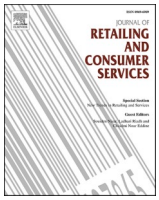
Journal of Retailing and Consumer Services 79 (2024) 103865

Contents lists available at ScienceDirect

Journal of Retailing and Consumer Services

journal homepage: www.elsevier.com/locate/jretconser



Predicting E-commerce customer satisfaction: Traditional machine learning vs. deep learning approaches

Maha Zaghloul *, Sherif Barakat, Amira Rezk

Department of Information System, Faculty of Computer and Information, Mansoura, P.O.35516, Egypt

ARTICLE INFO

Handling Editor: Prof. H. Timmermans

Keywords:
Customer satisfaction
Deep learning
E-commerce
Machine learning

ABSTRACT

The rapid growth of e-commerce has increased the need for retailers to understand and predict customer satisfaction to support data-driven managerial decisions. This study analyzes online consumer behavior through a comparative machine learning modeling approach to forecast future customer satisfaction based on review ratings. Using a large dataset of over 100 k online orders from a major retailer, traditional machine learning models including random forest and support vector machines are benchmarked against deep learning techniques like multi-layer perceptrons. The predictive models are assessed for their ability to accurately predict customer satisfaction scores for the next orders based on key e-commerce features including delivery time, order value, and location. The findings demonstrate that the random forest model can predict future satisfaction with 92% accuracy, outperforming deep learning. The analysis further identifies core drivers of satisfaction such as delivery time and order accuracy. These insights enable retail managers to make targeted improvements, like optimizing logistics, to increase customer loyalty and revenue. This study provides a framework for leveraging predictive analytics and machine learning to unlock data-driven insights into online consumer behavior and satisfaction for superior retail decision-making. The focus on generalizable insights across a major retailer enhances the practical applicability of the machine learning approach for the retail sector.

1. Introduction

Retailers use customer satisfaction as one of the main ways to measure how well their business is doing. Different studies have shown that overall customer satisfaction is strongly linked to a company's profits (Bernhardt et al., 2000). Customer satisfaction is the factor that determines whether a company's product or service matches customer needs and is a metric that can provide business owners with the current status of the business, allowing them to enhance profits and reduce marketing expenses (Gómez et al., 2004). Consumer feedback may assist in examining previously unconsidered elements, such as delivery, safe packaging, professional and available customer service specialists, and an informative website. Asking clients for their opinions and respecting their feedback is the only way to make them feel important. Customers are made to feel valued when they are asked for their feedback (Supriyanto et al., 2021). Online customer reviews help people who are thinking about buying a product, business, or service figure out how good it is. Reports show that online customer reviews have a big effect on a lot of people's decisions about what to buy (Riaz et al., 2021). Also, companies may learn more about their clients and improve their services as a result of feedback left on review sites (Zhao et al., 2019). In the real world, there is a significant increase in the quantity of reviews. According to data from the review platform Yelp, the number of written reviews exceeded 233 million by the fourth quarter of 2021, and this number is still increasing (Mewada and Dewang, 2023). There are advantages and disadvantages for both businesses and consumers in the rapidly expanding pool of customer evaluations available on the Internet (Bilal et al., 2021). Information overload from the variety and volume of Internet customer evaluations makes it difficult for potential buyers with restricted cognitive ability to identify useful reviews. Real, favorable evaluations boost business credibility. A negative review can highlight a company's customer service issues and suggest improvements (Hu and Krishen, 2019) (Roetzel, 2019). There are three parts to a customer's shopping experience: before the sale, in the store, and after the sale (Terblanche, 2018). In the pre-sale phase, the client establishes goals for using the service. The consumer leaves the store with the purchased goods or services, which is the second phase in the sale process. In the last," post-sale," stage, the buyer assesses how their complaints and

* Corresponding author.
E-mail addresses: mahafouad@std.mans.edu.eg (M. Zaghloul), sheiib@mans.edu.eg (S. Barakat), amira_rezk@mans.edu.eg (A. Rezk).

https://doi.org/10.1016/j.jretconser.2024.103865
Received 4 March 2024; Received in revised form 26 March 2024; Accepted 15 April 2024
Available online 26 April 2024

requests for help were resolved. In our research, we want to know what a customer thinks right after a shopping trip, so the second phase is the focus of this study. Machine learning algorithms are used in e-commerce to improve customer experience, increase sales, reduce costs, analyze customer data, and provide personalized recommendations for products and services. It can also be used to detect fraud and automate customer service tasks. Additionally, machine learning can be used to optimize pricing strategies, improve search engine results, and automate marketing campaigns (Pallathadka et al., 2023). Online businesses face the challenge of predicting customer satisfaction and review ratings before receiving feedback. Specifically, can businesses estimate the rating a customer is likely to provide in their next order review before it is submitted? Developing models to effectively forecast ratings enables proactive identification of customer sentiment, allowing businesses to address issues preemptively and improve experiences. This study aims to predict the next order review (positive or negative) based on historical data on customer orders using machine learning algorithms and deep learning, enabling businesses to understand satisfaction levels earlier and take corrective actions as needed. This data includes the date of the order, the product purchased, the customer's review rating, and other customer-related information. Then, we may use these models to identify clients who appear pleased and those who may be unsatisfied with a product or service. Marketers may utilize this data to adjust their strategies and produce more in line with what their customers want. Also, the current study aims to find the most important things that affect customer satisfaction in e-commerce and make a model that can predict whether a customer will be happy. The remaining sections of this work will be structured as follows: The relevant literature is covered in Section 2. The suggested model is provided in Section 3, experiments, and findings in Section 4, followed by a discussion in Section 5, and finally, Section 6 provides the conclusion and future work.

The following points summarize the principal contributions of this paper: proposed machine learning and deep learning models for the next order review prediction based on a historical dataset from the retail company. Contributions can be detailed in the following points: • A review of the most recent studies in review rating, classification, and prediction. • A proposal for a framework for classifying customer reviews based on customer historical data. • A comparison of different supervised machine learning algorithms, such as Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RS), Gradient Boost Classifier (GBC), and Multi-Layer Perceptron (MLP) as a model used for deep learning. • A framework model was made to solve a binary classification problem, A customer with scores of 1 and 2 will be classified as a not-satisfied customer (negative review) and a customer with review scores of 3,4 and 5 will be classified as a satisfied customer (positive review) • The framework included the study of the effects of creating new features, feature selection, and resampling techniques to improve the quality and efficiency of customer review classification. • An Identifying the key features and drivers of customer satisfaction aligned with prior research, such as delivery performance, pricing, and service experience elements. • Finally, this study indicates that the best classifier is (RF) with an accuracy of 92%, AUC-ROC 90%, weighted F1-score 92%, weighted recall 91%, and weighted precision 91%.

2. Related work

Related work will be divided into two sections. The first section will review previous studies using traditional algorithms of machine learning, and previous studies using deep learning algorithms will be in the second section.

2.1. Machine learning algorithms

Bansal and Srivastava (2018) used Nave Bayes, Logical Regression, Support Vector Machine, and Random Forest as e-commerce customer satisfaction classifiers. They found that Random Forest had the highest accuracy 90.66% and Nave Bayes had the lowest 54.84%. Similarly, Lin (2020) used XGBoost, LightGBM, Support Vector Machine, Random Forest, and Logistic Regression in the online women's clothing industry. LightGBM was the most accurate at 98%. Shah et al. (2021) analyzed Amazon product reviews using machine learning models (Support Vector Machine, Decision Trees, naive Bayes, and Logistic Regression). Random Forest with an accuracy of 93.17% and Logistic Regression of 90.88% performed better than the other models. W. Cao et al. (2021) developed consumer online purchase prediction models using Catboost, Random Forest, and Support Vector Machine. Catboost was the most accurate model at 98.38%. Sharma and Shafiq (2022) used online reviews to build models that can predict whether someone will buy something based on information about service experience. LightGBM outperformed Logistic Regression with an accuracy of 96.1% compared to 95.8%. Taking user experience into account when using machine learning can help predict user satisfaction and find the right direction for development in several areas. Gräßer et al. (2018) reviewed Drugs.com medication reviews for patients and health professionals. Each medicine evaluation has a 0–9 rating indicating patient satisfaction. Positive (7 ratings), negative (4 ratings), and neutral (3 ratings) reviews were grouped. Logistic regression sorted medication reviews with an accuracy of 92%. Hossain et al. (2021) predict Bangla product review scores from written text. SVM, Random Forest, XGBoost, and Logistic Regression with Term Frequency-Inverse Document Frequency (TF-IDF) Vectorizer were used to measure accuracy, precision, recall, and f1-score. SVM outperformed. SVM has 90% dataset accuracy. SVM precision, recall, and f1-score are 90%, 92%, and 91%, respectively. Fouad et al. (2023) proposed a model to predict customer review ratings in e-commerce based on historical transaction data using numeric features only. Framed this as a multi-class classification problem with five rating classes. Five classifiers - Decision Tree, Random Forest, XGBoost, SVM, and KNN are evaluated. Random Forest achieves the best performance with an F1-score of 0.67 after feature selection, outperforming prior literature models. Key drivers of satisfaction are identified as delivery timeliness, product value, and freight costs.

2.2. Deep learning algorithms

Colón-Ruiz and Segura-Bedmar (2020) predicted drug ratings using Drugs.com data and NN models. Patients rated their satisfaction from 0 to 9. The authors suggested numerous NN models, including BERT-LSTM, using two configurations (10-class and 3-class, the compact dataset). Despite its extensive training duration, BERT-LSTM performed best for the 3-class setup. Kurniasari and Setyanto (2020) applied RNN and Word2Vec to analyze Traveloka reviews. Positive and negative groups were employed. CNN, Naive Bayes, and RNN conv were evaluated for comparison in their study. CNN + Word2vec's accuracy is 89%, Naive Bayes 44%, RNN Conv 88%, and RNN + Word2vec 91%. Hameed and Garcia-Zapirain (2020) created a deep-learning binary sentiment classification algorithm. They employed three popular movie opinion databases. MR, IMDB, and SST2. They achieved 80%, 85%, and 90% accuracy on the MR, SST2, and IMDb datasets using one BiLSTM layer and global pooling. Ahmed and Ghabayen (2022) used deep learning to predict review ratings. The system uses deep learning bidirectional gated recurrent unit Bi-GRU model architectures to predict polarity and review rating from review content in two steps. Real-world Amazon and Yelp datasets were used. Precision: 0.72, recall: 0.72, f1-score: 0.69, RMSE: 0.63. Balakrishnan et al. (2022) suggest benchmarking deep learning models such as Convolutional Neural Networks, Recurrent Neural Networks, and Bidirectional Long Short-Term Memory to predict online consumer reviews. Based on BERT, FastText, and Word2Vec word embedding methods. Two datasets—original and augmented—were created using Easy Data Augmentation. 5-Class and 3-class (compressed) versions were evaluated. CNN-RN-BiLSTM had the highest accuracy 96% and F-score (91,1) among Word2Vec and Neural Network prediction models. RNN had 83.5% accuracy, whereas RoBERTa had 73.1%.

The majority of the existing research on predicting review ratings relies on sentiment analysis to determine the polarity of reviews as positive or negative. However, these studies typically do not factor in product attributes that may influence customer satisfaction. Our study aims to predict review ratings using an approach that excludes sentiment analysis of review text content. Instead, we aim to identify key variables related to the product, delivery, and order that drive satisfaction and ratings in e-commerce. By focusing on these variables over sentiment analysis, we can assist businesses in proactively estimating ratings before reviews are even submitted. It is worth noting that sentiment analysis approaches require extensive data cleaning and preprocessing to extract and summarize emotional signals from unstructured review text (Li et al., 2020). Our methodology avoids this labor-intensive text analysis in favor of predictive variables that can be more readily operationalized by businesses seeking to improve customer satisfaction.

3. Proposed model

This paper presents a model designed to predict customer reviews, to enhance the user experience in the e-commerce domain. The proposed model encompasses a comparative analysis between traditional machine learning models and a deep learning model (see Fig. 1). The delineation of the problem at hand involves framing it as a binary classification problem, where customers receiving scores of 1 and 2 are categorized as dissatisfied (indicating a negative review), while those with review scores of 3, 4, and 5 are classified as satisfied customers (indicating a positive review). The ensuing sections of this paper will delve into the details of the model, the experimental setup, and the findings derived from these comparative analyses.

3.1. Data collection

The research used 112,000 orders collected by Olist from the Brazilian e-commerce market over three years (2016–2018) (Olist and Sionek, 2018). Olist was chosen as the data source due to its status as a major Brazilian e-commerce platform offering a wide range of products. The three-year timeframe allowed capturing a large, representative sample of transactions. This duration encompassed seasonal variations and trends in consumer behavior over time. The dataset's breadth and longitudinal nature enhanced the robustness and generalizability of the findings. The data set consists of different datasets namely, the customers dataset, order reviews dataset, order payments dataset, geolocation dataset, order items dataset, orders dataset, product dataset, seller's dataset, and product category name translation. The collected data was reformatted from (CSV) into a tabular format and imported correctly. Data dimensions and column names were understood. Missing data, duplicates, and wrong data types were checked.

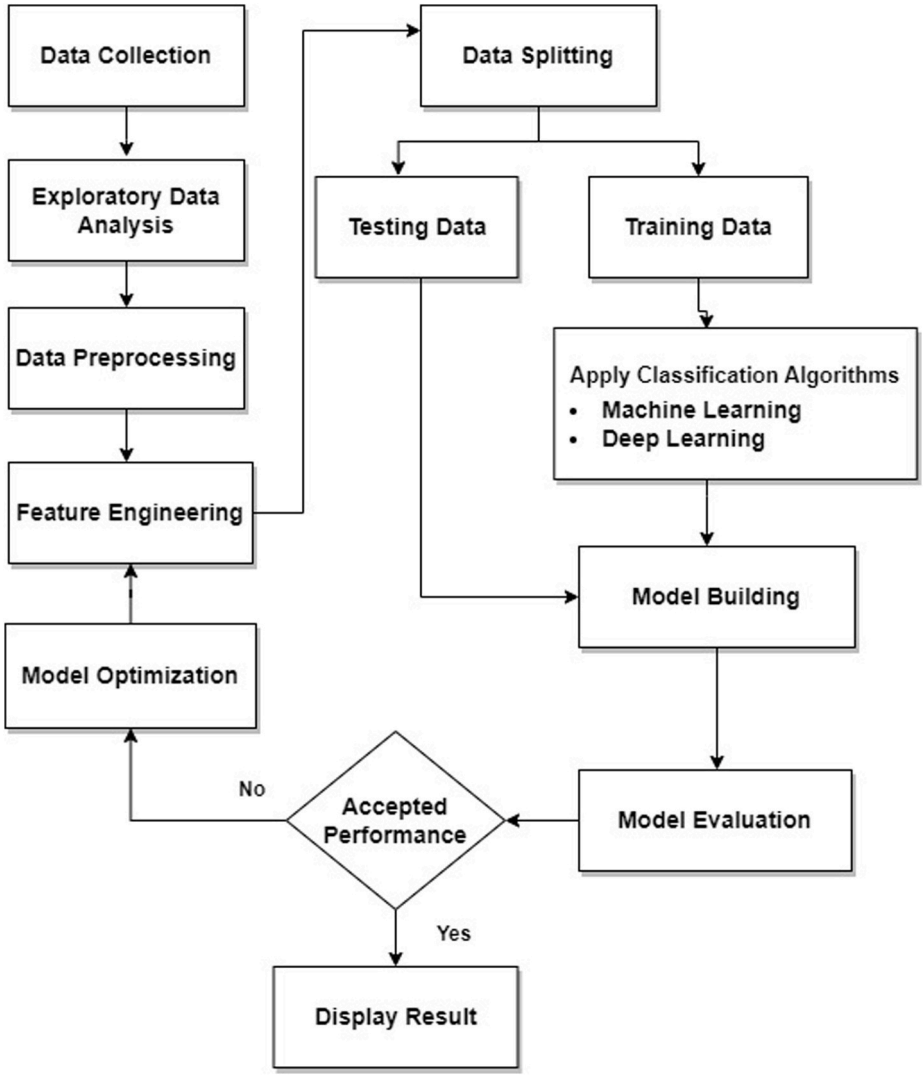


Fig. 1. The proposed model.

3.2. Exploratory data (EDA)

When establishing prediction models, it is crucial to perform preliminary data analysis (Xu-Ying Liu et al., 2009). As a foundation for model selection, this will disclose the data's fundamental information, integrity, distribution features, and factor correlation. Using Python libraries (EDA) was conducted for a deep understanding of the data. Functions like shape, summary, describe, isnull, info, data types, and more were used in the nongraphical method. Uni-variate analysis and bi-variate analysis were used to analyze attributes with the target attribute (review score) (Pajankar and Joshi, 2022a). The univariate analysis concluded that the data set is imbalanced. As a result, the models' performance metrics should be chosen carefully, and resampling techniques are needed to solve this problem.

3.3. Data preprocessing

The source files were all merged to form the dataset. Missing values (null entries) were identified across several features, including order_item_id, seller_id, product_id, and shipping_limit_date. To maintain data integrity and avoid potential biases introduced by imputation techniques, the decision was made to remove any rows containing one or more null values. This was achieved by applying the dropna function, which drops rows with any null values. Additionally, columns containing unique identifiers, such as order_id, customer_id, and product_id, were removed to avoid introducing redundancy and potential leakage in the analysis. Furthermore, any remaining duplicated rows were eliminated using the drop_duplicates function to ensure the uniqueness of each record in the dataset. While this stringent approach led to the loss of 2.6% of the original data, it was deemed necessary to ensure the reliability and consistency of the remaining 97.4% of records. Categorical features were then encoded using label encoding. The remaining instances after this preprocessing totaled 112,897. The data was partitioned into Training (67%) and Test (33%) sets using the train_test_split function from the model_selection class of the Scikit-learn Python library. The 67-33 split was chosen to balance training and evaluation, providing enough data for effective model learning while ensuring reliable performance assessment on unseen data. This widely adopted practice enhances model robustness by allowing thorough parameter tuning and evaluation, contributing to the overall efficacy of the preprocessing methodology. The StandardScaler function helped standardize the training data, ensuring that features were on the same scale and preventing any bias towards variables with larger magnitudes. The fit_transform method was applied to fit the scaler to the training data and transform the feature values into a standard form, as described by (Pajankar and Joshi, 2022b).

3.4. Feature engineering

Initial exploratory data analysis (EDA) indicated that many existing features were not sufficiently informative for classification. Consequently, additional features were introduced based on domain expertise, aiming to improve the model's performance. Subsequent analysis revealed that the correlation between the newly introduced features and the review score surpassed that of other features, suggesting a stronger relationship with the target variable. Through iterative experimentation, the optimal value of the parameter "k" in the SelectKBest function, representing the number of features selected, was determined. This process, utilizing mutual information regression (mutual_info_regression) as the criteria for feature selection, aimed at optimizing model performance by assessing the impact of different "k" values on predictive accuracy. The iterative approach ensured that the selected features effectively contributed to the model's predictive power while mitigating overfitting or underfitting risks, thereby enhancing the robustness and effectiveness of the feature selection methodology (Pajankar and Joshi, 2022b).

The mathematical operation for the process described can be represented as follows:

Given a dataset X with n features ($X = [X\_1, X\_2, ..., X\_n]$) and a target variable y, the goal is to select the k most informative features using mutual information regression.

The mutual information between a feature X_i and the target variable y is calculated as in Eq. (1):

$$I(X_i; y) = \int\int p(x_i, y) log(p(x_i, y) / (p(x_i)p(y)))dx_i dy \quad (1)$$

where p (x_i, y) is the joint probability density function of X_i and y, and p (x_i) and p(y) are the marginal probability density functions of X_i and y, respectively.

The SelectKBest function from scikit-learn computes the mutual information between each feature and the target variable using the mutual_info_regression scoring function.

For each value of k (the number of features to select), the function performs the following steps.

- Compute the mutual information score I (X_i; y) for each feature X_i.
- Sort the features in descending order based on their mutual information scores.
- Select the top k features with the highest mutual information scores.

The optimal value of k is determined by iteratively evaluating the model's performance (e.g., predictive accuracy) for different values of k and selecting the value that yields the best performance.

Mathematically, the process can be represented as.

1 For each $k \in [1, 2, ..., n]$
   a Select the top k features with the highest mutual information scores: $F_k = [X_{i_1}, X_{i_2}, ..., X_{i_k}]$
   b Train the model using the selected features F_k and the target variable y.
   c Evaluate the model's performance (e.g., predictive accuracy) on a validation set.
2 Select the value of k that maximizes the model's performance as in Eq. (2):

$$koptimal = argmaxk(performance(Fk, y)) \quad (2)$$

The mutual information regression scoring function captures the non-linear relationships

Between the features and the target variable, making it a suitable choice for selecting informative features in various machine learning tasks (Battiti, 1994; Ross, 2014).

3.5. Machine learning algorithms

Four distinct classification methods were employed in the field of machine learning: logistic regression (Das, 2021) (LR), Random Forest (RF) (Genuer and Poggi, 2020), Support Vector Machine (SVM) (Pisner and Schnyer, 2020), Gradient Boost Classifier (GBC) (Bahad and Saxena, 2020), and Multilayer Perceptron (MLP) a deep learning model (Izadkhah, 2022). These classifiers were chosen specifically for their effectiveness in handling imbalanced datasets, a prevalent characteristic in the research context. The simplicity and interpretability of logistic regression make it ideal for binary classification. Random Forest uses ensemble learning to combine decision tree predictions to manage imbalanced data and reduce overfitting. SVM is adept at finding ideal hyperplanes for class separation, making it useful in imbalanced classes. Gradient Boost Classifier iteratively improves performance by correcting misclassified instances and addressing class imbalance. Multilayer Perceptron's deep learning allows it to understand complex data patterns and manage imbalanced datasets hierarchically. Therefore, these classifiers were chosen for their proven capacity to handle imbalanced

datasets, ensuring robust model performance in the current study context.

3.6. Deep learning algorithms

Deep learning (DL) is a subset of machine learning (ML) that relies on artificial neural networks (ANNs) Among the various deep learning methods, the Multi-layer Perceptron (MLP) is widely used (Ravikumar et al., 2024). MLP, a versatile approach, is capable of both regression and classification tasks through non-linear approximation, given a set of input characteristics. Comprising the input layer for information reception, hidden layers for representation acquisition, and an output layer for prediction generation, MLP stands out as a key component of deep learning architectures. There are two primary reasons for the adoption of deep learning models: 1) The recent success of DL-based models is attributed to the availability of larger datasets for training. 2) The accessibility of powerful computing resources facilitates the development and training of intricate models, leading to substantial advancements (Dong et al., 2021) (Naskath et al., 2023).

The proposed model employs a deep neural network architecture to perform classification (see Fig. 2). The model is structured as a sequence of dense layers, each performing representation learning on the data. The first hidden layer has 128 nodes with ReLU activation (Si et al., 2018), extracting features from the input data. To regularize this layer and prevent overfitting, it is followed by a dropout layer which randomly drops 50% of the nodes during training (Garbin et al., 2020). The second hidden layer has 64 nodes and uses ReLU activation, followed by another 50% dropout layer. Then a third hidden layer extracts 32 feature representations of the data through its ReLU-activated nodes. Finally, the output layer is a single sigmoid-activated node to output a classification probability. With this sequential design of multiple hidden layers, the model can hierarchically extract nonlinear feature representations from the raw input data to perform the classification task. The depth of the network enables representing overly complex relationships between inputs and outputs that shallower models may not be able to capture. Regularization techniques like dropout are important to prevent overfitting given the model's high representational capacity. Through end-to-end training via backpropagation, this deep network can effectively learn the intricate mapping between inputs and outputs required for accurate classification, even for difficult problems. The total trainable parameters from the stacked dense layers enable the model to tune itself to the dataset and task.

4. Experiments and results

4.1. Experimental setup

Experiments were conducted using Python 3.7.3 in Jupyter Notebooks on a standard laptop equipped with an Intel i7 processor and 12 GB RAM. To build, evaluate, and visualize predictive models, specialized Python data science libraries such as TensorFlow, Scikit-Learn, Pandas, and Matplotlib were leveraged.

4.2. Evaluation performance

Because of dealing with classification datasets with imbalanced classes, care must be taken in selecting appropriate evaluation metrics that can provide insights into model performance on the minority class. Overall accuracy alone is insufficient, as high accuracy can mask poor predictive performance in rare cases (Branco et al., 2017). Alternative metrics like AUC-ROC as in Eq. (3), precision-recall AUC as in Eq. (4), weighted precision as in Eq. (5), weighted recall as in Eq. (6), and weighted F1-score as in Eq. (7) are better suited for imbalance (Buda et al., 2018). The AUC-ROC curve demonstrates discrimination between classes by the model irrespective of class distribution. However, performance differences in the positive class in skewered data are better captured by precision-recall AUC (Saito and Rehmsmeier, 2015). Weighted versions of precision, recall, and F1-score accounted for class imbalance by having the metrics computed separately for each class and then averaged, giving more emphasis to the minority class (Buda et al., 2018). Taken together, these metrics besides accuracy as in Eq. (8) provide greater insight into model capabilities on imbalanced data when used.

$$AUC - ROC = \int_0^1 TPR(FPR)\, d(FPR) \tag{3}$$

$$Precision - RecallAUC = \int_0^1 Precision(recall)\, d(Recall) \tag{4}$$

$$WeightedPrecision = \frac{\sum_{i=1}^{k} n_i \cdot Precision_i}{\sum_{i=1}^{k} n_i} \tag{5}$$

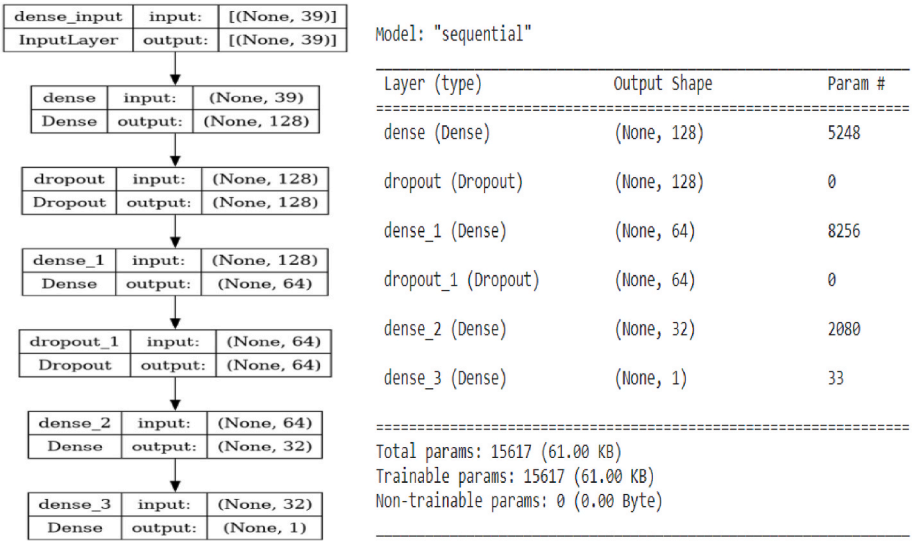


Fig. 2. Architecture of the proposed deep neural network model.

$$WeightedRecall = \frac{\sum_{i=1}^{k} n_i \cdot Recall_i}{\sum_{i=1}^{k} n_i} \quad (6)$$

$$Weighted F1-score = \frac{\sum_{i=1}^{k} n_i \cdot F1-score_i}{\sum_{i=1}^{k} n_i} \quad (7)$$

$$Accuracy = \frac{NumberofCorrectPredictions}{TotalNumberofPredictions} \quad (8)$$

4.3. Experimental results

To assess the efficacy of the proposed model, a series of four experiments were conducted. The next section will provide a comprehensive discussion of these experiments, elucidating the methodologies employed and presenting the findings to evaluate the model's performance.

4.3.1. Experiment 1: base model

In the initial experiment, a base model was implemented to establish a benchmark performance for the given problem. The role of a base model is to serve as a foundational reference point, providing a baseline for the subsequent development of more intricate and accurate models. This foundational model was constructed using logistic regression (LR), chosen for its simplicity, efficiency, and suitability for the nature of the current problem. The base model was trained using a dataset consisting of 75,640 samples, each featuring 39 characteristics. Subsequently, the following section will provide a comprehensive overview of the base model's performance metrics on the test set, comprising 37,257 samples (see Table 1).

The LR model achieved notable results on the dataset with an overall accuracy of 87%. The weighted precision of 85% indicates balanced accuracy in positive predictions, considering the imbalanced nature of the data. With a weighted recall of 87%, the model effectively identifies positive cases, considering class distribution. The weighted F1-score, a comprehensive metric balancing precision, and recall, is reported at 84%, affirming the model's robust performance across multiple evaluation criteria on the given dataset. Building upon this foundation with additional features and more advanced models is likely to further improve the classification capability of this imbalanced dataset.

4.3.2. Experiment 2: before feature selection (BFS)

The experiment was conducted on training data employed to train the base model comprising 75,640 samples, each characterized by thirty-nine features. The performance of classifiers on the testing dataset, which includes 37,257 samples, was evaluated and compared (see Table 2).

Looking at the model performance metrics before feature selection (see Table 2), GBC achieved the best overall results. It obtained an accuracy of 90% and AUC-ROC of 0.88 (see Fig. 3). GBC's weighted precision of 0.89, weighted recall of 0.90, and weighted F1-score of 0.89 showed strong predictive ability even before optimizing the feature space.

However, GBC took 78 s to train which was moderate. RF also performed well, with an AUC- ROC of 0.86 and an accuracy of 88%. Its weighted precision, recall, and F1 were all near 0.87, just slightly behind GBC. But RF required the longest training time at 191 s. Still, its metrics were excellent before reducing features. NN achieved decent but lower metrics than the top models, with an AUC- ROC of 0.82, an accuracy of 88%, a weighted precision of 0.87, a weighted recall of 0.88, and a weighted F1-score of 0.86. However, its training time of 83 s was reasonable. With tuning, NN could improve. SVM was the worst performer before the feature selection. It had an AUC-ROC of just 0.72 and an accuracy of 71%. While SVM took 118 s to train, its poor predictive performance outweighs the slightly faster training time. Finally, LR was the fastest at 7 s, but its metrics were also low, with an AUC ROC of 0.79 and an accuracy of 87%. Faster training did not mean better performance for LR. Results conclude that GBC and RF achieved the top metrics before feature selection but required more training time due to their model complexity. Simpler linear models like LR were quicker but less accurate. Selecting the right model involves balancing predictive power and training efficiency. The area under the precision-recall curve (AUC-PR) measures a model's ability to rank positive examples higher than negative ones (see Fig. 4). All the models achieved strong AUC-PR scores between 0.93 and 0.97 before feature selection. RF and gradient boosting classifier GBC models performed the best with AUC-PR scores of 0.96 and 0.97, respectively. This indicates their high capability of assigning higher ranks to positive examples during prediction. The logistic regression LR and neural network NN models achieved moderately high AUC-PR scores of 0.95 and 0.96, suggesting they could also reliably rank the positive class, though not as effectively as RF and GBC. Finally, the SVM model had the lowest AUC-PR of 0.93, though still respectable. This shows it had some difficulty reliably giving higher ranks to positive examples compared to the other model's pre-feature selection. Overall, the high AUC-PR scores across all models indicate that the data contained predictive signals enabling the ranking of positive examples. Feature selection and tuning may help further enhance the AUC-PR performance.

Table 1
Performance of base model.

Model	Accuracy	W-precision	W-Recall	W–F1-score
(LR)	0.87	0.85	0.87	0.84

Table 2
Performance of models before feature selection.

Model	Accuracy	W-precision	W-Recall	W–F1-score	Time in Seconds
LR	0.87	0.85	0.87	0.84	7
SVM	0.71	0.82	0.71	0.75	118
RF	0.88	0.87	0.88	0.85	191
GBC	0.90	0.89	0.90	0.89	78
NN	0.88	0.87	0.88	0.86	83

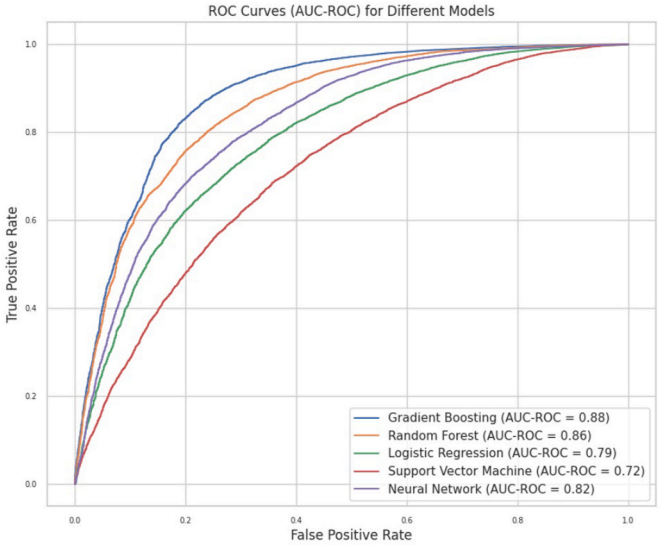


Fig. 3. Roc curves (auc-roc) BFS.

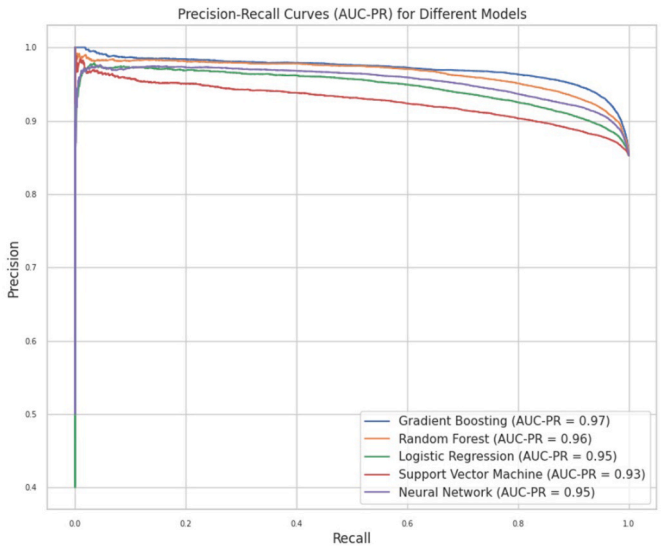


Fig. 4. Precision-recall curves (auc-pr) BFS.

4.3.3. Experiment 3: after feature selection (AFS)

This section aims to explore avenues for enhancing model performance by implementing feature selection techniques. The examination of feature selection arises from the prospect of augmenting the scalability of a system, positing that the system's efficacy can be maintained or even improved when achieving comparable or superior classification performance with a reduced set of features. Subsequently, experiment two was executed on a refined feature set, comprising twenty selected features following the feature selection process.

To further refine the model, hyperparameters were systematically tuned using a Grid Search algorithm (Khalid and Javaid, 2020). The optimal hyperparameter settings resulting from this process are presented in Table 3. Classifiers' results on testing data are presented in Table 4.

After reducing the feature space, the model performance and training times shifted noticeably. Looking at the model performance metrics before feature selection (see Table 4), GBC still achieved the best overall predictive performance, with an accuracy of 91%, weighted precision, and recall. All reaching 0.91, its metrics improving to an AUC-ROC of 0.90 (see Fig. 5). GBC's training time only increased slightly to 86 s after feature selection. (RF) saw a small drop-in training time to 183 s, but its metrics remained excellent with an AUC-ROC of 0.86, an accuracy of 88%, and weighted precision, recall, and F1 holding at 0.86–0.88. The NN model improved the most in terms of efficiency, with training time dropping from 83 to 47 s after feature selection. Its metrics also rose slightly, reaching an AUC-ROC of 0.83, accuracy of 88%, and weighted precision, recall, and F1 in the 0.86–0.88 range.

The performance of SVM model improved markedly when trained on the reduced set of features after feature selection. Its metrics jumped to an accuracy of 86%, and weighted precision, recall, and F1 of 0.81–0.86. Training time also dropped to 136 s. LR remained the fastest to train at 4

Table 3
Hyperparameters of models.

Model	hyperparameters
GBC	(n_estimators = 80, max_depth = 9, random_state = 33)
RF	(max_depth = 10, n_estimators = 800, min_samples_split = 2, criterion = 'entropy')
SVM	(C = 0.1, gamma = 'auto', kernel = 'poly')
LR	(penalty = 'l2', solver = 'sag', C = 1.0, random_state = 33)
NN	Activation Functions: ('relu' for hidden layers and 'sigmoid' for the output layer) Loss Function: ('binary_crossentropy'), Optimizer: ('adam')

Table 4
Performance of models after feature selection.

Model	Accuracy	W-precision	W- Recall	W–F1-score	Time in seconds
LR	0.87	0.85	0.87	0.84	4
SVM	0.86	0.85	0.86	0.81	136
RF	0.88	0.87	0.88	0.86	183
GBC	0.91	0.91	0.91	0.91	86
NN	0.88	0.87	0.88	0.86	47

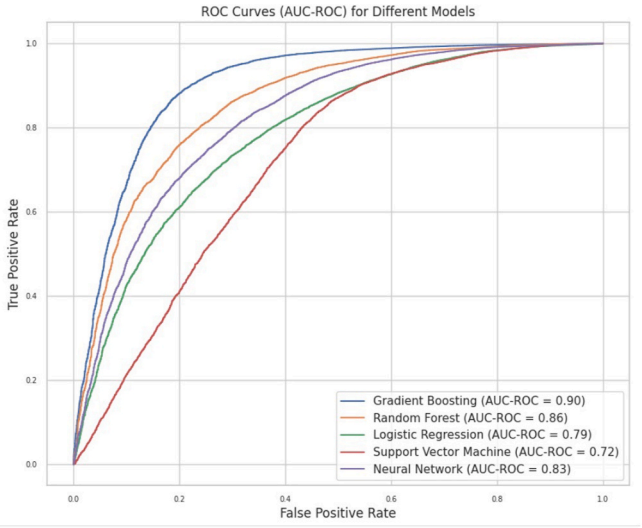


Fig. 5. Roc curves (auc-roc) afs

s. However, its metrics stayed unchanged, indicating the limitations of this linear model even with optimized features. (SVM) AUC-PR decreased from 0.93 to 0.92 after feature selection (see Fig. 6). LR and RF maintained the same AUC-PR of 0.95 and 0.96 respectively after feature selection. This suggests the initial feature set already contained the primary drivers of positive classification for a tree-based ensemble method. GBC again, the top AUC-PR of 0.98 was slightly higher than the 0.97 before feature selection. The marginal improvement indicates GBC was already highly performant before reducing features. Finally, NN achieved a higher AUC-PR of 0.96 than 0.95 before feature selection.

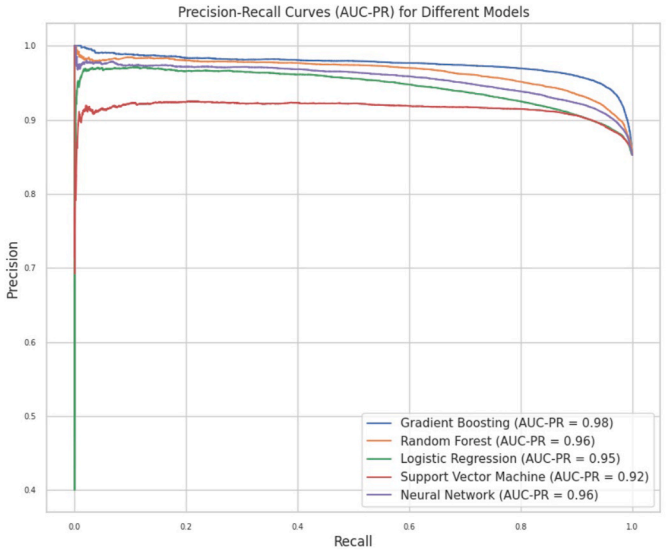


Fig. 6. Precision-recall curves (auc-pr) afs

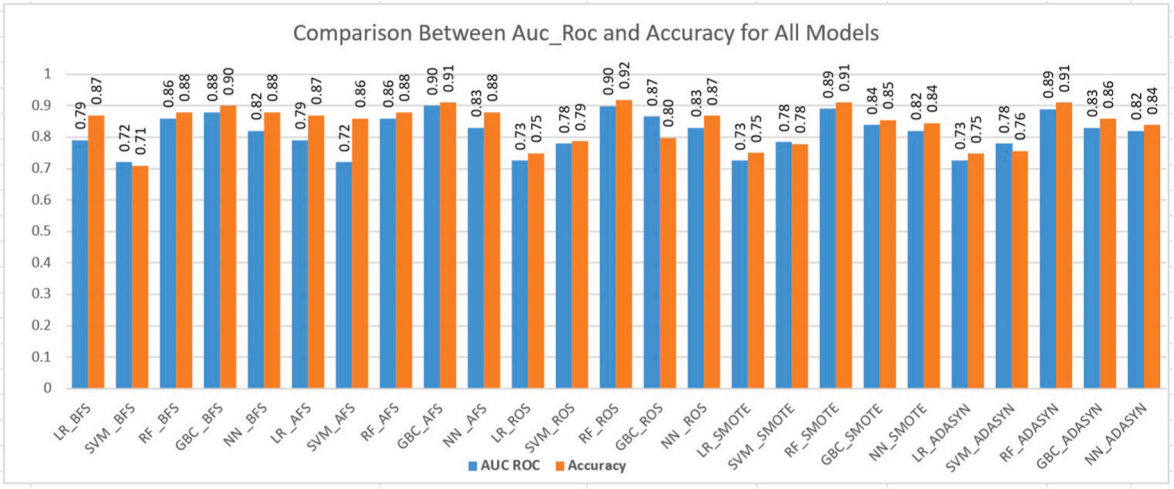


Fig. 7. Comparison between auc-roc and accuracy for all models.

GBC model retained the top predictive performance, while SVM saw major improvements. The RF model remained strong as well. This highlights the importance of proper feature engineering. Further resampling techniques could potentially yield additional minor improvements as evinced by the analysis delineated in Fig. 7.

4.3.4. *Experiment 4: after sampling*

This section will investigate the possibilities of improving performance by solving imbalanced data problems. Various methods have been developed to deal with the issue of imbalanced data (Yap et al., 2014) (Ghorbani and Ghousi, 2020). In this experiment's data, an oversampling strategy was employed to handle the unbalanced data because this technique is usually used more often than other methods as RundomOverSampler (Mesquita et al., 2021), SMOTE(Raghuwanshi and Shukla, 2020) and ADASYN(Y. Cao et al., 2018), and the models were retrained again. Resampling is performed via the imbalanced-learn Python library, which provides numerous resampling approaches. It is compatible for use with sci-kit-learn (Letteri et al., 2022). After applying the oversampling technique, the classifiers' results on testing data were compared (see Table 5).

After oversampling, the RF model achieved the best performance across all sampling methods, with AUC-ROC scores up to 0.90, accuracy around 0.91–0.92, and F1 reaching 0.92 with random oversampling. This indicates that RF was highly capable of distinguishing between positive and negative classes while maintaining strong predictive performance even with the imbalanced dataset. It also trained efficiently in under one hundred seconds. The SVM machine model saw improved AUC-ROC and accuracy scores of 0.78–0.79 with the oversampling techniques compared to 0.72 without sampling. This suggests that SVM's performance was hindered by class imbalance, and oversampling helped address this issue. However, SVM took longer to train than other models. LR and GBC performed well across metrics under different sampling conditions, though neither surpassed random forest. The NN model achieved moderate results overall, with AUC-ROC scores around 0.82–0.83, an accuracy of 0.84–0.87, and F1 reaching 0.87 with random oversampling.

This suggests that while NN could distinguish between positive and negative classes well, it did not surpass tree-based methods like random forest. Results Concluded that RF architecture was best suited to efficiently leveraging this dataset for accurate binary classification. The results highlight how performance can vary across machine learning algorithms given a certain dataset and problem. Oversampling proved useful in handling class imbalance.

5. Discussion

This section provides an overview of the experimental findings. Additionally, conduct a comparative analysis between the proposed model and prior research.

5.1. *Comparison of different models*

This section will compare all models in terms of AUC-ROC, accuracy, weighted F1-score, weighted precision, weighted recall, and time of

Table 5
Performance of models after OverSampling.

Sampler	Model	AUC-ROC	Accuracy	W-precision	W- Recall	W– F1-score	Time in seconds
RandomOverSampler	LR	0.73	0.75	0.78	0.82	0.75	16
	SVM	0.78	0.79	0.80	0.84	0.79	1286
	RF	0.90	0.92	0.91	0.91	0.92	78
	GBC	0.87	0.80	0.82	0.87	0.80	91
	NN	0.83	0.87	0.87	0.86	0.87	76
SMOTE	LR	0.73	0.75	0.78	0.82	0.75	16
	SVM	0.78	0.78	0.80	0.84	0.78	1198
	RF	0.89	0.91	0.91	0.91	0.91	99
	GBC	0.84	0.85	0.86	0.86	0.85	143
	NN	0.82	0.84	0.85	0.85	0.84	83
ADASYN	LR	0.73	0.75	0.77	0.82	0.75	16
	SVM	0.78	0.76	0.79	0.84	0.76	1815
	RF	0.89	0.91	0.91	0.90	0.91	97
	GBC	0.83	0.86	0.86	0.85	0.85	145
	NN	0.82	0.84	0.85	0.85	0.84	76

training.

The performance of previous classification models was evaluated on an imbalanced dataset, both before and after feature selection as well as with three sampling techniques. Overall, GBC achieved the strongest results across all metrics. Before reducing the feature space, GBC obtained an AU- ROC score of 0.88 and an accuracy of 90%. Its precision, recall, and F1-score on the weighted average of all classes were 0.89 each—the highest among all models evaluated. This indicates GBC was adept at handling the imbalance in the training data.

After feature selection, GBC saw further improvements, reaching an AUC-ROC of 0.90 and an accuracy of 91%. The precision, recall, and F1-score on a weighted average basis all rose to 0.91 as well. The results show proper feature engineering can further optimize GBC's predictive abilities. The RF model also performed well, achieving the second-highest scores overall. Before feature selection, RF obtained an AUC-ROC of 0.86 and an accuracy of 88%.

Its weighted precision, recall, and F1-score were each 0.87, reflecting impressive performance but just slightly behind GBC. After reducing the feature space, RF maintained an AUC-ROC of 0.86 and an accuracy of 88%, while improving the weighted F1-score to 0.88. When combined with sampling techniques like RandomOverSampler, RF achieved its best metrics of 90–92% for precision, recall, accuracy, and F1-score. This highlights the usefulness of sampling to help RF deal with imbalanced data. In contrast, NN model did not see the same benefits from feature selection and sampling. Before reducing features, NN achieved decent but lower metrics than GBC and RF, with an AUC-ROC of 82%, accuracy of 88%, precision of 0.87, recall of 0.88, and F1 of 0.86. These results lagged behind the top models. After feature selection, NN improved slightly but only to an AUC ROC of 83%. Accuracy, w-precision, w-recall, and w-F1-scores did not change. Additional sampling failed to boost NN's performance further. Overall, NN may require more hyperparameter tuning and architectural adjustments to better handle the class imbalance. In summary, tree-based ensemble models like GBC and RF achieved the top overall performance, indicating their suitability for imbalanced classification tasks. NN was decent but did not improve as much, signaling a need for further optimization. Proper feature selection and sampling techniques can help optimize predictive performance on skewed data.

The training time required for each classification model was evaluated before and after feature selection, as well as with three sampling techniques (see Fig. 10). Before reducing the feature space, LR model was the fastest, taking only 7 s to train. NN was also relatively quick, training in 83 s. However, SVM took 118 s, while RF and GBC were slower at 191 s and 78 s, respectively. After feature selection, training times improved for all models since the feature space was reduced. LR trained the quickest in just 4 s. NN took 47 s, while both SVM and RF saw decent drops in training time, to 136 s and 183 s each. GBC did not improve as much, requiring 86 s to train after feature selection. The sampling techniques increased training time across all models. With RandomOverSampler, LR took 16 s, while NN took 76 s. SVM and GBC were slower at 1286 s and 91 s, respectively. RF was the fastest with RandomOverSampler, training in just 78 s. The relative training times were similar for SMOTE Sampler and ADASYN, though SVM became exceptionally slow with these techniques, taking 1198 s and 1815 s, respectively. Simpler linear models like LR were the fastest to train overall. More complex models like SVM, RF and GBC were slower, while NN was somewhere in between. Feature selection improved training efficiency for all models by reducing the feature space. However, sampling techniques added significant computational costs, especially for SVM. The results highlight the trade-off between training time and model performance when dealing with imbalanced data. Fig. 8 compares the weighted F1-Score for all models. Fig. 9 compares the weighted precision and weighted recall for all models.

5.2. Top features

The top features listed are the names of the features that the Random Forest model has identified as the most important for making predictions (see Fig. 11). Each feature has an associated importance score, indicating its contribution to the model's decision-making process. Five Features were created in the feature engineering process: (wd_delivery_time_delta, wd_actual_delivery_time, payment_value, total_order_value, and order_freight_ratio).

Regarding the top 10 features from the RF model for predicting customer satisfaction, the presence of time-related features like wd_delivery_time_delta, wd_actual_delivery_time, and order_freight_ratio highlights the importance of delivery timeliness in driving customer satisfaction. This aligns with research showing delivery speed is a key driver of e-commerce satisfaction (Smith et al., 1999). The inclusion of review_comment_message indicates the textual content of customer reviews contains meaningful signals correlated with satisfaction. This suggests potential value in applying natural language processing techniques to extract insights from review texts. Payment_value and total_order_value being top features indicate the monetary amount of orders and payments impacts satisfaction levels. Higher values may be associated with increased expectations. Price ranking as a major feature suggests that perceived fairness or competitiveness of pricing influences satisfaction with the purchase. Freight_value surfacing as a top feature highlights the role shipping costs play in the customer experience. Higher freight values may negatively impact satisfaction. Geolocation

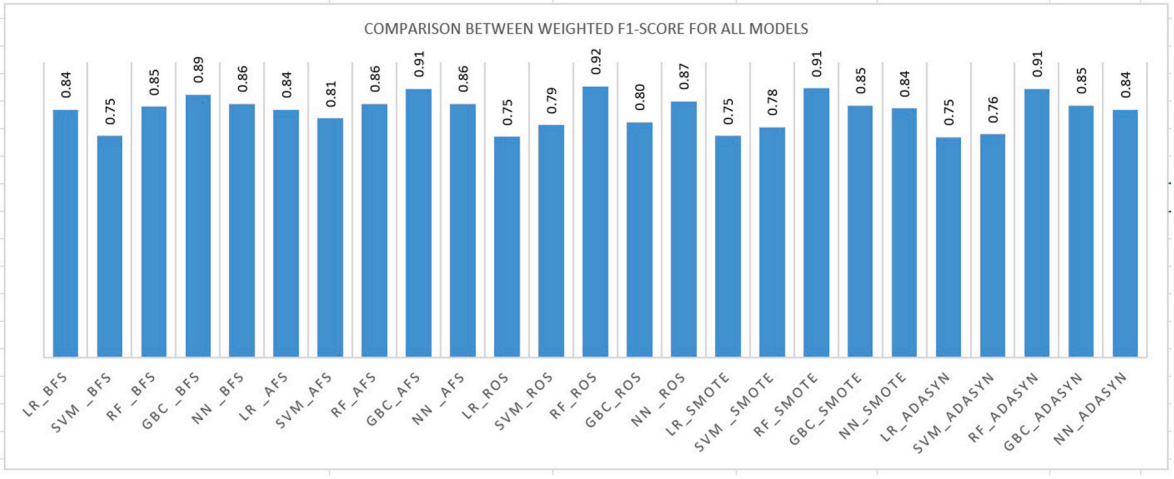


Fig. 8. Comparison between weighted F1-Score for all models.

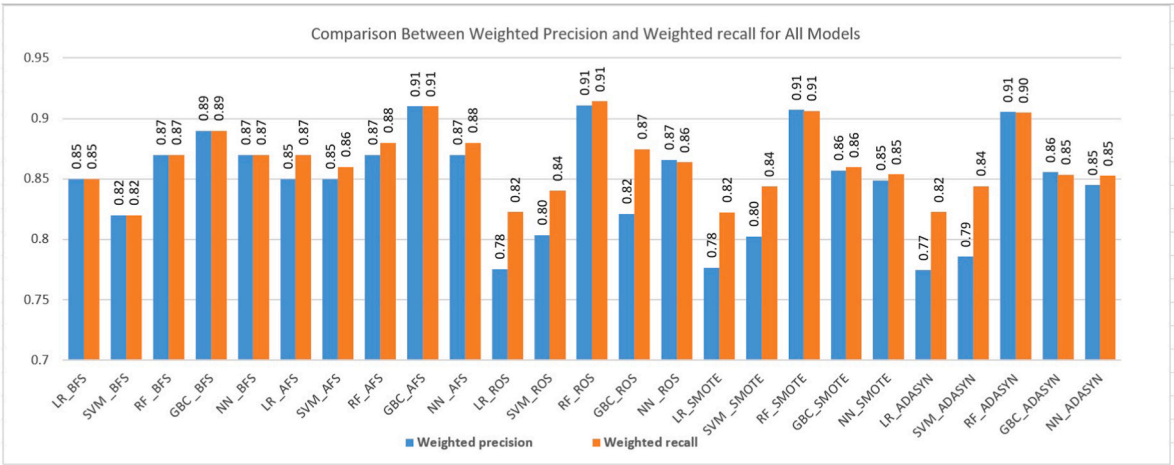


Fig. 9. Comparison between weighted precision and weighted recall for all models.

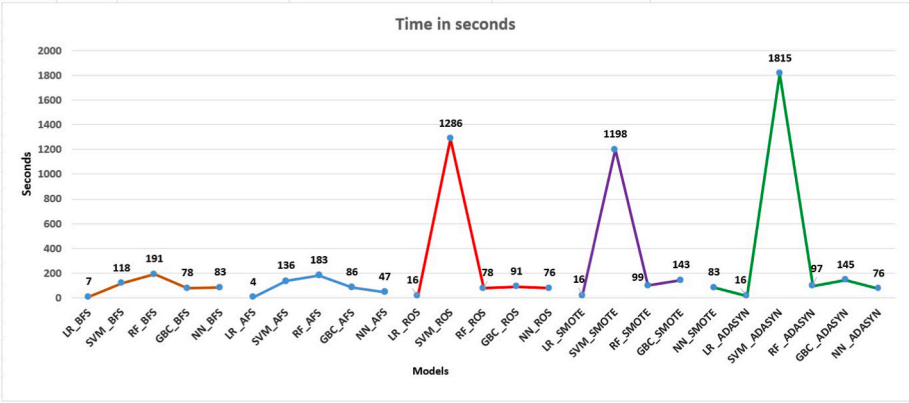


Fig. 10. Comparison between training time for all models.

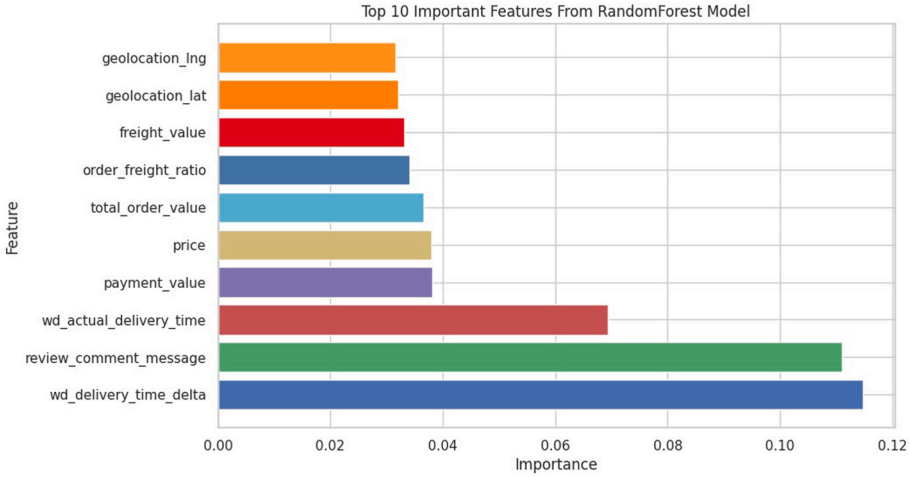


Fig. 11. The most important features.

features emphasize satisfaction levels may vary based on customer geographic location. This provides support for location-based personalization. Overall, the key drivers of satisfaction identified by the RF model align with established research around delivery performance, pricing, and service experience. The findings offer data-driven guidance for e-commerce providers to focus efforts on enhancing these key areas to improve customer satisfaction. Analyzing feature importance from predictive modeling provides targeted insights.

5.3. Comparison with literature studies

This section presents a comparison between the current and previous research using the E-commerce dataset (see Table 6).

Kumar et al. (2021)analyzed the text of low, neutral, and high ratings

Table 6
Comparison between previous researches.

REF	year	Data set	methods	results
Kumar et al. (2021)	(2021)	"Home and Kitchen" dataset from Amazon.	Combined model of (Naïve Bayes, SVM, and logistic Regression)	F1-score of Low rating 76.2 % F1-score of Neutral rating 69.2% F1-score of High rating 78.5%
Bappon et al. (2022)	(2022)	Bengali tech review data	Random Forest	Accuracy 86.28%.
Ahmed and Ghabayen (2022)	(2022)	two datasets (Yelp and Amazon datasets)	bidirectional gated recurrent unit Bi-GRU	precision 72% recall 72% f1-score 69%
Fouad et al. (2023)	(2023)	Olist Data set	(RF), (XGB), (SVM), (KNN), (DT)	F1- score 67%, 62%, 60%, 54%, 53%.
Current Study		Olist Data set	RF Accuracy 0.92, W-precision 0.91, W-recall 0.91, W-f1-score 0.92, AUC-ROC 0.90 (AOS). GBC Accuracy 0.91, W-precision 0.91, W-recall 0.91, W-f1-score 0.91, AUC-ROC 0.90 (AFS). NN Accuracy 0.88, W-precision 0.87, W-recall 0.88, W-f1-score 0.86, AUC-ROC 0.83 (AFS). LR A accuracy 0.87, W-precision 0.85, W-recall 0.87, W-f1-score 0.84, AUC-ROC 0.79 (AFS). SVM Accuracy 0.86, W-precision 0.85, W-recall 0.86, W-f1-score 0.81, AUC-ROC 0.72 (AFS).	

using Naïve Bayes, SVM, and logistic regression. They achieved F1 scores of 76.2%, 69.2%, and 78.5% for each rating class. Bappon et al. (2022)categorized Bengali reviews as good, negative, or neutral with random forest, achieving 86.28% accuracy. Ahmed and Ghabayen (2022) predicted ratings via Bi-GRU deep learning, obtaining 0.72 precision, recall and 0.69 F^1 score. Anonymous et al. predicted 5-class review ratings using a decision tree, random forest, XGBoost, SVM, and KNN on numeric order data. Random forest performed best with a 0.67 F^1 score after feature selection. The current study presented a model using RF, SVM, and other classifiers to predict positive/negative satisfaction through binary classification of reviews. RF achieved top performance with 0.92 accuracy, 0.90 AUC ROC, 0.91 precision/recall, and 0.92 F^1 score. Comparing the previous studies, the current study surpasses these previous benchmarks, especially in the binary classification task.

5.4. Limitations of the current study

While our study has yielded valuable insights, certain limitations that may affect the interpretation and generalizability of our findings must be acknowledged. Firstly, concerns about the representativeness of the data and its ability to capture the full diversity of the e-commerce market are raised due to the reliance on data from a single e-commerce platform. Furthermore, the reliability of our results could potentially be impacted by data quality issues, such as missing values or inaccuracies. Secondly, the applicability of our findings to other geographical regions or industry contexts may be limited due to the study's specific focus on the Brazilian market and the unique consumer behaviors and market dynamics present in that region. Thirdly, although multiple machine learning and deep learning models were evaluated, the possibility that other advanced techniques or ensemble approaches, which were not explored in our study, might yield superior performance in certain scenarios cannot be discounted. Moreover, the effectiveness of our predictive models is contingent upon the selection and engineering of relevant features, and important features may have been overlooked or complex relationships between variables may have failed to be captured, leading to suboptimal model performance. Furthermore, while techniques were employed to address class imbalance in our binary classification task, the effects of imbalanced data may not have been fully mitigated, potentially affecting the model's ability to accurately classify minority classes. Finally, external factors, such as macroeconomic conditions, competitive landscape changes, or unexpected events, that could influence customer satisfaction and impact logistics and customer experience were not accounted for in our study. By acknowledging these limitations, the aim is to provide context for our findings and encourage future research to address these constraints, ultimately strengthening the robustness and validity of conclusions in this domain.

6. Conclusion and future work

This study demonstrates a successful machine learning approach for predicting online retail customer satisfaction based on review ratings. The random forest model performed robustly on a large dataset from a major e-commerce retailer, achieving an accuracy of 92% in classifying satisfied customers. These predictive insights into future consumer behavior enable data-driven managerial decisions to improve customer loyalty and retail performance. The top features identified by the RF model align with established research, highlighting the importance of delivery performance metrics (e.g., delivery time), pricing factors (e.g., payment value, total order value), and service experience elements (e.g., review text, freight value) in driving customer satisfaction. The findings provide data-driven guidance for e-commerce companies to prioritize enhancements in these key areas. Future research directions include exploring advanced neural network architectures, such as deep learning models, coupled with extensive hyperparameter tuning to enhance their ability to handle class imbalance effectively. Incorporating natural language processing (NLP) techniques could exploit valuable insights from customer review text data, which emerged as a top predictive feature. Furthermore, investigating ensemble and stacking methods that combine multiple models may yield improved performance by leveraging the strengths of different algorithms. Expanding the methodologies to other domains beyond e-commerce, such as healthcare and finance, could validate the generalizability of the findings across industries with imbalanced data. Finally, integrating complementary data sources, such as social media sentiment and customer demographic information, could provide a more comprehensive understanding of the key drivers behind customer satisfaction.

Funding

No funding.

CRediT authorship contribution statement

Maha Zaghloul: Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization. **Sherif Barakat:** Writing – review & editing, Supervision. **Amira Rezk:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors confirm that they do not have any conflicts of interest to declare about this manuscript. They have no financial interests, affiliations, or other competing interests that could potentially bias my involvement in the publication process. Their contributions to this manuscript are made in good faith and with full transparency.

Data availability

Data will be made available on request.

Acknowledgments

This work is part of a doctoral dissertation that is now in its last stages of completion.

References

Ahmed, B.H., Ghabayen, A.S., 2022. Review rating prediction framework using deep learning. J. Ambient Intell. Hum. Comput. 13 (7), 3423–3432. https://doi.org/10.1007/s12652-020-01807-4.

Bahad, P., Saxena, P., 2020. Study of AdaBoost and Gradient Boosting Algorithms for Predictive Analytics, pp. 235–244. https://doi.org/10.1007/978-981-15-0633-8_22.

Balakrishnan, V., Shi, Z., Law, C.L., Lim, R., Teh, L.L., Fan, Y., 2022. A deep learning approach in predicting products' sentiment ratings: a comparative analysis. J. Supercomput. 78 (5), 7206–7226. https://doi.org/10.1007/s11227-021-04169-6.

Bansal, B., Srivastava, S., 2018. Sentiment classification of online consumer reviews using word vector representations. Procedia Comput. Sci. 132, 1147–1153. https://doi.org/10.1016/j.procs.2018.05.029.

Bappon, S.D., Mursalin, G.S.M., Khan, M.I., 2022. Sentiment analysis of Bengali texts on online tech gadget reviews using machine learning. In: 2022 25th International Conference on Computer and Information Technology (ICCIT), pp. 324–329.

Battiti, R., 1994. Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Network. 5 (4), 537–550. https://doi.org/10.1109/72.298224.

Bernhardt, K.L., Donthu, N., Kennett, P.A., 2000. A longitudinal analysis of satisfaction and profitability. J. Bus. Res. 47 (2), 161–171. https://doi.org/10.1016/S0148-2963(98)00042-3.

Bilal, M., Marjani, M., Hashem, I.A.T., Malik, N., Lali, M.I.U., Gani, A., 2021. Profiling reviewers' social network strength and predicting the "Helpfulness" of online customer reviews. Electron. Commer. Res. Appl. 45, 101026 https://doi.org/10.1016/j.elerap.2020.101026.

Branco, P., Torgo, L., Ribeiro, R.P., 2017. A survey of predictive modeling on imbalanced domains. ACM Comput. Surv. 49 (2), 1–50. https://doi.org/10.1145/2907070.

Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Network. 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011.

Cao, Y., Zhao, X., Zhou, Z., Chen, Y., Liu, X., Lang, Y., 2018. MIAC: mutual-information classifier with ADASYN for imbalanced classification. In: 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), pp. 494–498. https://doi.org/10.1109/SPAC46244.2018.8965597.

Cao, W., Wang, K., Gan, H., Yang, M., 2021. User online purchase behavior prediction based on fusion model of CatBoost and Logit. In: Journal of Physics: Conference Series, vol. 2003. IOP Publishing Ltd. https://doi.org/10.1088/1742-6596/2003/1/012011, 1.

Colón-Ruiz, C., Segura-Bedmar, I., 2020. Comparing deep learning architectures for sentiment analysis on drug reviews. J. Biomed. Inf. 110 https://doi.org/10.1016/j.jbi.2020.103539.

Das, A., 2021. Logistic regression. In: Encyclopedia of Quality of Life and Well-Being Research. Springer International Publishing, pp. 1–2. https://doi.org/10.1007/978-3-319-69909-7_1689-2.

Dong, S., Wang, P., Abbas, K., 2021. A survey on deep learning and its applications. Computer Science Review 40, 100379. https://doi.org/10.1016/j.cosrev.2021.100379.

Fouad, M., Barakat, S., Rezk, A., 2023. Effective E-Commerce Based on Predicting the Level of Consumer Satisfaction. https://doi.org/10.1007/978-981-99-4764-5_17, 261–278.

Garbin, C., Zhu, X., Marques, O., 2020. Dropout vs. batch normalization: an empirical study of their impact to deep learning. Multimed. Tool. Appl. 79 (19–20), 12777–12815. https://doi.org/10.1007/s11042-019-08453-9.

Genuer, R., Poggi, J.-M., 2020. Random Forests, pp. 33–55. https://doi.org/10.1007/978-3-030-56485-8_3.

Ghorbani, R., Ghousi, R., 2020. Comparing different resampling methods in predicting students' performance using machine learning techniques. IEEE Access 8, 67899–67911. https://doi.org/10.1109/ACCESS.2020.2986809.

Gómez, M.I., McLaughlin, E.W., Wittink, D.R., 2004. Customer satisfaction and retail sales performance: an empirical investigation. J. Retailing 80 (4), 265–278. https://doi.org/10.1016/j.jretai.2004.10.003.

Gräßer, F., Kallumadi, S., Malberg, H., Zaunseder, S., 2018. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In: Proceedings of the 2018 International Conference on Digital Health, pp. 121–125. https://doi.org/10.1145/3194658.3194677.

Hameed, Z., Garcia-Zapirain, B., 2020. Sentiment classification using a single-layered BiLSTM model. IEEE Access 8, 73992–74001. https://doi.org/10.1109/ACCESS.2020.2988550.

Hossain, M.I., Rahman, M., Ahmed, M.T., Rahman, M.S., Islam, A.Z.M.T., 2021. Rating prediction of product reviews of bangla language using machine learning algorithms. In: 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), pp. 1–6.

Hu, H., Krishen, A.S., 2019. When is enough, enough? Investigating product reviews and information overload from a consumer empowerment perspective. J. Bus. Res. 100, 27–37. https://doi.org/10.1016/j.jbusres.2019.03.011.

Izadkhah, H., 2022. Training multilayer neural networks. In: Deep Learning in Bioinformatics. Elsevier, pp. 95–111. https://doi.org/10.1016/B978-0-12-823822-6.00012-3.

Khalid, R., Javaid, N., 2020. A survey on hyperparameters optimization algorithms of forecasting models in smart grid. Sustain. Cities Soc. 61, 102275 https://doi.org/10.1016/j.scs.2020.102275.

Kumar, P., Dayal, M., Khari, M., Fenza, G., Gallo, M., 2021. NSL-BP: a meta classifier model based prediction of Amazon product reviews. Int. J. Int. Multimedia and Artificial Intelligence 6 (6), 95. https://doi.org/10.9781/ijimai.2020.10.001.

Kurniasari, L., Setyanto, A., 2020. Sentiment analysis using recurrent neural network. J. Phys. Conf. 1471 (1) https://doi.org/10.1088/1742-6596/1471/1/012018.

Letteri, I., Cecco, A. Di, Dyoub, A., Penna, G. Della, 2022. Imbalanced Dataset Optimization with New Resampling Techniques, pp. 199–215. https://doi.org/10.1007/978-3-030-82196-8_15.

Li, Z., Tian, Z.G., Wang, J.W., Wang, W.M., 2020. Extraction of affective responses from customer reviews: an opinion mining and machine learning approach. Int. J. Comput. Integrated Manuf. 33 (7), 670–685. https://doi.org/10.1080/0951192X.2019.1571240.

Lin, X., 2020. Sentiment analysis of E-commerce customer reviews based on natural language processing. In: ACM International Conference Proceeding Series, pp. 32–36. https://doi.org/10.1145/3436286.3436293.

Liu, Xu-Ying, Wu, Jianxin, Zhou, Zhi-Hua, 2009. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39 (2), 539–550. https://doi.org/10.1109/TSMCB.2008.2007853.

Mesquita, F., Mauricio, J., Marques, G., 2021. Oversampling techniques for diabetes classification: a comparative study. In: 2021 International Conference on E-Health and Bioengineering (EHB), pp. 1–6. https://doi.org/10.1109/EHB52898.2021.9657542.

Mewada, A., Dewang, R.K., 2023. A comprehensive survey of various methods in opinion spam detection. Multimed. Tool. Appl. 82 (9), 13199–13239. https://doi.org/10.1007/s11042-022-13702-5.

Naskath, J., Sivakamasundari, G., Begum, A.A.S., 2023. A study on different deep learning algorithms used in deep neural nets: MLP SOM and DBN. Wireless Pers. Commun. 128 (4), 2913–2936. https://doi.org/10.1007/s11277-022-10079-4.

Olist, Sionek, A., 2018. Brazilian E-Commerce Public Dataset by Olist. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/195341.

Pajankar, A., Joshi, A., 2022a. Hands-on Machine Learning with Python. Apress. https://doi.org/10.1007/978-1-4842-7921-2.

Pajankar, A., Joshi, A., 2022b. Introduction to machine learning with scikit-learn. In: Hands-on Machine Learning with Python. Apress, pp. 65–77. https://doi.org/10.1007/978-1-4842-7921-2_5.

Pallathadka, H., Ramirez-Asis, E.H., Loli-Poma, T.P., Kaliyaperumal, K., Ventayen, R.J. M., Naved, M., 2023. Applications of artificial intelligence in business management, e-commerce and finance. Mater. Today: Proc. 80, 2610–2613. https://doi.org/10.1016/j.matpr.2021.06.419.

Pisner, D.A., Schnyer, D.M., 2020. Support vector machine. In: Machine Learning. Elsevier, pp. 101–121. https://doi.org/10.1016/B978-0-12-815739-8.00006-7.

Raghuwanshi, B.S., Shukla, S., 2020. SMOTE based class-specific extreme learning machine for imbalanced learning. Knowl. Base Syst. 187, 104814 https://doi.org/10.1016/j.knosys.2019.06.022.

Ravikumar, N., Zakeri, A., Xia, Y., Frangi, A.F., 2024. Deep learning fundamentals. In: Medical Image Analysis. Elsevier, pp. 415–450. https://doi.org/10.1016/B978-0-12-813657-7.00041-8.

Riaz, M.U., Guang, L.X., Zafar, M., Shahzad, F., Shahbaz, M., Lateef, M., 2021. Consumers' purchase intention and decision-making process through social networking sites: a social commerce construct. Behav. Inf. Technol. 40 (1), 99–115. https://doi.org/10.1080/0144929X.2020.1846790.

Roetzel, P.G., 2019. Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. Business Research 12 (2), 479–522. https://doi.org/10.1007/s40685-018-0069-z.

Ross, B.C., 2014. Mutual information between discrete and continuous data sets. PLoS One 9 (2), e87357. https://doi.org/10.1371/journal.pone.0087357.

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10 (3), e0118432. https://doi.org/10.1371/journal.pone.0118432.

Shah, B.K., Jaiswal, A.K., Shroff, A., Dixit, A.K., Kushwaha, O.N., Shah, N.K., 2021. Sentiments detection for Amazon product review. In: 2021 International Conference on Computer Communication and Informatics, ICCCI 2021. https://doi.org/10.1109/ICCCI50826.2021.9402414.

Sharma, A., Shafiq, M.O., 2022. A comprehensive artificial intelligence based user intention assessment model from online reviews and social media. Appl. Artif. Intell. 36 (1) https://doi.org/10.1080/08839514.2021.2014193.

Si, J., Harris, S.L., Yfantis, E., 2018. A dynamic ReLU on neural network. In: 2018 IEEE 13th Dallas Circuits and Systems Conference (DCAS), pp. 1–6. https://doi.org/10.1109/DCAS.2018.8620116.

Smith, A.K., Bolton, R.N., Wagner, J., 1999. A model of customer satisfaction with service encounters involving failure and recovery. J. Market. Res. 36 (3), 356–372. https://doi.org/10.1177/002224379903600305.

Supriyanto, A., Wiyono, B.B., Burhanuddin, B., 2021. Effects of service quality and customer satisfaction on loyalty of bank customers. Cogent Business & Management 8 (1). https://doi.org/10.1080/23311975.2021.1937847.

Terblanche, N.S., 2018. Revisiting the supermarket in-store customer shopping experience. J. Retailing Consum. Serv. 40, 48–59. https://doi.org/10.1016/j.jretconser.2017.09.004.

Yap, B.W., Rani, K.A., Rahman, H.A.A., Fong, S., Khairudin, Z., Abdullah, N.N., 2014. An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets, pp. 13–22. https://doi.org/10.1007/978-981-4585-18-7_2.

Zhao, Y., Xu, X., Wang, M., 2019. Predicting overall customer satisfaction: big data evidence from hotel online textual reviews. Int. J. Hospit. Manag. 76, 111–121. https://doi.org/10.1016/j.ijhm.2018.03.017.