A Study on Quantitative Analysis of Customer Satisfaction and Its Impact on Retention Using Statistical and Machine Learning Models

SADIKA ISLAM TISHA

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	Sadika Islam Tisha
Date of Birth	:	
Title	:	Quantitative Analysis of Customer Satisfaction and Its Impact on Retention Using Statistical and Machine Learning Models
Academic Session	:	

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☐	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

_____	_____
(Student's Signature)	(Supervisor's Signature)
_____	_____
Student ID	Name of Supervisor
Date:	Date:

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

THESIS DECLARATION LETTER (OPTIONAL)

Librarian,
Daffodil International University,
Daffodil Smart City,
Ashulia.Dhaka,Bangladesh

Dear Sir,

CLASSIFICATION OF THESIS AS RESTRICTED

Please be informed that the following thesis is classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are as listed below.

Author's Name
Thesis Title

Reasons (i)

(ii)

(iii)

Thank you.

Yours faithfully,

_____
(Supervisor's Signature)

Date:

Stamp:

Note: This letter should be written by the supervisor and addressed to the Librarian, *Daffodil International University* with its copy attached to the thesis.

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and, in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.

(Supervisor's Signature)

Full Name : Mr. M Khaled Sohel

Position : Assistant Professor

Date : 26/4/25

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

Sadika

(Student's Signature)

Full Name :Sadika Islam Tisha

ID Number : 221-35-872

Date : 24-11-25

A study on Quantitative Analysis of Customer Satisfaction and Its Impact on Retention Using Statistical and Machine Learning Models

SADIKA ISLAM TISHA

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Non-Major)

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

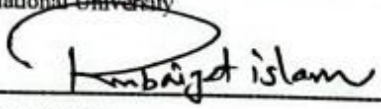
APPROVAL

This thesis titled on **"A Study on Quantitative Analysis of Customer Satisfaction and Its Impact on Retention Using Statistical and Machine Learning Models "**, submitted by **Sadika Islam Tisha (ID: 221-35-872)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

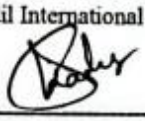
Chairman

Dr. A. H. M. Saifullah Sadi
Professor
Department of Software Engineering
Faculty of Science and Information Technology Daffodil International University



Internal Examiner 1

Dr. Rubaiyat Islam
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

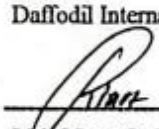


Internal Examiner 2

Dr. Md. Abdul Kader
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Nuruzzaman Faruqui
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



External Examiner

Md. Mostafiz Khan
Managing Director
Tecognize Solutions Limited

ACKNOWLEDGEMENTS

This thesis would not have been possible without the guidance, support, and patience of several people who have helped me throughout my academic journey. First and foremost, I would like to express my deepest gratitude to my thesis supervisor, **Mr. M Khaled Sohel.** Your invaluable expertise, critical feedback, and constant encouragement were instrumental in shaping this research. Your insightful questions pushed me to think deeper and refine my methodology, and your open door provided unwavering support. Finally, I could not have completed this journey without the unconditional love and support of my family. To my parents, thank you for your endless belief in me, for your patience, and for your constant encouragement, especially when the challenges seemed overwhelming. This accomplishment is as much yours, as it is mine.

DEDICATION

Dedicated to my parents.

ABSTRACT

Customer satisfaction and customer retention are essentially related to business strategy, but it is usually assumed without immediate quantitative proved. Subsequently, this problem is addressed in this research, which quantifies the effects of customer satisfaction and retention by employing a hybrid analysis. The present research makes use of a publicly available dataset on e-commerce offered by Kaggle and comprising 3,900 customer records. Preprocessing of the data involved imputing the missing values and also engineering the main variables; Retention (based on Subscription status) and satisfaction binary (based on binned Review rating scores). Two-way methodology was used. Statistically, firstly, a logistic regression model was applied. Because of this model, it was found that Retention is never significantly predicted by satisfaction binary (p-value = 0.631), which requires one to reject the primary hypothesis. On the contrary, the model established Gender as a very important factor that contributes to retention ($p < 0.001$) and Discount value as a slightly important factor ($p = 0.056$). Second, prediction and confirmation of a classification model was conducted using a random forest classification model. To verify the statistical results, the importance of measuring the model features, an analysis revealed that satisfaction binary was among the least significant features. The model was found to have Purchase Amount, Age, and Gender with the most significant predictors of retention. The conclusion of this thesis is that the general opinion that satisfaction can be the reason why a person retains is not supported in this dataset. Experiential factors (satisfaction) do not influence retention, but the demographic factor (Gender) and the financial factor (Purchase Amount, Discount value) factor. This is an indication that the business approaches of retention improvement should be directed at demographics-related segmentation and financial stimulation instead of being completely determined by maximizing customer satisfaction metrics.

Keyword: Customer Retention, Customer Satisfaction, Machine Learning, Logistic Regression, Random Forest, Feature Importance

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background and Context

Customer retention has become a key to the sustainable business success in the hyper-competitive world of modern e-commerce. The universally acknowledged marketing and business strategy law is that the retention of an existing customer is much more economical as compared to acquisition of a new customer. At the core of this retention philosophy, it is important to have the notion of customer satisfaction.

The popular business logic seems to have the direct, positive, and influential correlation: a satisfied customer is a loyal one. It is based on this assumption that massive corporate investment will be focused on improving the customer experience, user interface optimization, and quality improvement of the services. The perceived outcome of such efforts is customer satisfaction ratings (through reviews), and these efforts work on the supposition that it will directly and objectively increase retention, commonly as it relates to return-buy or repeat subscription purchases.

1.2 The Research Problem and Hypothesis

Although the relationship between satisfaction and retention has now gained broad acceptance, even quantification of the effect is an intricate and even situation-dependent issue. Most organizations are guided by this alleged correlation the actual extent of which is neither known nor clearly understood through data. This poses a severe business issue: Are the resources being utilized effectively?

Is the customer satisfaction the key factor influencing the choice of a customer to stay or is the effect indirect to other stronger influences that include price, financial incentive (discounts) or fundamental demographic characteristics? When the company spends millions of dollars on satisfaction improvement, and the retention rate does not

increase, then it may be using the incorrect lever. This work is a challenge to this issue. Its main aim is to shift to a narrow assumption to a narrow and quantitative analysis. This thesis will be informed using the following central hypothesis:

- **Hypothesis (H1):** Higher customer satisfaction, as measured by customer review ratings, has a direct and statistically significant positive impact on customer retention, as measured by subscription status.

1.3 Aims and Objectives

The main focus of the given research is to prove the given hypothesis and define the actual forces of customer retention in the given dataset. The specific objectives are:

i. To statistically measure the association between customer satisfaction and customer retention, and with logistic regression model.
ii. To create a predictive machine learning model (Random Forest) to keep their customers and test its precision in classifying their customers.
iii. To establish the driving forces behind the retention by comparing the predictive power (of the machine learning model) to the statistical significance (of the regression model) between these two factors.
iv. To compare the effect of customer satisfaction with other variables (e.g. demographics, financial incentives) to conclude on its relative significance in a multivariate model.

1.4 Significance of the Study

There are some practical, immediate visuals of the business strategy through the findings of this research. With scarce resources, the businesses have to make rational decisions based on the data to choose the place to invest to achieve the highest level of impact concerning retention. The given work gives a clear picture in answering a crucial question: Should retention programs be aimed at customer experience improvement (enhancing customer satisfaction), or some other, more powerful driver (such as pricing strategies, discounting program, or demographic-based operations). The findings of this thesis will help set more efficient and effective resource allocation towards customer

retention programs by empirically testing one of the basic, yet usually unchecked, business assumptions.

1.5 Chapter Outline

i. **Chapter 1: Introduction** provides the background, research problem, hypothesis, and aims of the study.

ii. **Chapter 2: Data and Methodology** details the dataset, the extensive data preprocessing and feature engineering steps, and the analytical setup for both the logistic regression and Random Forest models.

iii. **Chapter 3: Results and Analysis** present the complete findings, including the Exploratory Data Analysis (EDA), the results from the statistical (logistic regression) models, and the performance and feature importance from the machine learning (Random Forest) models.

iv. **Chapter 4: Discussion** interprets the findings presented in Chapter 3, discusses the (lack of) support for the initial hypothesis, and explores the significant and unexpected drivers that were uncovered.

v. **Chapter 5: Conclusion** summarizes the entire study, reflects on its key findings, discusses the practical implications for the business, acknowledges the limitations of the research, and suggests directions for future work.

1.6 Summary

This chapter has defined the context as well as the rationale of the study. It has defined the common business premise, customer satisfaction is equal to customer retention, and has come up with a definite testable hypothesis to question the premise. In its introduction, the actual objectives, the twofold methodology of analysis, and overall format of the thesis have also been elaborated. This hypothesis will now be tested in the following chapters to determine the actual underlying factors of customer retention in the dataset.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Customer satisfaction and customer retention have being one of the foundations of business strategy and marketing theory over the decades. The empirical and intuitively-supported idea that highly satisfied customers tend to become loyal and retained customers has been upheld by a large and overwhelming amount of literature. This loyalty, in its turn, is directly related to long-term profitability, decreased marketing expenses, and sustainable development of the business.

This chapter summarises all available studies into three imperative areas that constitute the basis of this thesis. It provides a review of the classical and modern theories which determine the causal relationship between customer satisfaction and customer retention in the first place. Second, it also looks at the development of the methodologies of measurement of these two abstract concepts shifting not only the traditional surveys but also the modern data-driven proxy.

Last and most importantly, this review examines how statistical and machine learning models are currently applied in this area. Although most studies have affirmed that there is a correlation, this section will explore how researchers have tried to measure it. This review also seeks to determine the gaps in current literature especially regarding the use of hybrid statistical (explanatory) and machine learning (predictive) models and this way establish the novel methodological contribution of this thesis.

2.2 Existing Study

Zhang et al. (2022) used the given prediction model in the investigation of the e-commerce customer retention to define the possible repeat buyers. The authors chose to tackle the critical issue of the low repurchase rates and suggested a multistage strategy

with reference to the real-world information provided by Tmall. The combination of extensive feature engineering, in which 147 features were constructed manually using user, merchant, and interaction portraits, and a severe class imbalance solution, which is the Synthetic Minority Oversampling Technique (SMOTE), was a large contribution to their methodology. As an attempt to predict buyer behavior, the authors ran classical models against ensemble methods, finally adopting a two-layer Stacking fusion model (with LightGBM, XGBoost and Random Forest as basic learners) which provided the most accurate predictive power (AUC 0.68406). The experiment proves a fused ensemble model supplemented by careful feature engineering and data imbalance pre-processing to be more accurate in recovery potential repeat customers as opposed to single model models.

Davoodi and Mezei (2022) comparative study chose to perform an evaluation of the performance of various machine learning models applied to e-commerce customer review sentiment analysis. The authors wanted to find out the effectiveness of modern transformer-based models (BERT and RoBERTa) in comparison with traditional models (Naive Bayes and Support Vector Machines). They used a manually annotated dataset of 3,500 reviews on Trustpilot to categorise the sentiment as positive, negative, or mixed. Their results showed that transformer models were much better and RoBERTa had the highest performance at more than 98% accuracy. One of the most valuable additions of the research was the point on the unreliability of using the star rating provided by the user, as a simple proxy of the sentiment, which often differs with the sentiment relayed by the text of the review. The paper finds that the contemporary transformer models form an almost perfect solution to sentiment classification in this field.

Barzizza et al. (2024) suggested an innovative, data-driven approach to the determination of the causes of customer satisfaction, especially when developing new products. Their machine learning-based method converts customer survey data (i.e. Likert scales) in a binary classification problem through a scoring system of "Top Box" (TB) vs. Other. The framework is used to give priority to the influence of particular product features because it compares and selects an optimal ML model to predict the top box satisfaction. One of the most important contributions of their approach is the localization of a space of improvement measure, one of the forecasts of how much in

general customer satisfaction would be obtained were a single aspect to be perfected. This is a well-presented methodology that when used with an earphone case example gives a clear analysis of how the R&D and marketing teams can give priority to improvements and gap analysis when comparing with those of competitors.

Ruder et al. (2016) tackled one of the frequent weaknesses of aspect-based sentiment analysis (ABSA) that review sentences are evaluated independently, without the consideration of the overall structure of the review. The authors postulated that sentences that make up a review are mutually reliant and such a context plays a key role in classification. They suggested hierarchical bidirectional LSTM (H-LSTM) model which performs at sentence (processing words) and review (processing sentences as a sequence) levels. It is hierarchical to enable this model to represent the argumentative flow and the sentential context of the whole review. Their H-LSTM generalized well, and in SemEval-2016 experiments showed higher accuracy than non-hierarchical and more basic alternatives (C CNN and a regular LSTM), as well as state-of-the-art results with a range of multilingual, multidomains datasets. One of its main contributions was that this had been accomplished without any hand-engineered aspects or without external lexicons and this shows how useful it is to model document structure to the sentiment task.

The article by Chebolu et al. (2023) has offered a survey of 98 publicly available datasets on Aspect-Based Sentiment Analysis (ABSA), with more than 25 domains and 21 languages. The major contribution of the study is the critical review of the existing corpora dedicated to training and evaluation of systems of ABSA. Authors listed different types of ABSA subtask (e.g., Aspect Term Extraction, Aspect Category Sentiment Analysis) and annotated existing datasets onto them. They have found considerable drawbacks in the existing situation, such as the spread of fragmented and small datasets, no standardisation of annotation types and excessive emphasis on sentence-level analysis, which does not reliably represent the entire context of a review. The paper proposes the combination of current datasets, manual annotation of the opinion phrases, the construction of larger review-level corpora and the increase of datasets to low-resource languages and new areas, beyond customer reviews.

Shah et al. (2023) reviewed the extensive literature on the use of Natural Language Processing (NLP) and Machine Learning (ML) in the process of contact center automation. The authors suggested that the COVID-19 pandemic made more urgent the necessity of corresponding innovative approaches to the field of solutions more advanced than traditional Interactive Voice Response (IVR) systems. The paper systematically reviewed the existing literature and listed the main applications of NLP to this area including: (1) customer sentiment/satisfaction analysis; (2) smart call routing; (3) customer-agent interaction optimization as a result of data analysis; and (4) the creation of customer service chatbots. One of the key discoveries of their review was that the biggest challenge and a gap in the research is the lack of large, high-quality, publicly available labeled datasets which is also complicated by the data privacy policies, and the low technical quality of call audio. The research ended with a recommendation on how these data problems can be solved in order to promote automation and efficiency of customer service.

As a method of getting finer details of customer satisfaction, Davoodi et al. (2025) suggested an aspect-based sentiment analysis (ABSA) system to derive insights on customer satisfaction in e-commerce reviews. To complete the aspect-specific feedback, the authors provided a special dataset by compiling the Trustpilot feedback on five large online stores (Zalando, Wish, etc.). They singled out 14 critical dimensions of the e-commerce experience (e.g., "Shipping," "Item quality," "Refund process" etc.) using a literature review and analyzing 3500 reviews. This was then manually annotated with the positive or negative sentiment to each aspect that was discussed in a review. The paper has compared the performance of a number of machine learning models in aspect-based sentiment classification and discovered that RoBERTa was the most performing model with an accuracy of more than 90%. The importance of the authors in the research lies not only in the development of this rich, manually annotated, multi-aspect dataset, but also in the fact that a high-accuracy model such as RoBERTa can give actionable suggestions to businesses, which reveals a significant difference between the perceived sentiment in the form of the star rating and the text.

Noori (2021) has made a comparative analysis of different machine learning algorithms in the task of classifying customer reviews. The participants involved in the

study are the 50,000 reviews of Amazon products, and they were dedicated to the influence of text preprocessing and feature extraction strategies, using TF-IDF, on the success of the models. Some of the classifiers that were tested in the research include Logistic Regression, Naive Bayes, Support Vector Machine (SVM) and the random forest. The most important was that the Logistic Regression model was repeatedly the best as it demonstrated the most accurate prediction (90.2) in the determination of the sentiment of the reviews. As mentioned in the research, it is essential to note that the Logistic Regression is a rather strong and effective baseline in binary sentiment classification exercises on large review datasets.

Lim, S. L., Foo, L. K., and Chua, S. L. (2023) made a comparative analysis of machine learning and deep learning methods in sentiment analysis of electronic product reviews on Amazon and BestBuy. The two studies examined, in particular, the effect of various sampling methods to train the training data and the influence of hyper-parameter optimization on the model performance. The analysis was made between a tuned machine learning and a deep learning model and a baseline. Their experimental results revealed that both of their tuned models performed better than the baseline in terms of accuracy and the F1-score indicating that given the correct data sampling and hyper-parameter optimization, both the deep learning and the traditional machine learning methods can be useful in capturing customer sentiment through reading product reviews.

Seymen et al. (2023) performed a comparative evaluation of Customer churn prediction task in terms of Ordinary Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN) in terms of performance. The objective of the study was to establish the relational effectiveness of a typical neural network in comparison to a deep learning arrangement (CNN) concerning this particular classification issue. The authors used a customer churn prediction dataset and discovered that, although both algorithms were highly successful, the CNN algorithm was proven to be more successful and, in comparison with the traditional ANN, which makes the CNNs appear as a viable method to use in this prediction activity.

2.3 The Research Gap

The studies consider the retention prediction and satisfaction analysis as independent parallel operations. Not a multitude of studies examined in this case, but the few that do so do not seem to directly quantify the effect of a measure of satisfaction (either a star rating or an anticipated sentiment score) upon a retention model. This thesis attempts to close this gap by bridging these two streams of research. It enquires explicitly the interrelationship between satisfaction and retention by posing the following question: Does satisfaction contribute to retention? It also utilizes a hybrid approach, where it is not only predictive ML models are used (as it is in Zhang et al., 2022, and Noori, 2021) but also a statistical inference model (Logistic Regression) is used to explicitly measure this correlation.

2.4 Conclusion

The literature is very explicit with two prongs in analyzing customer feedback and customer behavior. The former research possibility (Zhang et al., 2022; Seymen et al., 2023) is on retention/churn prediction. These works discuss retention as a pure classification problem, exploiting the benefit of the latest methods such as model stacking, deep learning (CNNs), and imbalance correction (SMOTE) to achieve a high level of predictive accuracy. The second stream of research, which is stronger (Davoodi and Mezei, 2022; Barzizza et al., 2024; Ruder et al., 2016; Noori, 2021), is concerned with the analysis of satisfaction. This work is mostly focused on sentiment analysis on the textual basis, where traditional models of this process (Logistic Regression, SVM) are opposed to modern deep learning models (LSTMs, BERT, RoBERTa) in order to select the most successful one in terms of classifying customer attitudes. The most crucial observation made by the second stream (Davoodi and Mezei, 2022; Davoodi et al., 2025) is the direct lack of reliability of user-submitted star ratings to reflect actual satisfaction. This observation is very pertinent since it offers a possible a priori reason why a model built on this kind of ratings could not succeed. Moreover, the recurrent problem of insufficient data is emphasized in the reviews of literature on the methodological front (Chebolu et al., 2023; Shah et al., 2023) and, consequently, the need to apply effective methods to whichever data is accessible appears to be a recurring problem.

CHAPTER 3

DATA AND METHODOLOGY

3.1 Introduction

This chapter explains the analysis model of testing the hypothesis presented in Chapter 1. It explains the data, the many beforehand processes of preparing the data that must be done to support the analysis process, and the application of the dual (statistical and machine learning) methodologies. The main objectives of this stage are to make the data clean, robust, and structured in a proper way to both infer a statistic and make a predictive model.

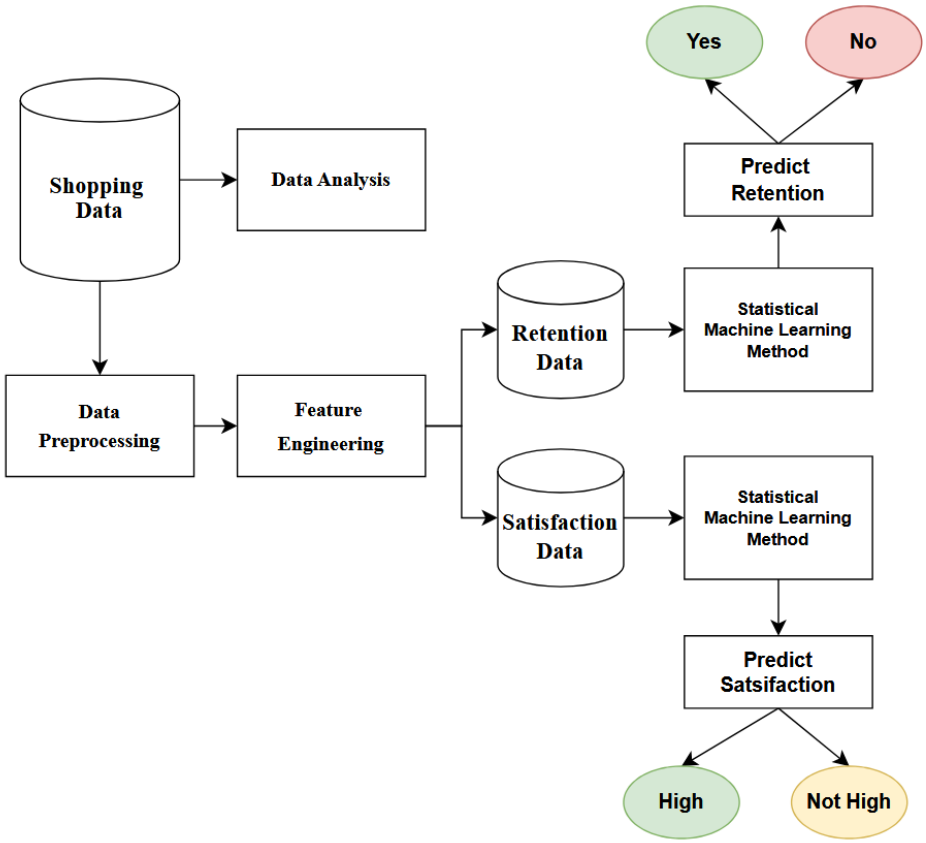


Figure 3.1: Methodology of our study.

3.2 Dataset Description

The data in this paper has been obtained with the help of the Customer Purchase Data publicly available on Kaggle (Insight Programme, 2023). This data has 3, 900 entries of customer transactions of a fictive e-commerce site. All records consist of a blend of customer demographic, transactional and post purchase feedback attributes:

- **Customer Demographics:** Age, Gender
- **Transaction Details:** Item Purchased, Category, Purchase Amount, Payment Method
- **Logistical Information:** Country, Location, Shipping Type
- **Marketing Information:** Promo Code Used, Discount value
- **Customer Feedback:** Review Rating
- **Retention Indicators:** Subscription Status, Previous Purchases

This richness of data will give all the variables needed to formulate and model the two satisfaction and retention models.

3.3 Data Pre-processing

The raw data were not in their final form, but they needed numerous processing procedures to address the absence of values, inaccurate records as well as the wrong type of data. A combination of these steps was brought together into one repeatable function to provide some consistency.

- Gender Standardization: The Gender feature contained inconsistent values ('Female', 'F', 'Male', 'M') and 195 missing entries. All values were standardized to 'F' and 'M'. The missing values were imputed using the majority class ('F'), a common method for handling missing categorical data.
- Discount Value Imputation: The Discount value feature had 1,450 missing entries. These were imputed using the dataset's median discount value. The median was chosen over the mean to avoid skewing the data, as discount values are often not normally distributed.

- Handling 'Previous Purchases': The Previous Purchases column contained a mix of numeric strings and the text 'nothing'. The 'nothing' entries were converted to the integer 0, and the entire column was cast to a numeric type to be used in analysis.
- Handling 'Review Rating': The Review Rating column, which serves as the basis for our satisfaction metric, was converted to a numeric type. Any rows where this conversion failed or was missing were dropped to ensure the integrity of the analysis.

3.4 Feature Engineering

Once the pre-processing was done, two key features were designed that would both be used as the major independent and dependent variables in the hypothesis test.

3.4.1 Dependent Variable (Y): Retention

- To test the hypothesis, a clear, binary measure of retention was required. The Subscription Status column ('Yes'/'No') was identified as the most direct and reliable indicator of active customer retention.
- A new column named Retention was created by mapping 'Yes' to 1 and 'No' to 0. This binary variable (0 or 1) serves as the dependent variable (y) for all subsequent retention models.

3.4.2 Independent Variable (X): satisfaction binary

- The Review Rating (a continuous scale from 2.5 to 5.0) served as the basis for the customer satisfaction variable.
- Initial analysis (as detailed in Chapter 3) revealed a severe class imbalance. The original "Low" satisfaction group (ratings 2.5) represented only 1.7% of the dataset, making it statistically too small to be modelled as a separate category.
- In order to fix this, a binary satisfaction variable, satisfaction binary, was developed. The ratings were divided into two categories, namely: 'Not High' (ratings 0-3.5) and 'High' (ratings 3.6-5.0). This binning formed two well balanced classes thereby making the analysis robust and stable. The hypothesis is going to be tested using this variable as the main independent variable (X).

3.5 Analytical Methodology

This is two-part research that uses the methodology outlined in Chapter 1. The implementation of each of the models technically is described below.

3.5.1 Part 1: Statistical Inference Model (Logistic Regression)

A logistic regression model was done to explain the relationship between satisfaction and retention. This approach has been selected as the dependent variable Retention is binary (0 or 1). The analysis was done in two phases:

i. **Bivariate Model**: A simple logistic regression was run with satisfaction_binary as the only independent variable to test the direct, isolated relationship on Retention.
ii. **Multivariate Model**: A full logistic regression model was built using all relevant features (e.g., Age, Gender, Discount value, satisfaction_binary, Category). This model's purpose is to test the hypothesis in a real-world context, determining if satisfaction_binary has any significant impact after controlling for all other factors.

In both models, the P>|z| value (p-value) for each coefficient is the primary metric. A p-value less than 0.05 is considered statistically significant, indicating that the variable has a real effect on retention.

3.5.2 Part 2: Predictive Model (Random Forest)

In order to anticipate retention, and validate the statistical results, the Random Forest classifier was constructed. The aim of such a model is not to explain, but to have the best predictive accuracy possible.

i. **Feature Selection**: All features from the dataset were used, except for Review Rating (which would cause data leakage, as it's the source of satisfaction_binary) and other identifiers.
ii. **Data Splitting**: The data was split into a training set (80%) and a test set (20%) to evaluate the model's performance on unseen data.

iii. **Preprocessing Pipeline**: A ColumnTransformer was used to apply StandardScaler to all numerical features and OrdinalEncoder to all categorical features.

iv. **Handling Class Imbalance (SMOTENC):** The variable of Retention is lopsided (73 percentage of NO and 27 percent Yes). To ensure that the model does not just get trained to make predictions based on the majority class, SMOTENC (Synthetic Minority Over-sampling Technique of Nominal and Categorical data) was used. Only the training data (included in an ImbPipeline) underwent this technique with the aim of generating synthetic representatives of the minority 'Retained' data type, which balanced the training set.

v. **Model Training**: A RandomForestClassifier was instilled with the class weight value of balanced. This environment also teaches the model to be extra vigilant to the minority group in the training process.

vi. **Evaluation:** The model's performance was evaluated on the untouched test set using a classification_report. The primary metric for confirmation of the hypothesis is the feature_importance_ plot, which shows which variables the model found most useful for making its predictions.

3.6 Summary

The whole methodology of the thesis has been outlined in this chapter. It has outlined how the raw shoppingdata.csv was changed into an analytical and ready-to-process dataset using preprocessing and feature engineering. The process of developing the important Retention and satisfactionbinary variables was described. Lastly, the technical configuration of both logistic regression (inference) and Random Forest (prediction) model was also described, which preconditions the analysis and findings within the next chapter.

CHAPTER 4

RESULTS AND ANALYSIS

4.1 Introduction

In this chapter, the entire results of the analyses in Chapter 3 are provided. The chapter has been separated into three parts based on the dual methodology. It will first give the results of Exploratory Data Analysis (EDA) that will give the first visual evaluation of the data and relations between important variables. Second, it provides the facts of inferences made in the statistical (logistic regression) models, the p-values, odds rations to test the thesis hypothesis formally. Third, it shows the performance of the predictive (Random Forest) models, classification performance, and, most importantly, feature ranks in their importance. The two-sided display enables a complete evaluation of the hypothesis in an explanatory and a predictive approach.

4.2 Exploratory Data Analysis (EDA)

A visualization of the data was done before modeling to see its properties. The most important findings are mentioned below.

4.2.1 Target and Predictor Variable Distributions

Retention variable, which is engineered of Subscription Status, is unbalanced. According to Figure 4.1, about 73 percent of the customers in the dataset fall in the Not Subscribed segment whereas 27 percent fall in the Subscribed (Retained) category. The machine learning approach to this imbalance was to apply SMOTENC and class weighting.

Review Rating was the main independent variable which produced the primary independent variable, satisfaction binary. The binned variable is fairly balanced, as 58%

of clients are rated as being of High satisfaction and 42% of the customers are rated as Not High.

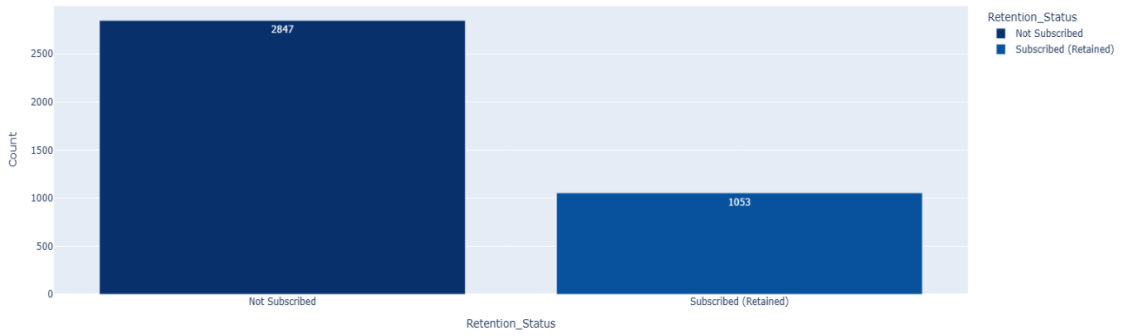


Figure 4.1: Bar chart showing the distribution of 'Subscribed (Retained)' vs. 'Not Subscribed'

4.2.2 Visualizing the Hypothesis and Key Drivers

In order to have a preliminary understanding of the relationships, the mean retention rate was plotted in relation to the significant variables.

- **Satisfaction vs. Retention (The Null Hypothesis):** The first big piece of evidence in terms of the thesis hypothesis is availed by figure 4.2. The customer retention rate of those who were not highly satisfied is average of 26.6 and that of highly satisfied customers is 27.2. The variation is insignificant to give an adequate visual cue that satisfaction does not correlate considerably with retention.
- **Gender vs. Retention (A Key Driver):** In stark contrast, Figure 3.3 shows a pronounced difference in retention rates based on Gender. The average retention rate for 'F' (Female) is 30.1%, while for 'M' (Male) it is 17.6%. This suggests Gender is a far more powerful explanatory variable than satisfaction.
- **Other Factors:** The EDA also revealed different distributions for Age and Purchase Amount (USD) when compared by retention status, suggesting they could be valuable for the predictive machine learning models.

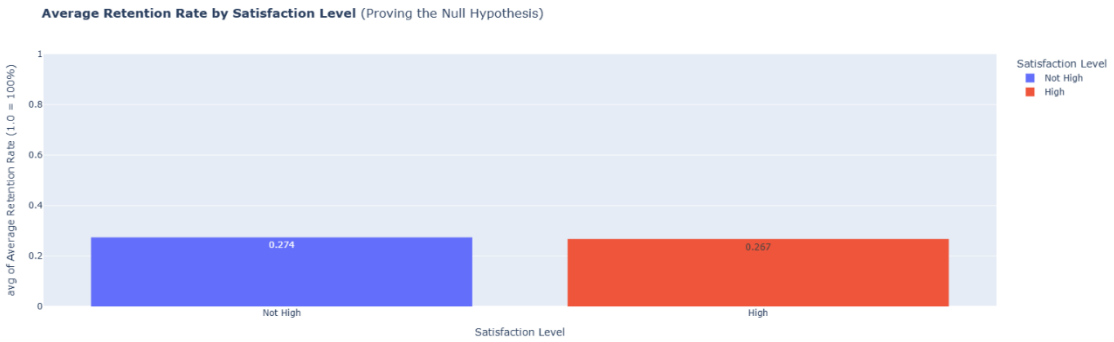


Figure 4.2: Bar chart of Average Retention Rate by Satisfaction Level, showing 'Not High' at 0.267 and 'High' at 0.274

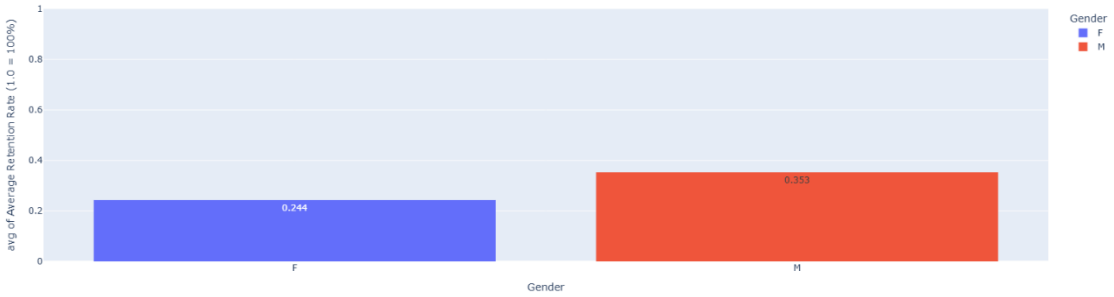


Figure 4.3: Bar chart of Average Retention Rate by Gender, showing 'F' at 0.244 and 'M' at 0.353

4.3 Part 1: Statistical Inference (Logistic Regression Results)

In order to test the hypothesis properly, statistical inference models were constructed.

4.3.1 Bivariate Analysis: Satisfaction -> Retention

To begin with, a basic logistic regression was performed where only the satisfaction binary noted as a predictor of Retention was used. It was aimed at testing the direct, independent effect of satisfaction.

The p-value of satisfaction-binary (P>|human|) is 0.631 as indicated in the output. This is significantly greater than 0.05 which is the significance level. This finding shows that the relations between satisfaction and retention when measured independent of each other are not statistically significant.

4.3.2 Multivariate Analysis: [All Variables] -> Retention

Then, the complete multivariate logistic regression model was applied to determine whether the satisfaction is noteworthy in the case when all the other variables are considered (e.g. Age, Gender, Discount value). The Table 4.1 provides the results.

Table 4.1: Multivariate Logistic Regression Results for Retention

Variable	Odds Ratio	P-Value (P>|z|)	Significant
num__Age	1.007029	0.847	No
num__Purchase Amount	0.977674	0.534	No
num__Discount value	1.103793	0.056	**Yes (Marginal)**
cat__Gender	1.714373	0.000	**Yes (High)**
cat__Category	1.009475	0.816	No
cat__Size	0.944151	0.144	No
cat__Season	0.960765	0.219	No
cat__Shipping Type	1.012505	0.561	No
cat__Payment Method	1.026591	0.163	No
cat__satisfaction_binary	0.960646	0.584	No

4.3.3 Interpretation of Statistical Findings

The statistical analysis provides two definitive conclusions:

i. **Hypothesis Not Supported:** The primary hypothesis is rejected. The p-value for cat_satisfaction_binary is 0.584, confirming it has no statistically significant effect on retention, even when controlling for all other factors.

ii. **Key Drivers Identified:** The model identified two significant drivers of retention.

- **cat_Gender** (**p < 0.001**): This is the single most significant driver. The odds ratio of 1.714 indicates that one gender (likely 'F', based on the EDA) is 71.4% more likely to be retained than the other, holding all else constant.

- **num_Discount value** (**p = 0.056**): This variable is marginally significant. The odds ratio of 1.104 suggests that for every one-unit increase in the standardized discount, the odds of retention increase by 10.4%.

4.4 Part 2: Predictive Modeling (Random Forest Results)

The second part of the analysis used a Random Forest machine learning model to predict retention and confirm the statistical findings from a non-linear, predictive standpoint.

4.4.1 Retention Model Performance

The model was trained on 80% of the data (using SMOTENC) and evaluated on the 20% test set. The overall accuracy was 63%. The detailed classification report (Table 3.2) shows that while the model is good at predicting the majority class ('Not Retained'), it struggles significantly with the minority class ('Retained').

Table 4.2: Classification Report for Retention Model

Class	Precision	Recall	F1-Score	Support
0 (Not Retained)	**0.73**	**0.79**	**0.76**	**569**
1 (Retained)	**0.26**	**0.19**	**0.22**	**211**
Accuracy			**0.63**	**780**
Macro Avg	**0.49**	**0.49**	**0.49**	**780**
Weighted Avg	**0.60**	**0.63**	**0.61**	**780**

This low F1-Score (0.22) for the 'Retained' class indicates that predicting retention is an inherently difficult task with the given data, as the factors separating the two classes are not overwhelmingly clear.

4.4.2 Retention Model Feature Importance (The Confirmation)

The most crucial output from this model is the feature importance plot, which ranks the variables by how useful they were to the model in making accurate predictions.

Figure 3.4: Feature Importance for Retention Prediction Model

The plot reveals two key findings:

i. **Hypothesis Confirmation:** cat_satisfaction_binary is ranked 9th out of 11 features, with a very low importance score (approx. 0.04). This confirms the logistic regression finding: satisfaction is not a useful or important driver of retention.

ii. **Key Predictive Drivers:** The model's top predictors are num_Purchase Amount (USD) (0.24) and num_Age (0.22). These are followed by cat_Shipping Type (0.10) and cat_Season (0.08).

A large discrepancy was realized: cat_Gender, the most important statistical force, is among the least important predictive features (ranked 10 th). This implies that Gender is statistically related, but it is not as practical in splitting logic of its tree-based model as the continuous variable such as Age and Purchase Amount.

4.4.3 Satisfaction Model Performance

Finally, the separate model to predict satisfaction itself was trained. The results (Table 3.3) show an overall accuracy of 54%, which is only slightly better than a 50/50 random guess.

Table 4.3: Classification Report for Satisfaction Model

Class	Precision	Recall	F1-Score	Support
Not High	0.45	0.44	0.45	329
High	0.60	0.61	0.61	451
Accuracy			**0.54**	**780**

This result indicates that the available features are poor predictors of a customer's satisfaction level.

4.5 Summary

This chapter presented the complete results of the analysis. The findings were consistent and clear across all models:

i. **Hypothesis Rejected:** The primary hypothesis was definitively not supported. The EDA (Figure 4.2) showed no visual correlation. The logistic regression confirmed this with a high p-value (0.584). The Random Forest model confirmed it again, ranking satisfaction as one of the least important features.

ii. **Statistical vs. Predictive Drivers:** A key finding emerged in the difference between the two models.

    1. The **statistical (inference) model** identified **Gender** ($p < 0.001$) as the only significant driver of retention.
    2. The **predictive (ML) model** identified **Purchase Amount (USD)** and **Age** as the most *useful* features for prediction.

iii. **Model Performance:** Both predictive models showed that the given features are weak predictors. The retention model struggled to identify 'Retained' customers (22% F1-score), and the satisfaction model was only 54% accurate.

These results—the rejection of the primary hypothesis and the discrepancy between the statistical and predictive drivers—will be interpreted in the following chapter.

CHAPTER 5

DISCUSSION

5.1 Introduction

This chapter just interprets the findings that were shown in Chapter 4 in the broader context of the study. There were three vital findings in the chapter preceding:

i. The primary hypothesis that customer satisfaction drives retention was definitively rejected by all analytical methods.
ii. A significant discrepancy emerged between the drivers identified by the statistical (inference) model and the predictive (machine learning) model.
iii. The overall performance of both predictive models was low, indicating that the available data provides weak predictive power.

This chapter will now discuss the implications of these findings, focusing on *why* these results may have occurred and what they mean for the business.

5.2 Discussion of Primary Finding: The Rejection of the Hypothesis

The focal and the most unexpected result of this work is the presence of the obvious absence of the statistically significant link between the customer satisfaction expressed through the Review Rating and customer retention through Subscription Status. The p-value on cat_satisfaction_binary was high (0.584) and the feature weight in the Random Forest model was minimal (ranked 9 th out of 11).

This null finding is not a failure of the analysis but an important finding that directly questions the assumption in the study. This can be interpreted in several ways:

- **Satisfaction as a "Hygiene Factor":** It is possible that satisfaction operates as a "hygiene factor," not a motivator. In this model, its presence (a high rating) does not motivate a customer to stay, but its absence (a hypothetical very low

rating, which was rare in the data) might have motivated them to leave. Because the data was skewed towards moderate-to-high ratings, this "punishment" effect is not visible.

- **The Wrong Proxy for Satisfaction**: Review Rating may be a poor proxy for the customer's true, overarching satisfaction. The rating might be tied to a specific product ("the sweater was great") rather than the overall service ("I love shopping here"). A customer can be satisfied with a product but have no intention of subscribing to the service.
- **Transactional vs. Relational Loyalty:** The model suggests retention for this company is not driven by relational loyalty (emotional connection, satisfaction) but by other factors. The low predictive power of all available features suggests the decision to subscribe may be driven by external variables not captured in the dataset, such as the perceived necessity of the subscription or the presence of a better competitor.

5.3 Interpreting the True Drivers of Retention

The multivariate analysis did not just reject the hypothesis; it also identified the variables that do matter.

5.3.1 The Statistical Driver: Gender

The logistic regression model identified Gender as the only highly significant predictor ($p < 0.001$), with an odds ratio of 1.714. Based on the EDA (Figure 3.3), which showed a 30.1% retention rate for 'F' vs. 17.6% for 'M', this finding suggests that female customers are 71.4% more likely to be subscribers than male customers, holding all other factors constant.

This is a powerful explanatory finding. It implies a fundamental difference in purchasing behavior or perceived value of the subscription between these two demographic groups. The subscription service may be inherently more appealing to, or marketed more effectively towards, female customers.

5.3.2 The Marginal Driver: Discount value

Discount value was marginally significant ($p = 0.056$). This finding is intuitive: financial incentives matter. The odds ratio of 1.104 suggests that customers who use larger discounts are slightly more likely to subscribe. This may indicate that price-sensitive customers are drawn into the subscription, or that the subscription itself offers discount-related perks that attract these users.

5.4 Statistical Significance vs. Predictive Importance: A Key Discrepancy

One of the most nuanced findings of this study is the discrepancy between the statistical and predictive models:

i. **Inference Model (Logistic):** Gender was the most significant driver. Age and Purchase Amount were not significant.

ii. **Prediction Model (Random Forest):** Age and Purchase Amount were the most important predictors. Gender was one of the least important.

This is not a contradiction; it highlights the different goals of the two models.

- **Why Gender was Significant but Not Predictive:** Logistic regression measures the consistency and statistical significance of a relationship across the entire dataset. The 71.4% increase in odds for 'Gender' is a consistent and statistically "real" effect. However, for a Random Forest model, Gender is a poor predictor because it's a binary variable that splits the data only once. It cannot create the complex, multi-level decision rules that continuous variables like Age and Purchase Amount can.

- **Why Age and Purchase Amount were Predictive but Not Significant:** These continuous variables are excellent for a Random Forest, which can find many different "split points" (e.g., "if Age > 45" or "if Purchase Amount < 30") to isolate small groups of customers. This makes them *useful for prediction*. However, in the logistic regression, their effect was not statistically significant. This means that while Age might be a useful predictor in a non-linear model, its overall (linear) effect on retention across the whole population isn't consistent enough to be statistically significant.

This demonstrates a key principle: A variable's explanatory power (statistical significance) is not the same as its predictive utility.

5.5 Discussion of Model Performance

A final, crucial finding is that all models performed poorly. The retention model's F1-score for the "Retained" class was only 0.22, and the satisfaction model's accuracy was 54%.

This indicates that the most important drivers of both satisfaction and retention are not present in this dataset. The available features (Age, Gender, Category, etc.) are simply weak predictors. The decision to subscribe or to be satisfied is likely driven by external factors (e.g., product quality, shipping speed, competitor pricing, customer service interactions, website usability) for which we have no data.

5.6 Summary

This chapter has streamlined the findings of Chapter 4, which revealed that the main hypothesis is not accepted. It is evident that the customer satisfaction as it is measured does not impact any significant customer retention. Rather, Gender and marginally Discount value are the major statistically significant explanatory variables of retention. Moreover, one major difference was discovered between the explanatory and predictive models: the most notable statistical driving factor (Gender) was not the most predictive characteristic. The significant features used in prediction (Age and Purchase Amount) were not significant. Lastly, the underperformance of all modeling programs is a clear indication that the most pivotal data required in finding the causes of retention and satisfaction must have been omitted in the data source. The conclusions will be explained in the concluding chapter.

CHAPTER 6

CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

This conclusion is the summation of the previous chapters on findings and discussions. It will start by giving an overview in a high level summary of the whole study, including the original hypothesis to the analytical results. It then makes the conclusive findings of this research.

Continued on these conclusions, this chapter offers a list of detailed business recommendations due to the findings contained in the data. Lastly, it concludes the study by discussing the weaknesses of such a study, and providing direction on future research that may offer the missing details in getting a complete picture of customer behaviour in this business setting.

6.2 Summary of the Study

This research aimed to calculate the effectiveness of customer satisfaction on retaining customers in a given e-commerce platform. The working hypothesis stipulated that Retention as assessed by Subscription Status would be a strong predictor of higher levels of satisfaction as assessed by Review Rating.

A dual methodology was employed to test this:

i. **Statistical Inference:** A logistic regression model was used to *explain* the drivers of retention and test the statistical significance of satisfaction.
ii. **Machine Learning:** A Random Forest model was used to *predict* retention and to confirm the inferential findings via feature importance.

The analysis led to three clear and consistent findings:

i. **Hypothesis Rejected:** Customer satisfaction was found to have no statistically significant impact on customer retention (p-value = 0.584) and was ranked as one of the least important features in the predictive model.

ii. **Key Drivers Identified:** The statistical model identified Gender ($p < 0.001$) as the only highly significant driver of retention. The predictive model identified Purchase Amount (USD) and Age as the most useful features for prediction.

iii. **Weak Predictive Power:** All predictive models demonstrated low performance (e.g., 22% F1-score for the "Retained" class), indicating that the available data is insufficient to accurately predict or explain the majority of customer behavior.

6.3 Final Conclusions

Based on this comprehensive analysis, the study concludes the following:

i. The common business assumption that improving customer satisfaction (as measured by product reviews) will directly improve subscription retention is not supported by this data.

ii. The decision to subscribe (retention) is not an experiential problem tied to satisfaction. Instead, it is a demographic and financial problem statistically linked to Gender and, to a lesser extent, Discount value.

iii. A variable's explanatory power (statistical significance) is not the same as its predictive utility (feature importance)**.** Gender was highly significant but a poor predictor, while Age was a strong predictor but not statistically significant. This highlights the importance of using a dual-methodology approach.

6.4 Business Recommendations

The concluding findings of this thesis result in a number of actionable suggestions to the business that the business should no longer concentrate on a hypothesis which was proven to be false but factor in on the factors that proved to be important.

i. **Recommendation 1: De-couple Satisfaction from Retention Strategy.**

Action: Stop investing in programs that attempt to increase review scores with the sole expectation that this will improve subscription rates. The data shows no link. Satisfaction and retention should be treated as separate business problems.

ii. **Recommendation 2: Launch a Strategic Investigation into the 'Gender' Discrepancy.**

Action: The fact that female customers are 71.4% more likely to subscribe is the most significant finding in this study. This should become the company's new focus.

Tactics:

- **Qualitative Research:** Conduct interviews and focus groups with both male and female customers (subscribed and unsubscribed) to understand why this discrepancy exists. Is the product selection, marketing, or perceived value of the subscription fundamentally different for these groups?
- **Marketing & Product:** Re-evaluate the subscription's value proposition. If it is inherently more valuable to female customers, marketing should be laser-focused on this segment. If the goal is to grow the male subscriber base, the offering itself may need to be redesigned to appeal to them.

iii. **Recommendation 3: Carefully A/B Test 'Discount Value' as a Retention Lever.**

Action: The marginal significance of Discount value ($p = 0.056$) suggests it is a potential, but not guaranteed, driver.

Tactics: Implement A/B tests offering different discount levels or promotions specifically aimed at increasing subscriptions (not just one-time sales) to quantify its true impact on retention.

6.5 Limitations of the Study

It is critical to acknowledge the limitations of this research, which are primarily related to the dataset itself.

i. **Weak Proxy for Satisfaction:** Review Rating represents a product level measure, that took the form of a proxy of customer level satisfaction. A customer

has the ability to love a product and not love their service. The null hypothesis was most probably brought about by this.

ii. **Missing Explanatory Data:** The most informative limitation is the poor performance results of all the models (54% accuracy on satisfaction, 63% on retention). It demonstrates the lack of the most significant aspects that can be used to explain why a customer is satisfied or subscribes to the dataset.

6.6 Suggestions for Future Work

Building on these limitations, future research should focus almost entirely on acquiring a more complete dataset.

- **Future Work 1: Implement a True Customer Satisfaction Metric.** The company needs to abandon using product reviews and adopt a real time and post interaction or a post-interaction customer satisfaction (CSAT) survey or a periodical Net Promoter Score (NPS) survey. This would give an actual gauge of the loyalty of customers to compare against.
- **Future Work 2: Expand Data Collection.** The highest priority should be to collect operational data. The inability of the models to predict retention or satisfaction indicates that the drivers are elsewhere. Future datasets must include:
  - **Logistical Data**: Shipping times (promised vs. actual), delivery success rates.
  - **Customer Service Data:** Number of support tickets, reason for contact, resolution time.
  - **Website/App Data:** Website usability, page load times, session engagement.
  - **Product Data:** Product quality/return rates (as distinct from review ratings).

Having a more enriched data set that incorporates all these experiential variables the analysis can be re-run again and give a real, 360 analysis of what actually drives a customer to stay.

REFERENCES

Barzizza, E., Campbell, S., Ceccato, R., Dobosz, A., Haag, M., Martins, R., & Salmaso, L. (2024). A data-driven approach to understanding customer satisfaction. *Journal of Machine Intelligence and Data Science (JMIDS)*, *5*. https://doi.org/10.11159/jmids.2024.001

Chebolu, S. U. S., Dernoncourt, F., Lipka, N., & Solorio, T. (2023). A review of datasets for aspect-based sentiment analysis. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 611–628). Association for Computational Linguistics.

Davoodi, L., & Mezei, J. (2022). A comparative study of machine learning models for sentiment analysis: Customer reviews of e-commerce platforms. In A. Pucihar, M. Kljajić Borštnar, R. Bons, A. Sheombar, G. Ongena, & D. Vidmar (Eds.), *35th Bled eConference: Digital Restructuring and Human (Re)Action*. https://doi.org/10.18690/um.fov.4.2022

Davoodi, L., Mezei, J., & Heikkilä, M. (2025). Aspect-based sentiment classification of user reviews to understand customer satisfaction of e-commerce platforms. *Electronic Commerce Research*, 1–43. https://doi.org/10.1007/s10660-025-09948-4

InsightProgramme. (2023). *Customer Purchase Data* [Dataset]. Kaggle. https://www.kaggle.com/datasets/insightprogramme/customer-purchase-data

Lim, S. L., Foo, L. K., & Chua, S. L. (2023). Comparing machine learning and deep learning based approaches to detect customer sentiment from product reviews. *Journal of System and Management Sciences*, *13*(2), 101–110. https://doi.org/10.33168/JSMS.2023.0207

Noori, B. (2021). Classification of customer reviews using machine learning algorithms. *Applied Artificial Intelligence*, *35*(8), 567–588. https://doi.org/10.1080/08839514.2021.1922843

Ruder, S., Ghaffari, P., & Breslin, J. G. (2016). A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 999–1005). Association for Computational Linguistics.

Seymen, O. F., Olmez, E., Dogan, O., Er, O., & Hiziroglu, A. (2023). Customer churn prediction using ordinary artificial neural network and convolutional neural network algorithms: A comparative performance assessment. *Gazi University Journal of Science*, *36*(2), 720–733. https://doi.org/10.35378/gujs.992738

Shah, S., Ghomeshi, H., Vakaj, E., Cooper, E., & Fouad, S. (2023). A review of natural language processing in contact centre automation. *Pattern Analysis and Applications*. https://doi.org/10.1007/s10044-023-01182-8

Zhang, M., Lu, J., Ma, N., Cheng, T.C.E., & Hua, G. (2022). A feature engineering and ensemble learning based approach for repeated buyers prediction. *International Journal of Computers Communications & Control*, *17*(6), 4988.

https://doi.org/10.15837/ijccc.2022.6.4988

221-35-872

ORIGINALITY REPORT

17%	15%	6%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	Submitted to Midlands State University Student Paper	3%
3	research.abo.fi Internet Source	1%
4	Ashok Kumar Jayaraman, Tina Ester Trueman, Gayathri Ananthakrishnan, Abirami Murugappan et al. "Aspect category and sentiment polarity detection using a permutation language-based transformer model", Procedia Computer Science, 2025 Publication	1%
5	internationalpolicybrief.org Internet Source	1%
6	Submitted to Universiti Malaysia Pahang Student Paper	1%
7	Submitted to Austin High School Student Paper	<1%
8	research.polyu.edu.hk Internet Source	<1%
9	Submitted to Colorado Technical University Online Student Paper	<1%
10	link.springer.com Internet Source	<1%

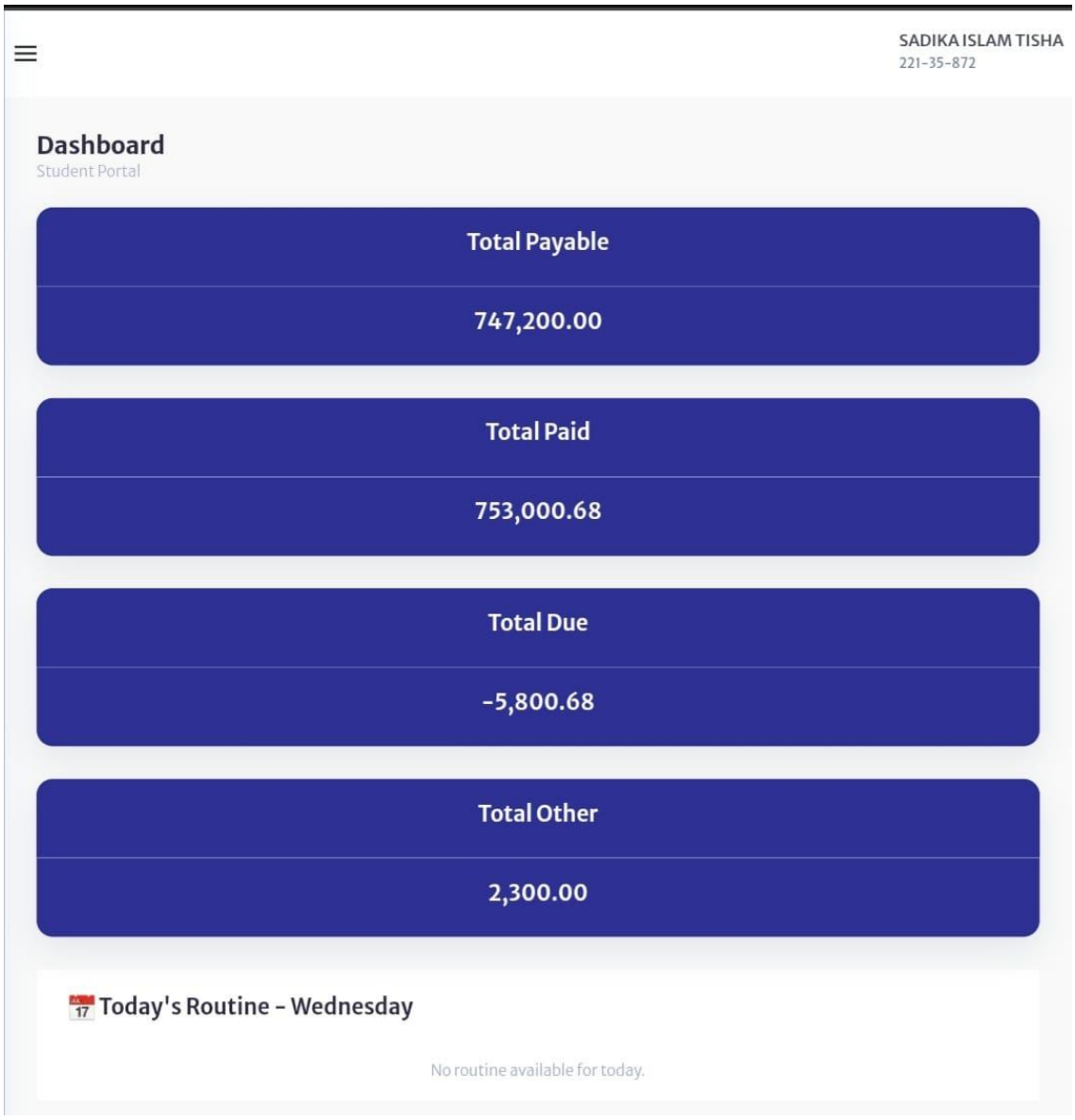
11	Submitted to Pace University Student Paper	<1 %
12	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	<1 %
13	Submitted to United International College Student Paper	<1 %
14	researchspace.ukzn.ac.za Internet Source	<1 %
15	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025 Publication	<1 %
16	tnsroindia.org.in Internet Source	<1 %
17	Submitted to University of Mauritius Student Paper	<1 %
18	thesai.org Internet Source	<1 %
19	uir.unisa.ac.za Internet Source	<1 %
20	doctorpenguin.com Internet Source	<1 %
21	scholarworks.waldenu.edu Internet Source	<1 %
22	file-thesis.pide.org.pk Internet Source	<1 %
23	speedypaper.x10.mx Internet Source	<1 %

24	Submitted to University of Essex Student Paper	<1%
25	globaljournals.org Internet Source	<1%
26	www.studymode.com Internet Source	<1%
27	github.com Internet Source	<1%
28	William B. Ware, John M. Ferron, Barbara M. Miller. "Introductory Statistics: A Conceptual Approach Using R", Routledge, 2013 Publication	<1%
29	dk.um.si Internet Source	<1%
30	aran.library.nuigalway.ie Internet Source	<1%
31	Kavuş, Helin Kardelen. "A New Wave of "High-Skilled" Migration From Türkiye: Call Centre Workers in Athens.", Middle East Technical University (Turkey) Publication	<1%
32	vital.seals.ac.za:8080 Internet Source	<1%
33	repository.psa.edu.my Internet Source	<1%
34	etd.uwc.ac.za Internet Source	<1%
35	Submitted to Dublin Business School Student Paper	<1%
36	Salawudeen, Iyabo Ajarat. "Conflict Management Dynamics and Employees	<1%

	Performance of Revenue Generating Agencies in North-Central Nigeria", Kwara State University (Nigeria), 2024 Publication	
37	Submitted to Sydney Institute of Technology and Commerce Student Paper	<1 %
38	ehealth-graz.at Internet Source	<1 %
39	hdl.handle.net Internet Source	<1 %
40	spectrum.library.concordia.ca Internet Source	<1 %
41	www.mdpi.com Internet Source	<1 %
42	psasir.upm.edu.my Internet Source	<1 %
43	researchwap.net Internet Source	<1 %
44	"Advances in Information Retrieval", Springer Science and Business Media LLC, 2018 Publication	<1 %
45	Arvind Dagur, Sohit Agarwal, Dhirendra Kumar Shukla, Shabir Ali, Sandhya Sharma. "Artificial Intelligence and Sustainable Innovation - Volume 1", CRC Press, 2026 Publication	<1 %
46	Ghita Regasse, Francesco Venier. "Implementing machine learning for predictive analytics: An empirical study of employee turnover", Next Research, 2025 Publication	<1 %

47	Li, Hanyan. "A Framework for Optimizing Public Transit Fleet Conversion to Alternative Fuels.", Georgia Institute of Technology Publication	<1%
48	Varolgunes, Uras. "Harnessing Graphs for Knowledge Representation in Natural Language Processing", New Jersey Institute of Technology, 2025 Publication	<1%
49	Zhaoguang Xu, Yifan Wu, Lin Tang, Shumeng Gui. "From user-generated content to quality improvement: A multi-granularity analysis of customer satisfaction and attention in new energy vehicles using deep learning", Computers in Industry, 2025 Publication	<1%
50	doria.fi Internet Source	<1%
51	etd.cput.ac.za Internet Source	<1%
52	fashiondocbox.com Internet Source	<1%
53	journals.sagepub.com Internet Source	<1%
54	www.dtic.mil Internet Source	<1%
55	Carol Xu, Xuan Luo, Dan Wang. "Chapter 8 MCPR: A Chinese Product Review Dataset forMultimodal Aspect-Based Sentiment Analysis", Springer Science and Business Media LLC, 2022 Publication	<1%


56 cdn.istanbul.edu.tr Internet Source <1%
57 docs.google.com Internet Source <1%
58 dokumen.pub Internet Source <1%
59 irf.fhnw.ch Internet Source <1%
60 publicacoes.riqual.org Internet Source <1%
61 repository.ulb.ac.id Internet Source <1%
62 vtechworks.lib.vt.edu Internet Source <1%
63 www.classace.io Internet Source <1%
64 www.grin.com Internet Source <1%
Exclude quotes Off
Exclude matches Off
Exclude bibliography Off


SADIKA ISLAM TISHA
221-35-872
Dashboard
Student Portal
Total Payable
747,200.00
Total Paid
753,000.68
Total Due
-5,800.68
Total Other
2,300.00
Today's Routine - Wednesday
No routine available for today.