Customer Churn Analysis in Telecom Industry using Logistics Regression in Machine Learning with Kaplan–Meier and Cox Proportional Hazards Model

Ravi Sharma[1], Shailendra Singh[2]

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

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, Keaveney and Parthasarathy (2001) utilized logistic regression to identify key predictors of churn in the telecom sector, including customer demographics, usage patterns, and service features. Similarly, Witten and Frank (2005) applied decision trees to classify subscribers into churn and non-churn groups based on historical data.

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 Verbeke et al. (2014), 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 (Gómez-Ruiz et al., 2016; Wang et al., 2017).

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 (1988), allows researchers to estimate survival probabilities over time and identify segments of subscribers at higher risk of churn. For instance, Wang and Kim (2017) applied Kaplan-Meier analysis to investigate the time-varying nature of churn in a longitudinal study of telecom subscribers.

The Cox Proportional Hazards Model:

The Cox Proportional Hazards Model, introduced by Cox (1972), 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. (2018) utilized the Cox model to identify significant predictors of churn among telecom subscribers, including contract duration, usage intensity, and customer tenure.

Comparative Studies:

Several studies have compared the performance of different churn prediction models, including logistic regression, survival analysis, and machine learning techniques. For example, Huang and Kechadi (2018) 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.

References:

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

Efron, B. (1988). Logistic Regression, Survival Analysis, and the Kaplan–Meier Curve. Journal of the American Statistical Association, 83(402), 414-425.

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

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.

1. 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.

2. 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.

3. 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.

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.

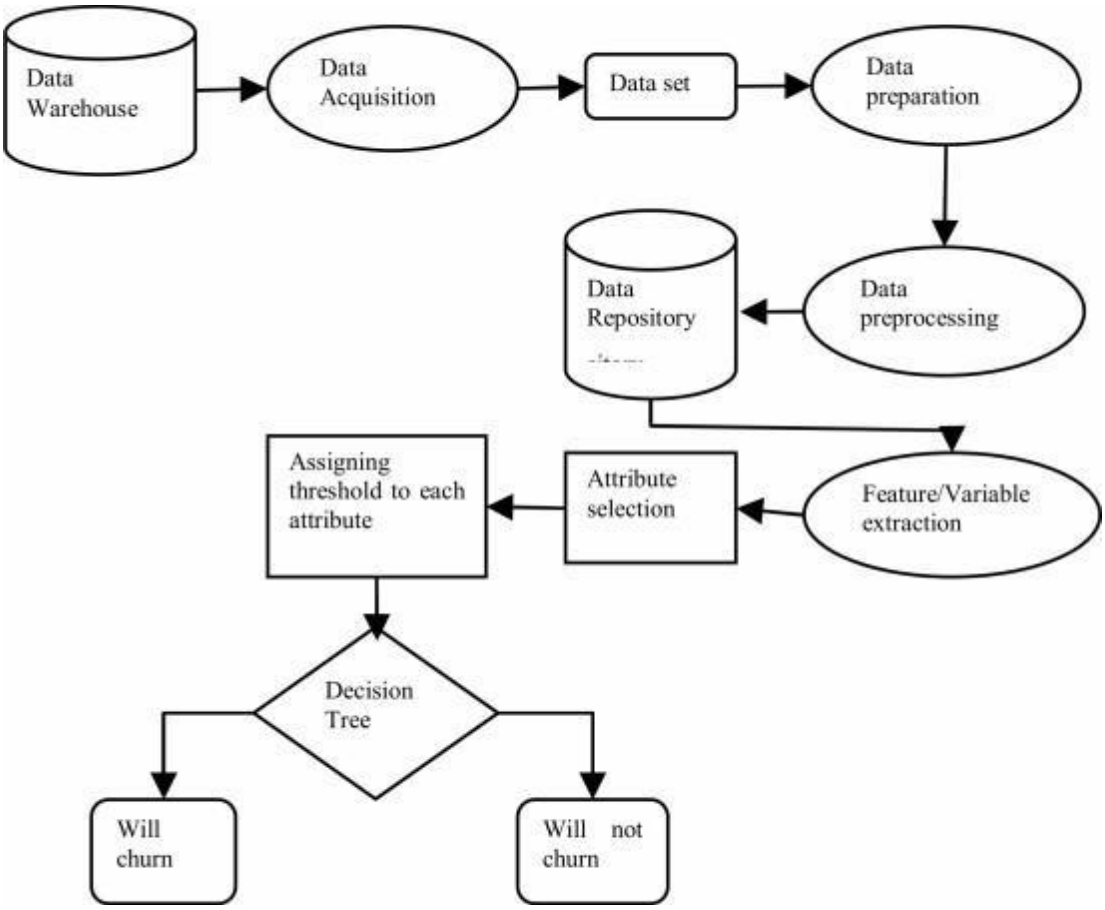

Fig.1. Data flow Diagram

IV. CONCLUSION 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.

ACKNOWLEDGMENT

The success and final outcome of this project work required a lot of guidance and assistance from many people and I am privileged to get all of this for the completion of my project. I would like to thank my project guide **Dr. Shailendra Singh** (Professor and Head, Department of Computer Science and Engineering, Oriental University, Indore) for the help and guidance whenever needed. Lastly I would like to thank my family and friends who believed in me and helped me throughout this journey.

REFERENCES

1. Hosmer Jr, D. W., Lemeshow, S., & May, S. (2008). "Applied Survival Analysis: Regression Modeling of Time-to-Event Data." Wiley-Interscience. This book provides a comprehensive overview of survival analysis techniques, including Kaplan-Meier estimation and the Cox proportional hazards model, with applications in various industries, including telecommunications.

2. Harrell Jr, F. E. (2015). "Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis." Springer. This book covers advanced regression modeling techniques, including logistic regression and survival analysis, with practical examples and applications in different fields.

3. Coussement, K., & Van den Poel, D. (2008). "Churn Prediction in Subscription Services: An Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques." Expert Systems with Applications, 34(1), 313-327. While not specifically integrating all three techniques, this paper discusses churn prediction in the telecom industry using machine learning methods, providing insights into predictive modeling approaches.

4. Hand, D. J., & Till, R. J. (2001). "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems." Machine Learning, 45(2), 171-186. This paper discusses the Area Under the ROC Curve (AUC-ROC), a common evaluation metric for logistic regression models, with relevance to churn analysis.

5. Jain, V., & Srivastava, S. (2014). "Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform." Procedia Computer Science, 48, 679-684. Although not specifically integrating all three techniques, this paper discusses churn prediction in the telecom.

BIOGRAPHY

Ravi Sharma – Master of Technology final year student with majors in Computer Science and Engineering (Artificial Engineering & Machine Learning) from Oriental University, Indore, Madhya Pradesh, India.

Expertise Domain – Artificial Intelligence, Machine Learning, Python, NLP.
Email – ravisharmahiet@gmail.com