Applied Information System and Management (AISM)
Volume 7, (2) 2024, p. 31–38
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v7i2.39326

Struggling Models: An Analysis of Logistic Regression and Random Forest in Predicting Repeat Buyers with Imbalanced Performance Metrics

Siska Farizah Mauludiah[1*], Yunifa Miftachul Arif[2], Muhammad Faisal[3], Dony Darmawan Putra[4]

***Abstract*—Predicting repeat buyers is essential for businesses seeking to improve customer retention and maximize profitability. This study examines the effectiveness of logistic regression and random forest algorithms in forecasting repeat buyers, utilizing an e-commerce dataset from Kaggle. Despite the theoretical strengths of these models, our results indicate significant performance challenges. Both models were evaluated on key metrics: accuracy, precision, recall, F1 score, and ROC-AUC. The findings revealed that the models logistic regression and random forest performed poorly, with accuracy hovering around 50%, precision and recall demonstrating imbalanced performance, and ROC-AUC scores barely exceeding random guessing levels. Such metrics highlight the limited discriminative power of these models in identifying repeat buyers. The analysis suggests that issues such as data quality, feature relevance, and class imbalance contribute to these shortcomings. Specifically, the models struggled to effectively learn from the data, leading to suboptimal predictions. These results underscore the need for enhanced feature engineering, better handling of class imbalance, and possibly exploring more advanced algorithms. This study provides a critical assessment of the limitations inherent in using Logistic Regression and Random Forest for predicting repeat buyers, hence implements feature engineering, SMOTE and hyperparameter tuning using RandomSearchCV to get better result.**

***Index Terms*—E-commerce, repeat buyers, customer retention, logistic regression, random forest, imbalanced performance metrics.**

Received: 8 June 2024; Revised: 3 July 2024; Accepted: 24 July 2024.
*Corresponding author
[1]Siska Farizah Mauludiah, Universitas Islam Negeri Maulana Malik Ibrahim Malang Indonesia (e-mail: siskafm@yahoo.com).
[2]Yunifa Miftachul Arif, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia (e-mail: yunif4@ti.uin-malang.ac.id).
[3]Muhammad Faisal, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia (e-mail: mfaisal@ti.uin-malang.ac.id).
[4]Dony Darmawan Putra, Institute of Communication Engineering, National Sun Yat Sen University Taiwan (e-mail: d093070002@nsysu.edu.tw).

I. Introduction

Predicting repeat buyers is a critical task for businesses aiming to enhance customer retention and maximize lifetime value. Accurate predictions enable targeted marketing strategies, personalized offers, and improved customer satisfaction. However, developing effective predictive models is fraught with challenges, particularly when dealing with complex datasets and imbalanced classes [5]. This study explores the efficacy of logistic regression and random forest algorithms in predicting repeat buyers. Despite their popularity and theoretical robustness, our findings reveal that both models struggle to achieve satisfactory performance metrics.

Machine learning techniques have become important tools in doing predictive analytic, offering sophisticated methods to analyze complex datasets and uncover hidden patterns. Among these techniques, Logistic Regression and Random Forest stand out due to their widespread use and complementary strengths. Logistic regression (LR), a linear model, is renowned for its simplicity, interpretability, and efficiency. It provides clear insights into the factors influencing repeat purchases, making it a valuable tool for initial analysis and hypothesis generation.

On the other hand, random forest, an ensemble learning method, excels in handling non-linear relationships and interactions between features. By aggregating the prediction of multiple decision trees, random forest offers enhanced predictive performance and robustness against over-fitting. This makes it particularly suitable for more complex datasets where interactions among variables are intricate and non-linear [12].

Using an e-commerce dataset from Kaggle which has around 250.000 records, we assess the accuracy, the precision, the recall, the F1 Score, and the ROC-AUC of the models. The results highlight significant performance issues, with metrics barely surpassing random guessing. An accuracy of approximately 50%, coupled with a low ROC-AUC score, indicates the models' limited discriminative power. Moreover,

imbalanced precision and recall further underscore the difficulties in capturing true repeat buyers.

These findings prompt a deeper examination of the factors contributing to poor model performance, including data quality, feature relevance, and class imbalance. By identifying these challenges, this research's goal is to provide insights into the limitations of traditional machine learning approaches in this context and suggest potential pathways for improvement. The goal is to enhance the predictive capabilities of models used for repeat buyer identification, ultimately aiding businesses in more effectively leveraging their customer data.

Predicting repeat buyers in the context of e-commerce is a complex task, primarily due to the inherent imbalance in the dataset where repeat buyers are significantly fewer compared to one-time buyers. Initially, logistic regression, a widely used and interpretable model, is often employed for this classification task. Logistic regression, despite its simplicity, provides a robust baseline for comparison [12]. However, its linear nature can limit performance when dealing with complex, non-linear relationships prevalent in e-commerce data. The primary challenge with logistic regression in this scenario is its sensitivity to class imbalance, often resulting in biased performance metrics favoring the majority class.

To address the limitations of LR, random forest (RF) is introduced as an alternative. Random Forest, an ensemble learning method, leverages multiple decision trees to improve predictive accuracy and robustness. Its ability to model non-linear relationships and handle high-dimensional data makes it a strong candidate for this task. Furthermore, Random Forest's inherent feature importance mechanism aids in understanding the significant predictors of repeat buying behavior [8]. However, the default Random Forest model, like logistic regression, can still struggle with class imbalance, leading to suboptimal performance metrics such as precision, recall, and F1-score for the minority class.

Feature engineering becomes crucial at this stage, transforming raw data into meaningful features that better capture the underlying patterns of repeat buying behavior. This involves creating new variables, such as recency, frequency, and monetary (RFM) metrics, customer demographics, and interaction behaviors, which enhance the model's predictive power. Additionally, employing SMOTE (Synthetic Minority Over-sampling Technique) addresses class imbalance by generating synthetic samples for the minority class. SMOTE effectively mitigates the skewness in the dataset, enabling the model to learn more effectively from underrepresented instances and improving the balance in performance metrics [6].

The journey from LR to RF, augmented by feature engineering and SMOTE, culminates in extensive hyperparameter tuning to optimize the model's performance. RandomizedSearchCV, a robust hyperparameter tuning technique, is employed to systematically explore a wide range of hyperparameters and identify the best combination for the Random Forest model. This process involves random sampling of hyperparameter values, which is computationally efficient and effective in finding near-optimal solutions compared to exhaustive grid search methods. Through this rigorous approach, the RF model achieves improved predictive accuracy and balanced performance metrics, demonstrating its superiority in handling imbalanced classification problems in predicting repeat buyers.

The integration of advanced techniques such as feature engineering, SMOTE, and hyperparameter tuning using RandomizedSearchCV marks a advancement in predictive modeling for repeat buyers. This methodology not only enhances model performance but also provides deeper insights into customer behavior, enabling businesses to devise targeted strategies for customer retention and engagement.

II. Related Work

Previous research has extensively explored the use of machine learning models to predict customer behaviors, particularly focusing on repeat purchases. Studies have shown that models such as logistic regression and random forest can effectively handle large and complex datasets, capturing intricate patterns in customer transactions. However, many researchers have also encountered significant challenges, including class imbalance and the variability of customer behavior, which often result in sub-optimal model performance. These issues necessitate further investigation and refinement of predictive techniques to enhance accuracy and reliability in identifying repeat buyers.

Prior researches had explored various dimensions of repeat buying, including the determinants of customer loyalty, the impact of customer satisfaction on repurchase intentions, and the predictive models that forecast repeat purchases. A study by Zhang and Dong proposed a better model utilizing a combination of the DeepCatboost, DeepGBM, and double attention BiGRU (DABiGRU) individual models with vote-stacking technique to model the discrete purchase records and historical behavior sequence data in order to increase the accuracy of repeat buyer prediction [2]. There was also a study to improve the prediction of repeat buyers and identify customers with the potential to return for more purchases, a comprehensive solution porposed by Liu and Song introduced the basic principles of the XGBoost algorithm, analyzing historical data from an e-commerce platform, and preprocessing the original data to construct a consumer purchase prediction model based on XGBoost [3]. In the other hand, Liu et al. discussed the drawbacks and difficulties of conventional online purchasing behavior prediction techniques by proposing an advanced online shopping behavior analysis and prediction system, offering a more accurate and robust framework for understanding and anticipating customer purchasing habits using logistic regression and decision tree based model [4].

Another research by Zang and Wang suggested a better deep forest model, and the original feature model was enhanced by the addition of the items' and user's interaction activity characteristics to forecast the e-commerce customers' repurchase behavior. The experiment shown that the model performs better overall and has higher accuracy when there are more interactive behavior features [5]. Meanwhile, a BERT-MLP prediction model is proposed by Dong et al., leveraging "large-scale data unsupervised pre-training"

Applied Information System and Management (AISM)
Volume 7, (2) 2024, p. 31–38
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v7i2.39326

followed by "fine-tuning with a small amount of labeled data." This innovative combination aims to optimize the model's ability to generalize from large datasets and refine its predictions with precise, labeled examples, ultimately leading to more accurate and reliable predictions of customer behavior [7]. A study by Suhanda et al., ascertained and analyzed client happiness, customer loyalty and customer trust, in order to make company monitoring of the customers easier using random forest [8]. According to Kuric et al.., there are a number of challenges in analyzing low-level interaction data that records atomic behavioral events like e-commerce users' clicks, scrolling, and motions in relation to the prediction of repeat buyer [12]. And collecting and integrating historical sales data from various e-commerce platforms is complex, requiring meticulous preprocessing to ensure data quality. This involves addressing issues like data heterogeneity, inconsistencies, and noise, which can significantly impact the development of an accurate and reliable prediction model, based on a study by Kc et al.. [13]. An alternative method was an ensemble strategy user by De et al. to more accurately anticipate client order patterns by combining XGBoost with a modified version of poisson gamma model.

Regarding imbalanced performance issues, some researches conducted investigation about it. A research conducted by Jeni et al. introduced modifications of the F-score and the MCC to robust performance metrics [1]. There was also a method introduced by Zang et al. to solve data imbalance issue and to improve the prediction performance using synthetic minority oversampling technique (SMOTE). The result showed that the are under curve improved by 0.01161 trough data imbalance and feature engineering [6]. SMOTE was also implemented by Halim et al. in their research regarding classification of imbalanced data using RF [16]. Meanwhile, Riyanto et al. implemented some methods to compare various evaluation metrics on imbalanced data using multinomial naive bayes, k-neirest neighbors, support vector machine, random forest and long short-term memory [9]. Owusu-Adjei et al.. also tried to determine the optimum prediction solution for performance metrics [10], while Hancock explored how to solve the gap by analyzing it on three big data classification tasks [11]. And a study by Hozmann and Klar introduced how to robust performance metrics using modifications of the F1 Score and the Matthews' Correlation Coefficient (MCC) [15].

As mentioned previously that there are challenges imbalanced performance metrics. In this study, the challenge also appear when working on repeat buyer case. Previous researches started from methods that can be used to find the best way to handle the imbalanced performance metrics. But in this study started from facing the unexpected performance metrics results while analyzing repeat buyer study. Therefore, encouraged us to try to find out the cause of performance metrics issue and how to improve the result.

III. Research Method

To investigate the predictive performance of logistic regression and random forest models in identifying repeat buyers, we employed a structured research methodology encompassing data collection, preprocessing, model implementation, and evaluation. The dataset utilized in this study comprised customer purchase histories as we can see in Fig. 1 and Fig. 2, enriched with features such as transaction frequency, purchase amounts, and time intervals between purchases. We focused on addressing common issues in predictive modeling, such as class imbalance and feature relevance, to ensure robust model training and assessment. By systematically applying LR and RF models and rigorously evaluating their performance metrics, we aimed to uncover the inherent challenges and limitation faced by these algorithms in the context of repeat buyer prediction.

Customer ID	Purchase Date	Product Category	Product Price	Quantity	Total Purchase Amount	Payment Method
46251	2020-09-08 09:38:32	Electronics	12	3	740	Credit Card
46251	2022-03-05 12:56:35	Home	468	4	2739	PayPal
46251	2022-05-23 18:18:01	Home	288	2	3196	PayPal
46251	2020-11-12 13:13:29	Clothing	196	1	3509	PayPal
13593	2020-11-27 17:55:11	Home	449	1	3452	Credit Card

Fig. 1. First part e-commerce dataset features.

Payment Method	Customer Age	Returns	Customer Name	Age	Gender	Churn
Credit Card	37.0	0.0	Christine Hernandez	37.0	Male	0.0
PayPal	37.0	0.0	Christine Hernandez	37.0	Male	0.0
PayPal	37.0	0.0	Christine Hernandez	37.0	Male	0.0
PayPal	37.0	0.0	Christine Hernandez	37.0	Male	0.0
Credit Card	49.0	0.0	James Grant	49.0	Female	1.0

Fig. 2. Second part e-commerce dataset features.

A. Repeat Buyer

Repeat buyers represent a cornerstone of sustainable business growth in the e-commerce landscape. These valuable customers are characterized by their tendency to make multiple purchases from the same seller or platform over time. Unlike one-time purchasers, repeat buyers have already demonstrated a level of trust and satisfaction with the brand or product, making them more likely to engage in future transactions. Their behavior often signifies a deeper level of brand loyalty and affinity, as well as a higher lifetime value to the business.

Understanding the motivations and preferences of repeat buyers is essential for businesses seeking to foster long-term relationships with their customer base. By analyzing patterns in their purchasing behavior, businesses can tailor marketing strategies, personalize product recommendations, and enhance the overall customer experience to incentivize further repeat purchases. Moreover, cultivating a loyal base of repeat buyers not only boosts revenue but also serves as a testament to the brand's reputation and customer satisfaction levels.

B. Logistic Regression

A statistical technique called logistic regression is used to simulate the likelihood of a binary outcome depending on one or more predictor factors. It is often employed in situations when there are two potential results for the dependent variable, such as "yes" or "no", "success" or "failure", or "0' or "1". The logistic function, often known as a sigmoid function, is the foundation of the model of logistic regression and has a value 0 to 1. It is hence appropriate for modelling probability.

Equation (1) expresses the formula of Logistic Regression:

$$P\,(Y = 1|X) = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\ldots+b_nX_n)}} \qquad (1)$$

where:
- ($P(Y=1|X)$ shows the likelihood that, given the values of the predictor variables (X), the dependent variable (Y), which is will equal 1.
- e is the the natural algorithm's base.
- b_0 which is referred to by intercept term
- $b_1, b_0, \ldots, b_n$ are the associated coefficients for every predictor variable $(X_1, X_2, \ldots, X_n)$.

In logistic regression, the coefficients $(b_0, b_1, \ldots, b_n)$ are estimated using a process called maximum likelihood estimation. The objectives is to identify the set of coefficients that, given the predictor factors, maximizes the probability of seeing the specified outcome. After the coefficients are computed, one can use them to forecast the likelihood that, with new data, the dependent variable will be 1. Typically, a threshold of 0.5 is used to classify observations into one of the two categories.

Logistic regression in marketing for repeat buyer analysis typically involves using historical customer data, which may include variables such as purchase frequency, total spend, engagement with marketing campaigns, demographics, and browsing behavior. These variables serve as predictors to estimate the likelihood of a customer becoming a repeat buyer.

To predict repeat buyer using Logistic Regression applies several steps as follows:

1) Data Collection: collect data on customer demographics, purchase history and other relevant information.

2) Data Preparation:
- Feature Engineering: create features that might influence repeat buyer behavior
- Labeling: define the target variable (1 for repeat buyers, 0 for non repeat buyers)
- Data Splitting: split the data into 20% for testing and 80% for training sets.
- Model Training: using the training set of data, train the Logistic Regression model.

C. Random Forest

One ensemble learning technique called random forest combines the predictions of several decision tress to increase accuracy and robustness. It generates a huge number of decision tress, or a forest of them, and trains them on a random selection of the characteristics and data before combining their predictions. Applying RF and leverage the historical customer data can help to get valuable insights for customer behavior.

Random forest has high effective algorithm to predict repeat buyers and able to handle large datasets with many features. Using feature selection and bootstrapping, a subset of the data is used to build each decision tree in the RF. Bootstrapping is taking a random sampling with replacement to create different training subsets. And feature selection is choosing a random subset of features at each split to determine the best split.

For a node *m,* select the split that minimizes the Gini Impurity or maximizes the information gain.

Gini Impurity:

$$Gini\,(D) = 1 - \sum_{i=1}^{C} P_i^2 \qquad (2)$$

where P_i is the probability of class *i* in dataset D.

Information Gain:

$$IG\,(D, A) = \sum_{v\epsilon Values(A)} \frac{|D_v|}{D} Entropy(D_v) \qquad (3)$$

where $Entropy(D) = -\sum_{i=1}^{C} P_i \log_2 (P_i)$

In a RF, every tress offers a prediction; in a classification instance, the majority vote determines the final prediction.

$$\hat{y} = mode(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n) \qquad (4)$$

where $\hat{y}$, is the prediction of *i*-th tree.

To predict repeat buyers using Random Forest involves several steps as follows:

1) Data Collection: collect data on customer demographics, purchase history and other relevant information.

2) Data Preparation:
- Feature Engineering: create features that might influence repeat buyer behavior
- Labeling: define the target variable (1 for repeat buyers, 0 for non repeat buyers)

3) Data Splitting: *split the data into 20% for testing and 80% for training sets.*

4) Model Training: use the training data to teach the Random Forest model. The number of tress (n_estimators) and the

Applied Information System and Management (AISM)
Volume 7, (2) 2024, p. 31–38
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v7i2.39326

maximum depth of each tree (max_depth) are two most important hyperparameters.

D. Data Preprocessing

A crucial component of machine learning is data processing, which transforms unstructured data into a clean, organized format fit for modelling. Proper data preprocessing makes the data more accurate, consistent and suitable to enhance model performance and reliability. In this study, there are several steps applies for data processing.

As we can see in Fig. 3, there are numbers of missing value. To fix the problem, several steps to clean the data as mentioned below need to be applied for data processing in order to make the data ready to use. Fig. 4 shows the data after preprocessing and the features that have been selected.

```
Customer ID                0
Purchase Date              0
Product Category           0
Product Price              0
Quantity                   0
Total Purchase Amount      0
Payment Method             0
Customer Age               1
Returns                32762
Customer Name              1
Age                        1
Gender                     1
Churn                      1
dtype: int64
```

Fig. 3. Numbers of missing values.

Data cleaning steps:

- Drop null values
- Checking missing values
- Dropping unnecessary columns
- Separate features and target

	Customer ID	Product Price	Quantity	Total Purchase Amount	Returns	Churn
0	46251	12	3	740	0.0	0.0
1	46251	468	4	2739	0.0	0.0
2	46251	288	2	3196	0.0	0.0
3	46251	196	1	3509	0.0	0.0
4	13593	449	1	3452	0.0	1.0

Fig. 4. Data after preprocessing.

E. Data Splitting

As shown bel, data splitting can be achieved by dividing the data into training sets (80%) and testing sets (20%). The model is fitted using a training set to identify underlying patterns in the data, and it modifies its parameters depending on this set to reduce prediction errors. The test set gives an objective appraisal of the model's capacity to generalize to new data and is used to measure the model's performance on unknown data. On this set, performance metrics are computed to determine the efficacy of the model.

Here is the code for splitting the data:

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test =
train_test_split(X,y,test_size=0.2, random_state=2)
print(x_train.shape)
print(x_test.shape)
print(y_train.shape)
print(y_test.shape)
```

F. Model Evaluation

Using the testing data, the models are assessed using typical classification metrics as accuracy, precision, recall, F1 score and ROC-AUC. In the context of classification tasks, evaluating model performance requires a comprehensive understanding of various metrics. Here, we explain the key metrics: accuracy, precision, recall, F1 score, and ROC-AUC, along with their equations.

1) Accuracy

Accuracy measures the proportion of correctly classified instances (both true positives and true negatives) out of the total instances. It is a general metric indicating overall performance, but it can be misleading in the presence of class imbalance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

where:
- TP = True Positives (correctly predicted positive instances)
- TN = True Negatives (correctly predicted negative instances)
- FP = False Positives (incorrectly predicted positive instances)
- FN = False Negatives (incorrectly predicted negative instances)

2) Precision

Precision, also known as positive predictive value, measures the proportion of true positive predictions out of all positive predictions. It indicates the accuracy of the positive class predictions.

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

3) Recall

Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions out of all actual positive instances. It indicates the model's ability to capture all relevant instances of the positive class.

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

4) F1 score

The F1 score is the harmonic mean of precision and recall. It provides a balanced measure, especially useful when dealing with imbalanced datasets, as it considers both false positives and false negatives.

$$F1\ Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \quad (8)$$

5) ROC-AUC

ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) measures the performance of a classification model at various threshold settings. The ROC curve plots the true positive rate (recall) against the false positive rate (FPR), which is defined as:

$$FPR = \frac{FP}{FP + TN} \quad (9)$$

The area under the curve (AUC) represents the degree of separability achieved by the model. A higher AUC indicates a better-performing model, with a value of 1 representing perfect classification and 0.5 representing a random guess.

$$A = \int_0^1 ROC\ (t)\ dt \quad (10)$$

These metrics collectively provide a comprehensive view of a model's performance. While accuracy is useful for general performance, precision, recall, and F1 score are crucial for evaluating models on imbalanced datasets. ROC-AUC further aids in assessing the discriminative power of the model across different thresholds.

IV. RESULT

After preprocessing the dataset, the first method that is implemented is Logistic Regression. Below codes show how to implement LR in Python.

```
from sklearn.linear_model import LogisticRegression
# Initialize the model
lr = LogisticRegression()
# Train the model
lr.fit(x_train, y_train)
```

The output from LR model evaluation that is shown in Table 1 indicates that the model is not performing well. The Accuracy and Precision are slightly better than random guessing which suggests that the accuracy of the model is not able to effectively differentiate between the two classes and it has high number of false positive in precision. Recall and F1 score are low and the ROC-AUC score indicates equivalent to random guessing.

Table 1.
The Performance Metrics of Logistic Regression

Accuracy	0.5009199466070204
Precision	0.5041222114451989
Recall	0.2990936555891239
F1 Score	0.3754401805869074
ROC-AUC	0.5054199951985573

The poor performance can happen because of some reasons. It can be because of imbalance issue, feature engineering or hyperparameters issue. To handle this, we try to implement RF as different model as we can see in the code below.

```
from sklearn.ensemble import RandomForestClassifier
# Initialize the model
rf = RandomForestClassifier(n_estimators=100,
max_depth=10, random_state=42)dy/dx
# Train the model
rf.fit(x_train, y_train)
```

The result of RF model's performance is also poor and quite similar with the previous model which the result can be seen in Table 2.

Table 2.
The Performance Metrics of Random Forest

Accuracy	0.5022547710956383
Precision	0.5057433541188053
Recall	0.33254208027621923
F1 Score	0.4012498372607733
ROC-AUC	0.502776793919683

To improve that model performance metrics, we explore the feature engineering by creating a new feature named `Days_Since_Last_Purchase` that might help the model to learn better patterns and restart to run LR model once again as shown below.

```
# Feature Engineering
df['Total_Spent'] = df['Product Price'] *
df['Quantity']
df['Days_Since_Last_Purchase'] =
(pd.to_datetime('today') -
pd.to_datetime(df['Purchase Date'])).dt.days
# Dropping the original columns after creating new
features
df = df.drop(['Product Price', 'Quantity', 'Purchase
Date'], axis=1)
```

And Fig.5 shows the complete features after creating a new feature Days_Since_Last_Purchase.

	Customer ID	Total Purchase Amount	Returns	Churn	Total_Spent	Days_Since_Last_Purchase
0	46251	740	0.0	0.0	36	1366
1	46251	2739	0.0	0.0	1872	823
2	46251	3196	0.0	0.0	576	743
3	46251	3509	0.0	0.0	196	1301
4	13593	3452	0.0	1.0	449	1285

Fig. 5. Features selection result.

After running the LR with the new features, the result shows that the feature engineering still indicates minimal improvement of the performance metrics of the LR model. The accuracy 0.4985 still slightly worse than random guessing (0.5) for a binary classification problem) and the model is not effectively differentiating between the two classes. The precision is 0.5 which means a high number of false positive and is equivalent to random guessing. Recall 0.2772 and F1

Applied Information System and Management (AISM)
Volume 7, (2) 2024, p. 31–38
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v7i2.39326

score 0.3566 are still remain low and the ROC-AUC is exactly 0.5 that indicates a model with no discriminative ability (equivalent to random guessing).

Another approach is improving the performance after creating a new feature, which are by doing resampling using syntetic minority over-sampling technique (SMOTE) and also doing the hyperparameter tuning using RandomizedSearchCV to find the best model parameter on Random Forest model. As the result as we can see in Table 3, the performance metrics from random forest model after SMOTE indicate some minor improvement in recall and F1 score, but the model still exhibits poor overall performance.

Table 3.
The Performance Metric With Smote and Hyperparameter Tuning

Accuracy	0.5022547710956383
Precision	0.50646682897139398
Recall	0.47662346960693064
F1 Score	0.4910921766072812
ROC-AUC	0.5034015190714021
Fitting 3 folds for each 0f 50 candidates, totalling 150 fits	
Best Model Accuracy	0.49363252642591726
Best Model Precision	0.4973418599275753
Best Model Recall	0.46216080761795664
Best Model F1 Score	0.4791063608698879
Best Model ROC-AUC	0.4951199584427234

The accuracy 0.5023 is only slightly better than random guessing which means the model is not significantly better at differentiating between the two classes. Precision 0.5065 is slightly better than random guessing that indicates a relatively high number of false positives. Recall, which measures the proportion of true positive predictions among all actual positive shows some improvement to 0.4766, indicating that the model is capturing more actual positive cases compared to previous metrics but is still not ideal. The F1 Score 0.491, which is the harmonic mean of Precision and Recall, reflects a better balance between Precision and Recall compared to previous attempts, though it is still not high. ROC-AUC Score, which assesses how well the model can differentiate between the positive and negative classes, is somewhat higher than 0.5, indicates a model with minimal discriminative ability, only marginally better than random guessing. And after hyperparameter tuning, the best model's performance metrics does not show significant improvement as can be seen in Table 4.

Table 4.
Complete Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score	ROA-AUC
Logistic Regression	0.5	0.504	0.299	0.37	0.505
Random Forest	0.502	0.505	0.332	0.401	0.502
Logistic Regression (New features)	0.4985	0.5	0.277	0.356	0.5
Random Forest (New features+SMOTE + RandomSearchCV	0.502	0.506	0.476	0.491	0.503
Best Model Random Forest (New features + SMOTE + RandomizeSearchCV	0.493	0.497	0.462	0.479	0.495

And the improvement chart of the whole performance metrics is shown in Fig. 6.

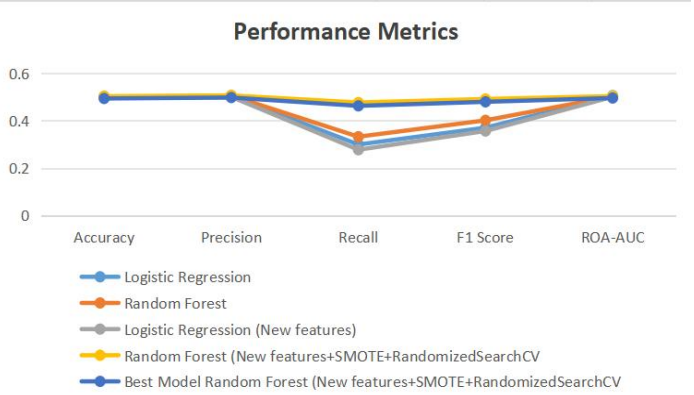

Fig. 6. Performane metrics chart.

V. CONCLUSION

Despite employing feature engineering, SMOTE (Synthetic Minority Over-sampling Technique), and extensive hyperparameter tuning using RandomizedSearchCV, both Logistic Regression and Random Forest models failed to deliver satisfactory performance in predicting repeat buyers. The performance metrics for Logistic Regression showed an accuracy of 0.4985, precision of 0.5, recall of 0.2772, F1 score of 0.3566, and ROC-AUC of 0.5000, indicating a model that performs no better than random guessing. Similarly, the Random Forest model, with the accuracy of 0.5023, the Precision of 0.5065, the Recall of 0.4766, the F1 score of 0.4911, and the ROC-AUC of 0.5034, demonstrated minimal improvement and still exhibited poor discriminative power.

From this result we can found that for the Accuracy of both models are very close to random guessing (0.5), but the Random Forest has a slightly higher accuracy even the difference is minimal. For the Precision, both models have very similar precision, indicating that around 50% of the positive predictions are correct and the difference is negligible. The Recall is higher, indicating it correctly identifies more actual positive cases compared to Logistic Regression. The Random Forest has a higher F1 score, suggesting a better balance between precision and recall compared to Logistic Regression. For both models, they have ROC-AUC values very to 0.5, indicating their ability to distinguish between classes is nearly random. The Logistic Regression has a slightly higher ROC-AUC, but the difference is very small. These results underscore the challenges faced when working with imbalanced datasets and highlight the limitations of these models in their current configurations.

The analysis suggests that further steps are necessary to improve model performance significantly. Advanced feature engineering, exploring more complex models such as gradient boosting machines or neural networks, and employing ensemble methods may offer potential solutions. Additionally, refining hyperparameter tuning and employing robust cross-validation techniques could lead to better outcomes. While SMOTE and RandomizedSearchCV provided some incremental benefits, the overall effectiveness of Logistic Regression and Random Forest in this context remains limited. This study underscores the importance of continuous experimentation and adaptation when dealing with imbalanced data and predictive modeling challenges.

REFERENCES

[1] L. A. Jeni, J. F. Cohn, and F. de La Torre, "Facing imbalanced data - Recommendations for the use of performance metrics," in *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 2013, pp. 245–251. doi: 10.1109/ACII.2013.47.

[2] H. Zhang and J. Dong, "Prediction of repeat customers on e-commerce platform based on blockchain," *Wireless Communications and Mobile Computing*, vol. 2020, no. 1, pp. 1-15, 2020, doi: 10.1155/2020/8841437.

[3] P. Song and Y. Liu, "An xgboost algorithm for predicting purchasing behaviour on e-commerce platforms," *Tehnicki Vjesnik*, vol. 27, no. 5, pp. 1467–1471, Oct. 2020, doi: 10.17559/TV-20200808113807.

[4] C. J. Liu, T. S. Huang, P. T. Ho, J. C. Huang, and C. T. Hsieh, "Machine learning-based e-commerce platform repurchase customer prediction model," *PLoS ONE*, vol. 15, no. 12, Dec. 2020, Art. no. e0243105.

[5] W. Zhang and M. Wang, "An improved deep forest model for prediction of e-commerce consumers' repurchase behavior," *PLoS ONE*, vol. 16, no. 9, Sep. 2021, Art. no. e0255906.

[6] M. Zhang, J. Lu, N. Ma, T. C. E. Cheng, and G. Hua, "A feature engineering and ensemble learning based approach for repeated buyers prediction," *International Journal of Computers, Communications and Control*, vol. 17, no. 6, 2022, Art. no. 4988.

[7] J. Dong, T. Huang, M. Liang, and W. Wang, "Prediction of online consumers' repeat purchase behavior via BERT-MLP model," *Journal of Electronic Research and Application*, vol. 6, no. 3, pp. 12–19, 2022, doi: 10.26689/jera.v6i3.4010.

[8] Y. Suhanda, L. Nurlaela, I. Kurniati, A. Dharmalau, and I. Rosita, "Predictive analysis of customer retention using the random forest algorithm," *TIERS Information Technology Journal*, vol. 3, no. 1, pp. 35–47, Jun. 2022, doi: 10.38043/tiers.v3i1.3616.

[9] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative analysis using various performance metrics in imbalanced data for multi-class text classification," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 6, pp. 1082–1090, 2023, doi: 10.14569/IJACSA.2023.01406116.

[10] M. Owusu-Adjei, J. ben Hayfron-Acquah, T. Frimpong, and G. Abdul-Salaam, "Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems," *PLOS Digital Health*, vol. 2, no. 11, Nov. 2023, Art. no. e0000290.

[11] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced Big Data," *Journal of Big Data*, vol. 10, no. 1, Art. no. 42, Dec. 2023, doi: 10.1186/s40537-023-00724-5.

[12] E. Kuric, A. Puskas, P. Demcak, and D. Mensatorisova, "Effect of low-level interaction data in repeat purchase prediction task," *International Journal of Human-Computer Interaction*, vol. 40, no. 10, pp. 2515–2533, 2024, doi: 10.1080/10447318.2023.2175973.

[13] R. Kc, S. Shandilya, and M. Shandilya, "Unlocking Future Transactions: Predicting Customer's Next Purchase in E-commerce through Machine Learning Analysis," *IJARIIE*, vol. 9, no. 3, pp. 1077–1081.

[14] T. S. De, P. Singh, and A. Patel, "A Machine learning and Empirical Bayesian Approach for Predictive Buying in B2B E-commerce," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jan. 2024, pp. 17–24. doi: 10.1145/3647750.3647754.

[15] H. Holzmann and B. Klar, "Robust performance metrics for imbalanced classification problems," Apr. 2024, [Online]. Available: arXiv:2404.07661.

[16] A. M. Halim, M. Dwifebri, and F. Nhita, "Handling Imbalanced Data Sets Using SMOTE and ADASYN to Improve Classification Performance of Ecoli Data Sets," *Building of Informatics, Technology and Science (BITS)*, vol. 5. no. 1, pp. 246−253, Jun. 2023, doi: 10.47065/bits.v5i1.3647.