ISSN 2349-6193
Impact Factor: 2.805

IJESMR

International Journal OF Engineering Sciences & Management Research

PREDICTING CUSTOMER CHURN IN DISTANCE LEARNING: A COMPARATIVE STUDY OF XGBOOST, BAGGING CLASSIFIER, AND SMOTE FOR ENHANCED MODEL PERFORMANCE AND GENERALIZATION

Paril Ghori
parilghori@gmail.com

ABSTRACT

The prediction of customer churn has become a central focus in managing customer relationships, especially for businesses in the distance learning sector. Churn prediction models aim to identify users with a high likelihood of attrition, enabling companies to enhance the effectiveness of their customer retention efforts and reduce churn-associated costs. Although the primary goal of these models is cost reduction and retention, their performance is often assessed using statistical metrics and computational tools, such as machine learning techniques. This study focuses on developing and validating churn prediction models using data from over 150,000 customers of an online education company. The objective was to compare the performance of various machine learning algorithms implemented for churn prediction. Thirteen variables were selected from the literature, and the models were developed through four main steps: (I) training on balanced and unbalanced datasets, (II) generalization/testing on an independent dataset, (III) statistical comparison of the algorithms, and (IV) evaluation of the models with the highest accuracy. The results showed that the best-performing model for unbalanced classes was XGBoost, with an average accuracy of 87.11% and an average AUC (Area Under the Curve) of 0.86. For the balanced classes, the Bagging Classifier performed the best, achieving an average accuracy of 77.34% and an average AUC of 0.83 during both testing and generalization phases.

KEYWORDS – Accuracy, Area Under the Curve (AUC), Bagging Classifier, Churn Prediction, Distance Learning, Machine Learning, Predictive Modeling, SMOTE, XGBoost.

1. INTRODUCTION

In the context of India's competitive market, the need for customer retention has become a critical focus for businesses. Scholars of relationship marketing emphasize that it is often more expensive to acquire new customers than to retain existing ones. This insight is particularly relevant in India, where customer loyalty can be volatile, and the competition is fierce across industries, ranging from telecommunications to e-commerce and education. Understanding the phenomenon of Churn—where customers stop interacting with a business or switch to competitors—is essential for companies aiming to sustain and grow their customer base. In the Indian context, factors such as diverse customer needs, price sensitivity, and regional differences further complicate churn management, making it a critical area of study for businesses operating in the country [1].

Churn, often defined as the number of customers who discontinue using a company's services over a given period, is a common challenge across industries. In India, businesses are grappling with a high churn rate due to factors like increasing price competition, rapidly changing customer expectations, and the ease of switching services offered by competitors. Understanding churn requires businesses to focus on multiple factors such as customer satisfaction, product relevance, and pricing strategies, all of which vary significantly across India's diverse population. By analyzing churn, companies can adopt strategies to retain their most valuable customers, ensuring long-term profitability.

In India's highly competitive market, churn can be a critical metric to understand customer dissatisfaction. It represents the interruption or termination of a customer's relationship with a company, often involving a shift to a competitor. Effective churn management aims to minimize this disruption by developing strategies to retain profitable customers and enhance their engagement. The churn rate can serve as an indicator of dissatisfaction due to factors like poor service, high prices, and better deals offered by competitors. By identifying and analyzing these churn indicators, businesses can craft strategies that improve customer loyalty and reduce churn [2].

To analyze churn effectively, businesses first need to classify their customer base, identifying those who have disengaged or are likely to disengage. Data analysis techniques, including machine learning models such as decision trees, regression analysis, and neural networks, have become integral tools for predicting churn. In India, where customer data can be vast and complex due to the sheer scale of the population, data mining and predictive analytics are vital tools for understanding consumer behavior and making informed decisions. These models often rely on historical data to predict future churn, requiring companies to first identify the relevant independent variables (such as customer demographics, behavior patterns, and transaction histories) that can predict customer attrition [3].

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

The importance of customer satisfaction cannot be overstated in this context. In India, customer satisfaction directly correlates with customer retention and loyalty. The authors of [4] argue that customer satisfaction is a key driver of loyalty, which in turn impacts a company's bottom line. The Indian market, with its vast and diverse consumer base, presents unique challenges for companies in maintaining consistent customer satisfaction. This is particularly relevant in sectors like telecommunications, e-commerce, and education, where the pressure to deliver value consistently is high.

While the concept of Customer Relationship Management (CRM) is widely acknowledged, there is no single, universally accepted definition of CRM. According to [5], CRM represents the values and strategies of relationship marketing applied through human action and technology. In India, CRM systems are increasingly reliant on technology to manage customer relationships, helping companies retain their most valuable customers. These systems, which utilize customer data to predict churn, play a significant role in improving customer retention in industries like telecom, banking, and education.

According to the authors of [6], the primary goal of a relationship management system is to identify potential churners within a large customer base and implement targeted retention strategies to maintain long-term customer loyalty. This approach is particularly relevant in the Indian market, where customers are highly price-sensitive and have access to a wide array of alternatives. By predicting customer churn in advance, companies can take proactive measures to mitigate the risk of losing valuable customers.

The need to adopt advanced data analytics and machine learning techniques to predict churn is especially pressing in India's education sector. With a rapidly growing number of institutions offering distance learning and online education, retaining students has become a key challenge. The education market in India is highly fragmented, with numerous private and public institutions competing for the same pool of students. As competition increases, so does the churn rate, making it essential for educational institutions to predict and prevent student dropouts. By applying machine learning models to predict churn, educational institutions can gain valuable insights into student behavior, offering personalized retention strategies and improving student engagement [7].

Given the growing importance of data-driven decision-making, several studies in India have begun focusing on using machine learning techniques to predict churn and enhance customer retention. These studies utilize large datasets from industries such as telecommunications, retail, and education to develop predictive models that help companies identify customers at risk of leaving. By leveraging predictive analytics, businesses in India can make timely interventions, offering tailored incentives or improvements to retain customers. This shift towards data-driven customer relationship management is transforming industries and is expected to play a key role in the future of business in India.

As an example, consider a company operating in the Indian distance education sector, which has a large base of over 250,000 students. The churn rate in this sector is influenced by factors such as changing student preferences, course relevance, and pricing. By analyzing historical data on customer behavior and transactions, businesses can predict which students are likely to discontinue their courses and take proactive steps to retain them. Such predictive models are invaluable in a market like India, where customer behavior can be unpredictable and highly influenced by external factors such as regional disparities, economic conditions, and cultural preferences.

This paper aims to develop and compare different predictive models using machine learning techniques to forecast churn in the Indian education sector, specifically in the context of distance learning. By utilizing customer behavior and transaction data, the study seeks to build predictive models that can help educational institutions anticipate student churn and implement retention strategies accordingly.

The primary objective of the research is to assess various machine learning algorithms' effectiveness in predicting churn, specifically within the Indian context. To achieve this, the study will:

- Identify relevant historical data on student behavior, purchases, and interactions to inform the development of predictive models.
- Evaluate the predictions made by different machine learning models, including decision trees, support vector machines, and neural networks, among others.
- Compare the accuracy and feasibility of implementing these models in the Indian educational sector, where issues such as high competition and varying student preferences can make churn prediction particularly challenging.

Through this study, we hope to contribute valuable insights into the application of machine learning in churn prediction, offering practical recommendations for educational institutions in India. By adopting such predictive models, institutions can not only reduce churn but also enhance student satisfaction, ultimately leading to better retention rates and improved educational outcomes. The findings of this research can serve as a foundation for further studies and practical applications in the growing Indian education sector.

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

2 LITERATURE REVIEW

2.1 Customer Relationship Management (CRM)

Customer Relationship Management (CRM) is a critical practice for organizations seeking to optimize customer satisfaction and loyalty. The concept of CRM is not solely viewed as a technology but rather as a strategic approach aimed at improving business processes through deeper insights into customer behavior. It encompasses activities designed to acquire, retain, and foster long-term relationships with customers. According to [5], CRM is a business strategy that maximizes profitability, revenue, and customer satisfaction by implementing customer-focused processes. These processes have both tactical and strategic implications and can influence a company's operations at all levels. In this context, CRM is seen as a tool to understand and manage customer needs effectively, anticipate their expectations, and nurture loyalty, ultimately contributing to enhanced retention and profitability [8].

At its core, CRM involves creating value for both the company and its customers by ensuring that the organization's offerings align with customer preferences, thereby boosting retention rates. The ultimate goal is to increase customer loyalty and long-term profitability through more personalized interactions and enhanced service quality, making CRM an indispensable part of modern business strategy.

2.2 Customer Satisfaction and Retention

Customer satisfaction is a vital metric in assessing an organization's performance. The authors of [9] suggests that organizations must regularly evaluate their performance against strategic goals, with customer satisfaction being a significant indicator of success. Satisfied customers are more likely to return, recommend the company, and exhibit loyalty. However, it is crucial to recognize that customer satisfaction is not just about meeting current expectations but also about continuously exceeding those expectations to build strong, long-lasting relationships. Creating long-term relationships hinges on the understanding that acquiring new customers is more costly than maintaining existing ones. Studies indicate that the cost of retaining a customer is often far less than acquiring a new one [10]. This principle underscores the importance of customer retention as a strategic priority. When organizations focus on maintaining and enhancing their existing customer base, they secure not only short-term revenue but also long-term sustainability and growth. Thus, customer satisfaction directly contributes to retention, which is a cornerstone of CRM.

2.3 Churn

Churn, often described as customer attrition, refers to the loss of customers or clients over a given period. It is closely related to customer retention and can be defined as the number of customers who stop interacting with a company during a specific time frame [2]. According to [5], reducing churn or increasing retention significantly enhances the active customer base. By managing the customer lifecycle effectively—understanding, acquiring, satisfying, and retaining the most profitable customers—organizations can achieve better results.

The authors of [11] outline three essential actions to reduce churn:

- Measure Retention Rate: Organizations must define and assess their retention rate to gauge customer loyalty and satisfaction.
- Identify Causes of Attrition: Identifying the reasons behind customer churn allows companies to manage the factors that lead to attrition. These causes could include poor service, better competitor offers, or unmet customer expectations.
- Compare Customer Lifetime Value (CLV) with the Cost of Retention: Organizations need to balance the cost of retention efforts with the potential revenue generated from retaining a customer. If the cost of preventing churn exceeds the potential gains from retaining a customer, the organization must reassess its retention strategies.

Churn management is a critical focus for businesses aiming to retain their valuable customer base, especially in competitive markets where customer switching behavior is high. By identifying and addressing churn proactively, companies can mitigate its negative impact and enhance long-term profitability.

2.4 Machine Learning and Algorithms

Machine learning, a subset of artificial intelligence, focuses on creating systems capable of learning and improving from data without explicit programming. Machine learning techniques draw heavily from statistical and computational methods to build predictive models. The two primary objectives of machine learning are to predict future outcomes and to automate the modeling process using observed data [12].

Machine learning can be broadly classified into three main categories:

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

- Unsupervised Learning: In this category, the learning algorithm is not provided with any labeled data or predefined classes. It identifies patterns within the data by finding correlations and structures on its own.
- Supervised Learning: Supervised learning involves learning from labeled data where each training example is paired with an output label. The algorithm learns the mapping function from inputs to outputs, making predictions based on this learned relationship.
- Reinforcement Learning: This type of learning involves agents making decisions in an environment to maximize a cumulative reward. The agent explores the environment and learns by receiving feedback from its actions, adjusting its behavior to achieve the best possible outcomes over time.

In addition to these categories, machine learning algorithms are often classified based on how they handle incoming data:

- Non-Incremental Algorithms: These require all training data to be available upfront. Once the training process begins, no new data can be added without retraining the entire model.
- Incremental Algorithms: These algorithms can update their models incrementally as new data becomes available. This allows them to adjust to changing patterns over time without the need to rebuild the entire model from scratch.

When working with machine learning in real-world applications, data is rarely perfect. The presence of "noise" in the data—i.e., when examples with identical features belong to different classes—can lead to inaccuracies. Such data imperfections must be addressed carefully during the modeling process to avoid misleading predictions [13]. Below are brief descriptions of the algorithms commonly used in churn prediction:

2.4.1 Logistic Regression

Logistic regression is a statistical method used for binary classification problems, such as predicting customer churn (whether a customer will leave or not). It is one of the simplest and most widely used techniques for churn analysis. Logistic regression models the probability of an event occurring, based on one or more predictor variables, making it a valuable tool in business intelligence and churn prediction [14].

2.4.2 Decision Tree

Decision trees are a well-known and widely used technique in machine learning. They organize the information from a training dataset into a hierarchical structure of nodes and branches, where each node represents a decision or condition, and the branches represent outcomes. Decision trees are valued for their simplicity and interpretability, making them one of the most transparent algorithms for churn analysis. They can handle both numerical and categorical data, which makes them highly versatile [15].

2.4.3 Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and aggregates their results to improve prediction accuracy. The primary advantage of Random Forest over a single decision tree is its ability to reduce overfitting and provide more stable predictions. This algorithm is particularly useful for problems with complex, non-linear relationships between variables [13] [16].

2.4.4 Bagging Classifier

Bootstrap Aggregation (or Bagging) is a powerful ensemble method designed to reduce the variance of high-variance algorithms, such as decision trees. Bagging works by generating multiple random subsets of the training data, training a model on each subset, and then combining their predictions. This method improves the accuracy and robustness of the model, making it less likely to overfit the training data [17].

2.4.5 AdaBoost Classifier

AdaBoost (Adaptive Boosting) works by combining weak learners (simple models) to create a strong predictive model. It sequentially adjusts the weights of misclassified examples, making the classifier focus more on harder-to-predict instances. AdaBoost improves accuracy by giving more weight to misclassified data points, progressively refining the model's predictions [18].

2.4.6 XGBoosting Classifier

Extreme Gradient Boosting (XGBoost) is a highly effective and scalable machine learning algorithm. It is an extension of the gradient boosting framework and incorporates regularization techniques to prevent overfitting. XGBoost operates by building a sequence of decision trees, where each tree corrects the errors of the previous

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

one. It is particularly useful for handling large datasets and complex problems, making it an excellent choice for churn prediction tasks [19].

3 PROPOSED METHODOLOGY

The methodology for this study follows a descriptive research approach, which aims to systematically describe the characteristics of the selected population or phenomenon. As outlined by [20], descriptive research involves using standardized data collection methods, such as questionnaires or systematic observation, to capture the relevant information. This study is also a quantitative classification research, utilizing documentary collection and statistical analysis techniques. The data for this research will be sourced from a customer database of an Indian-based online learning platform (for the sake of confidentiality, we will refer to this platform as "EduTech India"). The dataset contains information on customer purchases, subscriptions, and behaviors, stored in the company's internal database over a span of several years.

After data collection, the steps of selection, coding, tabulation, and analysis will be performed using Python programming language, alongside relevant libraries such as Pandas, Numpy, Matplotlib, and Scikit-learn. The Jupyter Notebook environment will be used for coding, data visualization, and running the machine learning models. This methodology ensures that the process of data analysis, model development, and evaluation is transparent, reproducible, and efficient.

3.1 Data Set and Preliminary Selection of Variables

The dataset for this study will be derived from EduTech India's customer records, ensuring the privacy and confidentiality of users who have subscribed or used the platform's products. The dataset consists of 500,000 subscription observations (rows) from 200,000 individual users, each representing a customer interaction with the platform over time. The dataset includes several attributes that help to describe the customer's profile and their interactions with EduTech India, including:

- Customer demographics: Age, gender, education level
- Subscription details: Subscription type (e.g., basic, premium), renewal frequency
- Product usage patterns: Frequency of login, courses accessed, modules completed
- Payment details: Payment methods, transaction amounts, payment history
- Geographical location: Region or state of residence
- Support interaction: Number of support tickets raised, type of issues reported

These variables are selected based on their relevance to churn prediction, as supported by existing churn models found in the literature [3]. Given the wide range of available features, care will be taken to select those that are most likely to affect churn behavior in an Indian market context.

From the 500,000 observations, 110,000 (22%) are churned subscriptions, where customers have either voluntarily canceled their subscriptions or had their access suspended due to payment issues. These cancellations will be treated as the target variable in the churn prediction model.

Through an in-depth analysis of existing churn models, particularly in the context of online education platforms, 12 risk factors have been identified for churn prediction. These factors include:

- Frequency of course usage
- Payment history (delayed payments, defaults)
- Length of time subscribed
- Customer satisfaction (from support ticket sentiment analysis)
- Usage of premium features (if available)

The selection process ensured that all customer information used is non-sensitive and pertains only to the interactions with EduTech India's services. Moreover, any variables with high cardinality or complexity (e.g., unique session IDs or IP addresses) will be excluded, as they could increase computational cost and complexity unnecessarily.

The dataset will be divided into two parts: 75% for training and 25% for testing. The training set will undergo preprocessing, which includes handling imbalanced classes (explained below) and transforming categorical variables. The final dataset will contain the 12 chosen features, including the target variable "Churn."

Table 1 and Table 2 provides more information about numerical and categorical variables.

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

Table 1: Numerical Variables

Variable	Minimum	Median	Mean	Maximum	Standard Deviation	Description
Income	20,000	45,000	39,500	100,000	15,000	Customer's income
Login Frequency	1	5	6.2	50	8.5	Number of logins in the past month
Time Subscribed	0	1	1.5	5	1.2	Length of time customer has been subscribed (in years)
Total Spent	1000	5,000	3,800	20,000	6,000	Total amount spent on subscriptions

Table 2: Categorical Variables

Variable	Number of Levels	Frequency per Level	Description	Possible Values
Subscription Type	2	1: 300,000, 2: 200,000	Type of subscription (basic, premium)	1: Basic, 2: Premium
Region	4	1: 120,000, 2: 150,000, 3: 100,000, 4: 130,000	Customer's geographical region	1: North, 2: South, 3: East, 4: West
Payment Method	3	1: 100,000, 2: 300,000, 3: 100,000	Payment method chosen by customer	1: Credit Card, 2: Debit, 3: UPI
Churn	2	0: 390,000, 1: 110,000	Whether the customer canceled the subscription	0: No, 1: Yes

Data Preparation

3.2.1 Transformation of Categorical Variables

Categorical variables (e.g., Subscription type, Region, Payment method) will be transformed using one-hot encoding or label encoding based on their characteristics. One-hot encoding will be used for variables with nominal categories (no inherent order), such as Region or Payment method, converting them into binary variables, as explained in the following steps:

For example, for the categorical variable Payment Method with three categories:

- Credit Card
- Debit Card
- UPI

This will be transformed into three binary columns:

Table 3: Transformation of categorical variables

Payment Method - Credit Card	Payment Method – Debit Card	Payment Method - UPI
1	0	0
0	1	0
0	0	1

This process will be applied to other variables like Subscription Type and Region to make them suitable for machine learning models.

3.2.2 Class Balancing

The EduTech India dataset suffers from class imbalance, with churned customers accounting for only 22% of the observations, while the remaining 78% are non-churned. This imbalance can result in biased predictive models, where the model predicts the majority class (non-churn) more often than the minority class (churn).

To address this, we will use the SMOTE (Synthetic Minority Over-sampling Technique) [21] to balance the classes. SMOTE generates synthetic samples for the minority class (churn) by creating new instances based on the nearest neighbors of existing minority class samples.

The mathematical procedure for SMOTE is as follows:

- Select a sample from the minority class (churn).
- Find the k nearest neighbors to this sample.

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

Generate synthetic samples by choosing a random neighbor, then create new samples by linearly interpolating between the original sample and the selected neighbor. For a sample x_1 and its neighbor x_2, a new sample x_{new} is calculated as:

$$x_{new} = x_1 + \lambda \cdot (x_2 - x_1) \quad (1)$$

Where λ is a random number between 0 and 1.
This approach helps ensure that the machine learning model is exposed to a more balanced dataset, improving its ability to predict churn.

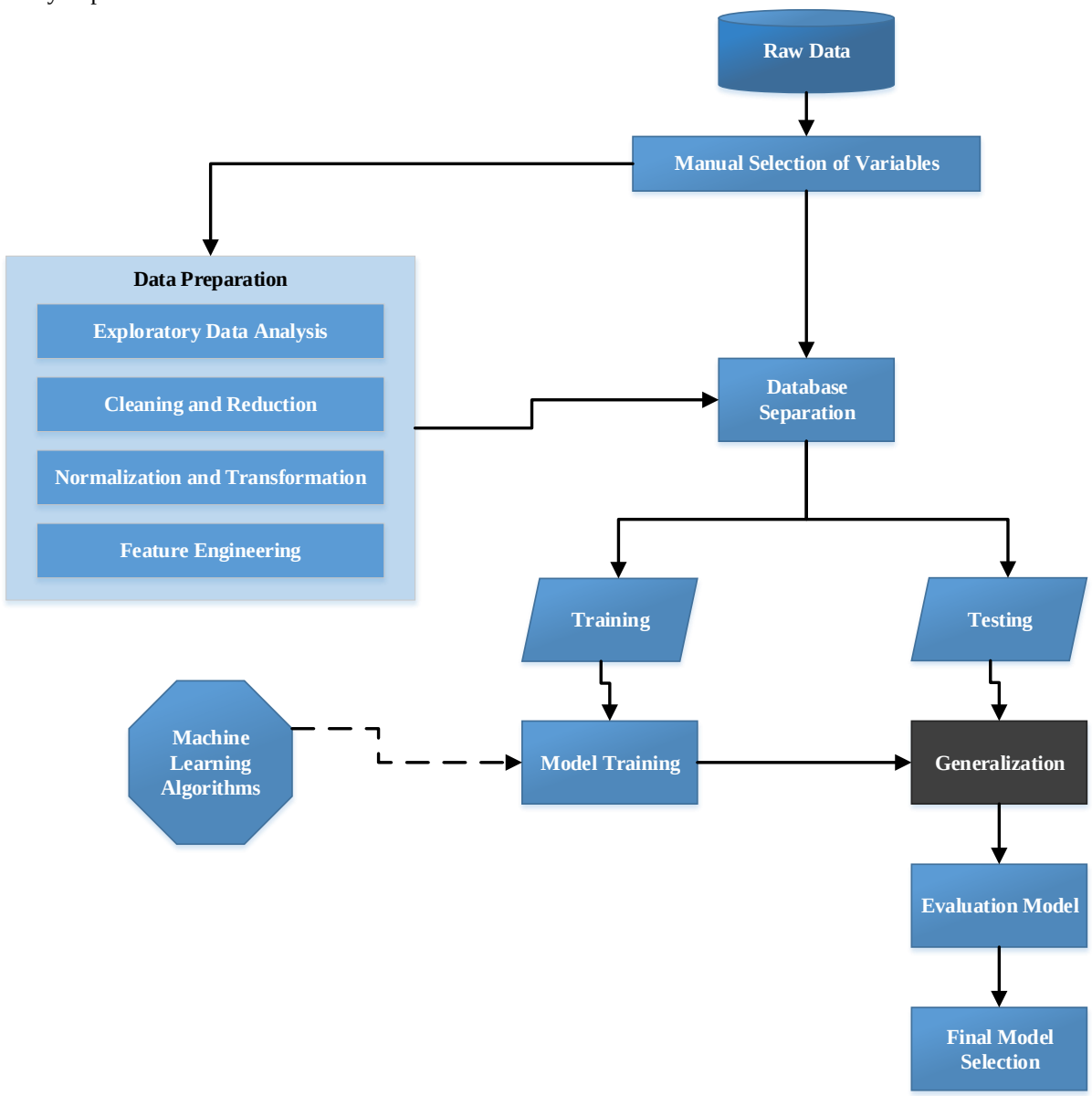

Figure 1: Machine learning model development and evaluation process

3.2.3 General Process of Model Construction and Evaluation

The predictive model construction will follow the steps outlined below:

- Training on Balanced and Unbalanced Datasets: The models will be trained both on the original (unbalanced) dataset and on the balanced dataset (using SMOTE) to analyze how class balancing impacts model performance.

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

- Generalization/Testing on Independent Dataset: After training, models will be tested on the separate 25% test set to evaluate their ability to generalize to new, unseen data. This will help assess the model's robustness and real-world applicability.
- Statistical Comparison of Algorithms: To compare the performance of the various models, accuracy will be the primary metric used. The formula for accuracy is:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Observations}} \quad (2)$$

  In addition to accuracy, other metrics such as precision, recall, and F1-score will also be evaluated to provide a comprehensive understanding of model performance.
- Evaluation of the Best Models: The final models will undergo in-depth evaluation, focusing on those that demonstrated the highest accuracy and generalization ability. These models will be assessed on additional metrics, including their confusion matrix, ROC curve, and AUC (Area Under the Curve) to further understand how well they perform across different thresholds.

The cross-validation technique with 10-folds will be used during the training process to minimize overfitting and ensure the results are consistent across different subsets of the data. The accuracy metric will be calculated for each fold, and the average accuracy will be reported to give a more stable evaluation of model performance.
To better understand the sequence of steps, the process of model construction, training, evaluation, and comparison is visually represented in Figure 1. The figure outlines the key phases, starting with the manual selection of variables, followed by data splitting into training and testing sets. The models undergo training with both balanced and unbalanced datasets and are then evaluated based on accuracy and other performance metrics. The final step involves in-depth evaluation of the best-performing models. This visual representation ensures clarity and provides a structured overview of the process flow.

4 RESULTS AND ANALYSIS

4.1 Dataset

The dataset used in this study consists of 500,000 observations from EduTech India, which includes 200,000 individual users who interacted with the platform between 2015 and 2020 [22]. These users' subscription activities, including purchases and cancellations, are tracked throughout their tenure on the platform. Automatic subscription renewals, recorded in the "Subscription Type" attribute of the dataset, account for the majority of customer interactions, comprising 53% of all subscriptions. Other key details regarding the sample and its distribution can be found in Table 1 and Table 2, which outline the breakdown of numerical and categorical variables, respectively.

4.2 Model Generalization Results

The results of this study are summarized in Table 4, following the methodology described in Section 3.2.3. For consistency, all models were trained on the same dataset split (75% training, 25% testing), ensuring that the models worked with identical information. Accuracy was used as the primary evaluation metric during model training and generalization. The models were first trained on the original unbalanced dataset and then on the balanced dataset using SMOTE to address class imbalance.
Overall, the models demonstrated satisfactory performance, with an average accuracy of over 80%. Specifically, models trained on the unbalanced dataset achieved an average accuracy of 87.11%, while models trained on the balanced dataset had a lower average accuracy of 74.0%. The best models selected for detailed comparison were XGBoost (trained on the unbalanced dataset with an average accuracy of 87.11%) and Bagging Classifier (trained on the balanced dataset with an average accuracy of 77.34%).
The performance results for each model are summarized in Table 4 below, showing the comparison between models trained on balanced and unbalanced datasets. The table presents the mean accuracy and standard deviation for each algorithm across both training conditions.

Table 4: Model Results

Algorithm	Class Balancing	Mean Accuracy	Standard Deviation
XGBoost (XGB)	No (Unbalanced)	87.11%	0.014
XGBoost (XGB)	Yes (Balanced)	72.42%	0.098
Logistic Regression (LR)	No (Unbalanced)	87.07%	0.013
Logistic Regression (LR)	Yes (Balanced)	70.38%	0.100
AdaBoost (ADA)	No (Unbalanced)	87.03%	0.013

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

AdaBoost (ADA)	Yes (Balanced)	70.36%	0.099
Random Forest (RFC)	No (Unbalanced)	85.94%	0.016
Random Forest (RFC)	Yes (Balanced)	76.88%	0.029
Bagging Classifier (BAGG)	No (Unbalanced)	85.75%	0.016
Bagging Classifier (BAGG)	Yes (Balanced)	77.34%	0.024
Decision Tree (CART)	No (Unbalanced)	84.76%	0.017
Decision Tree (CART)	Yes (Balanced)	76.64%	0.023

Figures 2 and 3 show the comparison of the prediction accuracies of each algorithm used in this study.
From the results, it is evident that there is greater variation in the prediction accuracies when using balanced classes, especially for Logistic Regression, AdaBoost, and XGBoost algorithms. This indicates that balancing the dataset can significantly impact the stability and performance of the models. However, a deeper evaluation using other metrics, beyond accuracy, is required to fully understand the model performance and its ability to generalize to new observations. The next section provides a detailed comparison of the algorithms that achieved the best results in the first evaluation.

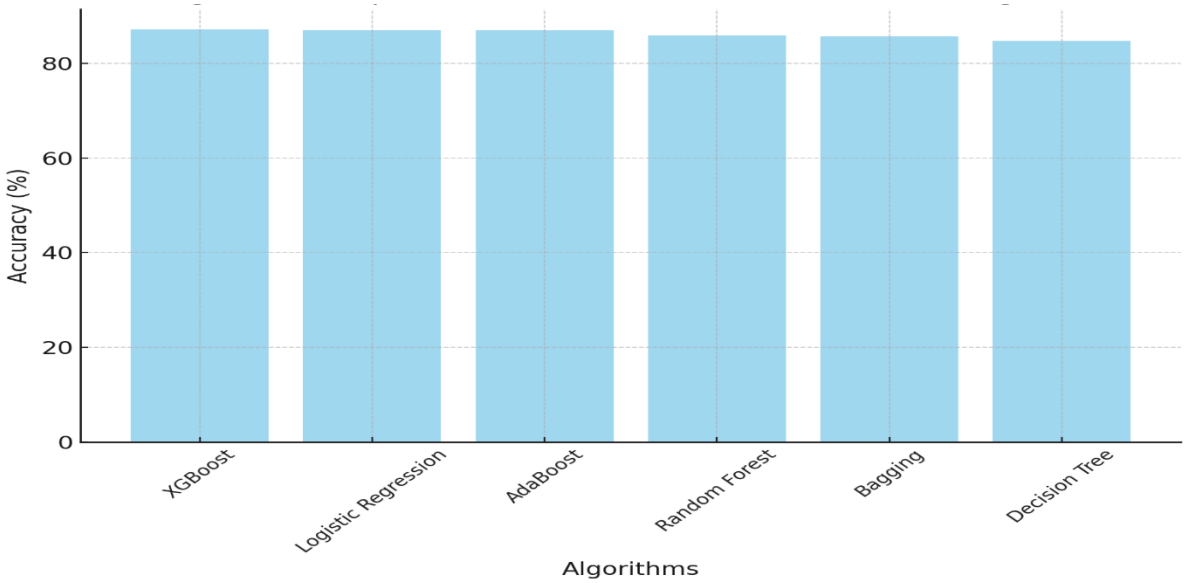

Figure 2: Comparison of Models with Unbalanced Training Data

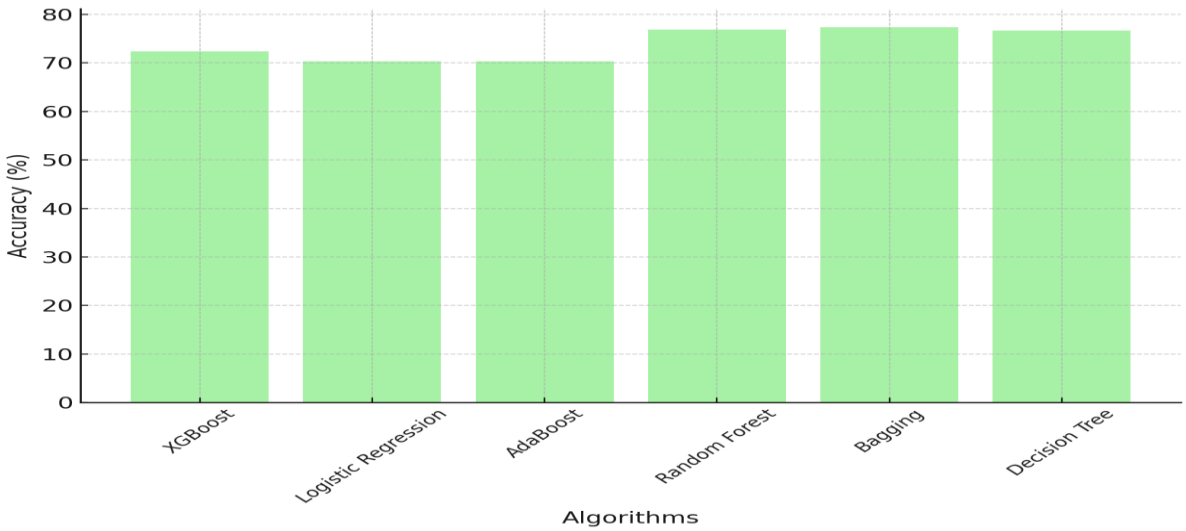

Figure 3: Comparison of Models with Balanced Training Data

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

4.3 Evaluating the Best Performing Models

The Area Under the ROC Curve (AUC) is an important performance metric for binary classification problems. AUC represents the model's ability to discriminate between positive and negative classes. An AUC of 1.0 represents perfect classification, while an AUC of 0.5 indicates random performance. The ROC curve is composed of two components:

- Sensitivity (True Positive Rate / Recall): The proportion of actual positive instances correctly predicted by the model.
- Specificity (True Negative Rate): The proportion of actual negative instances correctly predicted by the model.

Figures 4 and 5 compare the performance of the XGBoost algorithm, trained on unbalanced data, and the Bagging Classifier, trained on balanced data. From the graphs, it is clear that XGBoost outperforms Bagging, as indicated by the higher AUC value. Specifically:

- The AUC for XGBoost was 0.86, while for Bagging, it was 0.83.

Both algorithms showed relatively similar performance, but XGBoost proved to be more effective in predicting churn. It not only achieved the highest accuracy among all models but also had the highest AUC and demonstrated greater flexibility when working with imbalanced classes. This flexibility is important for real-world datasets, where class imbalance is often a challenge.

Although both Bagging and XGBoost performed well, XGBoost emerged as the superior model due to its higher accuracy, larger AUC, and better handling of imbalanced classes. Therefore, XGBoost is considered the optimal model for churn prediction in the context of EduTech India.

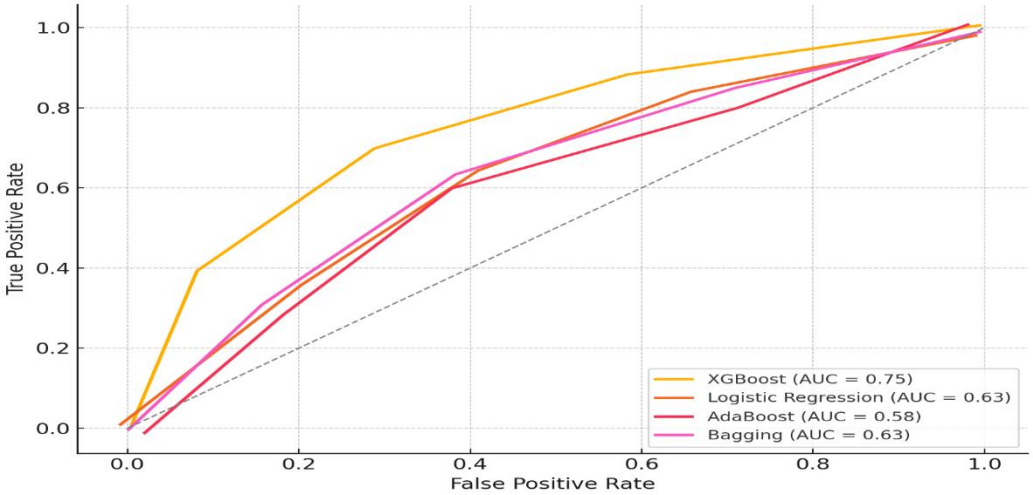

Figure 4: ROC Curve for XGBoost with Unbalanced Training Data

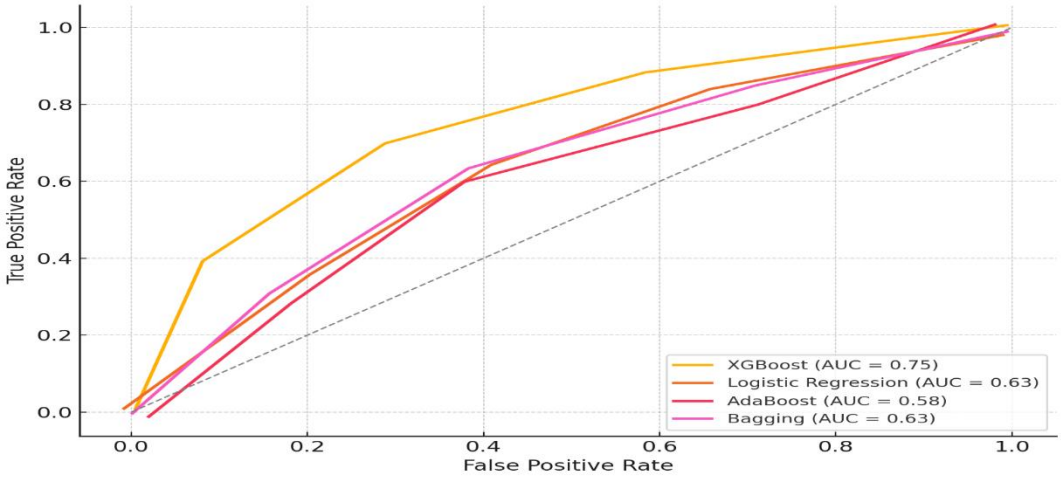

Figure 5: ROC Curve for Bagging Classifier with Balanced Training Data

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

5 CONCLUSION

The comparison between different algorithm families showed that all of them produced relatively similar results when using the same datasets, indicating the feasibility of churn prediction. The final predictive model for churn forecasting was XGBoost, which performed better without class balancing. However, all algorithms demonstrated an average accuracy of over 85%, based on at least one training dataset, suggesting that with more advanced preprocessing techniques and fine-tuning of hyperparameters, it is possible to implement a high-performing model for predicting churn in an online education company. In conclusion, the objectives of the research were successfully achieved with the data presented and analyzed. Given the time constraints, it was not possible to utilize all available resources for optimizing each applied model. Therefore, as a suggestion for future research, we propose parameter adjustments for each algorithm, along with a more robust transformation and treatment of the data, to improve and optimize predictions further.

REFERENCES

1. Lemmens, A. and Gupta, S., 2020. Managing churn to maximize profits. Marketing Science, 39(5), pp.956-973.
2. Khodabandehlou, S. and Zivari Rahman, M., 2017. Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. Journal of Systems and Information Technology, 19(1/2), pp.65-93.
3. Haenlein, M., 2013. Social interactions in customer churn decisions: The impact of relationship directionality. International Journal of Research in Marketing, 30(3), pp.236-248.
4. Khadka, K. and Maharjan, S., 2017. Customer satisfaction and customer loyalty. Centria University of Applied Sciences Pietarsaari, 1(10), pp.58-64.
5. Buttle, F. and Maklan, S., 2019. Customer relationship management: concepts and technologies. Routledge.
6. Akhila, V., Krithikaa, M., Pavithra, A.K., Adhipathy, K.G. and Pamina, J., 2019. Analysing the behaviour of customers to predict churn in telecom sector. International Journal of Emerging Technology and Innovative Engineering, 5.
7. Ullah, I., Raza, B., Malik, A.K., Imran, M., Islam, S.U. and Kim, S.W., 2019. A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. IEEE access, 7, pp.60134-60149.
8. Kumar, V. and Reinartz, W., 2018. Customer relationship management. Springer-Verlag GmbH Germany, part of Springer Nature 2006, 2012, 2018.
9. Rossomme, J., 2018. Customer satisfaction measurement in a business-to-business context: a conceptual framework. Journal of business & Industrial marketing, 18(2), pp.179-195.
10. Ascarza, E., Neslin, S.A., Netzer, O., Anderson, Z., Fader, P.S., Gupta, S., Hardie, B.G., Lemmens, A., Libai, B., Neal, D. and Provost, F., 2018. In pursuit of enhanced customer retention management: Review, key issues, and future directions. Customer Needs and Solutions, 5, pp.65-81.
11. Deepak, R.K.A. and Jeyakumar, S., 2019. Marketing management. Educreation Publishing.
12. Mullainathan, S. and Spiess, J., 2017. Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), pp.87-106.
13. Rigatti, S.J., 2017. Random forest. Journal of Insurance Medicine, 47(1), pp.31-39.
14. Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G. and Chatzisavvas, K.C., 2015. A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, pp.1-9.
15. El Naqa, I. and Murphy, M.J., 2015. What is machine learning? (pp. 3-11). Springer International Publishing.
16. Breiman, L., 2001. Random forests. Machine learning, 45, pp.5-32.
17. Zareapoor, M. and Shamsolmoali, P., 2015. Application of credit card fraud detection: Based on bagging ensemble classifier. Procedia computer science, 48(2015), pp.679-685.
18. An, T.K. and Kim, M.H., 2010, October. A new diverse AdaBoost classifier. In 2010 International conference on artificial intelligence and computational intelligence (Vol. 1, pp. 359-363). IEEE.
19. Quinto, B., 2020. Next-generation machine learning with spark: Covers XGBoost, LightGBM, Spark NLP, distributed deep learning with keras, and more. Apress.
20. Bryman, A., 2016. Social research methods. Oxford university press.

ISSN 2349-6193
Impact Factor: 2.805

IJESMR
International Journal OF Engineering Sciences & Management Research

21. Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, pp.321-357.
22. EduTech India Dataset. Available online at: https://www.kaggle.com/datasets/akshatsharma0610/edtech-dataset