Customer Satisfaction Prediction Using Machine Learning Techniques

by

Bauyrzhan Abdi

Submitted to the Department of Data Science
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

April 2025

Author ....

Department of Data Science
28.04.2025

Certified by ....

Michael Lewis
Associate Professor
Thesis Supervisor

Customer Satisfaction Prediction Using Machine Learning Techniques

by

Bauyrzhan Abdi

Submitted to the Department of Data Science
on 28.04.2025, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

Businesses increasingly focus on understanding customer behavior and experience, especially factors contributing to stronger customer satisfaction. The prediction of satisfaction is regarded as a complex, but important task due to its critical role in shaping loyalty and retention by increasing service quality. Despite extensive research on consumer behavior, satisfaction remains a relatively underexplored target in the context of accurate prediction for commercial organizations.

Data-driven approaches, specifically machine learning and deep learning strategies, have shown great potential in the analysis of customer feedback data, including techniques such as Logistic Regression, Decision Tree, Random Forest, XGBoost, SVM, ANN, LSTM, and etc.

This study focuses on using machine learning methods to predict contentment levels by analyzing publicly available datasets, relevant to existing works. The initial four methods discussed earlier are applied and compared. The process involves selecting suitable datasets across various industries, preparing features, training the models, and evaluating their performance. Previously published works have been successfully replicated, with explanations offered for any discrepancies in outcomes.

Keywords: *customer satisfaction, machine learning, predictive modeling, classification, customer experience*

Thesis Supervisor: Michael Lewis
Title: Associate Professor

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Understanding customer behavior is a vital component of business success, and yet it has long remained a challenging endeavor. Companies that succeed in recognizing the patterns and factors affecting consumer behavior and satisfaction, and leverage these insights to refine their product development strategies, would more likely thrive in a dynamic market environment. Customer contentment is inherently subjective, and is driven by a wide variety of personal expectations, experiences, numerous contextual indicators that are not easily measurable, including time of the day, economic conditions, client's mood, geographical locations, product quality, service delivery channel (digital or in-person), and others.

Historically, the challenge was primarily constrained by the technological development of those times and shortcomings in the availability, granularity, format, and amount of the data.

The computational capabilities have been extended significantly with the development of top-performance computing resources, particularly Graphics Processing Units (GPUs). According to [1], the GPU market was worth 52.1 billion US dollars in 2023 and is expected to grow by more than 27% each year from 2024 to 2032.

As regards consumer data, most of the customer feedback and transactional information used to be stored in paper formatted records or fragmented systems, which made predictive and searching analytics nearly impractical. The absence of computerized databases meant that companies had no tools to systematically store and process customer data. The landscape only began to shift with the introduction of the first digital databases in the 1960s, facilitating the era of structured data collection [2].

Following several decades, large-scale digitization has transformed how businesses interact with customers, enabling them to build detailed customer profiles. The use of digital touchpoints such as online surveys, reviews, mobile apps, transaction logs, and online payments has created room for myriads of consumer data.

A notable example of this trend can be observed in Kazakhstan, where 93.3% of all public services, including airline and e-commerce fields, are currently available, and 86% of them are optimized for mobile use [3]. Moreover, the share of online transactions in the state grew by 58% each year from 2018 to 2022, as illustrated in the figure below.

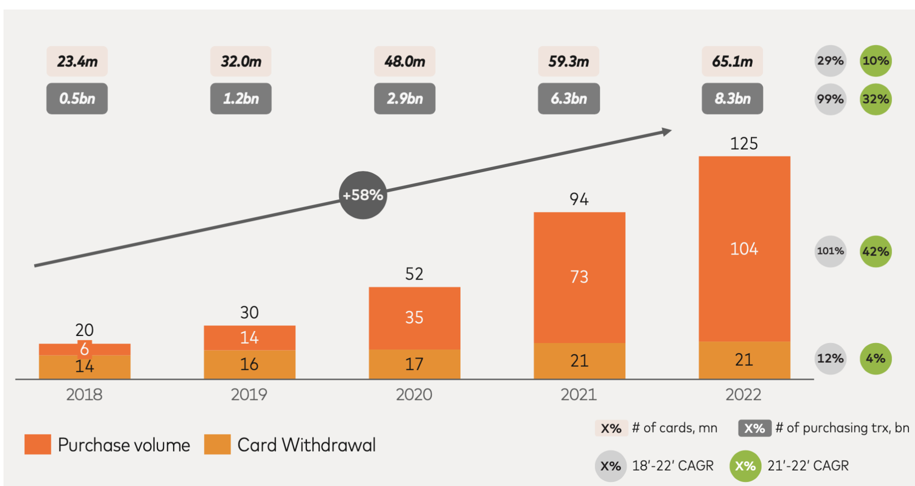


Figure 1-1: The growth of online transactions' volume between 2018 and 2022

The rise in the volume and availability of customer data, along with advancements in computational power primarily benefits data-driven approaches such as machine learning and deep learning algorithms, enhancing their predictive performance.

Businesses can now access and investigate large volumes of consumer feedback data, including geographical location, transaction records, purchasing histories (e.g. airline tickets or commercial products), and even experience ratings; and then use them to uncover valuable hidden insights that support satisfaction-focused decision-making. They are not only limited to predicting customer behavior, but, if necessary, are also able to estimate satisfaction and expectations to identify their strengths and possible levers of improvement.

Despite numerous studies involved in consumer behavior analysis, satisfaction as a predictive target remains relatively underexplored, particularly in the context of practical applications for commercial organizations. Many organizations still struggle to spot and predict which factors influence satisfaction and how they can be translated into actionable strategies. KPMG [4] finds that US companies experienced a 5% decline in customer satisfaction levels from 2022, achieving the lowest point since 2015. PwC [5] reports that 54% of consumers feel the customer experience provided by most global companies requires improvements, while only 38% believe their needs are adequately understood.

This thesis will explore commercial datasets capturing various aspects of customer experience records, and apply machine learning techniques to those datasets to predict individual consumer satisfaction levels.

1.2 Problem Statement and Objectives

Predicting customer satisfaction is not restricted to a single format as satisfaction is not inherently only a binary outcome and can be a continuous or multi-class variable. Machine learning and deep learning strategies serve as promising approaches for modeling such predictions. Nonetheless, there is no universally accepted method as the performance of each algorithm often varies depending on the dataset characteristics, such as the presence of noise, outliers, null values, or imbalanced classes due to the scarcity of satisfied customers in the data. In the context of smaller-scale datasets, machine learning (ML) methods often demonstrate superior performance compared to deep learning (DL) approaches, which tend to run a risk of poor generalization under such settings.

The primary goal of the study is to examine the machine learning techniques in predicting customer satisfaction and evaluate the performance of selected methods using frequent datasets referenced in the existing literature. Specific objectives include:

- to thoroughly inspect and select available customer satisfaction datasets for further analysis
- to pre-process, analyze, and model these datasets using chosen methods
- to compare and contrast the accuracy of different models (e.g., logistic regression, decision trees, random forest, and etc.)
- to assess the practical implications of using machine learning predictions for improving customer experience and business decision-making.

1.3 Scope of the Study

The scope of this study includes data pre-processing, feature engineering, model development, and performance evaluation. The research concentrates on technical aspects of prediction, but it does not aim to deliver a deployable software solution. For the purposes of this study, the prediction problem is approached as a binary classification task, distinguishing between satisfied and unsatisfied customers. Furthermore, the focus is limited to the data from selected datasets of some domains (e.g. aviation, e-commerce) and may not generalize across all industries and demographic contexts.

1.4 Prior Research

A detailed discussion of earlier methods to predict customer satisfaction is provided in Chapter 2 (Related Work), while the methods adopted in those studies are outlined in Chapter 4 (Methodology).

1.5 Proposed Approach

After careful examination of prior approaches, the algorithms that demonstrate the highest performance over chosen datasets and are widely used across existing works, are proposed. The proposed approach is detailed in Chapter 4 (Methodology).

Chapter 2

Related works

This chapter examines relevant studies in the area of customer satisfaction prediction, covering both conventional and state-of-the-art techniques.

Examination and maintenance of customer satisfaction has been acknowledged as a required, yet challenging issue in the businesses [6], [7], [8]. However, during the 1970s, US firms largely ignored the concept of consumer contentment analysis and supported the idea that clients derive delight more often in purchase rather than displeasure [9]. Despite meeting client needs via the establishment of product novelties in the following 20 years, the growing discontent amongst them drove public attention to the research in forecasting delight rates of buyers in the 1990s and early 2000s [10].

2.1 Traditional Approaches

Most of the approaches of that time were largely based on *statistical analysis* techniques. Commonly, *multiple regression analysis (MRA)* models were utilized to inspect the connection between customers' perspectives on specific features of the product and their overall satisfaction and loyalty. Simply, it is useful to determine what matters to the clients. Kristensen et. al [11] measured regression coefficients to extract four exogenous antecedents of user satisfaction for postal services in Denmark.

The authors discovered that the perceived image of the company far outweighs other attributes such as expectations, product quality, and value. Andaleeb and Conway [12] investigated factors affecting customer satisfaction in the restaurant industry. The study suggested that customers experienced an increased joy due to the responsiveness of employees in the first place with food price and quality also playing a slightly lesser role in it. However, the restaurant's appearance and design did not truly contribute to this feeling.

Dimitriades [13] reconnoitered how consumer satisfaction, loyalty, and commitment are understood and linked to each other in the Greek cultural and business environment using hierarchical regression and *t-testing* models. T-testing is a statistical tool that compares averages of two clusters to see if the difference between them is meaningful and helps to identify if the nature of the difference is real or random by chance [14]. The findings illustrated that satisfaction and loyalty are overlapping terms, whereas commitment is strongly linked to both of them. Moreover, gender or industry type has no effect on these relationships.

Some studies made use of other regression techniques such as *logistic regression* and *ordinary least squares (OLS)*, the former still being widely used today. Rust and Zahorik [15] investigated ways to improve customer retention using logistic regression and developed a mathematical framework for managers to evaluate client content levels. Lawson and Montgomery [16] applied various types of logistic regression for binary, nominal, and ordinal cases of customer survey data and portrayed their versatility and effectiveness in the scope of customer relationships. OLS had also proven itself well in determining whether poor service quality moderate discontent [17], [18].

Often, however, regression models face *multicollinearity issues* [11], [19], [20]. This occurs when intercorrelations between features, that contribute to satisfaction and loyalty, become too high such that results become unreliable to derive any statistical significance. Within the context of statistical modeling techniques, [19] established the use of bivariate regression analysis (BRA) in order to address these shortcomings. In comparison with predecessors, this model treats each feature separately in the customer satisfaction analysis. Additionally, regularization techniques such as a ridge, lasso or Bayesian regression were used as a countermeasure to events of collinearity [21], [22], [23].

Another common approach within customer satisfaction and loyalty measurements is *structural equation modeling (SEM)*. SEM is a powerful technique, that aids in examining how distinct components of the phenomenon are interconnected to one another in a causal manner [24]. SEM can analyze complex relationships and potentially circumvent the limitations of traditional regression models. [22] made use of the LISREL software package to discover the overall satisfaction levels of users. [25] and [26] applied the partial least squares (PLS) method of SEM to calculate national consumer contentment index models. [27] could distinguish the clear impact of customer satisfaction on loyalty and retention in the corporate environment via SEM methods. Certain SEM enthusiasts tried to embed neural networks into PLS and reduced prediction error, giving a more robust model for the estimation of satisfaction [28].

Some researchers used the *targeted bootstrapping* technique (TABOO) [29]. TABOO is specifically designed to generate 'what if' scenarios from customer satisfaction data. The authors demonstrated that the method overcomes regression limitations in two case studies. While these techniques improved predictive accuracy, they still required extensive manual feature engineering and domain-specific knowledge to yield meaningful insights.

2.2 Machine Learning Approaches

With the advancement of machine learning approaches in the late 2000s, researchers began digging into data-driven techniques for customer satisfaction prediction tasks. It has become difficult to handle the increasing number of complex relationships and the imbalanced nature of databases since common methods rely on analyzing existing patterns rather than extracting rules. Rules are insightful structures found by identifying relationships between attributes in data. [30] pinpointed that machine learning techniques are prevalent due to their increased accuracy and smaller errors.

Machine learning is a product of artificial intelligence studies. Feigenbaum stated that machines must possess knowledge in order to be considered intelligent [31]. El Naqa and Murphy [32] argue that machine learning is a set of algorithms that emulate human behavior by learning from neighboring environments.

There are various machine learning methods such as regression, decision trees, bayesian models, kernel methods, ensemble learning techniques, and others. Machine learning methods are already implemented and simple to use. However, it is of great significance to fine-tune parameters to achieve desired outcomes.

The user contentment issue plays a key role in the banking system. Banks are passionate about incorporating machine learning techniques to foresee and analyze customer behavior and feelings as large amounts of data are available to them. The advancement of online banking made it possible to acquire information about customer user experience, demographic data, and transaction history. While many users are happy about online banking services, third-party money laundering is not a rare occasion, which affects customer satisfaction and disappointment rates. Shetu et al. [33] evaluated various machine learning techniques such as Logistic Regression, Random Forest, Naive Bayes, SVM, decision tree, and ANN on real-life banking data of Bangladesh. KNN, Logistic Regression, and Random Forest correctly identified the proportion of satisfied and discontent users with an accuracy of 96%. The authors discovered that 86.66% of Bangladeshi bank users are content with their banking experience. Ligiarta and Ruldeviyani [34] researched online banking user experience in Indonesia after the COVID-19 pandemic using sentiment analysis via the SVM algorithm. SVM accurately provided insightful data with a 92% score on security, reliability, and usefulness issues of Indonesian banks from the collected data.

Eom et al. [35] used a dataset from the local water supply market to predict the customer satisfaction index. They employed techniques such as SVM and Logistic Regression as a baseline model. They analyzed customers' language, which words consumers use to fill out complaints, resulting in 49 critical words that affect customer satisfaction. The predictive model proved to be robust as SVM clearly outperformed

Logistic Regression, achieving 98.7% accuracy.

Gou et al. [36] used the XGBoost algorithm to estimate the consumer satisfaction index in the telecommunications field and compared the model with linear regression. XGBoost model is an ensemble learning approach and well-suited for capturing complex relationships that simple linear models often fail to represent [37]. Notably, the linear regression model reached 75.36% of prediction accuracy, whereas the latter significantly outperformed it by nearly 20%, reaching 94.97% of accuracy. In a subsequent study, authors enhanced their model with CatBoost, leading to slight improvements in predictive results [38]. Building on this, Wang et al. [39] evaluated the fusion model, which consists of XGBoost, LightGB, and a decision tree, achieving an impressive accuracy of 97% in the Chinese telecommunications market. [40] explored consumer contentment estimation in mobile telecommunication services using the Gradient Boosting decision tree (GBDT). Unlike XGBoost and CatBoost, which are known for their strong nonlinear pattern recognition capabilities, GBDT is especially efficient in selecting core features and handling both continuous and discrete data. Their study highlighted the reliability of GBDT in managing outliers, making it an ideal choice for satisfaction forecasting jobs where data variability is a big concern. Fei et al. [41] applied GBDT on the related dataset used in the previous study by [39], but with an extended approach combining it with quadratic dimensionality reduction techniques through systematic clustering and exploratory data analysis (EDA). Their enhanced model reduced data redundancy and optimized feature selection. They registered a strong predictive performance with 99% of accuracy, surpassing all prior research. Meanwhile, some researchers focused on the Indian market, emphasizing voice call quality as an essential aspect for the long-term survival of any telecommunication organization [42] on the [88] dataset. By comparing the Random Forest, attention model with Bi-GRU, CART decision tree, and ANN model, the research illustrated the effectiveness of Random Forest with 90% accuracy. [43] aggregated decision tree classifier and Random Forest prediction model to compute weights of the attributes affecting mobile user satisfaction with network performance. The score on the training set exceeded 95%.

The customer satisfaction prediction is applicable across many industries, including the airline sector. Airline firms are particularly interested in applying machine learning techniques to enhance customer loyalty to their companies. [44] investigated what features affect customer loyalty and happiness during the flight using XGBoost and bagging on [74] dataset. Bagging is a machine learning algorithm used to improve model stability by combining weak parameters into stronger ones. In the study, he found that in-flight internet service and online check-in are top-notch factors of traveler satisfaction, illustrating bagging is superior to the single decision tree due to their stable nature and reduced variance. However, bagging underperforms logistic regression since it has weaker performance on low-noise small data [45]. Yao [46] experimented with distinct depth configurations of a decision tree to streamline feature analysis on the same dataset [74]. His findings revealed that passenger's type of travel also plays a crucial role in satisfaction levels, which had been overlooked in the earlier research. Various researchers re-evaluated the existing dataset using different machine learning techniques, including Logistic Regression, Decision Tree, KNN, Random Forest, Naive Bayes algorithms, and Support Vector Machines (SVM). In the study by [47], they recorded an accuracy and f1-score of 0.94 for Random Forest. Meanwhile, [48] boosted the metrics up to 0.97 by integrating the GridSearch algorithm for hyperparameter optimization. Salah-ud-din et al. [49] used the airline dataset [74] and applied several models to predict customer satisfaction and Gradient Boosting achieved the highest performance. Hibovic and Smajic [50] utilized the same dataset [74] and the Naive Bayes algorithm had the best accuracy of 0.99, followed by Random Forest and Decision Tree with respective scores of 0.958 and 0.954. Some research investigated ways to help airline companies enhance their service quality and efficiently manage staffing during key business periods [51]. The researchers tested SVM with various kernel-type parameters (linear, polynomial, RBF), concluding the Gaussian RBF Kernel-based model is tailored for consumer satisfaction approximation in airline hospitality services. Jindal et al. [52] identified that most of the airline datasets have an imbalanced nature; therefore, authors applied the SMOTE technique to augment the data. Over-sampling contributed to metrics enhancements with the

Extra Trees algorithm being at the forefront of predictive analytics at 99.87% accuracy. Combining over-sampling techniques with XGBoost gave better results in another literature [53] on airline dataset [75].

Customer satisfaction forecasting is extremely important for e-commerce businesses as well since their brand loyalty is shaped by how well their needs are fulfilled. Wangkiat and Polprasert [54] investigated customer satisfaction scores on the [80] dataset provided by the major e-commerce retailer company in Brazil by employing Logistic Regression, Random Forest, and KNN algorithms. The authors used both oversampling and under-sampling techniques such as SMOTE and RandomUnderSampler to balance the data and GridSearchCV for hyperparameter tuning. Among these models, Random Forest combined with data augmentation techniques demonstrated the top performance in all metrics, including accuracy, precision, and f1-score. Zaghloul et al. [55] further extended the range of methods used in [54] by adding GBDT, SVM, and multi-layer perceptron (MLP) to test their performance using the same data. This study concluded that Random Forest is the most robust model for the e-commerce dataset [80], achieving an accuracy of 92%. It proved to be more effective than neural networks and big data techniques. Tukino and Maulana [56] researched consumer contentment of large phone accessories distributor products in Indonesia. Researchers applied the C4.5 decision tree algorithm using the WEKA application as a helper tool on the dataset based on questionnaire data, which consists of such attributes as availability, product quality, price, warranty, delivery rate, and promotion. The C4.5 algorithm is a type of decision tree model that minimizes bias towards attributes with many values, ensuring more balanced and reliable classifications. This is a pivotal point for the survey data in this study, which has uneven data distribution. They could identify that customer dissatisfaction is directly correlated with disruptions in cell phone accessories availability, challenging warranty policy, and mediocre goods standards. On the contrary, the high quality of items, their increased availability, and easy warranty contributed to a positive customer experience. Mookherje and Roy [57] researched satisfaction factors affecting repeat buying behavior in luxury watch retailer data. The researchers used the ID3 decision tree algorithm for customer satisfaction prediction tasks. They found that watch condition and retailer website performance are the most crucial factors, in deciding whether customers return to make the next purchase or not. The overall accuracy for the model constituted around 90%.

2.3 Deep Learning Approaches

Deep Learning is the next-stage development of machine learning that mimics human reasoning and the ability to process large amounts of data. Unlike machine learning, deep learning does not demand for manual feature selection, allowing the machine to learn patterns and recognize complex relationships from raw data. As an imitation of brain work, deep learning consists of neural networks of varying layers. There are several types of deep learning algorithms such as convoluted neural networks (CNN), recurrent neural networks (RNN), long short-term memory networks (CNN), and etc.

Deep learning models derive their strength from fine-tuning hyperparameters such as learning rates, number of hidden layers, batch sizes, activation functions, and dropout rates, ultimately optimizing their performance. Alkurd et al. [58] proposed a real-time customer satisfaction monitoring framework using a deep neural networks (DNN) approach. The DNN model was trained for 120 epochs, 4 hidden layers, 0.01 learning rate, and 27% dropout rate on [87] dataset. The authors contrasted their trained model with classic machine learning techniques such as KNN, Random Forest, and decision tree, illustrating that DNN outperforms all of them with 93.78% accuracy - nearly 20% higher than the next best-performing model. Gadri et al. [59] explored the application of neural networks in the tourism industry, investigating whether tourists are happy with hotel services during their travel or not using review data [86]. They tuned hyperparameters of their DNN model using 6 fully connected layers, 32 batch sizes, 15 iterations, ReLu activation function for input layers, sigmoid function for the outer layer, and Adam optimizer. Machine learning algorithms (LDA, KNN, CART, SVM) achieved approximately 68% accuracy, while the proposed technique delivered a phenomenal 99%.

Zhang and Guo [60] use deep neural networks (DNN) of two different modalities (unimodal, multimodal) to predict customer satisfaction in the e-commerce field on [81] dataset. Unimodal DNNs handle one type of data (i.e. text or image), whereas multimodal DNNs process multiple types. The study employs Adam and Root Mean Square Propagation (RMSProp) algorithms to help models train efficiently and adapt to complex data structures. In simple terms, these algorithms optimize the performance of neural networks. The authors measured the performance via mean absolute error (MAE) rates and achieved 0.4818 and 1.0887 on train and test sets for multimodal DNN with Adam optimizer, illustrating the superiority of multimodal methods over unimodal ones (1.2045 and 1.4277 MAE). Later, the authors could reach slight improvements in the accuracy of the same data [81] by increasing the number of epochs up to 2500 and applying the Keras framework [61].

Saleh and Abu-Soud [62] were concerned with the rising employee dissatisfaction among the population in Jordanian prestigious organizations. This issue largely stems from inefficient management of employee experience and satisfaction, which affects their loyalty to companies. The research exploited ANN to determine factors affecting employee contentment, which were found to be a comfortable working environment and salary satisfaction. The model reached 96.09% accuracy on the randomly split dataset. Prasanth et al. [63] applied ANN and SVM to measure student satisfaction with online learning in Sri Lanka. As per the results obtained, ANN resulted in 76.19% accuracy and identified a few attributes that contribute to student dissatisfaction, which are poor connectivity, limited access to laptops, and noisy environment. Sun [64] tested the capabilities of feed-forward ANN to the consumer content estimation problem. It is a type of neural network, that cyclically performs forward, and backward propagations and weight updates. The researcher confirmed the efficiency of ANN for user satisfaction analysis with 0.42566 MAE but noted overfitting and underfitting issues remain unsolved. Another study complemented a certain type of feed-forward ANN - backpropagation (BP) neural network with principal component analysis (PCA) [65]. PCA facilitated a thorough analysis of customer needs and habits, while the BP neural network was employed to predict customer satisfaction scores, eventually achieving a mean squared error of 2.1038 and a correlation coefficient of 0.84272.

Du [66] explored how the aviation service might be improved to attract more customers and raise satisfaction rates using a neural network. The study found that improvements in business class services lead to higher content levels while enhancing economy class quality helps minimize complaints and encourages passengers to return. Bhargav and Prabu [67] investigated the effectiveness of a hybrid neural network model (Random Forest and CNN) to determine if it outperforms a traditional ML approach (KNN). The results suggest the hybrid model delivers great forecasts with 90% accuracy, a significant improvement over traditional KNN's 75%. [68] used a variety of machine learning and deep learning algorithms to analyze airline customer propensities in South Korea. The authors compare KNN, Random Forest, XGBoost, CNN, and CNN-LSTM, and the latter surpassed models by 7-11% on average with 90% accuracy.

Similar to [68], some researchers focused on the use of LSTM, which is specifically designed for time-series and sequential data. Some datasets track records of customer behavior over time, and LSTM effectively recognizes patterns within them. Huang et al. [69] evaluated time-aware LSTM (T-LSTM) for handling irregular time intervals in public transportation; [70] assessed bidirectional LSTM (Bi-LSTM) to capture past and future records of touristic features and [71] benchmarked regular LSTM against decision tree for e-commerce dataset [80]. T-LSTM reached a root mean squared error (RMSE) of 11.40, whereas Bi-LSTM and LSTM demonstrated 0.0024 MAE and 0.89 accuracy, respectively.

Yin et al. [72] proposed the novel improved fully connected neural network (FCNN+) to predict mobile internet customer satisfaction, illustrating a high accuracy of 83.5%. Aftan and Habib [73] exploited the AraBERT algorithm to perform custom satisfaction forecasting in the telecom sectors of Saudi Arabia. AraBert is well-suited for the Arabic dataset since regular deep learning models (CNN, RNN) often struggle to learn patterns from the language. With an accuracy of 0.9433, AraBERT showed superior performance over CNN and RNN, which achieved 0.8834 and 0.9135.

Chapter 3

Dataset

This chapter outlines datasets commonly used in the literature or those containing features relevant to customer satisfaction prediction across various domains. Datasets play a key role in understanding customer satisfaction since they contain necessary information that helps recognize patterns and contributing factors. They are examined based on factors such as its content, size, features, collection, and satisfaction distribution. These datasets serve as a foundation for benchmarking models, analyzing trends, and validating predictions. Their impact on predictive performance will be further investigated in the results section. Datasets are grouped by categories, based on the industry of field type, featuring airline, e-commerce, telecommunications, and other sectors.

3.1 Airline Datasets

3.1.1 US Airlines

Dataset description

The US Airlines dataset [74] is a publicly available, open-source customer satisfaction dataset, which consists of passenger survey data from various aviation services in the USA. It contains 129,880 instances collected from 103,904 unique subjects, providing enough data for analysis.

There are 24 independent features that store the information encoded as ordinal, numerical, and categorical values, each representing the associated degree of record. To facilitate the smooth model training and evaluation, the dataset has been pre-divided by the author into train and test sets using an 80/20 split ratio.

The data collection consists of three main components:

1. Airline and passenger information

2. Service and experience ratings

3. Target variable, which represents the overall passenger satisfaction rate in a categorical manner as satisfied, dissatisfied, or neutral

Categorical variables such as gender, age, and customer type (loyal or disloyal) describe passengers and flight-related factors include type of travel (personal or business), class (class on board), flight distance, and departure and arrival delays in minutes. Notably, there are 13 numerical attributes where passengers provided ratings on a scale from 0 to 5, covering aspects such as inflight wifi service, departure/arrival time convenience, ease of online booking, gate location, food and drink, online boarding, inflight entertainment, on-board service, leg room service, baggage handling, check-in service, inflight service, and cleanliness. The target variable, satisfaction, groups dissatisfied, and neutral values as discontent and satisfied ones are classified as content.

Amidst the dataset, there are 43% satisfied passengers, and 57% fall under dissatisfied or neutral categories, which suggests that the dataset is slightly imbalanced, creating a 14% gap between the two classes. Though one class is not extremely dominant or skewed to another, it may still impact the predictive performance, highlighting the need for distribution adjustment techniques.

3.1.2 Invistico

Dataset description

The Invistico dataset [75] is an open-source passenger satisfaction dataset, widely available for the masses. This dataset was designed for the tech competition held in 2020 to give participants the opportunity to work on real-world data. The authors were specifically interested in determining the robustness of this dataset to predict future customer satisfaction based on several parameter values. Additionally, they sought to identify certain aspects of the services offered by the organization, which may generate more satisfied customers. After the end of challenge, it is now primarily utilized for model performance evaluation purposes.

The dataset comprises 129,880 unique data entries and 23 features, which include flight data, passenger information, and their ratings. Similar to [1], most of the features correlate with this dataset with exceptions for such attributes as online support and inflight service. There are 4 categorical values (sex, customer loyalty type, travel type, and flight class) and the rest is numerical, excluding the dependent variable. The dependent variable accepts either satisfied or dissatisfied values.

The dataset classifies 54.7% of passengers as satisfied and the remaining 45.3% as dissatisfied. There exists a minor imbalance in the dataset with around 9% difference, though it is not highly noticeable.

Dataset collection

The dataset is obtained from real-life data, provided by the airline organization. The actual name of the company remains undisclosed for a few reasons, which are to protect its identity and due to its intention to first evaluate the potential of machine learning in solving satisfaction prediction challenges before implementing these methods in its systems. A large amount of data (around 1,000,000 records) was gathered over time from passenger records of those who have flown with them. Then, they created 23 features and resampled the data into 129,880 instances, eventually anonymizing it.

3.1.3 Skytrax

Dataset description

This [76] is an airline review dataset across different aviation companies around the world. It is primarily applied in airline service quality analysis and customer delight estimation problems.

The dataset consists of 131,895 records and 16 independent and mostly numerical attributes, representing textual reviews, service rankings given by the customer, aircraft, and passenger information.

The dataset incorporates the following characteristics:

- Airline: the name of the airline company (i.e. Lufthansa, Egyptair)
- Aircraft: the make and brand of the plane
- Cabin: the class of the seat (business, economy, or other)
- Date flown: the date of the flight
- Author: the name of the passenger and review holder
- Review date: the date when the author left the review on the website
- Traveller type: a type of travel (i.e. business or leisure)
- Customer review: the review in the textual format
- Seat comfort: a comfort level of the seat in the plane
- Cabin service: a satisfaction with service in the cabin
- Food and beverages: an enjoyment with food and drinks
- Entertainment: content with amusement options on the board
- Ground service: an off-board service rating
- Value for money: a ranking of how price corresponds to in-flight service
- Overall rating: a total score left by the passenger
- Recommended: a binary value, indicating whether the customer recommends the airline or not

Around 30546 passengers favor the airline they flew with, while 33894 do not recommend it, with 47% and 53% satisfaction ratios, respectively. This illustrates that the dataset is fairly balanced, but it contains a large number of missing entry points, which should be further filtered out to avoid redundancy in the results.

Dataset collection

Skytrax, a UK-based consultancy organization, is popular for its airline ranking and review system. This dataset was web-scraped from Skytrax's website as part of a capstone project at MEF University Big Data Analytics program for 2018-2019 in the spring of 2019 using the Scrappy tool. Data encompasses customer reviews from 2006 to 2019 for well-known 82 airline organizations such as Turkish Airlines, Lufthansa, Air Canada, KLM Royal Dutch Airlines, Emirates, British Airways, Avianca, Korean Air, All Nippon Airlines, and etc.

3.1.4 Sky Insights

Dataset description

The dataset [77] is a real-life dataset on Kaggle, which provides a wealth of information about customer experience and airlines. It has a variety of applications, including hidden pattern identification and customer behavior analysis via sentiment analysis, time series analysis, and satisfaction prediction.

There are 23171 data points and 20 features, describing airline name, aircraft make, customer review and its date, verification status (whether the review is verified by the airline company or not), ranking scores for on-board service and amenities, and final overall rating given by the passenger.

50% of passengers rated their flight experience with a score of 1.0 out of 5.0, which suggests that data is imbalanced. In fact, 7807 customers recommended the flight, and the remaining 15364 reported discontentment with their trip.

Dataset collection

The aviation reviews dataset was collected from the AirlineQuality website. The author used several gathering methods, which are scraping, url formation, and review data scraping. The scraping process involved extracting the names of all airlines. URL formation was carried out by structuring each airline review page according to the website's format. Review data scraping was performed by retrieving customer reviews for each page. The data collection involved the following technology stack:

- Beautiful Soup - utilized for website data scanning and extraction
- Pandas - employed for data retention and CSV file generation
- Requests - applied to fetch web pages via HTTP calls
- Unicodedata - used to handle the processing of Unicode text formats

The data was collected across 497 airline organizations such as AB Aviation, Aer Lingus, Germanwings, Estonian Air, Lufthansa, Nordavia, and many more.

3.1.5 British Airways

Dataset description

The dataset [78] is a collection of customer feedback for the airline company in the UK, British Airways. Use cases of this dataset cover a wide range of applications for educational and research purposes, including:

- Sentiment analysis: customer feedback fosters insights into sentiments
- Service quality assessment: various services of the organization can be assessed
- Route performance: specific flight routes may be examined based on reviews
- Aircraft experience: customer reports provide opinions on certain aircraft.

It has 3701 records and 19 categorical and numerical features, which are customer reviews, name of the passenger, date of the feedback left, the status of the review, traveler type, seat class, flight destination, rankings on comfort and staff service, ratings for ground service and value, plane mark, internet connectivity scores, and etc. The prediction attribute is defined by the overall grade given by the customer. The total rating ranges from 1 to 10, where 1 represents the highest degree of discontent and 10 illustrates the utmost satisfaction level.

Table 1. Overall Rating Distribution

Overall Rating	**Number of Instances**	**Distribution**
1.0	865	23.372%
2.0	424	11.456%
3.0	406	10.970%
4.0	370	9.997%
5.0	330	8.916%
6.0	313	8.457%
7.0	310	8.376%
8.0	254	6.863%
9.0	234	6.322%
10.0	190	5.271%

There is a significant skew in the distribution (see Table 1). Low rankings (1.0-5.0) make up 60%, whereas high rankings (6.0-10.0) only form approximately 40%, making them largely underrepresented. The most dominant class is 1.0 rating, which accounts for about 23% of the data.

Dataset collection

The dataset was extracted from *AirlineQuality* website using a scraping technique. The authors web scraped all customer reviews of British Airways spanning 12 years, from October 9, 2011, to November 19, 2023. Afterward, all null variables have been deleted.

3.1.6 Airline Reviews

Dataset description

This dataset [79] comprises customer reviews of the 10 highest-ranked airline companies across the globe. It delivers insights into various aspects of airline customer satisfaction, featuring airline and passenger information, and estimation of service quality.

It covers data from such airlines as Singapore Airlines, Qatar Airways, All Nippon Airways, Emirates, Japan Airlines, Turkish Airlines, Air France, Cathay Pacific Airways, EVA Air, and Korean Air. The dataset contains 8100 instances and 16 attributes, except for the dependent variable. The prediction variable has a binary categorical nature with values yes and no. The satisfaction ratio is 53% to 47%, the former being those that are satisfied and the latter representing dissatisfied ones. This indicates that the dataset is only slightly imbalanced.

3.2 E-commerce Datasets

3.2.1 OlistBr

Dataset description

OlistBr [80] is a Brazilian e-commerce public dataset from a popular store in the area. Its attributes let to analyze an order from various perspectives: status of the order, price, payment, performance of freight forwarders, product features, and written customer reviews.

The data is separated into multiple smaller datasets such as reviews, orders, customers, products, sellers, and geolocation ones. The orders dataset contains features of items, price, and shipping, while the reviews one represents feedback from clients. Customer, products, and seller datasets have corresponding attributes: customer and seller name with addresses, and products presented in the store. The geolocation dataset comprises zip code details and coordinates to measure distances between retailers and recipients.

The dataset has 112,000 instances and 14396 reviews. The target variable is the reviews score, which is a numerical value given by the customer, ranging from 1 to 5. Higher values correspond to more satisfied clients. The satisfaction ratio constitutes 31.7% of dissatisfied consumers (scores 1, 2, and 3) and 68.3% of satisfied ones (scores

4 and 5), which illustrates an imbalanced nature of the data.

Dataset collection

Olist is the largest department store in Brazil, which allows leading merchants and small businesses to sell and ship their products through the company. After a customer purchases some product, the company delivers the order and notifies the customer to fulfill the survey data. The survey encourages consumers to share their purchase experience and indicate their satisfaction levels. Olist collected around 100,000 real commercial records of all orders made in different marketplaces around the country from 2016 to 2018. The customer information has been anonymized and all references to sellers have been replaced with synthetic data.

3.2.2 Dress Sales

Dataset description

Dress Sales [81] is the collection of dress sales records. The primary purpose of this dataset is to allow extraction of the most significant features by performing exploratory data analysis, though it may also be applied in the satisfaction prediction domain. The origin of the dataset remains unclear; however, records illustrate it was created for a Kaggle competition.

The data is split into two datasets: attributes and records. The attribute dataset contains 13 feature variables such as dress ID, style (casual, vintage, etc.), its categorical price (low, medium, high), size, neckline type, sleeve length type, the fabric of the attire, material, pattern, decoration, and recommendation. The recommendation is a dependent variable, which accepts only binary values: 0 as not recommended and 1 as suggested for purchase. This attribute defines the nature of customer satisfaction. The record dataset contains 24 columns, which consist of dress ID and dates ranging from August 29 to December 10, 2013.

Both datasets comprise 479 dress sales instances and all of them are unique values. There are 274 records labeled as not recommended, whereas the remaining 205 have high rating scores. These correspond to 57.2% and 42.8% of the entire dataset, which highlights the presence of subtle data imbalance.

Some rows hold NaN (not a number) values, indicating unrecorded data points in the dataset. Their removal may result in the shrinkage of the dataset size, especially considering its relatively small original volume.

3.2.3 Shopzilla

Dataset description

The dataset [82] covers a one-month period of real orders made at the e-commerce platform, Shopzilla. This data is particularly designed for conducting customer satisfaction forecasting, exploratory data analysis, and visualization.

There are 85907 unique instances of orders and 20 defining variables, which include channel name, category/sub-category of interaction, customer feedback, order identifier and date, customer name and location, item category and price, names of customer service agents and supervisors, customer satisfaction (CSAT) score.

Certain variables retain null entry points. The prediction variable is the CSAT score, ranging from 1 to 5. Scores of 4 and 5 indicate buyer contentment, while grades 1, 2, and 3 reflect a negative impression about the purchase. Around 82% of customers marked their experience in a positive way, whereas the remaining 18% experienced dissatisfaction. The data is too skewed, demonstrating the huge imbalance.

3.2.4 Amazon

Dataset description

This [83] is the collection of women's dress sales records on Amazon. Amazon is the largest e-commerce online platform, which sells products of various types, including dresses. This dataset is a selection of authentic customer reviews made on this platform.

It has 23486 entries and 11 primarily categorical attributes, encompassing customer age, type of apparel bought, garment category, class of the clothing, unique product identifier, text review left by the buyer, and numerical variables like the count of customers who share the same opinion on this feedback, the rating assigned to the product and purchase recommendation status (0 is no and 1 is yes).

There are 19314 positive and 4172 negative reviews, which roughly represent 82% and 18%. Following the pattern in [9], the data is notably unevenly distributed. Additionally, a small fraction of instances for variables "review title" and "review text" are missing, and their elimination is expected to have minimal influence on the dataset's balance.

3.2.5 CFS

Dataset description

Customer feedback and satisfaction (CFS) dataset [84] is a synthetically generated dataset, which is built to evaluate and forecast customer satisfaction through various demographic and behavioral patterns.

It holds data from 38444 buyers, capturing their reviews on products and services. There are 11 features, spanning customer identifier, age, sex, country of living, annual income, loyalty category, frequency of purchase in one year, ratings on product and service quality, overall feedback, and satisfaction score. The satisfaction score ranges from 1 to 100 percent. Around 30870 customers rated above 70%, 7048 gave a score between 40% and 70%, and 526 provided a score below 40%. Nevertheless, this does not provide information on data distribution as the threshold may be adjusted to reach data evenness, though low-level records are scarce.

3.2.6 CBSH

Consumer behavior and shopping habits (CBSH) dataset [85] provides insights into customer purchase behavior, patterns, trends, and preferences. It has a wide range of applications: exploratory data analysis, customer behavior analysis, consumer shopping pattern identification, and satisfaction estimation.

There are 3900 records of consumer reviews. The dataset contains a distinct range of attributes, encompassing consumer ID, age, gender, purchased product, category of the item, geographical location of the purchase made, size, color, season, history of purchase, payment method used, discounts, and promo codes applied, shipping methods, subscription status of the customer, and review rating from 1.00 to 5.00.

Table 2. Review Rating Distribution

Review Rating	**Number of Instances**	**Distribution**
2.50 - 2.75	379	9.718%
2.75 - 3.00	306	7.846%
3.00 - 3.25	471	12.077%
3.25 - 3.50	334	8.564%
3.50 - 3.75	461	11.820%
3.75 - 4.00	305	7.820%
4.00 - 4.25	500	12.820%
4.25 - 4.50	305	7.820%
4.50 - 4.75	461	11.820%
4.75 - 5.00	378	9.692%

According to Table 2, the threshold can be established to balance the data since it is more or less distributed evenly across various rating ranges. Also, there are no null entry points, which removes the necessity to filter the data. However, there is no information provided on the nature of the data - whether it is real or synthetic. Additionally, reviews for ratings below 2.5 are absent.

3.3 Other Datasets

3.3.1 Hotel Reviews

Dataset description

This [86] is a publicly shared dataset from a large travel platform. The dataset is directly connected to the customer satisfaction problem, yet it requires sentiment analysis to extract meaningful data from textual reviews.

The data consists of two datasets: offerings and reviews. The offerings dataset has

4333 instances and 9 attributes, describing the name of the hotel, its address, phone number, class, region, type (i.e. hotel or motel), link, and details of stay. Each hotel appears to be unique in the dataset. While most attributes give a detailed description of the hotel, some turn out to be useless in the context. Phone number and details attributes are missing for all hotels, whereas the hotel type attribute features only one type (hotel). The reviews dataset comprises 878561 reviews with the following variables: service ratings, textual review, author information, date of stay, and online reservation type (via mobile or not). Date stayed and date attributes provide the same information, the latter giving more details. Some reviews are written in the French language, which may limit their usefulness for global research.

Dataset collection

TripAdvisor is an online travel planning platform, which operates many trip agencies to help tourists scheme their journey. The platform provides information on flights, restaurants, hotels, travel activities, excursions, user experiences and reports. The dataset has been crawled from real-life reviews of customers by representatives of Carnegie Mellon University. It contains reviews from around 4333 hotels.

3.3.2 Wireless Networks

Dataset description

This [87] is a synthetic dataset created to measure AI-based personalized wireless network user satisfaction. The data consists of several datasets, and each dataset is generated for a one-year timeframe. The data is distributed among four different persons: a jobholder, a university student, a high school student, and a housekeeper. All of them have the same features: date of the record, time, time period, speed of the wireless network, minimum and maximum rates, customer location, activity type, user ID, and satisfaction score (from 0 to 5).

3.3.3 Voice Call

Dataset description

The dataset [88] is the collection of customer experience and satisfaction with voice call quality in India. The consumer feedback was collected by the application called TRAI MyCall from September to November of 2019. This app gathered reviews about consumer perception of the service in real-time.

It has 33812 records and 8 independent variables, including operator name, network type, caller coordinates, state, satisfaction with voice call service, and rating score. The data is quite imbalanced since a large proportion of users are satisfied with the service, whereas the number of unsatisfied people remains low.

3.4 Selected Datasets

For this study, the following datasets have been selected: US Airlines, Sky Insights, Airline Reviews, and OlistBr - from a broader pool to investigate customer satisfaction prediction trends.

The table below summarizes the implications of selecting datasets, their limitations, and additional information, including data size and field.

Table 3. Dataset Selection

Dataset	**Size**	**Field**	**Limitation**	**Selection Outcome**
US Airlines	129,880	Airline	The dataset collection methods remain unknown	Selected for this research since it has comprehensive reviews and is widely cited in many literature.
Invistico	129,880	Airline	No information on data provenance and the source of the dataset remains unknown	Removed from this research due to quite a similar resemblance of the previous dataset.

Skytrax	131,985	Airline	It contains a large proportion of missing entry points (over 50% for each column).	Removed from this selection.
Sky Insights	23,171	Airline	High skewness towards discontent customers with 66/34 ratio	Selected for this research.
British Airways	3,701	Airline	High skewness in the data with 60/40 ratio and relatively small dataset.	Removed from the selection.
Airline Reviews	8,100	Airline	The dataset spans a one year flight records	Selected for this research
OlistBr	112,000	E-commerce	It is an imbalanced dataset with 68/32 ratio.	Selected for its use in prior literature and as a model of the e-commerce industry.
Dress Sales	479	E-commerce	There are many unrecorded data points and it has a significantly small sample size	Removed from the selection.
Shopzilla	85,907	E-commerce	There is a substantial asymmetry towards satisfied customers with an 82/18 ratio. Over 80% of records on customer and product information are missing.	Removed from the selection.
Amazon	23,486	E-commerce	There is an extreme skewness towards satisfied customers with an 82/18 ratio. A number of features is too limited.	Removed from the selection.

CFS	38,444	E-commerce	The synthetic dataset does not represent real-life customer feedback for drawing real-world conclusions. Over 50% of its records consist of maximum satisfaction scores.	Removed from our selection.
CBSH	3,900	E-commerce	No information on data origin - whether synthetic or not. A relatively small dataset	Removed from the selection.
Hotel Reviews	878,561	Tourism	There are features with missing records and some reviews are written in French, making it too language-specific.	Removed from the selection due to computational limitations.
Wireless Networks	334,561	Telecom	It consists of many sub-datasets, and some of them are inaccessible. Also, the dataset is synthetic.	Removed from the selection due to computational reasons.
Voice Call	33,812	Telecom	The dataset is greatly imbalanced with many users being satisfied. It spans a relatively short timeframe (10 months). A number of features is too limited.	Removed from the selection due to its drawbacks.

3.4.1 Data Preprocessing

US Airlines dataset

- Features "Unnamed: 0" and "id" were removed as they are unique identifiers and does not include useful patterns. Including these features run a risk of over-fitting or poor generalization.
- Categorical features "Gender", "Customer Type", "Type of Travel", and "Satisfaction" were one-hot encoded, while feature "Class" was ordinally encoded due the to presence of hierarchical meaning in its values.
- Null values have been removed (393 rows from 129,880 records).
- Data normalization was performed using MinMaxScaler.

Sky Insights dataset

- Features "Unnamed:0", "Review", "Review Title", "Review Date", "Date Flown", and "Verified" were removed for not carrying meaningful information to the prediction. These features were either unique identifiers, unprocessed pure textual fields, or irrelevant to the target variable.
- Features "Aircraft", "Inflight Entertainment", and "Wifi & Connectivity" were dropped for the high presence of missing values (over 50%).
- Categorical features "Satisfaction" was one-hot encoded and "Type Of Traveller", "Origin", "Destination", and "Airline Name" were label encoded, while "Seat Type" was ordinally encoded due to the presence of hierarchical meaning in its values.
- Null values have been filled out using KNNImputer for columns "Seat Comfort", "Cabin Staff Service", "Food & Beverages", "Value For Money", and "Ground Service".
- Data normalization was performed using MinMaxScaler.

Airline Reviews dataset

- Features "Title", "Verified", "Reviews", and "Month Flown" were removed for not contributing meaningful information to the prediction. These features were either unprocessed raw textual entries, or unrelated to the outcome variable.

- Features "Aircraft", "Inflight Entertainment", and "Wifi & Connectivity" were dropped for the high presence of missing values (over 50%).
- Categorical features "Satisfaction" was one-hot encoded and "Type Of Traveller", "Origin", "Destination", "Airline", and "Name" were label encoded, while "Seat Type" was ordinally encoded due to the presence of hierarchical meaning in its values.
- Data normalization was performed using MinMaxScaler.

OlistBr dataset

- Target feature "satisfied" was created based on the categorical variable "review score" (1-3 indicated dissatisfaction and 4-5 illustrates satisfaction)
- Categorical features "satisfaction" was one-hot encoded and "customer_city", "customer_state", "seller_city", "seller_state" "order_status", "payment_type", "product_category_name_english" were label encoded.
- Data normalization was performed using MinMaxScaler.

Chapter 4

Methodology

This chapter provides a short overview of the methods used in the literature in the context of customer satisfaction prediction activity. Each method is briefly described in terms of its structure, implementation, practical application, operational logic, and limitations. Later, there is a section concluding the selected methods for this research.

The upcoming sections provide an examination of the following methods:

- Regression modeling (Linear Regression, Logistic Regression)
- Decision Tree (ID.3, C4.5, CART, and etc.)
- Instance-based algorithms (KNN)
- Kernel methods (SVM)
- Bayesian learning (Naive Bayes, LDA)
- Ensemble learning (Bagging, Random Forest, Gradient Boosting, XGBoost, AdaBoost, CatBoost)
- Artificial neural networks (ANN)

There are some methods that include a combination of the aforementioned algorithms (so-called hybrid models). Others deploy various modifications of neural networks to forecast consumer contentment on time series data (i.e., Long Short-Term Memory (LSTM))

4.1 Regression modeling

Regression modeling encompasses two major techniques: *linear regression* and *logistic regression*. Both approaches may be applied to prediction activities, though the latter is particularly appropriate for binary consumer satisfaction analysis [15].

4.1.1 Linear Regression

Linear regression is a supervised learning algorithm that estimates the relationship between input and target attributes by fitting a linear equation to data. Its primary objective is to locate the best-fit line to minimize the gap between the true and predicted values. The algorithm computes the line using the formula: $y = \theta_1 + \theta_2 x + \ldots + \theta_n x_n + \epsilon$

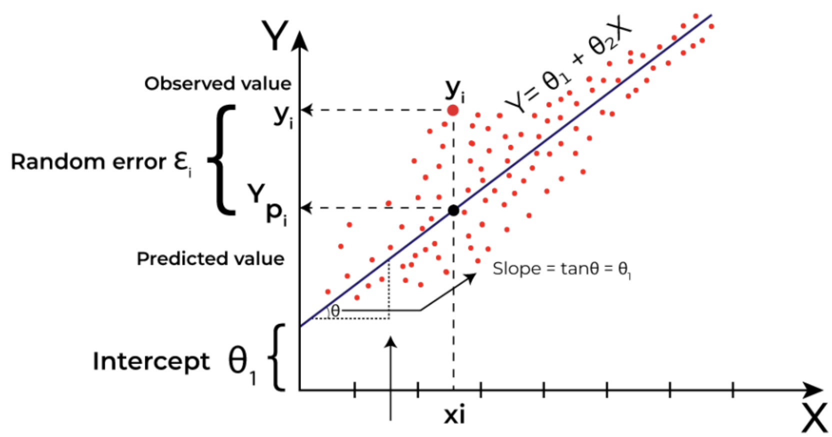


Figure 4-1. The architecture of linear regression

Here, y is a target variable, while x is a dependent attribute (also known as a predictor of y). The variable x may represent either a single input feature or a collection of multiple parameters, demonstrating the problem. θ_1 is an intercept, $\theta_2 ... \theta_n$ are coefficients of x and ϵ is some random error.

The algorithm is easy to understand and works well with linearly separable data. Especially, it is good at handling small datasets. However, it does not generalize well on large polynomial data. Additionally, the usage of linear regression is limited to the prediction of continuous values, as stated in [16]. So, it is of little value in the context of certain customer satisfaction estimation tasks (i.e. target feature may take categorical values such as true/false or 0/1 to showcase contentment experience.)

4.1.2 Logistic Regression

Logistic Regression is well-suited for the purpose of customer satisfaction prediction, illustrating its effectiveness in the works of [16] and [47]. Logistic regression accomplishes binary prediction challenges, where it is necessary to classify elements of a set into two groups. Its objective lies in the idea of determining the binary relationship between features. The algorithm delivers a dichotomous outcome: yes/no, true/false, or 0/1, as illustrated in the study led by [33]. It estimates the mathematical probability of whether the value corresponds to the category or not; therefore, the function is capable of forecasting satisfied and dissatisfied consumers. The logistic function is of the form:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)}} \tag{4.1}$$

where $p(x)$ is a probabilistic output between 0 and 1, x is an explanatory variable, β_0 is an intercept, and $\beta_1 \ldots \beta_n$ are slopes of the log-odds as a function of x.

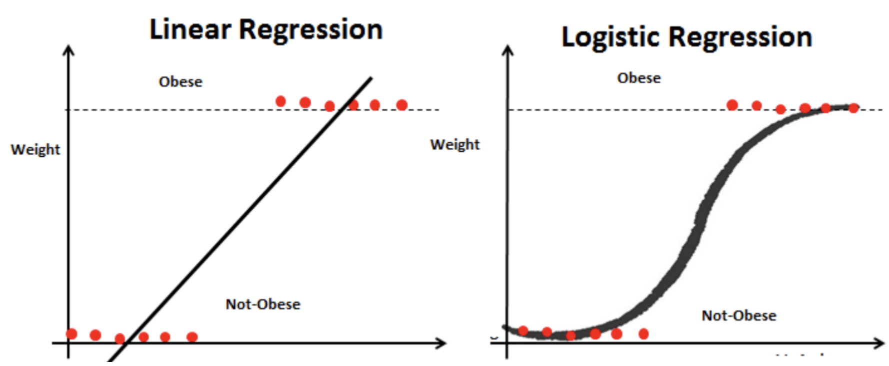


Figure 4-2. The comparison of linear regression and logistic regression

Figure 4-2 illustrates the logistic regression model in contrast to linear regression, highlighting the dominance of non-linear decision boundary over straight-line approximation in binary classification.

Logistic regression is one of the earliest machine learning methods widely used in consumer satisfaction prediction today, even with the rise of more advanced machine learning techniques. Its continued popularity is due to its computational efficiency, simplicity, and ease of interpretation. Meticulously, the coefficients are attributed to each feature, allowing us to smoothly estimate their significance to the target variable. This makes it a solid baseline choice with well-defined decision boundaries, which can be observed in [50], [54].

However, the performance of logistic regression depends on specific conditions — ironically, these same conditions also highlight its limitations:

- There should be no strongly influential outliers. Though some customer data have extreme values, it can be mitigated by either removing or standardizing them, which was done in the experiments by [52].
- The dataset size is preferred to be large. This does not seem to pose a challenge due to an abundance of customer satisfaction data and advancements in storage, which is supported by several research [47], [48], [55].
- The observations ought to be independent of each other, meaning one entry point should not affect another or contain duplicates. This requirement is not specific to the domain of this study, but if dependencies arise, they can be handled by either dropping one of the variables or merging them into a single attribute.
- No or low correlation between independent variables is expected. Multicollinearity makes the logistic model unstable and contributes to incorrect coefficient estimation, decreasing the overall accuracy. Often, features such as "on-boarding service quality" and "travel class" are highly collinear with each other. For instance, business class passengers receive high-end service compared to those who hold economy class tickets, as evidenced in the work of [44].
- The dependent feature must be categorical. The satisfaction feature usually takes binary values or ratings from 0 to 1, which can be modeled with logistic regression.

Most of the assumptions hold true to the domain of this research; though multicollinearity primarily impacts the predictive performance of logistic regression, causing it to underperform compared to more advanced models, as demonstrated in experiments held by [54], [55], [71].

4.2 Decision Tree

Often, regression models fail to capture complex nonlinear relationships in feature space. Although some might argue that logistic regression allows for some flexibility, it still relies on linear decision boundaries [44]. Decision Trees serve as an effective resolution to overcoming this limitation [57]. Decision Tree is a non-parametric supervised learning approach used for both classification and prediction activities. It has a tree-like hierarchical structure, which consists of root nodes, leaf or terminal nodes, internal nodes, and branches.

Figure 4-3 illustrates the structure of the decision tree. Each node serves a distinct role and acts as a decision point.

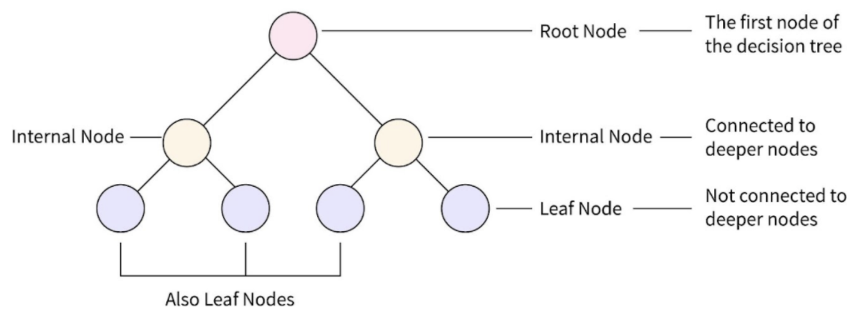


Figure 4-3. The structure of decision tree

The root node is a starting top-most node, which depicts the entire dataset that is selected based on the best feature. The root node is split further into internal nodes, which represent questions or conditions that are directed to specific paths. They are intermediary nodes between the root and leaf nodes. The leaf nodes are the end nodes, where the final predictions occur. They do not split up further and denote the predicted outcomes. Branches are connections between these nodes.

The primary goal of this approach is to generate a robust prediction model by learning simple intricate decision rules from data and splitting the data at various thresholds. In short, it performs a piecewise approximation at each step. At each stage, it computes impurity values to find the best split, depending on the nature of the target variable. It uses gini impurity (a measure of impurity) or entropy (a measure of disorder) for classification purposes and variance for continuous values. These mathematical scores dictate the decision-making process since a failure to exploit them led to decreased performance in [58].

Figure 4-4 portrays the example of the decision tree in the context of customer satisfaction estimation. In this flow chart, the level of satisfaction of the consumer is based on characteristics: quality and cleanliness of the bathroom.

Decision trees can be powerful models, but they heavily rely on specific critera, which

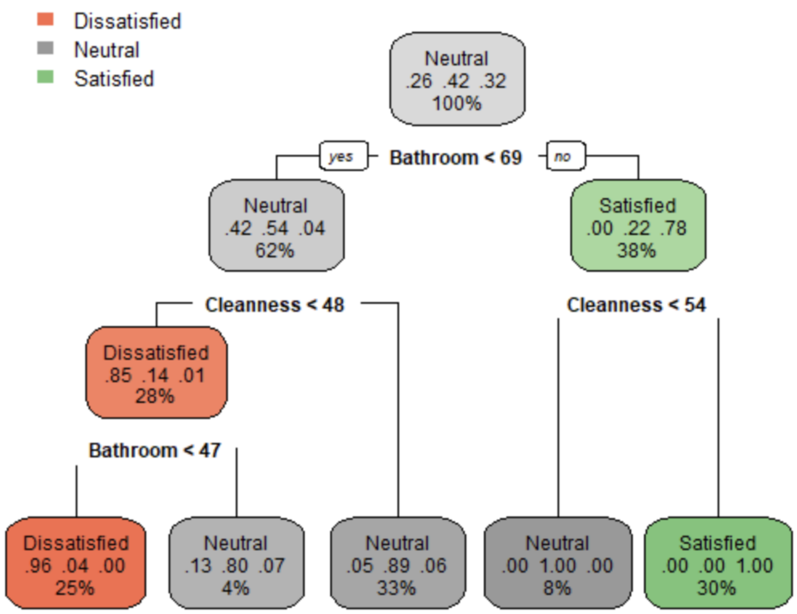


Figure 4-4. The example of decision tree in customer satisfaction domain

should hold true in order to get the desired performance:

- The depth of a tree should be regulated. Overly deep models can generate too complex trees, which recognize more noise than meaningful patterns and lead to over-fitting. Too shallow ones may not capture complex relationships in customer satisfaction data, leading to deteriorated performance. Methods such as pruning, setting a maximum depth of a tree (5-20), and installing a minimum number of samples (2-10) at the leaf node may resolve the issue.
- The dataset ought to be balanced since it may create biased trees. If there is an excessive number of satisfied customers than dissatisfied consumers or vice versa, then the model may favor the majority class. Based on some examples of literature [33], [48], [52], it is recommended to apply data balancing techniques such as over-sampling, under-sampling, and SMOTE if the class imbalance is severe.
- The dataset size has to be sufficiently large since a tree may become overly complex on too small data. Usually, the customer satisfaction dataset has enough samples. The datasets that are to be used in this research contain around 20,000 - 130,000 records ([74], [80], etc.)
- The number of features can significantly impact the predictive performance of a tree.

Too many features push the tree to split excessively. Some features may not provide meaningful information and should be ignored. It is a good practice to perform feature selection to avoid this issue.

- The missing values in the data have to be minimized. Unlike many methods, decision trees can actually handle missing data points. However excessive gaps may impact the performance; therefore, it is best to remove them. In this study, these values were deleted for most of the datasets, except for cases with smaller sample sizes, where KNNImputer or mean/median computation was applied.

- Trees are sensitive to outliers; hence, extreme values should be deleted. This was evident in the works of [53], [43], where the DT model struggled with unfiltered data.

- Decision outcomes should be yes/no outputs and proceed further. Satisfaction variables are mostly binary in nature, which is beneficial for decision-making.

- Outcomes vary from the selected mathematical measure. In this work, gini, entropy, and log-loss were used, following a similar approach in [50].

The advantages of a decision tree include its simplicity, ease of visualization, high interpretability, low cost, and good performance. Trees use a white box model; so, explanations for conditions are easily explained. Nevertheless, trees are not well-suited for linear relationships and they sometimes become unstable.

There exist several variations of decision trees such as CART (Classification and Regression Trees), ID3 (Iterative Dichotomiser 3) algorithm, C4.5 function, C5.0 algorithm, and etc. Commonly, CART is the default built-in choice for decision tree models, but C4.5 and ID3 are also widely recognized and applied, as seen in studies such as [56] for the former and [57] for the latter.

The trees are widely used in the customer satisfaction domain. In some studies such as [46] and [47], decision trees demonstrate a more robust performance over other models, but, in others, they are predominantly outperformed by ensemble learning methods such as Random Forest, XGBoost, and CatBoost (i.e., [53], [68], and etc.).

4.3 Instance-based algorithms

4.3.1 K-Nearest Neighbours (KNN)

K-Nearest Neighbours (KNN) is a supervised learning strategy, which can be employed for both classification and regression problems. It is one of the popular and simple methods utilized for customer satisfaction experiments, which is supported by [54] and [59]. This method relies on proximity-based on-the-fly decision-making rather than computed parameters or decision rules. Specifically, KNN does not perform a training stage; instead, it retains the entire dataset and retrieves relevant instances at the time of classification.

Typically, a class label is determined by the greatest number of assignments, where the label occurring most often amidst adjacent instances is chosen. However, a decision can be made even if the class receives around the quarter of the selections, depending on the number of categories, as stated by [95].

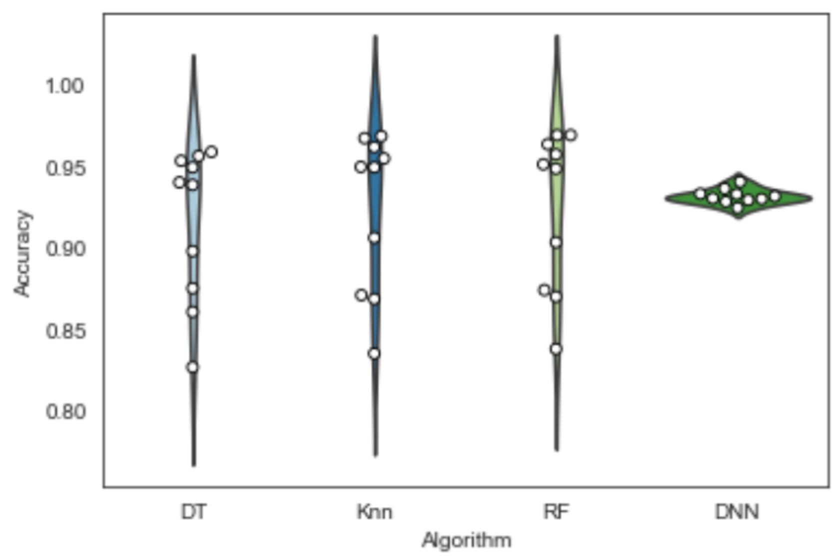


Figure 4-5. The comparison of accuracy scores across KNN and other models on the [87] dataset, as presented in the work of [58]

KNN determines proximity to classes by computing a distance from the target point to its neighbors. There exist several distance metrics, including Euclidean, Manhattan, and Minkowski distances. Euclidean distance measures the shortest distance between two points, while Manhattan distance calculates a total path along horizontal and vertical lines. Minkowski distance serves as a broader metric that blends these two. [48] and [67] used

Euclidean distance in their experiments, while [33] relied on Manhattan distance.

The advantages of KNN include simplicity, ease of implementation, absence of assumption about data distribution, and ability to handle multiple classes. However, the performance of the KNN algorithm heavily depends on meeting certain conditions. If the following assumptions are not met, KNN may suffer from poor accuracy and high computational costs:

- Sparse sample size may lead to failure to identify data patterns. One must ensure the data is sufficiently large and well-distributed. However, computational expenses grow with dataset size.
- The choice of a number of neighbors (k) is important since very small k makes KNN sensitive to outliers (over-fitting) and too large k leads to under-fitting. The cross-validation can help to select the right number, avoiding the curse of dimensionality. [33] and [58] utilized GridSearchCV to effectively identify the right k.
- The right distance metric matters. Euclidean distance generalizes well on continuous sets, while Manhattan distance is suited to grid-like pattern data. Minkowski distance can provide a best-fit line, but it requires some experiments.
- KNN requires a large storage space since KNN is memory-intensive.
- KNN is sensitive to outliers and noise in the data. Mislabeled points may distort predictions, leading to poor performance, as reported by [50] and [54], where KNN could not converge well.

Several studies, including [48], [52], [68] applied KNN to predict consumer satisfaction, but it did not outperform other models, suggesting that more advanced methods achieve higher accuracy, as observed in Figure 4-5. Nonetheless, KNN remained one of the top-scoring approaches.

4.4 Kernel methods

Kernel methods are a group of machine learning strategies that map data into multi-dimensional space to discern non-linear decision boundaries. Some datasets are not linearly separable, suggesting the sophisticated polynomial nature of the data. Linear models (i.e. Linear Regression, Linear SVC) often struggle with capturing complex relationships, as they rely on straight-line boundaries. On the other hand, kernel methods circumvent this limitation by extending models' capacity to identify intricate data structures [51].

4.4.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a kernel-driven approach employed for classification, regression, and outlier detection. The foundations of the method were laid by Vladimir N. Vapnik in the middle 1990s and maintained a strong appeal within researchers [34]. The method supports diverse data patterns, ranging from simple relationships to complex polynomial structures.

The aim of the SVM algorithm is to determine the separating hyperplane that ensures the widest distance between distinct classes (see Figure 4-6). The N-dimensional hyperplane is created by transforming data into high-dimensional space.

For linear data, SVM directly finds decision boundaries in the original space. Mathematically, the equation for linear hyperplane can be denoted as:

$$\omega^T x + b = 0 \tag{4.2}$$

ω is the vector perpendicular to the hyperplane and b represents how far the hyperplane is from the origin along the vector ω.

Meanwhile, non-linear data requires transformations on feature space [96]. The nature of transformation depends on kernel functions: linear, polynomial, and radial basis function (RBF) kernels. Linear kernel is used for linear separability, while polynomial function maps data into higher dimensions with the former being used by [49] and the latter being employed by [33]. RBF kernel (also known as Gaussian function) projects data into new space based on distance-dependent similarity, which was used by [34], [63].

SVM is effective in handling high-dimensional data and robust to extreme values in the data, as stated by [48]. Also, it is less prone to over-fitting issues in the data, confirming the trend identified in the research by [71], where evaluation metrics for training and test sets remained close. However, SVM becomes computationally expensive and slow for a larger sample size. It is sensitive to noise in the data and has a low interpretability. Additionally, the predictive performance of SVM highly depends on the choice of the right kernel function.

[52], [53], [55], [59] employed various SVM algorithms for customer satisfaction pre-

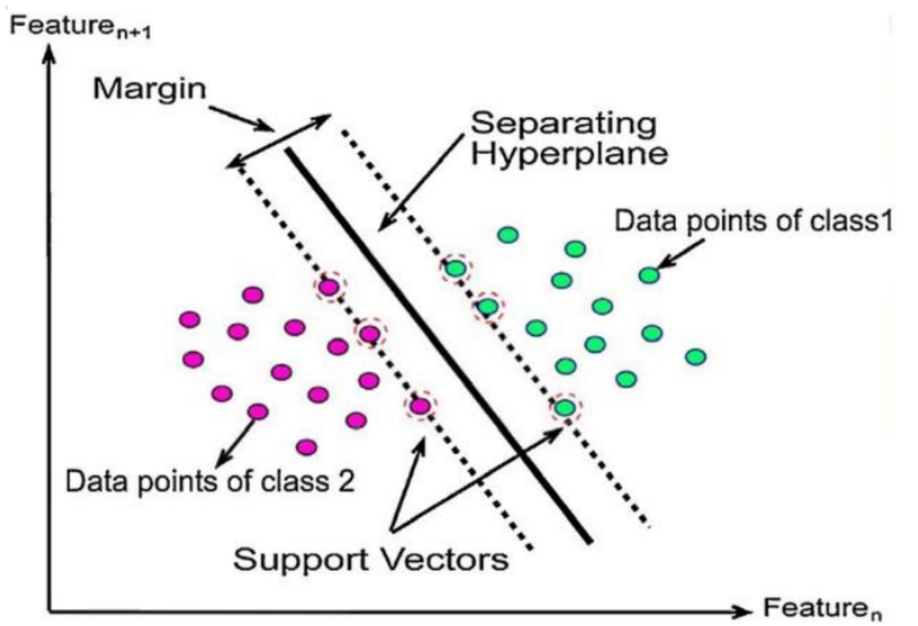


Figure 4-6. The architecture of SVM algorithm, as presented in the work of [55]

diction in their studies. It is evident that SVM is a strong state-of-the-art performer and can surpass most models when the right data transformations are applied, as shown in [34] and [51]. However, occasionally, it is outclassed by advanced Bayesian, ensemble, and deep learning methods.

4.5 Bayesian learning

4.5.1 Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is a parametric machine learning technique used for multi-class classification purposes. It reduces the dimensionality of data while keeping the most significant features. The algorithm tries to find a linear combination of features that best discriminates between classes. LDA works by widening the gap between groups and minimizing the variance within each category.

The class affiliation is determined based on predictions by using Bayes' theorem. Bayes computes conditional probabilities - a probability of an event given the fact that some other event occurred. Taking the logarithm of Bayes' theorem transforms the probability equation into a straight-line decision boundary. The decision boundary is formed at the point where class probabilities are equal.

For LDA to achieve high performance, some assumptions must hold true:

- Data should have Gaussian distribution. Otherwise, data re-shaping techniques (log transformation or square-root scaling) can help normalize the data.
- Covariance matrices of various classes should be equal.
- Data must be linearly separable. LDA does not perform well on non-linear data. Therefore, it is recommended to switch to non-parametric methods (Decision Trees, Random Forest, KNN, and etc.), which do not assume a certain data distribution.

[59] contrasted LDA with other approaches, where it achieved the lowest accuracy on customer satisfaction data, suggesting that exploring alternative methods could yield better results.

4.5.2 Naive Bayes

Naive Bayes (NB) is a generative machine learning algorithm, meaning that it determines the probabilistic likelihood of features appearing in each category and uses that information for classification. Unlike discriminative models (i.e. Logistic Regression, SVM), it does not discern feature importance to make predictions, as emphasized in the works of [49] and [71].

The name "Naive" comes from the assumption that there is at least one feature, which is independent of other parameters. The "Bayes" part refers to the use of Bayes' theorem as its foundation. The theorem facilitates the calculation of probabilities for predictions. NB algorithm assigns data points to classes by maximizing the posterior probability [48]. The mathematical equation of the NB algorithm can be represented as:

$$P(Y|X) = \frac{P(X \cap Y)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)} \tag{4.3}$$

Here, $P(Y|X)$ is a posterior probability, which is a probability of event Y given the observed event X. Y can denote the probability of the customer being satisfied and X is some factual event (for instance, flight distance of over 1000 km, arrival delay of 25 minutes, and business ticket class from [76] dataset). $P(X|Y)$ is a conditional or likelihood probability, which is a probability of event X (flight distance of over 1000 km, arrival delay of 25 minutes, and business ticket class) given the chance of event X (achieving of consumer contentment).

$P(Y)$ is a posterior likelihood of customer satisfaction and $P(X)$ is the probability of event X.

Some researchers [33], [49] used the Naive Bayes algorithm in their experiments on various domains to determine whether customers are satisfied or not, where more advanced methods such as Random Forest, XGBoost, and CatBoost outperformed NB. However, NB is proven to have the capability to show superior performance over these methods due to its simplicity, computational efficiency, and interpretability in customer satisfaction prediction, as illustrated by [50].

4.6 Ensemble learning

Ensemble learning is a machine learning technique, where two or more learners are trained on the data and their outputs are aggregated into one, producing reliable predictions. The learners are labeled as "weak" or "base" models, which yield a strong performance altogether. The core idea of ensemble learning is that a collectivity of multiple models leads to optimal bias-variance tradeoff by leveraging their strengths and compensating for individual weaknesses. In a real-life scenario, this is akin to asking a group of customers for their airline reviews rather than relying on single-person experience.

There are two major types of ensemble learning strategy: bagging and boosting. Figure 4-7 visually illustrates the operation flow of ensemble learning techniques.

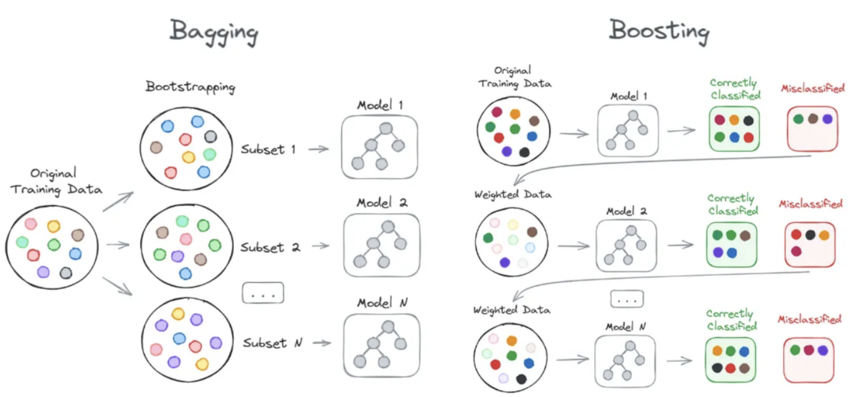


Figure 4-7. The flow of ensemble learning strategies

The bagging strategy is a parallel learning method, where each model undergoes independent training, and results are averaged. It uses replicates of original data (bootstrapped samples) to run multiple weak learners with the same algorithm and gets final results. Boosting strategy is a sequential method, where models are trained in a consecutive way, ensuring each learner builds on previous mistakes. Boosting samples the data, but prioritizes misclassified instances to create weighted data for the next model in order. Eventually, the combined weights of all learners yield end results.

The use of ensemble techniques leads to notable performance gains. Also, combining diverse ensemble techniques helps to reduce over-fitting issues without bias trade away. Indeed, an ensemble of distinct high-variance models (i.e. algorithms that tend to over-fit) has a superior performance over a single generalizable model. Likewise, this approach is able to address high-dimensional data issues and so offers a great substitute for dimensionality reduction techniques.

Despite its improved predictive performance, a few conditions should be met, for it to work effectively:

- Base models should not be identical; instead, they should have different perspectives on the data. If all models are highly correlated, then it might not enhance overall accuracy as they make similar errors. The solution is provided by the use of randomization, bagging, or boosting techniques.
- Ensemble learning requires a sufficiently large amount of training data. Models may over-fit on a smaller sample size. Considering the abundance of customer reviews, it should be an obstacle.
- Hyperparameters optimization matters for better overall accuracy. One could make a use of hyperparameter tuning techniques like GridSearchCV.

4.6.1 Bagging algorithm

Bagging is an ensemble learning algorithm, which can be used for both regression and classification. It works by dividing the data into smaller subsets and training a model on a random selection with replacement. The algorithm ensures out-of-bag (OOB) evaluation (sample exclusion) to estimate the model performance without the need for cross-validation. Either majority voting or averaging is applied to aggregate results, depending on the type of task. Afterward, bagging produces the final results for each record.

[44] used bagging for predicting airline passenger satisfaction, concluding the algorithm has immense power; yet, XGBoost outperformed it.

4.6.2 Random Forest

The first prototype of random decision forests was created in 1995 by Tin Ham Ko. Later, L. Breiman and A. Cutler released an extension of the algorithm using bagging and CART decision trees and registered it as "Random Forest" in 2006 [42].

Basically, Random Forest is a natural advancement of bagging and trees. Random Forest trains multiple tree models on bootstrapped data and combines their predictions for the ultimate outcome, maintaining diminished variance and over-fitting issues, a point referenced in several studies [50], [52], [53]. It also preserves robustness to noises and outliers in the data. Unlike linear models, Random Forest can draw non-linear decision boundaries, staying highly interpretable due to feature importance scores. This trend is consistent with conclusions drawn by [49], [55], [58], where Random Forest outclassed linear models by 10-20% on average.

Random Forest is regarded as a state-of-the-art technique for customer satisfaction estimation, making it a widely adopted method in the above-mentioned research as well as [67], [68], [71]. Among all studies, [33], [43], [47], [48], [54] illustrated the best performance for Random Forest, reaching between 80% and 95% depending on the dataset.

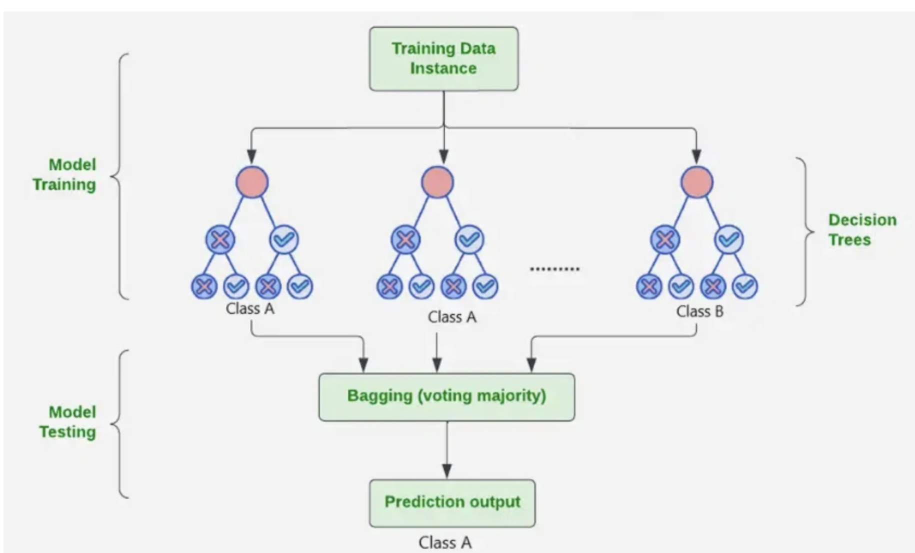


Figure 4-8. The structure of Random Forest algorithm

Figure 4-8 illustrates the Random Forest algorithm with several decision trees, which undergo a bagging process before it produces prediction output.

4.6.3 AdaBoost

The AdaBoost (Adaptive Boosting) was initially developed to enhance the effectiveness of binary classifiers in 1995 [89]. It follows an iterative approach, fusing multiple weak classifiers to generate a single strong classifier. AdaBoost uses a wide variety of models as base learners, ranging from Logistic Regression to Decision Trees.

AdaBoost algorithm starts by assigning equal weights to its base learners. During the training process, weights are adjusted based on performance with higher emphasis put on misclassified instances. Each new classifier learns from the mistakes of previous ones to improve accuracy. Ultimately, all models are amalgamated via weighted voting, assigning greater importance to more accurate learners, with the class receiving the highest votes being selected as the final output.

Owing to weighted voting and stage-wise optimization, AdaBoost resists over-fitting, making it robust. This is also driven by the algorithm being highly adaptable to changes in the learners. According to [52], the voting in AdaBoost naturally determines the feature importance, preventing irrelevant features from affecting the final outcome. Despite its robustness, AdaBoost is not immune to messy and noisy data. Additionally, it is highly sensitive to the learning rate. A high learning rate leads to data memorization, while a low learning rate demands more iteration to be effective. These challenges have already been discussed by [90].

[52] examined AdaBoost on customer satisfaction prediction, which demonstrated a strong performance; though; it did not surpass other ensemble learning methods.

4.6.4 Gradient Boosting

Gradient Boosting (GB) sequentially builds models that minimize the loss function (i.e. mean squared error or cross-entropy), turning them into a strong learner [91]. The error minimization follows an iterative optimization process using the gradient descent. The computation of the negative gradient of loss function for the current model helps to fine-tune the next weak model. The process continues until the slopping condition is satisfied.

Unlike AdaBoost, GB does not adjust the weight of training instances. Instead, each new predictor learns from previous models. This makes the algorithm significantly less sensitive to outliers, as demonstrated by [40] by exploiting Huber and Quantile loss functions in their study. Also, Gradient Boosting algorithm is more versatile as it can employ a variety of base models, including linear models and decision trees, which leads to stable error correction and learning.

Nonetheless, the effectiveness of GB can be hindered under specific conditions:

- GB has a tendency for over-fitting since it aggressively minimizes error. The issue can be addressed by limiting the tree depth or ensuring early stopping, as described by [55].
- It is highly sensitive to hyperparameters (learning rate, number of trees, tree depth or sample proportion, and so on). For instance, a model with a high learning rate may end up with a suboptimal solution, while a low learning rate slows the training phase. Tree depth increases the complexity, which can cause over-fitting. Therefore, it is of great importance to choose optimal hyperparameters. [41] employed a 0.1 learning rate and 0.8 sample proportion, which increased their prediction accuracy compared to the plain algorithm.
- The algorithm is sensitive to missing values and extreme values, however, data pre-processing resolves the issue. In the work of [49], authors cleaned the data from features that contain more than 5% of null records and removed 2 outliers before applying GBDT, increasing their performance by 4%.

These papers discovered a superior performance of the Gradient Boosting algorithm over other models to predict consumer satisfaction: [40], [41], [49].

4.6.5 XGBoost

XGBoost, short for Extreme Gradient Boosting, is an extended implementation of the Gradient Boosting algorithm. Following the principles of GB, XGBoost sequentially builds models by tuning previous learners based on errors. XGBoost uses multiple decision trees as its base learners. In addition to it, it supports built-in parallel processing for large datasets, which boosts the training speed, being reported by [68].

XGBoost attempts to minimize the objective function, which includes the loss function and regularization term. The loss function quantifies how accurately the model fits the data. Unlike Gradient Boosting, XGBoost introduces the regularization term, which simplifies overly complex trees by adding a penalty to the number and size of leaves in the tree. According to [71], this makes up for the class imbalance issue. Additionally, the algorithm performs tree pruning - a technique to monitor the growth of a tree. XGBoost employs pre-pruning, meaning it terminates a branch growth early when the improvement in the loss function falls below a predefined threshold. This helps to avoid excessive model complexity and over-fitting, improving the training speed. This statement aligns with the study of [39], reporting better generalization and fast training of XGBoost over simpler models. Moreover, XGBoost can deal with missing values by treating them as separate values to determine the ideal split, which was the reason why [36] skipped the cleaning of null values in their experiments.

Nonetheless, XGBoost can be computationally expensive and less interpretable. Also, it still requires a pre-processing stage to handle noise and outliers in the data. Despite this, XGBoost remains a go-to and state-of-the-art machine learning model with exceptional performance, which is tracked in works of [37], [38], reaching more than 95% accuracy.

Consistent with afore-mentioned papers, [44], [53] demonstrated the superiority of XGBoost over other machine learning methods as well.

4.6.6 CatBoost

CatBoost is an ensemble learning approach, which is based on the concept of the Gradient Boosting algorithm. Similar to GB, it sequentially builds decision trees to minimize errors and improve predictions based on gradient descent optimization [92]. What sets CatBoost apart is its specialization in handling both categorical and numerical variables without requiring data pre-processing (i.e. data encoding). The algorithm is particularly adapted to large datasets with a significant number of independent attributes, as noted by [93].

CatBoost employs an automatic regularization method that helps to prevent over-fitting of the data. Additionally, it uses a specific algorithm to handle missing values, making it prone to data inconsistencies. Moreover, CatBoost is highly interpretable, especially if variables are categorical. Another standout feature of this algorithm is the use of symmetric trees, where it performs the same feature splits across all nodes at each depth level. This speeds up the entire training process, which also occurs due to GPU support that CatBoost offers.

However, it is worth noting the large memory consumption of the CatBoost algorithm. Also, it requires a thorough hyperparameter tuning to find optimal solutions and avoid training computational costs. For instance, [38] selected a log-loss activation function and 500 iterations for CatBoost for their experiments.

In general, [38] investigated the effectiveness of CatBoost in estimating customer contentment, where it delivered a notable performance, indicating its potential in predictive activities.

4.7 Artificial Neural Networks

Artificial Neural Networks (ANN) process data by imitating the function and structure of the human brain using artificial neurons, which resemble biological neurons. Similar to the human brain, ANN learns patterns from the data during training, fine-tuning network weights to minimize errors and enhance productivity [58]. This forms an accumulated experience on ANN, which eliminates the need for manual rule adjustments.

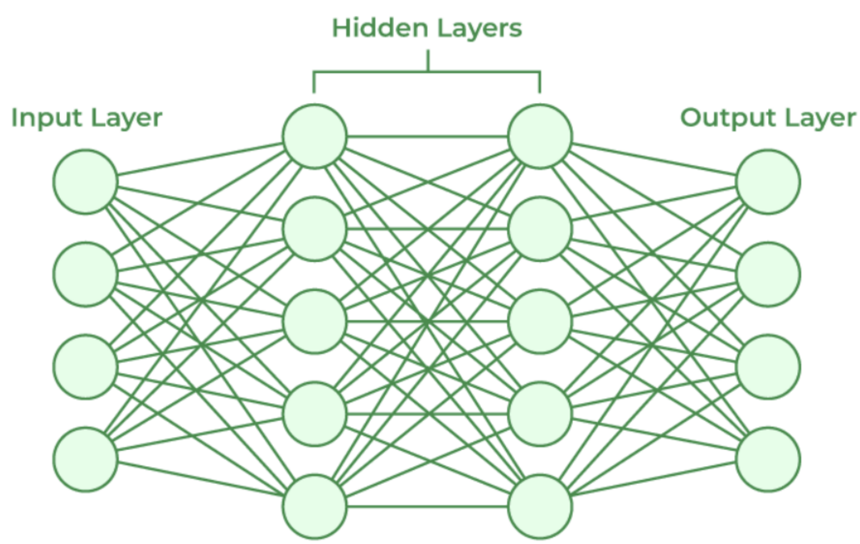


Figure 4-9. The basic architecture of ANN

ANN acts as a large computing system, which consists of many interconnected nodes or units (neurons) that are arranged in a series of layers. Each layer can contain a few dozen nodes or several millions of units, depending on the requirement for the complexity of neural networks to learn relationships in the data.

Commonly, there are 3 types of layers in ANN: input, hidden, and output layers [94] (see Figure 4-9). The input layer gathers the data from the external environment, providing the neural network with the source it necessitates to process and analyze. Next, this data flows through to one or multiple hidden layers, which perform feature extraction and data transformations before passing it to the output layer. The output layer receives the processed data, where it translates learned representations into meaningful outcomes by applying final classifications or predictions.

In a neural network, layers are linked, where each node is interconnected to each other. Each connection assigns weights to units, which determine their degree of influence between themselves. This helps the model refine its understanding, ultimately producing an output at the final layer.

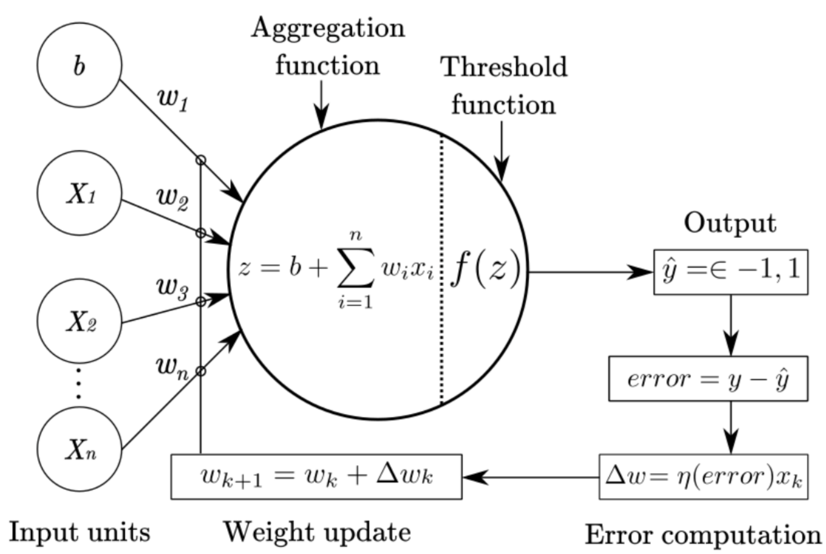


Figure 4-10. The flow of ANN

Figure 4-10 illustrates the function of ANN in more detail, where $X_1 ... X_n$ are inputs, b is a bias term and $\omega_1 ... \omega_n$ are their associated synaptic weights. The inputs get modified by their weights and sum up using the aggregation equation to generate the net input z:

$$z = b + \sum_{i=1}^{n} w_i x_i \tag{4.4}$$

The weighted sum of inputs z represents the total input signal received by the neuron. The threshold or activation function $f(z)$ is applied to identify if the neuron should activate, eventually computing the output and error between predicted and true values. If the output is incorrect or the error is too high, ANN performs a backpropagation. Backpropagation is used to update weight based on obtained results. The process continues until it produces a reasonable output with minimal possible error.

ANN is a powerful and robust technique, which is used across various domains, including feature recognition, classification, regression, and within the scope of customer satisfaction prediction, as seen in [59], [61], and etc. It can capture non-linear complex dependencies and process computations in parallel for large datasets. Also, ANN is resilient to noise, extreme, and missing values in data. More importantly, ANN is capable of learning on its previous mistakes and achieving better convergence as it processes more data.

Nevertheless, ANN can not learn hyperparameters from the data and they need to be adjusted by researchers. ANN requires hyperparameter tuning as these parameters shape the structure and performance of a neural network. Poor tuning can lead to inefficient training, low generalization, over-fitting, under-fitting, and high computational costs, as stated by [64]. Unfortunately, there is no universal one-choice-fits-all solution to hyperparameter selection. The optimal hyperparameters depend on the nature of the data, task, and model architecture. Given their influence on model behavior, certain hyperparameters must be carefully selected. The most influential ones include:

- **Learning rate**: determines how much the model adjusts its weights during training. If a researcher selects too high learning rate, it may lead to quick convergence and over-shoot the optimal solution. On the other hand, too low learning rate slows down the training process and a model may get stuck in a local minimum. In the study by [70], evaluation metrics varied from measures of learning rate used, demonstrating the need for proper adjustment of this parameter.

- **Number of hidden layers**: controls the depth of a neural network. An extremely deep model may capture complex features, but it may over-fit or induce deteriorated performance. A shallow network may miss significant patterns in the data. This behavior was noted in the work of [71], which was resolved by increasing the number of hidden layers up to five, ending up with three dense layers and 2 LSTM layers.

- **Number of neurons per layer**: defines the number of nodes in each hidden layer. A layer with few neurons may not identify enough features, while a layer with high number of neurons increases computational cost and likelihood of over-fitting.
- **Number of epochs**: illustrates the number of times the entire dataset passes through the network. In the presented literature, a number of epochs range widely - for example, [63] used 100, [66] applied 50, and [71] adopted 25.
- **Batch size**: represents the number of samples in one forward or backward pass. It is important to find the optimal batch size. A model with a small batch size generalizes better, but it may be slow. A large batch size increases the chance of over-fitting. [62] used a batch size of 128, achieving a high performance on customer satisfaction data.
- **Activation function**: defines the transformation of signals at each node. Common choices are ReLU, sigmoid, tanh, and softmax functions.
- **Optimizer function**: sets the weight update method (Adam, RMSprop, SGD). Each function in combination with other hyperparameters may lead to different outcomes, depending on the dataset. This behavior is consistent with conclusions drawn by [61], who used Adam optimizer with 500 and 2500 epochs, and RMSprop function with 500 epochs on the e-commerce dataset, demonstrating a wide variance in results.
- **Dropout rate**: determines the percentage of randomly dropping neurons to prevent over-fitting. Typical values are between 20% and 50%.
- **Regularization**: adds a penalty to extreme weights to prevent over-fitting. L1 (Lasso) regularization nullifies some weights and L2 (Ridge) regularization penalizes big weights. [69] employed ridge regularization to avoid over-fitting, while [67] reported the use of lasso method.

Artificial Neural Networks (ANN) is regarded as an advanced technique due to its capability to perform on unseen data and determine non-linear boundaries. Among its various architectures, researchers frequently use different versions of deep neural networks (DNN), which is a subclass of ANN with many hidden layers, to predict customer satisfaction. Commonly, they include *Convolutional Neural Networks (CNN)*, *Recurrent Neural Networks (RNN)*, and *Long-Short Term Memory (LSTM)* architectures, each designed to handle multiple aspects of consumer contentment estimation. CNNs are used to analyze both structured and unstructured data by capturing spatial and recurring patterns in customer feedback. RNNs identify sequential and time-dependent dependencies in consumer trends. LSTM is a special type of RNN, which is effective in analyzing customer behavior over time, overcoming the short-term memory limitation of RNN.

All of the above-mentioned studies used various versions of ANN to predict customer satisfaction, demonstrating their immense power and strong performance in the research.

4.8 Selected Methodology

In this study, multiple machine learning methods have been examined on selected public datasets, which are described in Chapter 4. The selected methods - *Logistic Regression*, *Decision Tree*, *Random Forest*, and *XGBoost* - were chosen based on their widespread use in research studies.

The tables below were prepared to demonstrate a transition over time in terms of what methods were used based on the literature and their corresponding datasets, where available. Taken together, tables span the period between 1990 and 2024.

Table 1. The overview of early methods in literature

Year	**Authors**	**Dataset**	**LR**	**DT**	**KNN**	**SVM**	**LDA**	**NB**	**Bagging**	**RF**	**GB**	**XGBoost**	**AdaBoost**	**CatBoost**	**ANN**	**CNN**	**RNN**	**LSTM**	**Other**
1993	[15]	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2006	[16]	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2019	[69]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
2020	[35]	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 1 represents early machine learning techniques employed between 1990 and 2020. During this period, the literature demonstrates a heavy reliance on Logistic Regression for customer satisfaction prediction. This trend reflects the dominance of traditional statistical models due to their simplicity and computational efficiency.

Table 2. The overview of methods in literature between 2021 and 2022

Year	**Authors**	**Dataset**	**LR**	**DT**	**KNN**	**SVM**	**LDA**	**NB**	**Bagging**	**RF**	**GB**	**XGBoost**	**AdaBoost**	**CatBoost**	**ANN**	**CNN**	**RNN**	**LSTM**	**Other**
2021	[33]	-	✓	✓	✓	✓	-	✓	-	✓	-	-	-	-	✓	-	-	-	-
2021	[56]	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2021	[58]	[87]	-	✓	✓	-	-	-	-	✓	-	-	-	-	✓	-	-	-	-
2021	[59]	[86]	-	✓	✓	✓	✓	-	-	-	-	-	-	-	✓	-	-	-	-
2021	[62]	-	-	✓	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-
2021	[63]	-	-	-	-	✓	-	-	-	-	-	-	-	-	✓	-	-	-	-
2021	[70]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
2022	[34]	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-
2022	[36]	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-
2022	[42]	[88]	-	-	-	-	-	-	-	✓	-	-	-	-	✓	-	-	-	-
2022	[50]	[74]	✓	✓	✓	-	-	✓	-	✓	-	-	-	-	-	-	-	-	-
2022	[68]	-	-	✓	✓	-	-	-	-	✓	-	✓	-	-	-	✓	-	✓	-

Table 2 illustrates how research in 2021 and 2022 gradually began adopting more advanced methods, including Decision Trees, KNNs, SVMs, Naive Bayes, and ANNs. The pattern signals a transitional phase from traditional ML methods towards more sophisticated and diverse approaches. As customer behavior became complex and varied, the progression into models that are capable of capturing non-linear patterns was both timely and justified.

Table 3. The overview of methods in literature between 2023 and 2024

Year	**Authors**	**Dataset**	**LR**	**DT**	**KNN**	**SVM**	**LDA**	**NB**	**Bagging**	**RF**	**GB**	**XGBoost**	**AdaBoost**	**CatBoost**	**ANN**	**CNN**	**RNN**	**LSTM**	**Other**
2023	[38]	-	-	-	-	-	-	-	-	-	-	✓	-	✓	-	-	-	-	-
2023	[39]	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-
2023	[40]	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-
2023	[41]	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-
2023	[43]	-	-	✓	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-
2023	[44]	[74]	-	-	-	-	-	-	✓	-	-	✓	-	-	-	-	-	-	-
2023	[46]	[74]	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Year	Authors	Dataset	LR	DT	KNN	SVM	LDA	NB	Bagging	RF	GB	XGBoost	AdaBoost	CatBoost	ANN	CNN	RNN	LSTM	Other
2023	[48]	[74]	✓	✓	✓	✓		✓	-	✓	-	-	-	-	-	-	-	-	-
2023	[54]	[80]	✓	-	✓	-	-	-	-	✓	-	-	-	-	-	-	-	-	-
2023	[57]	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
2023	[64]	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-
2023	[65]	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-
2023	[66]	[74]	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-
2023	[67]	[76]	-	-	✓	-	-	-	-	✓	-	-	-	-	-	✓	-	-	-
2023	[71]	[80]	✓	-	-	✓	-	✓	-	✓	-	✓	-	-	-	-	-	✓	-
2023	[72]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
2023	[73]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	✓	-	✓
2024	[37]	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	✓
2024	[47]	[74]	✓	✓	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	
2024	[49]	[74]	✓	-	-	✓	-	✓	-	✓	✓	-	-	-	-	-	-	-	✓
2024	[51]	[74]	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-
2024	[52]	-	✓	✓	✓	✓	-	-	-	✓	-	-	✓	-	-	-	-	-	✓
2024	[53]	[75]	-	✓	-	✓	-	-	-	✓	-	✓	-	-	-	-	-	-	-
2024	[55]	[80]	✓	-	-	✓	-	-	-	✓	✓	-	-	-	-	-	-	-	✓
2024	[60]	[81]	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-
2024	[61]	[81]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓

Table 3 portrays a further leap into advanced methodologies in the following two years, showcasing a strong emphasis on ensemble learning and deep learning algorithms with the former being the dominant strategy. Moreover, it is evident that research interest in customer satisfaction prediction saw a significant increase compared to previous years, as indicated by the growing number of studies. The ensemble methods observed include widely used Random Forest, Gradient Boosting, XGBoost, AdaBoost, and CatBoost algorithms, alongside deep learning techniques such as CNN, RNN and LSTM for time-series customer data. Notably, the number of studies incorporating ensemble methods increased considerably during this period, pointing to their growing influence in the field.

Overall, the analysis of existing literature reveals that Logistic Regression, Decision Tree, Random Forest, and XGBoost stand out as the most widely used machine learning choices across multiple studies. Even though these methods are commonly used, it does not imply one is superior to the rest in every case as each offers its own trade-offs that should be taken into account when selecting the right model.

Table 4. Selected methods

Method	**Advantages**	**Disadvantages**
Logistic Regression	Simple, interpretable, efficient for small datasets	Assumes linear relationship, not ideal for complex data
Decision Tree	Easy to interpret, handles both numerical and categorical data	Prone to overfitting, sensitive to small changes in data
Random Forest	Reduces overfitting, handles missing data well	Slower than single decision trees, less interpretable
XGBoost	Highly optimized, faster than standard random forest and gradient boosting, handles missing values well	Requires parameter tuning, can be complex to interpret

4.9 Performance metrics

Performance metrics play a crucial role in evaluating the effectiveness of machine learning methods. They deliver valuable insights into the predictive capabilities of models and provide a reliable measure of their performance across various activities. There are several evaluation tools, which are used to assess the models used in this study: confusion matrix, accuracy, recall, precision, F1-score, and AUC score.

4.9.1 Confusion Matrix

Confusion matrix is a tabular representation of the performance of the method, which shows correct and incorrect predictions. Confusion matrix is not really a performance metric but it is a basis for other evaluation metrics.

There are two categories in a confusion matrix - correct and incorrect labels. Correct predictions span true positives (TP) and true negatives (TN), while incorrect predictions consist of false positives (FP) and false negatives (FN). True labels are the actual values and false labels are misclassified instances.

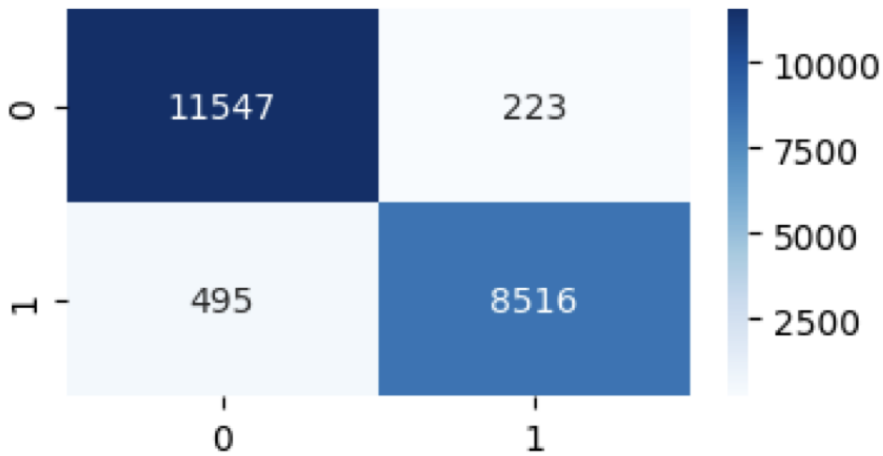


Figure 4-11. The confusion matrix of the Random Forest algorithm on the [74] dataset, as presented in the work of [48]

Each measurement (TP, FP, FN, TN) in the confusion matrix illustrates the number of instances that fall under that category. True positives are actual positive records, which were truly predicted (around 11547 out of 20781 records in Figure 4-11), True Negatives are records that the model correctly labeled as a negative class (about 8516 records), False Positives are negative instances that the model classified as positive (223 instances), and False Negatives are positives that are misclassified by the model, in this example - by Random Forest with 495 values. These measurements in the confusion matrix help to compute accuracy, recall, precision, and F1-score.

4.9.2 Accuracy

Accuracy is the most straightforward metric that evaluates the model performance. Accuracy depicts how often predictions are correct. It is calculated as the ratio of the number of correctly classified instances to the total number of records, expressed as a percentage. The formula of computation is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.5}$$

Despite its simplicity, accuracy can be misleading, especially in imbalanced datasets. The model may reach a high accuracy by mostly predicting the dominant class while failing to identify or ignore the minority class records. Therefore, it is of great importance to consider accuracy in conjunction with other performance metrics, as stated by [50].

4.9.3 Precision

Precision measures how many of the records classified as positive are essentially positive. It indicates the quality of the model's ability to identify positives. Precision is the ratio of actual positives to total predicted positives, including true and false labels:

$$Precision = \frac{TP}{TP + FP} \tag{4.6}$$

Precision is crucial in scenarios where minimizing false positives is a priority. High precision illustrates a high certainty in positive predictions, while low precision depicts a significant number of false positives in the model.

4.9.4 Recall

The recall is a measurement of how well the model identifies actual positive labels. It is essentially the ratio of true positives to the total true outcomes, including true and false positives. Recall is calculated using the following equation:

$$Recall = \frac{TP}{TP + FN} \tag{4.7}$$

A high recall rate signifies that the model did not misclassify any true positives. On the other hand, a low recall demonstrates the high presence of false negatives, which can be an outcome of imbalanced data or incorrectly tuned hyperparameters of the model.

4.9.5 F1-score

F1-score is a combination of precision and recall. It assumes both of them are equally important. A single metric provides a good recall-precision trade-off and extensive insight into the model's performance, particularly in imbalanced datasets. The following equation defines how the F1-score is computed:

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.8}$$

A high F1 score indicates high precision and recall, but a low score does not give much information. A weak recall means the model failed to catch many real positives. A lack of precision suggests that many of the predicted positive cases were misclassified. However, a low F1 score does not tell if precision or recall is low.

4.9.6 Area Under the Curve (AUC)

Area Under the Curve (AUC) represents the accumulated quantity of points the function describes over a certain range, which is used to measure the performance of models. Mathematically, AUC is an integral of a function, presented by the receiver operating curve (ROC), enclosed by fixed limits.

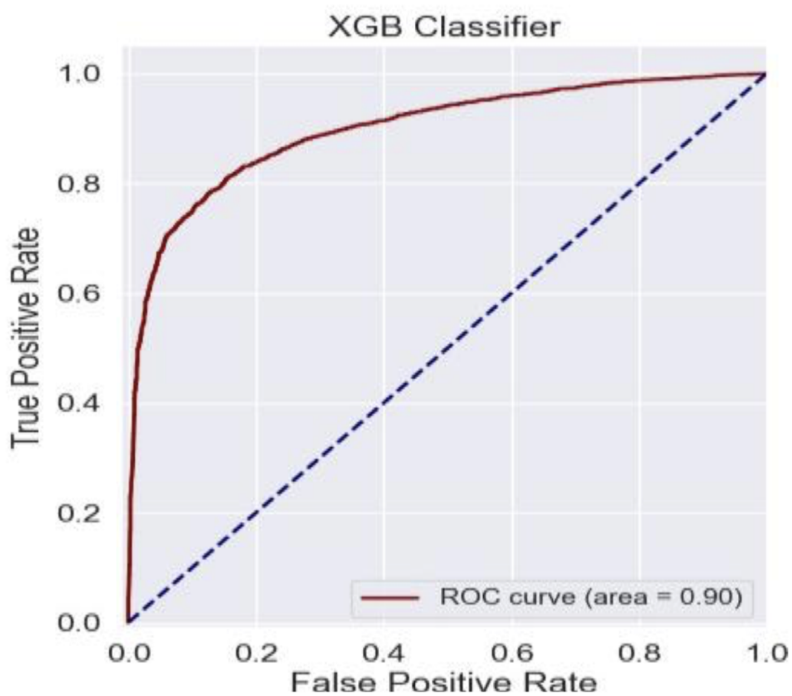


Figure 4-12. AUC score from ROC curve of the XGBoost algorithm on the [80] dataset, as presented in the work of [71]

AUC values range from 0.0 to 1.0. An AUC score of 1.0 indicates that the model makes perfect predictions of all instances. An AUC of 0.5 suggests that the model performs random guessing, correctly classifying only half of the data. A model completely misclassifies every instance or never predicts correctly, when AUC is 0.0. In the study led by [71], it is evident from Figure 4-12 that the AUC score for XGBoost reached 0.90, suggesting a strong performance of the algorithm.

4.10 Class imbalance

Class imbalance deeply affects the performance of models, which might favor a majority class by ignoring minorities. This typically occurs when one class in the dataset significantly outweighs the other, leading to an imbalanced ratio. In customer feedback analysis, an imbalanced dataset may cause a model to predict high satisfaction scores, overlooking customers who are unhappy. This misrepresentation can lead companies to ignore critical pain points in their services, impacting customer retention and engagement.

To address this issue, various techniques have been developed to circumvent the effects of class imbalance. The techniques usually fall into two categories: data-level (those that apply modification on the dataset) and algorithmic-level (those that adjust the sensitivity of models to minorities) approaches. Among the most widely used techniques are re-sampling methods, including under-sampling and over-sampling techniques.

4.10.1 Random Under Sampling (RUS)

Random Under Sampling (RUS) is an under-sampling technique, that involves randomly choosing instances of majority class to remove from the training set with or without replacement. RUS does not give priority to any data instance in the majority class; so, each data point in this category has an equal likelihood of selection. The sampling method is employed to balance binary datasets, generating a 50 by 50 class ratio.

In customer satisfaction data, the majority class is usually dissatisfied consumers; though, there are some exceptions, as presented in these datasets.

The primary advantage of this technique is its ability to reduce computational expenses by working with fewer samples than the original dataset. Also, it prevents over-fitting to the majority class by learning balanced decision boundaries between the two.

However, there is a potential risk of losing valuable information as removing random instances may discard critical patterns in the majority class. Additionally, RUS is ineffective

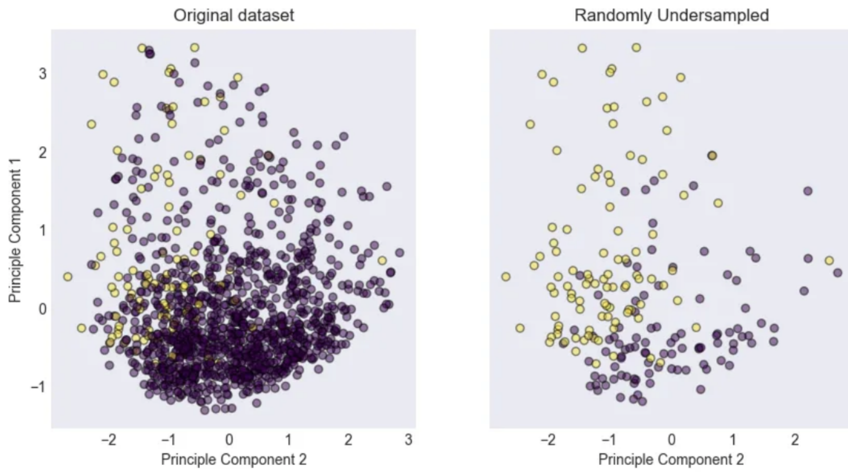


Figure 4-13. The illustration of RUS on data

for highly imbalanced datasets (i.e. 99:1 ratio), where it may cut down enormous chunks of data, leaving too few training samples.

4.10.2 Random Over Sampling (ROS)

Random Over Sampling (ROS) is an over-sampling technique, which includes randomly duplicating instances of minority class with replacements to add to the training set. Similar to RUS, it does not prioritize any instance in a minority class and assigns an equal chance of duplication for all data points. After over-sampling, there are more overlapping instances in the data. As a result, the minority class data points appear more concentrated and visually darker (see Figure 4-14).

The main advantage of ROS is the preservation of patterns and retention of information in the data, especially smaller ones. Increasing the number of instances in minority class improves the model's ability to learn decision boundaries; though, it may fail to do so when the sample size is size.

Nonetheless, ROS boosts computational costs by populating duplicated samples. If the gap between the two classes is large, it may significantly increase the sample size, which may slow down training and require more storage. Moreover, it enhances the risk of over-fitting on highly imbalanced data and subsequent poor generalization since the model may memorize specific patterns in the minority class.

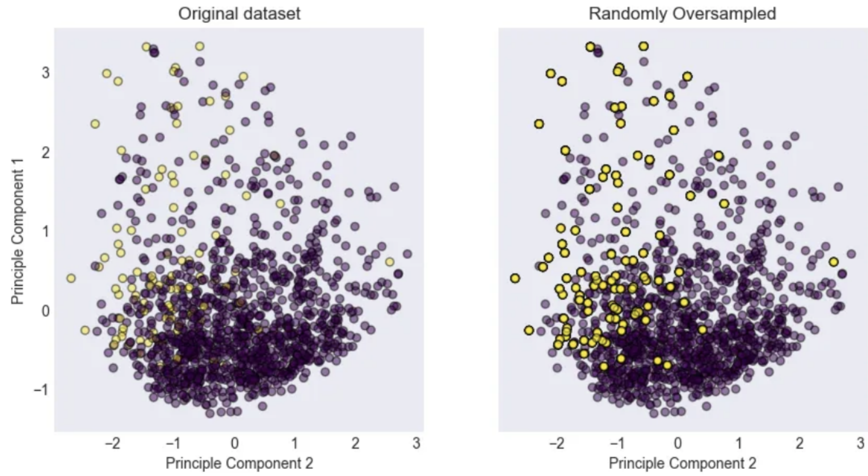


Figure 4-14. The illustration of ROS on data

4.10.3 Synthetic Minority Over-sampling Technique (SMOTE)

Synthetic Minority Over-sampling Technique (SMOTE) is another over-sampling technique, which produces synthetic data to add to the training set. The artificial data is created by leveraging patterns found in the existing data. SMOTE randomly selects instances in the minority class and identifies its k nearest neighbors. It adds new data by interpolating links between chosen and neighboring points. Pairs of neighboring data points of two classes are called Tomek links. By using TOMEK links, SMOTE refines synthetic data generation and removes overlapping majority class instances.

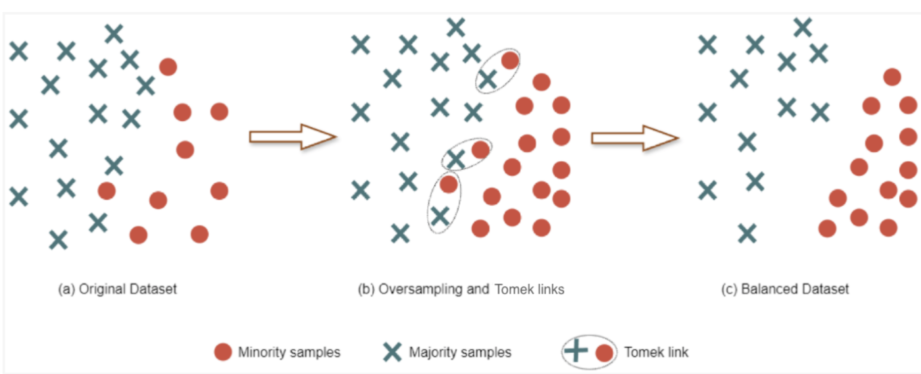


Figure 4-15. SMOTE

This technique reduces the likelihood of over-fitting and preserves majority class information. It is effective for highly imbalanced data (i.e. 99:1 ratio) where RUS would remove too many samples and ROS would lead to data memorization.

Nevertheless, it is slower than RUS and ROS since it heavily relies on the KNN algorithm. Also, SMOTE may introduce a synthetic noise if the minority class contains many outliers.

4.11 Hyperparameter tuning

Hyperparameter tuning is a process of finding optimal parameters for a model to gain the best performance. Hyperparameters are predefined settings that dictate the model's learning process. These parameters are set before the model begins training. Poorly chosen settings may induce suboptimal model performance, leading to over-fitting, under-fitting, or prolonged training times. Therefore, it is crucial to carefully select and fine-tune these values to achieve effective results. This can be done manually through step-by-step testing, but it does not eliminate a potential risk of human error and can be time-consuming. Instead, automated hyperparameter tuning techniques, such as GridSearchCV and RandomizedSearchCV, can be leveraged to systematically explore parameter combinations.

4.11.1 GridSearchCV

GridSearchCV is a hyperparameter tuning technique, which is often regarded as a brute-force optimization approach. It exhaustively explores all possible sets of specified hyperparameter settings for a given model. A grid of possible values is constructed and the model is trained for each combination using cross-validation. It logs the performance metrics of all trained models and selects the hyperparameter set, which has the highest accuracy or the lowest loss among all others.

This approach guarantees that the best-performing combination is found. The result remains reproducible and does not change when running it multiple times since it tests all sets of hyperparameters.

However, GridSearchCV may be extremely slow, especially for complex models, if there are many hyperparameters, which may bring significant computational expenses. Also, GridSearchCV can exhaust the entire memory resources even with a small set of values on large datasets.

Therefore, it is best to use this algorithm on small sample space with non-complex models and few settings.

4.11.2 RandomizedSearchCV

RandomizedSearchCV is another method for performing hyperparameter optimization. It works by randomly selecting a subset of hyperparameters as opposed to the pre-defined systematic combinations of settings used by GridSearchCV. The sampling can be performed with or without replacement. This approach leverages iterations to find an optimal solution, which is set by the user. At each iteration, RandomizedSearchCV attempts to train the model on various sets of hyperparameters and log performance measurements. After a certain number of iterations, it selects a combination that provides the most effective outcome.

The figure below depicts how two contrasting methods detect top-performing hyperparameters, following their distinct search strategies.

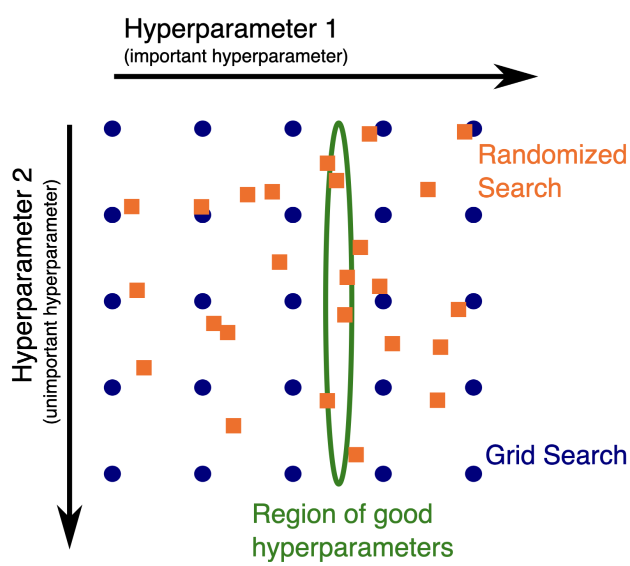


Figure 4-16. GridSearchCV and RandomizedSearchCV

This technique stands for its speed and relatively low computational cost, making RandomizedSearchCV an appealing choice for demanding activities and complex models (e.g. Random Forest, XGBoost). Moreover, it delivers highly competitive results and often finds a close-to-optimal solution, especially in high-dimensional hyperparameter spaces and large sample sizes.

However, RandomizedSearchCV runs a risk of missing an absolute best set of hyperparameters since not all given parameters are tried out. Additionally, results may vary between different runs, making it less deterministic.

Overall, this approach is practical for large datasets and search spaces with more complex models.

Chapter 5

Results

This chapter describes results obtained from the application of four of the most effective and widespread machine learning methods, defined by existing research, on the four datasets, which are described in Chapter 3.

Each method was implemented on four different variants of the same dataset, generated using distribution adjustment techniques: random over-sampling, random under-sampling, SMOTE, and the original dataset.

A total of 64 unique configurations were tested, spanning over 150 experimental runs on the baseline Google Colab environment. Execution time for each run differed depending on the configuration and available computational unit (CPU or GPU) - some runs finished within an hour or less, while others extended up to two days. While durations varied, the average duration was estimated at around 3-5 hours per run. The total computational time to conduct these experiments was approximately 500 hours spread across 70 days.

Hyperparameter tuning was initially processed by the grid search cross-validation method; however, due to its high computational costs and prolonged execution time, it was eventually replaced by the randomized search cross-validation technique. The number of iterations for the latter was set to balance the resource efficiency and predictive performance of models; though, it increased the number of runs to come close to optimal solutions.

The performance of models was evaluated based on key metrics of machine learning, namely accuracy, recall, precision, F1-score, and AUC score, which are later presented in tabular format. Additionally, for the sake of completeness, confusion matrices for all 64 model configurations are included in the Appendix.

Overall, the findings suggest that the results align with those reported in the prior literature.

5.1 US Airlines dataset

The table below represents the results of the experiments with the US Airlines dataset for the following methods: Logistic Regression, Decision Tree, Random Forest, and XGBoost algorithms.

The Logistic Regression model, which serves as a benchmarking baseline, reached a maximum accuracy of 86.67%, indicating that the model correctly predicts satisfied passengers most of the time while failing to do so for the remaining $\sim$13%. Recall is relatively high, which is due to the fact that logistic regression has no tendency for over-fitting, meaning it maintains a decent score; yet false negatives exist. Precision is slightly lower than recall, suggesting the model may miss some values from the majority class, which is the class of satisfied customers. AUC represents the model's ability to differentiate between satisfied and dissatisfied passengers, with a score of 92.38% confirming its strong classification ability. However, Logistic Regression demonstrated the lowest performance among others as it struggles with the non-linearity in the data and feature interaction handling.

Table 1. US Airlines dataset: Performance metrics

Method	**Accuracy**	**Recall**	**Precision**	**F1-score**	**AUC**
Logistic Regression	0.8667	0.8497	0.8445	0.8471	0.9238
Logistic Regression (RUS)	0.8665	0.8554	0.8748	0.8650	0.9268
Logistic Regression (ROS)	0.8657	0.8541	0.8744	0.8641	0.9257
Logistic Regression (SMOTE)	0.8660	0.8530	0.8758	0.8643	0.9258
Decision Tree	0.9525	0.9278	0.9616	0.9444	0.9788
Decision Tree (RUS)	0.9485	0.9349	0.9611	0.9478	0.9791
Decision Tree (ROS)	0.9567	0.9609	0.9529	0.9569	0.9567
Decision Tree (SMOTE)	0.9536	0.9369	0.9693	0.9528	0.9792
Random Forest	0.9614	0.9374	0.9728	0.9548	0.9936
Random Forest (RUS)	0.9609	0.9478	0.9732	0.9603	0.9946
Random Forest (ROS)	0.9684	0.9598	0.9766	0.9681	0.9961
Random Forest (SMOTE)	0.9650	0.9527	0.9768	0.9646	0.9955
XGBoost	0.9629	0.9405	0.9732	0.9566	0.9950
XGBoost (RUS)	0.9624	0.9591	0.9721	0.9620	0.9956

Method	Accuracy	Recall	Precision	F1-score	AUC
XGBoost (ROS)	0.9681	0.9626	0.9734	0.9680	0.9963
XGBoost (SMOTE)	0.9665	0.9545	0.9780	0.9661	0.9963

Decision Tree performed a big jump in terms of all performance metrics, reaching 95.25% accuracy and 97.88% AUC scores. It moderately surpassed the Decision Tree algorithm by [46] (94.41% and 94.37%). However, Random Forest and XGBoost algorithms outperformed it since a decision tree is prone to over-fitting issues. Random Forest improved over Decision Tree, reducing over-fitting and boosting precision up to nearly 12%. XGBoost provided the best generalization over the data with slight improvements over the previous base models. Base variants of Random Forest and XGBoost reached accuracies of 96.09% and 96.29%, providing a slightly better performance than in [49], [47], [48], and others.

The US Airlines dataset is not perfectly balanced; but, there is only modest skewness in the data towards the discontent category of passengers, which affects the performance of models. The ratio between two classes is 43%/57% with around 14% of the gap between themselves in favor of dissatisfied passengers. Therefore, resampling techniques, presented in Chapter 4, were employed to refine the robustness of models.

Random Under-Sampling (RUS) was used to reduce the size of the majority class by randomly removing some instances from the data to make it balanced. When running models on data processed by RUS, the overall recall, AUC, and F1-score saw slight improvements throughout all methods. However, precision remained consistent for Logistic Regression and Decision tree but declined for other models. This is an impact of loss of valuable information and increased misclassifications.

Similarly, the use of Random Over-Sampler (ROS) significantly improved recall and F1-score but downgraded precision. The precision score for the Decision Tree went from 96.16% to 95.29% as ROS increases the risk of over-fitting by adding duplicates.

SMOTE generates synthetic data instead of duplicating, which resulted in more balanced enhancements in recall and precision. SMOTE scaled up accuracy values from 95.25% to 95.36% for Decision Tree (the best for DT), 96.14% to 96.84% for Random Forest, and 96.29% to 96.65% for XGBoost.

Overall, Random Forest and XGBoost with ROS showcased the highest accuracy scores (96.84% and 96.81%) across all models, followed by the same methods enhanced with SMOTE (96.84% and 96.65%). These results suggest that the models are barely superior or broadly comparable to [50] with 95.8% for Random Forest and [44] with 96% for XGBoost. Furthermore, feature importance analysis based on the ROS-augmented Random Forest algorithm revealed key factors affecting consumer satisfaction, with 'Online boarding', 'Inflight wifi service', and 'Type of Travel' being the most significant contributors (see Figure 5-1).

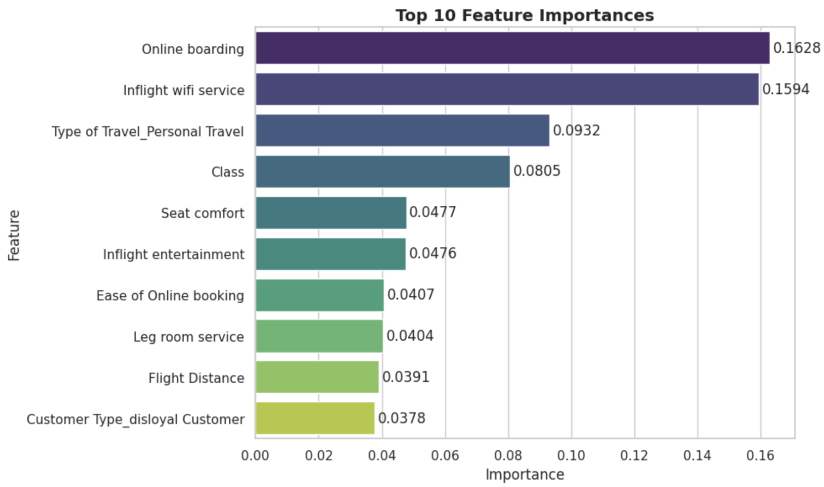


Figure 5-1. US Airlines dataset: Top 10 factors affecting customer satisfaction

5.2 Sky Insights dataset

In the table below, one can observe the results obtained for the Sky Insights dataset. Logistic Regression was employed as a benchmarking model for passenger satisfaction prediction and sampling was applied to boost its performance. The baseline model showed a strong performance - 94.41% accuracy and 98.34% AUC, suggesting the dataset mostly contains linear relationships as it could identify an optimal decision boundary between classes. Recall for content passengers is 95.25%, while precision lags behind at 89.45%, meaning it falsely predicts more customers as satisfied. Furthermore, sampling applied to this model resulted in a well-balanced recall-precision relationship with only around 2% difference at peak. The combination of sampling techniques and regression model consistently improved all evaluation metrics, with the SMOTE-based model achieving the highest accuracy and AUC at 95.16% and 98.59%. Such improvements could be explained by the largely imbalanced nature of this dataset, where 34% of customers are satisfied with airline services.

Decision Tree performed slightly better than the top-performing regression model at 95.85% accuracy and 98.34% AUC. It followed the same trend as previous models, with SMOTE yielding the strongest results. Random Forest saw moderate improvements over Decision Tree. With RUS the model's accuracy and AUC declined possibly due to information loss, while other metric values increased. ROS contributed to the highest performance (97.43%) among other Random Forest models, and it demonstrates comparable trends to prior research [48] (97.00%), [50] (97.02%) and others.

Table 2. Sky Insights dataset: Performance metrics

Method	Accuracy	Recall	Precision	F1-score	AUC
Logistic Regression	0.9441	0.9525	0.8945	0.9226	0.9834
Logistic Regression (RUS)	0.9486	0.9576	0.9407	0.9491	0.9839
Logistic Regression (ROS)	0.9488	0.9590	0.9399	0.9493	0.9834
Logistic Regression (SMOTE)	0.9516	0.9617	0.9426	0.9521	0.9859
Decision Tree	0.9585	0.9467	0.9353	0.9410	0.9857
Decision Tree (RUS)	0.9592	0.9589	0.9595	0.9592	0.9856
Decision Tree (ROS)	0.9597	0.9631	0.9565	0.9598	0.9898
Decision Tree (SMOTE)	0.9686	0.9686	0.9686	0.9686	0.9923
Random Forest	0.9623	0.9480	0.9443	0.9462	0.9912
Random Forest (RUS)	0.9595	0.9621	0.9572	0.9596	0.9883
Random Forest (ROS)	0.9743	0.9866	0.9630	0.9746	0.9949
Random Forest (SMOTE)	0.9683	0.9717	0.9651	0.9684	0.9943
XGBoost	0.9623	0.9403	0.9513	0.9457	0.9911
XGBoost (RUS)	0.9618	0.9640	0.9597	0.9619	0.9898
XGBoost (ROS)	0.9755	0.9859	0.9659	0.9758	0.9948
XGBoost (SMOTE)	0.9693	0.9690	0.9696	0.9693	0.9946

XGBoost outperformed all other models due to its enhanced capability to catch intricate minor non-linear patterns. ROS-based XGBoost yielded the best results with 97.55% accuracy and a nearly perfect AUC score of 99.48%.

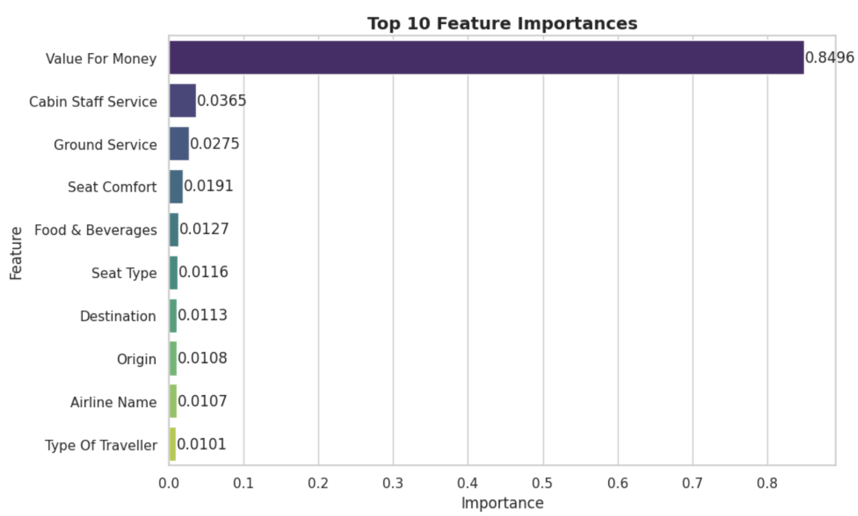


Figure 5-2. Sky Insights dataset: Top 10 factors affecting customer satisfaction

Overall, results suggest that SMOTE improved performance across all models, maintaining high precision values. SMOTE was particularly effective for simpler models such as Logistic Regression and Decision Tree. For ensemble models, ROS illustrated the best fit, handling imbalances well. Among all evaluated models, XGBoost proved to be the most effective model for this dataset. Based on this model, feature importance analysis identified factors such as 'Value for money', 'Cabin staff service', and 'Ground service' as the major determinants in shaping customer satisfaction outcomes, with the former being the predominantly impactful feature (see Figure 5-2).

5.3 Airline Reviews dataset

Logistic Regression, serving as a baseline model, delivered a robust result with 93.95% accuracy and an impressive 98.14% AUC, suggesting the data is well-separated. This is likely due to the strong correlation between higher ratings and satisfaction, implying linear trend in the data. Additionally, resampling methods (RUS, ROS, and SMOTE) demonstrated no drastic need for usage for the regression model as they provided nearly identical performance to the baseline approach. This can be explained by the dataset's roughly balanced class distribution with a 53%/47% satisfaction ratio in favor of content passengers.

Table 3. Airline Reviews dataset: Performance metrics

Method	**Accuracy**	**Recall**	**Precision**	**F1-score**	**AUC**
Logistic Regression	0.9395	0.9463	0.9397	0.9430	0.9814
Logistic Regression (RUS)	0.9423	0.9501	0.9354	0.9427	0.9854
Logistic Regression (ROS)	0.9364	0.9417	0.9319	0.9367	0.9821
Logistic Regression (SMOTE)	0.9341	0.9405	0.9286	0.9345	0.9819
Decision Tree	0.9253	0.8821	0.9742	0.9259	0.9726
Decision Tree (RUS)	0.9357	0.9436	0.9289	0.9362	0.9800
Decision Tree (ROS)	0.9335	0.9323	0.9345	0.9334	0.9802
Decision Tree (SMOTE)	0.9259	0.9242	0.9274	0.9258	0.9696
Random Forest	0.9407	0.9440	0.9440	0.9440	0.9839
Random Forest (RUS)	0.9409	0.9383	0.9433	0.9408	0.9869
Random Forest (ROS)	0.9487	0.9405	0.9561	0.9482	0.9841
Random Forest (SMOTE)	0.9434	0.9498	0.9378	0.9438	0.9834
XGBoost	0.9444	0.9440	0.9506	0.9473	0.9842
XGBoost (RUS)	0.9409	0.9383	0.9433	0.9408	0.9868
XGBoost (ROS)	0.9492	0.9405	0.9572	0.9488	0.9843
XGBoost (SMOTE)	0.9417	0.9405	0.9427	0.9416	0.9842

Even with a nearly balanced dataset, the Decision Tree is recommended to be more unstable compared to other models and prone to over-fitting. The main model reached 92.53% accuracy, 97.26% AUC, and very high precision of 97.24%, but recall dropped to 88.21%, suggesting the presence of misclassified discontent passengers. Re-sampling methods helped to increase recall significantly but at the cost of little precision drop (a sign of over-fitting).

Random Forest and XGBoost illustrated a strong performance, even without resampling with 94.07% and 94.44% accuracy. Both methods delivered around 98% AUC scores across all of their variants. RUS and SMOTE did not improve much, and ROS offered a minor increase in accuracy and precision at the expense of recall fall. Out of all methods, ROS-based XGBoost provided the highest performance (94.92%), which is roughly in agreement with existing studies [44], [53]. Based on this method, feature importance analysis identified 'Value for money', 'Airline', and 'Class' as principal drivers of customer satisfaction, highlighting the former attribute accounting for nearly 90% of the overall importance (see Figure 5-3).

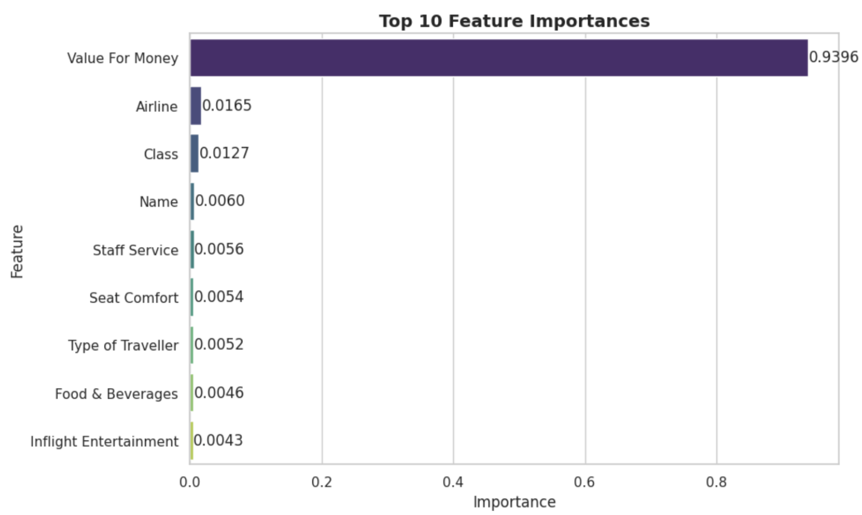


Figure 5-3. Airline Reviews dataset: Top 10 factors affecting customer satisfaction

Overall, ensemble methods exhibited the highest effectiveness, but the regression model remains a strong alternative to them with quite competitive results. Also, there is little necessity for the use of class imbalance handling techniques on this dataset across most models (except for the Decision Tree), given the balanced dataset.

5.4 OlistBr dataset

Table 4. OlistBr dataset: Performance metrics

Method	Accuracy	Recall	Precision	F1-score	AUC
Logistic Regression	0.6136	0.6753	0.7372	0.7049	0.5942
Logistic Regression (RUS)	0.5617	0.6589	0.5514	0.6004	0.5936

Method	Accuracy	Recall	Precision	F1-score	AUC
Logistic Regression (ROS)	0.5738	0.6833	0.5604	0.6158	0.6036
Logistic Regression (SMOTE)	0.5896	0.6574	0.5787	0.6156	0.6197
Decision Tree	0.7133	0.9476	0.7208	0.8188	0.6071
Decision Tree (RUS)	0.6155	0.6139	0.6156	0.6148	0.6155
Decision Tree (ROS)	0.8259	0.7402	0.8932	0.8095	0.8667
Decision Tree (SMOTE)	0.7226	0.6934	0.7362	0.7142	0.7660
Random Forest	0.7845	0.9709	0.7724	0.8603	0.7708
Random Forest (RUS)	0.6830	0.7517	0.6607	0.7033	0.7502
Random Forest (ROS)	0.9299	0.9046	0.9527	0.9280	0.9769
Random Forest (SMOTE)	0.8385	0.9178	0.7921	0.8504	0.9029
XGBoost	0.7880	0.9646	0.7783	0.8615	0.7602
XGBoost (RUS)	0.6796	0.7176	0.6667	0.6912	0.7501
XGBoost (ROS)	0.9137	0.8704	0.9529	0.9098	0.9744
XGBoost (SMOTE)	0.8515	0.9368	0.8002	0.8631	0.9109

Logistic Regression consistently underperforms compared to all other models. The highest accuracy and AUC for Logistic Regression were 61.36% and 61.97%, meaning the model struggles to capture complex relationships and separate classes. This suggests that data largely contains non-linear patterns, which regression models are not capable of identifying.

Decision Tree demonstrated better results but still had generalization issues with 60.71% AUC. When applying RUS, the model performance declined due to information loss but increased recall, which is expected since it balanced the dataset. SMOTE enhanced the efficiency of the model; though, it fell short of ROS with 82.59% accuracy and 86.67% AUC.

The base variant of Random Forest already outperformed Logistic Regression and Decision Tree with 78.45% accuracy and 77.08% AUC, which outperformed the analogous algorithm used by [54]. However, there was a big gap between recall and precision, suggesting it failed to balance their trade-off. ROS-based Random Forest reached the highest results across all models (92.99%), followed by the SMOTE version (83.85%). [55] reached 92% and 91% accuracy scores for the respective re-sampled Random Forest variants, which are closely comparable to outcomes of this study, aside from SMOTE with lower performance. This study did not employ XGBoost in their experiments.

The combination of XGBoost and ROS showcased a strong predictive performance with 91.37% accuracy and 97.44% AUC, which demonstrates a notable performance gain over [71] (83.01% and 90%, respectively).

Using the top-performing method (ROS-based RF), it was determined that customer satisfaction was strongly influenced by 'Payment value', 'Price', and 'Product length (cm)', emphasizing the significance of both pricing and product-related features in shaping satisfaction outcomes (see Figure

5-4). It is notable that the rest of the less dominant features remain at the similar level of importance at around 5%.

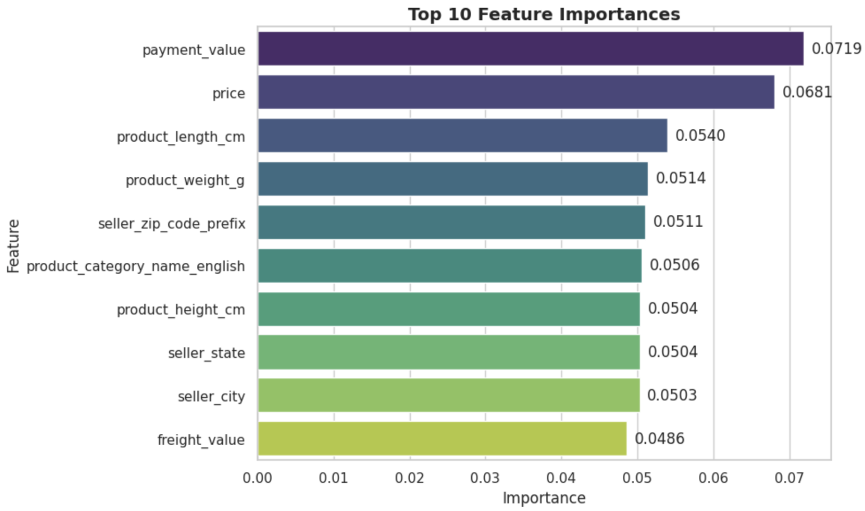


Figure 5-4. OlistBr dataset: Top 10 factors affecting customer satisfaction

Chapter 6

Conclusion and Directions for Future Improvement

Within the range of this work, several machine learning methods were explored to predict customer satisfaction across four various datasets. Following a careful examination of recent findings, this study focused on current methods and datasets used to evaluate their performance.

In this chapter, a summary of this work drawn from the experiments is provided and opportunities for future improvements based on the results are discussed.

6.1 Conclusion

Customer satisfaction prediction is a widely popular and common task in machine learning, which is driven by the necessity of businesses to cover consumer retention and understand customer reactions.

This thesis work was structured into six chapters, each covering a distinct aspect of the study.

The introduction chapter delivered the motivation for the thesis work, problem statement, objectives, and scope of this study.

Chapter two provided an overview of existing literature in the domain of consumer satisfaction estimation. The studies were categorized into three groups based on their relevance, effectiveness, and widespread use. The review highlighted that a significant proportion of state-of-the-art methods relied on machine learning techniques.

Chapter three outlined many datasets across multiple domains, which are relevant to the study. The data acquisition, characteristics of the datasets, and initial pre-processing needs were presented as well as their limitations found during analysis. Considering them, four distinct datasets were chosen for the experiments: US Airlines, Sky Insights, Airline Reviews, and OlistBr.

Chapter four delivered a detailed description of the architecture of the methodology used in the literature. Their workings, advantages, and common limitations were covered. Some methods were accompanied by the discussion of special conditions and requirements to explain factors affecting their effectiveness. From the range of methods explored, four were selected based on their frequent implementation in the literature: Logistic Regression, Decision Tree, Random Forest, and XGBoost. Likewise, there were subsections for class imbalance handling and hyperparameter tuning techniques to optimize model performance and address potential biases.

Chapter five portrayed results obtained from the chosen methods applied to the opted datasets. Initially, all base models with no re-sampling techniques demonstrated a moderate performance across the four datasets, as anticipated. This was a consequence of the highly imbalanced nature of the data. In order to deal with class imbalance, three common re-sampling techniques were introduced: Random Under Sampling (RUS), Random Over Sampling (ROS), and SMOTE. After applying either of them, a better balance between precision and recall was observed at the expense of accuracy, which slightly dropped compared to base models. From the results, it is noticeable that ROS and SMOTE-based Random Forest and XGBoost algorithms reached higher precision-recall, maintaining good accuracy scores.

6.2 Directions for Future Improvement

There are several possible enhancements that can be introduced to this project within a scope of future work:

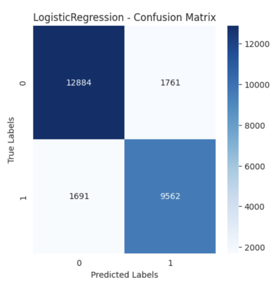
- Deep learning methods are a natural progression and extension to the prediction of customer satisfaction. As shown in Chapter 2, these methods illustrate top performance on consumer contentment datasets, which surpass machine learning techniques on some occasions.
- Fusion or hybrid models may be used to find out whether they lead to higher convergence or not.
- Feature selection techniques can contribute to performance boost as some variables may negatively affect the performance, while others help establish more refined rules and patterns understanding.
- Regression-based approaches represent a valuable addition to the existing work since customer satisfaction level is not only bound to a binary nature and some datasets include continuous measures (e.g. CSAT scores, ranging from 0.0 to 100.0).
- Some additional datasets from other domains may be tested out.
- Factors, contributing to customer satisfaction, could be thoroughly explored via exploratory

data analysis to investigate whether they contribute to a better customer behavior understanding.

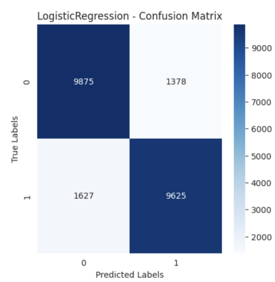
Appendix A

Confusion Matrices

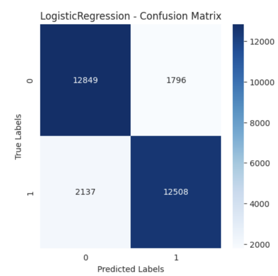
A.1 US Airlines Dataset



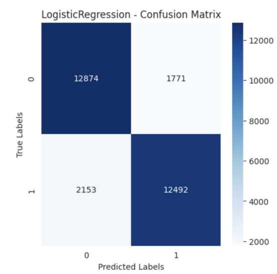
(a) LR



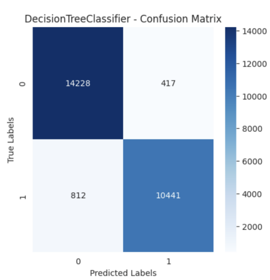
(b) LR - RUS



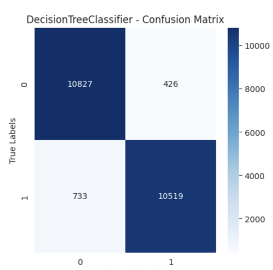
(c) LR - ROS



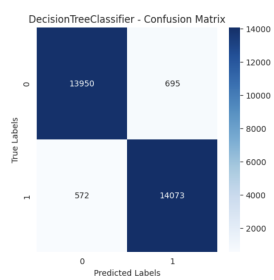
(d) LR - SMOTE



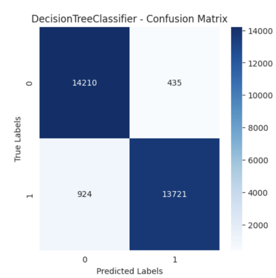
(e) DT



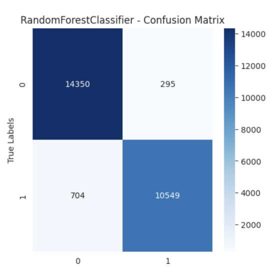
(f) DT - RUS



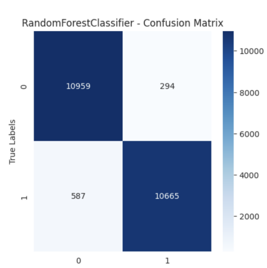
(g) DT - ROS



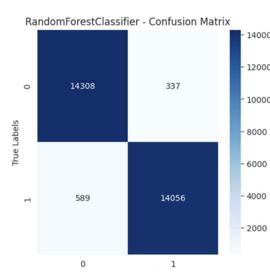
(h) DT - SMOTE



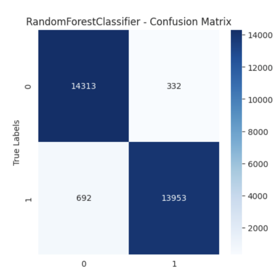
(e) RF



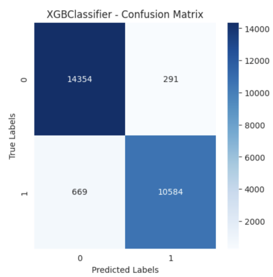
(f) RF - RUS



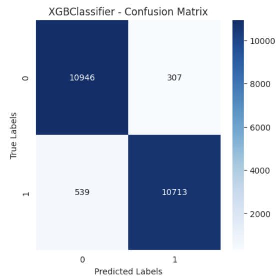
(g) RF - ROS



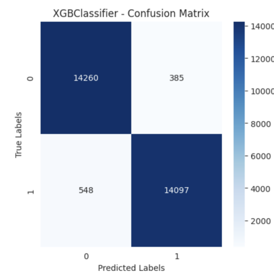
(h) RF - SMOTE



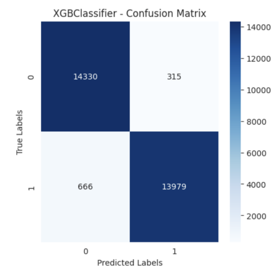
(e) XGB



(f) XGB - RUS



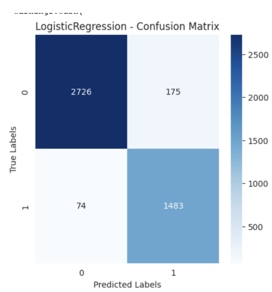
(g) XGB - ROS



(h) XGB - SMOTE

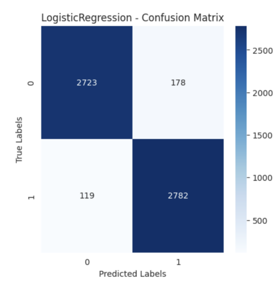
Figure A-1. US Airlines dataset: Confusion matrices

A.2 Sky Insights Dataset

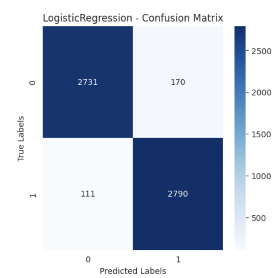


(a) LR

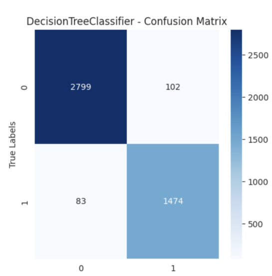
(b) LR - RUS



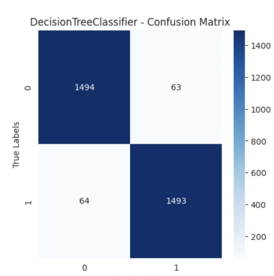
(c) LR - ROS



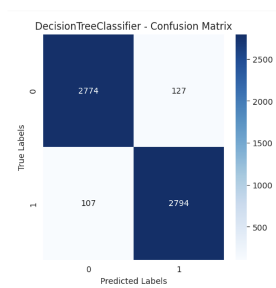
(d) LR - SMOTE



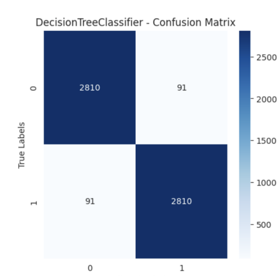
(e) DT



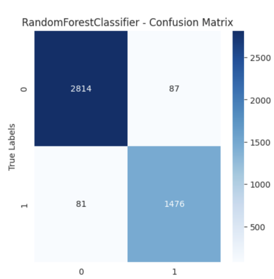
(f) DT - RUS



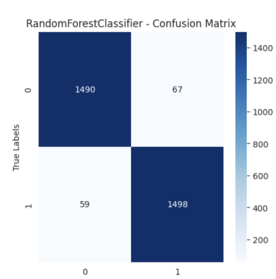
(g) DT - ROS



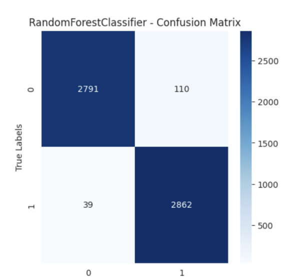
(h) DT - SMOTE



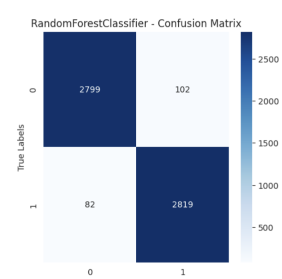
(e) RF



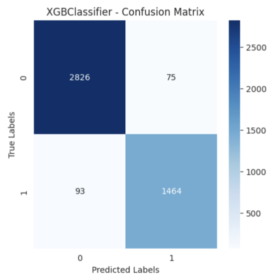
(f) RF - RUS



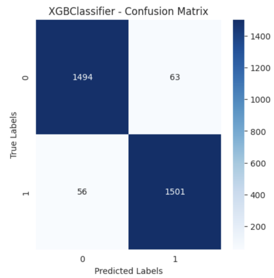
(g) RF - ROS



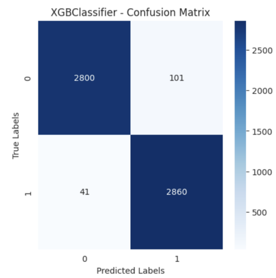
(h) RF - SMOTE



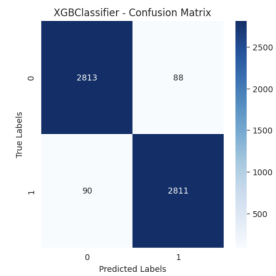
(e) XGB



(f) XGB - RUS



(g) XGB - ROS



(h) XGB - SMOTE

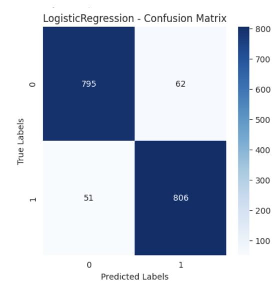
Figure A-2. Sky Insights dataset: Confusion matrices

A.3 Airline Reviews Dataset

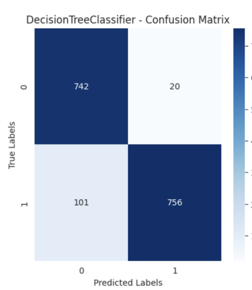
(a) LR

(b) LR - RUS

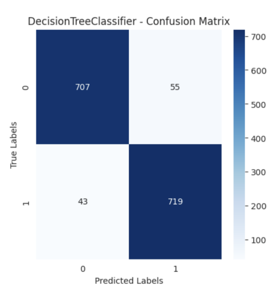
(c) LR - ROS



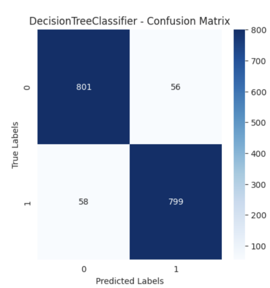
(d) LR - SMOTE



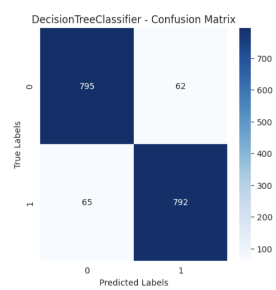
(e) DT



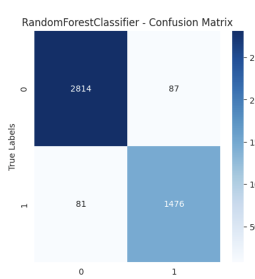
(f) DT - RUS



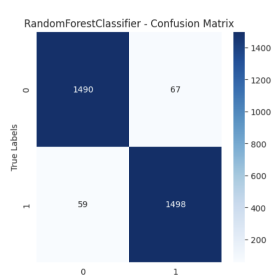
(g) DT - ROS



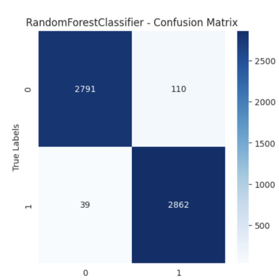
(h) DT - SMOTE



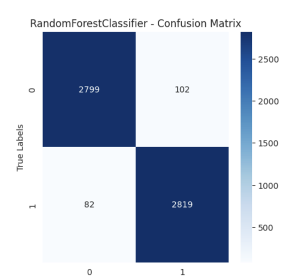
(e) RF



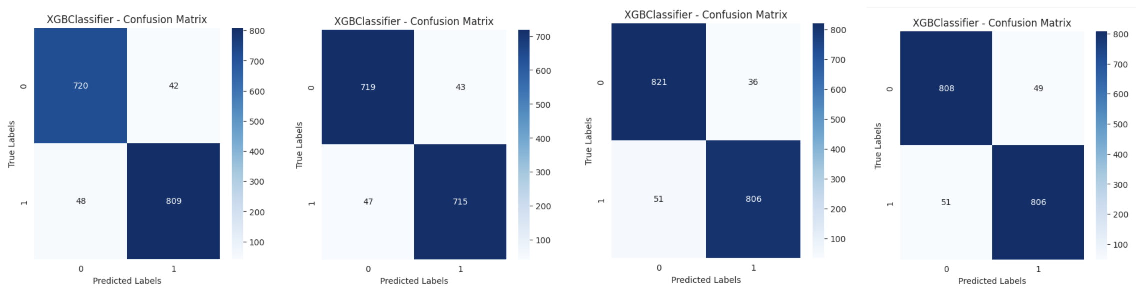
(f) RF - RUS



(g) RF - ROS



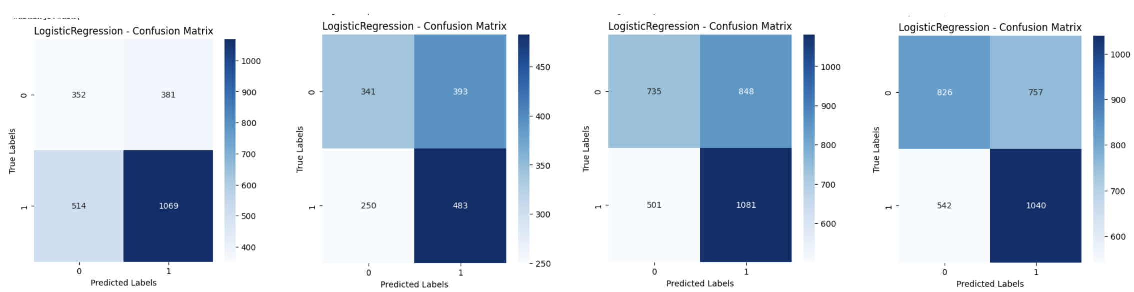
(h) RF - SMOTE



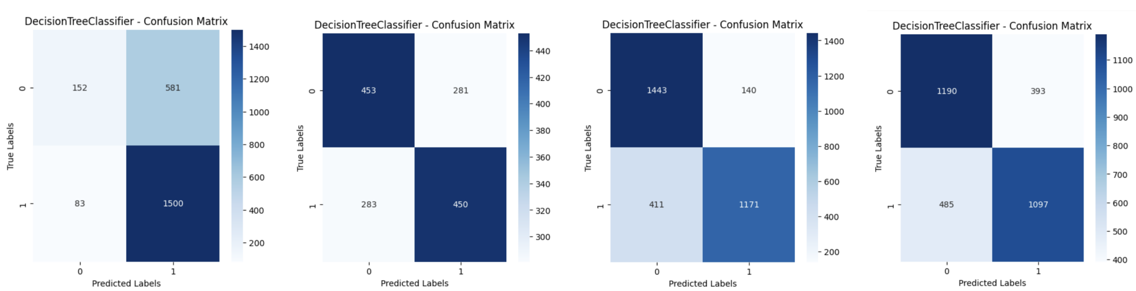
(e) XGB (f) XGB - RUS (g) XGB - ROS (h) XGB - SMOTE

Figure A-3. Airline Reviews dataset: Confusion matrices

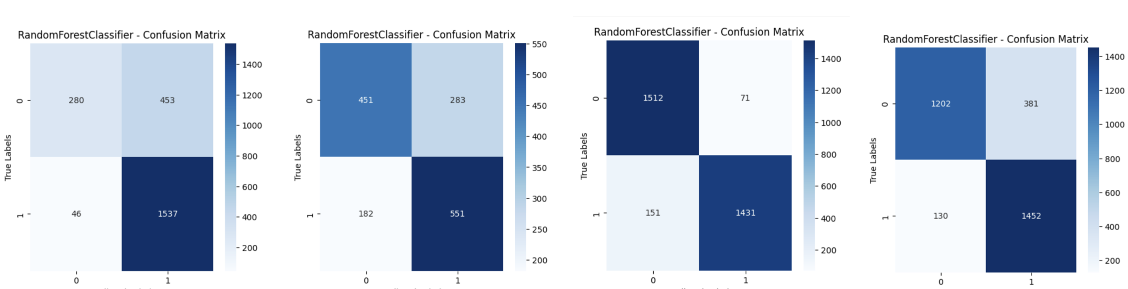
A.4 OlistBr Dataset



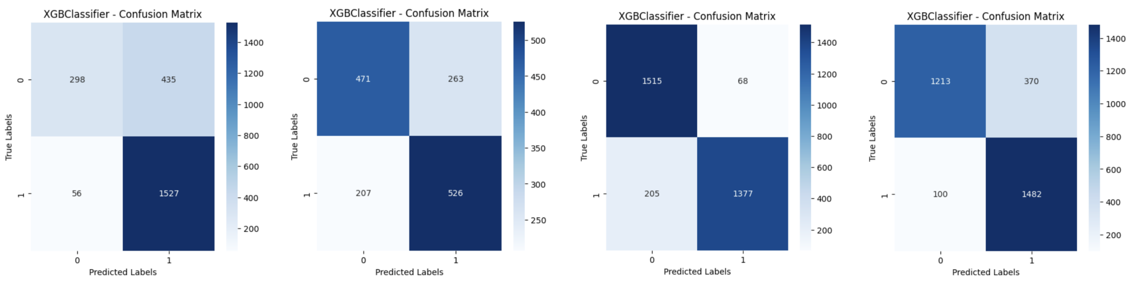
(a) LR (b) LR - RUS (c) LR - ROS (d) LR - SMOTE



(e) DT (f) DT - RUS (g) DT - ROS (h) DT - SMOTE



(e) RF (f) RF - RUS (g) RF - ROS (h) RF - SMOTE



(e) XGB (e) XGB - RUS (f) XGB - ROS (h) XGB - SMOTE

Figure A-4. OlistBr dataset: Confusion matrices

Bibliography

[1] Global Market Insights. Graphics Processing Unit (GPU) Market Size - By Component (Hardware, Software, Service), By Deployment Model (On-premises, Cloud), By End-user (Gaming, Design & Manufacturing, Automotive, Real estate, Healthcare), By Application & Forecast, 2024 – 2032. https://www.gminsights.com/industry-analysis/gpu-market, 2024. Accessed on September 6, 2024.

[2] Keith D. Foote. A Brief History of Database Management. https://www.dataversity.net/brief-history-database-management/, 2021. Accessed on October 25, 2021.

[3] Global CIO. Current aspects of digitalization in Kazakhstan. https://globalcio.com/articles/main/current-aspects-of-digitalization-in-kazakhstan/, 2024. Accessed on November 19, 2024.

[4] Destination CRM. U.S. Companies Fail to Meet Customer Expectations, KPMG Finds. https://www.destinationcrm.com/Articles/CRM-News/CRM-Featured-Articles/U.S.-Companies-Fail-to-Meet-Customer-Expectations-KPMG-Finds-162170.aspx?utm$_{s}ource = chatgpt.com, 2024. Accessed on November 28, 2024.$

[5] Melissa Scatena. Customer Experience Statistics: What the Numbers Reveal About CX in 2025. https://onramp.us/blog/customer-experience-statistics?utm_source=chatgpt.com, 2024. Accessed on December 19, 2024.

[6] Mary Jo Bitner. "Evaluating Service Encounters: The Effects of Physical Surroundings and Employee Responses." Journal of Marketing, 54(2):69-82, 1990.

[7] Richard L. Oliver and Wayne S. DeSabo. "Response Determinants in Satisfaction Judgments." Journal of Consumer Research, 14(4):495-507, 1988.

[8] A. Parsu Parasuraman, Valarie A. Zeithaml, and Leonard L. Berry. "SERVQUAL: A multiple Item Scale for measuring consumer perceptions of service quality." Journal of Retailing, 64(1):12-40, 1988.

[9] Jonathan D. Barsky and Richarcd Labagh. "Theory S: Total customer service." The Cornell Hotel and Restaurant Administration Quarterly, 31(1):88-95, 1992.

[10] Jonathan D. Barsky and Susan Dittmaim. "A Strategy for Customer Satisfaction." The Cornell Hotel and Restaurant Administration Quarterly, 33(5):32-40, 1990.

[11] Kai Kristensen, Anne Martensen, and Lars Gronholdt. "Customer satisfaction measurement at Post Denmark: Results of application of the European Customer Satisfaction Index Methodology." Total Quality Management, 11(7):1007-1015, 2000.

[12] Syed Saad Andaleeb and Carolyn Conway. "Customer satisfaction in the restaurant industry: an examination of the transaction-specific model." Journal of Services Marketing, 20(1):3-11, 2006.

[13] Zoe S. Dimitriades. "Customer satisfaction, loyalty and commitment in service organizations: Some evidence from Greece." Management Research News, 29(12):782-800, 2006.

[14] Rebecca Bevens. An Introduction to t Tests | Definitions, Formula and Examples. https://www.scribbr.com/statistics/t-test/, 2020. Accessed on January 31, 2020.

[15] Roland T. Rust and Anthony J. Zahorik. "Customer Satisfaction, Customer Retention, and Market Share." Journal of Retailing, 69(2):193-215, 1993.

[16] Cathy Lawson and Douglas C. Montgomery. "Logistic Regression Analysis of Customer Satisfaction Data." Quality and Reliability Engineering International, 22:971-984, 2006.

[17] Walfried M. Lassar, Chris Manolis, and Robert D. Winsor. "Service quality perspectives and satisfaction in private banking." Journal of Services Marketing, 14(3):244-271, 2000.

[18] Roger Hallowell. "The relationships of customer satisfaction, customer loyalty, and profitability: an empirical study." International Journal of Service Industry Management, 7(4):27-42, 1996

[19] Roland T. Rust, Anthony J. Zahorik, and Timothy L. Keiningham. *Service Marketing.* HarperCollins College Publishers, 1996.

[20] Terry G. Vavra. *Improving Your Measurement of Customer Satisfaction: A Guide to Creating, Conducting, Analyzing and Reporting Customer Satisfaction Measurement Program.* ASQ Quality Press, 1997.

[21] Dan Grove, Felician Campean, Janet Zeppenfeld, Neil Dixon, and Simon Robinson. "Automotive Customer Satisfaction Data Analysis Using Logistic Regression." In 2008 SAE World Congress & Exhibition, SAE Technical Papers, 2008.

[22] Subhash Sharma, Ronald W. Niedrich, and Greg Dobbins. "A Framework for Monitoring Customer Satisfaction: An Empirical Illustration." Industrial Marketing Management, 28(3):231-243, 1999.

[23] Lauren M. Trabold, Gregory R. Heim, and Joy M. Field. "Comparing e-service performance across industry sectors: Drivers of overall satisfaction in online retailing." International Journal of Retail & Distribution Management, 34(4):240-257, 2006.

[24] Joseph F. Hair Jr, Christian M. Ringle, Nicholas P. Danks, G. Tomas M. Hult, Marko Sarstedt, and Soumya Ray. *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R.* Springer, 2021.

[25] Michael J. Ryan, Robert Rayner, and Andy Morrison. "Diagnosing customer loyalty drivers: partial least squares vs regression." Marketing Research, 11(2):1926, 1999.

[26] Michael D. Johnson and Anders Gustafsson. *Improving Customer Satisfaction, Loyalty, and Profit: An Integrated Measurement and Management System.* University of Michigan Business School Management Series, 2000.

[27] Nha Nguyen and Gaston Leblanc. "The mediating role of corporate image on customers' retention decisions: an investigation in financial services." International Journal of Bank Marketing, 16:52–65, 1998.

[28] Yan Liu, Changfeng Zhou, and Yingwu Chen. "Measuring Customer Satisfaction based on Neural Networks Partial Least Squares Approach." In 2006 IEEE International Conference on Service Operations and Logistics, and Informatics, pp. 627-631. IEEE, 2006.

[29] Eric Wilson and Tim Wragg. "We cannot diagnose the patient's illness but experience tells us what treatment works." International Journal of Market Research, 43(2):189-215, 2001.

[30] Bayu Adhi Tama. "Data mining for predicting customer satisfaction in fast-food restaurant." Journal of Theoretical and Applied Information Technology, 75(1):18-24, 2015.

[31] Zhi-Hua Zhou and Shaowu Liu. *Machine Learning.* Tsinghua University Press, 2016.

[32] Islam El Naqa and Martin J. Murphy. *Machine Learning in Radiation Oncology.* Springer, 2015.

[33] Syeda Farjana Shetu, Israt Jahan, Mohammad Monirul Islam, Refath Ara Hossain, Nazmun Nessa Moon, and Fernaz Narin Nur. "Predicting Satisfaction of Online Banking System in Bangladesh by Machine Learning." In 2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST), pp. 223-228. IEEE, 2021.

[34] Mahardhian Anjar Ligiarta and Yova Ruldeviyani. "Customer Satisfaction Analysis of Mobile Banking Application Based on Twitter Data." In 2022 2nd International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS), pp. 322-327. IEEE, 2022.

[35] Junseok Eom, Taewook Choi, Yungsuk Yoon, and Kyung Geun Lee. "The Study of Customer Satisfaction Index Characteristic with Local Water Supply Service." In 2020 International Conference on Information and Communication Technology Convergence (ICTC), pp. 1886-1889. IEEE, 2020.

[36] Haosong Gou, Ling Su, Gaoyi Zhang, Wei Huang, Yunbo Rao, and Yang Yang. "A XGBoost Method Based on Telecom Customer Satisfaction Enhancement Strategy." In 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), pp. 209-213. IEEE, 2022.

[37] Jiao Zhu, Yeneng Chen, Fan Shi, Yimeng Song, Rui Li, and Xufeng Liu. "Customer Satisfaction Prediction Model Based on Big Data and XGBoost." In 2023 Signal and Information Processing, Networking and Computers (ICSINC), pp. 250-256. Springer, 2024.

[38] Haosong Gou, Ling Su, Gaoyi Zhang, Haicong Li, and Zhiheng Li. "Research on Telecom Customer Satisfaction Prediction Strategy." In 2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), pp. 593-598. IEEE, 2023.

[39] Chengjian Wang, Xiaowei Wang, Zhiyuan Hu, and Guiyin Lin. "Satisfaction prediction based on XGBoost classification and decision tree model." In 2023 International Conference on Electronics and Devices, Computational Science (ICEDCS), pp. 441-446. IEEE, 2023.

[40] Xiaotong Zhao, Nianwei Li, Wencheng Wang, and Xiaoxia Yang. "Research on mobile service satisfaction prediction model based on GBDT regression algorithm." In 2023 5th International Conference on Communications, Information System and Computer Engineering (CISCE), pp. 26-29. IEEE, 2023.

[41] Zeng Fei, He Yuqing, Yang Chengqin, Hu Xinkai, and Yuan Yining. "Mobile Customer Satisfaction Scoring Research Based on Quadratic Dimension Reduction and Machine Learning Integration." Applied Sciences, 13(17):9681, 2023.

[42] Khalid M. B. A. Joolfoo, Rameshwar A. Jugurnauth, and Muhammad B. A. Joolfoo. "Application of Machine Learning in Predicting Customer Satisfaction of Telecom Service Providers." In 2022 4th International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM), pp. 1-10. IEEE, 2022.

[43] Chuanshang Li, Haoge Pang, and Quanli Zhou. "Mobile User Satisfaction Analysis Based on Decision Tree and Random Forest Algorithm." In 2023 International Conference on Internet of Things, Robotics and Distributed Computing (ICIRDC), pp. 304-308. IEEE, 2023.

[44] Sinan Yang. "Machine Learning-Based Airline Passenger Satisfaction Evaluation." In 2023 International Conference on Industrial IoT, Big Data and Supply Chain (IIoTBDSC), pp. 11-15. IEEE, 2023.

[45] Claudia Perlich, Foster Provost, and Jeffrey S. Simonoff. "Tree induction vs. logistic regression: A learning-curve analysis." Journal of Machine Learning Research, 4(2):211–255, 2004.

[46] Junru Yao. "Feature Analysis and Evaluation of Airline Passenger Satisfaction Based on Decision Tree." In 2023 International Conference on Industrial IoT, Big Data and Supply Chain (IIoTBDSC), pp. 23-27. IEEE, 2023.

[47] Ashwin Kumar, Vatsal Mehta, Trupthi Rao, and Ashwini Kodipalli. "Forecasting Airline Customer Satisfaction through the Application of Diverse Machine Learning Algorithms." In 2024 Interna-

tional Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS), pp. 1–5. IEEE, 2024.

[48] Farjana Islam, Md. Farukuzzaman Faruk, S. M. Mahedy Hasan, Farzana Akter, Azmain Yakin Srizon, and Md. Rakib Hossain. "An Investigation of Hyperparameters Optimization and Feature Reduction Techniques: Predicting Airline Passenger Satisfaction Using Machine Learning Models." In 2023 26th International Conference on Computer and Information Technology (ICCIT), pp. 1-6. IEEE, 2023.

[49] Maham Salah-Ud-Din, Blessy Trencia Lincy S.S., and Hannah Al Ali. "Exploratory Data Analysis and Prediction of Passenger Satisfaction with Airline services." In 2024 New Trends in Civil Aviation (NTCA), pp. 295-302. IEEE, 2024.

[50] Lamija Hibović, Selma Smajić, and Emine Yaman. "Predicting Satisfaction of Airline Passengers Using Classification." In 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 939-946. IEEE, 2022.

[51] Adedayo Sobowale, Olukemi Osadare, Afeez A. Soladoye, and Peace O. Sobowale. "Prediction of Customer Satisfaction in Airline Hospitality services for improved service delivery using Support Vector Machine." FUOYE Journal of Engineering and Technology, 9(2):241-245, 2024.

[52] Krish Jindal, Disha Sharma, Pranjal Sharma, Akrit Gupta, and Mansha. "Airline Customer Satisfaction Prediction." International Journal For Multidisciplinary Research, 6(2):1-10, 2024.

[53] Ashraf Osman Ibrahim, Chiew Cheng Yi, Abubakar Elsafi, and Fuad A. Ghaleb. "Revolutionizing Airline Customer Satisfaction Analysis with Machine Learning Techniques." In 2023 Advances in Intelligent Computing Techniques and Applications (IRICT), pp. 141-152. Springer, 2024.

[54] Pongthanin Wangkiat and Chantri Polprasert. "Machine Learning Approach to Predict E-commerce Customer Satisfaction Score." In 2023 8th International Conference on Business and Industrial Research (ICBIR), pp. 1176-1181. IEEE, 2023.

[55] Maha Zaghloul, Sherif Barakat, and Amira Rezk. "Predicting E-commerce customer satisfaction: Traditional machine learning vs. deep learning approaches." Journal of Retailing and Consumer Services, 79:103865, 2024.

[56] Tukino and Algifanri Maulana. "C4.5 Algorithm Application For Prediction 0f Customer Satisfaction Accuracy In PT. Pico Jaya Telesindo." In 2021 International Conference on Computer Science and Engineering (IC2SE), pp. 1-6. IEEE, 2021.

[57] Arunava Mookherjee and Krishna Roy. "Predicting Customer Satisfaction of Luxury Watches Using ID3 Algorithm." In 2023 7th International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech), pp. 1-6. IEEE, 2023.

[58] Rawan Alkurd, Ibrahim AbuAlhaol, and Halim Yanikomeroglu. "User Satisfaction Prediction Framework for Personalized Wireless Networks: A DNN Approach." In 2021 IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1-6. IEEE, 2021.

[59] Said Gadri, Sara Ould Mehieddine, Khadidja Herizi, and Safia Chabira. "An Efficient System to Predict Customers' Satisfaction on Touristic Services Using ML and DL Approaches." In 2021 22nd International Arab Conference on Information Technology (ACIT), pp. 1-8. IEEE, 2021.

[60] Chunrong Guo and Xiaodong Zhang. "Research on Multimodal Prediction of E-Commerce Customer Satisfaction Driven by Big Data." Applied Sciences, 14(18):1-18, 2024.

[61] Chunrong Guo and Xiaodong Zhang. "Intelligent Prediction of Cross-Border E-Commerce Customer Satisfaction Using Deep Learning Embeddings." IEEE Access, 12:173268–173278, 2024.

[62] Lara Saleh and Saleh Abu-Soud. "Predicting Jordanian Job Satisfaction Using Artificial Neural Network and Decision Tree." In 2021 11th International Conference on Advanced Computer Information Technologies (ACIT), pp. 735-738. IEEE, 2021.

[63] Senthan Prasanth, Kuhaneswaran Banujan, Btgs Kumara, and Hiruni Rupasingha. "Prediction of Student Satisfaction on Online Learning during the COVID-19 Pandemic - A Case Study on Sri Lankan Universities." In 2021 21st International Conference on Advances in ICT for Emerging Regions (ICter), pp. 69-74. IEEE, 2021.

[64] Xueqing Sun. "User satisfaction prediction based on feed-forward neural network algorithm." In 2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE), pp. 379-384. IEEE, 2023.

[65] Zongxuan Zhang, Shuxin Wang, Guohao Zhou, and Di Sun. "Analysis and Prediction Model of Mobile User Experience Based on PCA Principal Component Analysis and BP Neural Network." In 2023 International Conference on Applied Physics and Computing (ICAPC). pp. 480-485. IEEE, 2023.

[66] Wangbin Du. "Neural Network in Aircraft Customer Satisfaction Prediction." Advances in Economics, Management and Political Sciences, 38:19-29, 2023.

[67] Bhargav Bhargav and R. Thandaiah Prabu. "Airline Passenger Satisfaction Prediction Using Novel Hybrid Random Forest Model Comparison with K-Nearest Neighbour Model." In 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), pp. 1-6. IEEE, 2023.

[68] So-Hyun Park, Mi-Yeon Kim, Yeon-Ji Kim, and Young-Ho Park. "A Deep Learning Approach to Analyze Airline Customer Propensities: The Case of South Korea." Applied Sciences, 12(4):1916, 2022.

[69] Zhao Huang, Qingquan Li, Jizhe Xia, and Fan Li. "A Passenger Satisfaction Prediction Method Based on Driving Time." In 2019 Computing, Communications and IoT Applications (ComComAp), pp. 310-315. IEEE, 2019.

[70] Haiyan Zhang, Ziye Wang, Mingxing Ke, Ming Cai, and Qiyao Sun. "Fine-grained Tourist Satisfaction Prediction Based on Deep Learning." In 2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC), pp. 30-35. IEEE, 2021.

[71] Trilok Nath Pandey, Anirudh Vasudev, Dennis Sagayanathan, Govind Anjan, Danish Arshad, and Sudhansu Shekhar Patra. "Predicting Customer Satisfaction in Brazil E-commerce: A Comparative Study of Machine Learning Techniques." In 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), pp. 505-510. IEEE, 2023.

[72] JunJie Yin, JingXia Jiang, LiHan Tong, and PeiYun Huang. "FCNN+: An Improved Fully Connected Neural Network High-accuracy Prediction Model." In 2023 8th International Conference on Information Systems Engineering (ICISE), pp. 539-542. IEEE, 2023.

[73] Sulaiman Aftan and Shah Habib. "Using the AraBERT Model for Customer Satisfaction Classification of Telecom Sectors in Saudi Arabia." Brain Sciences, 13(1):147, 2023.

[74] US Airlines dataset. https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction?resource=download, 2020.

[75] Invistico dataset. https://www.kaggle.com/datasets/sjleshrac/airlines-customer-satisfaction, 2020.

[76] Skytrax dataset. https://www.kaggle.com/datasets/efehandanisman/skytrax-airline-reviews, 2019.

[77] Sky Insights dataset. https://www.kaggle.com/datasets/juhibhojani/airline-reviews, 2023.

[78] British Airways dataset. https://www.kaggle.com/datasets/chaudharyanshul/airline-reviews, 2024.

[79] Airline Reviews dataset. https://www.kaggle.com/datasets/sujalsuthar/airlines-reviews, 2024.

[80] OlistBr dataset. https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce, 2022.

[81] Dress Sales dataset. https://www.kaggle.com/code/ggsri123/exploratory-data-analysis-on-dress-sales, 2022.

[82] Shopzilla dataset. https://www.kaggle.com/datasets/ddosad/ecommerce-customer-service-satisfaction, 2024.

[83] Amazon dataset. https://www.kaggle.com/datasets/surajjha101/myntra-reviews-on-women-dresses-comprehensive, 2024.

[84] CFS dataset. https://www.kaggle.com/datasets/jahnavipaliwal/customer-feedback-and-satisfaction, 2024.

[85] CBSH dataset. https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset?select=shopping_behavior_updated.csv, 2024.

[86] Hotel Reviews dataset. https://www.kaggle.com/datasets/joebeachcapital/hotel-reviews?select=reviews.csvs, 2023.

[87] Wireless Networks dataset. https://github.com/rawanalkurd/Personalization-Framework-Datasets, 2019.

[88] Voice Call dataset. https://www.kaggle.com/datasets/yashovardhan99/voice-call-quality-experience-for-november-2019, 2020.

[89] Vihar Kurama. A Guide to AdaBoost: Boosting To Save The Day. https://www.digitalocean.com/community/tutorials/adaboost-optimizer, 2024. Accessed on October 21, 2024.

[90] Samy Baladram. AdaBoost Classifier, Explained: A Visual Guide with Code Examples. https://medium.com/data-science/adaboost-classifier-explained-a-visual-guide-with-code-examples-fc0f25326d7b, 2024. Accessed on November 10, 2024.

[91] Tomonori Masui. All You Need to Know about Gradient Boosting Algorithm Part 2 Classification. https://medium.com/data-science/all-you-need-to-know-about-gradient-boosting-algorithm-part-2-classification-d3ed8f56541e, 2022. Accessed on February 8, 2022.

[92] Oluseye Jeremiah. CatBoost in Machine Learning: A Detailed Guide. https://www.datacamp.com/tutorial/catboost, 2024. Accessed on September 6, 2024.

[93] John T. Hancock and Taghi M. Khoshgoftaar. "CatBoost for big data: an interdisciplinary review." Journal of Big Data, 7:94, 2020.

[94] IBM. Ibm - What is a neural network?. https://www.ibm.com/think/topics/neural-networks, 2024. Accessed on February 25, 2024.

[95] IBM. Ibm - What is the k-nearest neighbors (KNN) algorithm?. https://www.ibm.com/think/topics/knn, 2024. Accessed on April 25, 2024.

[96] IBM. Ibm - What are support vector machines (SVMs)?. https://www.ibm.com/think/topics/support-vector-machine, 2024. Accessed on April 24, 2024.