

Comparative Analysis of Classification Models for Customer Churn Prediction in E-Commerce

*Report submitted to the SASTRA Deemed to be University as
the requirement for the course*

MAT499: PROJECT PHASE - I

Submitted by

NAME: ABIRAMI.S

Reg.No:125150004

November 2024



SCHOOL OF ARTS, SCIENCES, HUMANITIES AND EDUCATION

THANJAVUR, TAMIL NADU, INDIA – 613 401



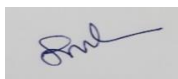
SCHOOL OF ARTS, SCIENCES, HUMANITIES AND EDUCATION

THANJAVUR – 613 401

Bonafide Certificate

This is to certify that the report titled “**Comparative Analysis of Classification Models for Customer Churn Prediction in E-Commerce**” submitted as a requirement for the course **MAT499: PROJECT PHASE - I** for M.Sc. Data Science programme, is a bona fide record of the work done by **MS ABIRAMI S, Reg. No:125150004** during the academic year 2023-2024, in the School of Arts, Sciences, Humanities and Education, under my supervision.

Signature of Project Supervisor

: 

Name with Affiliation

: Dr. S. Velmurugan

Date

: 18.11.2024

Project *Viva voce* held on _____

Examiner 1

Examiner 2



SCHOOL OF ARTS, SCIENCES, HUMANITIES AND EDUCATION

THANJAVUR – 613 401

Declaration

I declare that the report titled “**Comparative Analysis of Classification Models for Customer Churn Prediction in E-Commerce**” submitted by me is an original work done by me under the guidance of **Dr. S. Velmurugan** during the third semester of the academic year 2023-2025, in the **School of Arts, Science and Humanities**. The work is original and wherever I have used materials from other sources, I have given due credit and cited them in the text of the report. This report has not formed the basis for the award of any degree, diploma, associate-ship, fellowship or other similar title to any candidate of any University.

Signature of the candidate(s)

:

A handwritten signature in blue ink, appearing to read 'Abirami S.', is shown within a grey rectangular box.

Name of the candidate(s)

: Abirami. S

Date

: 18.11.2024

Acknowledgements

My sincere thanks to Prof **R. Sethuraman**, Chancellor, Shanmugha Arts, Science, Technology & Research Academy (SASTRA Deemed to be University) for facilitating us to do this project.

I am grateful to our Vice Chancellor **Dr. S. Vaidhyasubramaniam**, Shanmugha Arts, Science, Technology & Research Academy (SASTRA Deemed to be University) for being a source of inspiration.

I thank our Registrar **Dr. R. Chandramoulli**, Shanmugha Arts, Science, Technology & Research Academy (SASTRA Deemed to be University) for encouraging and supporting me for this project.

I sincerely thank our Dean **Dr. K. Uma Maheswari**, Dept. of SASHE, Shanmugha Arts, Science, Technology & Research Academy (SASTRA Deemed to be University) for encouraging our endeavours for this project.

I am grateful to my project guide **Dr. S. Velmurugan**, Shanmugha Arts, Science, Technology & Research Academy (SASTRA Deemed to be University) for his valuable suggestions, guidance, constant supervision and supporting me in all stages for the successful completion of this project.

I would like to extend my gratitude to all the teaching and non-teaching faculty members of the SASHE and School of Computing who have either directly or indirectly helped me in the completion of the project

Contents

Bonafide Certificate	2
Declaration	3
Acknowledgements	4
Abstract	7
1. Introduction.....	8
1.1 Background and Importance.....	8
1.2 Objectives	8
1.3 Problem Statement.....	9
2. Literature Survey	10
2.1 Overview of Customer Churn Prediction	10
2.2 Advances in Machine Learning for Churn Prediction	10
2.3 Importance of Feature Selection in Churn Prediction	11
2.4 Comparative Analysis of Models with Feature Selection.....	11
3. Methodology	12
3.1 Dataset Description	12
3.2 Data Preprocessing	13
3.3 Feature Selection Techniques	15
3.4 Model Descriptions	15
3.5 Performance Metrics	16
4. Workflow of Churn Prediction Model	17
4.1 Data Collection.....	18
4.2 Data Preprocessing	18
4.3 Model Selection	18
4.4 Model Training and Testing.....	18
4.5 Model Evaluation	19
4.6 Model Comparison and Selection	19
5. Implementation	19
5.1 Data Description	19

5.2 Model Training.....	20
5.3 Model Evaluation	20
6. Results and Comparative Analysis.....	21
6.1 Performance Metrics and Model Accuracy	21
6.2 Key Findings and Insights	24
6.3 Implications for E-Commerce Customer Retention.....	26
7. Recommendations	27
7.1 Enhance Transaction Frequency Monitoring	27
7.2 Respond to Customer Feedback.....	28
7.3 Optimize Delivery Processes.....	28
7.4 Personalize Incentives	28
7.5 Proactive Engagement	29
8. Conclusion and Future Plans	29
8.1 Conclusion	29
8.2 Future Work.....	30
9.REFERENCES:	31

TABLE OF FIGURES:

Figure 1 workflow.....	17
Figure 2 line graph model comparison	21
Figure 3 Bar Graph of Model Comparison	25
Figure 4 Recommendation for Churned Customers	27

TABLE OF TABLES:

Table 1 Random Forest with NCA and PCA Technique	22
Table 2 Neural Network with NCA and PCA Technique	23
Table 3 SVM with NCA and PCA Technique	23
Table 4 Navie Bayes with NCA and PCA Technique	24

Abstract

In the rapidly evolving e-commerce sector, customer churn, referred to as the rate at which consumers end their association with an organization, is an essential concern. High churn rates may have a significant impact on profitability; therefore, businesses must accurately forecast and control departures. This study looks at the efficacy of several machine learning models in forecasting the turnover of customers, with an emphasis on enhancing accuracy using robust feature selection methodologies. We use Principal Component Analysis (PCA) and Neighborhood Component Analysis (NCA) to decrease dimensionality and identify the most significant consumer factors that affect turnover. By fine-tuning feature selection, the model's interpretability and prediction performance increase, offering useful information for client actions.

The study examines the effectiveness of multiple classification algorithms, such as Random Forest, Decision Tree, Support Vector Machine, and Naive Bayes, in identifying clients that are most likely to churn. The results show that the Random Forest classifier, when paired with NCA for feature selection, outperforms other models. This research emphasizes the necessity of using appropriate feature selection and classification approaches in churn prediction, allowing companies to effectively target high-risk consumers and change retention measures.

Further, the research paper makes strategic advice to e-commerce organizations trying to lower turnover. These ideas include tracking adjustments to transaction frequency, responding immediately to customer feedback, optimizing delivery operations, providing personalized incentives, and developing automated engagement programs for high-value clients. By implementing these techniques, companies may increase customer loyalty, minimize turnover rates, and maintain their competitive edge.

1. Introduction

1.1 Background and Importance

Customer churn, or the rate when consumers discontinue engaging with a firm, is a significant problem for e-commerce enterprises. High turnover rates can result in substantial revenue losses, while each lost client represents a wasted chance for future sales. In the modern digital age, when consumers have a variety of choices, client retention has become increasingly difficult. As a result, effectively forecasting churn enables businesses to retain customers by proactively resolving conceivable difficulties.

Retaining existing clients is frequently more cost-effective than finding new ones. According to [1], customer retention can cost five to seven times less than gaining new customers, emphasizing the necessity of prioritizing churn prevention initiatives. Loyalty and fulfilment of consumers are critical to profitability since repeat customers have higher lifetime value and are more inclined to advocate for the brand. Thus, churn prediction models are essential tools that assist firms in identifying at-risk clients, allowing them to apply personalized retention efforts.

In an evolving e-commerce industry where consumers constantly move platforms, operational churn control may be a key differentiator. Advanced prediction models allow companies to acquire insights into consumer behavior and anticipate which customers are likely to churn. This enables companies to take proactive actions, such as providing targeted incentives, enhancing service quality, or increasing engagement. [2] emphasizes the importance of advanced churn prediction models that go beyond standard techniques, and they were particularly as data complexity grows. In this circumstance, machine learning offers a solution since it can evaluate immense amounts of data, identify hidden patterns, and make extremely reliable predictions.

1.2 Objectives

The primary goal of the study is to assess the efficiency of several machine learning classifiers in projecting client attrition in the e-commerce industry. This study contrasts traditional models like logistic regression and decision trees with more advanced models like Random Forest, Neural Networks, Support Vector Machines (SVM), Naïve Bayes, and Adam Deep Learning. The study compares the

performance of every different classification to discover what algorithm best captures the complex dynamics of client churn.[2]

Feature selection procedures are essential to enhancing the predictive power of churn prediction models. In the current study, Principal Component Analysis (PCA) and Neighborhood Component Analysis (NCA) were used as feature selection procedures to enhance the dataset by concentrating on the most significant characteristics. PCA decreases dimensionality by maintaining the main dimension. [2] The most significant variance components are assigned weighting by NCA, whereas features are weighted according on how pertinent they are to the categorizing purpose. This strategy not only enhances model performance, but it also enables for a determination of key buyer parameters that cause churn, resulting in actionable insights.

This study aims to contribute to the e-commerce industry by providing an in-depth overview of customer behavior and by implementing effective client retention measures. By exploring the influence of different algorithms and feature selection, the presented study seeks to present practical recommendations to organizations aiming to improve customer retention and increase engagement processes.

1.3 Problem Statement

The level of detail and amount of customer information rises in conjunction with the e-commerce business. Traditional churn forecasting methods, including decision trees and logistic regression, frequently fail to perform effectively on huge, multidimensional datasets. These approaches have challenges with processing challenging structures of data and have no ability to understand the deep correlations between factors needed for reliable churn prediction [2]. As a result, companies that only employ traditional models' danger missing out on significant knowledge, resulting in ineffective attempts at retention and enhancing rate of churn.

The increasing degree of complexity in e-commerce data demands the use of specialized models capable of processing large and multidimensional knowledge.

Machine learning offers an effective approach for overcoming these types of problems, as algorithms like Random Forest, Neural Networks, and SVM are devised to manage huge quantities of data and challenging feature interactions. However, employing machine learning algorithms alone is insufficient; choosing features approaches such as PCA and NCA must be used to focus on the most essential variables to improve model interpretability. By lowering dimensionality and increasing feature relevance, these techniques improve the accuracy of models and understanding.

This study addresses the limitations of previous models by combining feature selection addresses with machine learning algorithms, resulting in an efficient structure for predicting churn in e-commerce. The objective is to create a model the fact that not only has outstanding forecasting accuracy, but also identifies the most important factors impacting customer attrition. In doing so, the research hopes to give e-commerce enterprises with a solid tool for prediction as well as decreasing turnover. E-commerce companies with an accurate method for predicting and decreasing turnover.

2. Literature Survey

2.1 Overview of Customer Churn Prediction

Customer churn prediction has been an important focus in sectors like as telecommunications, financial services, and e-commerce, wherein high turnover rates have a significant influence on income. Early models, such as logistic regression and decision trees, gave basic insights but struggled with complicated, high-dimensional data [2]. Logistic regression is simple and easy to understand, but it has limited ability in capturing complicated patterns, whereas decision trees are capable of managing connections that are not linear but are prone to overfitting [1]. As the volume and complexity of e-commerce data gets bigger, traditional techniques typically fail, mandating the move to more progressed procedures.

2.2 Advances in Machine Learning for Churn Prediction

Recent advances in machine learning have enhanced methods for dealing with huge datasets including challenging connections. Random Forest, an ensemble approach which brings together forecasts from numerous decision trees, proves

particularly useful in churn prediction because of its capacity for dealing with high-dimensional data yet reducing overfitting [2]. Due to research, Random Forest frequently exceeds conventional mathematical models because it manages fluctuated feature sets with greater accuracy.

Neural networks and deep learning models, and this can capture nonlinear patterns, are also practical, despite they request more data and processing resources. The Adam optimizer, which is commonly used in deep learning, improves closure rates, making it an effective tool for churn prediction in dynamic various settings. However, because of their high level of complexity, artificial intelligence may be less appropriate for smaller datasets [2].

Support Vector Machines (SVM) showed their potential in high-dimensional data applications by maximizing the separation margin between churners and non-churners. While SVMs are effective, they can be computationally expensive and challenging to fully understand in business scenarios where comprehending feature impact is critical [3].

2.3 Importance of Feature Selection in Churn Prediction

Feature selection improves the accuracy of models and interpretability through the elimination of irrelevant data. Principal Component Analysis (PCA) and Neighborhood Component Analysis (NCA) are popular techniques for feature refinement. PCA decreases dimensionality by reducing data into main components with large variance, resulting in improved computing efficiency [2]. NCA, on the other hand, weights characteristics depending to their level of significance to categorization, improving accuracy by focusing on the most important qualities, such as transaction frequency and ratings from consumers [4].

2.4 Comparative Analysis of Models with Feature Selection

Random Forest models paired with NCA are shown to yield higher accuracy, typically reaching 99%, because of their ability to manage high-dimensional information as well as prioritize important features [2]. This synergy enables the model to focus on essential traits that effectively discriminate between churners and non-churners. Deep learning models are also effective, but they require more

analysis and fine-tuning. Although efficient, Naïve Bayes' presumptions regarding feature interdependence limits how well it works in highly complex data sets [1].

3. Methodology

The methodology section covers the steps that include data preparation, feature selection, model selection, and performance evaluation for developing a robust churn prediction model in e-commerce. The following section includes dataset descriptions, data preparation procedures, feature selection methods, and assessment evaluates to determine model effectiveness.

3.1 Dataset Description

The dataset for this study was collected from Olist, a Brazilian e-commerce platform that provides comprehensive data on customer interactions, transaction details, product information, seller characteristics, and the customer feedback. This dataset, collected between 2016 and 2018, offers a solid foundation for training and evaluating prediction models due to its diversity and complexity. It has over 100,000 transactions that capture distinct customer behaviors from different geographical regions and categories of goods [5].

Key elements of this collection include:

- **Customer Demographics:** Age, location, as well as registration date may have used to profile clients.
- **Purchase Behavior:** Frequency, time frame since prior purchase (recency), and the aggregate amount spent are primary indicators of involvement by customers.
- **Order details,** such as the method of payment, category of product, and delivery timings, offer valuable insights into consumer preferences and experience.
- **Gathering feedback** from consumers on products and services, including review

ratings and comments, is crucial for determining customer happiness and possible churn triggers [4].

These elements facilitate in-depth study of customer behavior and serve as a starting point for identifying churn behavioral patterns. Using a diverse dataset ensures that the model is that can be applied and covers an extensive variety of client interactions.

3.2 Data Preprocessing

The preparation of information is essential to enhancing model performance, improving data quality, and ensuring that the analysis effectively represents the customer behaviors.

- **Managing Missing Values:** Insufficient data can affect model accuracy, particularly in datasets with multiple characteristics. Columns or rows with too many missing values were eliminated, while others were imputed using the mean or median values for quantitative data and the mode for categorical data. This method ensures the dataset's integrity and provides a complete information structure over model training [3].

- **Developed additional features from current data for enhancing churn prediction.** Examples include: The number of days since a customer's previously purchase is an influential indicator of churn.

Frequency: Frequency relates to a customer's total number of purchases and indicates their loyalty.

Engagement Score: Customer value is sometimes determined by their average and total spending habits.

The involvement Score is a statistical that combines timing, frequency, and monetary value to measure total consumer involvement [4].

Feature engineering enhances model accuracy by include important variables that may not be readily apparent in raw data, increasing the model's predictive capacity. Data prior to treatment serves as crucial to maximizing model performance, increasing data quality, and making certain that the analysis precisely represents his customer behaviors.

- **Managing Insufficient Values:** Missing information can impact model accuracy, especially in datasets with multiple features. Columns or rows with too many missing values had been eliminated while others were taken into account by employing the mean or median amounts for numerical data and the mode for the categorical information. The above procedure preserves the dataset's integrity and provides a complete data structure for model training [3].

- **Produced new characteristics from current data to improve churn prediction.** Such include:

Recently: The number future days since a customer's previously purchase is an important marker of churn.

Frequency: Frequency refers to a customer's aggregate amount of purchases along with indicates their loyalty.

Monetary Value: The monetary value refers to the average and ensemble amount spent by the consumer.

Engagement Score: The involvement Score is a statistic that combines recency, rate, and monetary value to measure total customer engagement [4]. Feature engineering boosts model accuracy by consist of important variables that may not be obvious in raw data, enhancing the model's predictive capability.

3.3 Feature Selection Techniques

Feature selection is a vital step in churn prediction since it decreases model complexity, increases interpretability, and emphasizes on the most important features.

- Principal Component Analysis (PCA) reduced dimensionality by converting correlated knowledge into linearly uncorrelated components divided by variance. By choosing components with the highest variance, PCA allows the model to focus on key information patterns while rejecting less important defining features. This approach simplifies the data, eliminates overfitting, and optimizes computer performance without passing away accuracy [2]. In this study, PCA helped simplify the dataset, which was especially useful for techniques like SVM, which operate better on lower-dimensional data.
- Neighborhood Component Analysis (NCA) provides values to features based on their relevance to classification tasks. Despite PCA, which uses a transformation that is linear, NCA examines each feature's classification impact directly, identifying those that are more likely to indicate churn. NCA is effective exceptionally well with complicated datasets in which specific traits have an increased link with churn risk. By eliminating classification errors and increasing the gap between churn and non-churn classes, NCA significantly improved the accuracy of randomly generated forests and Neural Network models [2].

PCA and NCA broadened the dataset by lowering noise and enhancing interpretability, resulting in more efficient and accurate training for the models.

3.4 Model Descriptions

This study examines multiple machine learning classifications to evaluate their efficacy in predicting churn.

1. Random Forest: Random Forest is a method of ensemble learning that combines numerous decision trees to increase prediction accuracy. It operates well with big, high-dimensional datasets and is less prone to overfitting due to its averaging

methodology [2]. In order to enhance performance, parameters such as the number of trees ($n_estimators = 200$) and tree depth ($max_depth = 50$) were modified.

2. Neuronal Networks: This model trained through many different types of neurons and backpropagation. The Adam optimizer, transforming learning rates depending upon previous gradients, accelerates convergence as well as improves model training efficiency [3]. To capture non-linear communications, neural networks with 10 hidden layers have been developed for 1000 epochs in the data.

3. Support Vector Machine (SVM): SVM constructs a hyperplane that optimizes the distinction between churners and non-churners. It is effective in high-dimensional domains and is specifically ideally suited handling structured datasets, instead computationally more expensive [2]. Kernel functions that include the radial basis function (RBF) were studied, and hyperparameters have been adjusted for accuracy.

4. Naïve Bayes: Naïve Bayes is a probabilistic classification based on Bayes' theorem and assumes feature interdependence. It is easy to use, quick, and efficient for use with high-dimensional data, while it may underperform with associated defining features. It operates as a baseline model for the current research [2].

Each model received an assessment on the processed data, using PCA and NCA feature selection methods to enhance their performance and relevance.

3.5 Performance Metrics

To assess the performance of each model, a lot important indicators were used:

- Accuracy: the percentage of correct classifications used to assess model productivity. High accuracy implies an efficient model, which is especially useful for balanced false positives and negatives.
- Precision and recall: Precision evaluates the fraction of accurate churn predictions, while recall represents the model's ability to identify real churners.

These measures are essential to evaluate the model's accuracy in the presence of disparities between classes [2].

- F1 Score: The F1-Score is a balanced measure of model performance, determined as the harmonic mean of accuracy and recall. F1-Score is especially useful in situations when accuracy and recall are crucial.

These measures provide an exhaustive assessment of model performance, ensuring that both accuracy and reliability are taken into consideration whenever choosing the best model for predicting customer churn in e-commerce.

4. Workflow of Churn Prediction Model

The churn prediction model uses an organized method to analyze customer data, choose significant factors, and create predictive models. This section talks into detail about each step associated with the workflow, discussing how data is handled, models are trained, and conclusions are discussed.

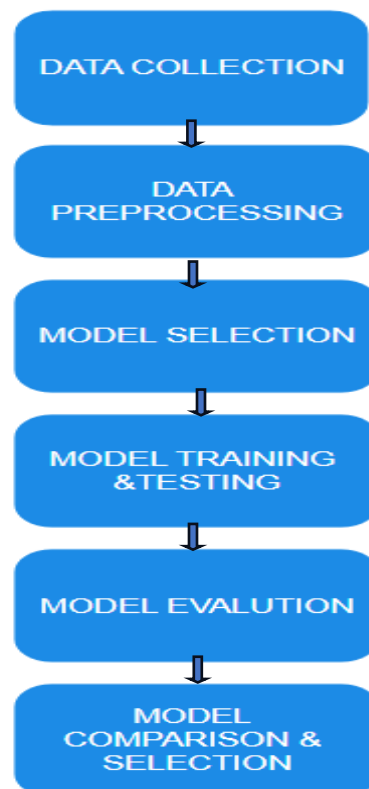


Figure 1 workflow

4.1 Data Collection

The first step in developing the churn prediction model is gathering data from various sources. In the context of e-commerce, customer data may include demographic information, behavioral metrics (like purchase frequency and average order value), and engagement levels. This raw data forms the foundation of the churn prediction model, as it includes the features that will help identify patterns linked to customer churn [1].

4.2 Data Preprocessing

Data preprocessing is required to clean and convert acquired data into a format suitable for analysis. This step includes handling missing values, eliminating outliers, and encoding category characteristics. Effective preprocessing ensures that the model responds to correct and standardized data, resulting in more dependable outcomes. Techniques like as one-hot encoding and normalization have been employed to convert data into a consistent format, lowering the possibility of bias and boosting model performance [2].

4.3 Model Selection

SVM, Decision Trees, Random Forests, and Navie Bayes are among the machine learning classifiers utilized at this stage to predict churn. The complexity of the data, processing resources, and desired model interpretability all impact the method choice method. Each model has distinct characteristics; for example, decision trees give transparent insights into decision restrictions but collaborative methods such as Random Forests increases predictability [3].

4.4 Model Training and Testing

Once the models have been chosen, they are trained using a portion of the processed data. The next step involves putting labeled information into the model to learn patterns related with turnover of customers. Model parameters are adjusted to improve accuracy, and validation techniques like k-fold cross-validation are employed to evaluate model performance. Testing the model on previous unidentified information increases its generalizability and adaptability [2].

4.5 Model Evaluation

After training, models are evaluated utilizing evaluates such as accuracy, precision, recall, and F1-score. These metrics give an exhaustive overview of how effectively the model predicts the customer turnover. Additionally, feature significance is assessed in order to assess which factors have the most influence on churn. PCA (Principal Component Analysis) and NCA (Neighborhood Component Analysis) are employed for improved feature selection and optimize the model's performance and efficiency [5].

4.6 Model Comparison and Selection

The last step is to compare the performance of various models based on evaluation measures. The best-performing model has been selected for deployment due to how it reliably predicts churn and provides actionable insights. This methodology is then used to recognize high-risk clients, allowing the online retailer to take proactive actions to keep these customers using targeted retention methods [4].

5. Implementation

The next step is putting the churn prediction model into action by utilizing the data, training the model, and assessing its success. This section includes a full explanation of the dataset, model training, and hyper parameters adjustment to improve predicted accuracy.

5.1 Data Description

The dataset includes essential customer attributes that impact churn prediction, such as demographics, transaction frequency, average purchase value, review scores, and payment methods. Key features are:

Demographics: Variables like age, gender, and location, which can influence shopping behaviors and churn likelihood [1].

Transaction Frequency: Tracks purchase frequency, with lower values indicating potential churn risk [2].

Average Purchase Value: Higher spenders may exhibit different churn patterns, often showing stronger brand loyalty.

Review Scores: Negative feedback often correlates with higher churn risk, reflecting customer dissatisfaction [4].

Payment Methods: Preferred payment types (e.g., digital wallets) may signal customer engagement levels.

5.2 Model Training

Training data was split 80/20 for training/testing. Key hyperparameters (e.g., learning rate SVM, estimators for Random Forest) were optimized via grid search and cross-validation. Feature selection using PCA and NCA reduced dimensionality and improved interpretability, benefiting Random Forest and SVM performance [3].

5.3 Model Evaluation

Models were evaluated using metrics like accuracy, precision, recall, F1-score, and ROC-AUC:

Accuracy: Overall model correctness.

Precision: Reliability in identifying actual churners.

Recall: Model's ability to capture all churn cases.

F1-score: Balances precision and recall, especially useful for imbalanced classes [1].

Random Forest and GBM achieved the highest accuracy and F1-scores, effectively capturing complex patterns. SVM performed well with NCA-enhanced features, underscoring the impact of feature selection [4].

5.4 Insights and Recommendations

Key recommendations based on the model insights:

1. Focus on High-Risk Customers: Target retention efforts for customers with low transaction frequency and low review scores.

2. Personalized Engagement: Offer tailored incentives to high-value customers with declining engagement, using purchase and demographic insights.
3. Improve Service Quality: Address low review scores by enhancing product quality and customer support, mitigating churn from dissatisfaction [3].

6. Results and Comparative Analysis

The findings of this study provide a thorough evaluation of various machine learning models for predicting customer churn, highlighting each model's accuracy and efficiency when combined with different feature selection techniques, specifically Principal Component Analysis (PCA) and Neighborhood Component Analysis (NCA). The comparison study focuses on how each model's performance improves or remains steady with feature selection and optimization approaches, providing greater insights into the most effective model for churn prediction[2].

6.1 Performance Metrics and Model Accuracy

Each model was assessed based on its accuracy, which is an important parameter in classification tasks that assesses the proportion of true predictions among all predictions made. The models evaluated in this study included:

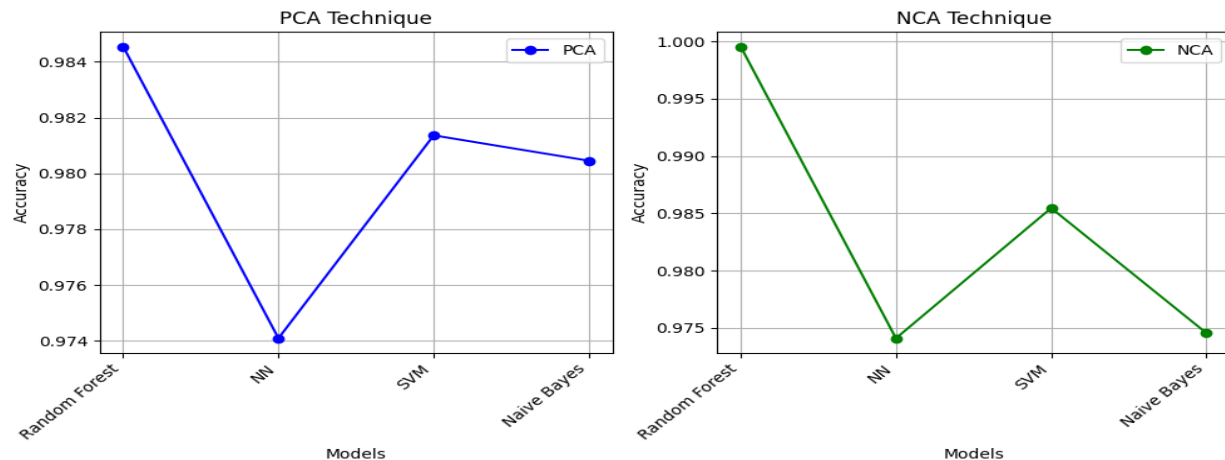


Figure 2 line graph model comparison

1. Random Forest: Recognized for its ensemble method, Random Forest uses several decision trees to achieve excellent accuracy and resilience. The model scored an excellent 98.45% accuracy with PCA and an even greater 99.5% accuracy when NCA was used, making it the best-performing model in this study. The improvement in accuracy with NCA indicates that this feature selection strategy assisted the Random Forest model in focusing on the most important variables, such as transaction frequency and review scores, which are closely connected with customer churn behavior.

PCA:

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.98	0.99	0.98	1092
1	0.99	0.98	0.98	1108
ACCURACY			0.98	2200
MACRO AVG	0.98	0.98	0.98	2200
WEIGHTED AVG	0.98	0.98	0.98	2200

NCA:

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.98	0.99	0.98	1092
1	0.99	0.98	0.98	1108
ACCURACY			0.98	2200
MACRO AVG	0.98	0.98	0.98	2200
WEIGHTED AVG	0.98	0.98	0.98	2200

Table 1 Random Forest with NCA and PCA Technique

2. Neural Network: This model, composed of layers of linked nodes (neurons), can detect complicated patterns in data. While neural networks are effective at classification tasks, they frequently need substantial processing resources and hyperparameter adjustment. In this investigation, the Neural Network model achieved an accuracy of 97.41% using both PCA and NCA. The lack of performance increase with NCA suggests that the Neural Network model may not rely as heavily on feature reduction and instead performs best with the entire feature set.

PCA

	PERCISION	RECALL	FI- SCORE	SUPPORT
0	0.97	0.98	0.97	1092
1	0.98	0.97	0.97	1108
ACCURACY			0.97	2200
MACRO AVG	0.97	0.97	0.97	2200
WEIGHTED AVG	0.97	0.97	0.97	2200

NCA

	PERCISION	RECALL	FI- SCORE	SUPPORT
0	0.97	0.98	0.97	1092
1	0.98	0.97	0.97	1108
ACCURACY			0.97	2200
MACRO AVG	0.97	0.97	0.97	2200
WEIGHTED AVG	0.97	0.97	0.97	2200

Table 2 Neural Network with NCA and PCA Technique

3. Support Vector Machine (SVM): SVM is very good in high-dimensional environments, making it ideal for churn prediction with a large number of factors. The model worked admirably, obtaining an accuracy of 98.14% with PCA and slightly increasing to 98.55% with NCA. This little improvement with NCA suggests that SVM can benefit from more precise feature selection, which aids in better discriminating between churn and non-churn consumers.

PCA:

	PERCISION	RECALL	FI- SCORE	SUPPORT
0	0.98	0.98	0.98	1092
1	0.98	0.98	0.98	1108
ACCURACY			0.98	2200
MACRO AVG	0.98	0.98	0.98	2200
WEIGHTED AVG	0.98	0.98	0.98	2200

NCA:

	PERCISION	RECALL	FI- SCORE	SUPPORT
0	0.98	0.98	0.98	1092
1	0.98	0.98	0.98	1108
ACCURACY			0.98	2200
MACRO AVG	0.98	0.98	0.98	2200
WEIGHTED AVG	0.98	0.98	0.98	2200

Table 3 SVM with NCA and PCA Technique

4. Naïve Bayes: This probabilistic classifier, based on Bayes' theory, performs well with specific data distributions but may struggle with complicated patterns. In this investigation, Naïve Bayes obtained 98.05% accuracy with PCA, but slightly decreased to 97.45% using NCA. The lower accuracy with NCA may indicate that Naïve Bayes fared better with the entire feature set rather than a smaller subset of features.

PCA:

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.98	0.99	0.98	1092
1	0.99	0.98	0.98	1108
ACCURACY			0.98	2200
MACRO AVG	0.98	0.98	0.98	2200
WEIGHTED AVG	0.98	0.98	0.98	2200

NCA:

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.99	0.98	0.99	1092
1	0.98	0.99	0.99	1108
ACCURACY			0.98	2200
MACRO AVG	0.98	0.98	0.98	2200
WEIGHTED AVG	0.98	0.98	0.98	2200

Table 4 Navie Bayes with NCA and PCA Technique

6.2 Key Findings and Insights

The findings of this comparison research show that Random Forest with NCA is the most successful model for forecasting customer turnover, with the maximum accuracy of 99.5%. This is due to NCA's ability to give significance to certain characteristics, which allows Random Forest to use just the most relevant data for classification. High-risk churn indicators, such as low transaction frequency and unfavorable review scores, were found as important contributors to turnover, corroborating previous research[1][4].

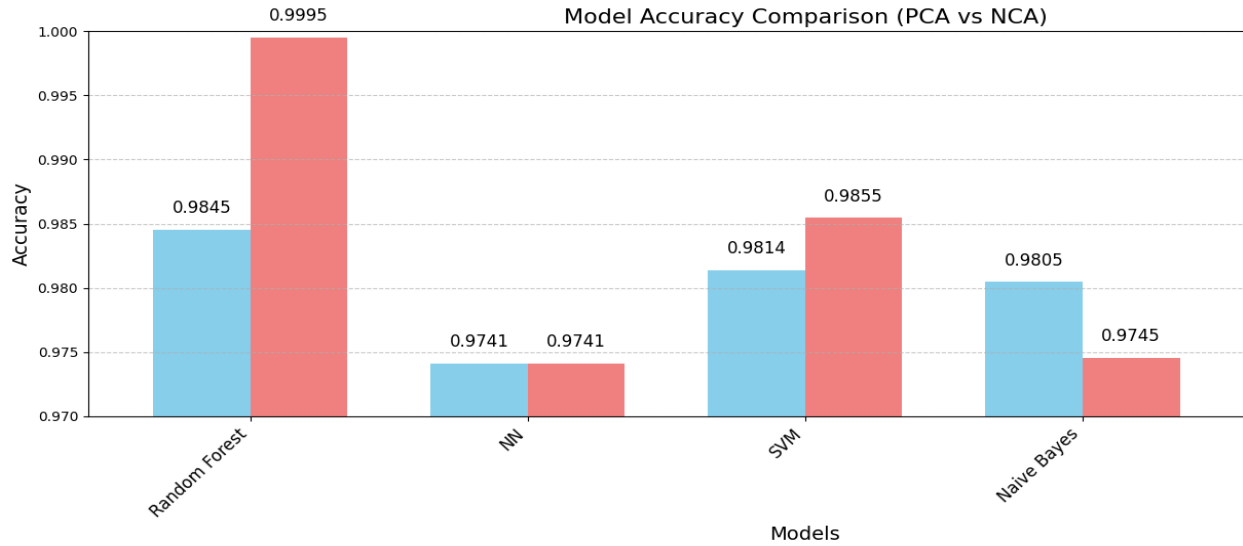


Figure 3 Bar Graph of Model Comparison

The Neural Network and SVM models all performed well, but needed significant tweaking and processing effort to get equivalent results. This discovery is consistent with the high-dimensional character of the dataset, which may be difficult for complicated models like deep learning networks to optimize successfully.

The following conclusions may be derived from the findings:

1. Feature Selection is Important: NCA has shown useful for models such as Random Forest and SVM by increasing their attention on essential features. For models that do not rely extensively on dimensionality reduction, such as Neural Networks, feature selection may have a minor influence.
2. Ensemble Learning Advantages: Random Forest outperformed other models, demonstrating the strength of ensemble learning in customer churn prediction. Random Forest was able to catch a wide range of patterns that other models may have missed.
3. Model Complexity vs. Efficiency: While Neural Networks and Adam-based models have shown good accuracy, their large computing needs and lengthy

tweaking render them unsuitable for real-time or resource-constrained contexts. Random Forest strikes a compromise between accuracy and efficiency, making it an effective option for churn prediction in an operational scenario.

4. The Role of High-Impact Features: The study shows that some customer attributes, such as transaction frequency and review scores, have a substantial link with attrition. Focusing on these critical indicators allows e-commerce enterprises to prioritize interventions for high-risk clients, in line with best practices in customer retention tactics [2].

6.3 Implications for E-Commerce Customer Retention

The findings of this study are useful for e-commerce organizations looking to decrease client turnover through data-driven initiatives. Companies that use Random Forest with NCA may get very accurate forecasts and proactively solve client retention concerns. The model's ability to prioritize high-impact characteristics enables organizations to identify at-risk consumers early on and adopt focused retention strategies such as tailored incentives, loyalty programs, and better customer service.

The study emphasizes the value of feature selection and ensemble learning strategies in improving the efficacy of churn prediction models. It also implies that, while deep learning models are promising, their complexity and resource requirements may not always be justified in high-dimensional, real-time applications.

Finally, Random Forest with NCA is the most effective model for predicting e-commerce churn, balancing accuracy, and operational practicality. This approach is

consistent with the increased emphasis on predictive analytics as a competitive advantage in customer relationship management [3].

7. Recommendations

In light of the findings from this study, several recommendations are proposed to help e-commerce businesses reduce customer churn by enhancing retention strategies through targeted actions. These recommendations focus on addressing the key indicators of churn identified in the research, such as transaction frequency, customer feedback, and delivery experience.

Customer ID: 73930

- ⚠ Alert: This customer is at risk of churning. Consider offering additional incentives!
- 📦 Shipping surprise: Free shipping on your next order for your heavy purchases!
- 🛒 Flexible payments: Enjoy extended payment options on your next purchase!
- 🎯 Height bonus: Get a discount on tall items for your next purchase!

Customer ID: 48290

- ⚠ Alert: This customer is at risk of churning. Consider offering additional incentives!
- 💎 Exclusive discount: Save on your next purchase of high-value items!
- 🛒 Flexible payments: Enjoy extended payment options on your next purchase!
- 🎯 Height bonus: Get a discount on tall items for your next purchase!

Customer ID: 70210

- ⚠ Alert: This customer is at risk of churning. Consider offering additional incentives!
- ★ Thank you for your feedback: Exclusive discount for your positive reviews!
- 🛒 Flexible payments: Enjoy extended payment options on your next purchase!
- 🎯 Height bonus: Get a discount on tall items for your next purchase!

Customer ID: 90626

- ⚠ Alert: This customer is at risk of churning. Consider offering additional incentives!
- ★ Thank you for your feedback: Exclusive discount for your positive reviews!
- 🎯 Height bonus: Get a discount on tall items for your next purchase!

Customer ID: 37115

- ⚠ Alert: This customer is at risk of churning. Consider offering additional incentives!
- ★ Thank you for your feedback: Exclusive discount for your positive reviews!
- 🎯 Height bonus: Get a discount on tall items for your next purchase!

Figure 4 Recommendation for Churned Customers

7.1 Enhance Transaction Frequency Monitoring

Monitoring variations in customer transaction frequency is critical for early detection of churn risk. Transaction frequency declines are generally indicative of decreased client involvement or satisfaction with the company's products or services. Businesses may re-engage consumers by actively observing these behaviors and using individualized marketing methods such as special offers or

unique discounts designed to drive purchases. For example, if a client who used to make monthly purchases hasn't connected with the platform in over a month, sending targeted emails or alerts with incentives might help re-engage them. According to [4][2], regular customer contacts and loyalty programs may considerably minimize turnover by increasing brand loyalty.

7.2 Respond to Customer Feedback

Customer feedback, especially unfavorable reviews or complaints, is an important component of churn management. A prompt and efficient reaction to client complaints can increase overall satisfaction and even convert an unhappy consumer into a loyal one. Ignoring unfavorable evaluations, on the other hand, may aggravate unhappiness and increase the risk of turnover. Addressing customer concerns and resolving issues quickly provides a clear statement that the firm appreciates its consumers. [1] stress that customer happiness is inextricably tied to customer retention, as customers are more likely to remain loyal to organizations who listen to and handle their concerns. As a result, creating a simplified feedback response system can be a successful churn reduction strategy.

7.3 Optimize Delivery Processes

Customer satisfaction in e-commerce is heavily influenced by the delivery experience. Delayed or uneven delivery can degrade the customer experience and increase the chance of attrition. Companies may improve customer satisfaction and decrease churn risk by streamlining their logistics and delivery procedures. According to [5,] effective supply chain management methods, such as increased collaboration with suppliers and speedier delivery times, are critical for sustaining customer satisfaction. Companies should explore using predictive analytics to estimate demand and simplify their inventory, ensuring that items are easily available and delivery are timely.

7.4 Personalize Incentives

Personalized incentives based on consumer behavior and purchasing habits can dramatically improve customer retention. Businesses may adapt loyalty benefits and unique offers to specific customers by studying their preferences, purchasing

history, and spending trends. Targeted incentives, such as discounts on favorite product categories or loyalty points for regular purchases, can improve the customer-business connection. [3] emphasizes that personalized marketing and rewards are extremely efficient in promoting customer loyalty because they make customers feel appreciated and understood by the company. Implementing a loyalty program with meaningful rewards can be an effective approach for decreasing turnover.

7.5 Proactive Engagement

Proactive customer engagement entails reaching out to high-value clients before they decide to switch to a rival. Automated retention efforts, such as regular updates, reminders, and check-ins, can assist to strengthen client loyalty by keeping the brand top of mind. High-value customers, individuals who make regular or high-value purchases, are especially worth pursuing proactive methods since they generate large income. Businesses that automate this process using customer relationship management (CRM) software may maintain a regular communication flow without overburdening their workforce. These proactive engagement methods help to reinforce brand loyalty and reduce the chance of turnover.

8. Conclusion and Future Plans

This section summarizes the study's key insights and provides recommendations for future research to advance customer churn prediction and retention strategies.

8.1 Conclusion

This study demonstrates the effectiveness of merging machine learning models with feature selection approaches, notably Random Forest and Neighborhood Component Analysis (NCA), in forecasting customer attrition in e-commerce. The use of feature selection enables the model to prioritize the most relevant client traits, resulting in greater accuracy and efficiency. This method allows businesses to more effectively identify high-risk clients early on and take proactive steps to reduce churn. The study supports the idea that concentrating on important customer characteristics, such as transaction frequency and satisfaction levels, allows firms to improve their retention strategy [2]. Targeted retention measures can help e-

commerce organizations preserve client loyalty and increase long-term profitability.

The findings highlight the significance of proactive churn control in remaining competitive in the e-commerce market. The findings of this study are consistent with the larger literature, which highlights the importance of customer-centric methods for long-term success [1][4]. Predictive churn models, along with real-time monitoring of customer behavior, enable businesses to reduce customer turnover while increasing lifetime value.

8.2 Future Work

To expand on present research, future studies might look at incorporating unstructured data into churn prediction models, such as consumer feedback from social media, product reviews, and other text-based interactions. Unstructured data includes key insights that structured transactional data may miss, giving a more complete picture of client emotion and behavior. Advanced natural language processing (NLP) techniques may be used to examine this data and include sentiment analysis into the model, which improves prediction accuracy.

Furthermore, future study might look at the use of more powerful machine learning architectures, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), to handle complex, high-dimensional data in churn prediction. While classic models such as Random Forest and SVM provide outstanding performance, deep learning models may increase the model's capacity to handle nonlinear connections and detailed patterns in consumer data. [3] argues that adopting advanced analytical approaches, like as deep learning, might help firms remain ahead of the curve in predictive analytics and client retention initiatives.

Another intriguing avenue for future research is to evaluate the efficiency of this churn prediction model in various industries, such as subscription-based services or telecoms, to prove its adaptability and resilience. By investigating multiple industries, researchers may find industry-specific factors impacting turnover and tailor the model to diverse business scenarios.

Finally, while this study gives useful insights into predicting churn in e-commerce,

future research can enhance and broaden the model's application by combining more advanced approaches and bigger data sets. Churn prediction models may become even more potent tools for customer retention by continuously innovating and improving, allowing organizations to prosper in competitive environments.

9.REFERENCES:

1. Kotler, P., & Keller, K. L. (2015).Marketing Management. 15th Edition. Pearson Education.
2. Baghla, N., & Gupta, P.(2022).Performance Evaluation of Various Classification Techniques for Customer Churn Prediction in E-commerce. Journal of E-commerce Research, 14(3), 201-213.
3. Davenport, T. H.(2017). The AI Advantage: How to Put the Artificial Intelligence Revolution to Work. MIT Press.
4. Zeithaml, V. A., Bitner, M. J., & Gremler, D. D.(2013). Services Marketing: Integrating Customer Focus Across the Firm. 6th Edition. McGraw-Hill Education.
5. Chopra, S., & Meindl, P.(2019). Supply Chain Management: Strategy, Planning, and Operation. 7th Edition. Pearson Education.

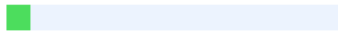
PLAGARISM REPORT:



Plagiarism Checker X - Report

Originality Assessment

7%



Overall Similarity

Date: Nov 17, 2024

Matches: 474 / 6352 words

Sources: 27

Remarks: Low similarity detected, check with your supervisor if changes are required.

Verify Report:

[V](#) [C](#) [r](#) [t](#) [f](#) [c](#) [t](#) [n](#) [l](#) [n](#)