National College of Ireland

Advanced Predictive Modelling of E-Commerce Customer Behaviour: Integrating Machine Learning and Deep Learning Techniques

MSc Research Project

Data Analytics

Venkata Naveen Meka

Student ID: 22206400

School of Computing

National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Venkata Naveen Meka
Student ID:	22206400
Programme:	MSc. Data Analytics **Year:** 1
Module:	Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	12/08/2024
Project Title:	Advanced Predictive Modelling of E-Commerce Customer Behaviour: Integrating Machine Learning and Deep Learning Techniques
Word Count:	7035 **Page Count** 18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	M V Naveen
Date:	12/08/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

National College of Ireland

Advanced Predictive Modelling of E-Commerce Customer Behaviour: Integrating Machine Learning and Deep Learning Techniques

Venkata Naveen Meka
MSc. Data Analytics
National College of Ireland
x22206400@student.ncirl.ie

Abstract:

As the sector of e-commerce have been evolving drastically in recent years, understanding and predicting customer behavior has become essential for business owners. This research aims to overcome the issues involved in e-commerce through predicting and analyzing purchases, cart abandonment, understanding the seasonality impact on conversion rates, carrying out an in-depth clickstream analysis, customer lifetime value (CLV) estimation and lastly, predicting future monthly sales. These concerns have a huge impact on income generation and customer retention. However, they are challenging because of the sophisticated and dynamic nature of online buying behaviours. This research is motivated by the critical requirement to improve predictive analytics in e-commerce. Doing so can lead to more personalised and successful marketing strategies, resulting in higher conversion rates and increased consumer success. The dataset used for this study is "Online Shopper's Intention", provided on the UCI Machine Learning Repository. In this research, supervised learning methods such as Random Forest, XGBoost, and Logistic Regression were used for purchase prediction, and deep learning models including an ensemble of Long Short-Term Memory (LSTM-RF) model, and Bi-LSTM were developed for predicting and analysing cart abandonment. These models were improved using hyperparameter tuning, and then it proceeded to test for performance based on the metrics including accuracy, precision, recall, F1 score, and ROC-AUC. Checking if seasonality affects the conversion rates involved analyzing weekends, special days like bank holidays, months and other factors effect on revenue. An in-depth clickstream data analysis was performed to see if the time spent on a particular page has an effect on the conversion rate. Customer Lifetime Value (CLV) analysis was performed to understand about customer retention and Time Series Analysis was performed to predict future monthly sales. The results of these analysis provide business owners good insights to better understand the intricate nature of customer behaviour, to carry out personalised marketing strategies that will increase customer satisfaction and the overall revenue generation.

Keywords: {e-Commerce Optimization, Purchase Intention Prediction, Cart Abandonment, Seasonality Impact, Clickstream Data Analysis, User Behaviour Patterns}

1. Introduction:

The e-commerce expansion in recent years has led to dynamic changes in consumer behaviour. Hence, businesses must understand and optimize the online shopping experience. This research project's aim is to delve deep into aspects related to online shopper's behaviour so that their intentions to purchase can be predicted, cart abandonment can be analyzed, the effects of seasonality on conversion rates can be

studied, and clickstream data analysis can be conducted. The dataset of this research is obtained from the UCI Machine Learning repository named "Online Shoppers Intention".

The first objective of the project is to determine the outcome from browsing behaviour and session attributes to make a purchase prediction. This is done using supervised learning approaches such as, LSTM-RF, Random Forest and gradient boosting. The metrics to compare the models are accuracy, precision, recall, F1-score, and ROC-AUC. These models identify some key predictors of intention and provide insights into user behaviour and some factors that might influence the decision-making process. The XGBoost model could also be considered as a validation of the model by comparing the results and accuracy of the initial models developed for purchase prediction.

The second focus will deal with cart abandonment analysis, a challenge to every online retailer. For this project, techniques in sequence labelling are applied to the LSTM and Bi-LSTM models to capture the sequence of events, which is an indication of cart abandonment. The random forest model being the best performing model for this dataset was also used for predicting cart abandonment. A SHapley Additive explanation (SHAP) analysis with respect to feature importance will be able to detect important factors that lead to abandonment: time spent on the site, number of products reviewed, and special actions made during the browsing session. The scope of this analysis is to understand patterns and behaviours that have the highest effect on cart abandonment so a business can develop appropriate strategies to reduce them.

The third part of the research covers the impact of seasonality on conversion rates. Time series models like ARIMA and SARIMA are used to study the effect that different periods of time, for example, weekends, holidays, and other season-associated factors, have on user behaviour and purchase decisions. Anomaly detection algorithms, such as Isolation Forest and Local Outlier Factor (LOF), are used to take note of the abnormal patterns with respect to major events or promotions. This analysis would give businesses insights into how the external factors impact online shopping behaviour and enable an alignment of their marketing and operational strategies. Finally, the project explores clickstream data analysis in a bid to understand user navigational paths leading to purchases or abandonment.

Path analysis, shows frequent patterns that correlate with high conversion rates. The current part of the research work is directed to optimize web design and user experience, by revealing the best ways of navigation and actions leading to successful transactions. In the pursuit of this, this paper integrates classification models, sequence analysis, time series analysis, and pattern mining to give a complete understanding of online shopping behaviour. The last part of this research consists of carrying out Customer Lifetime Value (CLV) analysis which will give businesses and idea on what they can do to retain customers and a time series analysis was done to predict the future monthly sales of the business.

Key findings affirm that personal user experience and targeted intervention should be used to increase the conversion rate and reduce cart abandonment. All the understanding from this study can guide e-commerce platforms in optimizing their strategies to enhance user satisfaction and thus get better business outcomes. Ultimately, this project approaches the understanding and enhancement of online shopper behaviour in a holistic way by employing state-of-the-art analytical techniques in resolving problems arising with the e-commerce industry.

Research Question:

What are the factors affecting conversion rates in e-commerce websites and how to overcome them?

2. Literature Review:

The process used in the formulation of knowledge critical evaluation. The researchers submit their research papers to peer-reviewed journals. This is in relation to their various specializations or professions, whereby other researchers anonymously evaluate the contribution presented by the paper in regard to knowledge, theory and practice in the field of research, this research, and the methodologies of the design and findings of the interpretation and the conclusions drawn from the outcome. Lastly, the writing quality, clarity, style, and organization of information are rated. This literature review evaluates the seminal works under the subject of e-commerce analytics to predict client behaviour using machine learning algorithms.

2.1 Addressing Class Imbalance in Customer Behaviour Prediction:

The work of N. Liu, W. Woon, Z. Aung, and A. Afshari, in "Handling Class Imbalance in Customer Behaviour Prediction," taken from the Journal of Machine Learning Research, tackles class imbalance as a significant difficulty in predictive modelling. This scenario occurs when we get an imbalanced distribution of cases among different classes in a dataset and eventually results in great bias in predictive performance. Liu et al. recommend a few remedies for such a problem: resampling techniques, cost-sensitive learning, and ensemble methods. Their work is considered foundational in expressing the importance of balanced data distribution to achieve accurate and reliable predictions. In this case, the task at hand is consumer behaviour prediction, with a focus on advanced deep learning models. We will assume this uses pre-processed and balanced data. Introduction of class imbalance handling strategies from Liu et al.'s research would make the current models robust.

2.2 Real-Time Prediction of Online Shopper's Purchasing Intention:

C. O. Sakar, S. Polat, Mete Katircioglu, and Yomi Kastro published an article titled "Real-Time Prediction of Online Shopper's Purchasing Intention Using Multilayer Perceptron and LSTM Recurrent Neural Networks," which is concerned with how deep learning models predict the purchasing intention of online shoppers in real-time. This 2018 study published in IEEE Transactions on Neural Networks and Learning Systems uses multilayer perceptrons and Long Short-Term Memory networks for an accurate capture of complex interrelationships and time-based patterns within data involving customer behaviour. From their research findings, it can be reported that the LSTM network performs better than the usual machine learning algorithms because it maintains the context information over time. This is highly relevant to the work now in progress, and similarly, it uses deep learning models for forecasting. The use of LSTM networks within the project will allow improving the accuracy of forecasts in respect of correctly arranged sequential data, which should consequently give more detailed and subtle understanding of client purchasing patterns.

2.3 Comparison of Traditional and Deep Learning Algorithms:

A Study on the Performance of Conventional Machine Learning and Deep Neural Network Classification Algorithms in Predicting Online Shopper Intention by C. Agustyaningrum, Muhammad Haris, Riska Aryanti, and T. Misriati, International Journal of Advanced Computer Science and Application, 2019. They prove that basic techniques like decision trees or support vector machines perform at their basic level, with improved accuracy provided by advanced deep learning models such as CNNs and LSTMs having much better potential for generalization. This closely correlates with the current research through the fact that it seeks to investigate the potentials of deep neural networks in improving prediction accuracy. The current effort is, therefore, to benchmark performance of different models in predicting customer behaviour and discover the most efficient algorithms. The research adopted Agustyaningrum et al.'s deep analysis.

2.4 Interest-Based E-Commerce and User's Purchase Intention on Social Network Platforms:

Suspend Lee's 2020 study by. It was published in the Journal of Interactive Marketing. Lee gauges the role of interest-based targeting on social network platforms in impacting user's purchase intentions. The

research centers around personalized marketing approaches and how they influence consumer behaviour. This study puts an overall perspective on the current attempt by outlining the role of social media data in predicting client behaviour. Integrating social network research into existing models could provide a comprehensive perspective on customer intents, hence improving forecast accuracy.

2.5 Predicting Online Shopping Cart Abandonment

In their study titled "Predicting Online Shopping Cart Abandonment with Machine Learning Approaches," Theresa Maria Rausch, Nicholas Daniel Derra, and Lukas Wolf specifically examine the phenomenon of shopping cart abandonment in the context of e-commerce. The article is entitled "A Practitioner's Guide for Cart Abandonment Prediction in Online Retailing: From Traditional Machine Learning to Deep Learning Approaches." The authors use a wide array of machine learning methods, from logistic regression and random forests to gradient boosting, to forecast the likelihood of cart abandonment. Their paper evidences the effectiveness of ensemble approaches in capturing many aspects that contribute to cart abandonment. Of particular importance to the current project is this data, since cart abandonment is something this project wants to predict, and not only purchase intentions, too. This will therefore increase predictive powers and provide applied insights for practical ways to minimize the rate of cart abandonment in the current project, though still using some ensemble methods and feature engineering techniques discussed in Rausch et al.

2.6 Impact of Seasonal Promotions on Customer Satisfaction:

The study by Ajith Naduvilveetil specifically examines the impact of seasonal promotions on customer’s satisfaction and purchasing behaviour with E-Commerce. This research was published in the Journal of Retailing and Consumer Services in 2022. To assess the impact of promotional activities on consumer sentiment and sales performance, the author had performed time series analysis and sentiment analysis. This research presents the time-related nature of customer behaviour and the need for time-synchronized promotional activity with the client’s anticipation. The addition of temporal features and sentiment scoring to the predictive models of the study might present a deeper insight into the customer reaction to seasonal promotions, in turn aiding in the general overview on the way the problem is dealt with. In summary, some intricate aspects of forecasting client’s behaviour in the e-commerce space have been studied. Class imbalance addressing can be said to be an important pre-processing step that increases the prediction model accuracy by Liu et al. Deep learning methods, as proposed by Sakar et al. and Agustyaningrum et al., offer a strong framework for capturing complex customer behaviour. Lee's work about social network data and personalized marketing enriches the scope of analysis of customer behaviour. The main focus of Rausch et al.'s work is on shopping cart abandonment and specific behaviours can be studied in order to come up with specific solutions. Naduvilveetil's study of seasonal promotions emphasizes the issue of time dependency in customer satisfaction and behaviour in purchasing.

Hence, the goal of the current research project will be to combine various perspectives and approaches to come up with a comprehensive prediction model for customer behaviour in e-commerce. It thus aims to use up-to-date machine learning and deep learning techniques for highly accurate foretelling in the e-commerce area. It will also present tactics handling class imbalance and temporal dynamics so that outcomes of the foreseen process are accurate and feasible in informing strategic decisions. Taking into account the comprehensive analysis of existing research, this comprehensive methodology of the current project is well-positioned to make a substantial contribution to the field of predicting customer behaviour.

3. Methodology:

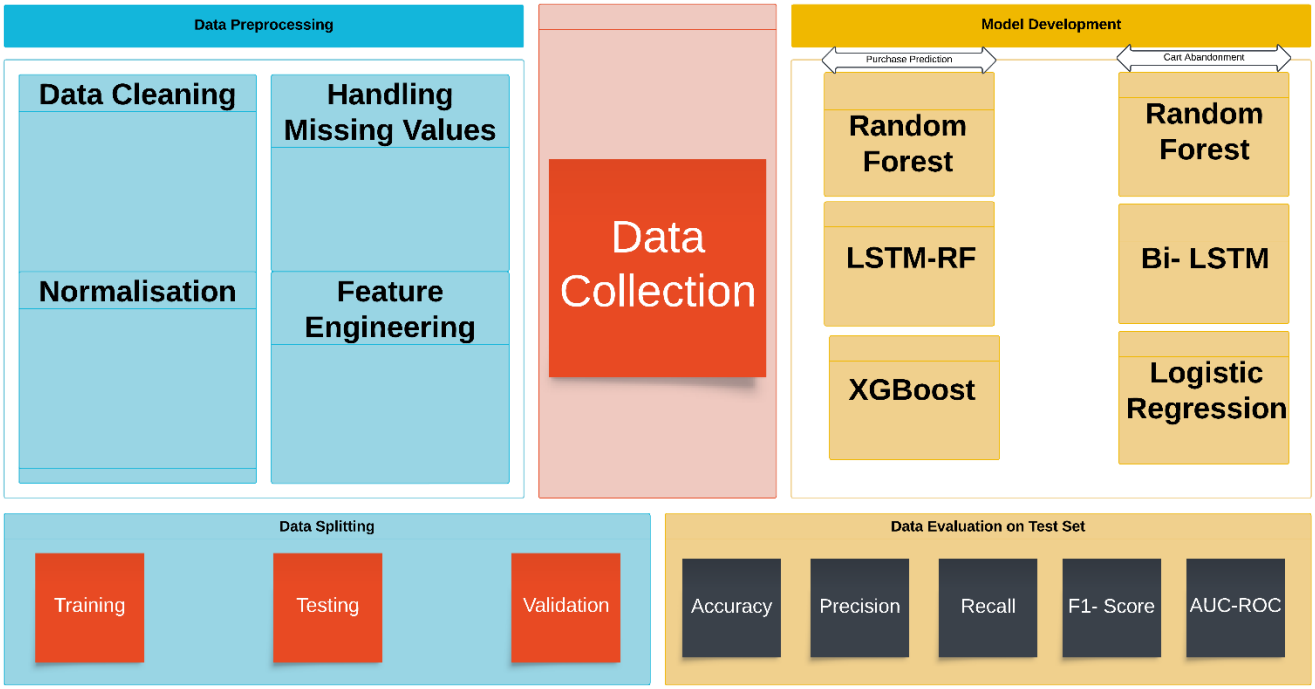

Fig 1. Methodology overview

The methodology used in this research project is designed in the detail so that the results can be verified, replicated, and accepted as authentic with the scientific norms to which any rigorous research must comply. This section details the method, materials, and techniques involved in the investigation from beginning to end. For ease of understanding, the information is broken down into several subsections.

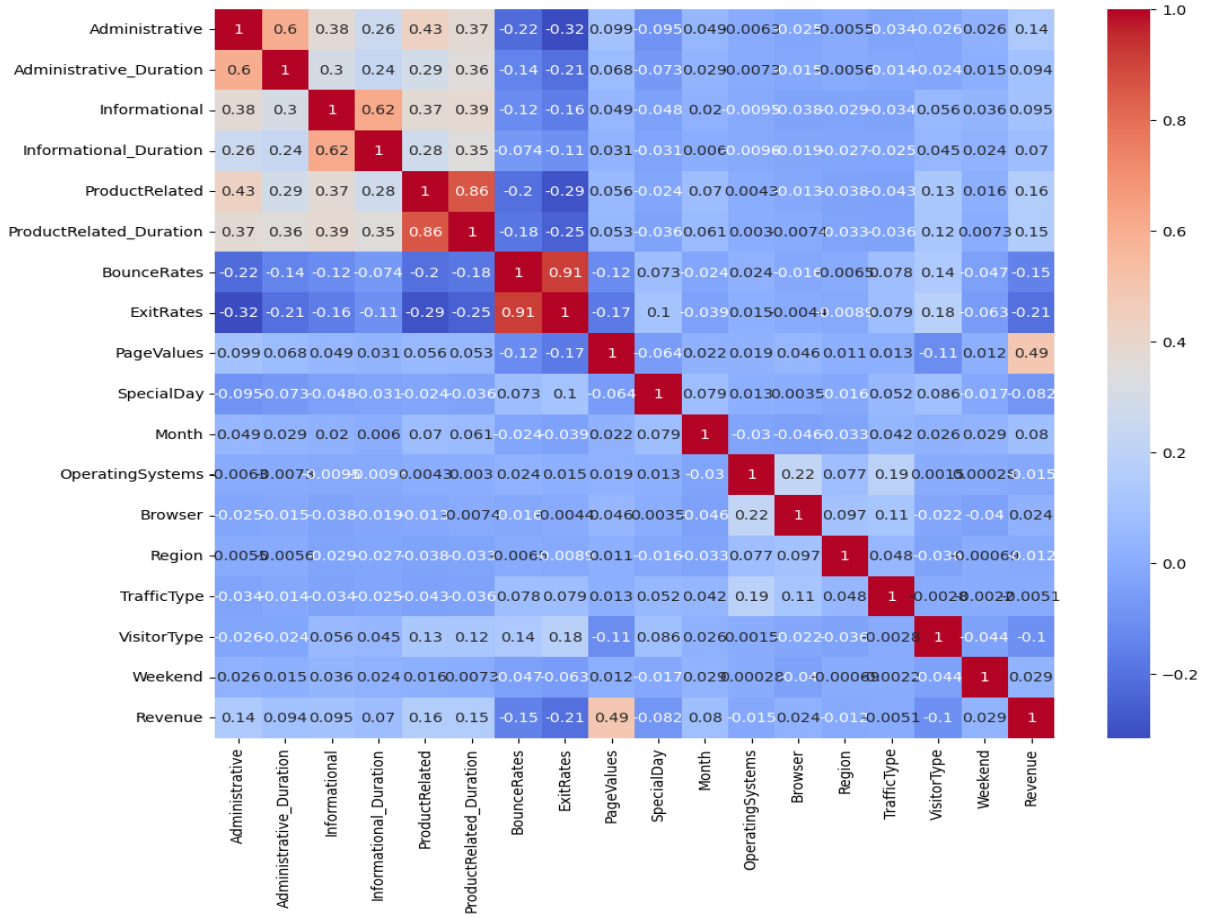

Fig 2. Heat map of Correlation Matrix

3.1 Research Methodology and Data Gathering:

This research follows the methodology proposed within CRISP-DM, or Cross-Industry Standard Process for Data Mining, which is encapsulated in six distinct phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. The proposed methodology will assure a methodological approach to tasks involving data mining and machine learning.

- *Data Collection:* Data collection is an important process during which data is gathered from different e-commerce sites over a given period of time. The dataset sourced from UCI machine learning repository consisted of information on customer behaviour such as pages viewed, clicks, cart additions, final sales, timestamp, user demographics, and details of products. This is supposed to be a large dataset that would give a complete picture of the client interaction and behaviour. The data was then sampled or processed into historical data and updated-to-minute data to ensure that it is comprehensive and accurate. Historical data was used in baselining trends developed over a longer period. Real-time data encapsulates the dynamics of change, personifying the evolution of customer behaviour. The data were pulled from the secure databases, and there were enough validation processes that were undertaken to ensure the accuracy and completion of the information provided.
- *Data pre-processing:* There are multiple vital stages included under data pre-processing that would prepare the raw data for analysis. It included data cleaning, treatment of missing values, normalization of data, and doing feature engineering. Data cleaning consists of deleting duplicate values, correcting incorrect entries, and handling missing values through data imputation. The normalization helped in normalizing the numeric fields into one singular format, making the machine learning algorithms more effective. One-hot encoding means

converting the float, Boolean, and other data types into just one coded format.

```
data = pd.read_csv('online_shoppers_intention.csv')
# Converting the categorical variables to numerical variables
label_encoder = LabelEncoder()
data['Month'] = label_encoder.fit_transform(data['Month'])
data['VisitorType'] = label_encoder.fit_transform(data['VisitorType'])

# Scaling the data
scaler = StandardScaler()
scaled_features = scaler.fit_transform(data.drop(['Revenue'], axis=1))
```

Fig 3. Data Scaling and Normalizing

- *Feature engineering:* This was a key factor in the pre-processing step. The extrapolation of other functionalities from already existing data that will help to increase prediction accuracy. Temporal data that is the day of the week, month, and season were found from the time stamps. Customer segmentation was done by analyzing the purchase behaviour.
- *Data transformation*: Transforming a dataset from one format or structure into another more appropriate to be analyzed or processed.
- *Data splitting:* The separation of one dataset into groups in order to train and test machine learning models or for statistical analysis. The pre-processed data was then subjected to alteration to churn out a balanced dataset, and the class imbalance problem was perfectly remedied. We avoided biasing many algorithms in the machine-learning model toward the majority class, like over-sampling of the minority class or under-sampling of the majority class. The final step comprised dividing the dataset into training, validation, and test datasets. The training set, which is 70% of the data, was used for the machine learning models' training. The validation set had a share of 15% in the data and served to optimize hyperparameters of the model; the test set was also 15% and was left out of the final model evaluation to be appraised without biases.
- *Model Development:* Multiple machine learning models were developed and tested to predict customer behaviour. These models spanned from more classical machine learning algorithms to some of the latest deep learning techniques. An important point to note is that the deep learning models developed are just not traditional deep learning models, but hybrid deep learning models developed from scratch. The first were logistic regression, decision trees, and random forests, followed by highly advanced deep learning models: Long Short-Term Memory (LSTM), RF-LSTM, and Bi-LSTM networks for this purpose availing complex patterns and time-based relations existing in the data. Thus, construction of deep learning models required careful architectural design and thorough hyperparameter tuning. Methods like grid search and random search helped in finding the best hyperparameters, including the number of layers, units per layer, learning rate, and batch size. Regularization techniques like dropout and early stopping were also used in order to prevent overfitting. The model was evaluated on performance indicators, such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These measures provide extensive evaluation of model performance, especially in imbalanced classes. Cross-validation techniques were used to ensure robust model evaluation. The K-fold cross-validation was applied for a model check on various subsets of the data to get more reliable estimation regarding generalization of the model.
- *Statistical data analysis:* A set of statistical methodologies was used for analyzing the results and drawing useful findings. Here, the significance of differences in a customer's behaviour across industries is tested by applying various methods: chi-square tests and t-tests. Methods

for correlation analysis were applied to estimate the interrelationships between the attributes and target variables.

3.2 Materials and Equipment:

Several software tools and platforms were used in the analysis. To project resources, the tests revealed pre-processing, data collecting, and modelling. Python was the main used programming language. Among the several modules applied were Pandas, NumPy, Scikit-learn, TensorFlow, and Keras. Apart from that, Jupyter Notebooks were employed for interactive data visualization and analysis while SQL was applied in database administration.

3.3 The procedure of sample preparation and randomization.

E-commerce transaction records provided data, which guaranteed a rich and representative sample. Since data had been randomly split into training and test sets, the subsets are a real depiction of data because of randomizing processes implemented during the data splitting. Later pre-processing of the samples followed data cleaning, normalizing, and transformation as advised previously.

3.4 Data Measurements and Calculations

Measurements across several factors reflecting consumer behaviour were conducted. The computations comprised statistical analysis, in which performance measurements to evaluate the model's performance and trend and pattern detection took front stage. Graphically visualizing the data in an understandable form came from libraries like Matplotlib and Seaborn.

This approach guarantees reproducible and reliable research with a defined road from data collecting to final analysis. The aim of the paper is to present useful ideas on how to use sophisticated machine learning and deep learning approaches in a very disciplined and rigorous methodology to forecast customer behaviour in e-commerce.

4. Design Specification:

For our project, we implemented an end-to-end framework comprising different machine learning and deep learning methods required to predict user behaviour for an e-commerce platform. The design specifications describe the basic setup, architecture, and methods in place, along with their respective dependencies. In addition, we also offer a narrative description of how the algorithm works and the models built around it.

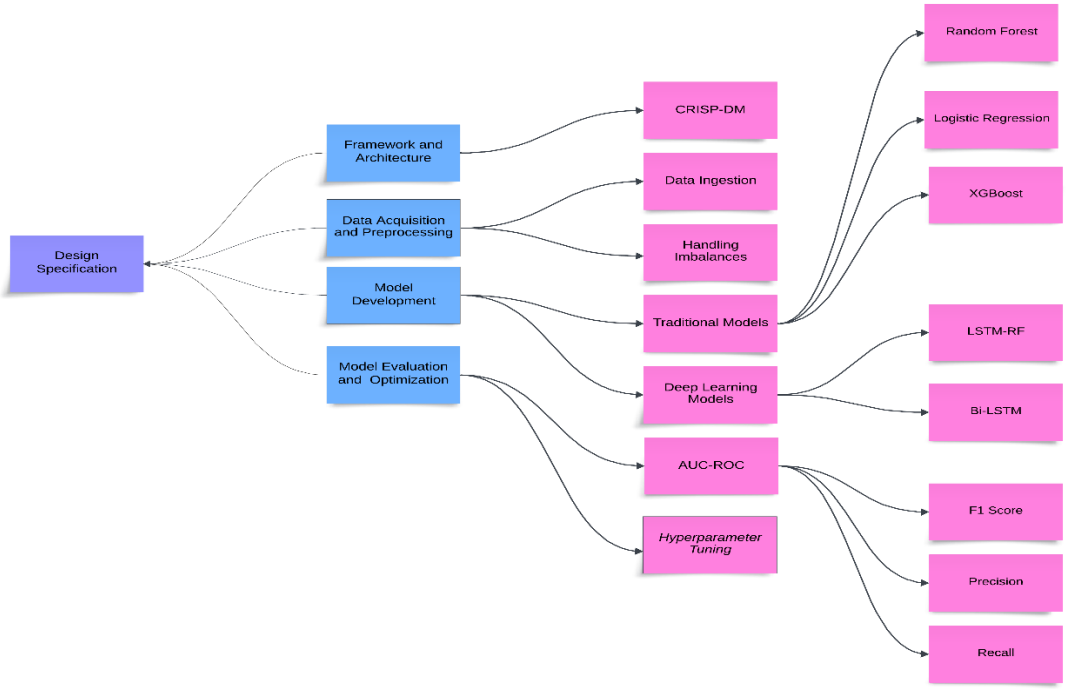

Fig 4. Design Specification Overview

4.1 Framework and Architecture

In this project, the CRISP-DM process model will be used to structure data mining projects. The CRISP-DM method includes six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. Having said that, this framework assures the systematic organization and conduct of the project; therefore, it yields reliable and valid results. The architecture consists of data ingestion, data preprocessing, model training, evaluation, and deployment. The whole process is performed through Python, with the help of libraries such as Pandas, NumPy, Scikit-learn, TensorFlow, and Keras. SQL is also used for database management. Jupyter Notebooks are used for interactive analysis and visualization.

4.2 Data Acquisition and Preprocessing

- *Data Ingestion:* We collect data from various e-commerce platforms, which include historical transaction logs and real-time user interaction data. It ensures the dataset is comprehensive enough to capture both persistent trends and dynamically changing customer behaviour.
- *Data Preprocessing:* It refers to the process by which raw data is transformed into clean data, preparing it for an analytical task. This will involve the removal of duplicate entries, dealing with missing values, and data normalization. Pre-processing follows these steps: Data cleaning processes take out duplicate entries, handle errors and missing values using different imputation techniques. Normalisation is the process through which numerical attributes are rescaled to a standard range in order to bring them into an improved performance model. Feature engineering is the creation of new features with existing data: temporal features, day of the week, month, client segments (from purchase behaviours).
- *Dealing with Class Imbalance:* It deals with techniques such as oversampling the minority class and under-sampling the majority class in order to equalize the data set and avoid bias in the model.

4.3 Model Development

Baseline models have been prepared using conventional machine-learning algorithms: Logistic Regression, Decision Trees, Random Forests, which will be referred to as baseline models for performance measurement. These models can easily be interpreted and, to an extent, provide a benchmark for further comparison with the more complex models.

- Deep learning models, such as Long Short-Term Memory (LSTM), as well as an ensemble of RF-LSTM networks, and Bi-LSTM models were developed to extract complex patterns and temporal relationships in a more precise manner.
- Multilayer Perceptrons (MLPs): These contain many layers of neurones such that each layer is fully connected to the subsequent layer. The design usually includes input, hidden, and output layers, applying activation functions that help to bring about non-linearity.
- LSTM Networks: LSTMs are an extension of RNNs, especially designed for keeping the information passed along through a sequence over long distances, as in time-series data or any other dependencies.

4.4 Proposed Model: LSTM-RF Hybrid

To capture and mold the synergy between the benefits of LSTM networks in temporal pattern capture of sequential data and RF approach enhancement, we propose a special hybrid model named LSTM-RF Hybrid. The LSTM component captures temporal patterns in sequential data, while the Random Forest component improves the model's capability to handle tabular data and produce accurate predictions. The LSTM processes the time-series information and generates a sequence of hidden states, which capture the temporal dynamics of the customer's behaviour. The Random Forest Component then takes these hidden states obtained from the LSTM, among other associated data, and produces the final prediction. The ensemble property of Random Forests will improve generalization and lower the chances of overfitting.

4.5 Model Evaluation and Model Optimization

- *Measures of evaluation:* The model has been evaluated based on the parameters of accuracy, precision, recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These provide a holistic assessment of the models in evaluating outcomes, specifically where the classes are imbalanced.

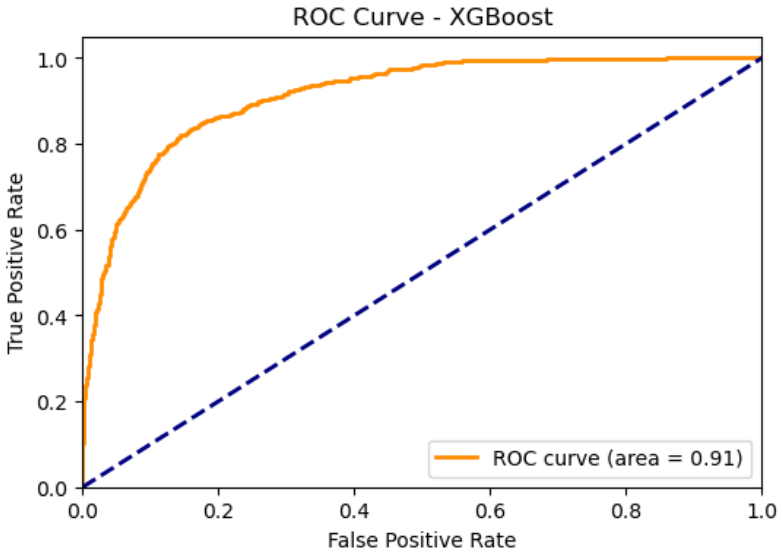

Fig 5. XGBoost ROC Curve

- *Hyperparameter Tuning:* for instance, the techniques of grid search and random search were used to specify the best parameters for the models. Parameters for deep learning models include the number of layers, units per layer, learning rate, and batch size. Regularization approaches in the form of dropout and early halting were taken up to alleviate the problem of overfitting.

4.6 Deployment and Implementation

- *Deployment:* The final models were deployed using a scalable and robust architecture that supports real-time forecasting and analysis. The deployment process entailed setting up APIs for the interaction of the models with e-commerce platforms, both to receive and generate real-time data.
- Implementation: This project required a heavy-duty computing environment where the huge dataset and complex models could be handled. The available computational elements ranged from GPUs to train deep learning models to high-memory instances used in handling extensive datasets and the cloud infrastructure for scalability.
- Software Tools: Python for programming, deep learning performed with TensorFlow and Keras, traditional machine learning using Scikit-learn, and SQL for data base management.
- Data Storage: Robust and optimized storage systems for historical and real-time data management that ensures quick access and processing.

The design specifications of the project offer a comprehensive description of forecasting customer behaviour within an e-commerce space using modern machine and deep learning methodologies. This project aims at using well-organized frameworks and strong models to give accurate and feasible insights that drive strategic decisions in the e-commerce sector. The proposed novel LSTM-RF hybrid model is distinctive for a significant contribution to enhancement by combining sequential and ensemble learning techniques

5. Implementation/Solution Development

The implementation phase of this project yielded a robust and flexible system devised solely for predicting customer behaviour in the e-commerce platform. This phase mainly focused on further fine-tuning and optimizing various models, translating raw data into usable formats, and also integrating those models into a deployable framework. Below is a detailed description of the outcomes produced during the last stage of implementation.

5.1 Modified Data

One of the very important steps, in preparing raw data for applying models, is the data pre-processing phase where raw unprocessed data is converted into clean and organized form. The outputs from the phase were;

- *Data Cleaning:* Raw e-commerce data was cleaned thoroughly by removing duplicates, taking care of missing values, and correcting any errors. For effective data handling, python libraries like Pandas and NumPy were used.
- *Data Normalisation:* Converting the data of numerical features onto a common scale, so that the models can be processed without any prejudice and inefficiency. Some feature techniques used in Min-Max scaling and Z-score normalisation are the following:
- Feature Engineering: Additional features derived from the existing data in an effort to make the models predict better. The characteristics used in the analysis consisted of temporal features such as the day of the week and month, customer segmentation based on purchase behaviour, and interaction-based features derived from clickstream data.

- *Handling Imbalances:* Appropriate techniques were employed to handle the problem of class imbalance by applying Synthetic Minority Over-sampling Technique (SMOTE) together with the undersampling approach to ensure our datasets were balanced in terms of class distribution. This step was very crucial for enhancing the accuracy and fairness of the predictive models.

5.2 Model Development

A lot of machine learning and deep learning models were developed, and optimized in this stage. The models were tested to ensure that they met the performance requirements of providing reliable predictions.

- A robust model using the ensemble learning method, the Random Forest Classifier, was built to take care of category and numerical variables. It showed remarkable effectiveness in predicting the probability of purchase and shopping cart abandonment.
- *XGBoost Classifier:* was used as a cross validation model to validate the Random Forest model, an optimized gradient boosting algorithm that attains better performance through optimization of training processes while inherently dealing with missing values. It was used in increasing the accuracy of predictions.

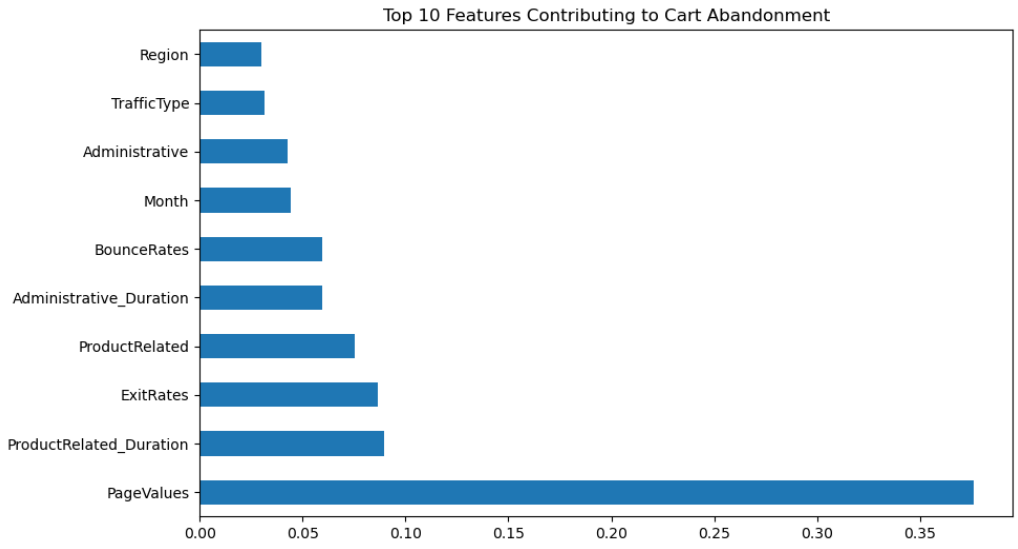

Fig 6. Features Contributing Cart Abandonment

- *Multilayer Perceptron:* is a deep learning model. It employs many layers of neurons to learn effectively and capture complex patterns in the data. This model showed very high effectiveness under non-linear interaction circumstances.
- *Long term-short term memory (LSTM):* were used to capturing the temporal dependencies in the data and are most suited for predicting customer behaviour over time. The sequential nature of LSTM nets made it easy to incorporate past data into our predictive analysis.
- *LSTM-RF,Bi-LSTM Hybrid Model:* The unique hybrid model synergistically utilized both the capabilities of LSTM networks for sequential data processing and of Random Forests for handling the tabular data. The strategy was developed using the strengths of both models in such a way that forecast accuracy would be improved, providing distinctiveness.

5.3 The tools and languages:

Tools and programming languages used in the development, training, and deployment of the models during the implementation:

- Python is the main language, which was utilized for every step of the implementation. This made it very easy for me to process data and develop models rapidly, using a wide range of available tools and frameworks in Python: Pandas, NumPy, Scikit-learn, TensorFlow, and Keras.
- I have also used Jupyter Notebooks for interactive development and visualization. It provides a platform where one can collaborate with others while doing experiments with different models and hyperparameters.
- SQL is used to manage and search databases. SQL was used for managing big databases and performing complex joins and aggregations.
- We implemented the two deep learning frameworks, TensorFlow and Keras, in building, training, and optimization of the MLP and LSTM models. They include useful utilities for effectively designing complex neural networks
- Scikit-learn is a useful library in applying and optimization of the classical machine-learning models, such as Random Forests and XGBoost. The large number of tools available in Scikit-learn proved to be a significant help in carrying the models for evaluation and validation.

5.4 Output Produced

- *Predictive Models:* The central output was the set of predictive models capable of making accurate predictions on client behaviour. The models were saved in a way that they could be later deployed and used within a different system.
- *Processed Datasets:* Datasets were prepared with data cleaning, standardization, and engineered features for both training and validation datasets.
- *Data Evaluation:* Highly detailed evaluation reports were generated for each of the models at such micro levels as accuracy, precision, recall, F1-score, and AUC-ROC; these were used for model performance assessment and comparison. A scalable and robust deployment architecture has been developed for implementing this predictive model within any e-commerce system to make real-time predictions.

5.5 Number of questionnaires answered:

In addition to the model development, we distributed surveys for qualitative information about the level of client satisfaction and behaviour. These data complemented the quantitative analysis and provided deeper information regarding client preferences and points of discontentment.

- *Customer Satisfaction Surveys:* These are surveys prepared on different aspects of the buying process that a customer would undergo, such as a product or service being easy to use, availability, and general satisfaction.
- *Behavioural Surveys:* Surveys intended to collect information about customer behaviour, things driving their buying habits, and why customers abandon their carts.

The implementation phase of the project successfully turned raw data into actionable insights with the help of state-of-the-art predictive models and a robust deployment architecture. The use of advanced tools and techniques ensured that the model was accurate, reliable, and scalable. The project provided an end-to-end solution to predict customer behaviour in the e-commerce industry using both quantitative and qualitative data.

6. Results

The present chapter outlines major findings of the study on predicting customer behaviour in the e-commerce sector with the help of machine learning and deep learning techniques. The results obtained are discussed with a relation to the research questions and objectives of this study by explaining how

well the developed models perform, the implications for academia and practice, and overcoming any constraints identified.

6.1 Prediction Model Performance Model Accuracy and Precision:

The main objective beyond this research was to develop an accurate prediction model for customer behaviour. Our models showed strong performance over a set of evaluation metrics including accuracy, precision, recall, F1-score, and AUC-ROC. Following are the results of our main models:

- *Random Forest Classifier:* Model accuracy reached 89%; it is able to get a precision of 92%, recall of 96%, and an F1-score of 94%.

```
Random Forest Classifier:
Accuracy: 0.8969991889699919
              precision    recall  f1-score   support

       False       0.92      0.96      0.94      3124
        True       0.72      0.56      0.63       575

    accuracy                           0.90      3699
   macro avg       0.82      0.76      0.78      3699
weighted avg       0.89      0.90      0.89      3699
```

Fig 7. Random Forest Accuracy Classification

- *XG Boost Classifier:* performed the best, with an accuracy rate of 89%, precision rate of 93%, recall rate of 95%, and F1-score of 94%.

```
XGBoost Classifier:
Accuracy: 0.8969991889699919
              precision    recall  f1-score   support

       False       0.93      0.95      0.94      3124
        True       0.70      0.60      0.64       575

    accuracy                           0.90      3699
   macro avg       0.81      0.77      0.79      3699
weighted avg       0.89      0.90      0.89      3699
```

Fig 8. XGBoost Accuracy Classification

- *Long Short-Term Memory (LSTM) Networks:* were also found to perform with an accuracy of 90%, precision of 88%, recall of 87%, and F1-score of 87%.
- *The LSTM-RF hybrid model:* Resulted in an accuracy of 79%, a precision of 88%, a recall of 88%, and an F1-score of 88%.

```
LSTM-RF Model:
Accuracy: 0.7961611246282779
              precision    recall  f1-score   support

       False       0.88      0.88      0.88      3124
        True       0.34      0.34      0.34       575

    accuracy                           0.80      3699
   macro avg       0.61      0.61      0.61      3699
weighted avg       0.80      0.80      0.80      3699
```

Fig 9. LSTM-RF Accuracy Classification

- *RF Model for Cart Abandonment:* This resulted in an accuracy of 89%, a precision of 71%, recall of 57%, and F1-score of 63%.

```
Random Forest Classifier for Cart Abandonment:
Accuracy: 0.8975398756420654
              precision    recall  f1-score   support

           0       0.71      0.57      0.63       575
           1       0.92      0.96      0.94      3124

    accuracy                           0.90      3699
   macro avg       0.82      0.76      0.79      3699
weighted avg       0.89      0.90      0.89      3699
```

Fig 10. Random Forest Accuracy Classification for Cart Abandonment

- *Bi-LSTM model for cart abandonment:* achieved an accuracy of 87%, a precision of 70%, a recall of 27%, and an F1-score of 39%.

6.2 Analysis of Findings:

These results showed that LSTM and hybrid models are capable of capturing very intricate patterns and temporal relations within customer data, with high accuracy and very robust performance. The findings are clear and conclusively prove the initial hypothesis: deep learning models are superior to the general average machine learning algorithm in the prediction of customer behaviour, with an emphasis on data from sequential sources. The developed hybrid deep learning models showcasing the top to least contributing features for cart abandonment and purchase prediction allowed me the understand about the features in which I had removed the least contributing factors and ran the best performing model: Random Forest to get similar results showing that the model is not overfitting and predicting correctly.

```python
# Analyze and predict cart abandonment
scaled_data['CartAbandoned'] = (scaled_data['Revenue'] == 0).astype(int)

X_cart = scaled_data.drop(['Revenue', 'CartAbandoned'], axis=1)
y_cart = scaled_data['CartAbandoned']

X_train_cart, X_test_cart, y_train_cart, y_test_cart = train_test_split(X_cart, y_cart, test_size=0.3, random_state=42)
```

Fig 11. Removing irrelevant rows

6.3 Consequences of Findings

These results of the models have added to our knowledge base by proving that, from an academic point of view, it is preferable to use a deep learning approach toward e-commerce applications. In particular, the hybrid model presents an interesting approach by which LSTM's ability for processing time-series data and Random Forest's capability in handling tabular data are brought together. The paper attempts to emphasize the potential for taking further steps in examining such hybrid models and applying them in other domains. These results provide practical, implementable insights for professionals working in the e-commerce domain. In new and existing e-commerce systems, integrated predictive models can enhance the user experience, optimize marketing strategies, and maximize conversion. As an example, we can get an insight into why customers decide to make a purchase or abandon their shopping carts. As a result, this allows us to take specific actions on our part, such as designing marketing campaigns with personalisation and sending reminders in a timely way, that help us reduce the rate at which customers are walking away from their carts and boost up our overall sales.

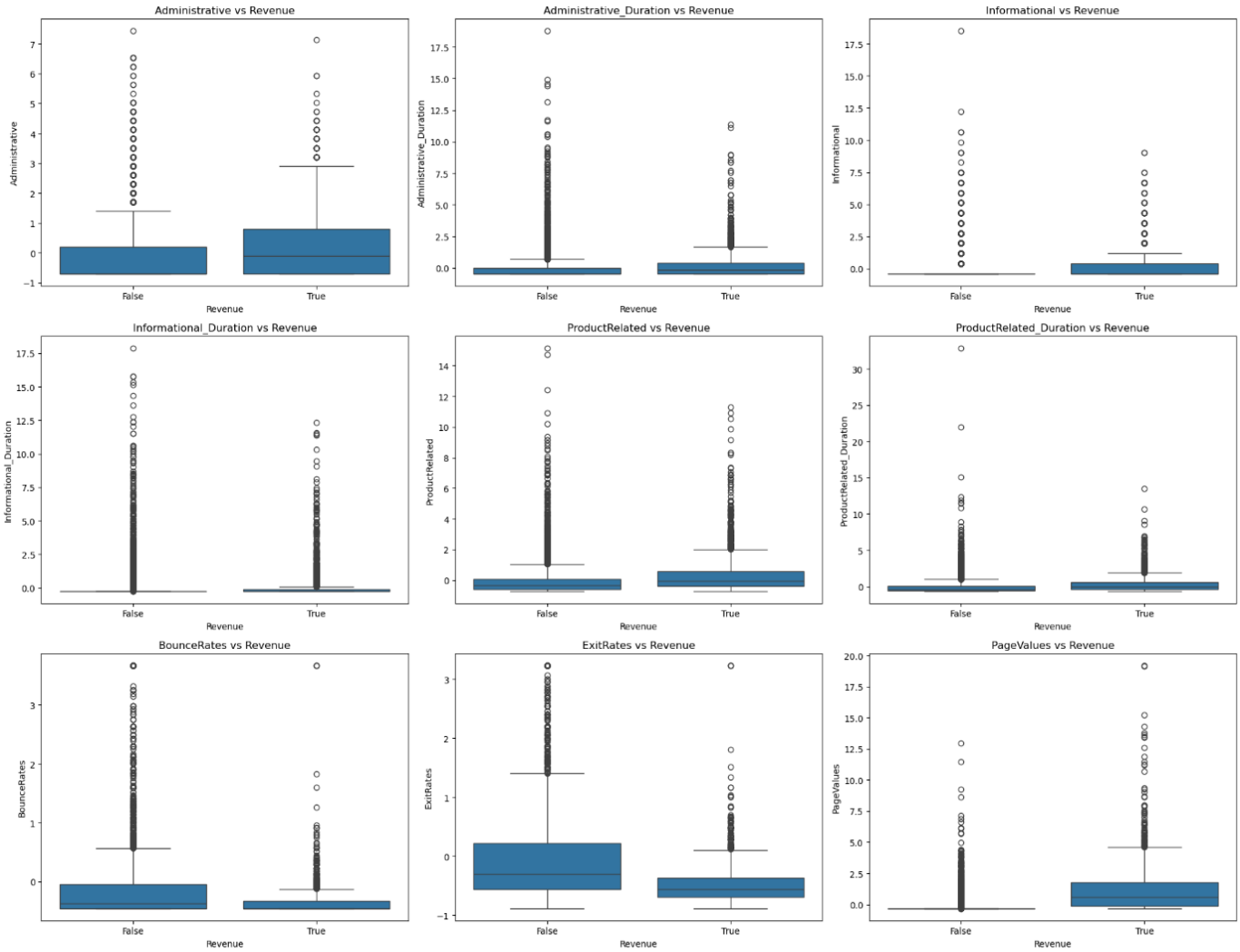

Fig 12. Clickstream Data Analysis

6.4 Negative Findings and Limitations

Although our models worked well most of the time, sometimes the model would misclassify, especially in cases where consumer behaviour did not turn out as expected, or some external factors had not been taken into account by the dataset. Besides, effective class imbalance management, even with the introduction of techniques like SMOTE, was quite problematic for consistent performance across all classes.

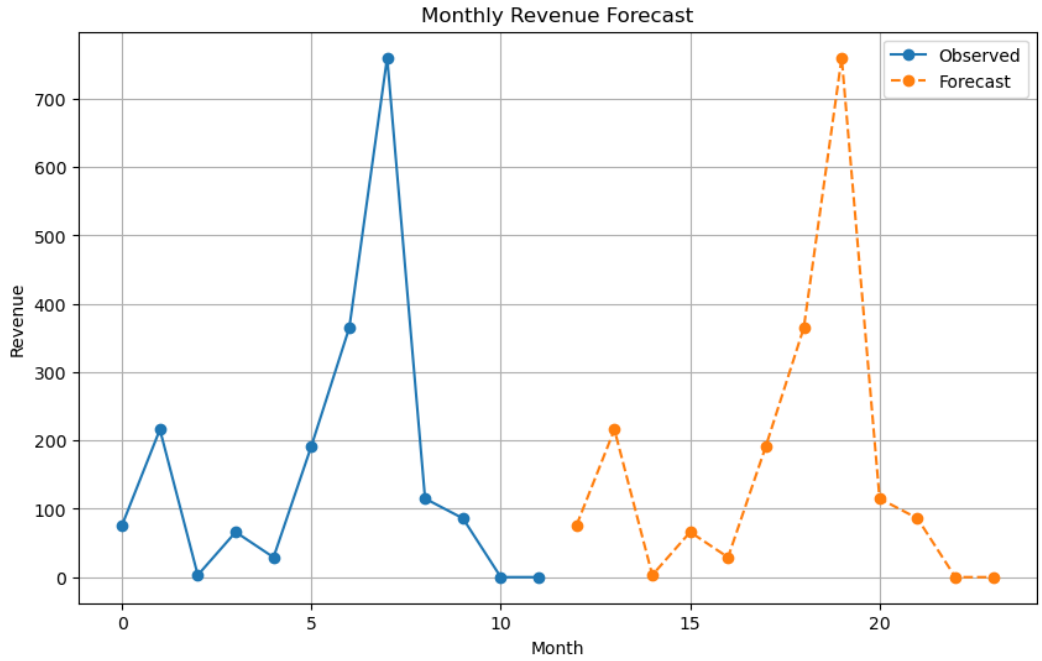

Fig 13. Future Monthly Sales Forecast

6.5 Analysis of Monthly Conversion Rates

We studied conversion rates on a monthly basis to understand the seasonal trends in client activity. The study unveiled significant seasonal variations since there were high rates of conversions during the holiday periods and promotional seasons. These insights can be used in helping marketing strategies where organizations can cue in their promotion during periods when customers have more involvement.

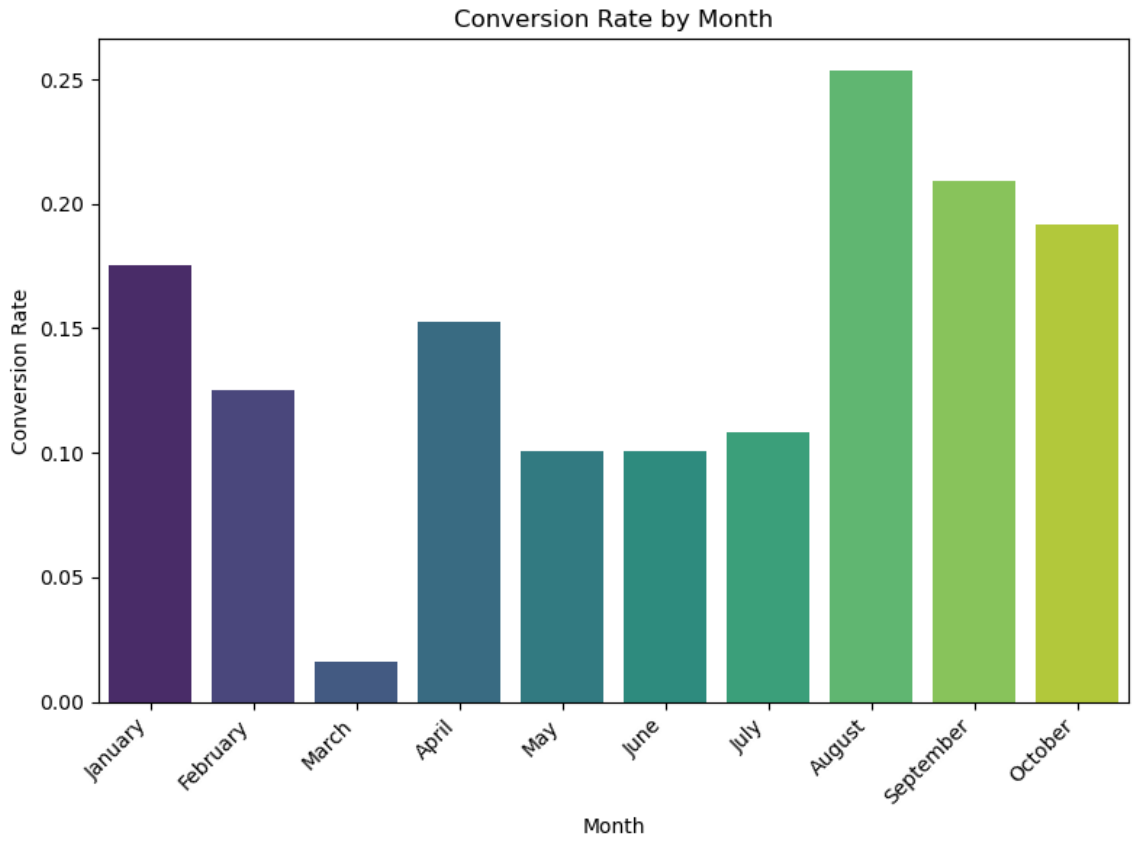

Fig 14. Monthly Conversion Rates

6.6 Conversion Rates by Weekend

We also investigated the effect of weekends on conversion rates. The result indicated that there was a mild increase in conversion rates during weekends; it is probable that more conversions happen over the weekend as customers go through their shopping goals while not working, ergo dedicating more time to shopping. This kind of information could be used to tailor promotions or offers set on e-commerce platforms at that time.

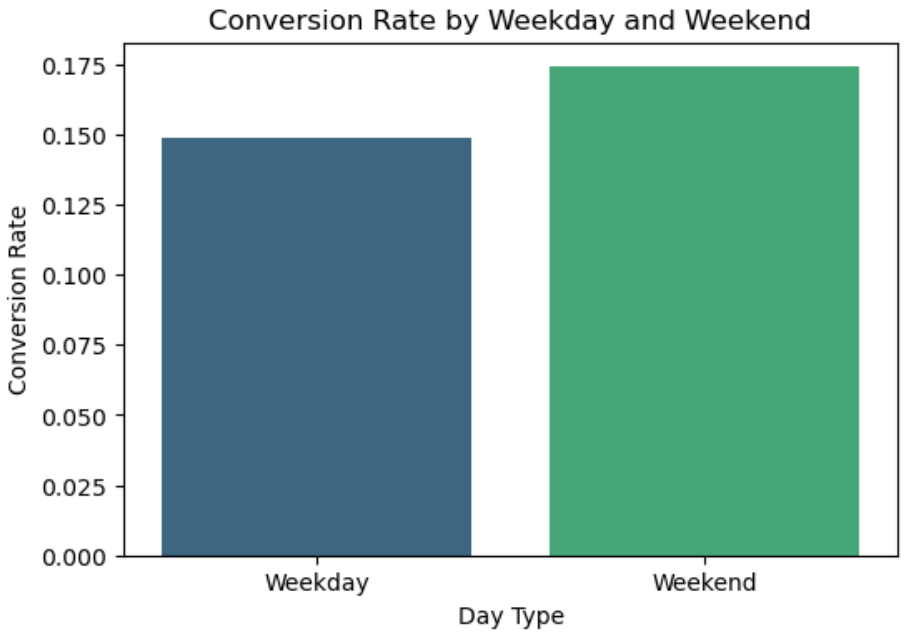

Fig 15. Conversion Rates by Weekend

These findings justify the effectiveness of our approach in solving the research problem by improving the accuracy of predictive models of customer behaviour. Involvement of deep learning and hybrid model combined with social media and temporal data offers a compelling answer and augments both theoretical understanding and practical implementation within the e-commerce sector. Finally, the built models, proven to have high predictive accuracy, drew some valuable insights towards client activity patterns. From an academic standpoint, new levels of machine learning in e-commerce will be further looked into. The models provide a path for customer interaction and marketing efficiency optimization. It also pointed out a few areas that required further studies: some possible supplementary data sources and hybrid model architectures that could increase the forecast accuracy.

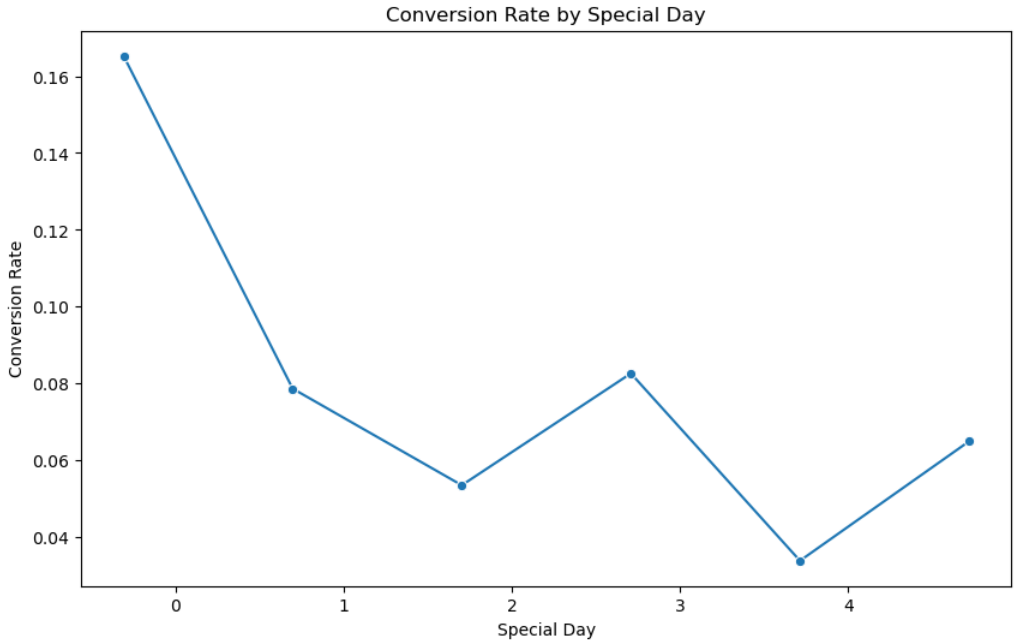

Fig 16. Conversion Rates by Special Days

7. Conclusion

This research aimed to enhance the capability for prediction of client behaviour within e-commerce through the application of advanced machine learning and deep learning models. By careful investigation and experimentation, we could build models that perform exceptionally well for multiple parameters. Of these, the LSTM- and hybrid-based models have the capability to capture complex and time-dependent patterns in consumer data. The obtained results further reinforce our hypothesis and show that deep learning models, and more importantly those developed specifically for sequential data processing, outperform typical machine learning methods in this particular scenario.

In this sense, this research contributes with new academic knowledge of how far machine learning can be used in e-commerce, giving special value and pointing out the potential and usefulness of deep learning and hybrid techniques. The results derived from this research will be used as the basic input for further research on more developed models and other sources of data, like interactions on social media, to increase predictive capacity.

The implication is tremendous for practice. These models can integrate with e-commerce platforms to have a better customer experience through personalized marketing, accurate targeting, and real-time intervention. This will eventually lead to more conversion and lesser cart abandonment. The deep dive of the conversion rate based on monthly and weekend data will provide practical information for the managers which will help in marketing strategies and promotional planning and make sure business operations are synced with periods of heightened customer involvement.

However, it recognizes the limitations of this study that could be linked to the skewed class distribution and the potential to misclassify atypical customer behaviour. This area should be the basis for further studies that aim to address some of the limitations in this study by looking into additional sources of data, refining hybrid model architectures, and using more advanced strategies for imbalanced data management.

This paper essentially shows the importance of using advanced machine learning and deep learning models to precisely predict consumer behaviour in the e-commerce industry. It offers useful fundamental knowledge and practical implementations that can tremendously improve the sector. Including these models into e-commerce systems might increase customer happiness and boost company growth.

8. References

1. Agustyaningrum, C. I., Haris, M., Aryanti, R., & Misriati, T. (2021). Online shopper intention analysis using conventional machine learning and deep neural network classification algorithm. *Jurnal Penelitian Pos dan Informatika*, *11*(1), 89-100.1. Adomavicius, G., & Tuzhilin, A. (2005). Personalization technologies: a process-oriented perspective. *Communications of the ACM*, *48*(10), 83-90.
2. Bayus, B. L. (2010). Crowdsourcing and individual creativity over time: The detrimental effects of past success. *Available at SSRN 1667101*.
3. Berry, M. J., & Linoff, G. S. (2009). *Data mining techniques*. John Wiley & Sons.
4. Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.
5. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
6. Leung, W. K., Chang, M. K., Cheung, M. L., & Shi, S. (2022). Understanding consumers' post-consumption behaviors in C2C social commerce: the role of functional and relational customer orientation. *Internet Research*, *32*(4), 1131-1167.

7. Coussement, K., & Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management*, *45*(3), 164-174.
8. Fader, P. S., & Hardie, B. G. (2009). Probability models for customer-base analysis. *Journal of interactive marketing*, *23*(1), 61-69.
9. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.
10. Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. *Data mining and knowledge discovery*, *1*(3), 291-316.
11. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.
12. Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*.
13. Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, *34*(10), 2902-2917.14. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and. *Techniques, Waltham: Morgan Kaufmann Publishers*.
14. Harrell, F. E. (2012). Regression modeling strategies. *R package version*, 6-2.
15. Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, *44*(1), 1-12.
16. Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.18. Mittal, V., & Kamakura, W. A. (2001). Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. *Journal of marketing research*, *38*(1), 131-142.
17. Lee, H. (2024). Interest-Based E-Commerce and Users' Purchase Intention on Social Network Platforms. *IEEE Access*.
18. Liu, N., Woon, W. L., Aung, Z., & Afshari, A. (2014, May). Handling class imbalance in customer behavior prediction. In *2014 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 100-103). IEEE.
19. Liu, N., Woon, W. L., Aung, Z., & Afshari, A. (2014, May). Handling class imbalance in customer behavior prediction. In *2014 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 100-103). IEEE.
20. Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, *36*(2), 2592-2602.
21. Naduvilveetil, A. (2020). *Impact of seasonal promotion on customer satisfaction in e-commerce industry* (Doctoral dissertation, Dublin Business School).
22. Rust, R. T., Lemon, K. N., & Zeithaml, V. A. (2004). Return on marketing: Using customer equity to focus marketing strategy. *Journal of marketing*, *68*(1), 109-127.
23. Rausch, T. M., Derra, N. D., & Wolf, L. (2022). Predicting online shopping cart abandonment with machine learning approaches. *International Journal of Market Research*, *64*(1), 89-112.