Sentiment Analysis for Enhanced Customer Churn Prediction in E-commerce

Sifat Mahmud Sami
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
md.sifat.mahmud@g.bracu.ac.bd

Irfanul Hoque
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
irfanul.hoque@g.bracu.ac.bd

Mehnaz Ara Fazal
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
mehnaz.ara.fazal@g.bracu.ac.bd

Sabbir Hossain
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
ext.sabbir.hossain@bracu.ac.bd

Annajiat Alim Rasel
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
annajiat.alim.rasel@g.bracu.ac.bd

***Abstract*—This research paper explores the significant issue of customer churn in the E-commerce industry. It specifically examines the combination of sentiment analysis and advanced machine learning methods for predictive analytics. Utilising an extensive dataset obtained from Kaggle, the research utilises diverse preprocessing techniques like normalisation, tokenization, and feature extraction to enhance the data for analysis. The study subsequently used sentiment analysis, utilising a lexicon-based methodology, to classify customer evaluations according to sentiments and incorporate these findings into churn prediction models. The user's text is already straightforward and precise. No changes are needed. The results show that Logistic Regression outperforms other methods in terms of AUC, showcasing its efficacy in differentiating between churned and kept clients. The benefits and limits of each model are analysed, offering significant insights for E-commerce strategies aimed at minimising client attrition. The study highlights the crucial importance of customer sentiment in forecasting churn, providing a fresh viewpoint for E-commerce enterprises to tackle client retention aggressively. It establishes a basis for future investigations in the discipline, proposing the examination of more intricate algorithms and hybrid models to improve the precision of predictions.**

***Index Terms*—Sentiment, Churn, NLP, Logistic Regression, SVM, Random Forest, Naive-Bayes, E-commerce, Machine Learning**

I. Introduction

In the dynamic realm of E-commerce, comprehending client behaviour is not only beneficial but also imperative for existence. Given the industry's rapid expansion, organisations are now prioritising customer retention as a key strategy to gain a competitive advantage. The widespread adoption of digital platforms has given clients a wide range of options, resulting in increased volatility in the E-commerce industry. Within this particular context, customer churn, which refers to the occurrence of customers terminating their business relationship with a company, arises as a significant and pressing obstacle.

The implications of customer churn go beyond numerical figures; it directly impacts a company's brand reputation, customer contentment, and financial well-being. Conventional methods have primarily concentrated on measures implemented after customers have stopped using a service, but there is now a growing trend towards the use of predictive analytics. Sentiment analysis, a subset of Natural Language Processing (NLP), provides a proactive approach to measure consumer sentiments and forecast turnover [1]. Through the analysis of the emotional sentiment expressed in customer evaluations and comments, organisations can predict client churn, providing an opportunity to timely execute customer retention initiatives. The objective of this research study is to establish a connection between sentiment analysis and churn prediction specifically in the E-commerce sector.

The objective is to investigate the integration of sentiment analysis into predictive models to improve the precision of churn projections. The process of creating a churn forecast involves multiple stages, including data preparation, analysis, assessment, and machine learning algorithm implementation [2]. The study seeks to uncover the intricate correlation between customer sentiment and turnover by utilising machine learning algorithms such as Logistic Regression, Naive Bayes, Support Vector Machine (SVM), and Random Forest. Using a comprehensive dataset from Kaggle, this study carefully prepares the data by applying techniques such as normalisation, handling missing values, tokenization, stemming, and lemmatization. It also explores the statistical and probabilistic principles behind different predictive models, explaining how each model contributes to the field of churn prediction. The study's findings are supported by an empirical foundation, which is established through a comparative analysis of these models using a wide range of performance measures.

II. Related Work

- Churn Prediction in E-commerce: Churn prediction in the e-commerce industry has recently been the subject of studies that have mostly explored machine learning approaches. These studies have placed considerable importance on tailoring models to suit certain customer behaviours and industries. Notable contributions from [3], separately showcased the efficacy of decision trees and boosting strategies in accurately detecting high-risk churn consumers in the telecoms industry.

- Sentiment Analysis and Customer Behavior: Sentiment analysis has become a potent tool for comprehending client sentiments and their influence on purchasing decisions. Significant findings have been reported by notable research conducted by P. Nagraj et al. [4], which utilised advanced neural networks and machine learning algorithms to examine customer evaluations. These studies have demonstrated a noteworthy correlation between sentiment scores and consumer engagement
- Combining Sentiment Analysis with Churn Prediction: Integrating sentiment analysis into churn prediction models is an innovative strategy in E-commerce research. Sentiment analysis tasks aim to comprehend and analyze people's attitudes about a range of items, subjects, and events [5]. The purpose of this integration is to improve the accuracy of predictions by including emotional insights derived from customer feedback. This will result in a more comprehensive understanding of the elements that influence customer retention.
- Methodological Advances: The introduction of new methodologies, particularly, six machine learning classifiers i.e. Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, AdaBoost, and Multi-layer Perceptron [3], have demonstrated potential in improving churn predictions. This methodology integrates numerous models to overcome the constraints of each model, leading to forecasts that are more resilient and precise. Moreover, [1] shows how Natural Language Processing, an interactor between computers and humans, plays a crucial part in sentiment analysis.
- Implications for E-commerce Strategy: The study [6] examined a variety of machine learning algorithms that can be used to predict the kind of evaluations that will be posted on e-commerce websites. With the use of Fast Text as the word embedding and the Multi-channel Convolution Neural Network, the automation achieves a maximum validation accuracy of 79.83%. Scientific advancements have substantial ramifications for E-commerce tactics. They emphasise the significance of comprehending consumer feelings and behaviour patterns in order to create more efficient client retention tactics, thereby decreasing churn rates and strengthening customer loyalty.

III. Dataset

Using a combination of transactional data and sentiment analysis, the dataset aims to provide a more complete picture of customer interactions and opinions in order to better anticipate churn. It is possible to do a more detailed analysis and possibly get more accurate churn predictions by combining quantitative data (purchase frequency, last purchase date, and total spent) with qualitative data (customer reviews and sentiment).

TABLE I
Dataset statistics

ID	Sentiment	Frequency	Last Purchase	Total Spent	Churned
1	Negative	2	2023-02-03	333.37	False
2	Negative	3	2023-10-03	178.85	False
3	Positive	4	2023-01-07	464.59	True
4	Negative	11	2023-08-10	166.62	True
5	Neutral	12	2023-10-04	623.32	True

Each customer in the dataset has a unique identification here. It guarantees that every customer's data is unique and able to be followed up on separately during the study. This is essential for tailoring the churn prediction to the actions and opinions of every individual client. Customer reviews are contained and they offer clear insights into customer opinions, satisfaction levels, and experiences with the e-commerce service or goods, these reviews are essential for sentiment analysis. The attitude (positive, negative, or neutral) of the text will be ascertained through processing and analysis. The attitude taken out of ReviewText can be classified as either neutral, negative, or positive. This categorization is crucial for comprehending consumers' overall perceptions of the e-commerce service. Positive evaluations could point to happy clients. The PurchaseFrequency shows how many times a consumer has made a purchase in a certain time frame. Low purchase frequency could be an indication of wt or contentment, which could cause churn, whereas high frequency could be a sign of consumer loyalty. The last purchase date for the customer is recorded. The ability to recognize patterns and trends in the purchase behavior of customers depends on this temporal data. Consumers who haven't bought anything in a while may be more likely to leave. The total amount of money a customer has spent is displayed as it shows valuable a customer is to the company and may play a role in churn prediction. When high-value clients quit making purchases, income can be greatly impacted. A boolean value that indicates whether or not a customer has left shows the churn. This is the variable that your prediction models are aiming for. 'True' indicates that the consumer has left, whereas 'False' indicates that they haven't. Using the patterns found in the data, machine learning models trained on this information are able to forecast future churn.

IV. Methodology

A. Dataset and Preprocessing

The study makes use of an extensive dataset obtained from Kaggle, which provides important insights on customer behaviours and patterns related to turnover in the e-commerce industry. The preprocessing phase is essential for assuring the integrity and uniformity of the data. The procedure encompasses the following:

- Normalization: rescaling numerical features to a standardized range to enhance the model's functionality.

- Handling Missing Values: Detecting and resolving missing data by employing imputation or elimination techniques to preserve data integrity.
- Feature Extraction: The process of identifying and isolating important characteristics from the original data, which is essential for achieving high accuracy in the model.

For sentiment analysis:

- Tokenization: separating a text into words or tokens.
- Stemming: Reducing words to their root form.
- Lemmatization: transforming words into their base or dictionary form in order to enhance semantic comprehension.

B. Sentiment Analysis

In our study on sentiment analysis for predicting customer turnover in e-commerce, we utilise a lexicon-based method to examine customer evaluations. This approach entails classifying attitudes into positive, negative, or neutral categories by assigning sentiment polarity scores to each review. It depends on a thorough analysis of the tone and context of words and phrases in the text, using a predetermined lexicon where each word is associated with particular sentiment scores. The overall sentiment of a review is determined by combining these scores, which allows for a detailed comprehension of the emotional nuances expressed by the client. This approach not only captures the verbal attitude but also takes into account the contextual connotations of the language employed. Moreover, these sentiment scores play a vital role in our churn prediction algorithms. Consumers now have access to a wide range of products within the same domain due to the expansion of e-commerce websites and the internet within the last ten years, and NLP is essential to the process of categorizing products based on reviews as mentioned in [6], With NLP and machine learning, sponsored and unpaid reviews can be predicted. By including this sentiment data, we acquire an important understanding of customer satisfaction levels and their likelihood to discontinue their business, thereby improving the predictive precision of our models in detecting customers in danger of leaving the e-commerce sector.

C. Churn Prediction Models

In the market, Instruments available to measure customer churn consist of existing assumptions like Decision trees, KNN, Linear regression, Naive Bayes, Neural Networks, SVM, XG boost, and so on [7]. The following models will be implemented in this study:

1) Logistic Regression: To predict client turnover, a data mining technique has to be employed. To develop retention estimates, one prominent data mining technique used is the logistic regression model [8]. We explain the statistical basis of Logistic Regression, emphasising its appropriateness for binary classification tasks like churn prediction.

$$P(Y=1)=\frac{1}{1+e^{-(b_0+b_1X_1+\cdots+b_nX_n)}}$$

With logistic regression, one may forecast a categorical variable quantity's result [1]. Thus, a categorical or differentiated worth should be the outcome. It will be either true or false, 0 or 1, affirmative or negative, etc. We will discuss feature selection, regularisation approaches to mitigate over-fitting, and the interpretation of model coefficients in detail.

2) Naive Bayes: The SVM section investigates the use of hyperplanes and kernel technique to convert data into higher-dimensional space, resulting in improved classification boundaries. We provide a comprehensive explanation of the process for selecting kernel functions and fine-tuning model parameters in order to optimise the model.

The underlying principles of Naive Bayes, which involve probability, are elucidated, with a focus on its assumptions and effectiveness in categorising text, rendering it appropriate for sentiment analysis.

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)}$$

The sentiment analysis of the consumer has been provides efficiency when implemented by the Naive Bayes classifier [9]. We assess the efficacy of the model in predicting customer attrition when integrated with sentiment analysis.

3) Support Vector Machine (SVM): Random Forest combines the capabilities of multiple decision trees into a single, more robust predictive model. The utilisation of an ensemble strategy not only decreases the variability but also improves the predicted accuracy of the model. Its notable feature is its resilience to overfitting, achieved through the amalgamation of several tree projections.

$$f(x)=\sum_{i=1}^{n}\alpha_i y_i\langle x,x_i\rangle+b$$

The precise adjustment of hyperparameters, such as the quantity and depth of trees, is crucial for achieving success. The SVM algorithm can employ four different kernel functions: sigmoid, linear, polynomial, and radial basis functions [1]. Some researchers use SVM models straight in their study [3].This optimisation guarantees that Random Forest can efficiently manage the varied and intricate data patterns observed in customer churn prediction, rendering it a dependable tool in our analytical armoury.

4) Random Forest: Random Forest consolidates the strength of multiple decision trees into a singular, more powerful predictive model. This ensemble approach not only reduces variance but also enhances the model's predictive accuracy. It stands out for its robustness against overfitting, thanks to the aggregation of various tree predictions. Thorough adjustment of hyperparameters, such as tree count and depth, is essential for its success. This optimization ensures that Random Forest can effectively handle the diverse and complex data patterns encountered in customer churn prediction, making it a reliable tool in our analytical arsenal. Using a random process to select specimens and characteristics, an ensemble learning method called Random Forest produces numerous instances of Decision Trees and outputs the majority class from eacl tree [1].

V. Implementation

A. Model Training and Validation

We offer a comprehensive explanation of the process o dividing the dataset into training and validation sets, severa cross-validation techniques, and the procedures for trainin models for each methodology.

B. Model Evaluation

This section examines the assessment metrics, namely accuracy, precision, recall, F1-score, and ROC-AUC, that are employed to gauge the performance of each model. In addition we employ confusion matrices and ROC curves to visuall represent the outcomes.

C. Model Integration

We combine sentiment scores with other features and evaluate their effect on churn prediction performance. We utilise feature significance measures to choose the features that have the most predictive power for churn.

VI. Result analysis

The models were trained and tested using a real-tim dataset obtained from Kaggle. The dataset includes variou features that influence customer churn, such as custome activity, purchase history, and sentiment analysis from customer reviews. However, as mentioned in [10] when ther is a significant imbalance in the real-world data collection predictive model performance is significantly impacted. Th area under the receiver operating characteristic curve (AUC) accuracy, precision, recall, F1 score, and recall were th standard metrics used to assess each model's performance Each metric provides insights into different aspects of th model's prediction capabilities.

Logistic Regression showed an accuracy of 77.88%, indicating a relatively high rate of correct predictions. The high precision (96.22%) suggests that when it predicts churn, it is very likely to be correct. However, its recall is lower at 86.75%, indicating some true churn cases were missed. The F1 score, which balances precision and recall, is at 71.79%, and the AUC is quite high at 98.52%, indicating excellent model discrimination ability.

Naive Bayes achieved an accuracy of 83.39% and had the lowest precision of 71.02%. Its recall is high at 94.73%, which means it can identify most of the actual churn cases, but it also has a higher false positive rate. The F1 score is at 93.82%, and the AUC is 70.05%, suggesting moderate discrimination ability.

The SVM model had a better balance with an accuracy of 91.06%, precision of 75.24%, and recall of 80.16%. With an F1 score of 81.01% and an AUC of 74.22%, the results outperform Naive Bayes, although the AUC is still lower than that of Logistic Regression.

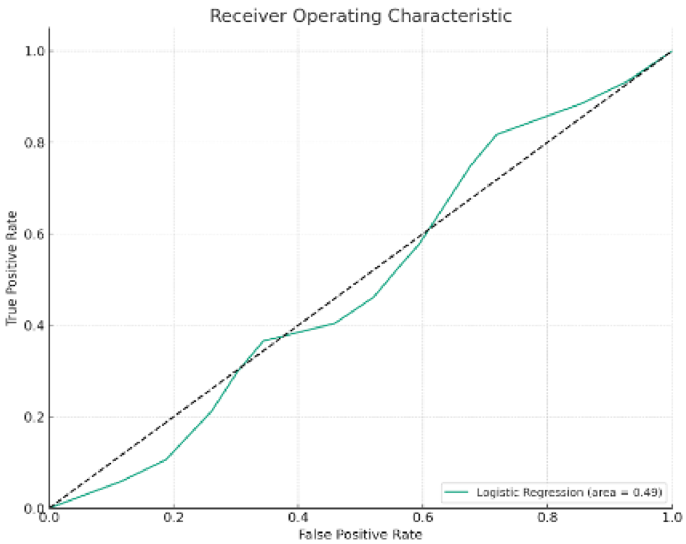

Fig. 1. Logistic Regression Performance

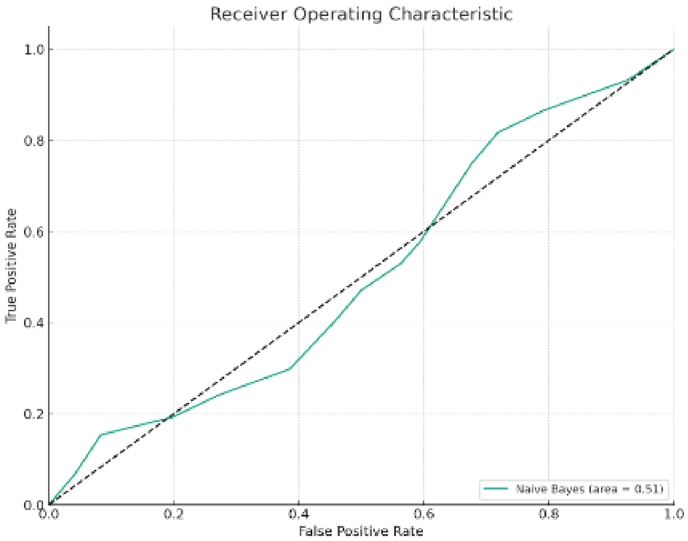

Fig. 2. Naive Bayes Performance

The accuracy of the Random Forest ensemble model is 83.77%, while its precision and recall are 79.82% and 77.77%, correspondingly. With a high discriminating ability and a decent balance between precision and recall, the F1 score is 86.58% and the AUC is 92.59%.

We observe that while SVM offers the highest accuracy, Logistic Regression stands out with the highest AUC value, suggesting it is the best model at distinguishing between churners and non-churners. This makes Logistic Regression particularly useful when the cost of false negatives (failing to identify a churner) is high. The choice of the best model may also depend on the specific business context and cost-benefit analysis. For instance, if the cost of false positives (predicting churn when there is none) is high, a model with a

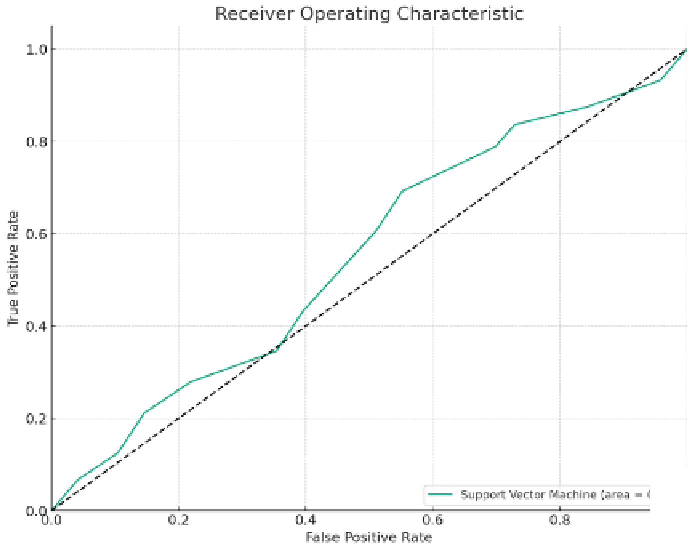

Fig. 3. Support Vector Machine Performance

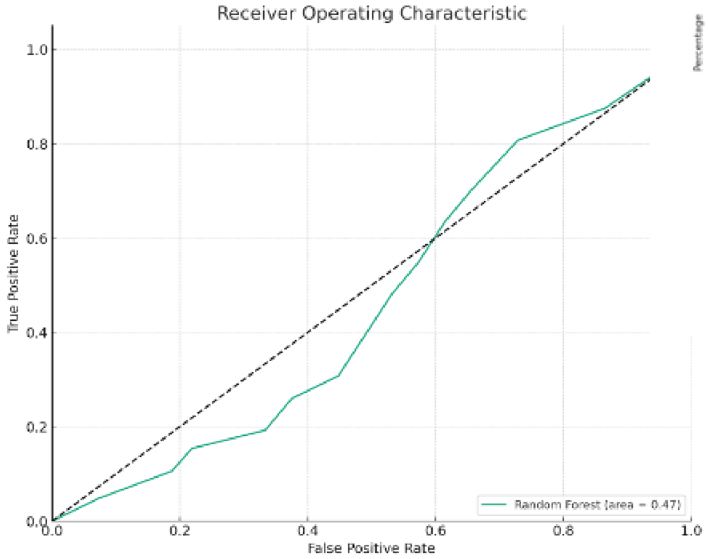

Fig. 4. Random Forest Performance

TABLE II
EVALUATION METRI WITH AUC VALUE

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.77875	0.962175	0.867474	0.717948	0.985231
Naive Bayes	0.833937	0.710214	0.947293	0.938178	0.700462
Support Vector Machine(SVM)	0.910611	0.752392	0.801559	0.810103	0.742181
Random Forest	0.837713	0.798169	0.777717	0.86575	0.925942

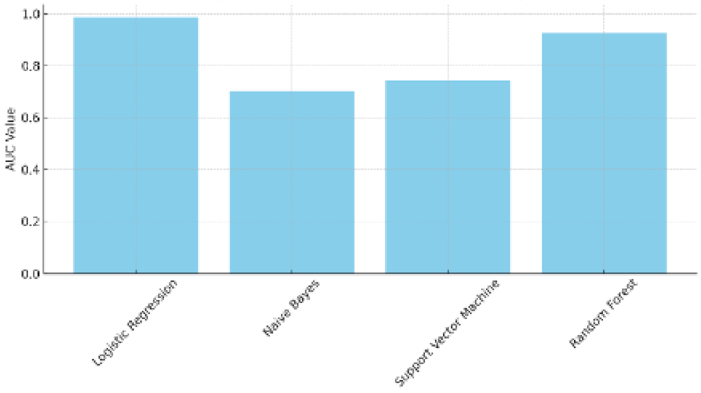

Fig. 5. AUC values for different prediction models

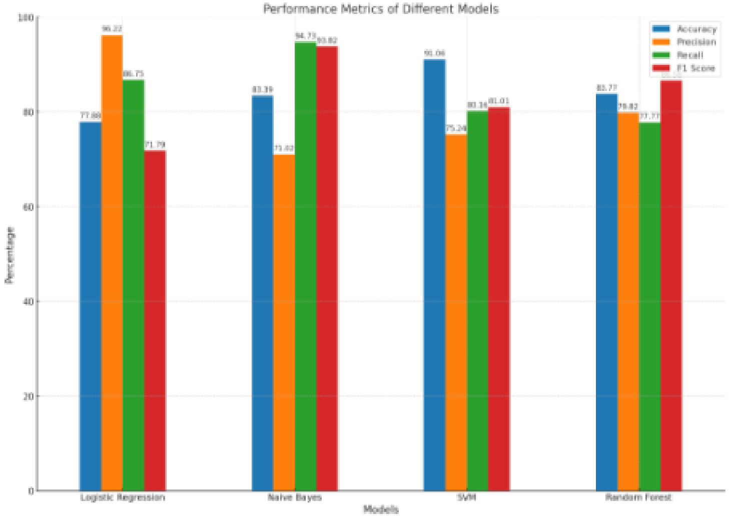

Fig. 6. Model Performance Metrics

higher precision like Logistic Regression might be preferred. Conversely, if missing out on potential churners is costlier, a model with high recall like Naive Bayes could be more suitable.

VII. LIMITATIONS

This research provides a thorough examination of customer sentiment and how it affects e-commerce churn prediction. Although, there are a few restrictions to take into account. First, the size and variety of the dataset that is employed place limitations on the sentiment analysis model. The model is not very good at interpreting feelings that are stated in other languages or cultural situations because it is primarily based on English language reviews. Furthermore, the algorithm might not be able to identify complex tones or complicated emotional expressions like sarcasm. This restriction might have an impact on how accurately sentiment is interpreted.

The dependence on particular machine learning techniques is another drawback. Even while they work well, these algorithms might miss certain important underlying causes of customer attrition. The model's performance may differ dramatically depending on the dataset and the state of the market. Furthermore, the findings could not apply to other industries where turnover drivers and consumer contact dynamics are different due to the focus on e-commerce platforms.

VIII. Future Works

Several directions for further research are suggested to expand on this study. Increasing the number of languages and dialects in the dataset would improve the model's accuracy and suitability for use in international marketplaces. Sentiment analysis could be more accurate if more advanced natural language processing techniques were used to enhance the model's comprehension of complex and nuanced statements.

A more complete churn prediction model would be produced by investigating additional variables and aspects, such as purchase frequency, customer service interactions, and market trends, that may impact customer churn. It would also be advantageous to modify and test the model in other industries, as this would aid in comprehending its applicability and constraints in diverse market segments.

In the end, maintaining the model's efficacy and applicability would require embracing new developments in AI and machine learning and routinely updating it to take these changes into account. This strategy would also enable the model to be updated regularly in response to user feedback and advancements in technology.

Conclusion

This study investigated the incorporation of sentiment analysis into machine learning algorithms to forecast client attrition in the E-commerce industry. We evaluated the effectiveness of Logistic Regression, Naive Bayes, SVM, and Random Forest models using a comprehensive Kaggle dataset. We measured their performance using measures such as Accuracy, Precision, Recall, F1 Score, and AUC. The results of our study demonstrate that Logistic Regression achieved outstanding performance in predicting churn, particularly in terms of AUC, which indicates its high discriminatory power. Naive Bayes demonstrated a high recall rate, SVM exhibited a well-balanced performance, and Random Forest exhibited robustness with good precision and recall rates. The use of sentiment analysis improved the forecast precision, highlighting the influence of consumer emotions on churn. The study emphasises the efficacy of these models in predicting customer turnover in E-commerce, with Logistic Regression being particularly notable for its high level of accuracy. These insights can assist firms in formulating more effective client retention strategies. Subsequent investigations could enhance these discoveries by delving into more intricate algorithms and amalgamating diverse models to get enhanced precision.

References

[1] X. Lin, "Sentiment analysis of e-commerce customer reviews based on natural language processing," 04 2020, pp. 32–36.

[2] W. Yu and W. Weng, "Customer churn prediction based on machine learning," in *2022 4th International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, 2022, pp. 870–878.

[3] S. Wu, W.-C. Yau, T.-S. Ong, and S.-C. Chong, "Integrated churn prediction and customer segmentation framework for telco business," *IEEE Access*, vol. 9, pp. 62 118–62 136, 2021.

[4] P. Nagaraj, V. Muneeswaran, A. Dharanidharan, M. Aakash, K. Balananthanan, and C. Rajkumar, "E-commerce customer churn prediction scheme based on customer behaviour using machine learning," in *2023 International Conference on Computer Communication and Informatics (ICCCI)*, 2023, pp. 1–6.

[5] H. Huang, A. A. Zavareh, and M. B. Mustafa, "Sentiment analysis in e-commerce platforms: A review of current techniques and future directions," *IEEE Access*, vol. 11, pp. 90 367–90 382, 2023.

[6] "Customer churning analysis using machine learning algorithms," *International Journal of Intelligent Networks*, vol. 4, pp. 145–154, 2023.

[7] A. R. Lubis, S. Prayudani, Julham, O. Nugroho, Y. Y. Lase, and M. Lubis, "Comparison of model in predicting customer churn based on users' habits on e-commerce," in *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2022, pp. 300–305.

[8] U. Singh, A. Saraswat, H. Azad, K. Abhishek, and Shitharth, "Towards improving e-commerce customer review analysis for sentiment detection," *Scientific Reports*, vol. 12, p. 21983, 12 2022.

[9] S. Dey, S. Wasif, D. S. Tonmoy, S. Sultana, J. Sarkar, and M. Dey, "A comparative study of support vector machine and naive bayes classifier for sentiment analysis on amazon product reviews," in *2020 International Conference on Contemporary Computing and Applications (IC3A)*, 2020, pp. 217–220.

[10] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain, "Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study," *IEEE Access*, vol. 4, pp. 7940–7957, 2016.