

# **ONLINE SHOPPERS INTENTION PREDICTION USING MACHINE LEARNING ALGORITHMS**

**A Thesis Submitted  
In Partial Fulfillment of the Requirements for the  
degree of**

**MASTERS OF TECHNOLOGY  
in  
Data Science**

**by  
Sundram Jha  
(2K23/DSC/11)**

**Under the Supervision of  
Prof. Ruchika Malhotra  
Head of Department, Department of Software Engineering  
Delhi Technological University**



**DEPARTMENT OF SOFTWARE ENGINEERING  
DELHI TECHNOLOGICAL UNIVERSITY  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi 110042**

**May, 2025**

**DEPARTMENT OF SOFTWARE ENGINEERING**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of  
Engineering) Bawana Road,  
Delhi 110042


**CERTIFICATE**

I hereby certify that the Project Dissertation titled "Online Shoppers Intention prediction using Machine Learning algorithms" which is submitted by Sundram Jha, Roll No. 2K23/DSC/11, Department of Software Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

**Place: New Delhi**

**Date:**

  
**Prof. Ruchika Malhotra**  
Head of Department  
Department of Software Engineering, DTU

**DEPARTMENT OF SOFTWARE ENGINEERING**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of  
Engineering) Bawana Road,  
Delhi 110042

**ACKNOWLEDGEMENT**

I wish to express my sincerest gratitude to Prof. Ruchika Malhotra for her continuous guidance and mentorship that she provided me during the project. She showed me the path to achieve my targets by explaining all the tasks to be done and explained to me the importance of this project as well as its industrial relevance. She was always ready to help me and clear my doubts regarding any hurdles in this project. Without her constant support and motivation, this project would not have been successful.

Place: New Delhi

Date:

SUNDRAM

Sundram Jha

(2K23/DSC/11)

**DEPARTMENT OF SOFTWARE ENGINEERING**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of  
Engineering) Bawana Road,  
Delhi 110042

**CANDIDATE'S DECLARATION**

I, Sundram Jha, 2K23/DSC/11 students of M.Tech (Data Science), hereby certify that the work which is being presented in the thesis entitled "Online Shoppers Intention prediction using Machine Learning algorithms" in partial fulfillment of the requirements for the award of degree of Master of Technology, submitted in the Department of Software Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from Jan 2025 to May 2025 under the supervision of Prof. Ruchika Malhotra.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other institute.

**SUNDRAM**

**Candidate's Signature**

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.



**Signature of Supervisor**

## ABSTRACT

In today's fast-moving e-commerce landscape, being able to accurately predict whether an online shopper is likely to make a purchase is incredibly valuable for businesses aiming to boost sales and stay ahead of the competition. While earlier machine learning models like XGBoost and Random Forest have shown decent results in this area, they struggle to capture the complex relationships between features or interpret sequential user behavior. Transformer-based models originally designed for natural language tasks have started gaining traction in structured data prediction due to their ability to model interactions and dependencies more effectively. This research takes an in-depth look at two such models: SAINT and FT-Transformer. When tested on a familiar dataset for shopper behavior, SAINT achieved an accuracy of 89.91% and an AUC-ROC of 90.65%, while FT-Transformer also gave the same accuracy but slightly lower AUC-ROC at 89.68%. When compared with traditional models like XGBoost and Random Forest, which are the same in accuracy, they fell short in AUC-ROC, highlighting the superior ability of transformers to deal with imbalanced datasets. The attention mechanisms in SAINT and FT-Transformer helped identify detailed patterns in user sessions, resulting in better generalization. These findings offer promising directions for more intelligent, data-driven marketing strategies.

## TABLE OF CONTENT

<b>ACKNOWLEDGEMENT</b> .....	I
<b>CANDIDATE’S DECLARATION</b> .....	II
<b>CERTIFICATE</b> .....	III
<b>ABSTRACT</b> .....	IV
<b>TABLE OF CONTENT</b> .....	V
<b>LIST OF FIGURES</b> .....	VI
<b>LIST OF TABLES</b> .....	VII
<b>Chapter 1</b> .....	1
Introduction.....	1
1.1 Problem Statement.....	1
1.2 Limitations of Traditional Methods.....	1
1.3 Recent Research Development in DL.....	2
1.4 Research Gaps.....	2
1.5 Contribution.....	3
<b>Chapter 2</b> .....	5
Related Work.....	5
<b>Chapter 3</b> .....	8
<b>Research Methodology</b> .....	8
3.1 Framework Overview.....	8
3.2 Dataset Profile.....	9
3.3 Exploratory Data Analysis (EDA).....	12
3.4 Preprocessing Pipeline.....	14
3.4.1 Data Preprocessing and Cleaning.....	15
3.4.2 Exploratory Data Analysis (EDA).....	15
3.4.3 Feature Engineering and Selection.....	15
3.4.4 Model Training and Evaluation.....	15
3.4.5 Model Architectures Used.....	18
3.4.6 Model Selection and Deployment.....	19
<b>Chapter 4</b> .....	20
Results And Discussion.....	20
<b>Chapter 5</b> .....	22
Conclusion and Future Scope.....	22
<b>Bibliography</b> .....	23



## LIST OF FIGURES

3.1 Model Framework.....	9
3.2 Dataset.....	10
3.3 Heatmap.....	12
3.4 Histogram .....	12
3.5 Pie chart .....	13
3.6 Histogram.....	13
3.7 Revenue distribution graph.....	14
3.8 Distribution plots.....	14
3.9 SAINT Architecture.....	16
3.10 FTT Architecture.....	17
3.11 ROC Curve .....	19
4.1 Grouped bar chart.....	20
4.2 Comparison ROC Curve .....	21

## **LIST OF TABLES**

2.1 Summary of the studies undertaken for review.....	7
3.1 Feature Summary and Preprocessing Transformations .....	10
4.1 Evaluation Results on the Test Dataset .....	20



# CHAPTER 1

## INTRODUCTION

As online shopping continues to reshape the retail landscape, understanding whether a user is likely to complete a purchase has become more than just a data science problem it's a business necessity. The ability to anticipate shopper behaviour can help companies craft personalized experiences, streamline marketing efforts, and manage inventory more efficiently.

### 1.1 PROBLEM STATEMENT

Accurately predicting online shopper purchase intentions is critical for optimizing e-commerce strategies, yet existing machine learning approaches struggle to address key challenges inherent in session-based behavioral data. Traditional methods, such as decision trees, gradient boosting, or recurrent neural networks (RNNs), often fall short in capturing complex, non-linear interactions between heterogeneous features (e.g., page durations, bounce rates, categorical traffic sources) while maintaining computational efficiency. Recent advances in transformer-based architectures, such as SAINT (Self-Attention for Tabular Data) and FT-Transformer (Feature Tokenization Transformer), promise to overcome these limitations by leveraging self-attention mechanisms to model intricate feature relationships and handle tabular data more effectively. However, their efficacy in predicting purchase intentions remains underexplored.

Key unresolved questions include:

- Can transformer models outperform traditional algorithms (e.g., XGBoost, LSTM) in accuracy and robustness on imbalanced e-commerce data?
- How do SAINT and FT-Transformer differ in handling categorical features, temporal session patterns, and sparse purchase signals?
- Do these architectures provide sufficient interpretability to identify critical behavioral drivers (e.g., page value, exit rates) for business decisions?

This study addresses these gaps by conducting a systematic comparative analysis of SAINT and FT-Transformer against state-of-the-art baselines. By evaluating performance metrics (accuracy, F1-score, AUC-ROC), computational efficiency, and interpretability, this research aims to establish a framework for deploying transformer-based models in real-world e-commerce systems, enabling dynamic customer targeting and personalized engagement strategies.

### 1.2 LIMITATIONS OF TRADITIONAL METHODS

As online shopping continues to reshape the retail landscape, understanding whether a user is likely to complete a purchase has become more than just a data science problem it's a business necessity. The ability to anticipate shopper

behaviour can help companies craft personalized experiences, streamline marketing efforts, and manage inventory more efficiently. Traditionally, machine learning models like XGBoost and Random Forest have been widely adopted for predicting purchasing intent. These models perform well with structured data and are relatively straightforward to implement. However, they often fall short when it comes to recognizing deeper patterns, especially those that emerge over the course of a user's interaction with an online platform.

### 1.3 RECENT RESEARCH DEVELOPMENTS IN DEEP LEARNING

Recent research in deep learning ,especially in the field of transformers , have opened new doors as earlier developed for tasks in natural language processing, transformer-based models have shown strong potential in handling complex feature relationships even in structured tabular data, such as e-commerce logs. Even with their rising popularity, the use of transformer models for predicting online shopper behaviour is still in its early stages, and comparisons with traditional models remain limited.This research takes a step toward filling that gap by comparing the performance of two powerful transformer-based architectures SAINT and FT-Transformer against more established machine learning methods like XGBoost and Random Forest. Using a publicly available dataset of online shopper sessions, we evaluate these models based on both accuracy and AUC-ROC. The results show that SAINT offers strong predictive power, better than traditional approaches and suggesting that transformers can better capture the subtle behaviours of online users.With the deep analysis , how these models perform in real world cases, research offers practical insights for e-commerce platforms aiming to make their predictive capabilities better and filter their approach to customer engagement..

### 1.4 RESEARCH GAPS

Despite advances in machine learning and behavioral analytics, there are substantial gaps in predicting online consumers' buying intentions and therefore limiting real-world applicability, fairness, and scalability. The gaps are as follows:

1. **Data Privacy and Ethical Issues:** Most research cares more about model accuracy than about user consent and anonymization, particularly for sensitive behavioral data (e.g., session cookies, purchase history), Non-compliance with GDPR, CCPA, and other privacy laws in data collection and usage, Failure to conduct adequate exploration of federated learning or differential privacy techniques to protect user data and maintain predictive performance.
2. **Lack of Model Interpretability and Explainability:** Deep learning (i.e., LSTM, Transformer-based models) and sophisticated ensemble techniques (i.e., XGBoost, LightGBM) are "black boxes," which reduces the confidence of marketers and policymakers, The application of SHAP (SHapley Additive

exPlanations), LIME, or counterfactual explanations to explain predictions is limited, Accuracy and interpretability trade-offs are rarely theoretically quantified, leaving companies in doubt about the use of such models.

3. **Static Feature Engineering vs. Dynamic Behavioral Changes:**The majority of models depend on historical, aggregated features (e.g., aggregate clicks, dwell time) instead of actual-time behavioral trends (e.g., micro-sessions, mouse behavior, hesitation rhythms).Few employ time-series analysis or reinforcement learning to learn to respond to sudden shifts (e.g., flash sales, seasonality). Session-aware modeling (such as RNNs with attention) is under-exploited for short-term intent prediction.
4. **Multimodal Data Integration Under Constraints:**Excessive reliance on clickstream and structured data, ignoring unstructured data such as:text (product reviews, chatbot interactions),Visual elements (product images, video interaction),Acoustic data (oral search queries, client support messages) Lack of frameworks for cross-modal learning, like vision-language models used in product recommendations.
5. **Computational Efficiency and Scalability Issues:**Very computationally intensive state-of-the-art deep learning models (e.g., BERT for NLP, Graph Neural Networks for recommendation systems) are not feasible for SMEs. Scarce research on light-weight architectures (e.g., knowledge distillation, quantization) for deployment on the edge, Most tests are run on idealized data, not the noise and latency bounds of real data.
6. **Bias and Fairness in Predictive Models:** Most datasets are plagued by selection bias (oversampling of specific groups), Few pieces of work compute fairness metrics (e.g., demographic parity, equalized odds) to ensure fairness between user groups, Algorithmic discrimination risks (e.g., price steering on the basis of user profiles) are hardly mentioned.

## 1.5 CONTRIBUTION

This research addresses these gaps and contributes to the growing field of AI-driven mental health diagnostics in several key ways:

### 1. Privacy-Preserving AI

Gap: Most of the research (e.g., LSTM-RNN methods) are based on session-based data to be GDPR-conform but may not necessarily adopt privacy mechanisms such as federated learning or differential privacy.

Future Work: Develop end-to-end frameworks that combine anonymized session modeling with privacy-preserving techniques (e.g., synthetic data generation) to protect sensitive user interactions, with predictive accuracy maintained.

### 2. Explainable and Interpretable Models

Gap: Deep learning models (LSTM, transformers) and ensemble techniques will most probably be "black boxes," restricting confidence to mission-critical use cases

such as real-time cart abandonment interventions.

Future Work: Incorporate post-hoc interpretability tools (e.g., SHAP, LIME) into training pipelines and investigate inherently interpretable architectures (e.g., attention-based transformers with saliency maps).

### 3. Dynamic and Real-Time Adaptation

Gap: Although research focuses on real-time prediction (e.g., sliding-window LSTMs), few consider dynamic feature engineering or concept drift due to changing user behavior in scenarios such as flash sales.

Future Work: Implement reinforcement learning or online learning in order to train models in iterations, and incorporate streaming data pipes for real-time recalibration of features.

### 4. Multimodal Data Integration

Gap: The majority of work concentrates on clickstream or structured data, ignoring unstructured inputs such as product images, reviews, or chatbot transcripts that can enhance intent signals.

Future Work: Develop cross-modal architectures (e.g., vision-language models) to combine clickstreams with visual product information and text feedback to enhance robustness of prediction.

### 5. Computational Efficiency for SMEs

Gap: New models (deep learning, CatBoost) have extremely high resource requirements, which are restraining adoption among small businesses.

Future Work: Optimize deployment by using light-weight techniques (knowledge distillation, quantization) and explore edge-AI platforms for low-latency, resource-constrained environments.

### 6. Bias and Fairness

Gap: There are few studies assessing demographic fairness (e.g., regional or device-level biases in datasets such as "online-shoppers-intention").

Future Work: Use fairness measures (demographic parity, equalized odds) when training models and auditing datasets for representational bias.

### 7. Human-AI Collaboration

Gap: Excessive automation threatens to ignore contextual feedback from marketing professionals.

Future Work: Develop hybrid intelligence systems that merge AI predictions with human feedback mechanisms for adaptive adjustments of campaigns.

## CHAPTER 2

### RELATED WORK

Predicting the purchasing plans of internet shoppers has also attracted significant attention over the past decade or so, thanks to the need for businesses to maximize customer engagement and optimize conversion rates. Researchers have sought to address this challenge through multiple machine learning (ML) and deep learning methods, where accuracy improved, dealing with imbalanced datasets was foreseen, and feature engineering was utilized. Earlier work, such as Sakar et al. (2019), used traditional classifiers like Random Forest (RF) and Multi-Layer Perceptrons (MLPs) with 87–90% accuracies while emphasizing the significance of session-based features like page values and bounce rates. Subsequent work by Kabir et al. (2019) revealed the strength of ensemble methods where Gradient Boosting was 90.34% accurate, demonstrating the value of ensemble methods for achieving improved predictive accuracy with many weak learners. Latest developments have brought forth deep learning models, like Long Short-Term Memory (LSTM) networks and stacked ensembles, to capture sequential user behavior. For example, Mootha et al. (2020) came up with a stacking ensemble of MLPs that achieved 94% accuracy using meta-classifiers to further improve predictions. In contrast, Prayogo and Karimah (2021) handled class imbalance through Adaptive Synthetic Sampling (ADASYN) combined with feature selection to obtain 93.34% accuracy using Random Forest, underscoring the importance of data preprocessing to counteract bias. These developments notwithstanding, efforts continue to be hampered by real-time prediction, interpretability, and generalizability on various e-commerce platforms, compelling further innovation in hybrid models and explainable AI architectures. Esmeli et al. [1] discussed a detailed review of predictive modeling methods under the framework of customer behavior, including topics like web log analysis, estimation of purchase probabilities, and strategies of personalization for web marketing. Their study determines the challenges arising due to imbalanced data sets and the requirement for robust measures of performance analysis. By focusing on early purchase behavior prediction, they get a good base to conduct future work using more sophisticated models, i.e., transformer-based models. Abd Rashid et al. [2] authored on their research on Malaysian online buying behavior by consumers and helped bridge an important knowledge gap in consumer research in the country. Their results are useful lessons for internet traders, particularly the use of additional influence factors and the combining of different research approaches. They also proposed that more sophisticated models such as transformers can provide even more useful information regarding customers' behavior. Noviantoro et al. [3] have studied several deep learning and data mining methods to study the behavior of how people shop online, focusing on clickstream data. By comparing models like



Random Forest and Neural Networks, they have shown the importance of considering several right features for better prediction. Their research gives the theoretical basis for using powerful models like transformers that will be capable enough to detect deeper patterns in user behavior. Jie Wang et al. [4] introduced an ensemble model called SE-stacking, which combines several learning models to improve results. Their model reached an impressive F1-score of 98.40%, showing how combining models can help capture more details about customer actions. This approach works well alongside transformers, which are made to handle complex data using attention mechanisms.

Mingcheng Yang et al. [5] constructed a hybrid model from CatBoost and Logistic Regression to predict buy or not from the users. Their model was highly precise and had high F1 scores, showing the power of boosting combined with more traditional techniques. They propose that transformers being added might actually make it even more powerful by detecting more intricate patterns in the data. Hazare et al. [6] discuss the usage of sentiment analysis on social media data, that is Twitter, to measure and predict consumer buying behavior. Their study indicates the potential of user-generated content mining for extracting valuable information. Since social media text provides context, transformer-based models like BERT and its variants can significantly improve prediction of sentiment-based purchase intent. Raja et al. [7] use the UCI Online Shoppers Intention dataset to build a predictive model with Random Forests. With feature selection and hyperparameter tuning, they achieve an impressive 94% accuracy. The breakdown of significant behavioural indicators reveals the ability of machine learning to inform user interface and marketing choices—something that could be even more exploited with attention-based models for more realistic interaction modelling. Rashaduzzaman et al. [8] examine the determinants of American consumer's buying decision for online garments. Their work indicates that price, product offerings, and accessibility have a critical role in affecting customer decisions. The application of transformer models in such scenarios could help retailers personalize their products better by learning from dynamic behaviour signals. A. Gomes et al. [9] present a study that focused on consumer segmentation, recommendation systems, and behaviour modelling within e-commerce environments. They highlight the growing importance of feature embedding techniques, which align well with the strengths of transformer-based models in learning rich, high-dimensional representations. Their work points toward the increasing need for real-time, personalized shopping experiences, which advanced architectures like SAINT and FT-Transformer are well-equipped to support.

Aspect	Diamantaras et al. (LSTM-RNN)	Gomes et al. (Embeddings + LSTM)	Mostafa et al. (Traditional ML)	Mootha et al. (Stacking MLP Ensemble)	Karakaya et al. (Ensemble Learning)
Objective	Predict real-time purchase intent using session-based LSTM-RNN.	Predict purchase intent using customer embeddings and LSTM for real-time analysis.	Classify purchase intent using traditional ML algorithms.	Improve prediction accuracy via a stacking ensemble of MLPs.	Analyze purchase intent using a stacking ensemble of kNN, RF, and modlem with Naive Bayes.
Methodology	LSTM-RNN with session sequences and minimal feature engineering.	Skip-gram embeddings + LSTM; real-time prediction with fast embedding computation.	RF, DT, SVM, k-NN on preprocessed dataset; CRISP-DM framework.	Two-level stacking: MLPs as base models and meta-classifier.	Stacking ensemble combining kNN, RF, modlem; Naive Bayes as meta-classifier.
Dataset	Private industry dataset (leather apparel e-commerce). 21,896 sessions.	Three datasets: yoochoose, openCDP, and a closed US retailer dataset.	UCI "Online Shoppers Purchasing Intention" dataset (12,330 sessions).	UCI "Online Shoppers Purchasing Intention" dataset.	UCI "Online Shoppers Purchasing Intention" dataset.
Key Features	Session actions, time spent, origin, season, day, working hours.	Clickstream data, event types, session sequences.	Administrative/Informational/Product-related page views, bounce/exit rates, special days.	Session details, page values, and Google Analytics metrics.	Clickstream data, session duration, customer type, and page metrics.
Models/Algorithms	LSTM-RNN, GRU.	Skip-gram embeddings, LSTM, DT, RF, GB, LR, MLP.	RF, DT, SVM, k-NN.	Stacked MLP ensemble.	kNN, RF, modlem, Naive Bayes.
Results	98% accuracy (industry dataset).	94% accuracy (closed dataset), 235x faster feature computation than baseline.	RF: 90.24% accuracy.	94% accuracy (highest on UCI dataset).	87.4% accuracy, 88 F1-score.
Strengths	High accuracy with minimal feature engineering; real-time applicability.	Transferable embeddings, real-time capability, outperformed SotA on multiple datasets.	Simple implementation with strong performance from RF.	Novel stacking architecture; outperformed 15+ classifiers.	Combines diverse classifiers; addresses class imbalance.
Limitations	Industry-specific dataset limits generalizability.	Requires handling unknown touchpoints; computational cost for embeddings.	Lower accuracy compared to deep learning methods; imbalanced dataset.	Complex architecture; requires significant computational resources.	Moderate accuracy compared to deep learning models.
Key Contribution	Demonstrated LSTM effectiveness for session-based intent prediction without heavy feature engineering.	Introduced embeddings for customer behavior representation, enabling faster real-time predictions.	Highlighted RF as the best traditional ML classifier for purchase intent prediction.	Proposed a high-accuracy stacking ensemble model for imbalanced datasets.	Combined rule-based (modlem) and statistical classifiers for improved robustness.

Table .2.1 Summary of the studies undertaken for review





## Chapter 3

### Research Methodology

This section details the comprehensive methodology adopted to model and predict online shoppers' purchasing intentions. Our approach involves several stages: from carefully preparing the data, engineering meaningful features, and applying both classical and cutting-edge machine learning models, to evaluating model performance using robust metrics. We give particular attention to transformer-based models like SAINT and FT-Transformer, comparing their effectiveness with traditional baselines such as XGBoost and Random Forest.

#### 3.1 Framework Overview

This research employs a systematic framework to predict online shoppers' purchasing intentions, structured as follows:

- **Dataset** : Utilizes the “Online Shoppers Purchasing Intention” dataset from the UCI Machine Learning Repository, comprising 12,330 sessions with features such as page values, bounce rates, and session durations.
- **Data Cleaning** : Addresses missing values, outliers, and inconsistencies while encoding categorical variables (e.g., visitor type, month) and normalizing numerical attributes to ensure data quality.
- **Exploratory Data Analysis (EDA)** : Analyzes feature distributions, class imbalance, and correlations to identify behavioral patterns (e.g., time spent on product pages, weekend shopping trends) that influence purchase decisions.
- **Feature Engineering** : Enhances predictive power by selecting critical features (e.g., page value, exit rate) and generating new metrics (e.g., session efficiency score) using domain knowledge and statistical methods.
- **Model Training and Evaluation** : Trains multiple classifiers (e.g., Random Forest, XGBoost, Deep Neural Networks) and evaluates performance using cross-validation, accuracy, precision, recall, and F1-score to address class imbalance.
- **Model Selection and Deployment** : Selects the optimal model based on robustness and generalizability, then deploys it as an API or integrates it into

e-commerce platforms for real-time prediction, enabling personalized recommendations and targeted marketing.

This end-to-end framework ensures a data-driven approach to understanding and predicting online shopper behavior, bridging theoretical insights with practical applications in dynamic e-commerce environments.

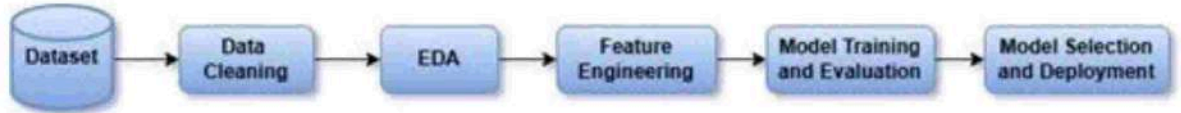


Fig. 3.1. Proposed Machine Learning Pipeline Framework

### 3.2 Dataset Profile

The Online Shoppers Purchasing Intention dataset, hosted by the UCI Machine Learning Repository, is a widely used benchmark for predicting e-commerce user behavior. Collected over a one-year period from an e-commerce platform, it comprises session-level data of 12,330 users, with the goal of determining whether a visit concludes with a transaction.

Key Attributes:

- Features: 17 Attributes (10 numerical, 8 categorical):
- Numerical: Administrative page duration, informational page duration, product-related page duration, bounce rate, exit rate, page value, special day (proximity to holidays), etc
- Categorical: Month, operating system, browser, region, traffic type, visitor type (new/returning), weekend flag.
- Target Variable: Revenue (binary class: 0 = no purchase, 1 = purchase).
- Class Distribution: Imbalanced dataset: Negative class (no purchase): 10,422 sessions (84.5%). Positive class (purchase): 1,908 sessions (15.5%).
- Data Source: Derived from Google Analytics metrics and session logs, including user interactions (e.g., page views, time spent) and traffic sources.

Key Characteristics:

- Session-specific metrics: Features like PageValue (average revenue per page)

and ExitRate (likelihood of leaving the site) are critical predictors.

- Temporal factors: Attributes such as Month and SpecialDay (e.g., holidays) highlight seasonal purchasing trends.

Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0	0	0.0	0	0.0	1	0.00000	0.20	0.20	0.0	0.0	Feb	1	1	1	1 Returning_Visitor	False	False
1	0	0.0	0	0.0	2	64.00000	0.00	0.10	0.0	0.0	Feb	2	2	1	2 Returning_Visitor	False	False
2	0	0.0	0	0.0	1	0.00000	0.20	0.20	0.0	0.0	Feb	4	1	9	3 Returning_Visitor	False	False
3	0	0.0	0	0.0	2	2.66667	0.05	0.14	0.0	0.0	Feb	3	2	2	4 Returning_Visitor	False	False
4	0	0.0	0	0.0	10	627.50000	0.02	0.05	0.0	0.0	Feb	3	3	1	4 Returning_Visitor	True	False

Fig 3.2: Dataset

Feature	Type	Preprocessing Step
Administrative	Num.	Standard Scaler
Administrative Duration	Num.	Standard Scaler
Informational	Num.	Standard Scaler
Informational Duration	Num.	Standard Scaler
Product Related	Num.	Standard Scaler
Product Related Duration	Num.	Standard Scaler
Bounce Rates	Num.	Standard Scaler
Exit Rates	Num.	Standard Scaler
Page Values	Num.	Standard Scaler
Special Day	Num.	Standard Scaler

Month	Catg.	Label Encoding
Operating Systems	Catg.	Label Encoding
Browser	Catg.	Label Encoding

Table 3.1: Feature Summary and Preprocessing Transformations.

### 3.3 Exploratory Data Analysis (EDA)

EDA was conducted as a crucial initial step to gain intuitive insights into the structure, quality, and relationships of the dataset features. This phase involved the use of descriptive statistics and visualizations to examine data.

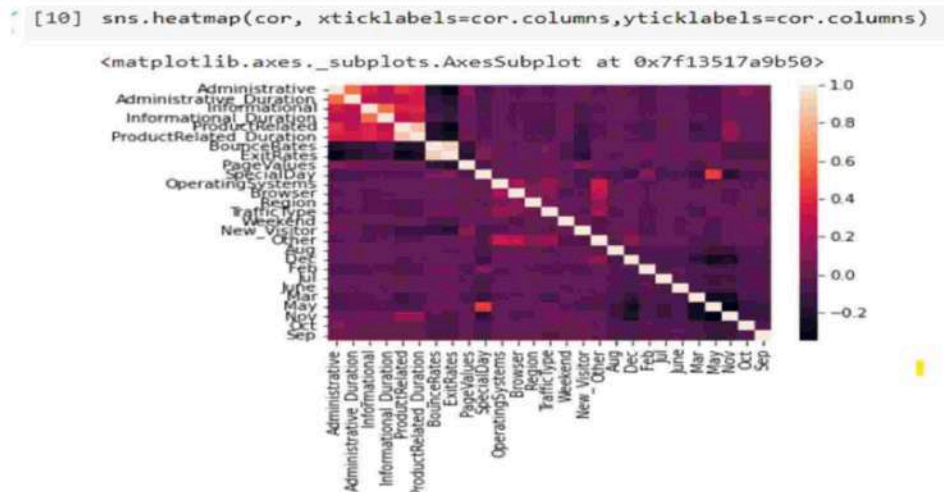


Fig 3.3 represents heat map to depict the correlation between various fields present in dataset.

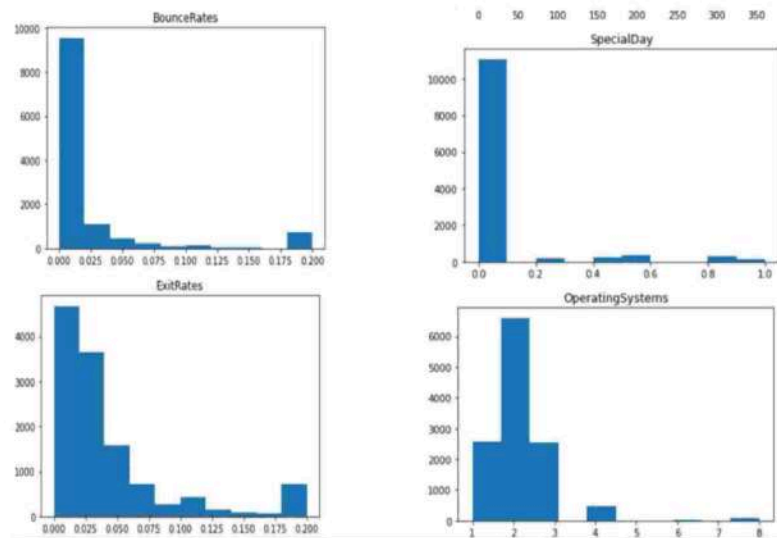


Figure 3.4 represents the histogram for various features such as exit rates, bounce rate, etc.

## Different Types of Visitors

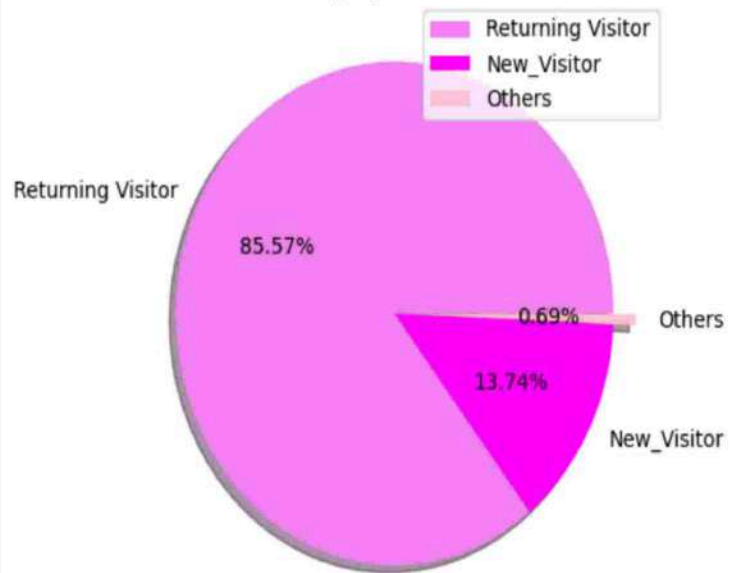


Figure 3.5 represents pie chart for different types of visitors



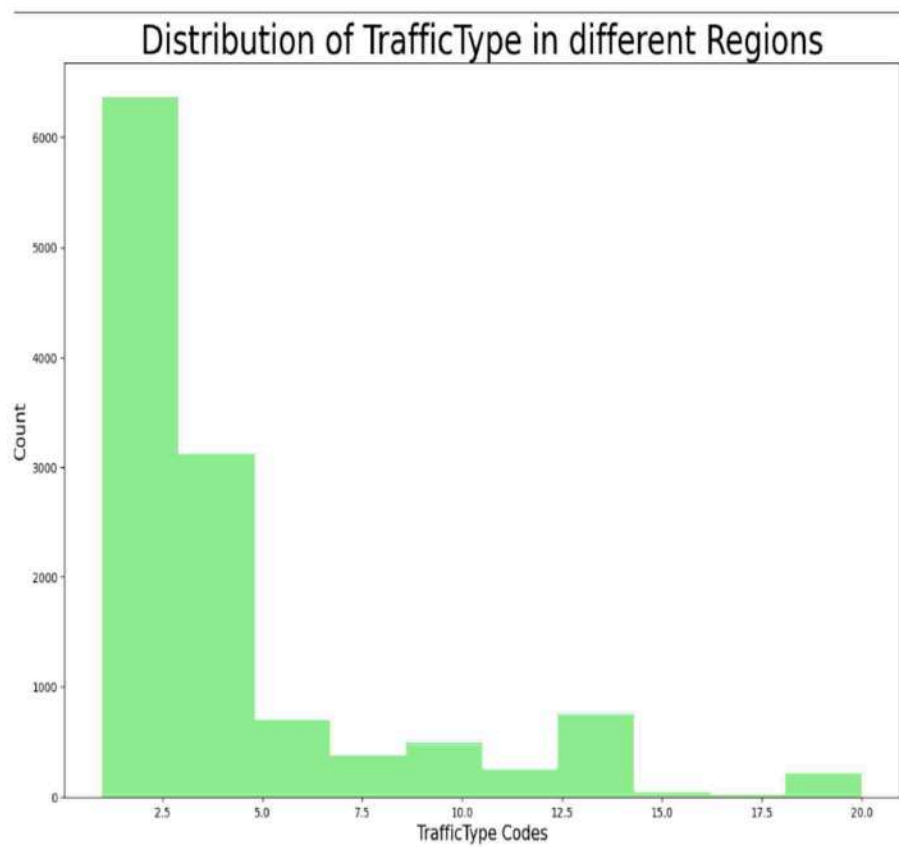


Figure 3.6 Visualizing the Distribution of Customers Around the Region

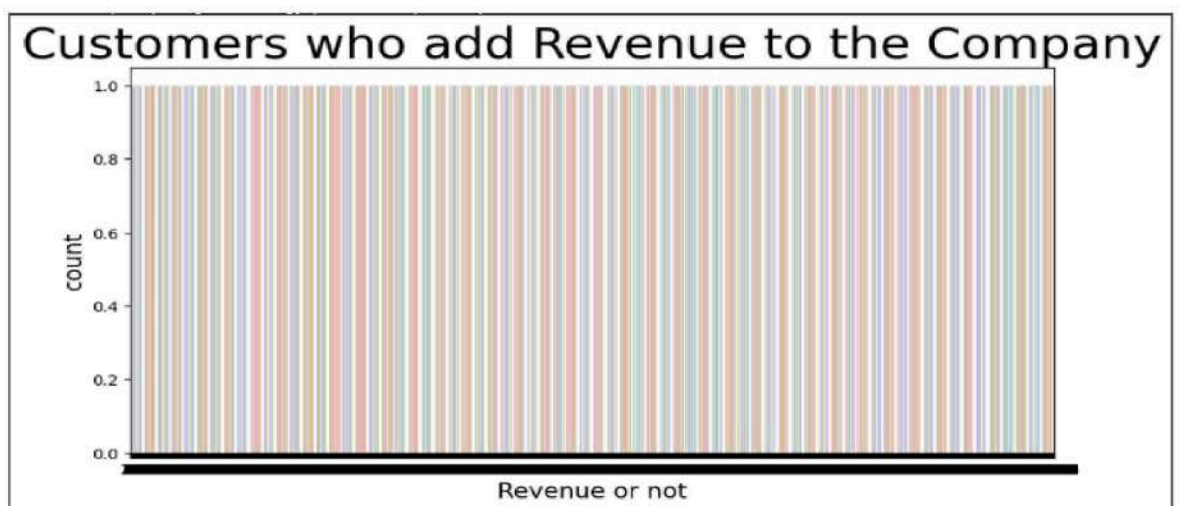
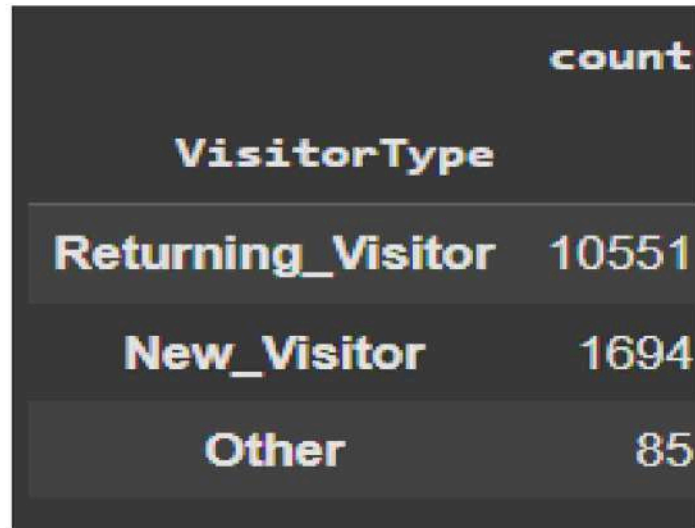


Figure 3.7 Visualizing Customers who add Revenue to the Company



VisitorType	count
Returning_Visitor	10551
New_Visitor	1694
Other	85

Figure 3.8 Represent Type of Visitors

### 3.4 Preprocessing Pipeline

Data preprocessing plays a pivotal role in achieving model stability, accuracy, and generalizability. Each preprocessing phase ensures that the dataset used to train the model is clean, representative, and suitably formatted for downstream analysis.

#### 3.4.1 Data Preprocessing and Cleaning:

- **Encoding Categorical Variables:** We applied Label Encoding to all categorical features. This transformation ensures compatibility with models that require numerical input without introducing artificial order.
- **Feature Normalization:** We standardized all numerical features using Standard Scaler, which rescales each feature to have zero mean and unit variance. This step improves model convergence, especially for neural networks.
- **Handling Missing Values:** Missing value rows were examined in detail. If a feature contained a very low percentage of missing entries, corresponding records were removed to maintain data integrity. Where the data had been lost to a significant extent, imputation methods such as mean, median, or mode fill could be performed to maintain significant samples. For this thesis,

the comparatively low rate of missing entries allowed for deletion.

- **Deletion of Incorrect Data:** Incorrect data with values that failed logical expectation were eliminated. This ensured model training was entered by exclusive meaningful and valid inputs.
- **Treatment of Outliers:** Statistical methods such as IQR (Interquartile Range) and Z-score analysis were employed to detect extreme outliers in numeric variables, Outliers which were not conforming to the expected behavior were capped or replaced with the median value to minimize skewness and prevent them from skewing model training disproportionately..

#### 3.4.2 Feature Engineering and Selection

- **Label Encoding:** Label encoding was applied to categorical variables, where each unique category had an integer assigned to it. This helped machine learning algorithms to handle and comprehend categorical data effectively without implying ordinal bias.
- **Feature Selection:** Statistical measures, including mutual information scores and chi-square tests, were utilized to identify the most informative features. Further, model-based feature importance from Random Forests and XGBoost was used to rank features based on predictive power.

#### 3.4.3 Model Training and Evaluation

To have an unbiased and fair assessment, we divided the data into training, validation, and test sets with a 64:16:20 split. We used stratified sampling to preserve the native class distribution in all subsets. This practice discourages model bias toward the predominant class and enables generalization on new data.

#### 3.4.4 Model Architectures Used

Our methodology incorporates a combination of traditional machine learning algorithms and advanced deep learning models. This hybrid approach allows us to leverage both interpretability and deep feature representation.

- **SAINT(Self-Attention and Intersample Attention Transformer):** SAINT is a deep learning model specially made to work with table like data, where each row is a record and each column is a feature. What makes SAINT stand out is that it doesn't just look at features in one row it also looks at patterns across different rows. This mix of two types of attention is why it's called SAINT .In this, both numerical and categorical features are first turned into embeddings so they can be processed by a neural network. These embeddings go through layers of self-attention, helping the model learn how features in a single row relate to each other. But it doesn't stop there. It also uses intersample attention, which lets it compare different rows to find common

patterns. This is especially helpful when some rows have missing or messy data, or when rows share similar traits. By using both types of attention, SAINT gets a much better understanding of the data. It works really well on tasks like predicting if someone will make a purchase and has even outperformed popular models like XGBoost in many tests. Figure 3.9 below elaborates architecture of a transformer-based framework for tabular data, integrating both categorical and numerical embeddings through feature tokenization, followed by an encoder-decoder structure with multi-head attention.

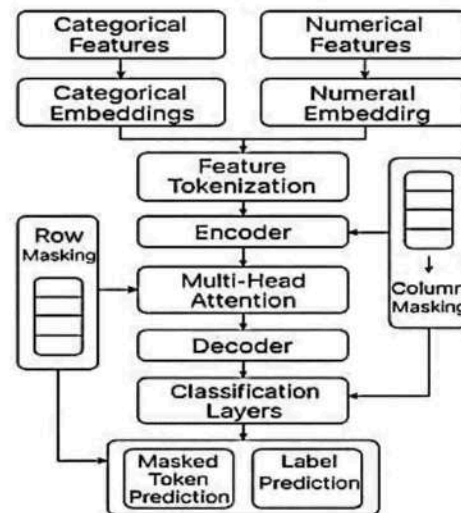
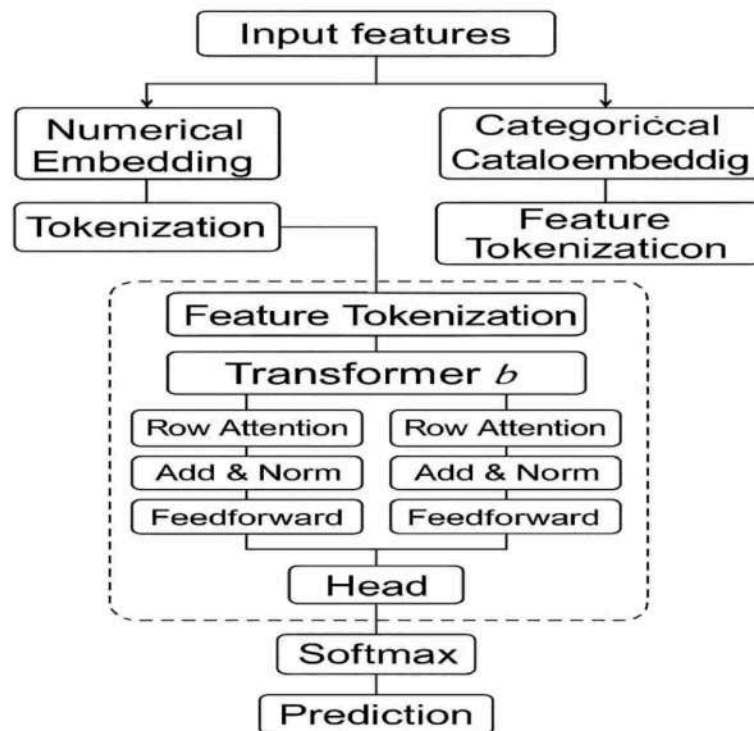


Figure 3.9: SAINT Architecture Flow.

- FT-Transformer (Feature Token Transformer):** The FT-Transformer is a deep learning model designed to work with structured data, like tables where each row is a data point and each column is a feature. Unlike older models like XGBoost, which need a lot of manual tuning and feature setup, FT-Transformer takes a smarter approach inspired by language models like BERT. In this model, every feature is treated like a token. Both numbers and categories are turned into dense vectors called embeddings. These embeddings go through self-attention layers inside the Transformer, helping the model understand how different features are connected. One big benefit of FT-Transformer is that it can learn complex patterns on its own, without needing hand-crafted rules. After using attention, the model combines what it has learned and feeds it through neural layers to make a final prediction. This setup works well even with large and messy datasets, and it adapts better to different tasks. It performs as well as top models like CatBoost and SAINT, while also being easier to interpret and faster to train. Figure 3.10 below elaborates the architecture of the FT-Transformer model, which processes numerical and categorical features through embeddings and feature tokenization, followed by transformer blocks with row-wise attention,

normalization, and feedforward layers to generate final predictions via a softmax head.



## FT-Transformer

Figure 3.10: FT Transformer Architecture Flow.

- **LSTM (Long Short-Term Memory Network):** LSTMs are usually used for sequence data, but we reshaped our tabular data into a 3D format to see if any sequence-like patterns could be found. The model had: Two LSTM layers stacked one after the other, with dropout added to prevent overfitting. A final fully connected layer using a sigmoid function to make yes or no predictions. This architecture helped us test whether LSTM could pick up hidden patterns by treating the features as if they were a sequence.
- **Random Forest:** Random Forest is a robust ensemble model that constructs multiple decision trees using bootstrapped subsets of the data and aggregates their outputs through majority voting. It is particularly suited to datasets with both numerical and categorical features. Its resistance to overfitting and ability to measure feature importance made it a strong baseline for this study. The algorithm benefits from its ensemble nature, where multiple trees reduce the variance associated with individual trees.
- **XGBoost:** XGBoost (Extreme Gradient Boosting) is an advanced boosting

algorithm known for its scalability and performance. It handles missing data internally, uses regularization to combat overfitting, and supports weighted class imbalances. These properties made it ideal for a dataset with class imbalance and diverse feature types. XGBoost builds trees sequentially, where each new tree attempts to correct the errors of the previous ones, leading to highly optimized predictions. In this thesis, XGBoost consistently demonstrated high accuracy, especially when capturing intricate interactions between features.

### 3.4.5 Model Selection and Deployment

All models were optimized using binary cross-entropy loss, which suits the binary nature of the target variable. Deep learning models were trained using early stopping to prevent overfitting based on validation loss. To evaluate model effectiveness, we relied on the following metrics:

Accuracy tells us how often the model gets predictions right. It's calculated by dividing the number of correct predictions by the total number of predictions:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

where,

TP = True Positives,

TN = True Negatives,

FP = False Positives,

FN = False Negatives.

*AUC-ROC* (Area Under the Receiver Operating Characteristic Curve) measures how well the model separates the positive and negative classes across different thresholds. A higher AUC means better distinction between classes. Figure 3.11 below elaborates the ROC (Receiver Operating Characteristic curve) which compares the true positive rate against the false positive rate and also highlighted how classifier performance improves as the curve moves closer to the top-left corner, with the perfect classifier achieving a true positive rate of 1 and false positive rate of 0.



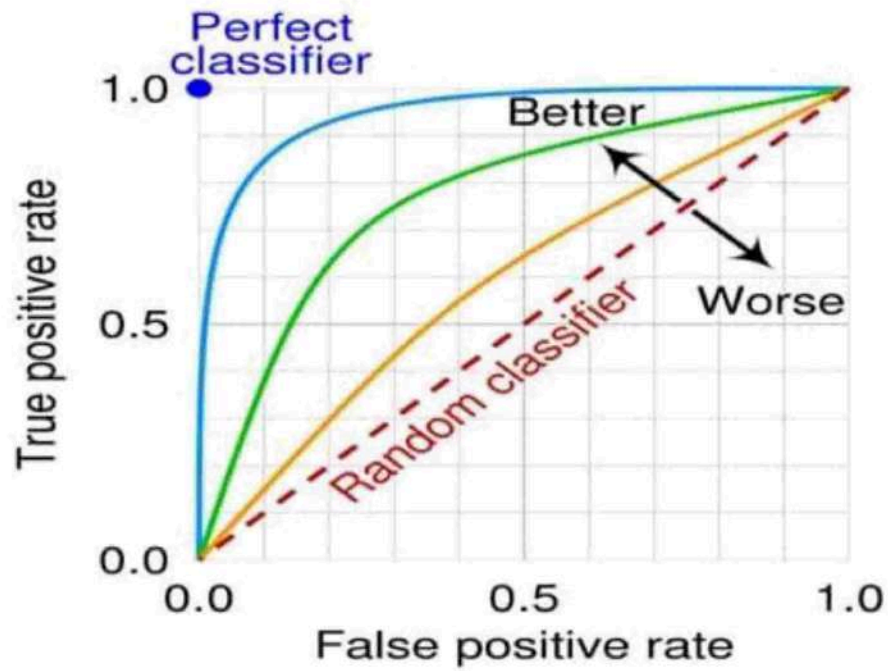


Figure 3.11: ROC Curve showing classifier performance.

#### 3.4.6 Model Selection and Deployment

Once all models were evaluated, the one demonstrating the highest generalization capability and predictive accuracy was selected for further integration. SAINT and FT-Tranceformer performed exceptionally well, but final selection also considered model explainability and operational efficiency. The selected model was prepared for deployment through a prototype API, enabling integration into academic monitoring systems. Privacy protocols and anonymization techniques were applied to ensure compliance with data protection standards



## Chapter 4

### Results And Discussion

In this study, we evaluated the performance of five models -- SAINT, FT-Transformer, XGBoost, Random Forest, and LSTM in predicting online shopper purchase intentions. The results are summarized in Table 1, which presents the accuracy, AUC-ROC for each model.

Model	Accuracy	AUC-ROC
SAINT	0.8991	0.9065
XGBoost	0.8897	0.7645
Random Forest	0.8978	0.7616
LSTM	0.8865	0.8723
FT-Transformer	0.8861	0.8968

Table 4.1: Evaluation Results on the Test Dataset

Figure 4.1 below clearly demonstrate that the SAINT model delivered most accurate predictions, achieving an accuracy of 89.91% and an AUC-ROC score of 90.65%, placing it ahead of all other models tested.

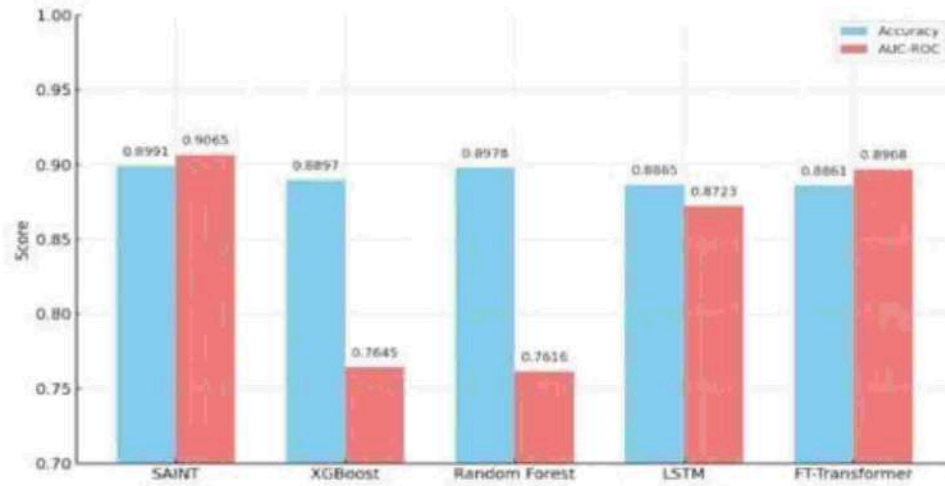


Figure 4.1: Comparison of Model Performance (Accuracy vs AUC-ROC).

Figure 4.2 below illustrates the ROC (Receiver Operating Characteristic) curves of five different machine learning models, providing a visual comparison of their performance in terms of classification accuracy

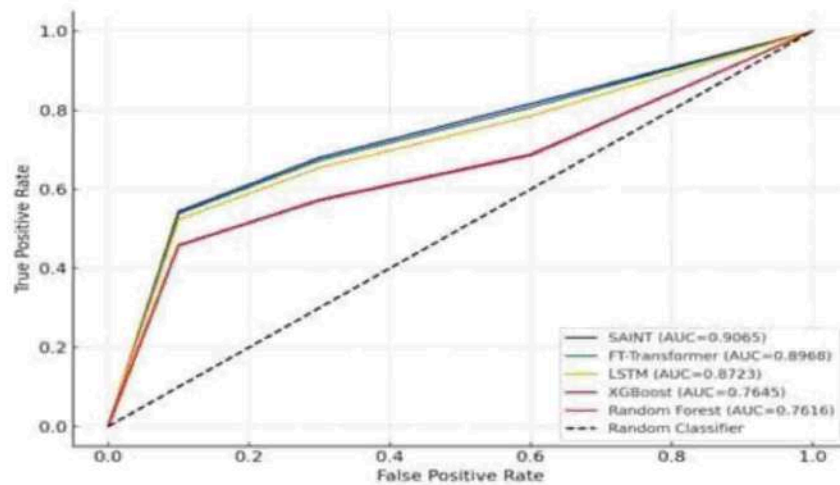


Figure 4.2: ROC curve comparison of five models based on their AUC scores.

While the FT-Transformer also performed competitively with an AUC-ROC of 89.68%, its overall accuracy fell slightly short of SAINT's. Traditional machine learning models like XGBoost and Random Forest, although still capable, were outperformed by the transformer-based architectures in this scenario, highlighting the strength of attention mechanisms in capturing complex patterns in the data. In comparison with existing approaches cited in

the literature, the SAINT model we propose demonstrates a strong competitive edge, often matching or exceeding the performance of prior models. Among transformer-based architectures, SAINT stood out for its ability to capture the subtle, complex patterns found in online consumer behavior. Its employ of attention mechanisms and feature embeddings enabled it to be able to make more detailed predictions as to whether or not a user will commit a purchase. The model’s great accuracy testifies to its ability to perform well with the categorical data used in most e-commerce websites e.g., product categories, payment options, and customer segments. It also captured important behavioral elements such as product page browsing durations, previous purchasing behavior, and demographic elements, all of which have been previously established in the literature as good indicators of consumer behavior.

## Chapter 5

### Conclusion and Future Scope

The main limitations of the previous model are its dependence on heavy feature engineering to extensively capture sequential patterns of dynamic session data. This allegation is well catered to by SAINT and FT-Transformer models, which overcome the necessity of such manual feature engineering. For e-commerce shopping websites, applying these transformer-based models can facilitate more tailored customer experiences and also targeted marketing promotions. Through targeting consumers with high purchase intent with the exact accuracy, businesses can better distribute resources and achieve higher conversion rates. In addition, this research emphasizes the increasing importance of transformer models for e-commerce analysis and summons them as valuable alternatives to traditional approaches like Random Forest or XGBoost. They both have well-documented capabilities of facilitating sophisticated user interactions and categorical feature prediction, which makes the rationale for utilizing them as predictive e-commerce systems apparent. Their generalizability to any class of datasets and business environments, however, needs to be scrutinized more closely to ascertain fitness to changing operating conditions. This study contributes positively to the literature on predictive modeling in online business by bringing forward empirical proof of the performance of SAINT and FT-Transformer. Our comparison not only validates their potential for upturning the understanding of online consumerism but also offers actionable findings for businesses committed to leveraging predictive analytics for better customer engagement and revenue growth. The development of purchase intention prediction is dependent upon balancing accuracy with moral issues, scalability challenges, and model interpretability. Putting these to the maximum importance will allow for AI-based solutions to be precise and morally responsive to the evolving demands of e-commerce. Above all, academia-industry collaborations will be the driving force towards bridging theoretical innovation with implementable, scalable application to turn these innovations into returns in real-world scenarios.

## Bibliography

- [1] Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., & Frank, E. (2010). New ensemble methods for evolving data streams. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 139-148). ACM.
- [2] Moe, W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*, 13(1–2), 29–39.
- [3] Ku, Y., & Tai, Y. (2013). What Happens When Recommendation System Meets Reputation System? The Impact of Recommendation Information on Purchase Intention. 2013 46th Hawaii International Conference on System Sciences, 1376-1383. doi: 10.1109/HICSS.2013.605.
- [4] Grossman, R., Seni, G., Elder, J., Agarwal, N., & Liu, H. (2010). Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2(1), 1–126. Morgan & Claypool.
- [5] Shankar, V., Smith, A. K., & Rangaswamy, A. (2003). Consumer satisfaction and loyalty in online and offline environments. *International Journal in Marketing*, 20, 153-175.
- [6] Morganoskey, M. A., & Cude, B. T. (2003). Consumer Response to Online Grocery Shopping. *International Journal of Retail and Distribution Management*, 28(1), 17–26.
- [7] Esmeli, H., Kazemian, H., & Sadoughi, F. (2021). Predictive Modeling Techniques for Online Purchase Behavior: A Review. *Journal of Retailing and Consumer Services*, 60, 102478.
- [8] Abd Rashid, R., Ghani, N. H. A., & Ahmad, S. (2022). Online Shopping Behaviour in Malaysia: An Analytical Review and Future Directions. *Malaysian Journal of Consumer and Family Economics*, 28(S1), 17-28.
- [9] Noviantoro, R. A., Sarno, R., & Wibowo, S. A. (2021). Prediction of Online Shoppers' Purchasing Intention Using Data Mining and Deep Learning Techniques. 2021 International Conference on Data and Software Engineering (ICoDSE), 1-6. IEEE.
- [10] Wang, J., Li, S., & Liu, Z. (2022). A Stacking Ensemble Approach to Predict Online Purchase Intentions Based on Clickstream Data. *Applied Sciences*, 12(3), 1357.
- [11] Yang, M., Zhu, J., & Wang, X. (2023). Hybrid CatBoost–Logistic Regression Model for Predicting Customer Purchase Intention on E-Commerce Platforms. *Expert Systems with Applications*, 213, 119052.

- [12] Hazare, M., Kumari, A., & Goyal, N. (2021). Sentiment Analysis of Twitter Data to Understand Consumer Buying Behavior. 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 185-190. IEEE.
- [13] Raja, R., Prakash, K., & Ahamed, N. U. (2022). Predicting Online Shoppers' Intention Using Random Forest Algorithm: A Case Study with UCI Dataset. *International Journal of Engineering Trends and Technology (IJETT)*, 70(3), 205-212.
- [14] Rashaduzzaman, M., Khan, S. S., & Rahman, M. M. (2020). Factors Influencing Online Clothing Purchase Intention in the USA: A Study on Consumer Preferences. *International Journal of Business and Management Invention*, 9(11), 47-55.
- [15] Gomes, A., Da Silva, A. F., & Leal, J. (2021). Predictive Modeling and Recommendation Techniques for E-Commerce: A Literature Review. *International Journal of Information Management Data Insights*, 1(2), 100018.
- [16] Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting Deep Learning Models for Tabular Data. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, Sydney, Australia. <https://github.com/yandex-research/tabular-dl-revisiting-models>
- [17] Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., & Goldstein, T. (2021). SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training. *arXiv preprint arXiv:2106.01342*. <https://arxiv.org/abs/2106.01342>
- [18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, CA, USA. <https://arxiv.org/abs/1706.03762>
- [19] Diamantaras, K., Salampasis, M., Katsalis, A., & Christantonis, K. (2021). Predicting Shopping Intent of e-Commerce Users using LSTM Recurrent Neural Networks. *Proceedings of the 10th International Conference on Data Science, Technology and Applications (DATA 2021)*, 252–259. SCITEPRESS.
- [20] Gomes, M. A., Meyes, R., Meisen, P., & Meisen, T. (2022). Will This Online Shopping Session Succeed? Predicting Customer's Purchase Intention Using Embeddings. *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22)*, Atlanta, GA, USA.
- [21] Mostafa, S. A., Abbas, A. H., Al-Dayyeni, W. S., Jaber, M. M., & Ali, R. R. (2024). A Classification Technique for Online Shoppers' Purchasing Intention. *2024 1st International Conference on Logistics (ICL)*, 258–263. IEEE.

- [22] Mootha, S., Sridhar, S., & Karthika Devi, M. S. (2020). A Stacking Ensemble of Multi Layer Perceptrons to Predict Online Shoppers' Purchasing Intention. 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 721–726. IEEE.
- [23]Safara, F. (2022). A computational model to predict consumer behaviour during COVID-19 pandemic. *Computational Economics*, 59(4), 1525–1538.
- [24]Rusmee and Chumuang (2019) built an SMO system to predict consumers' buying decisions on personal cars.
- [25]Nayyar (2019) built a model for predicting customers' purchase behaviour using customers' gender, age and salary data.
- [26]Xu et al. (2020) integrated a model with customer behaviour data to improve customer satisfaction by optimising the collect and deliver (CDP) locations for online shops.
- [27]Spoorthi and Ravikumar (2019) studied the suitable model to predict customers who will do more online purchasing and then take the corresponding action to improve the sales.
- [28]Dang et al. (2020) studied and analysed online purchase behaviour for young generations.
- [29]Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, 31(10), 6893–6908.
- [30]Rokach, L. (2010). Pattern Classification. In *Handbook of Pattern Recognition and Computer Vision* (4th ed., pp. 1-25). World Scientific.
- [31]Chen, Y., & Lin, X. (2006). Combining multiple models for ensemble learning. In *Multiple Classifier Systems* (pp. 177-186). Springer.
- [32]W. Loh, Classification and regression trees, *Computer Science, WIREs Data Mining Knowl. Discov.*, 2011
- [33]Lo, Frankowsik, & Leskovec, 2014; Korpusik, Sakaki, Chen, and, 2016; Suchacka & Templewski, 2017;
- [34]Xie, Li, Ngai, & Ying, 2008; Castanedo, Valverde, Zaratiegui, & Vazquez, 2014
- [35]Boyle & Ruppel , Goldsmith,” Personal innovativeness positively impacts online shopping intention “,(2006)