**A SYNOPSIS ON**

**“Customer Churn Prediction and Analysis Using Machine Learning”**

**SUBMITTED TO**

**Rashtrasant Tukadoji Maharaj Nagpur University in the partial fulfillment of the requirement for the degree of Master of Business Administration (M.B.A)**

**SUBMITTED BY**

**Om Kale**

**GUIDED BY**



**Wainganga College of Engineering & Management,**

**NAGPUR**

**2024-2025**

**Table of Content**

|  |  |
| --- | --- |
| **Sr.no** | **Title** |
| 1 | TITLE OF THE PROJECT |
| 2 | INTRODUCTION |
| 3 | LITERATURE REVIEW (IN 3 TO 4 Research Papers) |
| 4 | OBJECTIVE |
| 5 | HYPOTHESIS |
| 6 | RESEARCH METHODOLOGY |
| 7 | DATA COLLECTION METHOD |
| 8 | DATA ANALYSIS METHOD |
| 9 | BIBLIOGRAPHY AND REFERENCES |

**TITLE OF THE PROJECT**

**Customer Churn Prediction and Analysis Using Machine Learning**

**INTRODUCTION**

Customer churn is a significant challenge for businesses, particularly in e-commerce and subscription-based services. Retaining customers is often more cost-effective than acquiring new ones, making churn prediction an essential aspect of business strategy. Churn occurs when customers discontinue their association with a company, leading to revenue loss and increased marketing expenses to attract new users. By proactively identifying potential churners, businesses can implement targeted interventions to enhance customer retention.

In today’s data-driven landscape, machine learning (ML) plays a crucial role in predicting churn by analyzing large datasets and uncovering hidden patterns in customer behavior. Advanced ML algorithms can process vast amounts of structured and unstructured data, providing businesses with predictive insights that enable proactive decision-making. These models consider multiple factors such as transaction history, engagement levels, customer service interactions, and subscription status to assess the likelihood of churn.

This study focuses on understanding customer behavior, identifying churn patterns, and leveraging machine learning algorithms to predict customer attrition. By analyzing real-world e-commerce datasets, this research aims to provide insights into customer interactions, purchase history, and engagement levels. Businesses can use these insights to develop targeted retention strategies and enhance customer satisfaction, ultimately improving profitability.

Moreover, predictive churn models empower businesses to personalize customer experiences by offering tailored recommendations, discounts, and loyalty programs. A well-defined churn prediction system enables companies to segment customers based on their risk levels and allocate resources efficiently to retain high-value customers. This proactive approach not only minimizes revenue loss but also strengthens long-term customer relationships, fostering brand loyalty.

Given the competitive nature of the e-commerce industry, businesses that successfully implement predictive churn models gain a strategic advantage. This research will explore how various ML techniques, such as Decision Trees, Random Forest, and Logistic Regression, can be employed to forecast churn with high accuracy. By integrating visualization tools such as Tableau and Python-based libraries, this study aims to present actionable insights that businesses can leverage to refine their retention strategies.

Overall, this study underscores the importance of churn prediction in business decision-making, emphasizing how data-driven approaches can transform customer retention strategies. The findings will help businesses move from reactive to proactive engagement, ultimately leading to improved customer satisfaction and sustainable growth.

**LITERATURE REVIEW (IN 3 TO 4 Research Papers)**

Several studies have examined customer churn prediction using machine learning approaches. Key findings from previous research include:

**• Predictive Modeling:**   
Research has demonstrated that decision trees, logistic regression, and ensemble learning methods such as Random Forest and Gradient Boosting perform effectively in identifying potential churners. These models provide varying degrees of interpretability and accuracy, with ensemble techniques often achieving superior performance. However, deep learning models, despite their high accuracy, pose challenges in terms of explainability.

**• Feature Engineering:**   
Feature selection plays a crucial role in improving churn prediction models. Studies indicate that incorporating customer demographics, transaction history, product usage patterns, and behavioral data significantly enhances predictive accuracy. Additionally, advanced techniques such as feature transformation and dimensionality reduction further refine the model’s efficiency.

**• Real-time Analytics:**   
Traditional churn prediction models rely on historical data analysis, which, while effective, lacks adaptability to real-time customer behavior. Recent research highlights the importance of real-time predictive analytics, leveraging streaming data to capture changes in customer activity. Integrating live customer data with machine learning models can enhance the responsiveness of churn mitigation strategies by enabling proactive interventions.

**• Gaps in Existing Research:**   
Despite significant advancements in churn prediction, existing studies often focus solely on prediction rather than providing actionable insights for business decision-makers. Many models fail to explain why a customer is likely to churn, limiting their practical application. Additionally, real-time data processing remains a challenge, as many frameworks lack the infrastructure for dynamic model updates. This study aims to address these gaps by offering interpretable results, visualizations, and practical recommendations for customer retention strategies.

**OBJECTIVE**

**The primary objectives of this study are:**

* To develop a machine learning-based predictive model to identify customers at risk of churn.
* To analyze key factors contributing to customer churn and provide insights into customer behavior.
* To utilize machine learning techniques such as decision trees and Random Forest to improve predictive accuracy.
* To integrate business intelligence tools like Tableau for enhanced visualization and reporting.
* To provide actionable recommendations that businesses can implement to reduce churn and increase customer retention.

**HYPOTHESIS**

**The study is based on the following hypotheses:**

* Customers with lower engagement, fewer purchases, and shorter subscription durations are more likely to churn.
* Machine learning models such as Decision Trees and Random Forest significantly improve churn prediction accuracy compared to traditional statistical methods.
* High accuracy in churn prediction allows businesses to deploy targeted retention strategies, leading to improved customer retention and increased revenue.

**RESEARCH METHODOLOGY**

**6.1 Universe**

The research focuses on e-commerce customers, particularly those who engage in online shopping and subscription-based services. The data includes various factors such as:

* Customer demographics (age, gender, location)
* Purchase history (order frequency, transaction value, time since last purchase)
* Engagement metrics (website visits, customer support interactions, loyalty program participation)

**6.2 Sampling Plan**

A dataset comprising thousands of customer records from an e-commerce platform is used. To ensure a diverse and representative dataset, stratified sampling is applied. The data is split into training and testing sets to evaluate model performance effectively.

**6.3 Sampling Techniques**

* **Random Sampling**: Applied to select a balanced subset of customers from different engagement levels.
* **Data Preprocessing**:   
  Includes handling missing values, data normalization, and outlier detection to ensure model accuracy.
* **Feature Selection**:   
  Identifying the most critical variables that impact churn, such as purchase frequency, last transaction date, and discount usage.

**DATA COLLECTION METHOD**

The dataset is sourced from e-commerce transaction logs and includes:

• Customer purchase behavior over a specified time period, tracking buying frequency, cart abandonment rates, and total spending.   
  
• Interaction data from customer service, chatbots, and email campaigns, capturing response times, complaint resolutions, and engagement levels.   
  
• Subscription history and membership status, indicating renewal patterns, upgrade or downgrade tendencies, and duration of service.   
  
• Anonymized customer identifiers to ensure data privacy and compliance with ethical standards, following GDPR and other regulatory requirements.   
  
• External data sources, such as market trends, seasonal effects, and promotional campaign impacts, to enhance model accuracy.

The dataset undergoes rigorous pre-processing, including handling missing values, normalizing numerical attributes, and encoding categorical variables, ensuring consistency and reliability for analysis.

**DATA ANALYSIS METHOD**

The data analysis process includes:

**Exploratory Data Analysis (EDA):**

* Understanding the distribution of variables using statistical summaries and visual tools.
* Detecting patterns, anomalies, and trends that may impact customer retention.
* Visualizing correlations among variables through heat maps, scatter plots, and histograms to identify influential factors.

**Feature Engineering:**

* Transforming raw data into meaningful features such as customer lifetime value (CLV), frequency of purchases, and sentiment scores from customer interactions.
* Applying techniques like dimensionality reduction and feature scaling to optimize model efficiency.

**Model Selection:**

* Training and evaluating multiple machine learning models, including Decision Trees, Random Forest, and Logistic Regression, to identify the best-performing algorithm.
* Utilizing techniques such as cross-validation and hyper parameter tuning to improve predictive accuracy.

**Model Evaluation:**

* Using confusion matrices to understand prediction errors.
* Assessing performance using accuracy scores, precision-recall metrics, F1-score, and ROC curves.
* Comparing model predictions with actual churn occurrences to validate effectiveness.

**Visualization:**

* Leveraging Python libraries (matplotlib, seaborn) and Tableau for intuitive and interactive data representation.
* Creating dashboards to illustrate churn trends, customer segmentation, and predictive insights for business decision-making.

**8.1 Decision Tree Visualization**

A decision tree model is used to classify customers into churn and non-churn categories. The tree structure provides transparency in decision-making, highlighting critical factors influencing churn.

• The tree diagram showcases how different customer attributes (e.g., frequency of purchases, response to marketing emails, and service complaints) contribute to churn probability.  
• By analyzing the tree's splits, businesses can identify high-risk customers and implement targeted retention strategies.   
• The model's interpretability allows for easy integration with business operations, enabling decision-makers to develop proactive measures for customer retention.

**8.2 Confusion Matrix Analysis**

A confusion matrix evaluates model performance by comparing actual vs. predicted values. It measures classification effectiveness using four key metrics:

**• True Positive (TP):**   
Correctly identified churn cases, ensuring the model recognizes at-risk customers accurately.   
  
 **False Positive (FP):**   
Customers incorrectly predicted to churn, which may lead to unnecessary retention efforts.   
  
**• True Negative (TN):**   
Accurately classified non-churn customers, confirming model reliability in identifying loyal customers.   
  
**• False Negative (FN):**   
Churners mistakenly predicted as retained customers, which could result in revenue loss if not addressed.

• Precision and recall metrics are analyzed to balance false alarms and missed churn cases.   
• The ROC curve is used to evaluate the trade-off between sensitivity and specificity, ensuring optimal model performance.  
• Insights from the confusion matrix guide strategic adjustments, such as personalized marketing interventions and proactive customer engagement initiatives.

**BIBLIOGRAPHY AND REFERENCES**

The following sources provide foundational knowledge and support for this study:

* Research papers on machine learning applications in customer churn prediction.
* Case studies and industry reports on e-commerce customer retention strategies.
* Documentation for Python libraries including pandas, scikit-learn, matplotlib, and seaborn.
* Business intelligence reports on churn analytics and its impact on profitability.