IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, April 2026
Impact Factor: 8.2

Customer Churn Prediction for Telecommunication

Akshaya Devi P[1], Jayashree J[2], Parameshwari T[3] ,Sathiya S[4]
Students, Department of Computer Science and Engineering[123]
Assistant Professor, Department of Computer Science and Engineering[4]
Annamalai University, Annamalai Nagar, Tamil Nadu, India

Abstract: *Customer churn is a persistent challenge in the telecommunications industry, where the cost of customer acquisition significantly outweighs retention expenses. This research presents a robust hybrid predictive framework using the IBM Telco Customer Churn dataset to identify at-risk users and estimate their remaining tenure. To handle the inherent class imbalance and noise in telecom data, we integrated the SMOTE-ENN (Synthetic Minority Over-sampling Technique - Edited Nearest Neighbours) algorithm, ensuring a high-quality, balanced dataset for model training. The methodology compares a Decision Tree Classifier with an ensemble Random Forest approach. Experimental results show that while the Decision Tree achieved a strong accuracy of 93%, the Random Forest model outperformed it with a superior classification accuracy of 95%. Moving beyond binary "Yes/No" predictions, the framework incorporates the Cox Proportional Hazards (CPH) Model for survival analysis. By calculating Hazard Ratios and survival probabilities, the system provides a Mapped Retention Timeline (e.g., 1-2 months, 3-5 months, or 6+ months), offering a temporal window for intervention. This dual-layered approach-combining high accuracy classification with "Time-to-Event" insights-empowers telecom providers to deploy targeted, data-driven loyalty programs, effectively reducing revenue loss and enhancing long-term customer relationships.*

Keywords: SMOTE-ENN, Random Forest, Cox Proportional Hazards Model, Retention Timeline Mapping

I. INTRODUCTION

The rapid evolution of the telecommunications industry has shifted the business focus from aggressive customer acquisition to strategic customer retention, as the cost of securing new subscribers remains significantly higher than maintaining existing ones. Customer Churn, the process where users terminate their services, poses a major financial threat to providers worldwide. While traditional machine learning approaches offer binary predictions to identify potential churners, they often struggle with highly imbalanced datasets and fail to provide a temporal dimension—answering when a customer is likely to leave. This research addresses these gaps by proposing a robust hybrid framework using the IBM Telco Customer Churn dataset. By implementing the SMOTE-ENN technique, the system effectively balances the data and removes noise, ensuring high-precision mode modelling. A comparative analysis between a Decision Tree and a Random Forest Classifier was conducted, with the latter achieving a superior accuracy of 95%. Beyond simple classification, this study integrates the Cox Proportional Hazards (CPH) Model to perform survival analysis, enabling the estimation of a Mapped Retention Timeline. This comprehensive approach not only identifies at-risk individuals but also provides a critical window for intervention, allowing telecom companies to transition from reactive measures to proactive, data-driven loyalty management.

II. LITERATURE SURVEY

In[1], **Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024)** Proposed a machine learning–based churn prediction system for the telecom sector, demonstrating that ensemble

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, April 2026 Impact Factor: 8.2

models, particularly Random Forest combined with data balancing and survival analysis, achieve high accuracy and support effective customer retention strategies. In[2], **vWagh, S. K., et al. (2024)** Proposed a churn prediction system using Decision Tree, Random Forest, and survival analysis. The Random Forest model achieved over 99% accuracy, and survival analysis helped estimate churn timing for proactive retention strategies. In[3], **Srinivas, S., & Reddy, L. S. (2023)** Proposed a deep learning–based churn prediction model using neural networks, achieving better performance for large and complex telecom datasets. In[4], **Amin, A., et al. (2023)** Proposed an adaptive learning-based churn prediction approach using evolutionary computation and Naïve Bayes, achieving improved accuracy and adaptability. In[5], **Lalwani, P., Mishra, M. K., & Chadha, J. S. (2022)** Conducted a comparative study of multiple machine learning models including Random Forest, XGBoost, AdaBoost, and CatBoost. Ensemble models outperformed traditional classifiers in churn prediction. In[6], **Berrevoets, J., & Verbeke, W. (2022)** Emphasized profit-driven evaluation of churn models and showed that predictive performance alone is insufficient without business impact analysis. In[7], **Amin, A., et al. (2022)** Applied data transformation techniques for cross company churn prediction, addressing data scarcity problems in telecom industries. In[8], **Nadeem, A. N., Umar, S., & Shahzad, M. S. (2021)** Presented a review of data mining techniques for churn prediction, highlighting the importance of ensemble and hybrid models in telecom applications. In[9], **Devriendt, F., Berrevoets, J., & Verbeke, W. (2021)** Introduced uplift modeling instead of traditional churn prediction to maximize profit. Their approach focused on identifying persuadable customers rather than only churn-prone users. In[10], **Zhao, M., et al. (2021)** Developed a churn prediction model considering customer value using logistic regression. The study emphasized prioritizing high-value customers in retention campaigns.

III. METHODOLOGY

The proposed methodology predicts customer churn in the telecom sector by combining machine learning classification with temporal survival analysis. Historical customer data from the IBM Watson Telco Churn dataset is collected, including demographics, service usage, billing records, and contract details. Preprocessing addresses missing values in the **Total Charges** column, removes irrelevant features like **CustomerID,** and converts categorical variables into numerical form using Label Encoding. To tackle class imbalance, the SMOTE-ENN hybrid approach is applied: SMOTE generates synthetic samples for the minority class, while ENN removes noisy and overlapping points. This balanced dataset is then used to train Decision Tree and Random Forest models, which capture complex patterns to distinguish churners from non-churners. Model performance is evaluated using Accuracy, Precision, Recall, and F1-score. Beyond binary classification, Survival Analysis with the Cox Proportional Hazard Model estimates when churn is likely to occur by analysing hazard ratios and survival probabilities. Integrating outputs from both models provides each customer's churn status along with an estimated retention timeline (e.g., 1–2 months, 3–5 months). This dual-layered approach enables management to design time-sensitive retention strategies and personalized offers, ultimately reducing churn and enhancing long-term profitability.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, April 2026
Impact Factor: 8.2

PROPOSED SYSTEM ARCHITECTURE

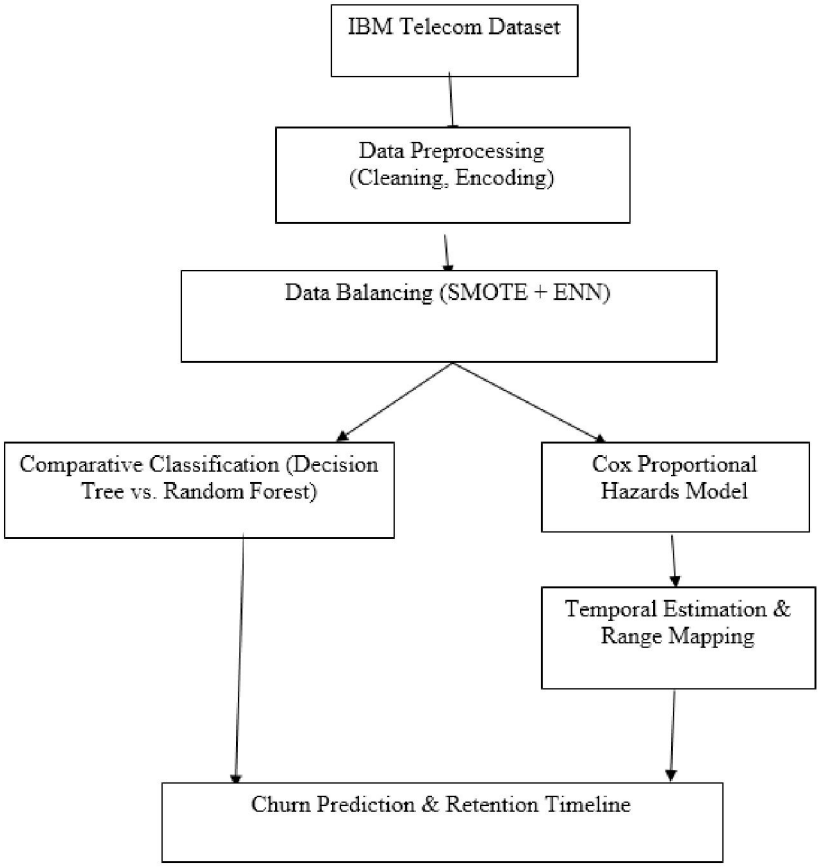

Figure 1. Block Diagram

MODEL DESCRIPTION:

1. Data Preprocessing

At first the raw dataset is refined by handling missing values in the TotalCharges column and performing feature selection to remove noncontributing columns like CustomerID. Categorical features are converted into numerical formats using Label Encoding to make the data compatible with machine learning algorithms.

2. Data Balancing (SMOTE + ENN)

To address the class imbalance where non-churners significantly out number churners, we apply a hybrid technique of SMOTE and ENN. SMOTE generates synthetic data for the minority class (churners), while ENN removes noise and overlapping data points to sharpen the class boundaries..we show in Figure 2.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, April 2026
Impact Factor: 8.2

```
Before Balancing:
Churn
0    5174
1    1869
Name: count, dtype: int64
After Balancing:
Churn
1    3248
0    2660
Name: count, dtype: int64
```

Figure 2.Data Balancing

3. Machine Learning Models (Decision Tree & Random Forest)

Figure 3, shows the comparison of the models .This stage involves training classification models to predict the likelihood ofchurn. We utilize Decision Tree for transparent rule-based logic and RandomForest, an ensemble method, to achieve high prediction accuracy by aggregating
multiple decision paths.

4. Survival Analysis(Cox Proportional Hazards Model)

Beyond binary classification, this stage implements the Cox Proportional Hazards model to analyze the 'time-to-event' or duration before a customer leaves. Itidentifies how specific factors like contract type and monthly charges influence the expected remaining tenure of a customer. we show in Figure 4.

5. Temporal Estimation & Range Mapping

In this stage, the survival model's predictions are translated into actionable timewindows. Each 'at-risk' customer is mapped into specific retention categories,such as 1-2 months or 6+ months, based on their predicted remaining lifetime. we show in Figure 5.

6. Final Output Generation

The results from the classification models and survival analysis are merged intoa single comprehensive report. This final document provides a holistic view of each customer, combining their churn status with a precise timeline for business intervention. we show in Figure 6.

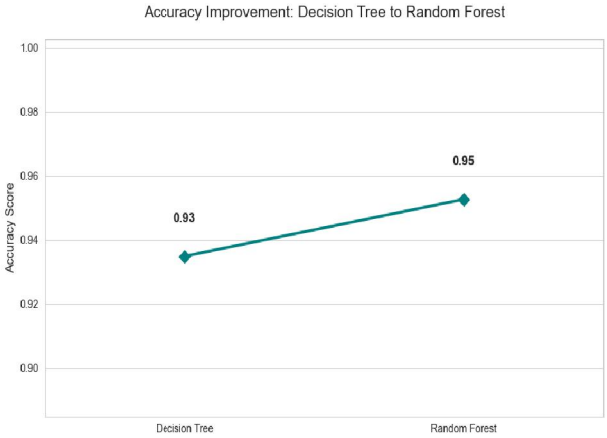

Figure 3. Comparative Performance Report

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, April 2026
Impact Factor: 8.2

```
--- 5. SURVIVAL ANALYSIS OUTPUT ---
                   coef  exp(coef)             p
covariate
gender        -0.046569   0.954499  2.261284e-01
SeniorCitizen  0.012692   1.012773  8.128793e-01
Partner       -0.283063   0.753472  5.894308e-08
Dependents    -0.365988   0.693511  2.162582e-06
PhoneService  -0.159128   0.852887  8.655382e-02
```

Figure 4. Survival Analysis

```
--- 6. TEMPORAL ESTIMATION OUTPUT ---
      Raw_Expected_Tenure Mapped_Retention_Timeline
1539             28.228777                6+ Months
2607             70.371377                6+ Months
3685              5.571653                6+ Months
4308              1.139358               1-2 Months
696              71.976215                6+ Months
889              21.543846                6+ Months
4443             10.537752                6+ Months
5599              2.052420               3-5 Months
683              58.530268                6+ Months
296              71.996729                6+ Months
```

Figure 5. Temporal Estimation

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, April 2026 Impact Factor: 8.2

	CustomerID	ChurnPrediction	ExpectedChurnTime
0	7590-VHVEG	YES	6+ Months
1	5575-GNVDE	NO	6+ Months
2	3668-QPYBK	YES	6+ Months
3	7795-CFOCW	NO	6+ Months
4	9237-HQITU	YES	6+ Months
5	9305-CDSKC	YES	3-5 Months
6	1452-KIOVK	YES	6+ Months
7	6713-OKOMC	YES	6+ Months
8	7892-POOKP	YES	6+ Months
9	6388-TABGU	NO	6+ Months
10	9763-GRSKD	NO	6+ Months
11	7469-LKBCI	NO	6+ Months
12	8091-TTVAX	NO	6+ Months
13	0280-XJGEX	YES	6+ Months
14	5129-JLPIS	NO	6+ Months
15	3655-SNQYZ	NO	6+ Months
16	8191-XWSZG	NO	6+ Months
17	9959-WOFKT	NO	6+ Months
18	4190-MFLUW	NO	6+ Months
19	4183-MYFRB	YES	6+ Months
20	8779-QRDMV	YES	6+ Months
21	1680-VDCWW	NO	6+ Months
22	1066-JKSGK	YES	6+ Months
23	3638-WEABW	NO	6+ Months
24	6322-HRPFA	NO	6+ Months
25	6865-JZNKO	NO	6+ Months
26	6467-CHFZW	YES	6+ Months
27	8665-UTDHZ	YES	6+ Months
28	5248-YGIJN	NO	6+ Months
29	8773-HHUOZ	YES	6+ Months
30	3841-NFECX	NO	6+ Months

Figure 6. Final Output Generation

IV. EXPERIMENTAL RESULT AND ANALYSIS

The machine learning models are trained and evaluated using the processed dataset.The performance of the models is analyzed using different evaluation metrics such as Accuracy, Precision, Recall, and F1Score.Two classification models are compared: Decision Tree and Random Forest. The dataset is first balanced using SMOTE-ENN and then models are trained on the balanced dataset.The results show that Random Forest performs better than Decision Tree.because,it combines multiple decision trees and reduces overfitting.Additionally, Survival Analysis provides insights into when a customer is likely to churn, helping businesses plan retention strategies effectively.

DECISION TREE - Accuracy: 0.93, and
RANDOM FOREST - Accuracy: 0.95

From the results we observe that Random Forest provides better accuracy andoverall performance compared to the Decision Tree model.

V. PERFORMANCE MEASURES

Accuracy: Measures the proportion of correct predictions made by the model.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, April 2026 Impact Factor: 8.2

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision: Measures how many predicted churn customers actually churned.

Precision = TP / (TP + FP)

Recall: Measures how many actual churn customers were correctly identified.

Recall = TP / (TP + FN)

F1 Score: Harmonic mean of precision and recall.

F1 Score = 2 × (Precision × Recall) / (Precision + Recall)

SMOTE Formula:

$$x_{new} = x_i + lambda * (x_{zi} - x_i)$$

Cox Proportional Hazards Model:To predict **when** a customer will churn, we used the Cox Model.

The Hazard Function: Predicts the risk of churn at time t based on customer features (X).

$$h(t, X) = h_0(t) * exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)$$

P-Value Calculation: The P-value is then derived from the z-score using the Standard Normal Distribution formula:

$$P - \boldsymbol{value} = 2 \times [1 - \Phi(|z|)]$$

Expected Tenure:

$$E[X] = \int_0^\infty S(X)dt$$

VI. CONCLUSION

This research successfully developed a hybrid framework for customer churn prediction in the telecommunications sector. By utilizing the IBM Telco dataset and addressing the class imbalance issue through SMOTE-ENN, the model achieved a robust foundation for training. The comparative analysis proved that the Random Forest Classifier outperformed the Decision Tree with a superior accuracy of 95%. Furthermore, the integration of the Cox Proportional Hazards Model added a vital temporal dimension, allowing the system to estimate the "Time-to-Churn." This comprehensive approach enables telecom providers to identify high-risk customers and implement proactive retention strategies, effectively reducing revenue loss and improving customer loyalty.

REFERENCES

[1]. BM Dataset: IBM Business Analytics, "Telco Customer Churn Dataset," Kaggle Repository, [Online].
[2]. Available: https://www.kaggle.com/datasets/blastchar/telco-customer-churn
[3]. .SMOTE-ENN: G. Batista, R. Prati, and M. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20–29, 2004.
[4]. Random Forest: L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[5]. Survival Analysis: D. R. Cox, "Regression Models and Life-Tables," Journal of the Royal Statistical Society: Series B (Methodological), vol. 34, no. 2, pp. 187–202, 1972.
[6]. Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. Results in Control and Optimization, 14, 100342.
[7]. Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest. IEEE Access, 7, 60134–60149.
[8]. Ahmed, A. A. Q., & Maheswari, D. (2017). Churn prediction on huge telecom data using hybrid firefly based classification. Egyptian Informatics Journal, 18(3), 215–220.