will be used to conduct EDA and gather insights, which will help with developing a model for predictions. The data set includes information about:

- Those who departed in the previous month are categorized under "Churn."
- The services each customer has subscribed to include phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
- Customer account details encompass their tenure, contract, payment method, paperless billing, monthly charges, and total charges.
- Demographic information about customers includes gender, age range, and whether they have partners and dependents.

Earlier work in this field has entailed machine learning [19] and more traditional ways to segment and recognize patterns. For example, prediction methods based on cluster stratified sampling logistic regression ([21-22]), using random forest for factor identification [7], etc.

In the article [23] extreme gradient boosting trees are used to predict profit-driven customer churn. Bayesian method is used to optimize the class weights and hyperparameters. Later empirical analysis is implemented which uses the real data sets received from service providers. This application leads to higher profits as compared to the benchmark model implementation.

In the paper [24] the efficiency of tree-based classifiers is computed with diverse computational properties in CCP. Also, the data issues like CIP (class imbalance problem) on the prediction are computed.

Article [25] figures out the number of leave subscriptions and their respective reason for leaving. The paper aims towards the analysis of various Machine learning algorithms utilized to create churn prediction models. It also provides retention plans by identifying their churn reasons using Random Forest (RF), machine learning techniques such as KNN, and decision tree Classifier.

Paper [6] proposed a customer churn prediction method based on cluster stratified sampling logistic regression model with parameters estimated methods suitable for an imbalanced data set. Using UCI and Orange to experiment on representative public data sets, with ROC curves and AUC value as the evaluation index of experiments, comparing the experimental results shows that the presented method for telecom customer churn prediction has a stable promotion effect.

Research [7], presents a churn prediction model employing classification and clustering techniques to identify and analyze factors contributing to customer churn in the telecom sector. Feature selection involves using information gain and correlation attribute ranking filters. The model initially classifies churn customer data, with the Random Forest (RF) algorithm exhibiting a notable 88.63% accuracy in classification.

III. Methodology

A. *Research Design*

The study adopts a combined approach of experimental and analytical methods, with the primary rationale being the evaluation of changes in the F1 score as a critical performance metric. Assess insights on churn behavior of subscribers; and use the information to strategize new marketing initiatives.

B. *Data Collection*

A publicly available dataset comprising 7000+ records of customers in a Telecom Company serves as the primary data source.

C. *Dataset Preparation*

The dataset required for this project would require to have user data, service data, company's user acquisition data, payment data, economic data as well as churning data. For the same, tangible datasets have been found, for example, the 'WA_Fn-UseC_ dataset', which will allow minimum data preprocessing.

Here, the initial intuition upon observing the dataset is that there are no missing values, as shown in Figure 1. However as observed in Figure 7, the data type for Total Charges is an object, whereas it should be a numeric type. Upon changing into numeric type, and checking for null value, missing values are observed.

D. *Data Cleaning*

It is crucial to the project's accuracy and model's efficacy that the data in the dataset be cleaned appropriately, and be consistent and uniform, void of any null objects that may disrupt the model. It should also be free of unnecessary information to ensure minimum data preprocessing. This step also divides the customers according to their tenures, with a tenure of <12 months being assigned in groups 1-12, a tenure of 1-2 years as 13-24, and so on. The value counts for each such group is obtained as shown in Table 1:

Table 1

Tenure Groups Value Counts

Tenure Group	Count
1-12	2175
13-24	1024
25-36	832
37-48	762
49-60	832
61-72	1407

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

Figure 1

Dataset Info used for preparing

It is also made sure to drop unwanted columns such as customerID and tenure (now replaced with tenure groups).

Furthermore, the Target Variable 'Churn' is converted into a binary numeric variable, i.e., Yes = 1 and No = 0.

The next step is to convert the categorical variables found in the dataset into dummy variables. In this, variables (such as Partner) with categorical values (Yes and No) are split into multiple binary numeric variables (Partner_Yes and Partner_No).

E. *EDA*

Once a clean dataset is obtained, data analysis begins and valuable insights are sought from it. For the same, python and its libraries are used, to query the dataset, draw informative conclusions, and identify crucial parameters for successful prediction. Data like, gender distribution of churning users, payment methods, contract distribution, services, security, and much more are looked at. Tools like Power BI will come in extremely handy during this step.

F. Basic Model Building

After having derived insights from EDA, the model building took place. The process started with a basic Decision Tree Classifier and Compared its results. Since it was observed in EDA that the Dataset is imbalanced, it was known that the accuracy of such testing would be "cursed", so to follow up, upsampling with the help of SMOTEENN was performed. Having Performed Model Building with the Decision Tree Classifier, to compare this model, the next step was to build a comparison with the Random Forest ([3,14]) Classifier.

G. Further Models Built

It is known that different classifier models work on different principles and have a performance variable. Thus, to make sure the optimum model for prediction is obtained, it becomes crucial to test multiple models. This is what has been focused on in this section of this model deciding step.

The models that have been considered for this study's use case after thorough research are:

- Linear Regression [6]
- Lasso [19]
- Ridge [19]

- K-Neighbours Classifier [14]
- Decision Tree [26]
- Random Forest Classifier [7]
- AdaBoost Regressor [26]

For this approach's purpose, a function was defined, which iterates through each model one by one and calculates its performance parameters based on the provided set, i.e., the Precision, Recall, and F1 Score.

H. *Performing Hyperparameter Tuning*

Now that the process has concluded the model, the next step is to further optimize this model by performing hyperparameter tuning as shown in Figure 2.

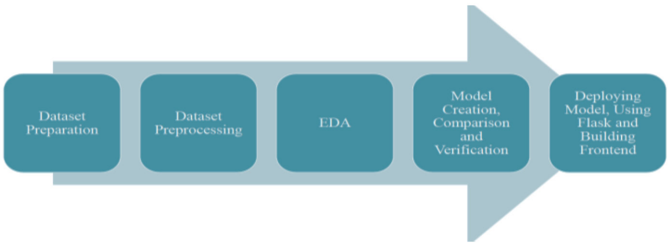

Figure 2

Research Process

III. Results and Discussion

The results of EDA, Model Building and Comparison, and Hyperparameter Tuning are obtained and analyzed in this section.

As seen in Figures 3, and 4, it is observed that a higher churn rate is obtained where monthly charges are very high. A surprising insight is drawn: higher churn at lower Total Charges. This can be attributed to the logic that if a customer is paying a large sum, he or she is more likely to withdraw from the service provider sooner, rather than later.

A. Yet, when considering the interplay of three factors—Tenure, Monthly Charges, and Total Charges—a clearer perspective emerges. Specifically, a higher Monthly Charge coupled with lower tenure leads to a decrease in Total Charges. Consequently, the correlation between Higher Monthly Charge, Lower Tenure, and Lower Total Charge becomes evident, signifying a connection to High Churn.

A. Results of EDA-Univariate Analysis

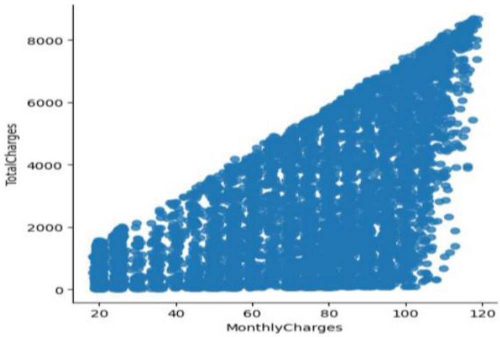

Figure 3

Relation between Total Charges and Monthly Charges

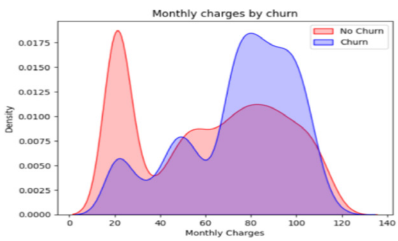

Figure 4(a)

Monthly Charges vs Churn and Total Charges vs Churn

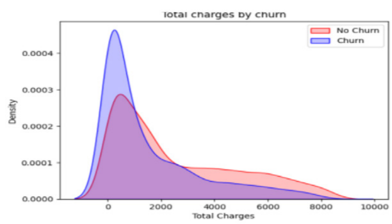

Figure 4(b)

Monthly Charges vs Churn and Total Charges vs Churn

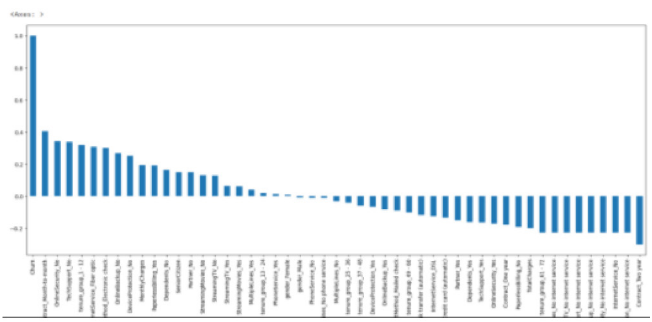

Figure 5

Correlation w.r.t. Churn

As seen in Figure 5, Elevated churn rates are observed with month-to-month contracts, absence of online security and tech support, first-year subscriptions, and Fiber Optics Internet. Conversely, low churn rates are noted in long-term contracts, subscriptions without internet service, and among customers engaged for 5+ years. Factors such as gender, availability of phone service, and the number of multiple lines have minimal to no impact on churn.

Figure 6 depicts the correlation matrix.

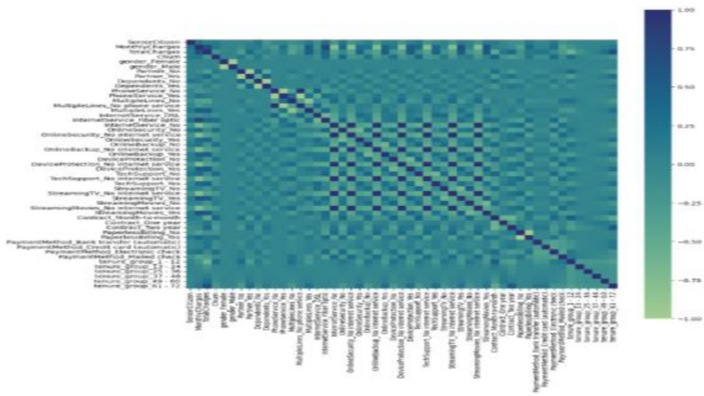

Figure 6

Heat Map / Correlation Matrix

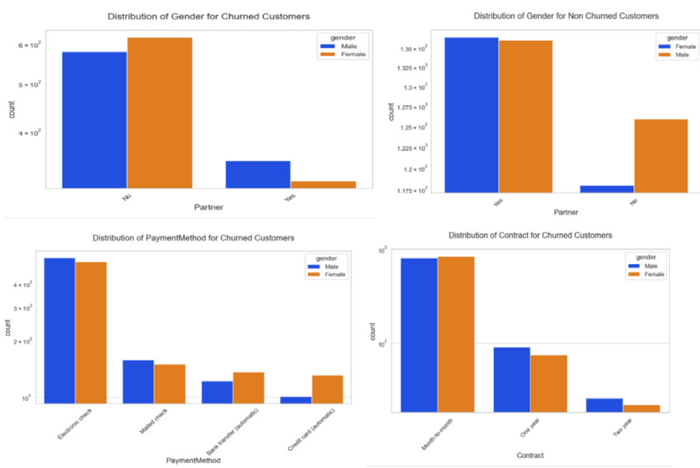

Figure 7

Various Bivariate EDA results

B. *Results of EDA-Bivariate Analysis*

After gathering insights from Univariate EDA, Bivariate analysis is performed as well to gather further insights and knowledge as to the connections between two variables taken together. Numerous insights were drawn from these charts. This study's main criteria of evaluation were based on different classes along with gender distribution. Some insights drawn are depicted in Figure 7.

C. *Conclusion of EDA*

- Electronic check transactions exhibit the highest churn rates.
- Monthly customers, lacking contract terms, are more prone to churn as they have the freedom to switch.
- Categories without online security and tech support experience higher churn rates.
- Non-senior citizens are more likely to churn.
- Males with partners are much more likely to churn than females with partners, whereas males without partners are even highly likely to not churn in comparison to females without partners.
- Churners that use credit cards as payment methods are much more likely to be females than males.

D. *Results of Basic Model Building*

As seen in Table 2, the accuracy is quite low, and as it's an imbalanced dataset, it is unwise to consider accuracy as the metric to measure the model, as accuracy is cursed in imbalanced datasets. Hence, there is a need to check recall, precision, and F1 scores for the minority class, and it's quite evident that the precision, recall, and F1 scores are too low for Class 1, i.e., churned customers. Hence, moving ahead to call SMOTEENN (UpSampling + ENN). The results are seen in Tables 3 and 4.

Now quite better results can be seen, i.e., Accuracy: 93%, and a very good recall, precision, and F1 score for the minority class.

As seen in Table 5, the accuracy for a simple Random Forest Algorithm is also poor, similar to a Decision Tree because of being fitted over an imbalanced dataset. This is overcome similarly with SMOTEENN. The observations are depicted in Tables 6, 7.

- Results for Basic Decision Tree Classifier Model

Table 2

Classification report of simple Decision Tree

	Precision	Recall	F1 Score	Support
0	0.82	0.92	0.87	1044
1	0.65	0.44	0.52	363
Accuracy			0.80	1407
Macro Avg	0.74	0.68	0.70	1407
Weighted Avg	0.78	0.80	0.78	1407

Table 3

Classification report of simple Decision Tree after upsampling

	Precision	Recall	F1 Score	Support
0	0.82	0.92	0.87	1044
1	0.65	0.44	0.52	363
Accuracy			0.80	1407
Macro Avg	0.74	0.68	0.70	1407
Weighted Avg	0.78	0.80	0.78	1407

Table 4

Confusion Matrix of simple Decision Tree after upsampling

495	41
35	595

- Results for Random Forest Classifier Model

Table 5

Classification report of imbalanced Random Forest

	Precision	Recall	F1 Score	Support
0	0.83	0.94	0.88	1044
1	0.72	0.46	0.56	363
Accuracy			0.82	1407
Macro Avg	0.78	0.70	0.72	1407
Weighted Avg	0.80	0.82	0.80	1407

Table 6

Classification report of simple Decision Tree after upsampling

	Precision	Recall	F1 Score	Support
0	0.96	0.91	0.94	542
1	0.93	0.97	0.95	628
Accuracy			0.94	1170
Macro Avg	0.95	0.94	0.94	1170
Weighted Avg	0.95	0.94	0.94	1170

Table 7

Confusion Matrix of simple Decision Tree after upsampling

495	47
18	610

With the Random Forest Classifier also, quite good results are obtained. The results observed are better than those of the Decision Tree. The better Accuracy is obtained at 95%.

G. *Results of Other Models Building and Their Comparison*

As seen in Table 8, the accuracy is quite low, and as it's an imbalanced dataset, it is improper to consider Accuracy as the metric to measure the model, as Accuracy is cursed in imbalanced datasets. Hence, the need to check the recall, precision & f1 score for the minority class arises, and it's quite evident that the precision, recall & f1 score is too low for Class 1, i.e. churned customers. Hence, moving ahead to call SMOTEENN (UpSampling + ENN). Results are stored in Table 9.

Here the observations can see the improvement in every aspect, after performing an upsampling [27] of the dataset with ENN with the help of SMOTEENN. The study can comment upon the precision, callback, as well as F1 scores of different models and make the following deductions firstly, the Lasso Model has been the weakest fit for the dataset in both cases. Secondly, both Random Forest Classifiers, as well as K-Neighbours Classifiers edge out the other models after upsampling, observing a significant gain in performance. Decision Tree loses its edge over Adaboost Regressor after upsampling. It is evident from Table 10 that these readings are cursed and don't offer any clarity and indication as to which model works best, simply because of an unbalanced dataset. Thus it is after comparing the results of upsampling + ENN, that the proposed approach begins to see the true performance of the models. Now from the observed results and Table 11 showing the performance rankings, the best model after training is found to be the Random Forest Classifier Model. As a further part of the process to finalize this model for the application, this model is tested using the test partition: xr_test and yr_test. The model score comes out to be a strong **0.9596219931271478.3**

Table 8

Tabulating the performance scores of various models when fitted to the dataset

Selected Model	Precision	Recall	F1 Score
Ridge	0.7154	0.4872	0.5797
Linear Regression	0.7154	0.4872	0.5797
Random Forest Classifier	0.6290	0.4541	0.5274
Decision Tree	0.5128	0.5102	0.5115

Contd...

K-Neighbours Classifier	0.5890	0.4388	0.5029
Adaboost Regressor	0.7234	0.3469	0.4690
Lasso	0.7095	0.3240	0.4448

Table 9

Tabulating the performance scores of various models when fitted to the dataset after SMOTEENN

Selected Model	Precision	Recall	F1 Score
Random Forest Classifier	0.9564	0.9680	0.9622
K-Neighbour Classifier	0.9604	0.9589	0.9596
Ridge	0.9525	0.9452	0.9488
Linear Regression	0.9525	0.9452	0.9488
Adaboost Regressor	0.9198	0.9604	0.9397
Decision Tree	0.9334	0.9391	0.9363
Lasso	0.9069	0.8600	0.8828

Table 10

Performance Ranking of various models trained on an unbalanced dataset

Rank	Model Name	F1 Score
1	**Ridge**	0.57966
2	**Linear Regression**	0.57966
3	**Random Forest Classifier**	0.527407
4	**Decision Tree**	0.511509
5	**K-Neighbours Classifier**	0.502924
6	**Adaboost Regressor**	0.468966
7	**Lasso**	0.444834

Table 11

Performance Ranking of various models trained after SMOTEENN

Rank	Model Name	F1 Score
1	**Random Forest Classifier**	0.962179
2	**K-Neighbours Classifier**	0.959634
3	**Ridge**	0.948816

Contd...

4	**Linear Regression**	0.948816
5	**Adaboost Regressor**	0.939687
6	**Decision Tree**	0.936267
7	**Lasso**	0.882812

F. *Results of Hyperparameter Tuning*

From the results of performance scores of various models, before and after upsampling, the study concludes that the optimum model for this study's use case is the Random Forest Classifier Model. The next step now is to perform hyperparameter tuning, to further improve its score.

The hyperparameters which are considered and tried to optimize are:

- n_estimators = [50, 100, 200]
- max_depth = [None, 10, 20]
- min_samples_leaf = [1, 2, 4]
- min_samples_split = [2, 5, 10]

After using these values for the process and 100 iterations later, the process arrives at optimized values for each of these parameters. These are:

- n_estimators = 200
- max_depth = None
- min_samples_leaf = 1
- min_samples_split = 5

Now that the optimum hyperparameters have been obtained, these must be used with the model. This will result in an improved score and increase the probability of making accurate predictions. The new score as per the Random Forest Classifier Model after Hyperparameter Tuning is:

Table 12

Performance Score of the Random Forest Classifier Model after performing Hyperparameter Tuning

Precision	Recall	F1 Score
0.964939	0.963470	0.964204

Here, from Table 12, it is seen that the overall performance of the tuned model is better in terms of Precision and F1 Score but dips very slightly in Recall. However, since the overall F1 Score and Precision are better, this study will pursue this model and save it for use in the application.

IV. Implications For The Telecom Industry/ Conclusion

The following will be the implications for the telecom industry owing to the results of this study's research and implementation:

- Ability to effectively manage customer churn through preventive management using leading and lagging indicators of churn.
- Ability to identify and save customers who are about to churn.
- Assess insights on churn behavior of subscribers; and use the information to strategize new marketing initiatives.
- Identify patterns in customer behavior of potential churners and initiate proactive measures to reduce churn.
- Capabilities in driving analytics-led campaign/marketing initiatives from Predictive Model experience.
- Reduction in the campaign spend by targeting fewer subscribers, exhibiting churn behavior, rather than targeting larger ones based on gut feel.
- Report revenue loss affected due to churn.
- Classification of subscribers as voluntary/involuntary churners.
- Cost savings from retention as opposed to re-acquisition.

V. Future Work

In future inquiries, this research endeavors to further explore both eager learning and lazy learning approaches to improve the precision of churn prediction. Additionally, there is the prospect of expanding the scope of this study to scrutinize the evolving behavioral patterns of churn customers, utilizing Artificial Intelligence techniques for predictive analysis and trend identification. Note that this study was greatly limited due to the limitations of the training dataset, with approximately 7000 records. A larger dataset with higher computing power can explore patterns in much more depth and derive better and more consistent insights.

References

[1] Bhale, U. A., & Bedi, H. S. Customer Churn Construct: Literature Review and Bibliometric Study. *Management Dynamics*, 24(1), 1.

[2] L.Zhou, "Performance of corporate bankruptcy prediction models on the imbalanced dataset: The effect of sampling methods," Knowledge-Based Systems, vol. 41, pp.16-25 (2013).

[3] Y. Bharambe, P. Deshmukh, P. Karanjawane, D. Chaudhari and N. M. Ranjan, "Churn Prediction in Telecommunication Industry," 2023 International Conference for Advancement in Technology (ICONAT), Goa India, pp. 1-5 (2022), doi: 10.1109/ICONAT57137.2023.10080425.

[4] Ali Yari, M., Modiri, M., Khalili Damghani, K., & Fathi Hafshjani, K. Measuring Customer Satisfaction Using Multi-Criteria Analysis Model of Customer Satisfaction to Evaluate Product Lines (Case Study: Kaveh Glass Industrial Group). *Journal of Information and Organizational Sciences*, *47*(2), 283-304 (2023).

[5] M. Hassouna, A. Tarhini, T. Elyas, and M. S. AbouTrab, "Customer churn in mobile markets a comparison of techniques," arXiv preprint arXiv:1607.07792, 2016.

[6] Venkatesh, S., & Jeyakarthic, M. Artificial fish swarm algorithm-based multilayer perceptron model for customer churn prediction in IoT with cloud environment. *International Journal of Business Information Systems*, 44(3), 442-465 (2023).

[7] N.Lu, H.Lin, J.Lu, G.Q.Zhang, "A customer churn prediction model in telecom industry using boosting," *IEEE Transactions on Industrial Informatics*, vol. 10, pp.1659-1665 (2014).

[8] H.F.Qin, "The Application of Data Mining in Telecommunication Churn Customer," *Research Journal of Applied Sciences*, vol. 38, pp.1054-1057 (2012).

[9] A.Idris, A.Khan, Y.S.Lee ,"Genetic Programming and Adaboosting based churn prediction for Telecom," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 20, pp. 1328-1332 (2012).

[10] G.Q.Li,X.Q.Deng, "Customer Churn Prediction of China Telecom Based on Cluster Analysis and Decision Tree Algorithm," *Communications in Computer and Information Science*, vol. 315, pp. 319-327 (2012).

[11] G.L.Nie, R.Wei, L.L.Zhang, et al, "Credit card churn forecasting by logistic regression and decision tree," *Expert Systems with Applications,* vol. 38, pp. 15273-1528 (2011).

[12] W.Wei,J.Li , L.Cao, et al. "Effective detection of sophisticated online banking fraud on extremely imbalanced data," *World Wide Web*, vol.18, pp. 1-27 (2012).

[13] J.Burez,D.Van den Poel, "Handling class imbalance in customer churn prediction ," Expert Systems.

[14] M. Kaur, K. Singh, and N. Sharma, "Data Mining as a tool to Predict the Churn Behaviour among Indian bank customers," *International Journal on Recent and Innovation Trends in Computing and Communication*, 1.9, pp.720-725 (2013).

[15] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman, "Telecommunication subscribers' churn prediction model using machine learning," In Eighth International Conference on Digital Information Management (ICDIM), IEEE, pp. 131-136 (2013).

[16] W. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building comprehensible customer churn prediction models with advanced rule induction techniques," *Expert Systems with Applications*, 38(3), pp. 2354-2364 (2011).

[17] G.King,M.Tomz,L.C.Zeng, "Relogit:Rare Events Logistic Regression," *Journal of Statistical Software,* vol. 8, pp. 84-113 (2003).

[18] Vairavan. S, Eshelman. L, Haider. S, Flower. A, Seiver. A, "Prediction of mortality in an intensive care unit using logistic regression and a hidden Markov model," *Computing in Cardiology*, vol. 39, pp-393-396 (2012).

[19] Liu, Z., Jiang, P., De Bock, K. W., Wang, J., Zhang, L., & Niu, X. Extreme gradient boosting trees with efficient Bayesian optimization for profit-driven customer churn prediction. *Technological Forecasting and Social Change*, 198, 122945 (2024).

[20] Usman-Hamza, F. E., Balogun, A. O., Nasiru, S. K., Capretz, L. F., Mojeed, H. A., Salihu, S. A., ... & Awotunde, J. B. Empirical analysis of tree-based classification models for customer churn prediction. *Scientific African*, 23, e02054 (2024).

[21] Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. Customer churn prediction in telecom sector

using machine learning techniques. *Results in Control and Optimization*, 14, 100342 (2024).

[22] Vairavan. S, Eshelman .L, Haider .S, Flower .A, Seiver .A, "Prediction of mortality in an intensive care unit using logistic regression and a hidden Markov model," *Computing in Cardiology*, vol. 39, pp-393-396 (2012).

[23] N. Lu, H. Lin, J. Lu, G. Q. Zhang, "A customer churn prediction model in telecom industry using boosting," *IEEE Transactions on Industrial Informatics*, vol. 10, pp. 1659-1665 (2014).

[24] Usman-Hamza, F. E., Balogun, A. O., Nasiru, S. K., Capretz, L. F., Mojeed, H. A., Salihu, S. A., ... & Awotunde, J. B. Empirical analysis of tree-based classification models for customer churn prediction. *Scientific African*, 23, e02054 (2024).

[25] Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization*, 14, 100342 (2024).

[26] A. D. Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," *European Journal of Operational Research*, 269.2, pp.760-772 (2018).

[27] Mohan, M., & Jadhav, A. Predicting Customer Churn on OTT Platforms: Customers with Subscription of Multiple Service Providers. *Journal of Information and Organizational Sciences*, 46(2), 433-451 (2022).