VIETNAM NATIONAL UNIVERSITY, HANOI
INTERNATIONAL SCHOOL

GRADUATION PROJECT

PROJECT NAME

AIRLINE PASSENGER SATISFACTION PREDICTION USING MACHINE LEARNING

Student's name

Nguyen Minh Hieu

Hanoi - Year 2023

VIETNAM NATIONAL UNIVERSITY, HANOI

INTERNATIONAL SCHOOL

GRADUATION PROJECT

PROJECT NAME

AIRLINE PASSENGER SATISFACTION PREDICTION USING MACHINE LEARNING

SUPERVISOR: Dr. Tran Thi Oanh

STUDENT: Nguyen Minh Hieu

STUDENT ID: 18071471

COHORT: MIS2018A

MAJOR: Management Information System

Tran Thi Oanh

Hanoi - Year 2023

ACKNOWLEDGEMENT

I worked really hard on this project. It would not have been possible, however, without the kind support and cooperation of many individuals and organizations. I would want to offer my deepest appreciation to each and every one of them.

I am deeply indebted to Dr. Tran Thi Oanh, who was my instructor throughout my graduation thesis. She guided and supported me a lot in the process of completing the project report.

I would also like to express my sincerity and appreciation to the member of International School - Vietnam National University and Office of Student Affairs for particular for information and their kind of cooperation and encouragement that help me in the completion of this project.

Due to the time limitation and lack of expertise, the report cannot avoid flaws, and I am hoping for feedback from professors so that I could learn from it and strengthen my abilities.

Hanoi, 27th January 2023

Student

Hieu

Nguyen Minh Hieu

DECLARATION

I hereby declare that the Graduation Project “Airline Passenger Satisfaction Prediction Using Machine Learning” is the result of my personal study and has never been published in anyone else's work. Throughout the implementation process of this project, I took research ethics seriously; all findings of this project are the results of my own study and surveys; and all references in this project are properly referenced in accordance with rules.

I accept full accountability for the precision of the numbers and statistics, as well as the other elements of my graduation project.

Hanoi, 27th January 2023

Student

Hieu

Nguyen Minh Hieu

TABLE OF NOTATIONS AND ABBREVIATIONS

Abbreviation	Meaning
AdaBoost	Adaptive Boosting
ANN	Artificial Neural Network
AUC	Area Under The Curve
CATBoost	Category Boosting
CNN	Convolutional Neural Network
GPU	Graphics Processing Unit
ID3	Iterative Dichotomiser 3
LightGBM	Light Gradient-Boosting Machine
ML	Machine Learning
ROC	Receiver Operating Characteristics
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

ABSTRACT

The aviation industry is a highly competitive sector that has expanded quickly during the last two decades and customer satisfaction is a major problem for the airline industry. In this study, I worked on a dataset containing airline passenger satisfaction surveys collected from an anonymous airline. The implementation process starts from using data preprocessing techniques to clean the data and transform the data accordingly for analysis. The analysis was carried out using 7 different classification strategies: Decision Tree, Random Forest, SVM, XGBoost, LightGBM, AdaBoost and CatBoost. The classifiers were trained using 80% of the data and tested using the remaining 20% data. The outcome of the test set is the satisfaction of the passenger (satisfied/neutral or dissatisfied). Based on the acquired results, the accuracy is calculated to make a comparison between each classification technique in order to discover the best approach.

OVERVIEW

1. The necessity of the thesis

Today, in an increasingly competitive market, customer satisfaction is extremely important for business, determining the success or failure of a business in the market school. Improving service quality and customer satisfaction helps companies to maintain existing customers, attract new customers, improve customer loyalty, maintain and enhance competitiveness. So in today's business activities, customer satisfaction becomes the center of the business strategy of enterprises.

Studying customer satisfaction helps businesses understand factors affecting customer satisfaction and satisfaction level and from there, it is possible to assess the competitiveness and business efficiency of the company implementing policies to correct and improve customer satisfaction customers for their products and services.

In the aviation market in the world today, the level of competition is increasing. With developments and changes in the aviation industry, the role and power of customers are getting bigger and bigger. If the customers are not provided with the service they expect, they will easily change to another airline.

In such a competitive aviation market, understanding factors affecting satisfaction and satisfaction level of customers, from which there are policies to overcome and improve customer satisfaction has a very important role for the airlines in general.

2. Research purposes

Identify factors that are highly correlated to a satisfied (or unsatisfied) passenger with an airline's flight experience and service.

3. Structure of the thesis

In addition to the abstract, overview, and list of references, the graduation thesis consists of 4 chapters:

Chapter 1: Introduction to machine learning and passengers' satisfaction prediction

An overview of machine learning, its concepts, types, strengths and weaknesses, its application in practice and also identify the problem to be solved.

Chapter 2: Methodology to predict passengers' satisfaction

List the steps to solve the problem and demonstrate the methods used.

Chapter 3: Model building and evaluation

Presenting insights after performing EDA and training, testing the prediction models as well as model evaluation on a public dataset.

Chapter 4: Conclusion

Summarize the research results, achievements gained from the study and its limitations

CHAPTER 1. INTRODUCTION TO MACHINE LEARNING AND PASSENGERS' SATISFACTION PREDICTION

In this chapter, I will provide a brief overview of machine learning, its concepts, classifications, benefits and drawbacks, practical applications, and the problem to be addressed along with related work and the contribution of the study.

1.1. Some machine learning concepts

Machine learning is the process of programming computers to maximize a performance criterion based on example data or previous experience. We've defined a model up to some parameters, and learning is the execution of a computer program to optimize the model's parameters using training data or prior experience. The model may be predictive in order to make future forecasts, or descriptive in order to gather information from data.

Machine learning is a topic of research dealing with the subject of how to build computer algorithms that automatically improve with experience [1].

1.2. Classification

Machine learning techniques are grouped into four broad groups, which are as follows, based on the nature of the learning signal or response accessible to a learning system [1]:

1.2.1. Supervised learning

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. The given data is labeled. Both classification and regression problems are supervised learning problems [1].

1.2.2. Unsupervised learning

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. In unsupervised learning algorithms, classification or categorization is not included in the observations [1].

1.2.3. Semi-supervised learning

Semi-supervised learning is an approach to machine learning that combines small labeled data with a large amount of unlabeled data during training. Semi-supervised learning falls between unsupervised learning and supervised learning [1].

1.2.4. Reinforcement learning

Reinforcement learning is the problem of getting an agent to act in the world so as to maximize its rewards. A learner is not told what actions to take as in most forms of machine learning but instead must discover which actions yield the most reward by trying them [1].

1.3. Advantages and disadvantages

1.3.1. Advantages

- It happens automatically: Machine learning automates the whole data interpretation and analysis process. There is no need for males to intervene in data prediction or interpretation. The whole machine learning process begins with the machine learning and predicting the algorithm or program to get the best results [2].
- It can handle a wide range of data: It can manage a wide range of data even in an unpredictable and dynamic environment. It is multifaceted and multitasking [2].
- Scope of innovation: Just as people grow with experience, machine learning improves itself and becomes more precise and efficient at work. This resulted in better judgments [2].

1.3.2. Disadvantages

- More data is required: The more data a machine receives, the more accurate and efficient it becomes, requiring more data to be supplied to the computer for improved predicting or decision-making. Nevertheless, this is not always achievable. Also, the data must be neutral and of high quality. Data needs might be challenging at times [2].
- Time-consuming and more resources needed: There may be instances when the machine's learning process takes a long time since efficacy and efficiency can only be gained via experience, which again takes time. Also, the resources required are greater, such as extra computers [2].
- Imprecision of interpretation of data: As we've seen, a little modification or biased data may lead to a long-drawn mistake chain, thus there's a potential of inaccuracy of interpretation as well. Data that is error-free may be interpreted incorrectly by the machine if the data supplied earlier does not meet all of the machine's requirements [2].

1.4. Application

Machine learning is widely employed in a variety of fields., and one of them can be mentioned as:

1.4.1. Image recognition

One of the most popular uses of machine learning is image recognition. It is used to identify items, people, places, digital photographs, and so forth [3].

1.4.2. Product recommendations

Machine learning is widely used by various e-commerce and entertainment companies such as Amazon, Netflix, etc…, for product recommendations to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser, and this is because of machine learning.

Google understands user interest using various machine learning algorithms and suggests the product as per customer interest.

As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning [3].

1.4.3. Email spam and malware filtering

Whenever we receive a new email, it is filtered automatically as important, normal, and spam. We always receive important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is machine learning [3].

1.4.4. Online fraud detection

Machine learning is making our online transactions safe and secure by detecting fraudulent transactions. Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place such as fake accounts, fake ids, and stealing money in the middle of a transaction [3].

1.4.5. Medical diagnosis

In medical science, machine learning is used for disease diagnoses. With this, medical technology is growing very fast and is able to build 3D models that can predict the exact position of lesions in the brain [3].

1.5. The problem of customers' satisfaction prediction

1.5.1. Customer satisfaction concept

Customer satisfaction is defined as a measurement that determines how happy customers are with a company's products, services, and capabilities. Customer satisfaction information, including surveys and ratings, can help a company determine how to best improve or changes its products and services [4].

For any business, customer satisfaction is a key factor of repurchase, word-of-mouth, and loyalty, and is strongly tied to a company's long-term earnings. While knowing customer behaviors is essential to help businesses improve their customer experience, the complexity of data collected via social media, call centers, websites, and other channels growing on a daily basis keeps businesses far-reaching from valuable customer insights [5].

1.5.2. Problem definition and motivation

The problem to be solved in this study is to predict the level of customer satisfaction (satisfied and neutral/dissatisfied) in airline services.

The motivation for me to do this research is that I want to identify the influencing factors and measure the influence of these factors on customer satisfaction in the hope that the solutions I offer can help improve customer satisfaction customers towards the airline's service and experience in general.

1.5.3. Related work and contribution

- **Related works**

In the world today there are many studies on the topic of predicting customers' satisfaction in many different fields. There are many studies that contribute great results in the field of Science and Technology.

In 2019, there was a study by Sachin Kumar and Mikhail Zymbler and in this paper they present a machine learning strategy to analyzing airline tweets in order to improve the customer satisfaction. Word embedding using Glove dictionary technique and n-gram approach were used to extract features from tweets. Moreover, SVM (support vector machine) and multiple ANN (artificial neural network) architectures were investigated to construct a classification

model that categorizes tweets as positive or negative. Convolutional neural networks (CNN) were also built to categorize tweets, and the results were compared to the best accurate model among SVM and multiple ANN designs. CNN was shown to outperform SVM and ANN models. Finally, association rule mining was done on several types of tweets in order to map the link with sentiment categories. The findings indicate that fascinating relationships were discovered, which would undoubtedly aid the airline industry in improving the consumer experience. (Sachin Kumar, Mikhail Zymbler, 2019)

There was a paper published in 2021 by Luciano Cavalcante Siebert et al. about forecasting client happiness for distribution firms using machine learning. This study aims to assist power distribution businesses in assessing and forecasting consumer happiness. To evaluate customer happiness from service data, power outage data, and reliability indices, the created approach chooses and uses machine learning techniques such as decision trees, support vector machines, and ensemble learning. The expected key indicator findings differed from the survey results with firm customers by only 1.36 percent. (Luciano Cavalcante Siebert, José Francisco Bianchi Filho, Eunelson José da Silva Júnior, Eduardo Kazumi Yamakawa, Angela Catapan, 2021)

Also in 2021, an article in estimating consumer experience for purchases made from the Brazilian e-commerce site Olist was published in Towards Data Science by Paritosh Mahto. The goal is to forecast the rating of customer satisfaction for a particular order based on the information provided, such as pricing, product description, on-time shipment, delivery details, and so on.

Different classification models are used such as Logistic Regression, Linear Support Vector Machine, Decision Tree Classifier, Random Forest Classifier, Boosting Classifier (XGBoost, LightGBM, AdaBoost, CATBoost), Stacking/Voting Ensemble techniques and the performance of the models is evaluated based on f1-score and confusion matrix. The obtained results show that the best model based on the test f1-score is Voting Classifier with score of 0.8068, the highest of all the models tested above. (Paritosh Mahto, 2021)

- **Contributions**

The research articles presented above demonstrated several approaches to the problem of customer satisfaction, from simple machine learning classification models to complex deep learning models to find the best model from which to enhance the customer experience to the services of different businesses.

For my research paper, there are also differences and improvements compared to the above-mentioned research papers. After a period of research, this thesis also has certain contributions as follows:

- Created various experiments such as evaluating the model results before and after optimization.
- There is an additional analysis of model errors, really beneficial for future development which other studies do not have.
- Based on the research results, the airline can have new policies and adjustments accordingly to increase the customer experience to the services.

CHAPTER 2. METHODOLOGY TO PREDICT PASSENGERS' SATISFACTION

In this chapter, I will present the steps to solve the classification problem, introduce the dataset and describe the methods used.

2.1. Predicting customers' satisfaction as a classification problem

The first thing that I need to do when starting a machine learning project is to identify the problem. The problem I need to solve in this graduate thesis is predicting airline passenger satisfaction.

As for data collection, the data set that was used in this project is survey data by an anonymous airline about its customers. So I basically got a dataset in CSV format without having to go do the actual data collection, which can be very time consuming and quite inconvenient.

With a set of raw data in hand, the first thing any data maker needs to do is preprocess the data. In this step I perform basic operations such as cleaning the data and transforming the data for later analysis.

To gain a deeper understanding of my dataset, I perform exploratory data analysis using statistical and visualization techniques.

Finally, I come to the most difficult but equally interesting part of building predictive models, presenting and interpreting the meaning of the experimental results.

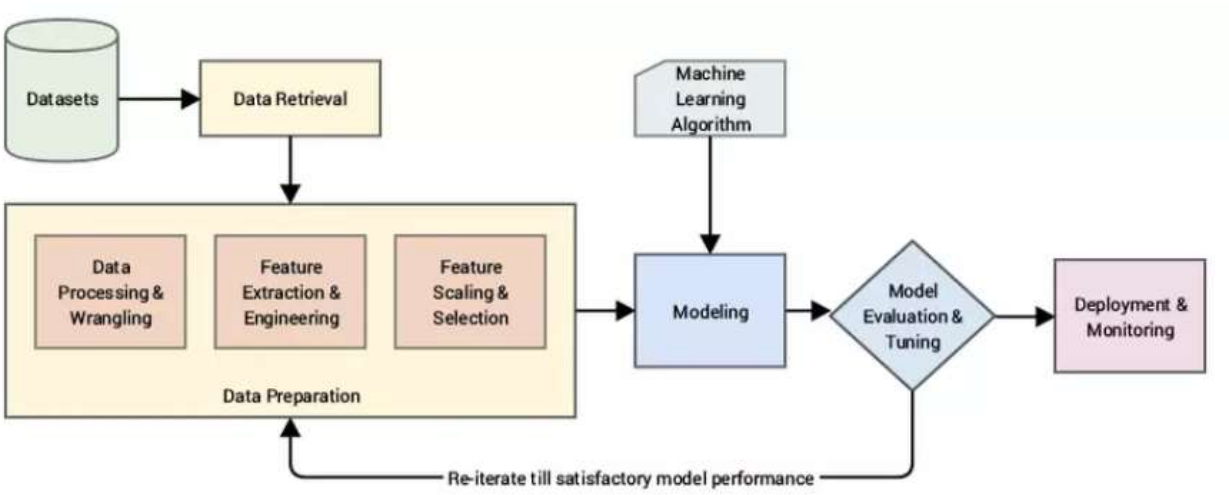

Figure 1. The standard process for deploying a machine learning model

A typical standard process based on the industry standard model CRISP-DM (cross-industry standard process for data mining) is depicted as shown above.

Any intelligent system basically consists of steps starting from inputting raw data, using techniques to organize, process, and design meaningful features and attributes from the data. Then, we often use techniques such as statistical modeling or machine learning models to build models for the purpose of solving the posed requirement.

2.2. Introduction to the dataset

The dataset was originally uploaded on kaggle.com by TJ Klein and it includes a poll of airline passengers' satisfaction. This dataset includes 2 files: train.csv (80% of full dataset to use for training) and test.csv (20% of full dataset to use for testing). It has a total of 23 attributes, namely:

Table 1. List of attributes and its definitions in the dataset

No	Attribute	Description
1	Gender	Gender of the passengers (Female, Male)
2	Customer Type	The customer type (Loyal customer, disloyal customer)
3	Age	The actual age of the passengers
4	Type of Travel	Purpose of the flight of the passengers (Personal Travel, Business Travel)
5	Class	Travel class in the plane of the passengers (Business, Eco, Eco Plus)
6	Flight distance	The flight distance of this journey
7	Inflight wifi service	Satisfaction level of the inflight wifi service (0: not rated; 1-5)
8	Departure/Arrival time convenient	Satisfaction level of departure/arrival time convenient (0: not rated; 1-5)
9	Ease of Online booking	Satisfaction level of online booking (0: not rated; 1-5)
10	Gate location	Satisfaction level of gate location (0: not rated; 1-5)
11	Food and drink	Satisfaction level of food and drink (0: not rated; 1-5)
12	Online boarding	Satisfaction level of online boarding (0: not rated; 1-5)

13	Seat comfort	Satisfaction level of seat comfort (0: not rated; 1-5)
14	Inflight entertainment	Satisfaction level of inflight entertainment (0: not rated; 1-5)
15	On-board service	Satisfaction level of on-board service (0: not rated; 1-5)
16	Leg room service	Satisfaction level of leg room service (0: not rated; 1-5)
17	Baggage handling	Satisfaction level of baggage handling (0: not rated; 1-5)
18	Check-in service	Satisfaction level of check-in service (0: not rated; 1-5)
19	Inflight service	Satisfaction level of inflight service (0: not rated; 1-5)
20	Cleanliness	Satisfaction level of cleanliness (0: not rated; 1-5)
21	Departure Delay in Minutes	Minutes delayed when departure
22	Arrival Delay in Minutes	Minutes delayed when arrival
23	Satisfaction	Airline satisfaction level (Satisfaction, neutral or dissatisfaction)

2.3. Algorithms used in predictive modeling

For this research, I use 7 algorithms for machine learning model and these algorithms include decision tree, random forest, support vector machine, extreme gradient boosting, light gradient-boosting machine, adaptive boosting and category boosting.

The selection of the above algorithms is based on the popularity of the algorithms used in similar studies and the suitability to the problem that this paper is trying to solve.

We will go into detail to understand the concept, how it works as well as the benefits and limitations of each algorithm.

2.3.1. Decision tree

- **Definition**

A decision tree is a non-parametric supervised learning approach that may be used for classification as well as regression applications. It features a tree-like structure with a root node, branches, internal nodes, and leaf nodes [6].

- **How it works**

There are other techniques for constructing a decision tree, however for this project, I choose ID3 (Iterative Dichotomiser 3) - an algorithm developed by Ross Quinlan in 1986 for generating a decision tree from a dataset. The ID3 method builds decision trees by traversing the space of potential branches in a top-down Greedy Search Approach that does not allow for backtracking. When we employ greedy algorithms, we always make the option (choose decision nodes) which appears as the optimal at the time. [7].

ID3 algorithm steps:

Step 1: It starts with the original root node, S.

Step 2: The algorithm iterates over the set S's most underused characteristic, calculating Entropy(H) and Information gain(IG) for each iteration.

Step 3: It then selects the attribute which has the smallest Entropy or largest Information gain.

Step 4: The set S is then split by the selected attribute to produce a subset of the data.

Step 5: The algorithm continues to recur on each subset, considering only attributes never selected before.

- **Benefits and limitations**

  Benefits

  - Easy to interpret: Decision trees are simple to grasp and consume because of their Boolean logic and visual representations. A decision tree's hierarchical architecture also makes it easier to identify which traits are most significant, which isn't always evident with other methods, such as neural networks [6].
  - Data preparation is minimal to non-existent: Decision trees have several qualities that make them more adaptable than other classifiers. It can handle a variety of data formats, including discrete and continuous values, and continuous values may be transformed to categorical values using thresholds. It can also handle missing values, which may be troublesome for other classifiers such as Naïve Bayes [6].
  - More adaptable: Decision trees may be used for classification and regression problems, making them more adaptable than other algorithms. It's also indifferent to underlying correlations between qualities; for example, if two variables are highly connected, the algorithm will only split on one of them [6].

Limitations

- Prone to overfitting: Complicated decision trees are prone to overfitting and may not generalize well to fresh data. This situation can be prevented by using pre- or post-pruning techniques. When there is insufficient data, pre-pruning stops tree growth, but post-pruning eliminates subtrees with insufficient data after tree formation. [6].
- High variance estimators: Slight differences in data might result in a drastically different decision tree. Bagging, or averaging estimates, can be used to reduce the variation in decision trees. This strategy, however, has limitations since it might result in strongly linked predictors [6].
- More expensive to train: Since decision trees use a greedy search method during construction, they might be more challenging to train than other algorithms [6].

2.3.2. Random forest

– **Definition**

Random forest is a popular supervised machine learning technique for classification and regression tasks. It constructs decision trees from several samples and uses their majority of votes for classification and average for regression.

One of the most essential characteristics of the random forest technique is that it can handle data sets with both continuous and categorical variables, as in regression and classification. It produces better results for classification tasks [8].

– **How it works**

The steps in the random forest algorithm [8]:

Step 1: In Random forest, n random records are chosen at random from a data collection of k records.

Step 2: For each sample, a unique decision tree is built.

Step 3: Each decision tree will deliver an outcome.

Step 4: For classification and regression, the final result is dependent on majority vote or average.

- **Benefits and limitations**

  Benefits

  - Overfitting risk is reduced: Decision trees are prone to overfitting because they tend to tightly fit all samples within training data. Nevertheless, when a random forest has a large number of decision trees, the classifier will not overfit the model since averaging uncorrelated trees reduces overall variance and prediction error [9].
  - Provides flexibility: Because random forest can handle both regression and classification problems with high accuracy, it is a popular approach among data scientists. Since it retains accuracy when a portion of the data is absent, feature bagging makes the random forest classifier a useful tool for guessing missing values [9].
  - Simple to assess feature significance: Random forest makes it simple to assess variable relevance, or contribution, to the model. There are several methods for determining the relevance of a characteristic. Gini importance and mean drop in impurity (MDI) are commonly used to quantify how much the model's accuracy decreases when a certain variable is removed. Permutation importance, also known as mean decrease accuracy (MDA), is another measure of significance. By randomly permuting the feature values in oob samples, MDA determines the average loss in accuracy [9].

Limitations

- Process takes time: Since random forest algorithms can handle enormous data sets, they can produce more accurate predictions, but they can be sluggish to process data because they compute data for each individual decision tree [9].
- More resources are required: Because random forests analyze bigger data sets, more resources are required to store that data [9].
- More complex: A single decision tree's forecast is easier to comprehend than a forest of them [9].

2.3.3. Support vector machine

- **Definition**

  Support vector machines are a type of supervised learning algorithms used for classification, regression, and detecting outliers [10].

  We may utilize certain forms of SVM for specific machine learning issues, such as support vector regression (SVR), which is an extension of support vector classification (SVC).

- **How it works**

  SVM works by mapping data to a high-dimensional feature space in order to categorize data points that are otherwise not linearly separable. A separator between the categories is discovered, and the data are processed so that the separator may be drawn as a hyperplane. Following that, fresh data features may be utilized to determine the group to which a new record should belong [11].

– **Benefits and limitations**

Benefits

- When there is an understandable margin of dissociation between classes, the support vector machine performs reasonably well [12].
- It is more efficient in high-dimensional spaces [12].
- It is useful when the number of dimensions is greater than the number of specimens [12].
- Memory systematic support vector machine is comparable [12].

Limitations

- For huge data sets, the support vector machine approach is unsuitable [10].
- It performs poorly when the data set has more noise, i.e. target classes overlap [12].
- When the number of attributes per each data point exceeds the number of training data specimens, the support vector machine will perform poorly [12].
- There is no probabilistic justification for the classification because the support vector classifier operates by positioning data points above and below the classifying hyperplane [12].

2.3.4. Extreme gradient boosting

– **Definition**

XGBoost is a machine learning technique that focuses on model performance and computation speed. The technique is designed to operate with big and complex datasets and may be used for both regression and classification problems [13].

– **How it works**

Consider a function or estimate. To begin, we create a series based on the function gradients. The equation below represents a specific type of gradient descent. The represents the loss function to minimize, hence it indicates the direction in which the function declines. The rate of change fitted to the loss

function is comparable to the gradient descent learning rate. is intended to accurately replicate the loss's behavior.

To iterate over the model and find the optimal definition we need to express the whole formula as a sequence and find an effective function that will converge to the minimum of the function. This function will serve as an error measure to help us decrease the loss and keep the performance over time. The sequence converges to the minimum of the function. This particular notation defines the error function that applies when evaluating a gradient boosting regressor [14].

- **Benefits and limitations**

  Benefits

  - Gradient Boosting is an easy-to-read and comprehend method, which makes most of its predictions simple to manage [13].
  - Boosting is a sturdy and strong approach that quickly avoids and reduces over-fitting [13].
  - XGBoost performs exceptionally well on medium, tiny, data with subgroups, and structured datasets with a limited number of characteristics [13].
  - It is an excellent method to take because the vast majority of real-world issues include classification and regression, two tasks for which XGBoost reigns supreme [13].

  Limitations

  - XGBoost performs poorly on sparse and unstructured data [13].
  - Gradient Boosting is particularly sensitive to outliers since each classifier is pushed to correct the errors made by the previous learners [13].
  - Overall, the approach is not scalable. This is due to the estimators basing their accuracy on earlier predictors, making the operation difficult to simplify [13].

2.3.5. Light gradient-boosting machine

- **Definition**

  Light GBM is a high-performance gradient boosting framework based on the decision tree technique that may be used for ranking, classification, and a variety of other machine learning applications [15].

- **How it works**

  LightGBM employs a histogram-based approach in which data is bucketed into bins based on the distribution's histogram. Instead of using each data point, the bins are utilized to iterate, calculate the gain, and partition the data. This approach can also be optimized for sparse datasets [16].

- **Benefits and limitations**

  Benefits

  - Faster training speed and higher efficiency: LightGBM implements a histogram-based technique, which buckets continuous feature values into discrete bins, hence speeding up the training operation [15].
  - Reduced memory utilization: Converts continuous data to discrete bins, resulting in reduced memory usage [15].
  - Better than any other boosting method in terms of accuracy: It generates far more complicated trees by using a leaf-wise split strategy rather than a level-wise split approach, which is the primary contributor in obtaining greater accuracy [15].
  - Compatible with huge datasets: It can perform similarly well with huge datasets while requiring far less training time than XGBoost [15].

  Limitations

  - Overfitting: When LightGBM splits the tree leaf-wise, it might result in overfitting because it creates more complicated trees.
  - Compatibility with datasets: Since LightGBM is sensitive to overfitting, it can easily overfit tiny datasets.

2.3.6. Adaptive boosting

- **Definition**

  AdaBoost, standing for Adaptive Boosting, is a statistical classification meta-algorithm developed in 1995 by Yoav Freund and Robert Schapire, who were awarded the Gödel Prize in 2003 for their work. It may be used with a variety of different types of learning algorithms to increase performance. The output of the other learning algorithms ('weak learners') is blended into a weighted sum that reflects the boosted classifier's final output. AdaBoost is often used for binary classification, although it may be extended to multiple classes or limited intervals on the real line [17].

- **How it works**

  Step 1: Based on the weighted samples, a weak classifier (e.g., a decision stump) is built on top of the training data. The weights of each sample reflect how critical it is to be accurately identified in this case. For the first step, we assign identical weights to all samples [18].

  Step 2: We design a decision stump to every feature and examine how effectively each stump classifies data into their respective target groups [18].

  Step 3: Additional weight is applied to wrongly categorized samples in order for them to be correctly categorised in the next decision stump. Weight is also allocated to each classifier depending on its accuracy, thus high accuracy equals high weight [18].

  Step 4: Repeat step 2 until all data points have been accurately categorised or the maximum iteration level is reached [18].

- **Benefits and limitations**

  Benefits

  - Unlike SVM, it is simpler to use and requires less adjusting of settings [19].
  - It is not prone to overfitting [19].
  - It may be used to increase the accuracy of weak classifiers, making it more adaptable [19].

Limitations

- It requires a high-quality dataset [19].
- When implementing an Adaboost algorithm, it is necessary to prevent noisy data and outliers [19].

2.3.7. Category boosting

– **Definition**

Yandex's CatBoost is a recently open-sourced machine learning algorithm. It integrates easily with deep learning frameworks such as Google's TensorFlow and Apple's Core ML. It can operate with a variety of data formats to assist organizations tackle a wide range of challenges. It also has the highest accuracy in its class [20].

– **How it works**

CatBoost is built on decision trees that have been gradient boosted. During training, a sequence of decision trees is created. Each succeeding tree is created with less loss than the prior ones.

The number of trees is determined by the initial settings. Use the overfitting detector to avoid overfitting. When it is activated, tree construction is halted [21].

– **Benefits and limitations**

Benefits

- It gives us great results for categorical data [22].
- It can train our model on GPU that significantly increase the speed of learning [22].

Limitations

- It performs only better than other algorithms only when we have categorical data [22].
- Can perform very bad if the variables are not properly tuned [22].

CHAPTER 3. MODEL BUILDING AND EVALUATION

In this chapter, I will first present insights from data through statistical techniques and data visualization. I will then perform the data preparation methods and finally build, test, evaluate the model and analyze the errors from the results obtained.

3.1. Exploratory data analysis

Before preprocessing data and building the model, it is preferable to use EDA to analyse, examine, and summarize their primary traits so that we can better understand the insights of the data before making any assumptions.

First, we need to check whether the dataset's target value is balanced.

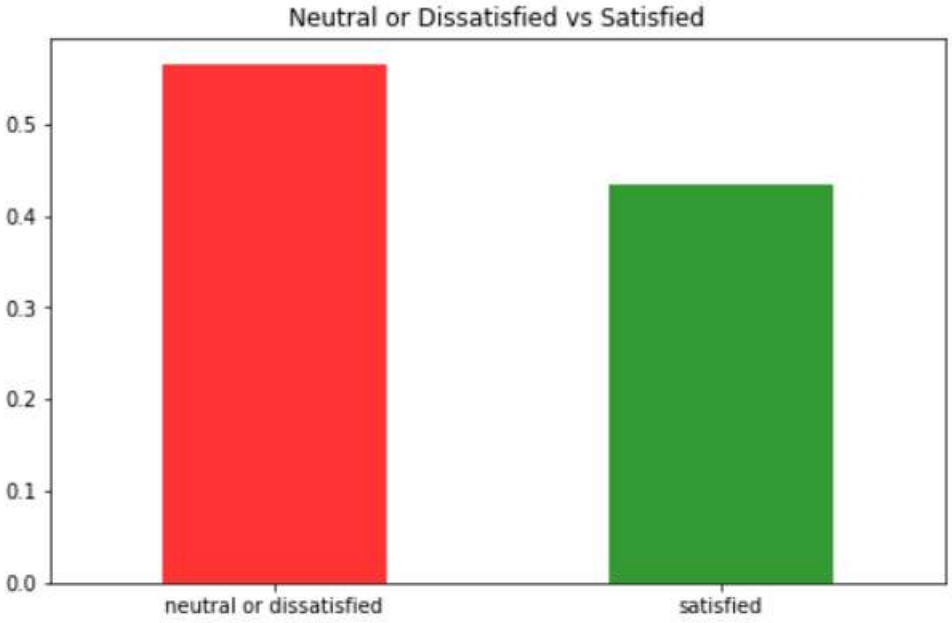

Figure 2. Ratio of satisfied vs neutral or dissatisfied customers

The above graphic illustrates a roughly 55%:45% split between neutral/dissatisfied passengers and satisfied passengers. As a result, the data is well-balanced and does not require any extra treatment or resampling.

3.1.1. Visualizing categorical features

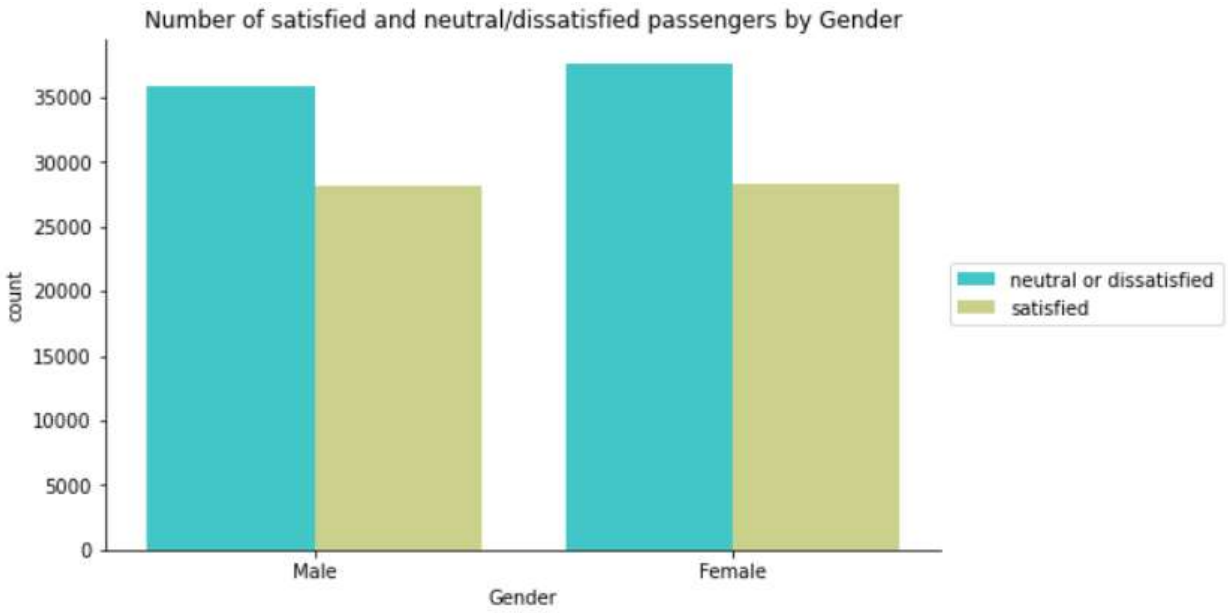

Figure 3. Number of satisfied and neutral or dissatisfied passengers by Gender

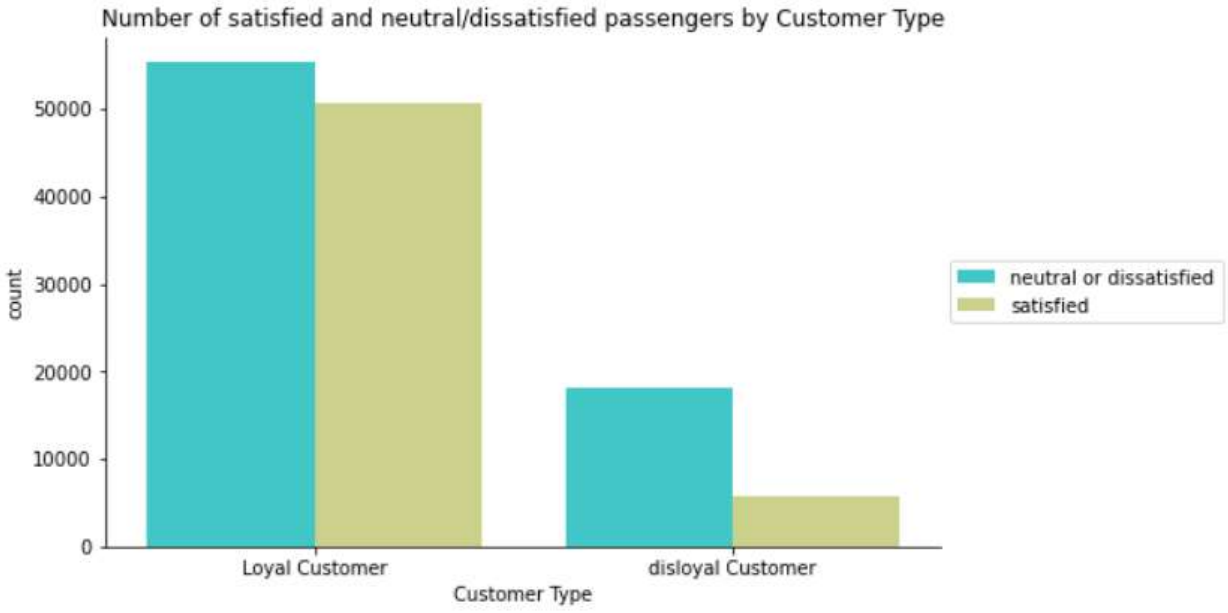

Figure 4. Number of satisfied and neutral or dissatisfied passengers by Customer Type

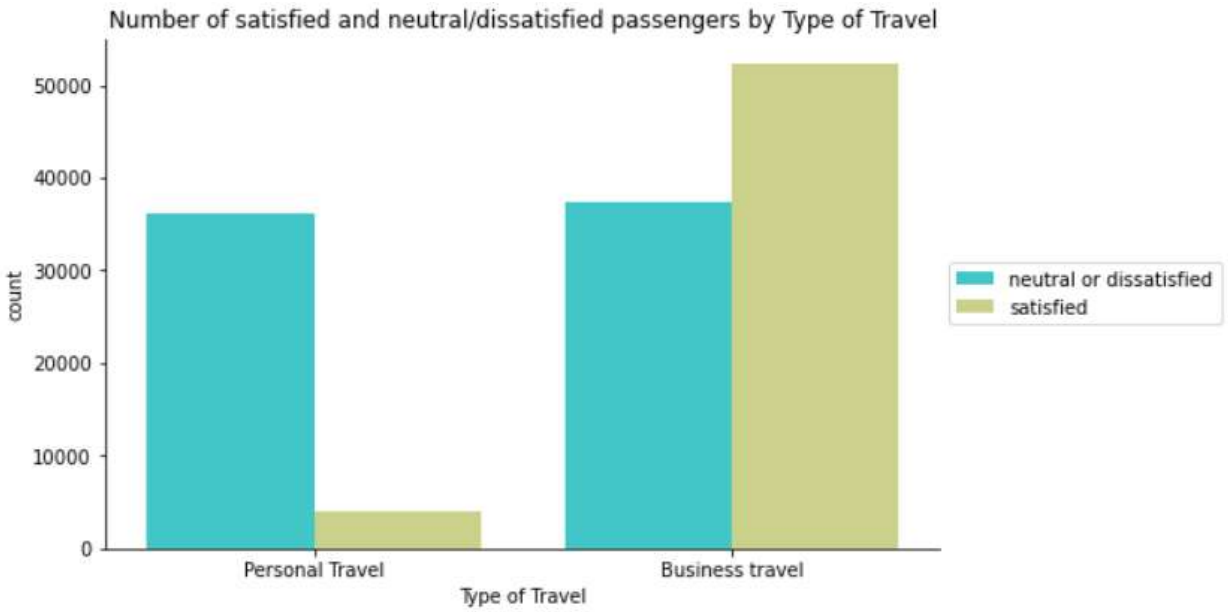

Figure 5. Number of satisfied and neutral or dissatisfied passengers by Type of Travel

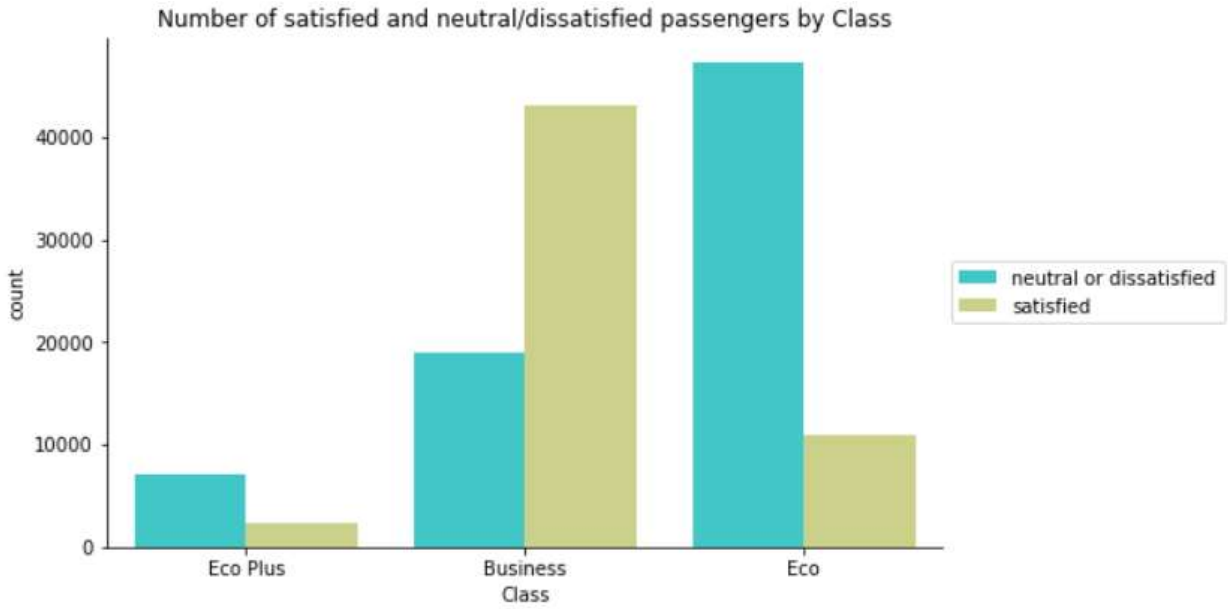

Figure 6. Number of satisfied and neutral or dissatisfied passengers by Class

Some conclusions from the visualization of categorical features:

- Gender does not appear to have a significant impact on satisfaction, since men and women appear to be equally concerned with the same issues.
- The vast majority of the airline's passengers are loyal customers. However, regardless of loyalty, the level of dissatisfaction is high.
- Business travelers seem to be happier with the flight than personal travelers.
- People in business class seem to be the most satisfied, while those in economy class appear to be the less happy.

3.1.2. Visualizing numerical features

Due to the relatively large number of images in this section, only a few images are included for representation.

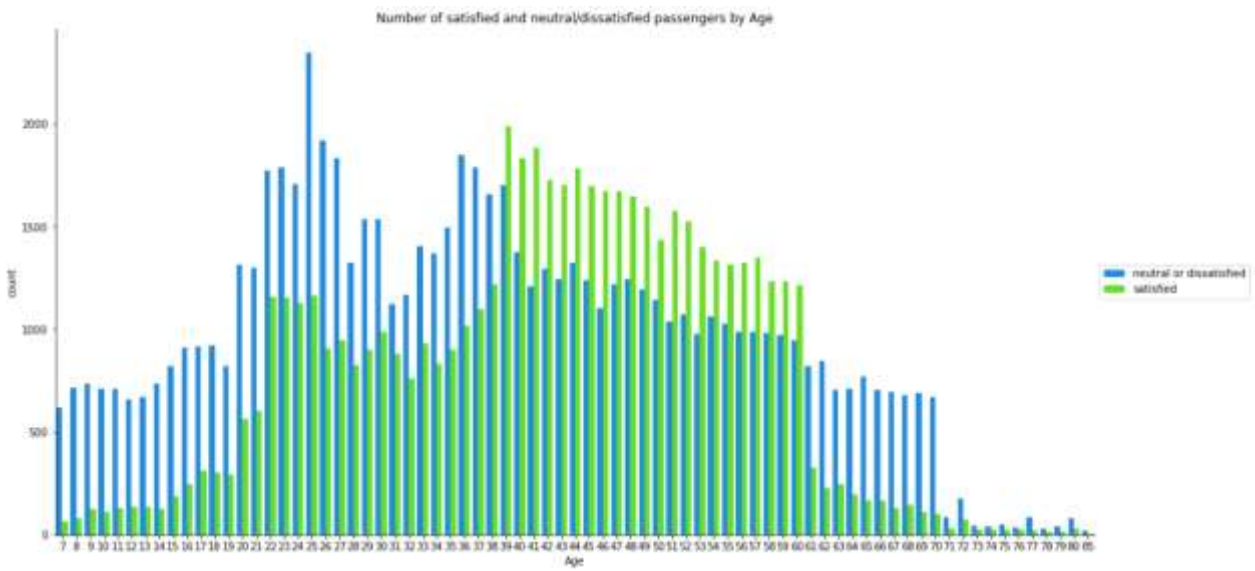

Figure 7. Number of satisfied and neutral or dissatisfied passengers by Age

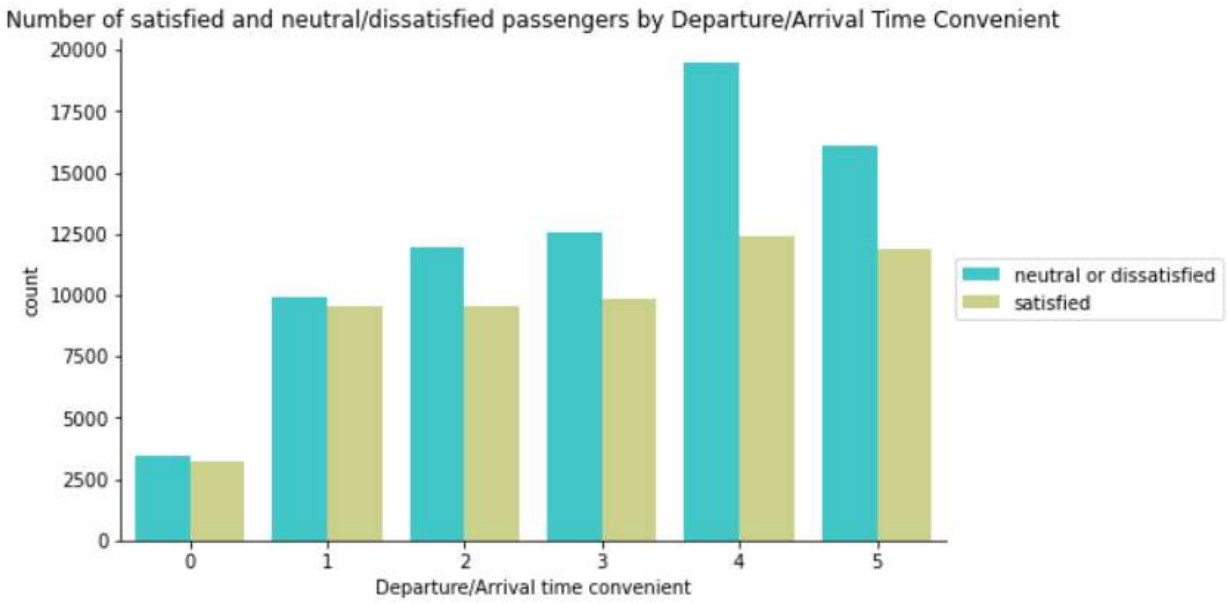

Figure 8. Number of satisfied and neutral or dissatisfied passengers by Departure/Arrival Time Convenient

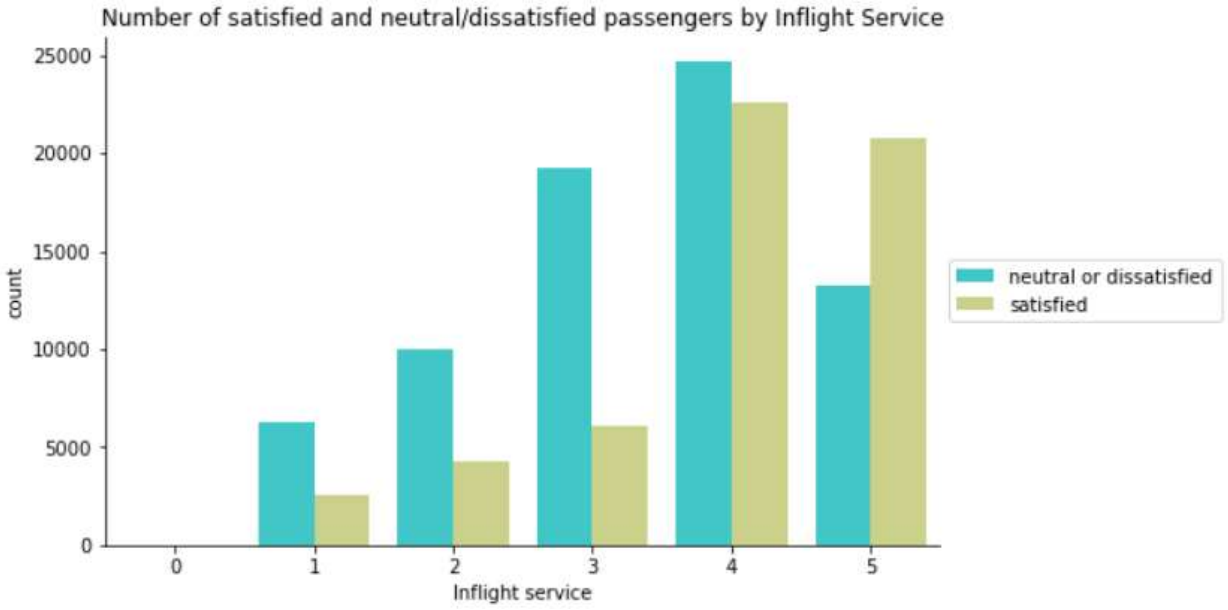

Figure 9. Number of satisfied and neutral/dissatisfied passengers by Inflight Service

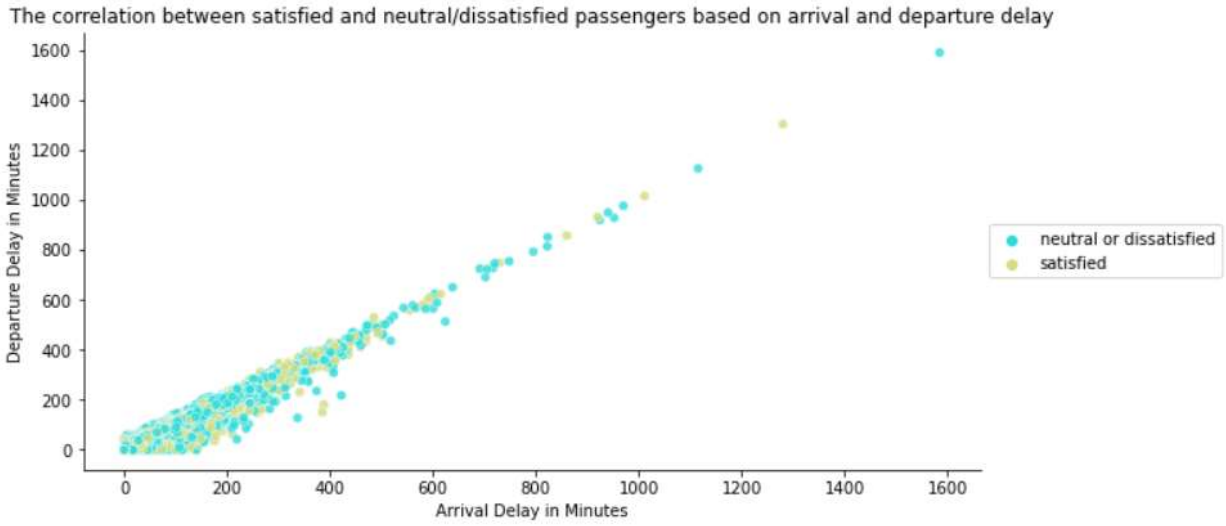

Figure 10. The correlation between satisfied and neutral or dissatisfied passengers based on arrival and departure delay

Some conclusions from the visualization of numerical features:

- Between the ages of 7 to 38, and 61 to 79, the proportion of unsatisfied travelers is relatively significant. In contrast, the proportion of satisfied passengers is greater than the proportion of unsatisfied passengers in the age range 39-60.
- For pre-flight as well as in-flight services such as online booking, luggage packing, Wifi, entertainment, cleaning, food, etc., most passengers rate 4 and 5 points. feel satisfied with the flight. In contrast to the services rated from 1 to 3, the number of customers who are not satisfied with the flight is quite large.
- Looks like Departure/Arrival Time Convenient does not accurately reflect the level of customer satisfaction because regardless of the rating, the number of unhappy consumers is always greater than the number of pleased customers.
- The arrival and departure delays appear to be linear, which makes perfect sense and there was one client who was pleased despite the 1300-minute delay.

3.2. Data preprocessing

Data preprocessing is a method we must apply to convert raw data into a usable format. It is regarded as a critical phase in data mining since real-world data is frequently incomplete, inconsistent, or includes several mistakes. So, when dealing with the dataset of airline passengers, we typically perform certain data pretreatment techniques.

First, let's take a look at some information about the dataset.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 129880 entries, 0 to 25975
Data columns (total 25 columns):
 #   Column                            Non-Null Count   Dtype
---  ------                            --------------   -----
 0   Unnamed: 0                        129880 non-null  int64
 1   id                                129880 non-null  int64
 2   Gender                            129880 non-null  object
 3   Customer Type                     129880 non-null  object
 4   Age                               129880 non-null  int64
 5   Type of Travel                    129880 non-null  object
 6   Class                             129880 non-null  object
 7   Flight Distance                   129880 non-null  int64
 8   Inflight wifi service             129880 non-null  int64
 9   Departure/Arrival time convenient 129880 non-null  int64
 10  Ease of Online booking            129880 non-null  int64
 11  Gate location                     129880 non-null  int64
 12  Food and drink                    129880 non-null  int64
 13  Online boarding                   129880 non-null  int64
 14  Seat comfort                      129880 non-null  int64
 15  Inflight entertainment            129880 non-null  int64
 16  On-board service                  129880 non-null  int64
 17  Leg room service                  129880 non-null  int64
 18  Baggage handling                  129880 non-null  int64
 19  Checkin service                   129880 non-null  int64
 20  Inflight service                  129880 non-null  int64
 21  Cleanliness                       129880 non-null  int64
 22  Departure Delay in Minutes        129880 non-null  int64
 23  Arrival Delay in Minutes          129487 non-null  float64
 24  satisfaction                      129880 non-null  object
dtypes: float64(1), int64(19), object(5)
```

Figure 11. Information about datasets

3.2.1. Data cleaning

One of the most critical tasks we must do before proceeding is to clean the data to ensure that there are no missing or duplicate values in the dataset we are given. So we will examine the dataset's basic information to see whether it contains any missing or duplicated data.

The first two factors are meaningless and have no bearing on the classification, thus they should be removed.

Check for duplicate values:

```
data.duplicated().any()
```

```
False
```

Figure 12. Check duplicate values

The result is False which means there are no duplicate values in the data set.

Check for missing values:

```
data.isna().sum()
```

```
Gender                                0
Customer Type                         0
Age                                   0
Type of Travel                        0
Class                                 0
Flight Distance                       0
Inflight wifi service                 0
Departure/Arrival time convenient     0
Ease of Online booking                0
Gate location                         0
Food and drink                        0
Online boarding                       0
Seat comfort                          0
Inflight entertainment                0
On-board service                      0
Leg room service                      0
Baggage handling                      0
Checkin service                       0
Inflight service                      0
Cleanliness                           0
Departure Delay in Minutes            0
Arrival Delay in Minutes            393
satisfaction                          0
dtype: int64
```

Figure 13. Check missing values

We can see that the column corresponding to the Arrival Delay in Minutes feature has 393 missing values, therefore we have to replace the missing data.
Because the missing values belong to a numerical variable, I will use the median function to replace the missing data.

```
data['Arrival Delay in Minutes'].fillna(data['Arrival Delay in Minutes'].median(axis = 0), inplace = True)
```

Figure 14. Replace missing values

3.2.2. Data transformation

We all know that machine learning models are built on mathematical equations, and it is easy to see how keeping non-numerical data in the equations might cause issues. We are creating models with ML, yet the computer only understands and processes numeric data types. That is why the non-numerical/categorical variables must be encoded.
In this dataset, we will transform all data types into numerical data to help the model with consistent data types and make it easy for us to develop any models so far.

First we will use the Label Encoder method for all the category variables.

```
label_encoder = preprocessing.LabelEncoder()
cat_cols = ['satisfaction', 'Gender', 'Customer Type', 'Type of Travel', 'Class']
for cols in data[cat_cols]:
    data[cols] = label_encoder.fit_transform(data[cols])
```

Figure 15. Label encoding

Then we will use the get_dummies method for the remaining variables.

```
data = pd.get_dummies(data)
```

Figure 16. Get dummies

To get better prediction results, we need to remove outliers in the dataset.

```
Q1 = data.quantile(0.25)
Q3 = data.quantile(0.75)
IQR = Q3 - Q1
data = data[~((data < (Q1 - 1.5 * IQR)) |(data > (Q3 + 1.5 * IQR))).any(axis=1)]
data.shape
```

```
(76442, 23)
```

Figure 17. Outliers detection and removal

After removing outliers, the number of records decreased from 129880 to 76442 records.

The last stage that we will perform in the data transformation process was to scale the dataset with the function StandardScaler. This step was added since it would improve algorithm normalization. Because the range of values in certain raw data differs greatly, it may impact the ML algorithm and lead it to predict inaccurately.

```
X = data.drop('satisfaction', axis=1)
y = data['satisfaction']

scale = StandardScaler()
X = scale.fit_transform(X)
```

Figure 18. Scale dataset

3.3. Model building

Before proceeding to develop the model, the dataset must be divided into training and testing sets using the function train_test_split.

We will split the data at the rate of 80% for the train set and 20% for the test set as the original data set provided.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Figure 19. Split dataset

After referring to similar studies as well as experiments on different algorithms, the following models are selected:

Model 1: Decision Tree

Model 2: Random Forest

Model 3: SVC

Model 4: XGBoost

Model 5: LightGBM

Model 6: AdaBoost

Model 7: CatBoost

3.4. Model evaluation

A classification model can be evaluated in a variety of ways. We employ several strategies depending on the nature of the challenge. For a problem where the predicted outcome is 2 labels, the usual methods used are: Accuracy score, Precision, Recall, F1 score and ROC AUC.

We will perform two experiments, one is compare the effectiveness of models without feature selection and with feature selection, the other is compare the effectiveness of the models without hyperparameter tuning and with hyperparameter tuning.

Experiment 1: Compare the effectiveness of models without feature selection and with feature selection.

For the selection of features as input to the model, I have generated a heatmap, a helpful way to understand the correlation between variables, and from there I can remove those variables that are not highly correlated with the target variable.

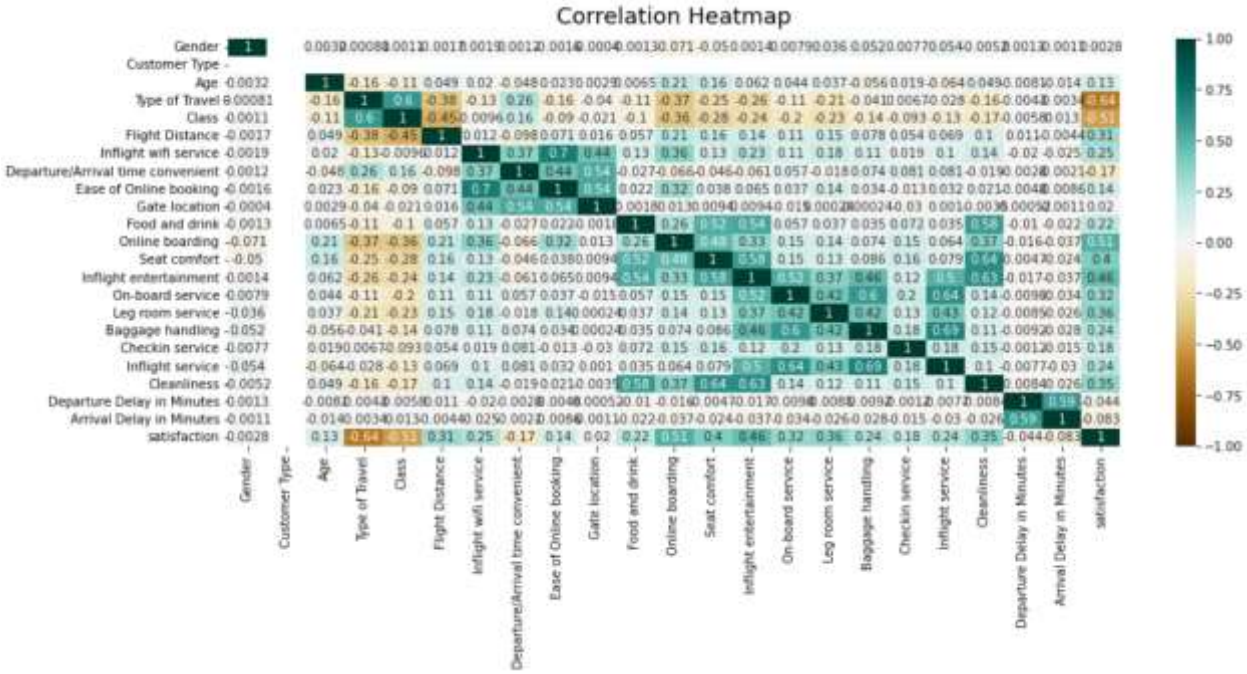

Figure 20. Correlation heatmap

As we can see from the heatmap, there are two variables that are least correlated with the predictor is Gender (0.0028) and Gate location (0.02) so we should remove them to get more accurate results from the predictive model.

Table 2. Experimental result without feature selection (%)

Model	Predicted labels	Precision	Recall	ROC AUC
Decision Tree	satisfied	94.17	93.84	94.66
	neutral or dissatisfied	95.23	95.49	
Random Forest	satisfied	97.56	93.84	96.01
	neutral or dissatisfied	95.36	98.18	
SVC	satisfied	96.16	93.17	95.14
	neutral or dissatisfied	94.82	97.11	
XGBoost	satisfied	97.15	94.23	96.04
	neutral or dissatisfied	95.62	97.85	
LightGBM	satisfied	97.82	93.95	96.16
	neutral or dissatisfied	95.44	98.38	
AdaBoost	satisfied	93	90.95	92.81
	neutral or dissatisfied	93.08	94.70	
CatBoost	satisfied	97.20	93.67	95.79
	neutral or dissatisfied	95.21	97.90	

Table 3. Experimental result with feature selection (%)

Model	Predicted labels	Precision	Recall	ROC AUC
Decision Tree	satisfied	95.54	95.88	95.64
	neutral or dissatisfied	95.75	95.40	
Random Forest	satisfied	98.18	95.46	96.82
	neutral or dissatisfied	95.46	98.18	
SVC	satisfied	97.15	94.86	96
	neutral or dissatisfied	94.85	97.14	
XGBoost	satisfied	97.62	95.81	96.70
	neutral or dissatisfied	95.77	97.60	
LightGBM	satisfied	98.65	95.40	97.03
	neutral or dissatisfied	95.43	98.66	
AdaBoost	satisfied	93.80	93.66	93.65
	neutral or dissatisfied	93.50	93.63	
CatBoost	satisfied	98.24	95.21	96.73
	neutral or dissatisfied	95.23	98.25	

It is not surprising that the results of models with selected input attributes give higher accuracy than models that keep all attributes as input.

Experiment 2: Compare the effectiveness of the models without hyperparameter tuning and with hyperparameter tuning.

The hyperparameter tuning for the model is inspired by research papers on the same topic and this public dataset. There are two most commonly used methods to adjust parameters for the model is grid search and random search. After the model is tuned using the above two techniques, the appropriate hyperparameters for each type of model are obtained. From there, the most common hyperparameters with each model type are taken to adjust for the models in this study.

Table 4. Experimental result without hyperparameter tuning (%)

Model	Predicted labels	Precision	Recall	ROC AUC
Decision Tree	satisfied	94.78	95.43	95.02
	neutral or dissatisfied	95.27	94.60	
Random Forest	satisfied	98.10	95.48	96.79
	neutral or dissatisfied	95.48	98.10	
SVC	satisfied	97.16	94.93	96.03
	neutral or dissatisfied	94.90	97.15	
XGBoost	satisfied	98	95.61	96.80
	neutral or dissatisfied	95.60	97.99	
LightGBM	satisfied	98.92	95.31	97.12
	neutral or dissatisfied	95.36	98.94	
AdaBoost	satisfied	93.80	93.52	93.58
	neutral or dissatisfied	93.36	93.65	
CatBoost	satisfied	98.17	95.62	96.90
	neutral or dissatisfied	95.62	98.17	

Table 5. Experimental result with hyperparameter tuning (%)

Model	Predicted labels	Precision	Recall	ROC AUC
Decision Tree	satisfied	95.54	95.88	95.64
	neutral or dissatisfied	95.75	95.40	
Random Forest	satisfied	98.18	95.46	96.82
	neutral or dissatisfied	95.46	98.18	
SVC	satisfied	97.15	94.86	96
	neutral or dissatisfied	94.85	97.14	
XGBoost	satisfied	97.62	95.81	96.70
	neutral or dissatisfied	95.77	97.60	
LightGBM	satisfied	98.65	95.40	97.03
	neutral or dissatisfied	95.43	98.66	
AdaBoost	satisfied	93.80	93.66	93.65
	neutral or dissatisfied	93.50	93.63	
CatBoost	satisfied	98.24	95.21	96.73
	neutral or dissatisfied	95.23	98.25	

Looking at the results obtained when comparing the accuracy between hyperparameter tuned models and models with default hyperparameters, not all tuned models have higher accuracy than models with default hyperparameters. Only 3 out of 7 models, decision tree, random forest, and adaboost, performed better when adjusted. The remaining 4 models svc, xgboost, lightgbm and catboost give better results with default hyperparameters. From the above, it is necessary to have more optimal tuning methods for the models to bring the best results.

3.5. Error analysis

Based on the test results from the models, I analyzed the error sets extracted from the models.

Errors are divided into two categories:

False positive: This implies that it is, in fact, a neutral or dissatisfied customer but mistakenly guesses that it is a satisfied customer.

False negative: This implies that it is, in fact, a satisfied customer, but mistakenly guesses that it is a neutral or dissatisfied customer.

After performing analysis on the above two types of errors, I come up with a few observations:

- Most of the false predictions come from surveys of loyal customers. This is also understandable when the number of surveyors who are loyal customers outnumbers disloyal customers.
- In terms of the flight's purpose, it seems to have little effect on the model's inaccurate predictions because the rate of false predictions between passengers traveling for personal and business purposes is the same.
- Passengers in the eco plus class are less likely to be incorrect in terms of class. In contrast, the model produced more mistakes in the eco class passengers.
- The model made false predictions mainly in the passengers who rated the airline's services at ratings 4 and 5, in which rating 4 accounted for the majority.

CHAPTER 4. CONCLUSION

In the last chapter, I will summarize the research study outcomes, the challenges I experienced, and the lessons and experiences I learned after completing the graduation thesis.

1. Conclusion

In order to be able to solve the problem of predicting airline passenger satisfaction, I had to approach related research papers as well as the advice of the instructor to solve the problem in the most reasonable and effective way.

Luckily I got a pre-collected data set from kaggle because it is really hard to get a survey of each customer using a particular airline.

After many times of testing with different models with different methods, there are seven machine learning models selected to deploy and among them, the predictive model that gives the best results is LightGBM when given input attributes and using default hyperparameters with a roc auc score of 97.12% and precision and recall of 98.92% and 95.31% respectively with satisfied labels, 95.36% and 98.94% with neutral or dissatisfied labels.

2. Limitations of the study

In addition to the objectives that the research topic has achieved, the study also has some limitations as follows:

- The study cannot avoid flaws due to the restricted time and lack of knowledge.
- The study only conducts customer satisfaction research for a specific airline, but has not conducted research for other airlines in the market, so it is not possible to evaluate the comparison between airlines with each other.
- Methods to deploy and evaluate the model are still limited.

3. Experiences and lessons learned from the study

During the process of doing my graduation thesis and after completing it, I realized that I have learned and improved a lot of different skills from researching, reading and understanding research papers on the same topic, how to see and solve problems from different perspectives, develop more communication skills and social network. Most importantly, I gained numerous professional knowledge and specialized skills from my instructor to better my expertise and abilities. In brief, both professional knowledge and soft skills are greatly enhanced.

REFERENCES

[1] An introduction to Machine Learning (2017, August 24). Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/introduction-machine-learning/

[2] Prasanna (2022, April 22). Advantages and Disadvantages of Machine Learning | Pros and Cons of Machine Learning, Drawbacks and Benefits. Retrieved from A Plus Topper: https://www.aplustopper.com/advantages-and-disadvantages-of-machine-learning/

[3] Applications of Machine Learning (n.d.). Retrieved from javatpoint: https://www.javatpoint.com/applications-of-machine-learning

[4] What is Customer Satisfaction? (n.d.). Retrieved from ASQ: https://asq.org/quality-resources/customer-satisfaction

[5] 4 Ways Machine Learning can Enhance Customer Experience (n.d.). Retrieved from adnovum: https://www.adnovum.com/blog/4-ways-machine-learning-can-enhance-customer-experience

[6] What is a Decision Tree (n.d.). Retrieved from IBM: https://www.ibm.com/topics/decision-trees

[7] Nagesh Singh Chauhan (n.d.). Decision Tree Algorithm, Explained. Retrieved from KDnuggets: https://www.kdnuggets.com/decision-tree-algorithm-explained.html

[8] Sruthi E R (2021, June 17). Understanding Random Forest. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[9] What is Random Forest? (n.d.). Retrieved from IBM: https://www.ibm.com/topics/random-forest

[10] Milecia McGregor (2020, July 1). SVM Machine Learning Tutorial – What is the Support Vector Machine Algorithm, Explained with Code Examples. Retrieved from freeCodeCamp: https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/

[11] How SVM Works (n.d.). Retrieved from IBM: https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works

[12] Support vector machine in Machine Learning (2020, December 20). Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/support-vector-machine-in-machine-learning/

[13] Sumit Saha (2022, July 22). XGBoost vs LightGBM: How Are They Different Retrieved from neptune.ai: https://neptune.ai/blog/xgboost-vs-lightgbm

[14] Aymane Hachcham (2022, July 21). XGBoost: Everything You Need to Know. Retrieved from neptune.ai: https://neptune.ai/blog/xgboost-everything-you-need-to-know

[15] Erlich_bachman Khandelwal (2017, June 12). Light GBM vs XGBOOST: Which algorithm takes the crown. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/

[16] How LightGBM algorithm works (n.d.). Retrieved from ArcGIS Pro: https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-lightgbm-works

[17] AdaBoost (2016, May 1). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/AdaBoost

[18] A Guide To Understanding AdaBoost (2020, February 23). Retrieved from Paperspace Blog: https://blog.paperspace.com/adaboost-optimizer/

[19] Dhanya Thailappan (2021, June 1). AdaBoost: A Brief Introduction to Ensemble learning. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/06/adaboost-a-brief-introduction-to-ensemble-learning/

[20] Sunil Ray (2017, August 14). CatBoost: A machine learning library to handle categorical (CAT) data automatically. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/08/catboost-automated-categorical-data/

[21] How training is performed. (n.d.). Retrieved from CatBoost: https://catboost.ai/en/docs/concepts/algorithm-main-stages

[22] Mohammad Yawar (n.d.). CatBoost-ML. Retrieved from Code Studio: https://www.codingninjas.com/codestudio/library/catboost-ml

[23] Paritosh Mahto (2021, June 11). Customer Satisfaction Prediction Using Machine Learning. Retrieved from Towards Data Science: https://towardsdatascience.com/customer-satisfaction-prediction-using-machine-learning-66240e032962

[24] Sachin Kumar & Mikhail Zymbler (2019, July 17). A machine learning approach to analyze customer satisfaction from airline tweets. Retrieved from SpringerOpen: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0224-1

[25] Cavalcante Siebert, L., Bianchi Filho, J.F., Silva Júnior, E.J.d., Kazumi Yamakawa, E. and Catapan, A. (2021), "Predicting customer satisfaction for distribution companies using machine learning", International Journal of Energy Sector Management, Vol. 15 No. 4, pp. 743-764. https://doi.org/10.1108/IJESM-10-2018-0007