National College of Ireland

Smart Supermarkets: A Unified Approach to Predicting Customer Churn with Machine Learning and Deep Learning

MSc Research Project
Data Analytics

Sherin Parakkalayil Mathew
Student ID: X22128859

School of Computing
National College of Ireland

Supervisor: Musfira Jilani

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Sherin Parakkalayil Mathew……

Student ID: ……X22128859……

Programme: MSc Data Analytics…… **Year:** ……2023……

Module: …MSc Research Project ……

Supervisor: ……Musfira Jilani……

Submission Due Date: ……14/12/2023……

Project Title: … Smart Supermarkets: A Unified Approach to Predicting Customer Churn with Machine Learning and Deep Learning……

Word Count: ……6606…… **Page Count**……24……

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: ……

Date: ……14/12/2023……

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):	

Abstract

Consumer churn, a key issue in many businesses, including supermarkets, necessitates the use of predictive algorithms to anticipate and reduce possible customer departures. We aim to comprehensively understand and forecast consumer turnover in the supermarket arena. This work is intended to construct a robust prediction system by leveraging a wide ensemble of machine learning and deep learning models such as SVM, Adaboost, Random Forest, XGBoost, CNN LSTM, and GRU BiLSTM. The models were built using a dataset that included customer characteristics, purchase behavior, and churn labels. The project concluded with the creation of a web application interface for predicting attrition by inputting client information. With an accuracy score of 80%, the GRU BiLSTM model emerged as the most accurate performer, proving its usefulness in grasping sequential consumer behavior. However, there were limits to real-time adaptation and dataset comprehensiveness. Future development will include incorporating real-time data sources and improving the user interface for better use. This research lays the groundwork for more robust and adaptive customer attrition prediction systems in the supermarket business.

Keywords: Customer Churn Prediction, Machine Learning Models, Deep Learning Models, Adaboost Classifier

Chapter 1 Introduction

1.1 Background

Customer churn, defined as customer turnover or attrition, is a critical problem in a variety of businesses, including retail. Customer churn is defined differently in various businesses. It is broadly classified into two types: active customer churn and passive customer churn (Hu et al., 2020). Supermarkets, as key components of the retail environment, are subject to the impact of customer turnover, demanding proactive methods for identifying and mitigating it. Supermarkets have implemented predictive techniques to project probable client departures, prioritizing the maintenance of existing customers over obtaining new ones. Customers are the most valuable assets of any firm (Kim and Lee, 2022). Keeping a loyal client is six times less expensive than acquiring new consumers. The Grocery Consumer Churn Prediction research tries to dive into the subtle patterns of customer behavior within the supermarket domain in this context. The goal is to create a strong prediction model capable of detecting clients at risk of churn by combining sophisticated machine learning and deep learning techniques. The motivation for this undertaking stems from the realization that understanding and anticipating customer turnover not only helps to minimize revenue loss but also enables focused initiatives for client retention and happiness.

1.2Aim of the study

The major goal of this project is to conduct a thorough investigation of customer turnover phenomena in the supermarket sector and to develop an effective forecasting model. The study uses machine learning and deep learning approaches to evaluate historical customer data, concentrating on characteristics, purchasing habits, and sequential patterns, in order to determine the underlying causes of customer turnover. The research intends to construct a strong predictive model capable of forecasting probable instances of customer attrition by identifying and analyzing these churn sources. The ultimate objective is to provide supermarkets with proactive customer retention methods by identifying who is most likely to churn in the future. The study's goal in developing this predictive system is to bridge the gap in supermarket-specific churn analysis and contribute to improving customer relationship management, decreasing revenue loss, and supporting long-term company success in the retail sector.

1.3 Research Objectives

The research objectives of this report are:

1. To build and analyze several machine learning and deep learning models to predict customer turnover in supermarkets, such as SVM, Adaboost, Random Forest, XGBoost, CNN LSTM, and GRU BiLSTM.
2. To train and assess prediction models, historical customer data, comprising traits, purchase activities, and churn indicators, must be analyzed.
3. To evaluate and analyze the performance of several models based on accuracy, ROC AUC, and other important metrics, select the most effective model for churn prediction.
4. To provide a user-friendly online application interface that allows stakeholders to enter customer information and generate churn forecasts based on the best-performing model.

1.4 Research Questions

The research questions for this report are:

1. Which machine learning or deep learning model exhibits superior performance?
2. What customer attributes and behaviors correlate with churn?
3. How Accurate Are the Predictive Models in Identifying Potential Churn Cases?
4. How Can Predictive Insights Aid in Proactive Customer Retention Strategies?

1.5 Research Gaps

There is a shortage of substantial research concentrating exclusively on customer churn prediction within the supermarket domain in the context of the Supermarket Customer Churn Prediction Project. Existing research frequently caters to wider sectors, demanding more focused studies on supermarket-specific customer habits and turnover trends. Existing predictive models may be lacking in real-time adaptation, making them difficult to deploy in dynamic retail situations. Improving models to accept real-time data updates and scalability for varied supermarket sizes is still a work in progress. The models' predictive powers may be limited if other data sources, like market trends or competition assessments, are not fully integrated with customer-centric data. Investigating methods to include varied external data sources might improve the prediction system's accuracy and resilience.

Chapter 2 Literature Review

2.1 ML Models for Supermarket Customer Churn Prediction

In recent years, firms in a variety of industries have faced persistent difficulty with client turnover. The telecoms industry has experienced the direct impact of turnover on revenue, motivating the hunt for reliable predictive models. (De et al., 2021) conducted a thorough assessment of 55 publications published between 2004 and 2020, classifying them as feature selection, class imbalance handling, experimentation with machine learning techniques, hybrid, and ensemble models in churn prediction. Their results emphasized the need for lowering churn in order to maximize revenues, emphasizing developments in machine learning techniques in industries such as telecommunications, finance, e-commerce, and energy. Meanwhile, (Odegua et al., 2020) investigated supermarket sales forecasting by applying machine learning algorithms such as K-Nearest Neighbor, Gradient Boosting, and Random Forest to solve shortcomings in traditional statistical approaches. Their research stressed the need for precise sales forecasting for supermarkets and demonstrated the Random Forest algorithm's superiority in this respect. Simultaneously, (Lavanya et al., 2023) investigated cost-sensitive learning (CSL) inside machine learning algorithms for predicting telecom customer turnover. Their strategy is intended to link models with business goals, greatly boosting forecast accuracy over prior methodologies. (Hooda and Pooja Mittal, 2023), on the other hand, introduced the OKMSVM model to solve customer churn analysis difficulties in telecoms. Their method included feature extraction with KPCA and instance selection with ALO optimization, which resulted in remarkable prediction performance and outperformed existing models. Finally, (Singh et al., 2020) emphasized the importance of integrating machine learning into customer relationship management (CRM), emphasizing the importance of using supervised and unsupervised techniques for customer identification, attraction, retention, and development in industries such as e-commerce and telecommunications.

Table 2.1: Comparative Analysis on ML Models for Supermarket Customer Churn Prediction

Study	Industry/Domain	Approach/Model	Techniques/Algorithms Used	Main Findings/Results
De et al., 2021	Various (Telecom, Banking, E-commerce, Energy)	Churn Prediction	Feature selection, Class imbalance handling, Experimentation with ML algorithms, Hybrid & Ensemble models	Highlighted significance of reducing churn for profit maximization; Advanced ML techniques in various sectors
Odegua et al., 2020	Supermarket/Sales Forecasting	Machine Learning for Sales Estimation	K-Nearest Neighbor, Gradient Boosting, Random Forest	Random Forest outperforms for supermarket sales prediction
Lavanya et al., 2023	Telecommunications	Cost-Sensitive Learning for Churn Prediction	ML with cost-sensitive learning, Feature selection	Accuracy: 91.05%; AUC Score: 85.76%; RMSE Score: 2.838
Hooda & Pooja Mittal, 2023	Telecommunications	OKMSVM Model for Churn Analysis	KPCA for feature extraction, ALO optimization	Accuracy: 91.05%; AUC Score: 85.76%; RMSE Score: 2.838
Singh et al., 2020	Various (E-commerce, Telecom)	Integrating ML with CRM	Supervised & Unsupervised ML techniques	Emphasized using ML for customer identification, attraction, retention, and development

(Günese et al., 2021) presented a study spanning 2018 and 2019 into forecasting churn behavior among valued current consumers across various FMCG brands in the Turkish industry.

(Baderiya and Chawan, 2018) introduced "CRetention," a predictive model suited for inventory-led e-commerce enterprises that uses machine learning and deep learning technologies to identify churners and manage inventory while providing individualized marketing tactics. Similarly, (Abbas et al., 2022) addressed retail sector churn by employing multiple machine learning techniques such as logistic regression, random forest, decision tree, K nearest neighbors, and XGboost, achieving a 73 percent accuracy rate in churn prediction and emphasizing XGboost's potential as an efficient classifier. Concurrently, (Suhanda et al., 2022) investigated customer retention techniques, emphasizing the importance of service quality, pricing, satisfaction, and trust, and proposed a predictive customer retention strategy based on data mining and the random forest algorithm. (Mallawarachchi et al., 2022) created a web-based prediction model called

"CRetention" specifically designed for retail food stores. This technology examined data, controlled inventories, performed market basket research, and made individualized suggestions, earning high praise and approval from retail supermarket operators. The Logit Leaf Model (LLM) was created by Caigny et al. (2018). It is a hybrid method that combines decision rules and models built on segmented data to get around the problems that decision trees and logistic regression have when trying to predict churn. Their research demonstrated greater prediction performance as compared to standard models. Furthermore, (Calli and Kasim, 2022) investigated B2B customer attrition in an ERP firm with a SaaS model, employing machine learning methods to detect prospective churners and emphasizing the importance of product quantity and customer attributes in forecasting churn in this business environment.

Table 2.2: Comparative Analysis on ML Models for Supermarket Customer Churn Prediction

Study	Industry Focus	Methodology	Key Findings
Baderiya and Chawan (2018)	E-commerce (Inventory)	Machine Learning, Deep Learning	"CRetention" predictive model for personalized strategies, inventory management
Abbas et al. (2022)	Retail	Various Machine Learning Techniques	Achieved 73% accuracy in churn prediction, XGboost as efficient classifier
Suhanda et al. (2022)	Various Industries	Data Mining, Random Forest	Emphasized service quality, Random Forest for retention strategies
Mallawarachchi et al. (2022)	Retail Grocery	Web-Based Predictive Model	High recommendation among retail grocery owners
Caigny et al. (2018)	Various Industries	Hybrid Model (LLM)	LLM outperformed traditional models in predictive performance
Calli and Kasim (2022)	B2B (ERP, SaaS)	Machine Learning Algorithms	Identified product quantity, customer features as churn predictors

2.2 DL Models for Supermarket Customer Churn Prediction

Churn studies have developed as essential tools for firms seeking to retain prized consumers while increasing income sources. Researchers have increasingly turned to sophisticated techniques, notably deep learning and machine learning algorithms, to untangle the complexity of consumer behavior and anticipate churn trends in industries such as retail, rental services, and electronics. Each of (Seymen et al., 2020), (Hiziroglu et al., 2020), (Mudassar & Byun, 2019), (Motevali, 2023), and (Suh, 2023) made significant contributions to this domain. Seymen et al. (2020) demonstrated the superiority of a deep learning model over logistic regression and neural networks in forecasting customer attrition in the retail industry. (Hiziroglu et al., 2020) expanded this investigation by using recurrent neural networks to solve traffic patterns and optimize parking management in grocery store parking lots. (Mudassar and Byun, 2019) investigated the use of RNNs to estimate traffic flow and improve parking tactics in a grocery store parking lot. (Motevali, 2023) dealt with customer turnover in a water purifier rental business by using machine learning algorithms to create a churn prediction model and enable targeted retention marketing efforts. (Suh, 2023) expanded on this knowledge by analyzing customer behavior data in the water purifier rental business, attaining excellent predicted accuracy, and allowing rental care customer management personnel to apply individualized retention tactics.

Table 2.3: Comparative Analysis on DL Models for Supermarket Customer Churn Prediction

Study	Industry Focus	Methodology Used	Key Findings
Seymen et al. (2020)	Retail	Deep Learning vs. Logistic Regression vs. Neural Networks	Deep learning outperforms other methods in churn prediction.
Hiziroglu et al. (2020)	Grocery Store Parking	Recurrent Neural Networks (RNNs)	RNNs predict customer flow for parking optimization.
Mudassar & Byun (2019)	Grocery Store Parking	Recurrent Neural Networks (RNNs)	RNNs used to forecast traffic flow in parking lots.
Motevali (2023)	Water Purifier Rental	Machine Learning Algorithms	Achieved high accuracy in predicting customer churn.
Suh (2023)	Water Purifier Rental	Machine Learning Algorithms	High accuracy in churn prediction, enabling tailored retention strategies.

Chapter 3 Research Methodology

3.1 CRISP-DM Methodology

The CRISP-DM (Cross-Industry Standard Process for Data Mining) approach divides the data mining process into six main phases, each playing a crucial role in gaining insights and developing predictive models. In the context of a customer attrition prediction project in a supermarket, here's a description of each phase:

1. **Business Understanding:** The initial phase focuses on comprehending the business objectives and needs. The goal of a supermarket is to understand the causes of consumer turnover. Knowing this helps to link data mining aims with business goals, ensuring that future studies and models address the organization's primary concerns, such as customer retention and profitability.
2. **Data Understanding:** This phase begins with data collection and exploration to determine the dataset's structure, content, and quality. In the context of customer churn, this stage entails gathering information about customer demographics, purchasing habits, and churn indicators. Understanding the data allows you to uncover possible churn drivers and evaluate their significance for constructing predictive models.
3. **Data Preparation:** Cleaning, preparing, and converting data to make it appropriate for analysis and modeling is what data preparation entails. This stage is essential in a customer churn project since it entails dealing with missing values, encoding categorical variables, scaling numerical features, and dealing with unbalanced classes. Data preparation guarantees that the dataset is of high quality and suitable for modeling approaches.
4. **Modelling:** During the modeling step, relevant algorithms are chosen, and predictive models are built and trained using the prepared dataset. To forecast customer turnover, you can use various machine learning algorithms, such as logistic regression, decision trees, random forests, or gradient boosting. Based on past data, this phase assists in finding trends and anticipating prospective churners.
5. **Evaluation:** The evaluation step evaluates the performance of the models created in the preceding phase. To assess how successfully the models forecast customer turnover, metrics like accuracy, precision, recall, and ROC-AUC are used. Evaluating the models aids in picking the most successful one for deployment and, if required, fine-tuning.
6. **Deployment:** The last stage is implementing the chosen model in the corporate context. In the case of a supermarket, adopting a churn prediction model may include connecting it with customer relationship management tools. Using the model's predictions, proactive actions are taken to retain customers who have been identified as prospective churners.

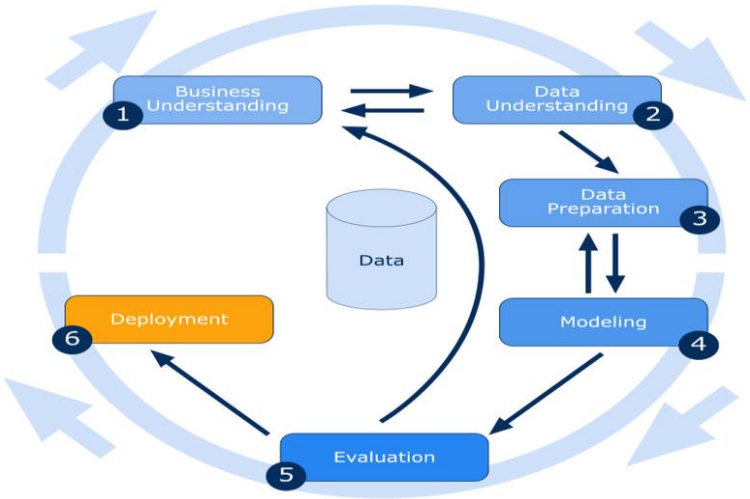

Figure 3.1: CRISP-DM architecture

3.2 Proposed Workflow Methodology

1. The workflow of the proposed methodology involves the following stages:
    1. **Data Collection and Understanding:** The data for the Supermarket Customer Churn Prediction project will be obtained from https://data.world/dradar/customerchurnpredictiononsupermarketdata, which will provide access to a dataset comprising information about customer turnover in supermarkets. This dataset includes a variety of variables with the primary goal of analyzing customer churn, which refers to client leaving or attrition. After importing the dataset into a Panda DataFrame, the first step will be to understand the characteristics included inside, including the target variable 'Churn.' This stage is critical for learning about the elements that may impact consumer exit from supermarkets. Understanding the qualities and their correlations, as well as the relevance of churn rates in the context of companies such as supermarkets, will create the framework for later activities such as data pretreatment, visualization, model construction, and prediction.
    2. **Importing Libraries and Loading Data:** NumPy for numerical operations, Pandas for data processing, Matplotlib and Seaborn for graphical representations, and Plotly for interactive visualizations including express, figure factory, and graph objects are loaded during the startup phase. This offers a wide range of functions necessary for data exploration, visualization, and analysis. After that, we import the dataset from the supplied link into a Panda DataFrame, enabling us to examine and manipulate the supermarket customer churn statistics. This stage lays the groundwork for extensive data cleansing, investigation, visualization, and subsequent

analysis to obtain relevant insights about consumer behavior and churn trends in the supermarket sector.

3. **Data Cleaning:** A first analysis of the dataset's columns was performed during the data cleaning process, exposing elements relating to customer information, purchase details, and the goal variable 'customer churn.' To simplify the dataset for analysis, superfluous columns such as 'row number,' 'invoice id,' and 'customer id' were removed because they had no bearing on churn prediction. Following that, a thorough examination of the dataset's structure and data types was carried out, verifying the lack of null values across all columns. Statistical summaries were prepared to comprehend the distribution and range of values within the dataset, including descriptive statistics that highlight numerical aspects such as age, credit score, and product-related variables.
4. **Data Visualization and Analysis:** During the data visualization and analysis phase, the dataset was thoroughly explored to get insights into numerous properties and their correlations. Initial studies included counting the number of unique values in each column and calculating value counts for categorical columns like 'branch,' 'gender,' and 'customer type.' We constructed visual representations such as count plots, bar plots, box plots, histograms, scatter plots, pie charts, and funnel area charts to explore deeper into the dataset. These visualizations depicted the distribution of parameters such as age, credit score, and product categories, as well as their relationships with the goal variable 'customer churn.' Exploring the influence of factors such as total amount spent, tax amounts, and product categories on turnover rates yielded insights on customer churn behavior. Furthermore, comparisons were done across consumer categories, genders, and branches to identify potential trends impacting customer attrition within the supermarket sector.
5. **Data preprocessing:** During the data preparation step, categorical columns like 'branch,' 'gender,' 'customer type,' and 'product category' were recognized and converted into numerical representations using LabelEncoder, making them easier to include in machine learning models. This change guarantees that the algorithms can successfully process these category variables. A correlation matrix was also generated and shown using a cluster map to provide insights into feature correlations and the intensities of linkages. This visualization style aids in comprehending the interdependence of several variables within the dataset. Furthermore, though not explicitly demonstrated in the provided code snippet, standard preprocessing steps such as data normalization using StandardScaler to bring features to a common scale and the use of SMOTE (Synthetic Minority Over-sampling Technique) for dealing with imbalanced target classes could be implemented.

6. **Feature Importance:** An Extra Trees Regressor model was used in the feature importance analysis to analyze the value of various attributes in predicting customer turnover inside the supermarket domain. The model trained on the dataset and identified the relevance of each feature in the prediction process. This research aids in determining the most relevant factors influencing customer attrition. The feature importance was calculated and shown using a bar plot, which demonstrated the relative importance of each characteristic.
7. **Splitting Data:** During the data splitting step, the dataset was separated into training and testing sets using the sklearn library's 'train test split' function. The data was divided into training and testing sets using a 90:10 ratio, assigning 90% of the data for training the prediction models and 10% for assessing model performance. Furthermore, the parameter'stratify=y' was used in both the training and testing sets to retain the proportionate distribution of the goal variable 'y' (indicating customer turnover). Ensuring an even distribution of the target variable's classes across the train-test split is crucial for maintaining model correctness and performance during assessment. Following that, the function 'plot roc curve' was written to show the Receiver Operating Characteristic (ROC) curve, a graphical depiction of a binary classification model's performance.

This project has created a web app with a GUI using Flask. So basically, let me explain the basic explanation for this web app. The customer churn prediction (real-time) is based on customer detail provided by the supermarket dataset. To build the supermarket customer churn prediction using customer churn prediction supermarket data from dataworld.com.

Chapter 4 Design Specification

The Design Specification chapter embodies a comprehensive blueprint outlining the technical architecture, functionalities, and methodologies deployed in the Customer Churn Prediction Web Application. Rooted in the critical understanding that churn represents the departure of clients or customers, this project addresses the common industry challenge across banks, mobile phone companies, internet service providers, and supermarkets where churn analysis serves as a crucial business metric. This metric underscores the economic advantage of retaining existing customers over acquiring new ones, given the higher purchasing potential of existing clientele. The primary objective of this project revolves around deciphering the reasons for customer churn in supermarkets and devising a predictive model capable of anticipating potential future churners. The application, developed using Flask, facilitates churn prediction by allowing users to input pertinent details and providing a seamless click-based prediction interface that categorizes churn likelihood into 'yes' or 'no'. The

design and functionality of the application encompass a robust backend infrastructure, incorporating data ingestion mechanisms, preprocessing pipelines, and model training frameworks.

This includes data cleansing, encoding categorical variables, and implementing predictive modeling algorithms like logistic regression, decision trees, or ensemble methods to discern churn patterns. Rigorous evaluation, utilizing metrics such as accuracy, precision, recall, and ROC-AUC, ensures the model's reliability and effectiveness. The user interface design emphasizes intuitiveness, featuring interactive elements that enable users to input crucial information for churn prediction effortlessly. The GUI offers a streamlined experience with clear prompts and feedback mechanisms. Additionally, the design specifications integrate considerations for scalability, security protocols, and deployment strategies within existing supermarket infrastructures. This comprehensive blueprint delineates a roadmap for system architects and developers, detailing system requirements, technical intricacies, and user interaction elements, ultimately aiming to empower supermarkets with a proactive tool for managing customer churn and enhancing retention strategies.

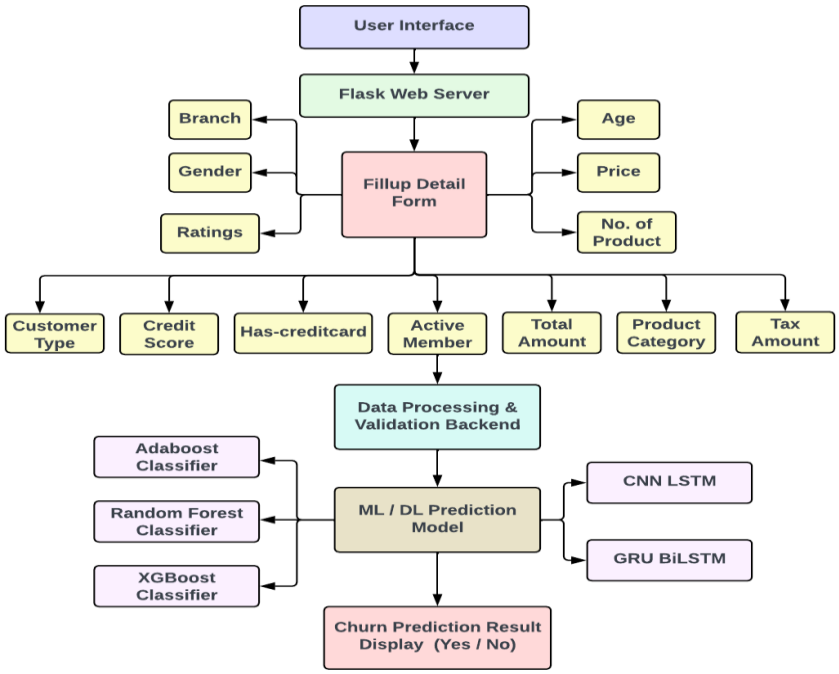

Figure 4.1 Block Diagram for the proposed system

Moreover, the predictive model development process involves the selection and implementation of appropriate machine learning algorithms and deep learning algorithms, emphasizing interpretability and accuracy. Techniques such as SVM Classifier, Adaboost Classifier, Random Forest Classifier, Xgboost Classifier, CNN Lstm, and Gru Bilstm are considered to build a predictive model capable of identifying customers most likely to churn in the future. The model's performance is rigorously evaluated using established metrics like accuracy, precision, recall, and ROC-AUC to ensure its efficacy and reliability in real-world scenarios.

Chapter 5 Implementation

The Implementation chapter provides a road map for the Customer Churn Prediction System's implementation. It includes the implementation of machine learning and deep learning models, beginning with the selection of varied classifiers: SVM, Adaboost, Random Forest, and XGBoost for machine learning, and CNN LSTM and GRU BiLSTM for deep learning. Extensive model training, validation, and assessment using relevant metrics are required throughout the deployment phase to ensure accuracy and resilience in forecasting customer turnover. This chapter also describes how to integrate these models into a Flask-based web application, allowing users to input critical customer variables such as branch selection, demographic information, credit scores, product preferences, and ratings to forecast attrition likelihood. Backend data processing, validation methods, and the creation of an intuitive user interface to promote smooth interaction are also part of the implementation. Thorough testing verifies the web app's performance and dependability, giving a platform for users to get forecasts about future customer churn in a supermarket scenario, eventually helping companies to make educated decisions about client retention methods.

5.1 Environmental Setup

The section then digs into the software requirements, covering the system's important software components and versions. Python programming language, libraries such as Pandas, NumPy, Scikit-learn for machine learning, TensorFlow or PyTorch for deep learning, Flask for web application development, and model-specific libraries such as XGBoost or Keras are all part of this.

5.2 List of Models

The List of Models section lists and describes the machine learning and deep learning models used in the Customer Churn Prediction System. It starts with a comprehensive list of predictive models, such as SVM Classifier, Adaboost Classifier, Random Forest Classifier, XGBoost Classifier, CNN LSTM (Convolutional Neural Network Long Short-Term Memory), and GRU BiLSTM (Gated Recurrent Unit Bidirectional LSTM).

A brief but instructive discussion of the underlying methodology, strengths, and relevance to customer churn prediction supports each model. The Support Vector Machine (SVM) is a machine learning classifier that finds the best hyperplane to separate various classes of data. The Adaboost Classifier, for example, progressively combines weak learners to generate a robust prediction model, whereas the Random Forest Classifier builds several decision trees to increase accuracy and avoid overfitting. For improved efficiency and accuracy, the XGBoost Classifier uses an advanced gradient boosting approach with parallel processing.

The section goes on to explain the two deep learning models provided, CNN LSTM and GRU BiLSTM. CNN LSTM takes into account both space and time in sequence prediction tasks by combining convolutional neural networks with long short-term memory units. GRU BiLSTM is a combination of Gated Recurrent Units and Bidirectional LSTMs that can collect long-term dependencies and use information from both past and future contexts in a sequence.

Readers gain a comprehensive understanding of the different methodologies used for predictive analysis within the Customer Churn Prediction System as each model's unique characteristics, benefits, and potential applications in predicting customer churn in a supermarket setting are thoroughly discussed.

5.3 Data Visualization (Data Analysis)

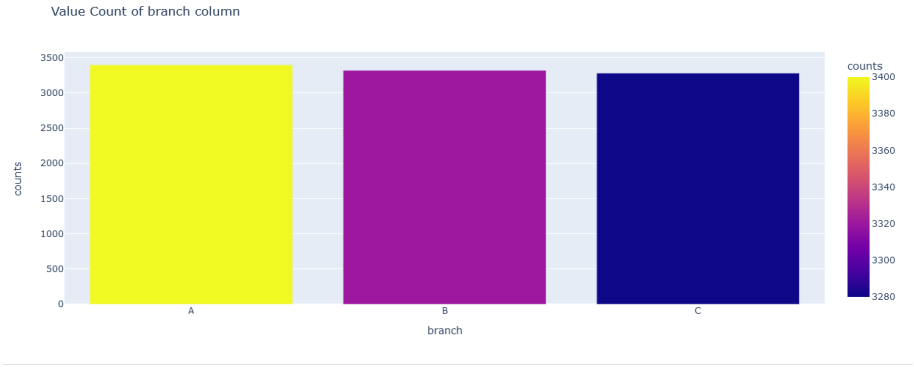

Figure 5.1: Bar Graph Illustrating Value Counts of Branches (A, B, C) in the Dataset

Figure 5.1 is a bar graph indicating the distribution of A, B, and C branches across the dataset. The graph's x-axis depicts the several branches, A, B, and C, while the y-axis displays the counts corresponding to each branch, ranging from 0 to 3500.

Figure 5.2: Histogram of Total Amount Distribution by Customer Churn (0 in Blue, 1 in Red)

Figure 5.2 depicts a histogram of the total amount distribution depending on customer turnover types. The total amount is represented by the x-axis, which displays the range or bins of total amounts seen in the dataset. Meanwhile, the y-axis shows the number of occurrences in each total amount bucket. To distinguish between customer churn categories, the histogram is colour-coded, with customer churn identified as 0 in blue and 1

in red. This colour differentiation provides for a visual comparison of the total amount distribution for customers who churn (labelled as 1) vs those who do not (labelled as 0).

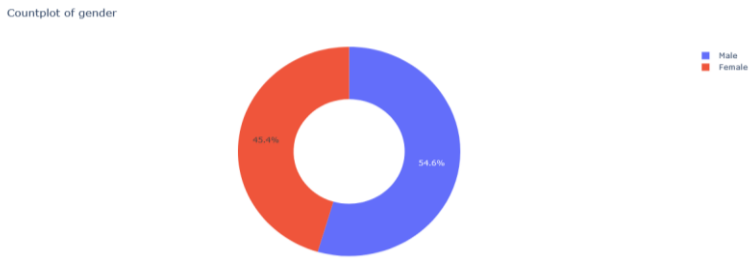

Figure 5.3: Pie Chart Illustrating Gender Distribution (Male in Blue: 45.4%, Female in Red: 54.6%)

A pie chart depicts the gender distribution within the dataset in Figure 5.3. The pie chart is divided into two portions that graphically show the percentage of men and women. The male part is shown in blue, accounting for 45.4 per cent of the total, while the female section is shown in red, accounting for 54.6 percent of the total.

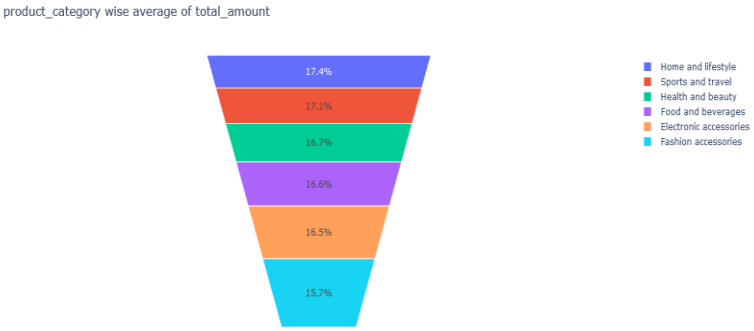

Figure 5.4: Funnel Chart of Product Category-wise Average Total Amount Distribution

A funnel chart or diagram depicts the product category-wise average distribution of total amount within the dataset in Figure 5.4. The funnel chart depicts numerous product categories, each marked by a different hue and a different proportion. Home and lifestyle are shown in blue, accounting for 17.4 percent of the average total amount. Sports and travel are highlighted in red, accounting for 17.1 percent, while health and beauty are highlighted in green, accounting for 16.7 percent. Furthermore, food and beverages are depicted in purple, accounting for 16.6 percent of the average total amount. Electronic and lifestyle accessories are shown in in orange at 16.3 percent, while fashion accessories are featured in sky blue at 15.7 percent.

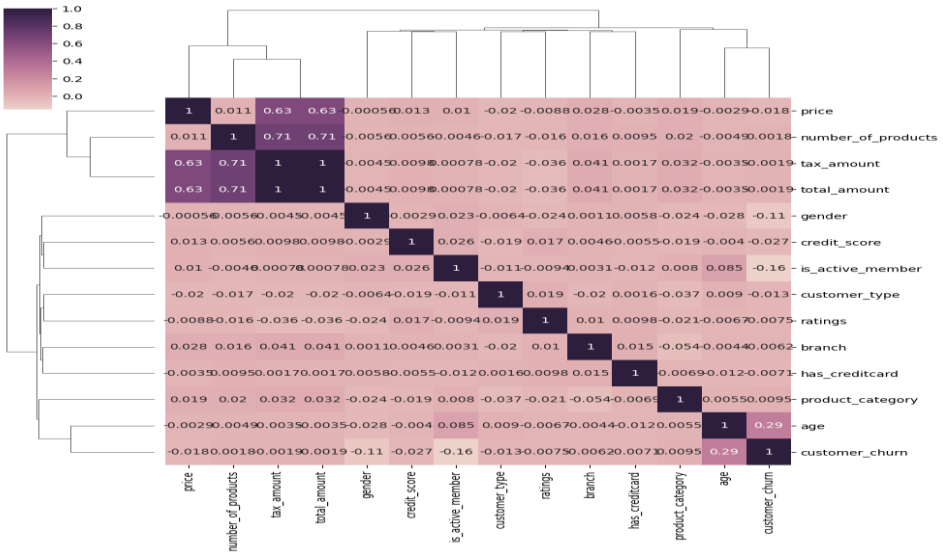

Figure 5.5: Correlation Matrix

The Seaborn cluster map displays the dataset's correlation matrix in Figure 5.5. Relationships between characteristics that affect the forecast of customer churn in the supermarkets industry are displayed visually. Positive correlations are represented by warm hues, and negative correlations by cool tones. A hierarchical clustering of related features is provided by the dendrograms, which shed light on possible patterns. Correlation strength is quantified by numerical annotations, which facilitate the analysis of inter-variable interactions. Understanding how different elements interact and affect the predictive model's evaluation of customer churn is made easier with the help of this visualization.

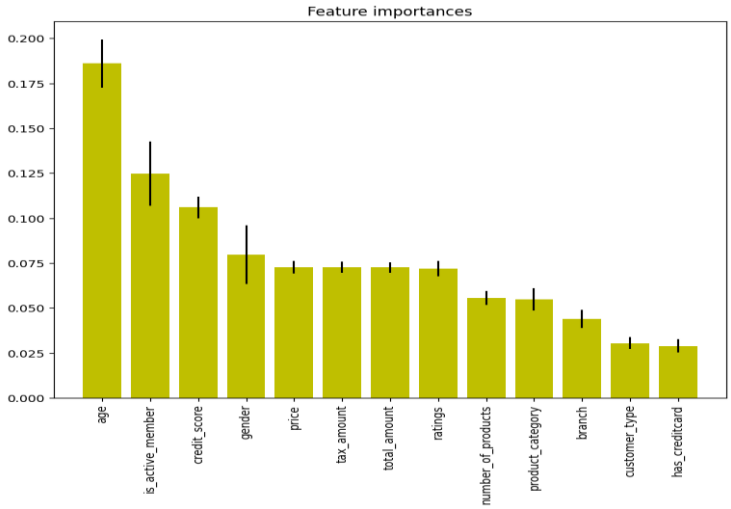

Figure 5.6: Feature Importance Graph

The bar graph in Figure 5.6 depicts the feature importance for forecasting customer attrition in the supermarket sector. The graph depicts the importance of several factors in the prediction model, such as age, price, gender, branch, customer type, ratings, and credit score, among others. Each characteristic is awarded a numerical

relevance value ranging from 0.0 to 0.2, reflecting its impact on customer churn predicting. A higher value on the graph indicates that that specific element is more important or relevant in evaluating whether a client is likely to churn or not. This graph assists in understanding the relative importance of various attributes in the model's decision-making process. For example, variables with higher importance scores (closer to 0.2) have a greater effect on forecasting churn, whereas those with lower importance scores (near to 0.0) have a lower significance in the prediction model.

Chapter 6 Evaluation

6.1 SVM Classifier

In the supermarket dataset, the Support Vector Machine (SVM) model is a key component of the predictive model ensemble and plays a major role in predicting consumer attrition trends. The SVM has strong prediction skills with an accuracy score of 0.77. Additionally, the SVM's high Area Under the Curve (AUC) score of 0.85 indicates great discriminating ability and highlights the manner in which it can distinguish between examples that are churn and those that are not. The usefulness of the SVM in the customer churn prediction system is highlighted in this study.

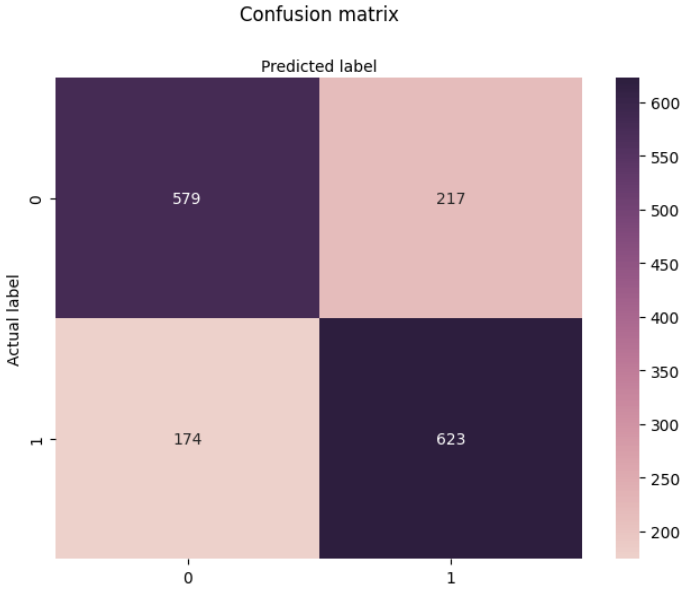

Figure 6.1 Confusion Matrix

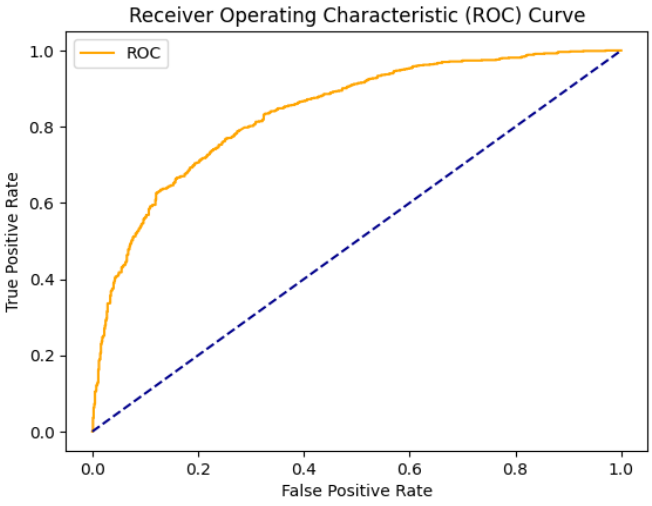

Figure 6.2 ROC Curve

6.2 Adaboost Classifier

The Adaboost Classifier, a key component in this report's predictive model ensemble, is critical in identifying customer churn trends in the supermarket dataset. Adaboost works by combining numerous weak learners into a powerful prediction model repeatedly. Throughout this investigation, Adaboost makes a major contribution to the correct prediction of customer turnover by leveraging the collective expertise of these weak learners. The Adaboost Classifier produced an accuracy score of 0.7746 in this scenario, indicating its ability to properly forecast customer turnover cases. Furthermore, the AUC score, a statistic that quantifies the model's

discriminating performance, was revealed to be 0.85. This high AUC value indicates the Adaboost Classifier's resilience in discriminating between churn and non-churn cases, further supporting its usefulness in the customer churn prediction system in this study.

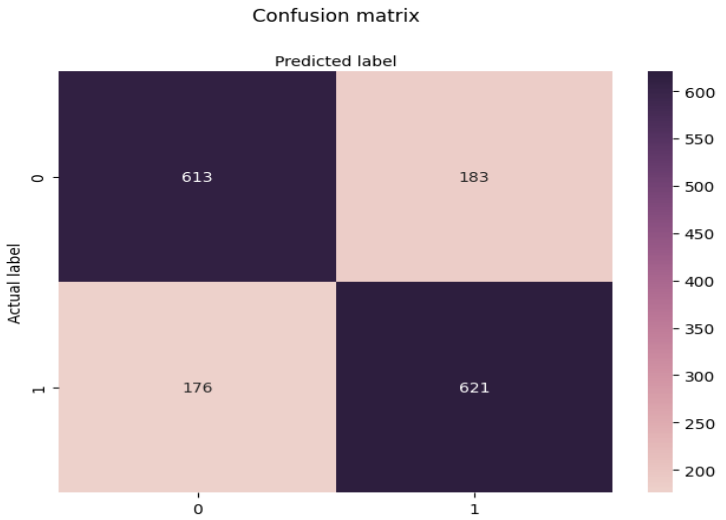

Figure 6.3 Confusion Matrix

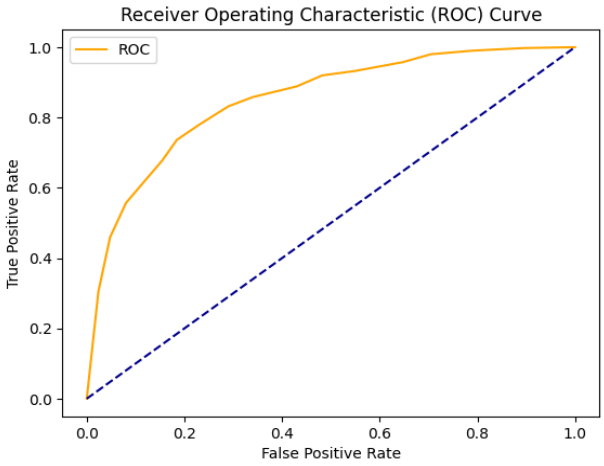

Figure 6.4 ROC Curve

6.3 Random Forest Classifier

This classifier works by building many decision trees and aggregating their results to make accurate predictions. The Random Forest Classifier greatly contributes to accurate customer churn forecasts throughout this investigation, employing its ensemble of decision trees to improve predictive accuracy. The Random Forest Classifier achieved an accuracy score of 0.7696 in this scenario, indicating its ability to effectively predict occurrences of customer turnover. Furthermore, the AUC score, which indicates the model's ability to distinguish between churn and non-churn cases, was found to be 0.84.

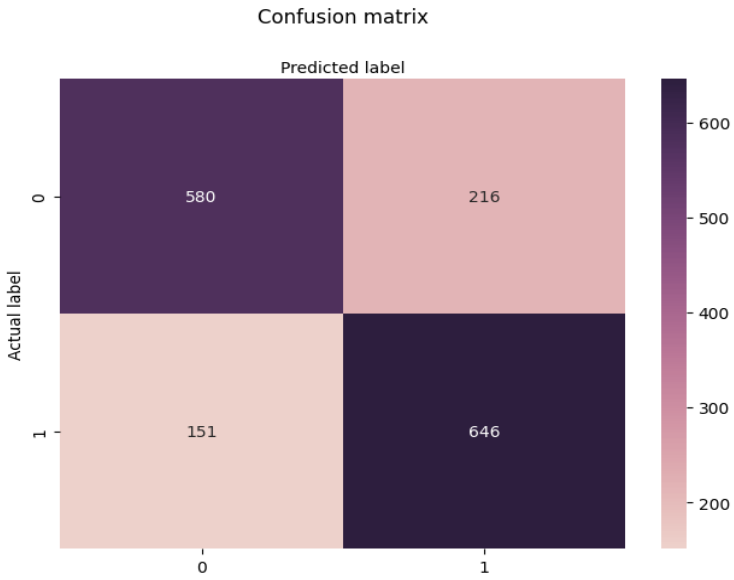

Figure 6.5 Confusion Matrix

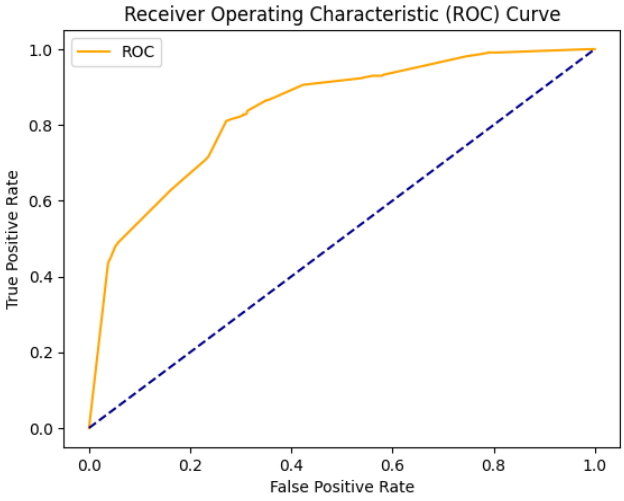

Figure 6.6 ROC Curve

6.4 XGBoost Classifier

XGBoost, known for its resilience, works by iteratively improving weak learners to create a powerful prediction model. XGBoost considerably helps to accurate customer churn estimates throughout this investigation by utilising the collective knowledge of these poor learners. The model achieved an accuracy score of 0.737, demonstrating its ability to properly identify instances of customer turnover. In addition, the AUC score, which measures the model's discriminating abilities, was a remarkable 0.80. These numbers highlight the XGBoost Classifier's capacity to discriminate between churn and non-churn cases, confirming its critical role in this report's customer churn prediction system.

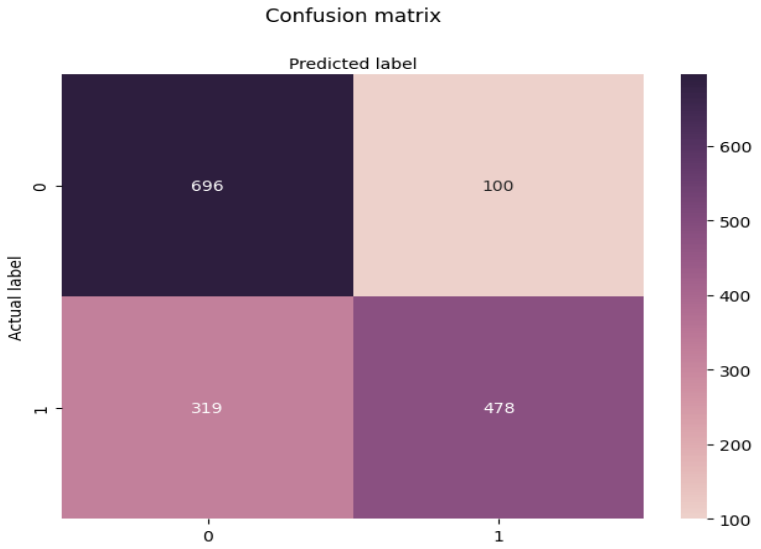

Figure 6.7 Confusion Matrix

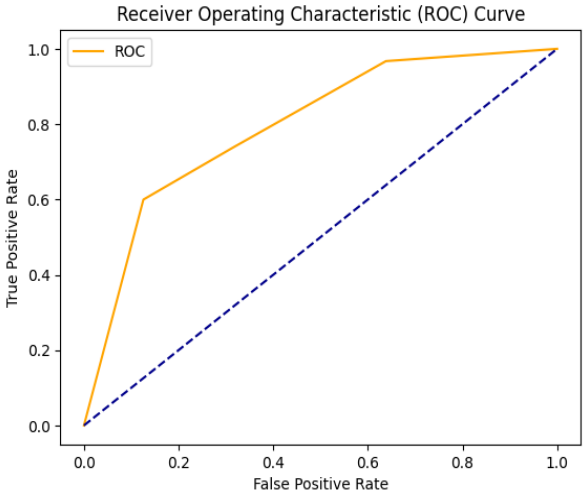

Figure 6.8 ROC Curve

6.5 CNN LSTM

The use of the CNN LSTM model in this Supermarket Customer Churn Prediction project is a conscious decision to take advantage of its abilities in sequence prediction problems. Combining Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) units, this model is capable of capturing both spatial and temporal patterns within sequential data, making it suited for studying consumer behaviour sequences. In this case, the CNN LSTM model was used to understand the sequential nature of consumer interactions and purchase habits in the supermarket dataset. With an accuracy score of 0.739 and an AUC score of 0.82, this model performed admirably in properly forecasting customer attrition.

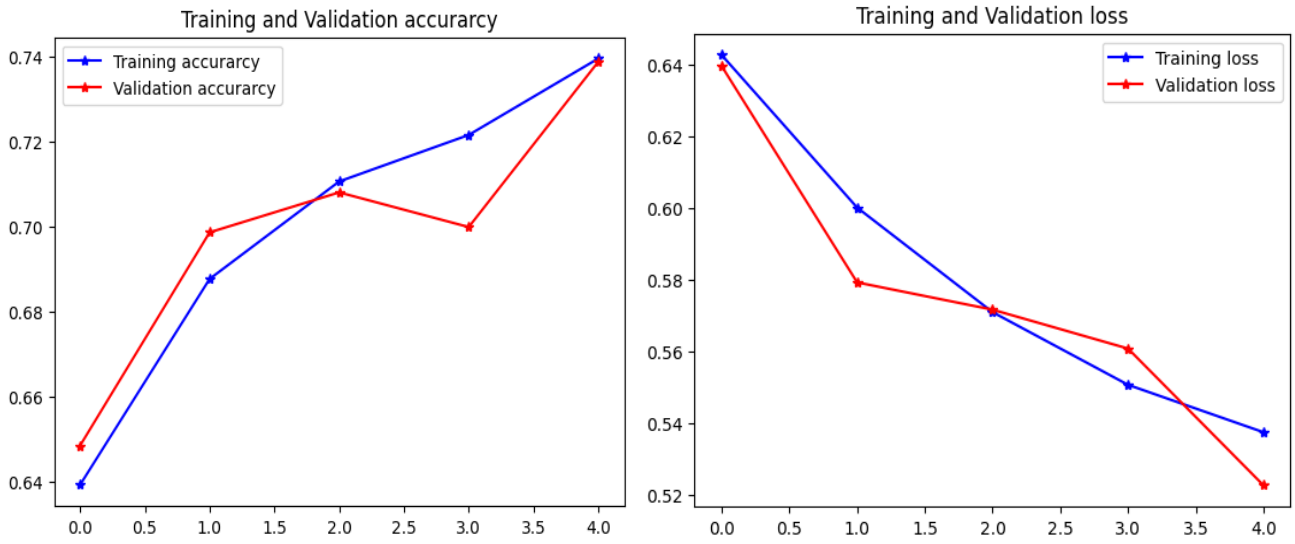

Figure 6.9: Accuracy & Loss Graph

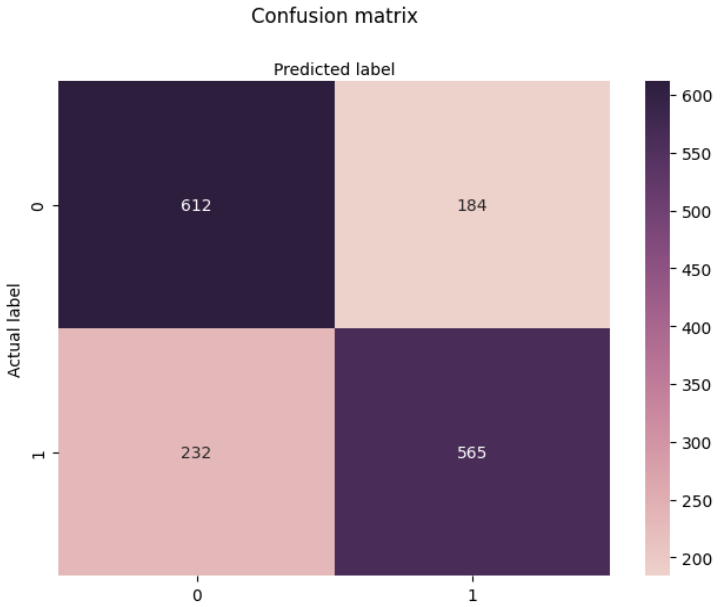

Figure 6.10: Confusion Matrix

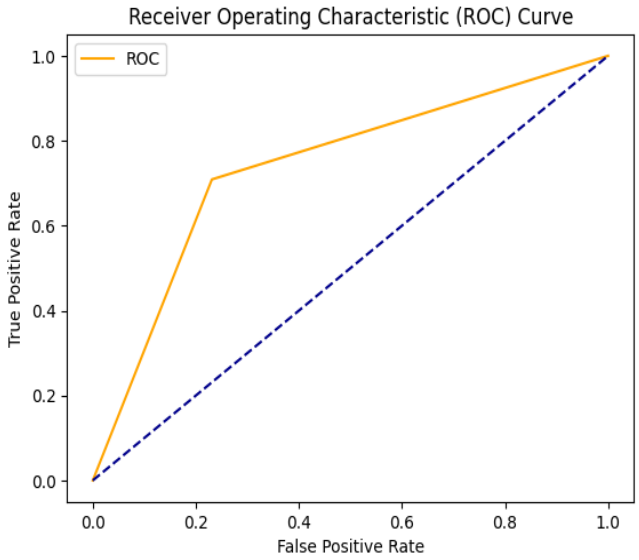

Figure 6.11: ROC Curve

6.6 GRU BILSTM

This model excels at capturing long-range relationships and subtle subtleties within sequential data by combining Gated Recurrent Units (GRUs) with Bidirectional Long Short-Term Memory (BiLSTM) layers. The GRU BiLSTM model was used in this project because of its ability to understand the temporal order and context of customer behaviors. This model beat others in correctly detecting instances of customer turnover within the supermarket dataset, with an accuracy score of 80% and an AUC score of 0.88. Its capacity to evaluate the sequence of client contacts and identify detailed temporal patterns greatly contributed to its effectiveness in anticipating probable instances of churn. As a result, the GRU BiLSTM model emerged as the best-performing model in this context, demonstrating its better predictive skills and establishing its position as the most effective tool for predicting customer attrition in the supermarket sector.

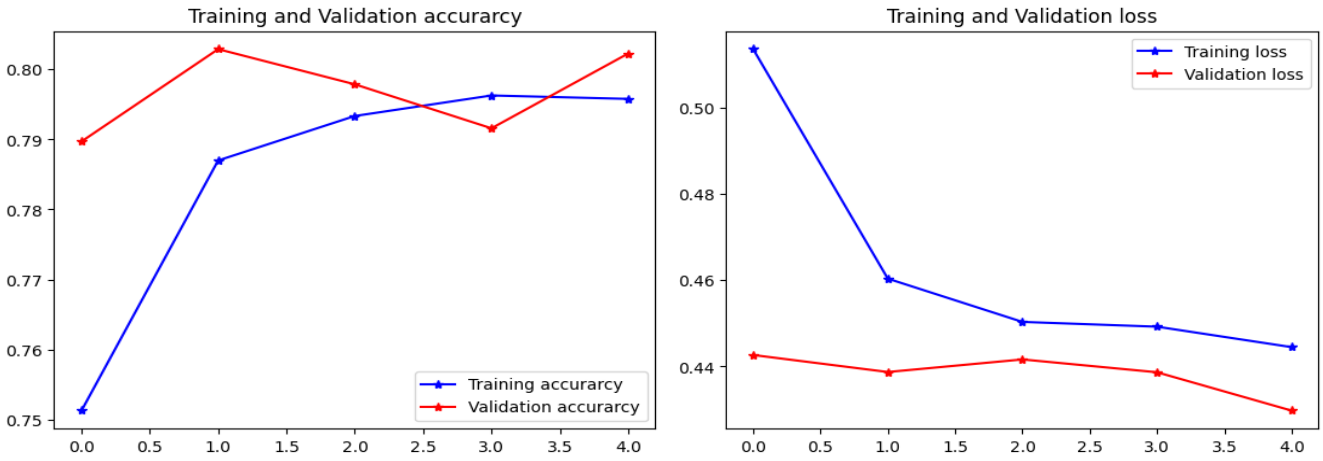

Figure 6.12: Accuracy & Loss Graph

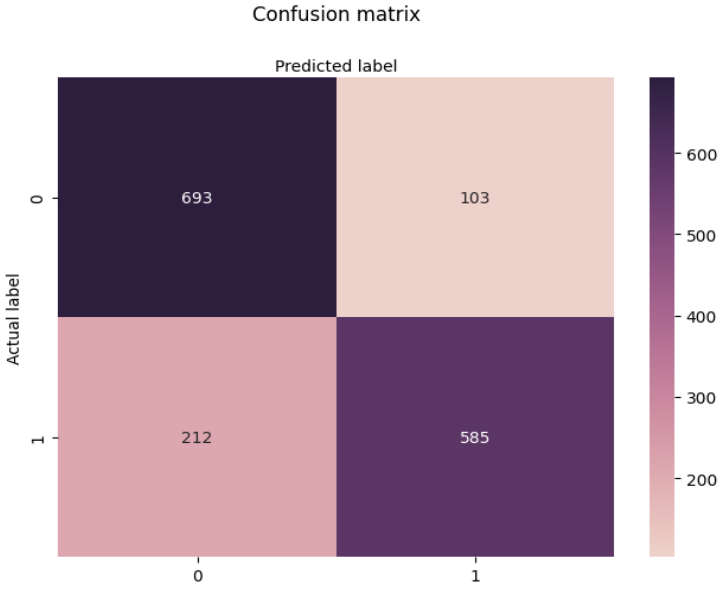

Figure 6.13: Confusion Matrix

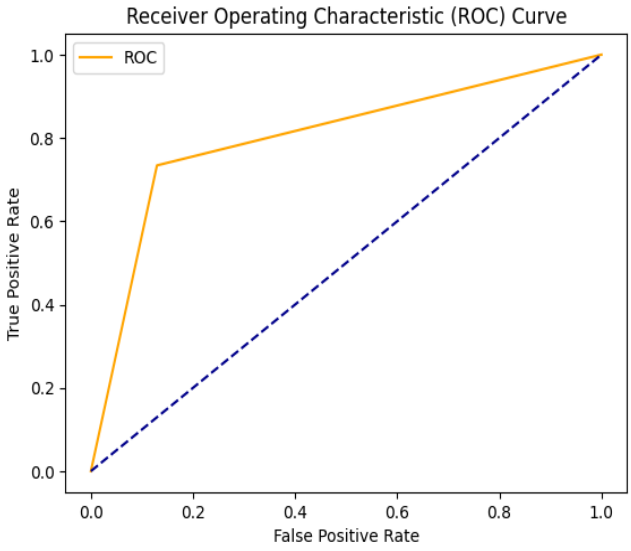

Figure 6.14: ROC Curve

Table 6.1: Model Performance Comparison: Accuracy Scores and ROC AUC Scores

Model	Accuracy Score	ROC AUC Score
Adaboost	0.7746	0.85
Random Forest	0.7696	0.84
XGBoost	0.737	0.80
CNN LSTM	0.739	0.82
GRU BiLSTM	0.80	0.88

Chapter 7 Conclusion and Future Works

7.1 Conclusion

This web application for supermarket consumer churn prediction has offered useful insights into customer behavior in the supermarket arena. We've developed considerable predictive skills in identifying possible instances of customer churn by utilizing several machine learning and deep learning models such as SVM, Adaboost, Random Forest, XGBoost, CNN LSTM, and GRU BiLSTM. With their various techniques, these models demonstrated variable degrees of accuracy and efficiency in forecasting churn. Notably, the GRU BiLSTM model performed best, with an accuracy of 80%, proving its ability to recognize sequential consumer behavior.

7.2 Limitations

This predictive approach, however, is not without flaws. The quality and completeness of the dataset may have an impact on the performance of the models. Furthermore, the predictive capability of the online application is mainly based on past data, which may not capture real-time changes or unanticipated market variables that affect customer attrition. Furthermore, although the accuracy attained is significant, it may not suffice in every real-world circumstance, and ongoing progress is necessary for strong forecasts.

7.3 Future Work

In the future, this prediction system might be improved by obtaining more comprehensive and real-time information to increase model accuracy and generalizability. External elements like economic statistics, consumer feedback, or seasonality trends might improve predictive power. Furthermore, improving the user interface and introducing more user-friendly features into the online application may improve its usability and accessibility to stakeholders. To react to changing consumer behavior patterns and market dynamics, continuous model review and modification are required.

References

1. Hu, X., Yang, Y., Chen, L. and Zhu, S., 2020, April. Research on a customer churn combination prediction model based on decision tree and neural network. In *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)* (pp. 129-132). IEEE.
2. Kim, S. and Lee, H., 2022. Customer churn prediction in influencer commerce: An application of decision trees. *Procedia Computer Science*, *199*, pp.1332-1339.
3. Odegua, R., 2020. Applied Machine Learning for Supermarket Sales Prediction. *Project: Predictive Machine Learning in Industry*.

4. De, S., Prabu, P. and Paulose, J., 2021, November. Application of machine learning in customer churn prediction. In *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)* (pp. 1-7). IEEE.
5. Lavanya, K., Aasritha, J.J.S., Garnepudi, M.K. and Chellu, V.K., 2023, February. A Customer Churn Prediction Using CSL-Based Analysis for ML Algorithms: The Case of Telecom Sector. In *International Conference on Innovative Computing and Communication* (pp. 789-804). Singapore: Springer Nature Singapore.
6. Hooda, P. and Mittal, P., 2023. IMPLEMENTATION AND PERFORMANCE ENHANCEMENTS OF OPTIMISED KERNEL MSVM MODEL FOR EARLY CHURN PREDICTION IN TELECOM SECTOR. *Semiconductor Optoelectronics*, *42*(1), pp.280-298.
7. Singh, N., Singh, P. and Gupta, M., 2020. An inclusive survey on machine learning for CRM: a paradigm shift. *Decision*, *47*(4), pp.447-457.
8. Günesen, S.N., Şen, N., Yıldırım, N. and Kaya, T., 2021, January. Customer churn prediction in FMCG sector using machine learning applications. In *IFIP International Workshop on Artificial Intelligence for Knowledge Management* (pp. 82-103). Cham: Springer International Publishing.
9. Suhanda, Y., Nurlaela, L., Kurniati, I., Dharmalau, A. and Rosita, I., 2022. Predictive analysis of customer retention using the Random Forest algorithm. *TIERS Information Technology Journal*, *3*(1), pp.35-47.
10. Abbas, W., Usman, M. and Qamar, U., 2022, December. Churn Prediction of Customers in a Retail Business using Exploratory Data Analysis. In *2022 International Conference on Frontiers of Information Technology (FIT)* (pp. 130-135). IEEE.
11. Baderiya, M.S.H. and Chawan, P.M., 2018. Customer buying Prediction Using Machine-Learning Techniques: A Survey.
12. Mallawarachchi, S.N., Rodrigo, M.N.D., Gunaratne, M.A.S.N., Gamage, M.P. and Qamra, N.N., A Web Application to Support Customer Churn Management for Retail Grocery Stores.
13. De Caigny, A., Coussement, K. and De Bock, K.W., 2018. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*(2), pp.760-772.
14. ÇALLI, L. and KASIM, S., 2022. Using Machine Learning Algorithms to Analyze Customer Churn in the Software as a Service (SaaS) Industry. *Academic Platform Journal of Engineering and Smart Systems*, *10*(3), pp.115-123.
15. Seymen, O.F., Dogan, O. and Hiziroglu, A., 2020, December. Customer churn prediction using deep learning. In *International Conference on Soft Computing and Pattern Recognition* (pp. 520-529). Cham: Springer International Publishing.
16. Hiziroglu, A., Customer Churn Prediction using Deep Learning.
17. Mudassar, L. and Byun, Y., 2019. Customer prediction on parking logs using recurrent neural network. *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pp.123-136.
18. Motevali, M., 2023. Predicting Customer Churn Based on Deep Learning, Neural Networks and Logistic Regression. *Mathematical Statistician and Engineering Applications*, *72*(1), pp.2180-2190.
19. Suh, Y., 2023. Machine learning based customer churn prediction in home appliance rental business. *Journal of big Data*, *10*(1), p.41.