Proceeding of International Conference on Advanced Technologies, Innovations and Science
Vol 1, No 1, June 2026, Page 537-545
ISSN XXXX-XXXX (Media Online)
https://journals.adaresearch.or.id/icatis

Hyperparameter Optimization of Logistic Regression Using GridSearchCV for Customer Churn Prediction

Saffira Izzati*, Dedy Hartama, Agus Perdana Windarto, Putrama Alkhairi, Irfan Sudari Damanik

Information Systems Study Program, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia
Jenderal Sudirman Street, Block A No. 1/2/3, Pematangsiantar 21127, North Sumatra, Indonesia
Email: [1,*]vira59645@gmail.com, [2]dedyhartama@amiktunasbangsa.ac.id, [3]agus.perdana@amiktunasbangsa.ac.id, [4]putramaalkhairi@amiktunasbangsa.ac.id, [5]irfansudahri@amiktunasbangsa.ac.id
Correspondence Author Email: vira59645@gmail.com
Submitted: **01/06/2026**; Accepted: **12/06/2026**; Published: **30/06/2026**

Abstract−Customer churn refers to a condition in which customers discontinue using a company's services, potentially resulting in decreased revenue and customer loyalty. Customer churn prediction on the Telco Customer Churn dataset was performed using the Logistic Regression algorithm optimized through GridSearchCV. The research process consisted of data preprocessing, categorical attribute transformation, feature scaling, splitting the dataset into training and testing sets, baseline model development, hyperparameter optimization, and model evaluation using accuracy, precision, recall, F1-score, confusion matrix, ROC curve, and AUC score metrics. The evaluation results showed that the baseline model achieved an accuracy of 81.61%, precision of 68.25%, recall of 55.13%, F1-score of 60.99%, and an AUC score of 86.00%. Meanwhile, the optimized model using GridSearchCV achieved an accuracy of 75.59%, precision of 52.05%, recall of 81.41%, F1-score of 63.50%, and an AUC score of 86.00%. The increase in recall from 55.13% to 81.41% and F1-score from 60.99% to 63.50% indicates that hyperparameter optimization improved the model's ability to identify customers who are likely to churn. However, these improvements were accompanied by decreases in accuracy and precision, indicating a trade-off between churn detection capability and positive prediction accuracy. Overall, the optimized model is considered more suitable for customer churn prediction tasks that prioritize the identification of customers at risk of churning

Keywords: Customer Churn; GridSearchCV; Hyperparameter Optimization; Logistic Regression; Machine Learning

1. INTRODUCTION

The development of information technology and digital transformation has encouraged companies to utilize data as a basis for making business decisions more effectively and systematically. One of the common challenges faced by subscription-based service companies is customer churn, which refers to a condition in which customers decide to stop using a product or service within a certain period (Iskandar et al., 2023). A high customer churn rate can lead to decreased company revenue, increased customer acquisition costs, and reduced customer loyalty (Loukili et al., 2022). Therefore, the ability to identify customers who are likely to churn is essential in helping companies develop customer retention strategies more proactively and effectively.

One of the classification methods widely used in customer churn prediction is Logistic Regression (Maisat et al., 2023). This method offers several advantages, including relatively fast computational processes, ease of interpretation, and effectiveness in binary classification problems such as churn and non-churn. Furthermore, Logistic Regression can describe the relationship between independent variables and the probability of an event occurring (Arivazhagan & Subramanian, 2020).

However, this method has several limitations, such as sensitivity to imbalanced data, suboptimal performance in handling complex and non-linear relationships, and dependency on the selection of model parameters. Therefore, an approach is needed to obtain a more optimal parameter configuration in order to improve the model's classification capability.

Various studies have been conducted to apply machine learning methods in customer churn prediction within the telecommunications industry. The implementation of Logistic Regression and Logit Boost algorithms has been reported to effectively identify customers with a high likelihood of churning . In addition, various classification algorithms have been employed to improve customer churn prediction accuracy, demonstrating that the selection of an appropriate method significantly affects model performance (Kavitha, 2020).

Other studies have shown that hyperparameter optimization in supervised learning algorithms can enhance customer churn prediction performance through the selection of more optimal parameter combinations (Loukili et al., 2022). Furthermore, the application of various machine learning algorithms has been reported to improve the identification of potential churn customers, thereby supporting more effective customer retention strategies (Maan & Maan, 2023). Previous research findings indicate that the implementation of machine learning techniques and hyperparameter optimization contributes to improving customer churn prediction performance.

Nevertheless, the performance of classification models is still influenced by the selection of model parameters and the characteristics of the data used. In addition, data distribution imbalance between churn and non-churn classes can affect the model’s ability to perform optimal classification. Although studies on hyperparameter optimization for customer churn prediction have been conducted, research specifically analyzing the impact of hyperparameter optimization using GridSearchCV on the performance of the Logistic Regression algorithm using the Telco Customer Churn dataset remains relatively limited (Hemlata Jain, Ajay Khunteta, 2020). This condition indicates that opportunities still exist to improve the model’s capability in detecting customers who are likely to churn (Kumari et al., 2025).

Based on these issues, this study applies hyperparameter optimization using GridSearchCV to the Logistic Regression algorithm for customer churn prediction (Wagh et al., 2024). GridSearchCV is employed to obtain the optimal parameter combination through a cross-validation process in order to enhance the model's classification performance. This study aims to analyze the effect of hyperparameter optimization using GridSearchCV on the performance of the Logistic Regression algorithm in predicting customer churn using the Telco Customer Churn dataset (Shrestha & Shakya, 2023). Model evaluation is conducted using accuracy, precision, recall, F1-score, and AUC score metrics to compare model performance before and after optimization. This research contributes by providing an analysis of the impact of hyperparameter optimization using GridSearchCV on the performance of Logistic Regression in customer churn prediction.

2. RESEARCH METHODS

This study was conducted to analyze the effect of hyperparameter optimization using GridSearchCV on the performance of the Logistic Regression method in predicting customer churn. The dataset used is the Telco Customer churn dataset obtained from the Kaggle platform (Autom & Geoespaciales, 2025).

The study began with the process of dataset collection, data preprocessing, categorical attribute encoding, training and testing data division, training the Logistic Regression baseline model, hyperparameter optimization using GridSearchCV, and evaluating model performance (Ehsani & Hosseini, 2024b). The research methodology was systematically structured to obtain a classification model that has optimal customer churn prediction capabilities (Rathore, 2025).

2.1 Research Dataset

The data used in this study comes from the Telco Customer Churn dataset obtained through the Kaggle platform. The dataset contains information about telecommunications company customers used to predict the likelihood of customers ceasing to use the service (customer churn). The variables in the dataset consist of numeric and categorical attributes such as tenure, MonthlyCharges, TotalCharges, InternetService, and PaymentMethod, with the churn variable as the classification target (Ehsani, 2022).

The data used in this study was 2,993 customer data consisting of 19 predictor attributes and 1 target attribute. The dataset was selected because it has suitable characteristics for applying the classification method to the case of customer churn, and includes various customer information that can influence customer decisions in using or discontinuing the company's services (Alamsyah et al., 2025).

Table 1. Customer churn Dataset Information

Attribute	Information
Number of Data	2993
Number of Variables	20
Target Data	Churn
Predictor Attributes	19
Data Type	Numeric and Categorical
Source	Kaggle

Based on Table 1, the Customer Churn dataset used in this study consists of 2,993 customer data with 20 variables, including 19 predictor attributes and one target attribute (churn). The dataset contains a combination of numeric and categorical attributes that are used as input in the customer churn classification process. These characteristics indicate that the dataset is suitable for use in developing and evaluating the classification model in this study.

Table 2. Dataset Class Distribution

Class	Total	Percentage
Non-churn	2211	74%
churn	777	26%

Based on Table 2, the class distribution in the dataset shows an imbalance between churned and non-churned classes. The non-churned class accounts for 74% of the data, while the churned class accounts for 26%. This indicates that the dataset is considered imbalanced data, potentially impacting the classification model's performance in optimally identifying minority classes.

2.2 Research Framework

The research framework is used to describe the research stages carried out systematically in the customer churn prediction process using the Logistic Regression method with hyperparameter optimization using GridSearchCV. The research stages start from the process of dataset collection, data validation, preprocessing, training and testing data division, baseline modeling, hyperparameter optimization, and model evaluation (Alotaibi, 2024).

Proceeding of International Conference on Advanced Technologies, Innovations and Science
Vol 1, No 1, June 2026, Page 537-545
ISSN XXXX-XXXX (Media Online)
https://journals.adaresearch.or.id/icatis

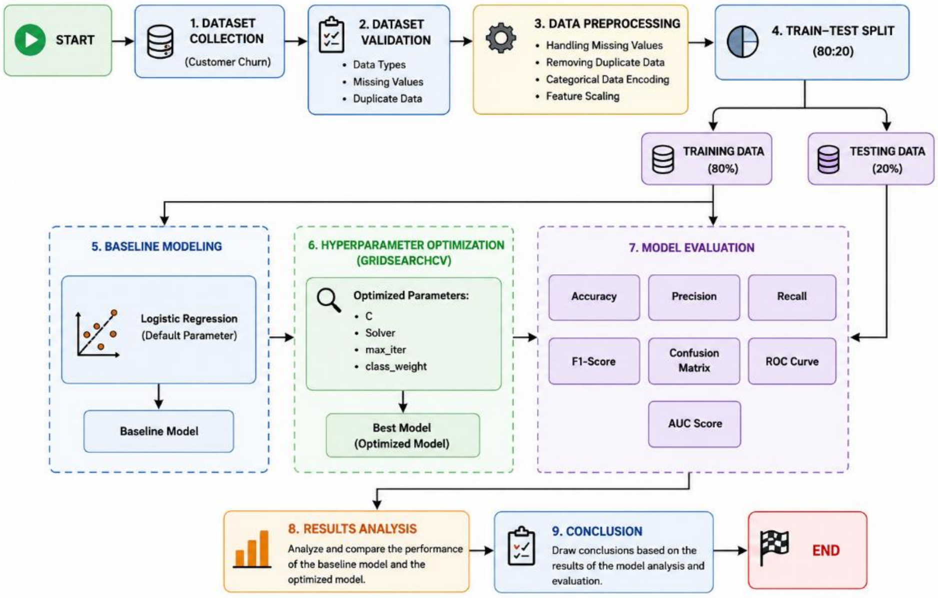

Figure 1. Customer Churn Prediction Research Framework

Based on Figure 1, the study began with the collection of the Telco Customer Churn dataset obtained from Kaggle and continued with the data validation process through checking data types, missing values, and duplicate data. The preprocessing stage was then carried out by handling missing values, removing duplicate data, encoding categorical data, and feature scaling to improve data quality (Preet et al., 2024). Next, the dataset was divided into training data and testing data using the train-test split method with a ratio of 80:20.

The training data was used to build a baseline Logistic Regression model, then hyperparameter optimization was performed using GridSearchCV to obtain the best parameter combination through cross-validation. The baseline model and the optimized model were then evaluated using the metrics of accuracy, precision, recall, F1-score, confusion matrix, ROC curve, and AUC score to analyze the effect of hyperparameter optimization on customer churn prediction performance.

2.3 Data Preprocessing

The data preprocessing stage was performed to improve the quality of the dataset before it was used in the Logistic Regression model training process. This process included checking and handling missing values, removing duplicate data, encoding categorical data, and feature scaling (Alotaibi, 2024). Missing value handling and duplicate data removal were performed to reduce errors, data redundancy, and potential bias that could affect the classification results. Next, the categorical data was transformed into numeric form using encoding techniques, and the numeric attributes were normalized using feature scaling to optimize and stabilize the model training process. After the preprocessing stage was completed, the number of data used in the modeling process was reduced from 2,993 to 2,988 due to the removal of duplicate data and data with missing values.

2.4 Logistic Regression

The classification method used in this study is Logistic Regression. Logistic Regression is a machine learning method used to solve classification problems by predicting the probability of data belonging to a particular class (Pan, 2025). In this study, the Logistic Regression method is used to predict the likelihood of customer churn based on the attributes contained in the dataset. Logistic Regression works using the sigmoid function to convert the results of linear combinations into probability values with a range of 0 to 1. These probability values are then used to determine the predicted class of churned or non-churn customers. The sigmoid function equation in Logistic Regression is shown in Equation (1).

$$P(y) = \frac{1}{1+e^{-z}} \tag{1}$$

Description:

P(y) = probability of classification result
e = exponential constant

z = linear combination of input variables

In this study, Logistic Regression was used as a baseline model before hyperparameter optimization using GridSearchCV. The baseline model was built using default parameters from the scikit-learn library to determine the model's initial performance in predicting customer churn.

Proceeding of International Conference on Advanced Technologies, Innovations and Science
Vol 1, No 1, June 2026, Page 537-545
ISSN XXXX-XXXX (Media Online)
https://journals.adaresearch.or.id/icatis

2.5 Hyperparameter Optimization with GridSearchCV

Hyperparameter optimization was performed to improve the performance of the Logistic Regression model in predicting customer churn using the GridSearchCV method (Ehsani & Hosseini, 2024a). This method systematically searches for the best parameter combination through a cross-validation process by testing several parameter values such as C, solver, max_iter, and class_weight. These parameters were chosen because they influence the regularization process, optimization algorithm, the number of training iterations, and handling of class distribution imbalances in the dataset (Bhatnagar & Srivastava, 2025). The best parameter combination was determined based on the highest evaluation value obtained during the cross-validation process, resulting in a more stable model with a lower risk of overfitting. The optimized model was then compared with the baseline model to analyze the effect of hyperparameter optimization on customer churn prediction performance.

2.6 Model Evaluation

Model evaluation was conducted to measure the performance of the Logistic Regression model in predicting customer churn (Chajia, 2024). In this study, evaluations were conducted on the baseline model and the hyperparameter optimization model using several evaluation metrics, namely accuracy, precision, recall, F1-score, confusion matrix, ROC curve, and AUC score. Accuracy is used to measure the model's accuracy in classifying data. The accuracy equation is shown in Equation (2).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

Description:
TP = True positive
TN = True negative
FP = False positive
FN = False negative

Precision is used to measure the precision of positive predictions produced by the model. The precision equation is shown in Equation (3).

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \tag{3}$$

Recall is used to measure the model's ability to detect positive data as a whole. The recall equation is shown in Equation (4).

$$Recall = \frac{\text{TP}}{\text{TP+FN}} \tag{4}$$

F1-score is used to measure the balance between precision and recall. The F1-score equation is shown in Equation (5).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

In addition to using accuracy, precision, recall, and F1-score metrics, this study also used a confusion matrix to assess the number of correct and incorrect predictions generated by the model (Maan & Maan, 2023). The ROC curve was used to analyze the model's ability to distinguish between positive and negative classes, while the AUC score was used to measure model quality based on the area under the ROC curve (Imani, 2023). The evaluation results of the baseline model and the hyperparameter optimization model were then compared to determine the effect of hyperparameter optimization on improving model performance in predicting customer churn.

3. RESULTS AND DISCUSSION

3.1 Data Preprocessing Results

The preprocessing stage was carried out to improve the quality of the dataset before the classification process was carried out. The dataset used consisted of 2993 customer data with 20 variables. The preprocessing stage included checking for missing values, removing data with incomplete values, removing duplicate data, encoding categorical data, and feature scaling. The difference in the amount of data between the initial dataset and the data used in the modeling stage occurred due to the preprocessing process, especially during the data cleaning stage. The initial dataset consisted of 2993 customer data, while after removing duplicate data and data with incomplete values (missing values), the number of data used became 2988 data.

Categorical attributes were converted to numeric form using encoding techniques so that they could be processed by the Logistic Regression algorithm. Next, feature scaling was performed using StandardScaler to normalize the range of numeric attribute values so that the model training process could run more optimally. After preprocessing was completed, the data was divided into training data and testing data with a ratio of 80:20.

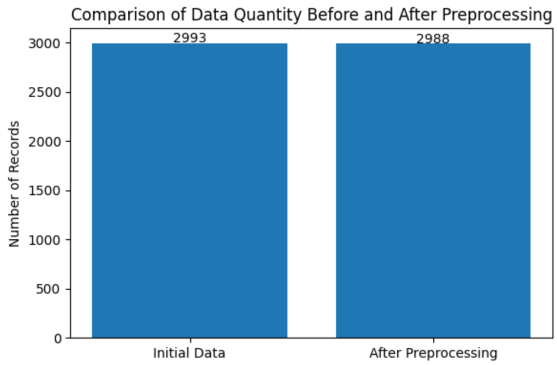

Figure 2. Comparison of the amount of data before and after preprocessing

Based on Figure 2, the number of data points decreased from 2,993 to 2,988 after preprocessing. This reduction was due to the removal of missing values, resulting in better quality data used in the modeling phase.

3.2 Data Sharing Results

Data splitting was performed to separate the model training and testing processes so that performance evaluation could be conducted objectively. The cleaned dataset was then divided into training and testing data using a train-test split method with a ratio of 80:20. Data splitting was performed to ensure the model could be trained using the training data and tested using the testing data, allowing for objective evaluation of model performance. The training data was used in the Logistic Regression model development process, while the testing data was used to measure the model's ability to predict data that had not been used in the training process. In this study, data splitting was performed using a stratified split method to maintain a balanced distribution of churned and non-churned classes in both the training and testing data.

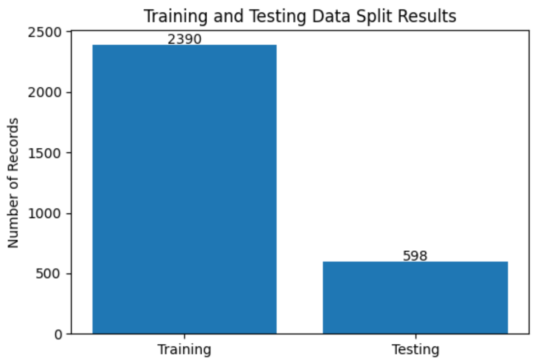

Figure 3. Results of dividing training and testing data

Based on Figure 4, the dataset that has gone through the preprocessing stage is divided into two parts, namely training data and testing data using the train-test split method with a ratio of 80:20. The results show that 2390 data points are used as training data and 598 data points are used as testing data. The training data is used for the Logistic Regression model training process so that the model can learn patterns from the data, while the testing data is used to test the model's ability to make predictions on data that has never been used in the training process. The data division is carried out using a stratified split technique so that the proportion of churned and non-churn classes in the training and testing data is maintained and representative of the original data distribution.

3.3 Baseline Logistic Regression Model Results

The baseline model is used as an initial reference to determine the basic performance of the Logistic Regression algorithm before the hyperparameter optimization process using GridSearchCV is carried out. At this stage, the model is built using the default parameters from the scikit-learn library without any additional parameter adjustments. The model training process is carried out using the training data from the previous dataset division, while the testing process is carried out using the testing data. The baseline model evaluation is carried out using several evaluation metrics, namely accuracy, precision, recall, and F1-score to determine the model's ability to classify customer churn.

Table 3. Baseline Logistic Regression Model Evaluation Results.

Evaluation Metrics	Before Optimization
Accuracy	81.61%
Precision	68.25%
Recall	55.13%
F1-score	60.99%

Proceeding of International Conference on Advanced Technologies, Innovations and Science
Vol 1, No 1, June 2026, Page 537-545
ISSN XXXX-XXXX (Media Online)
https://journals.adaresearch.or.id/icatis

Based on Table 3, the baseline Logistic Regression model was able to classify customer churn with an accuracy value of 81.61%. This value indicates that the model was able to make predictions with a high level of accuracy, demonstrating good classification performance across the entire test data. The precision value of 68.25% indicates that most of the customer churn predictions generated by the model were included in the correct classification. Meanwhile, the recall value of 55.13% indicates that the model still has limitations in detecting all customers who actually churned. Furthermore, the F1-score value of 60.99% indicates a balance in performance between precision and recall in the baseline model. The evaluation results were then used as a reference for comparison with the hyperparameter optimization model using GridSearchCV to determine the effect of optimization on improving the performance of the customer churn classification model.

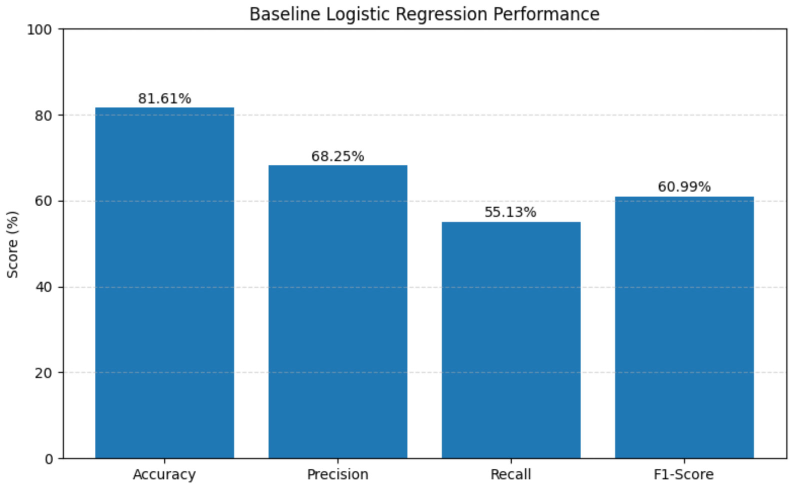

Figure 4. Baseline Logistic Regression Model Evaluation Results

According to Figure 4, the Baseline Logistic Regression model demonstrated fairly good classification performance. The highest accuracy value indicates that the model performed well in overall predictions. However, the recall value, which was still lower than other metrics, indicates that the model was not yet able to optimally detect all customers who were likely to churn. Therefore, hyperparameter optimization using GridSearchCV was performed to improve the model's ability to identify churned customers.

3.4 Hyperparameter Optimization Results with GridSearchCV

After evaluating the baseline Logistic Regression model, the next step is to optimize hyperparameters using the GridSearchCV method. Optimization is performed to obtain the best parameter combination that can improve the model's performance in predicting customer churn. In this study, the optimization process was carried out by testing several parameter combinations in the Logistic Regression model, namely the C parameter, solver, max_iter, and class_weight. The process of finding the best parameters was carried out using cross-validation techniques to obtain the parameter combination with the best evaluation score. The results of the hyperparameter optimization produced the best parameter combination used to build the optimized Logistic Regression model. The best parameters obtained from the GridSearchCV process are shown in Table 4.

Table 4. Best GridSearchCV Parameter Results

Parameter	Best Value
C	1
solver	liblinear
max_iter	1000
class_weight	balanced

The optimized model was then tested using the same testing data as the baseline model. The evaluation was conducted using accuracy, precision, recall, and F1-score metrics to determine the effect of hyperparameter optimization on model performance. The results of the optimization model evaluation were then compared with the baseline model to determine the performance improvements resulting from hyperparameter optimization.

3.5 Analysis of Hyperparameter Optimization Results

Classification analysis was performed to evaluate the performance of the Logistic Regression model in predicting customer churn. In this study, a confusion matrix and ROC curve were used to compare the performance of the baseline model with the hyperparameter-optimized model using GridSearchCV. This analysis was used to determine the model's ability to distinguish between churned and non-churned customers and to evaluate the effect of hyperparameter optimization on classification performance.

Proceeding of International Conference on Advanced Technologies, Innovations and Science
Vol 1, No 1, June 2026, Page 537-545
ISSN XXXX-XXXX (Media Online)
https://journals.adaresearch.or.id/icatis

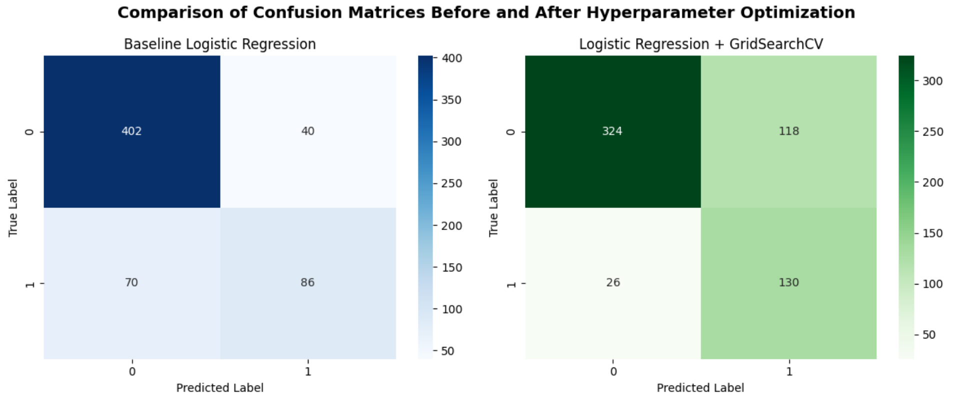

Figure 5. Comparison of confusion matrix before and after optimization

According to Figure 5, the GridSearchCV optimized model shows an improved ability to detect customer churning. The number of True Positives increased from 86 to 130 data points, and False Negatives decreased from 70 to 26 data points. This is in line with the increase in recall from 55.13% to 83.33%. Despite the increase in False Positives, the optimization results indicate that the model is more effective in identifying customers who are likely to churn than the baseline model.

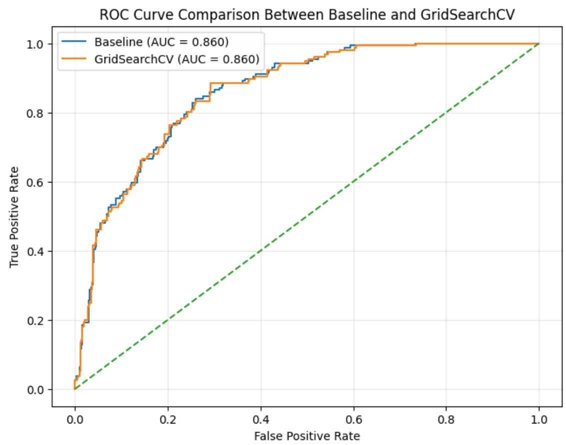

Figure 6. ROC curve Comparison Between Baseline and GridSearchCV

Based on Figure 6, the baseline Logistic Regression model and the optimized GridSearchCV model have the same AUC value, which is 0.860. This value indicates that both models have good classification ability in distinguishing churned and non-churn customers. Although the AUC value did not increase, the confusion matrix results indicate that the optimized model has better ability to detect churned customers, which is reflected in the increase in recall value. Overall, the results of the confusion matrix and ROC curve analysis indicate that hyperparameter optimization using GridSearchCV has an impact on the performance of the Logistic Regression model. Although the accuracy and precision values decreased, the increase in recall and F1-score values indicates that the optimized model is more effective in identifying customers who have the potential to churn. In the case of customer churn prediction, the ability to detect customers who are at risk of churning is an important aspect, so the optimized model is considered more suitable for use.

3.6 Comparison of Baseline Model and Optimization Model Performance

After evaluating the baseline model and the hyperparameter optimization model using GridSearchCV, the next step was to compare the performance of the two models based on the evaluation metrics used. This comparison was conducted to determine the effect of hyperparameter optimization on improving the performance of the Logistic Regression model in predicting customer churn. The results of the comparison of the performance of the baseline and optimization models are shown in Table 5.

Table 5. Comparison of Baseline Model and Optimization Model Performance

Evaluation Metrics	Baseline	GridSearchCV Optimization
Accuracy	81.61%	75.59%
Precision	68.25%	52.05%
Recall	55.13%	81.41%
F1-score	60.99%	63.50%
AUC score	86.00%	86.00%

Based on Table 5, the results of hyperparameter optimization using GridSearchCV resulted in changes to the performance of the Logistic Regression model. The recall value increased significantly from 55.13% to 81.41%, indicating that the optimized model was better able to detect customers who actually churned. However, the increase in recall was accompanied by a decrease in accuracy and precision. Accuracy decreased from 81.61% to 75.59%, while precision decreased from 68.25% to 52.05%. This indicates that the optimized model produced more false positive predictions. Furthermore, the F1-score increased from 60.99% to 63.50%, indicating that the optimized model had a better performance balance between precision and recall compared to the baseline model.

The baseline model performed better on accuracy and precision metrics. However, an increase in recall from 55.13% to 81.41% and F1-score from 60.99% to 63.50% was achieved after hyperparameter optimization using GridSearchCV. This improvement indicates that the model's ability to identify customers with the potential to churn and reduce the number of undetected churned customers has improved. Therefore, despite the decrease in accuracy and precision, the optimized model is considered more suitable for customer churn prediction cases that prioritize the ability to detect customers at risk of churning.

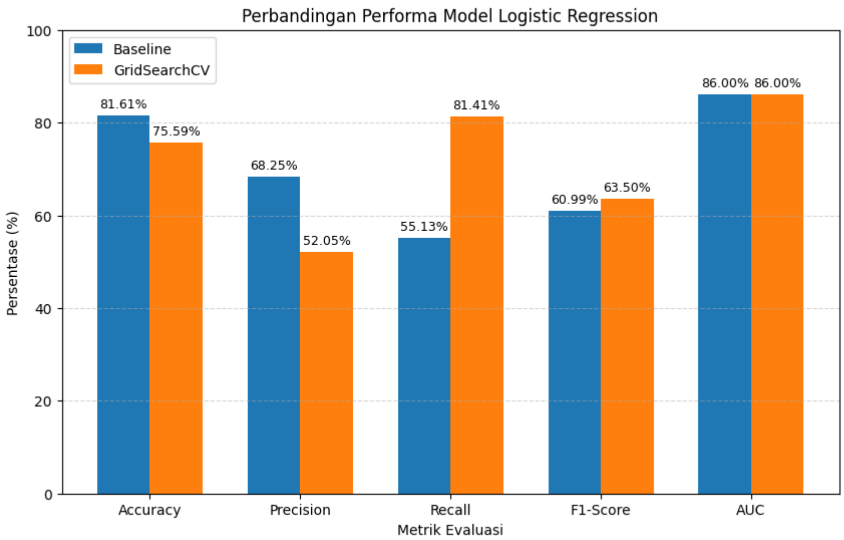

Figure 7. Comparison of the Performance of the Baseline Model and the GridSearchCV Optimization Model

Based on Figure 7, it can be seen that the optimized model experienced a significant increase in recall and F1-score values compared to the baseline model. However, the accuracy and precision values decreased after hyperparameter optimization. The AUC values for both models showed relatively similar results, so hyperparameter optimization had a greater effect on the model's ability to detect customer churn than on improving the model's overall discrimination ability. Overall, the optimized Logistic Regression model using GridSearchCV was considered more effective in predicting customer churn because it had a higher churn detection ability than the baseline model.

3.7 Discussion of Research Results

Based on the results of the research that has been conducted, the application of hyperparameter optimization using GridSearchCV on the Logistic Regression model has an impact on the performance of customer churn classification. The optimization process is carried out by finding the best combination of parameters in the Logistic Regression model so that the model can produce more optimal classification performance compared to the baseline model. The evaluation results show that the baseline model has an accuracy value of 81.61%, a precision of 68.25%, a recall of 55.13%, and an F1-score of 60.99%. After hyperparameter optimization using GridSearchCV, the recall value increased to 81.41% and the F1-score increased to 63.50%. This increase indicates that the optimized model has a better ability to detect customers who have the potential to churn. Although the accuracy and precision values decreased after the optimization process, the increase in recall is one of the important results in the case of customer churn prediction.

This is because the main goal of churn prediction is to identify as many customers as possible who are likely to stop using the company's services. By increasing the recall value, the number of churned customers that the model fails to detect can be reduced. The confusion matrix results also show that the optimized model is able to reduce the number of false negatives compared to the baseline model. The decrease in false negatives indicates that the model is more sensitive in detecting churned customers, allowing the company to implement customer retention strategies more quickly and accurately. Based on the ROC curve analysis, the baseline model and the optimized model have relatively similar AUC values, namely 0.860. This value indicates that both models have good classification capabilities in distinguishing churned and non-churn customers. However, hyperparameter optimization still provides an increase in churn detection capabilities, as reflected in the increase in recall and F1-score values. Thus, the model's ability to detect customers who are likely to churn is improved through hyperparameter optimization using GridSearchCV, despite a decrease in accuracy and precision. The results of this study indicate that optimal parameter selection contributes to improving the performance of the machine learning model.

4. CONCLUSION

Based on the research results, the Logistic Regression algorithm can be used to classify customer churn in the Telco Customer Churn dataset. Hyperparameter optimization using GridSearchCV has an impact on model performance, especially in improving the ability to detect customers who have the potential to churn. Although the baseline model produces higher accuracy and precision values, hyperparameter optimization using GridSearchCV successfully increased the recall value from 55.13% to 81.41% and the F1-score value from 60.99% to 63.50%. The increase in recall value indicates that the model's ability to identify customers who have the potential to churn has improved, so that the number of undetected churned customers can be reduced. Therefore, the optimized model is considered more suitable for implementing customer churn predictions oriented towards customer retention strategies. Further research can be conducted by comparing other classification algorithms and applying data imbalance handling techniques to obtain more optimal model performance.

ACKNOWLEDGMENT

The authors would like to thank STIKOM Tunas Bangsa, especially the Information Systems Study Program, for the support provided in this research

REFERENCES

Alamsyah, N., Setiana, E., & Fahmi, F. M. (2025). *Optimizing Logistic Regression for Digital Marketing Campaigns : Insights from Hyperparameter Tuning with Optuna*. *4*(1), 35–50.

Alotaibi, M. Z. (2024). *Customer Churn Prediction for Telecommunication Companies using Machine Learning and Ensemble Methodshttps*. *14*(3), 14572–14578. https://doi.org/https://doi.org/10.48084/etasr.7480

Arivazhagan, B., & Subramanian, R. S. S. (2020). *Customer Churn Prediction using Logistic Regression with Regularization and Optimization Technique*. *3075*(9), 334–339. https://doi.org/10.35940/ijitee.I7219.079920

Autom, A., & Geoespaciales, D. (2025). *Machine Learning Models Based on Geospatial Data for Customer Churn Prediction in ISP*. *9*, 1–32.

Bhatnagar, A., & Srivastava, S. (2025). *Customer Churn Prediction using Machine Learning Approach : A Comprehensive Study*. *10*. https://doi.org/10.52783/jisem.v10i25s.3944

Chajia, M. (2024). *Customer Churn Prediction Approach Based on LLM Embeddings and Logistic Regression*. https://doi.org/https://doi.org/10.3390/fi16120453

Ehsani, F. (2022). *Customer churn prediction from Internet banking transactions data using an ensemble meta- classi er algorithm*.

Ehsani, F., & Hosseini, M. (2024a). *Customer churn analysis using feature optimization methods and tree-based classifiers*. *April*. https://doi.org/10.1108/JSM-04-2024-0156

Ehsani, F., & Hosseini, M. (2024b). *Customer purchase prediction in electronic markets from clickstream data using the Oracle meta - classifier*.

Hemlata Jain, Ajay Khunteta, S. S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *PROCS*, *167*(2019), 101–112. https://doi.org/10.1016/j.procs.2020.03.187

Imani, M. (2023). *Hyperparameter Optimization and Combined Data Sampling Techniques in Machine Learning for Customer Churn Prediction : A Comparative Analysis*.

Iskandar, M. A., Latifa, U., Teknik, F., Karawang, U. S., & Churn, C. (2023). *Website Prediksi Customer Churn Untuk Mempertahankan Pelanggan Pada Perusahaan Telekomunikasi*. *7*(2), 1308–1316.

Kavitha, V. (2020). *Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms*. *9*(05), 181–184.

Kumari, D., Singh, S. K., Katira, S. S., & Srinivas, I. V. (2025). *Telecom Customer Churn Forecasting Using Machine Learning : A Data- Driven Predictive Framework*.

Loukili, M., Messaoudi, F., & Ghazi, M. El. (2022). *Supervised Learning Algorithms for Predicting Customer Churn with Hyperparameter Optimization*. *14*(3). https://doi.org/10.15849/IJASCA.221128.04

Maan, J., & Maan, H. (2023). *Customer Churn Prediction Model using Explainable Machine learning*. *11*(1), 33–38.

Maisat, Z., Darmawan, E., & Fauzan, A. (2023). *Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM*. *13*(1), 8–15.

Pan, H. (2025). *Logistic Regression for Customer Churn Analysis : The Role of Data Proportion ans Class Balancing in Model Performance*. *Vol. 1*(5), 1–4. https://doi.org/https://doi.org/10.61173/x28wbd81

Preet, P., Islam, F., Senapati, R., Sinha, A., & Sakib, N. (2024). Investigating customer churn in banking : a machine learning approach and visualization app for data science and management. *Data Science and Management*, *7*(1), 7–16. https://doi.org/10.1016/j.dsm.2023.09.002

Rathore, P. S. (2025). *Predicting Customer Churn in Telecom : A Review of Advanced ML-Based Methods and Applications*. *1*, 44–49.

Shrestha, S. M., & Shakya, A. (2023). ScienceDirect A Customer Customer Churn Churn Prediction Prediction Model Model using using XGBoost XGBoost for for the the Telecommunication Industry in Nepal Telecommunication Industry in Nepal. *Procedia Computer Science*, *215*, 652–661. https://doi.org/10.1016/j.procs.2022.12.067

Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization*, *14*(November 2023), 100342. https://doi.org/10.1016/j.rico.2023.100342