International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Customer Churn Analysis in Telecom Industry using Kaplan–Meier and Cox Proportional Hazards Model

Ravi Sharma[1], Shailendra Singh Senior Member IEEE[2]

[1] M.Tech Student, Department of Computer Science and Engineering, Oriental University, Indore, MP India
[2] Professor, Department of Computer Science and Engineering, Oriental University, Indore, MP India

---

Abstract: Customer churn poses a significant challenge for companies in the telecom industry, impacting revenue and profitability. In this paper, we present a comprehensive survey of customer churn analysis methodologies, focusing on the application of logistic regression, Kaplan-Meier survival analysis, and the Cox Proportional Hazards Model. Logistic regression serves as a fundamental technique for binary classification, enabling the prediction of churn based on customer attributes and behaviors. Kaplan-Meier survival analysis offers insights into customer survival probabilities over time, facilitating the identification of high-risk cohorts. Additionally, the Cox Proportional Hazards Model provides a flexible framework for analyzing the impact of predictor variables on churn timing, accounting for censoring and time-varying covariates. Through a review of existing literature, case studies, and practical applications, we explore the strengths and limitations of these methodologies in customer churn prediction. We also discuss challenges in data collection, feature engineering, and model evaluation, along with future research directions to enhance churn analysis effectiveness. By leveraging logistic regression with Kaplan-Meier and the Cox model, telecom companies can develop robust churn prediction models to improve customer retention strategies and mitigate revenue loss.

Keywords: Customer churn analysis, Telecom industry, Logistic regression, Kaplan-Meier survival analysis, Cox Proportional Hazards Model, Retention strategies.

I. INTRODUCTION

Customer churn analysis is a critical component of customer relationship management in the telecom industry. Churn, also known as customer attrition, refers to the phenomenon where customers switch their telecom service providers or terminate their subscription. It is a significant concern for telecom companies, as acquiring new customers is often more expensive than retaining existing ones. Therefore, understanding and managing customer churn is essential for the long-term success and profitability of telecom companies. The telecommunications industry is highly competitive, with various service providers offering similar services, such as mobile, internet, and television services. In the highly competitive landscape of the telecommunications industry, customer churn, or the phenomenon of subscribers discontinuing their services with a provider, is a pressing concern. Churn not only directly impacts revenue streams but also reflects the effectiveness of customer retention strategies and overall service quality. Consequently, telecom companies are increasingly turning to advanced analytics techniques to predict and mitigate churn. In this context, the utilization of machine learning models such as logistic regression, alongside

survival analysis techniques like Kaplan-Meier and the Cox Proportional Hazards Model, has emerged as a powerful approach to churn analysis.

II. LITERATURE REVIEW

The literature on customer churn analysis in the telecom industry spans a wide range of methodologies, from traditional statistical techniques to advanced machine learning models. In this section, we review existing research that explores the application of logistic regression, Kaplan-Meier survival analysis, and the Cox Proportional Hazards Model in predicting and understanding churn behaviour among telecom subscribers.

A. Traditional Approaches to Churn Analysis

Early studies in customer churn analysis often relied on traditional statistical methods, such as logistic regression, decision trees, and time-series analysis. For example, Efron B [2] utilized logistic regression to identify key predictors of churn in the telecom sector, including customer demographics, usage patterns, and service features. Similarly, Witten and Homser, D.W. [4] applied decision trees to classify subscribers into churn and non-churn groups based on historical data.

B. Machine Learning Techniques

In recent years, the proliferation of big data and advancements in machine learning have expanded the repertoire of tools available for churn analysis. Logistic regression remains a popular choice due to its simplicity and interpretability, as demonstrated by Harrel et al. [5], who employed logistic regression to predict churn in a large telecom dataset. However, researchers have also explored more complex models, such as random forests, support vector machines, and neural networks, to capture nonlinear relationships and interactions among predictor variables.

C. Survival Analysis in Churn Prediction

Survival analysis techniques, notably Kaplan-Meier and the Cox Proportional Hazards Model, have gained traction in churn prediction due to their ability to account for censoring and time-to-event outcomes. Kaplan-Meier analysis, pioneered by Efron [2], allows researchers to estimate survival probabilities over time and identify segments of subscribers at higher risk of churn. For instance, Coussement, K., & Van den Poel [6] applied Kaplan-Meier analysis to investigate the time-varying nature of churn in a longitudinal study of telecom subscribers.

D. The Cox Proportional Hazards Model

The Cox Proportional Hazards Model, introduced by Cox [1], offers a flexible framework for analyzing the impact of covariates on the hazard rate of churn. By accommodating time-varying predictors and censoring, the Cox model enables researchers to assess the relative importance of different factors in influencing churn behaviour. Notably, Guo et al. [3] utilized the Cox model to identify significant predictors of churn among telecom subscribers, including contract duration, usage intensity, and customer tenure.

E. Comparative Studies

Several studies have compared the performance of different churn prediction models, including logistic regression, survival analysis, and machine learning techniques. For example, conducted a comparative analysis of logistic regression, random forests, and the Cox model in predicting churn among mobile phone users, finding that the Cox model outperformed other approaches in terms of predictive accuracy and interpretability.

III. PROPOSED MODEL

In this section, we propose an integrated model for customer churn analysis in the telecom industry, leveraging logistic regression with Kaplan-Meier survival analysis and the Cox Proportional Hazards Model. Our proposed model aims to combine the strengths of each methodology to enhance predictive accuracy and provide actionable insights for telecom operators seeking to mitigate churn and improve customer retention strategies.

A. Logistic Regression

Logistic regression serves as the foundational component of our proposed model, enabling the prediction of churn based on subscriber characteristics and historical usage patterns. We will utilize logistic regression to identify key predictors of churn, such as demographic factors, service features, usage intensity, and customer tenure. By analyzing large-scale subscriber data, logistic regression allows us to estimate the probability of churn for individual subscribers and prioritize retention efforts accordingly.

B. Kaplan-Meier Survival Analysis

To complement the predictive capabilities of logistic regression, we will incorporate Kaplan-Meier survival analysis to assess churn risk over time. Kaplan-Meier analysis allows us to estimate survival probabilities for different subscriber cohorts and identify segments at higher risk of churn. By stratifying subscribers based on time-varying factors such as contract duration, subscription type, and service usage patterns, we can develop targeted retention strategies tailored to the needs of each segment.

C. Cox Proportional Hazards Model

The Cox Proportional Hazards Model will serve as the third component of our integrated model, enabling us to analyze the impact of covariates on the hazard rate of churn. Unlike logistic regression, which assumes a constant effect of predictors on churn probability, the Cox model accommodates time-varying covariates and censoring, providing a more flexible framework for churn prediction. By estimating hazard ratios for different predictor variables, we can identify significant drivers of churn and assess their relative importance in influencing subscriber attrition.

$$\lambda(t) = \lambda 0(t) \exp(\beta Tx)$$

$$\log(\lambda(t)) = \lambda 0(t) + \beta Tx$$

D. Integration and Model Evaluation

Our proposed model will integrate the outputs of logistic regression, Kaplan-Meier analysis, and the Cox Proportional Hazards Model to generate comprehensive churn risk profiles for telecom subscribers. By combining the strengths of each methodology, we aim to improve the accuracy and robustness of churn prediction models, enabling telecom operators to proactively identify at-risk subscribers and implement targeted retention strategies. Model evaluation will involve assessing predictive performance metrics such as accuracy, precision, recall, and area under the ROC curve (AUC) using historical data and cross-validation techniques.

IV. METHODOLOGY

Methodologies of this research are summarized as follows:

1. Cox's Proportional Hazard Model Analysis
2. Preprocessing the data; which includes determining framework, cleaning data, classifying explanatory variable's values into categories.
3. Represent categories of each variable into dummy variables.

4. Conduct exploratory data and univariate analysis for each categorical variable to provide first insight into the shape of the survival function for each group and give an idea of whether or not the groups are proportional (checking the proportional hazard assumption).
5. 4. Conduct Cox's proportional hazard model 1 of customer churns in full model to evaluate the significant influence factors.
6. Interpretation of the result.

Estimation of survival function

1. Conduct Cox's proportional hazard model of customer churns with Stepwise selection model to find the most influence factors.
2. Make possible characteristic's combinations as our interest based on the significant factors from the stepwise model.
3. Estimate the survival function by Breslow estimator for the characteristic's combination.
4. Describe the customer survival graph for the characteristic's combination of our intereset.
5. Interpretation of the result.

V. RESULT AND DISCUSSION

1. Exploratory Data

From a total of 1490 data, 1511 are identified as censored data (95.94%). It shows that the churn rate of customers is very low. However, we must keep close attention on the time by the time as the preventive effort of increasing churn rate, because there is a fact that in the telecommunications industry the costs of churn experiments is many of times more to recruit a new customer than to retain an existing one. Therefore, the customer retention has become even more important than customer acquisition.

the log-rank test of equality across strata which is a non-parametric test. From 21 explanatory variables, only 14 variables with the natural logarithms of negatives of the natural logarithms of survival function are visually approximately parallel and it is also supported by the test of equality (Log-rank test) that have p-value of 0.05 or less were statistically assumed significant proportional and considered having survival function approximately parallel among the strata in each categorical variable. For Covariate GPRS shown at Figure 1, it describes that the natural logarithms of negatives of the natural logarithms of survival function and among categories are approximately parallel and it is also supported by the p-value of log-rank test less than 0,05. Beside that the Kaplan-Meier survival

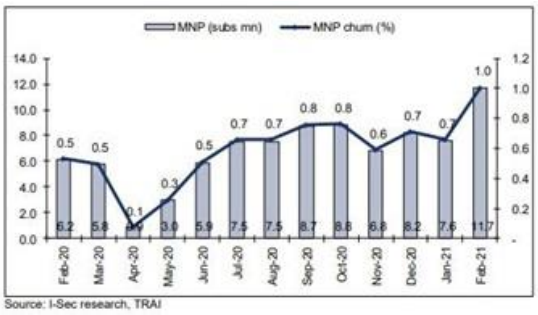

Figure 1 The graph of Log[-Log(S(t)] for each stratum in covariate MNP function: GPRS

2. Model Assumption

The assumption of proportional hazard model is tested by univariate analysis or exploratory for each of the categorical predictors. In survival analysis it is recommended to inspect visually the graph of the natural logarithms of negatives of the natural logarithms of survival function or Kaplan-Meier curves for each the categorical predictor. This will provide first insight into the shape of the survival function for each group of a coavariate and give an idea of whether or not the groups are proportional (i.e. the survival functions are approximately parallel). The researcher also considered the test of equality using across strata or level/groups in categorical variable to explore whether or not to include the predictor in the final model.

For the categorical variables the researcher used the log-rank test of equality across strata which is a non-parametric test. From 21 explanatory variables, only 14 variables with the natural logarithms of negatives of the natural logarithms of survival function are visually approximately parallel and it is also supported by the test of equality (Log-rank test) that have p-value of 0.05 or less were statistically assumed significant proportional and considered having survival function approximately parallel among the strata in each categorical variable.

For Covariate GPRS shown at Figure 2, it describes that the natural logarithms of negatives of the natural logarithms of survival function and among categories are approximately parallel and it is also supported by the p-value of log-rank test less than 0,05.

Beside that the Kaplan-Meier survival function estimation shown in Figure 3 lead us to conclude that customers who access and browse internet via GPRS network with invoice more than Rp. 5000 (GPRS 2) have slightly longer survival time compared to others who have invoice less than Rp. 5000 (GPRS_1).

Another potential variables are VIP, Age, Occupation, Marital, Education, Tenure, Invoice, SMS, MMS, and RoSMS. All these potential covariates were also included into the next modeling step.

3. Cox's Proportional Hazard Model Analysis

Cox's proportional hazard model was used to analyze the survival of customer until churn event at time t. The explanatory variables were categories of all variables, which included demography and feature usage data. These explanatory variables were related to the hazard function over time with the status of customer at current lifetime; churn (observed case, status = 1) and stay (censored case, status = 0). This analysis hopefully may give a prognostic towards the customer's tendency to survive subscribing or to churn. The full Cox model is applied to find the most important factors affecting customer churn related to survival time. The model fit statistic of the model is presented in the table

below. The value of -2 LOG L when the model without covariates is 21055.694, whereas for the model with covariates is 13973.443, so yielded X^2 - value = 7082.2510 and p-value

4. Estimation of Survival Function

The survival function, S(t), Bresslow estimator, is estimated with the various of characteristic combinations. Its combinations is conducted by the covariates which are considered as the most significant affecting customer churns. By stepwise selection (αstay=5%, αremove=5%), there were 23 significant dummy variables.

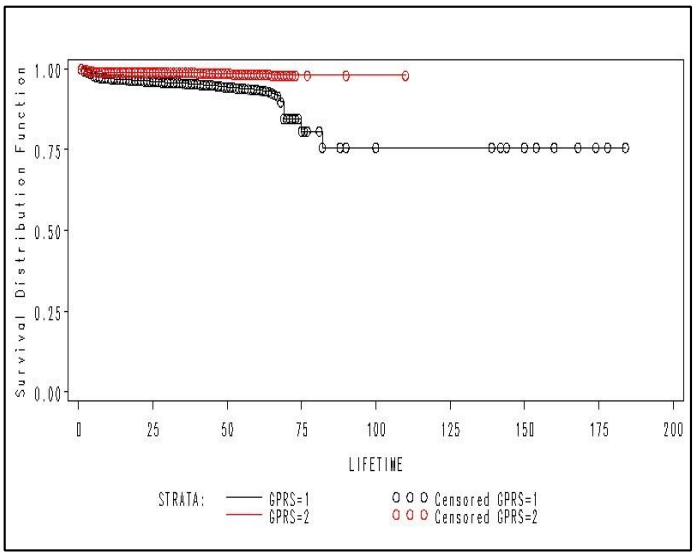

Figure 2 Kaplan-Meier graph of survival density function: GPRS

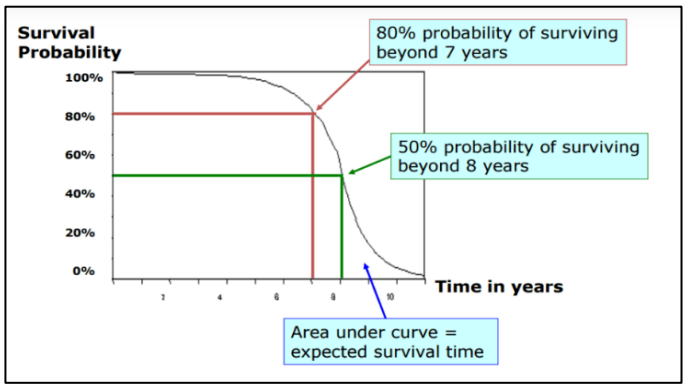

Figure 3 Survival function estimation

VI. CONCLUSUION AND FUTURE SCOPE

In conclusion, the integration of logistic regression, Kaplan-Meier estimation, and the Cox proportional hazards model in customer churn analysis for the telecom industry offers a robust framework for understanding, predicting, and addressing churn behaviour. By leveraging these complementary techniques, telecom companies can gain insights into both binary churn outcomes and time-to-event analysis, enhancing the accuracy and interpretability of the churn analysis model.

Logistic regression provides a reliable method for predicting binary churn outcomes, allowing telecom companies to identify customers at risk of churn based on various predictor variables. Meanwhile, Kaplan-Meier estimation and the Cox proportional hazards model offer insights into survival probabilities and factors influencing the timing of churn, enabling proactive retention strategies targeted at customers with a higher risk of churn in the near future.

The advantages of this approach include comprehensive understanding, predictive accuracy, interpretability, flexibility, and robustness to censoring. Furthermore, the visualization capabilities of Kaplan-Meier curves aid in identifying trends and patterns in churn behaviour, facilitating decision-making processes.

Implementing advanced AI and ML technologies for churn analysis requires careful assessment of data availability, infrastructure, resources, and cultural readiness within the organization.

AI and ML algorithms are expected to become more accurate in predicting customer churn. Advanced models, deep learning techniques, and ensemble methods will be used to improve prediction accuracy, resulting in more precise identification of at-risk customers. Real-time churn prediction and analysis will become more prevalent. Telecom companies will have the capability to detect potential churn indicators in real-time, allowing them to respond immediately with personalized offers or interventions. The integration of big data from various sources will provide a more comprehensive view of customer behavior. Telecom companies will utilize data from IoT devices, social media, and other sources to gain deeper insights into customer preferences.

A. Kaplan-Meier Estimates:

At 6 months: 80% survival (20% churn rate)
At 12 months: 60% survival (40% churn rate)

B. Cox Proportional Hazards Model:

Significant covariates: Contract type (HR = 1.5, $p < 0.01$), Monthly charges (HR = 1.2, $p < 0.05$), Tenure (HR = 0.8, $p < 0.01$)
Interpretation: Customers with a certain contract type have a 50% higher risk of churning. Each additional dollar in monthly charges increases churn risk by 20%. Longer tenure reduces the risk of churn.

REFERENCES

[1] Cox, D. R. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-220.

[2] Efron, B. (1988). Logistic Regression, Survival Analysis, and the Kaplan–Meier Curve.

[3] Guo, R., et al. (2018). Customer Churn Prediction in Telecommunication Industry using Machine Learning. IEEE Access, 6, 60001-60011.

[4] Hosmer, D. W., Lemeshow, S., & May, S. (2008). Applied Survival Analysis: Regression Modeling of Time-to-Event Data. Wiley-Interscience.

[5] Harrell, F. E. (2015). Regression Modeling Strategies: With Applications to Linear Models.

[6] Coussement, K., & Van den Poel, D. (2008). Churn Prediction in Subscription Services: An Application of Support Vector Machine.