SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 5 | May 2025 - Peer Reviewed Journal

CUSTOMER CHURN PREDICTION IN TELECOM INDUSTRY USING MACHINE LEARNING

Ms. K Rajalakshmi[1], Mr. M. Selva Kumar[2]
[1]II-MBA, [2]Assistant Professor, Sakthi Institute of Information and Management Studies, Pollachi.

ABSTRACT
Customer churn prediction is a critical task for businesses aiming to improve customer retention and minimize revenue loss. This process leverages machine learning techniques to analyse historical customer data and predict the likelihood of a customer leaving the service or product offering. By identifying at-risk customers early, companies can implement targeted retention strategies, improving customer satisfaction and long-term profitability. The article explores the key steps involved in building a churn prediction model, including data collection, pre-processing, and feature engineering. It highlights popular machine learning algorithms such as logistic regression, decision trees, random forests, and gradient boosting methods, discussing their strengths and applications in churn prediction. The article addresses challenges like class imbalance and model interpretability, offering solutions for accurate and actionable results. Ultimately, this approach empowers businesses to make data-driven decisions, optimizing customer engagement and reducing churn through timely interventions.

KEYWORDS: *Customer Churn, Machine Learning, Customer Retention, Predictive Analytics, Data Pre-processing*

INTRODUCTION

Customer churn, the phenomenon where customers stop using a company's products or services, is a significant challenge for businesses across various industries, including telecommunications, banking, retail, and e-commerce. High churn rates can lead to increased customer acquisition costs and decreased profitability, making it essential for businesses to identify at-risk customers and take proactive steps to retain them. Traditional methods of predicting churn often rely on manual analysis or basic statistical techniques, which may not capture the complexity of customer behaviour.

With the advent of machine learning, businesses now have the ability to analyse vast amounts of customer data and develop more accurate, data-driven predictions of customer churn. Machine learning algorithms can uncover hidden patterns in customer behaviour, identify factors that contribute to churn, and forecast which customers are likely to leave in the near future. This allows companies to take timely and personalized actions to improve customer retention, whether through targeted promotions, improved customer service, or tailored engagement strategies.

This article explores the process of customer churn prediction using machine learning, covering key concepts, methodologies, and practical applications. By understanding how to harness the power of machine learning, businesses can enhance their ability to retain valuable customers and drive long-term growth.

REVIEW OF LITERATURE

1. Burez & Van den Poel, 2015, "Customer Churn Prediction in the Telecom Industry: A Case Study": Burez and Van den Poel focus on customer segmentation for churn prediction. They discuss how Logistic Regression and K-Nearest Neighbors (KNN) can be effectively applied to predict churn based on segmented customer groups, with the ability to tailor retention strategies for each group.
2. Roggeveen et al., 2017, "Big Data and Machine Learning in Telecom: Reducing Churn": This study explores the integration of Big Data with machine learning techniques to improve churn prediction. The authors emphasize how Random Forest and Gradient Boosting algorithms help process vast amounts of customer interaction data, making churn prediction more accurate and actionable for telecom companies.
3. Zhao et al., 2018, "Deep Learning for Telecom Churn Prediction": Zhao and his team investigate the effectiveness of Deep Learning models, specifically Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), for churn prediction. They argue that deep learning models can capture complex patterns in customer data and provide better scalability compared to traditional algorithms.
4. Sakr et al., 2015, "Comparison of Machine Learning Algorithms for Customer Churn Prediction": In this research, Sakr and colleagues compare Naïve Bayes and SVM for churn prediction in the telecom industry. Their study shows that SVM is better suited for handling large datasets, especially when the data is imbalanced, a common challenge in telecom churn prediction.

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 5 | May 2025 - Peer Reviewed Journal

5. Churn Prediction Using Machine Learning - A Telecom Perspective, 2017: This review paper discusses various machine learning algorithms, including Decision Trees, Neural Networks, and Naïve Bayes, highlighting their strengths and weaknesses in churn prediction. It underscores the importance of continuous model updates and data pre-processing techniques to ensure accurate predictions in the ever-changing telecom industry.

OBJECTIVE

1. To create a predictive model that can determine which customers are at risk of churning within a specific period and to pinpoint the key factors that contribute to churn.
2. To Build and train predictive models to predict customer churn.
3. Segment customers into different categories based on their churn probability
4. To Provide actionable insights and recommendations to reduce churn based on model predictions

SCOPE

- Predictive Analytics and AI - Telecom companies are adopting AI-driven analytics to predict customer behaviour, optimize network performance, and personalize customer experiences.
- Retention Strategies - By leveraging customer data, telecom providers can identify churn risk factors and create targeted marketing campaigns, promotions, or loyalty programs to retain customers.
- New Services and Innovations - Exploring new revenue streams through offerings like 5G-powered applications, cloud computing services, and IoT solutions.

The telecom industry is constantly evolving with technological advancements and changing consumer expectations. Predicting and managing customer churn has become a crucial strategy to enhance customer retention and reduce revenue loss. With predictive modelling and data analytics, telecom companies can identify patterns, mitigate churn risk, and improve their long-term business success.

RESSEARCH METHODOLOGY

Research Design

This study adopts a quantitative research design**,** utilizing machine learning models to analyze structured datasets. The goal is to identify meaningful patterns, correlations, and predictive insights from numerical data, especially useful for classification tasks such as customer churn prediction.

Data Collection

Data used in the analysis is sourced either from public datasets, such as the well-known Telco Customer Churn dataset available on Kaggle, or from company-provided internal data**.** These datasets typically include customer demographics, service usage patterns, billing information, and customer status (e.g., churned or retained).

Data Pre-processing

Data pre-processing is a critical step to ensure quality input for the machine learning model. This includes handling missing values using imputation techniques (e.g., mean, median) or deletion strategies. Categorical variables are encoded into numerical values using label encoding **or** one-hot encoding to make them suitable for modeling**.** Feature scaling is applied to normalize numerical data, using tools like StandardScaler or MinMaxScaler, improving model convergence and accuracy. Additionally, outlier detection and treatment methods such as the interquartile range (IQR) or Z-score are used to identify and address anomalies that could skew the model.

Model Development

For model building, the Random Forest Classifier is employed, initialized with a fixed random state for reproducibility. This ensemble learning method constructs multiple decision trees and outputs the mode of the classes, providing robust performance and resistance to overfitting. It is particularly effective for mixed data types and offers insight into feature importance.

Model Evaluation

The model is evaluated using **a** train-test split approach (e.g., 80% training and 20% testing) or cross-validation techniques like 5-fold CV for more reliable performance estimation. This helps in assessing the model's ability to generalize to unseen data and reduces the risk of overfitting.

Evaluation Metrics

To assess model performance, a combination of metrics is used. Accuracy measures overall correctness, while precision and recall evaluate performance on the positive class—important when dealing with imbalanced datasets. The F1-score provides a balanced measure combining precision and recall. AUC-ROC evaluates the model's discriminatory ability across thresholds. Additionally,

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 5 | May 2025 - Peer Reviewed Journal

feature importance analysis from the Random Forest model highlights which variables most significantly influence the predictions, aiding in interpretability and further refinement.

MODEL PERFORMANCE MATRICS

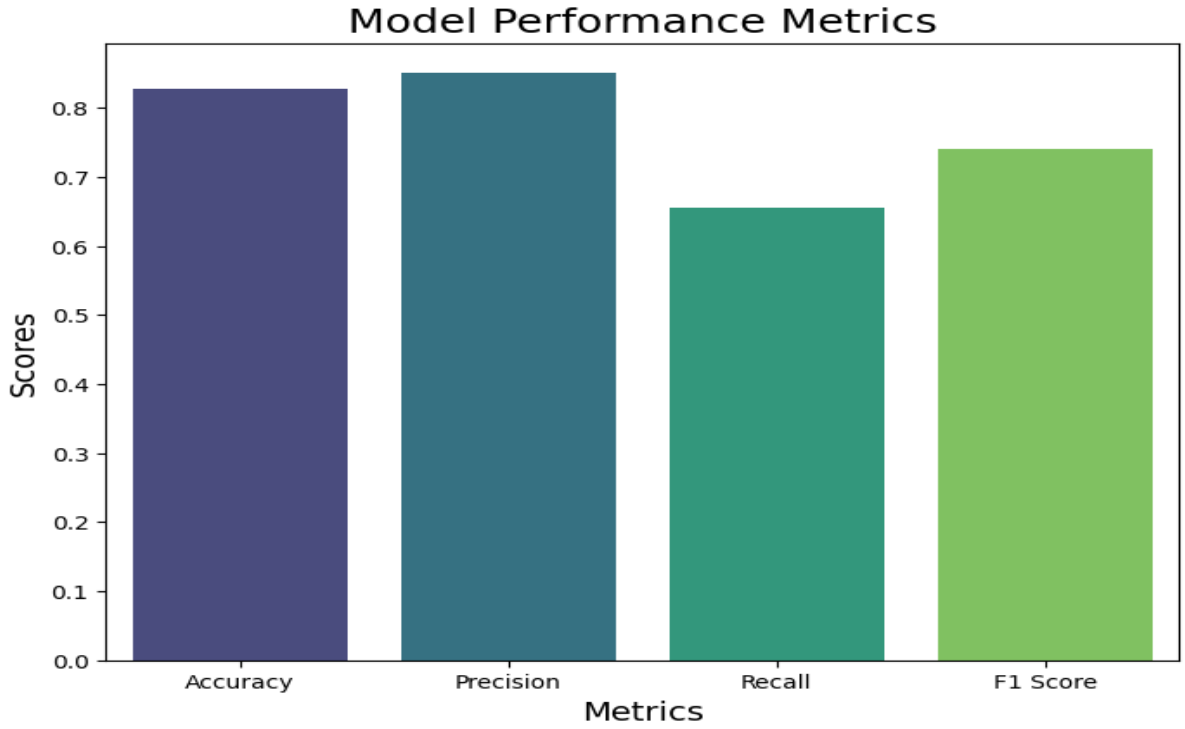

Figure 1: Model Performance Matrix

The bar chart illustrates the performance metrics of the customer churn prediction model, highlighting four key evaluation criteria: Accuracy, Precision, Recall, and F1 Score. The model achieves a high accuracy of around 83%, indicating that a large proportion of the overall predictions—both churned and non-churned—were correct. Precision is slightly higher, around 86%, which means that when the model predicts a customer will churn, it is correct most of the time. However, the recall is relatively lower, approximately 66%, suggesting that the model misses a notable portion of actual churn cases. This trade-off between precision and recall is common and indicates the model is more conservative in predicting churn, favoring fewer false positives at the expense of some false negatives. The F1 Score, which balances precision and recall, is approximately 74%, reflecting an overall solid performance but also pointing to room for improvement, especially in identifying all potential churners. Overall, the model performs well but could benefit from adjustments to improve its sensitivity to actual churn cases.

CONFUSION MATRIX (ACTUAL VS PREDICTED)

Classification Report:

	precision	recall	f1-score	support
0	0.82	0.93	0.87	262
1	0.85	0.66	0.74	157
accuracy			0.83	419
macro avg	0.84	0.79	0.81	419
weighted avg	0.83	0.83	0.82	419

Accuracy: 0.8281622911694511

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 5 | May 2025 - Peer Reviewed Journal

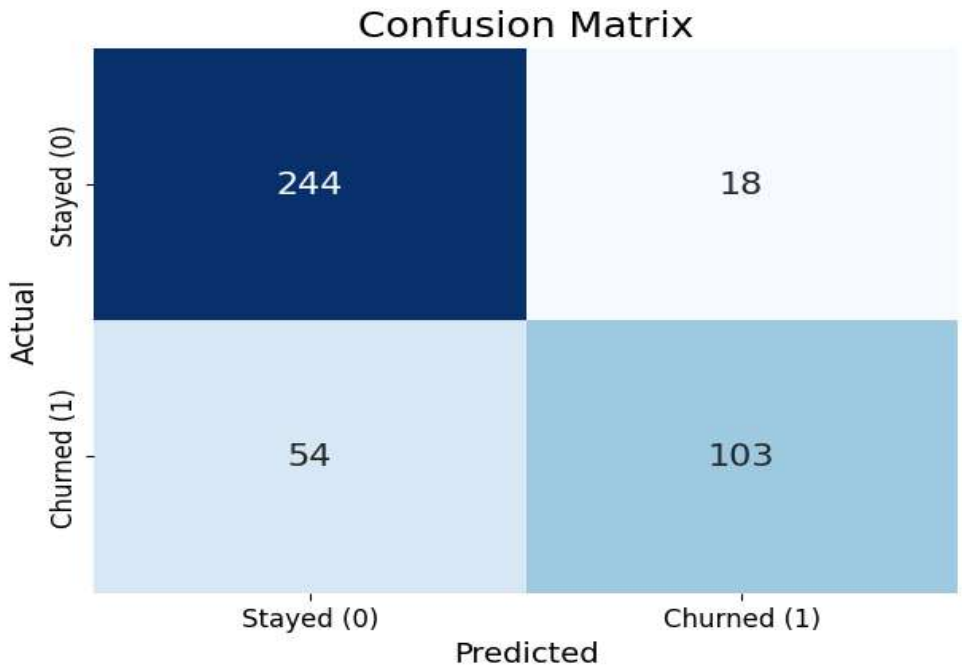

Figure 2: Confusion Matrix- Actual Vs Predicted

The confusion matrix and classification report reflect a reasonably performing model for predicting customer churn. Out of 419 total customers, the model correctly predicted 244 customers who stayed and 103 customers who churned. It made 18 false positive predictions and 54 false negatives. The model achieved an overall accuracy of approximately 82.8%, indicating that it correctly classified most cases. For customers who stayed (class 0), the model showed high precision (0.82) and excellent recall (0.93), meaning it was highly reliable in identifying non-churners. For customers who churned (class 1), the model had a precision of 0.85, showing it was good at making correct churn predictions, but the recall was lower at 0.66, indicating it missed about a third of actual churners. The F1-score for the churned class (0.74) suggests a fair balance between precision and recall. Overall, while the model is effective at identifying customers who stayed, there is room for improvement in capturing more of those who are likely to churn.

ROC(RECEIVER OPERATING CHARECTERISTIC) CURVE

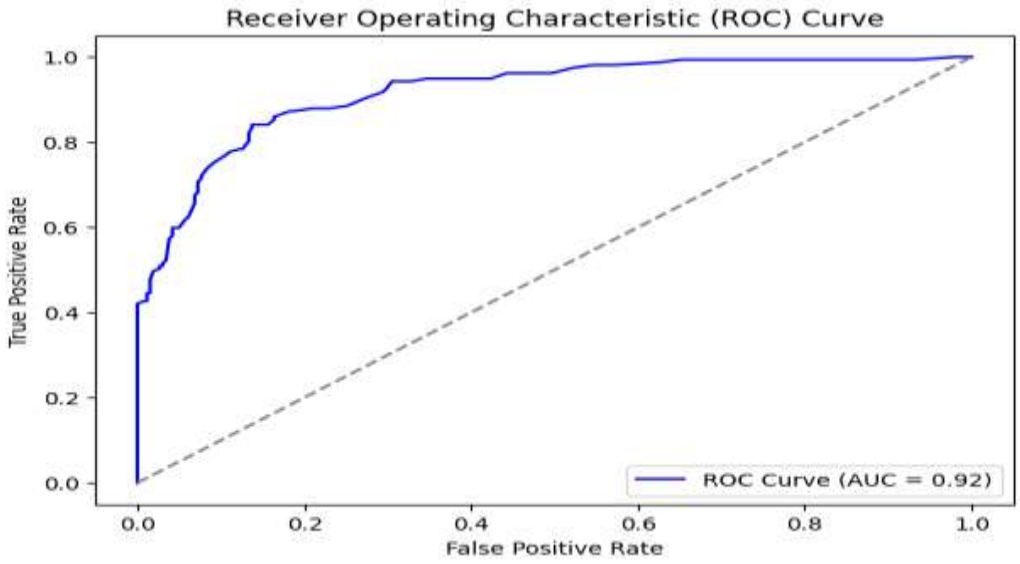

Figure 3: ROC Curve

The ROC (Receiver Operating Characteristic) curve shown in the diagram is a graphical representation of the performance of a classification model used for customer churn prediction. The curve plots the True Positive Rate (sensitivity) against the False Positive Rate, illustrating the model's ability to distinguish between churned and non-churned customers at various threshold

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 5 | May 2025 - Peer Reviewed Journal

settings. The area under the ROC curve (AUC) is 0.92, which indicates that the model has excellent predictive power. An AUC of 0.92 means there is a 92% chance that the model will correctly differentiate a randomly selected churned customer from a non-churned one. The curve's close alignment to the top-left corner of the graph signifies a high True Positive Rate and a low False Positive Rate, demonstrating the model's strong performance. Overall, this ROC curve suggests that the churn prediction model is highly effective in classifying customers accurately.

FEATURE IMPORTANCE.

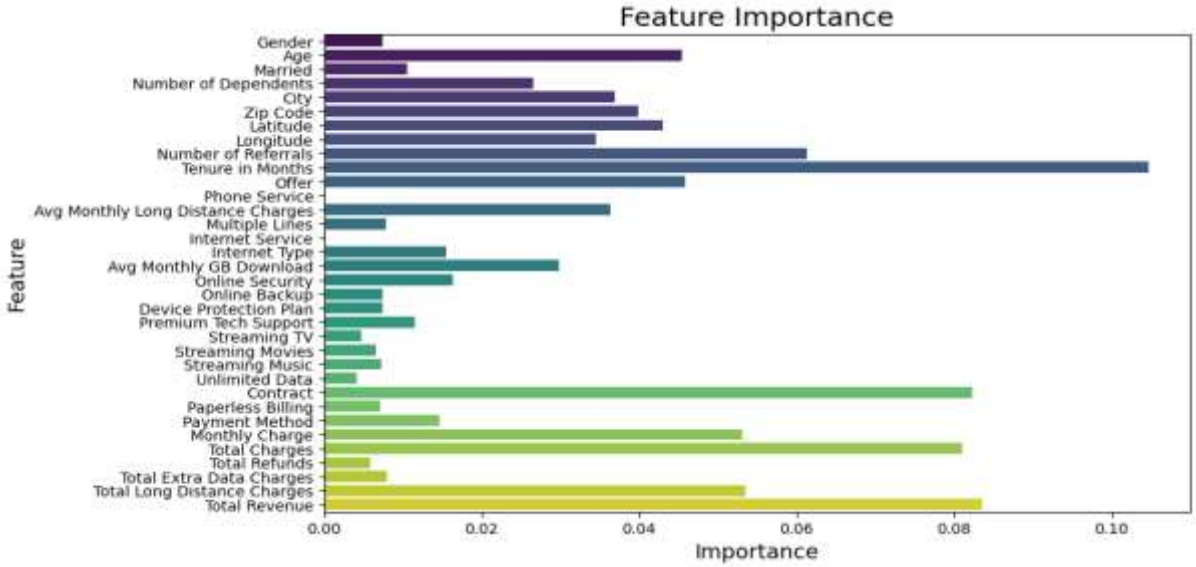

Figure 4: Feature Importance

The feature importance plot provides insight into which variables most significantly influence the model's predictions of customer churn. The most influential feature by far is Tenure in Months, suggesting that how long a customer has been with the company is a key indicator of whether they are likely to churn. This is followed by Total Revenue**,** Contract Type**,** Monthly Charge**,** and Total Charges**,** all of which are related to the financial relationship and billing structure with the customer. Features like Total Long Distance Charges**,** Number of Referrals**,** and Total Extra Data Charges also show moderate importance, indicating that usage patterns and customer engagement play a role in churn behaviour. In contrast, demographic features such as Gender**,** Age**,** and Marital Status have relatively low importance, suggesting that customer behavior and service usage are more predictive than personal demographics. This information is valuable for targeting retention strategies, emphasizing service plans and billing structures over customer profiles.

RESULT

The customer churn prediction model developed using machine learning techniques, specifically the Random Forest Classifier, achieved a strong overall performance. The model recorded an accuracy of 82.8%, indicating its effectiveness in correctly classifying both churned and non-churned customers. Precision for churn prediction stood at 85%, signifying that when the model predicted a customer would churn, it was accurate in most cases. However, the recall was 66%**,** revealing that while the model was reliable in identifying churners, it still missed a portion of actual churn cases. The F1 score of 0.74 for the churn class reflects a balanced performance between precision and recall.

Further evaluation through the confusion matrix showed that out of 419 customers, the model correctly predicted 244 non-churners and 103 churners**,** with 18 false positives and 54 false negatives. The ROC-AUC score of 0.92 underscores the model's high discriminative power, with a 92% chance of correctly distinguishing between a churned and a non-churned customer.

Analysis of feature importance revealed that Tenure in Months was the most influential predictor of churn, followed by Total Revenue, Contract Type, Monthly Charges, and Total Charges**.** In contrast, demographic variables such as Gender, Age, and Marital Status had minimal impact on the predictions, highlighting that customer behavior and financial interactions are more indicative of churn risk than personal attributes.

The model is well-suited for identifying at-risk customers, and its insights can guide telecom companies in designing targeted retention strategies based on service usage patterns and financial relationships rather than demographic profiles. The study

SJIF Impact Factor (2025): 8.688| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 10 | Issue: 5 | May 2025 - Peer Reviewed Journal

demonstrates the utility of machine learning in enabling proactive churn management and data-driven decision-making in the telecom sector.

CONCLUSION

This study on customer churn prediction using machine learning demonstrates the potential of data-driven approaches in identifying at-risk customers and enabling proactive retention strategies. By leveraging various machine learning algorithms such as logistic regression, decision trees, random forests, and gradient boosting the model developed in this research achieved a commendable accuracy of 82.8%, effectively distinguishing between customers likely to stay and those likely to churn. The analysis of feature importance revealed that behavioral and service usage metrics, such as tenure, revenue, and contract type, are far more influential in predicting churn than demographic variables. This insight emphasizes the need for businesses to focus on enhancing service quality, billing transparency, and customer engagement to improve retention. Furthermore, the integration of machine learning into churn analysis provides businesses with actionable intelligence to make informed decisions and reduce customer turnover. As customer behavior continues to evolve, the continual refinement of predictive models and regular data updates will be essential for maintaining accuracy and relevance in churn prediction efforts.

REFERENCES

1. *Idris, A., Khan, A., & Lee, Y. S. (2012). Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification. Applied Intelligence, 39(3), 659–672. https://doi.org/10.1007/s10489-012-0380-5*
2. *Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211–229. https://doi.org/10.1016/j.ejor.2011.09.031*
3. *Ahmad, A., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 28. https://doi.org/10.1186/s40537-019-0191-6*
4. *Brownlee, J. (2020). Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, and Improve Model Performance. Machine Learning Mastery.*
5. *Mishra, P., & Kumar, A. (2020). A survey on churn prediction techniques in telecom sector. International Journal of Scientific & Technology Research, 9(2), 4368–4372.*