Enhanced Customer Behaviour Analysis Using Stacking Classifiers for Churn Prediction

MSc Research Project
Data Analytics

Nayana Palathinkal Shibu
Student ID: x23174960

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Nayana Palathinkal Shibu		
Student ID:	x23174960		
Programme:	Data Analytics	**Year:**	2024
Module:	MSc Research Project		
Supervisor:	Noel Cosgrave		
Submission Due Date:	12/12/2024		
Project Title:	Enhanced Customer Behaviour Analysis Using Stacking Classifiers for Churn Prediction		
Word Count:	7804	**Page Count:**	22

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: Nayana Palathinkal Shibu

Date: 11/12/2024

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):	

Enhanced Customer Behaviour Analysis Using Stacking Classifiers for Churn Prediction

Nayana Palathinkal Shibu
x23174960

Abstract

Customer behaviour is defined as how individuals or customers make decisions regarding the use and purchase of products or any other services. This study aims to predict customers' behaviour and to use machine learning algorithms like Logistic Regression, AdaBoost Classifier, XGBoost Classifier and Stacking Classifier, to analyze their performance on the dataset used. Among the first three models, the XGBoost Classifier stood high in terms of accuracy and therefore required a more detailed analysis using hyperparameter tuning and GridSearchCV was applied to XGBoost Classifier. Among all the proposed algorithms, the Stacking Classifier developed using three base models, namely optimized XGBoost Classifier, Logistic Regression and AdaBoost Classifier, with XGBoost Classifier as the meta-classifier worked the best with an accuracy of 88% to predict customer behaviour. The performance of this system was further evaluated by comparing it with another methodology which used the dataset, "Ecommerce Customer Churn Analysis and Prediction". When the models were retrained on that same set of data that was used in the referenced study, the Stacking Classifier achieved precisely 99% accuracy, surpassing the 98% of the referenced study while the XGBoost achieved an average accuracy of 93%, with an even better 97% after cross-validation and 98% after hyperparameter tuning compared to 91% in that study. These results confirm the effectiveness of the proposed system, showing how it can be generalized across datasets and how it outperforms other models.

Keywords: Customer Behaviour, Machine Learning, Predictive Modelling, Customer Churn, Hyperparameter Tuning, Stacking Ensemble Model

1 Introduction

Customer behaviour prediction is an important area of the data analysis field that deals with the identification, study and modelling of customer's behaviour in buying or consuming goods or services (Zhang and Chang, 2021). It involves using some statistical methodologies to discover information about the use of products, services or brands by the consumers. Business organizations apply this information to fashion ways of improving customer satisfaction, commitment and involvement. This includes usage of psychological theory, economic theory, social theory, computational theory and employs modern computational methods such as machine learning and statistical modelling. Consumer behaviour prediction is the process of using historical data like purchase history, demographic data and web behaviour to improve marketing, product development and operational decisions. One major purpose of analytics in the business field is to forecast consumer needs and expectations so that the organizations can provide clients with tailored approaches, goods and services

(Mariani and Wamba, 2020), while controlling stock levels and increasing revenues. This prediction process normally involves the application of machine learning algorithms to understand patterns and make predictions such as the probability of purchase, churn or reaction to a promotional campaign. Prediction of customer behaviour is crucial in the current world of digital globalization because it helps to solve the problems of customer loyalty, recommending goods or services and using advertisement spaces. For example, subscription-based e-commerce organizations can determine their churn rate by using churn prediction and generate recommendations on the products to be offered to customers (Anudeep et al., 2024). Given the growing competition and rapidly changing consumer attitudes, companies experience a greater need for behaviour prediction.

The motivation for undertaking this study arises from the realization that, in the current dynamic market environment, customer retention has become an important strategic business goal than customer acquisition. Predicting customer behaviour related to churn provides insights about how customers behave and how a business can retain them to reduce the churn rate. Therefore, this research proposes the use of machine learning tools to predict which customers are at risk of attrition, thereby helping the organizations to improve the retention rates of the customers. By predicting churn, marketing resources can be utilized more efficiently and can also be used to understand the needs of the customers to increase loyalty towards the company. This study makes use of data analytics to assist various business units in the process of decision-making in today's competitive economic environment.

This research develops better customer churn prediction models by employing ensemble machine learning methods. Earlier works mainly applied separate models such as Logistic Regression or tree-based algorithms which although provided some satisfactory results but were not fully capable of handling intricate patterns. By using algorithms such as Extreme Gradient Boosting Classifier and Stacking Classifier, this work increases the accuracy of depicting inherent relations in the data and increases both the precision and recall rates for both the churn and non-churn outcomes. In addition to that, the use of GridSearchCV to fine-tune the models for the specific dataset alleviates the problem of setting inadequate hyperparameters as seen in some other works. Moreover, this work successfully fills a gap in comparative effectiveness by comparing the proposed system to another research that used a different dataset.

The research question for this research are as follows:

- What is the comparative effectiveness of ensemble machine learning models such as Stacking and XGBoost Classifier compared to other traditional models in predicting customer churn, in terms of accuracy?

The objectives of this research are as follows:

1. To compare the performance of various machine learning algorithms like Logistic Regression, Adaptive Boosting Classifier, Extreme Gradient Boosting Classifier and Stacking Classifier in predicting customer churn and to determine the best model.

2. To implement hyperparameter tuning techniques like GridSearchCV to increase the performance of the Extreme Gradient Boosting Classifier model.
3. To create and evaluate a Stacking Classifier model which combines different machine learning algorithms to improve prediction accuracy.

This report is organized into seven sections after the abstract, each addressing the key aspects of the customer churn prediction research. The first section introduces the background, motivation, research questions, research objectives and identifies the research gap. The second section presents a literature review which studies the previous research and methodologies in the field of churn prediction. The next section outlines the methodology describing the dataset, preprocessing steps and modelling. In the fourth section, the design specification is discussed which explains the workflow of the proposed system. The fifth section focuses on the implementation of the various machine learning models like Logistic Regression, Adaptive Boosting Classifier, Extreme Gradient Boosting Classifier and Stacking Classifier. The sixth section presents the results and evaluation which analyses the outcomes and compares the performance of the models, which includes a comparative analysis with another methodology. The final section concludes this study by summarizing the findings of this study, addressing the limitations and proposing directions for future research.

2 Related Work

2.1 Understanding Customer Behaviour Prediction

Customer behaviour prediction encompasses the way consumers decide to make choices and their usage patterns of a product or service (Sundararaj and Rejeesh, 2021). Some of the aspects featured in this process include buying behaviour which is defined as the customers' buying power, attitude and the propensity to buy (Chaudhuri et al., 2021), which is influenced by demographic influence, psychological influence and other factors like efforts by the marketers. When using this approach, firms can use past performance data to predict future behaviour and group customers in terms of their purchasing tendencies or likelihood of abandoning the business. Machinery like data mining and machine learning produce large datasets and look for the trends that support organizational management in its decision-making system. Moreover, analysis of the customer behaviour not only helps to increase customer satisfaction and loyalty (Sani and Febrian, 2023), but also helps with enhancing marketing advertising and customer segmentation process. As customer behaviour patterns expand and become more diverse, businesses can learn from these patterns and improve their marketing strategies (Haleem et al., 2022). By understanding the customer’s actions, long-term cooperation and stable development can be achieved by the organization.

2.2 Machine Learning Models for Churn Prediction in Customer Behaviour

There is a study which was proposed by Jain et al. (2020), who rightly pointed out one of the major problems, which is the customer churn problem in the highly competitive telecommunications industry where companies must constantly monitor customer behaviour for their retention. To predict customer churn, the model utilized two machine-learning techniques, Logistic Regression and Logit Boost. The results were depicted using several metrics and it shows that both the models had an accuracy of 85% which confirms the usefulness of the approach introduced for churn forecast as well as customer retention.

Another research done by Sunarya et al. (2024), examines a comparative analysis of Logistic Regression and Random Forest algorithms in customer churn prediction in the e-commerce customer dataset. A major issue that was encountered during the analysis was the lack of flexibility in identifying a broad range of factors that are likely to lead to customer churn. The evaluation metrics suggest that although Logistic Regression model had an accuracy of 90%, paralleled by some promising qualities in terms of precision and recall, Random Forest had a high accuracy of 95% as well as better precision, recall, F1-Score and AUC-ROC values. Apart from comparing the predictiveness and appropriateness of each model for customer churn, the study improves knowledge of machine learning in e-commerce contexts and this generalization is significant for stakeholders interested in the advanced understanding of customer retention. In addition to that, the work creates a research avenue for subsequent studies on the virtual social factors that impact customer behaviour, given the constantly shifting digital environment.

The third study (Thete and Sharma, 2023) worked on the telecommunications industry choosing the Logistic Regression analysis to identify complex relations affecting customer churn. The objectives of the study were to investigate the refined nature of customer behaviour and to determine which features directly lead to churn. To this effect, a Logistic Regression model has been used to establish a multifaceted correlation between the different influencing factors of customer churn. One of the issues that the research encountered is to model customer churn as a complex construct in the best way.

The study given by Senthan et al., (2021) suggested a forecasting model that aimed at identifying telecommunication consumers who are likely to leave the business and move to another service provider. The study used a dataset of 10000 post-paid customers' data containing 20 variables to measure customer characteristics and churn risk. Different algorithms of supervised machine learning including Decision Tree, Logistic Regression, Support Vector Machine and Artificial Neural Networks were used. In addition to that, ensemble algorithms like Random Forest, XGBoost, AdaBoost were also included in this study. The challenge of this study was to predict churn while keeping the bias and variance of the model in check, using the given dataset. From the results obtained, the XGBoost model gave the highest accuracy of 82.90% and further enhanced the accuracy level to 83.13% by hyperparameter tuning of the model.

Lastly, there is a study presented by Kim and Lee (2022) who investigated a model for estimating the rate of customer attrition which is relatively emerging and a novel field of electronic commerce and relates to the product promotion and selling of products through social media accounts. In this study, the author used a Decision Tree (DT) algorithm to predict customer churn by analyzing the results of the Rapidminer software program. The results using the F-measure came up to 90% of maximum prediction accuracy, which is good in terms of customer churn prediction from the influencer's standpoint and acts as the contribution to the further enhancement of customer service within this segment of the market.

2.3 Ensemble Methods and Stacking Techniques for Churn Prediction

Stacking techniques have been proven as an effective form of ensemble method for improving prediction accuracy across different fields. In the telecommunications industry, there is a study which was completed by Awang et al., (2021) and proposed a stacking ensemble model using six different machine learning algorithms as the base-classifiers to enhance the classification of customer churn. By incorporating the six base-classifiers with five meta-model classifiers, this approach was able to efficiently overcome the problem of low accuracy of the separate single classifiers, as well as depicted improvements in other aspects of prediction including accuracy, precision, recall as well as ROC.

Similarly, another study given by Hamza et al. (2024), developed a Heterogeneous Multi-layer Stacking Ensemble (HMSE) method tailored for customer churn prediction in the telecommunications sector, and it incorporated five machine learning classifiers which are Random Forest, Bayesian Network, Support Vector Machine, K-Nearest Neighbours and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) and addressed the problem of class imbalance in churn datasets by utilizing the method of Synthetic Minority Oversampling Technique (SMOTE) and compared the HMSE and the SMOTE integrated HMSE (S-HMSE) and both performed well in identifying churners over the traditional models and other methodologies, with a better prediction performance seen in S-HMSE method.

In the study given by Oladimeji et al. (2023), the authors improved customer retention in the banking sector with an improved stacking ensemble method based on the K-Nearest Neighbours, CART and Naïve Bayes as base classifiers and a logistic regression model as the meta classifier. This model had a good accuracy of 83%, much higher than similar research that had an accuracy ranging from 79% to 81%.

There is a study (Awasthi, 2023) which focused on customer churn prediction and intended to meet the most important problem of customer retention in the e-commerce organization by identifying the factors that can lead to churn. The proposed approach used a Stacking Classifier, which used ensemble learning with a meta-classifier that learned from the different base-classifiers such as KNN, Support Vector Machine, Random Forest and Decision Trees Classifier. This approach was designed to predict the customer churn using the features such as Gender, Tenure and other important attributes connected with customers. As seen in this research, the Stacking Classifier outperforms the individual machine learning models in terms of accuracy by 98.2%, Area under the ROC curve with 98.1% and F1-Score of 95%. In comparison to the Stacking Classifier, XGBoost had a lower accuracy of 91%. While implementing the Stacking Classifier, the model had a high predictive accuracy, but the technique might be able to overcome by the problem of high computational cost because of the usage of base learners, which might reduce the efficiency of this classification model in big data.

In conclusion, these investigations present that the stacking ensemble method works effectively in handling complicated prediction problems and enhances the churn categorization in various fields, providing a credible benchmark for future works in customer behaviour prediction.

3 Research Methodology

This research will focus on making the customer behaviour prediction using the "Customer Churn Dataset" from Kaggle by the application of CRISP-DM model. The first step is the business understanding phase which sets the goals for the project which is to make the customer churn predictions, and it involves defining customer attrition rates, segmenting customer behaviour and learning about the customer attrition profile. The next step is data understanding, where the dataset is analyzed by observing the structure of the data, assessing the quality and the completeness of the data provided in the dataset. Frequency distributions and other types of graphs and charts are used to get insights into customers regarding age, gender, account information and activity patterns. The data preparation phase involves eliminating irrelevant features, converting categorical features into numeric values using OneHotEncoder and oversampling techniques such as SMOTE for handling the imbalanced classes in the target variable, Exited. The presence of duplicate rows and any outliers were also checked. Finally, the dataset is split into training and testing datasets in a ratio of 90:10.

In the modelling phase, different types of machine learning algorithms are used to predict. Logistic Regression, AdaBoost Classifier, XGBoost Classifier and Stacking Classifier were used in this study due to its strength and capabilities to address the aspect of customer churn prediction. Logistic Regression can make clear predictions when working with binomial outcomes. It helps in relating input features to the likelihood of churn. AdaBoost Classifier is effective in enhancing weak learners and XGBoost minimizes the over-fitting of data, thereby contributing to the performance of the models. The Stacking Classifier was integrated to make use of the performance of each of the classifiers. As the base models, XGBoost, Logistic Regression and AdaBoost create different perspectives, while the meta-model XGBoost enhances the mean classifier's performance in Stacking Classifier to enhance interpretability, accuracy and precision. In this final phase of Evaluation, the models which have been developed are assessed based on the evaluation metrics to establish their capability of predicting customer behaviour. Evaluation of the model reliability and predictive powers are achieved using accuracy scores, a confusion matrix and the classification report. In addition to that, a comparative analysis of the models' outcomes is conducted to understand the advantages and limitations of the models.

3.1 Dataset Used in This Research

The dataset "Customer Churn Dataset" used for customer churn prediction includes all the necessary information about the customer's profile to analyze and predict churn. It consists of 10000 rows and 14 columns that define a customer's demographic data, monetary information and data on the customer's interaction with the business. RowNumber, CustomerId and Surname are used as key labels in the dataset but are not considered to predict churn. The CreditScore of a customer and Geography, or the customer's country, is useful to look for patterns of churn depending on the location. Demographic information can be inferred based on the Gender and Age columns, while the Tenure column states the period for which the customers have been associated with it. Balance and EstimatedSalary are two useful descriptive variables connected with customers' financial situation, and NumOfProducts measures how involved the customer is by analyzing the number of products purchased. Other features of behaviour such as HasCrCard and IsActiveMember can help in relating customers to their trend in activity level. Due to the variety of features available in this dataset, this dataset is well suited for modelling customer behaviour, determining potential churn drivers and designing customer retention strategies.

3.2 Data Cleaning and Data Preprocessing

While cleaning the dataset before using it for customer churn prediction model, some columns that do not have much value for the prediction are dropped, which includes attributes such as RowNumber, CustomerId and Surname as these truly represent IDs rather than elements that would help in predicting customer churn. After that, to find out the structure of the dataset and the data types of all the remaining columns, the information method, info() is used which shows the null and the data types for the whole dataset. This study also checked duplicate rows, and the outliers were checked using some visualizations and it was found that there were no duplicate rows and outliers.

During data preprocessing, the dataset is prepared and is made ready for model training for interpretation by machine learning models. Firstly, the categorical variables in the dataset, Geography and Gender are transformed into numerical values using OneHotEncoder from sklearn.preprocessing, which creates binary columns for each of the categorical columns depending on the number of unique features in that column so that there is no assumption of ordinality. This results in new features like Geography_France, Geography_Germany and Geography_Spain for the column, Geography whereas Gender_Female and Gender_Male for the Gender column. Moreover, the issue of class imbalance in the target variable, Exited was addressed using SMOTE from imblearn to balance the dataset which increased the number of rows in the dataset, ensuring both the classes had equal representation, which helps in reducing biased predictions during model training.

3.3 Data Visualizations

Data visualizations were done using Seaborn, while for more customized plots, matplotlib were utilized and for interactive plots, where one can zoom in and out to gain more insight, plotly was used.

The clearly labelled pie chart in Figure 1, "Count Plot of the column Exited" which demonstrates the count of the two classes with 0 and 1 values in the target column. The pie chart shows that the major portion of the pie chart which is coloured with blue and marked '0' occupies 79.6% of the data whereas the second and the small portion coloured with red, marked '1' occupies 20. 4% of the total data. This visualization shows that there are far more instances of class 0 as compared to class 1 (exited customers), with a difference of approximately four times. Similar distribution is seen in most real-life classification problems that include cases of customer churn, frauds or risks where an event of interest is rare, which is typically represented as 1, then an opposite, which is a normal case marked as 0.

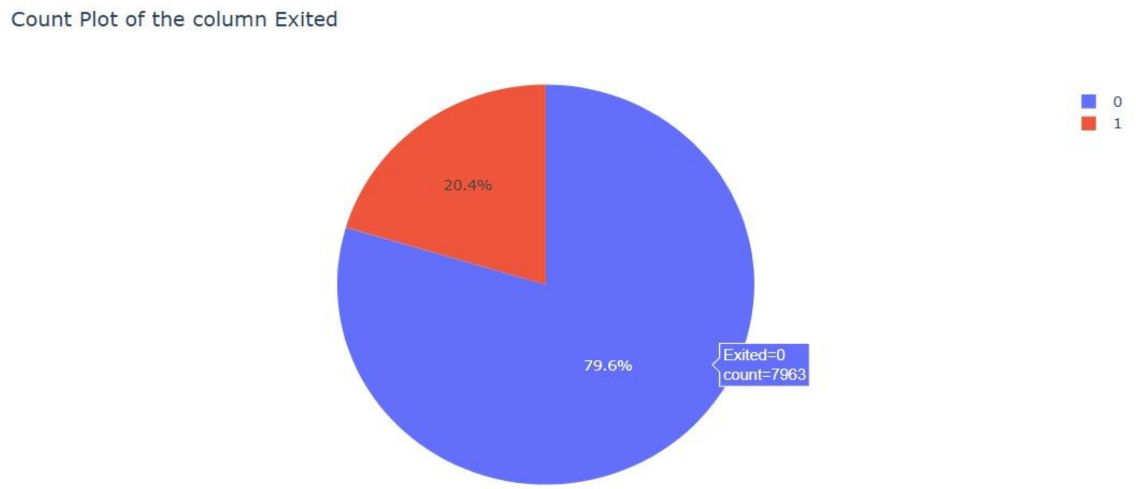

Figure 1: Pie Chart of the column 'Exited'.

The pie chart in Figure 2 is the Count Plot of Target Column after SMOTE Oversampling which shows how SMOTE has balanced the dataset by over-sampling the minority class of the target column. The current balance between the two classes can be seen as they are depicted with class 0 in red, and the class 1 in blue and they are taking a 50/50 share of the total amount of data in the dataset.

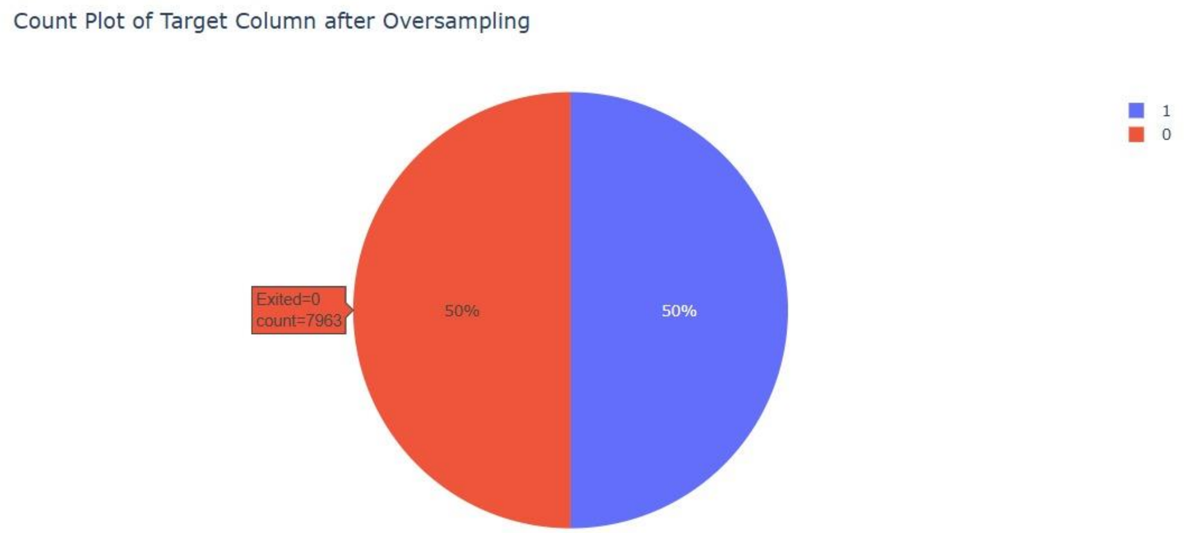

Figure 2: Pie Chart of the column 'Exited' after Oversampling using SMOTE.

Figure 3 analyses the count of customers exited based on the active membership status. From the figure, the non-active members, or the customers at 0 on the x-axis, means 3547 customers did not churn, whereas the green bar, about 1302 of the same customers churned. Among active members (1, on the x-axis), about 4416 customers were still loyal to the company while about 735 customers churned. This pattern clearly shows that the proportion of active members yields a higher retention rate, and a lower exit rate compared to the non-active customers, therefore it is understood that active membership status has escalated customer loyalty and retention rate.

Figure 3: Bar Plot comparing the columns 'IsActiveMember' and 'Exited'.

A histogram demonstrating the distribution of the column, Age has been presented in Figure 4 which gives the age distribution of the customers in the dataset. The histogram is designed as a series of horizontal bands that transition gradually from dark blue representing the youngest ages to yellow for the older ages. Age is categorized in the X-axis in years, between twenty and ninety years, while the Y-axis exhibits the count of the customers, with counts to five hundred in the central region which is 30 to 40 years of age. The distribution generally reflects right-skewed distribution; however, the largest group of customers lies in the 30 to 40 years of age with the highest density at 35 to 37 years with about 450-500 people. From the age twenty onwards, the level of frequency increases gradually, reaching a peak, and then declines and continues to decline in the older age groups. The data illustrates that participation sharply declines at 60 years and there are nearly none in the 70 to 90 age range.

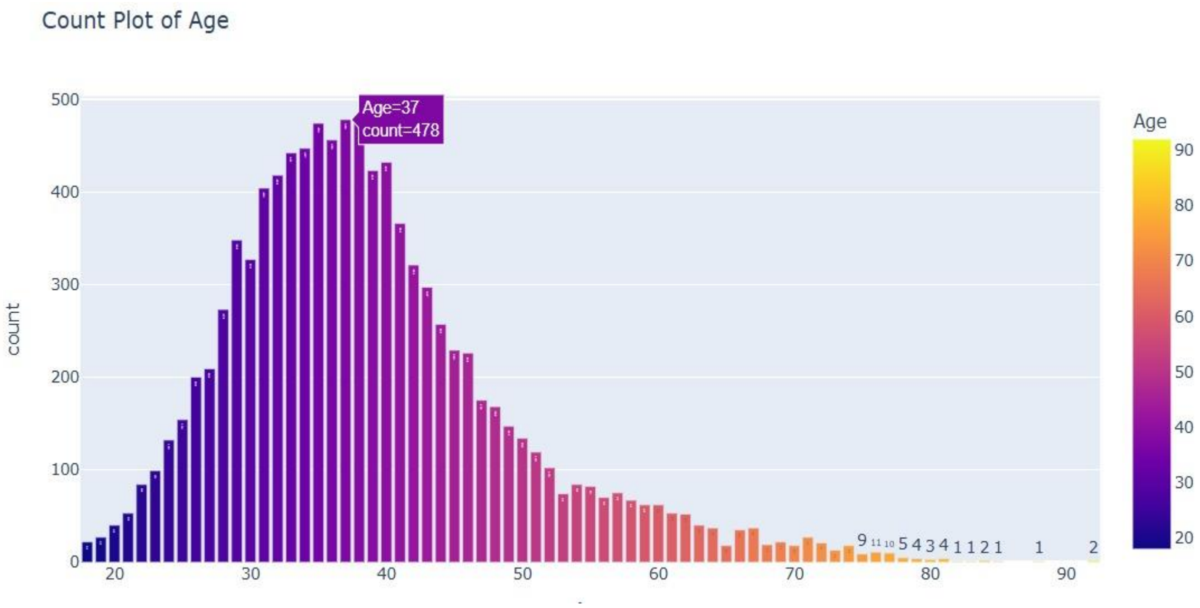

Figure 4: Bar Chart of the Column 'Age'.

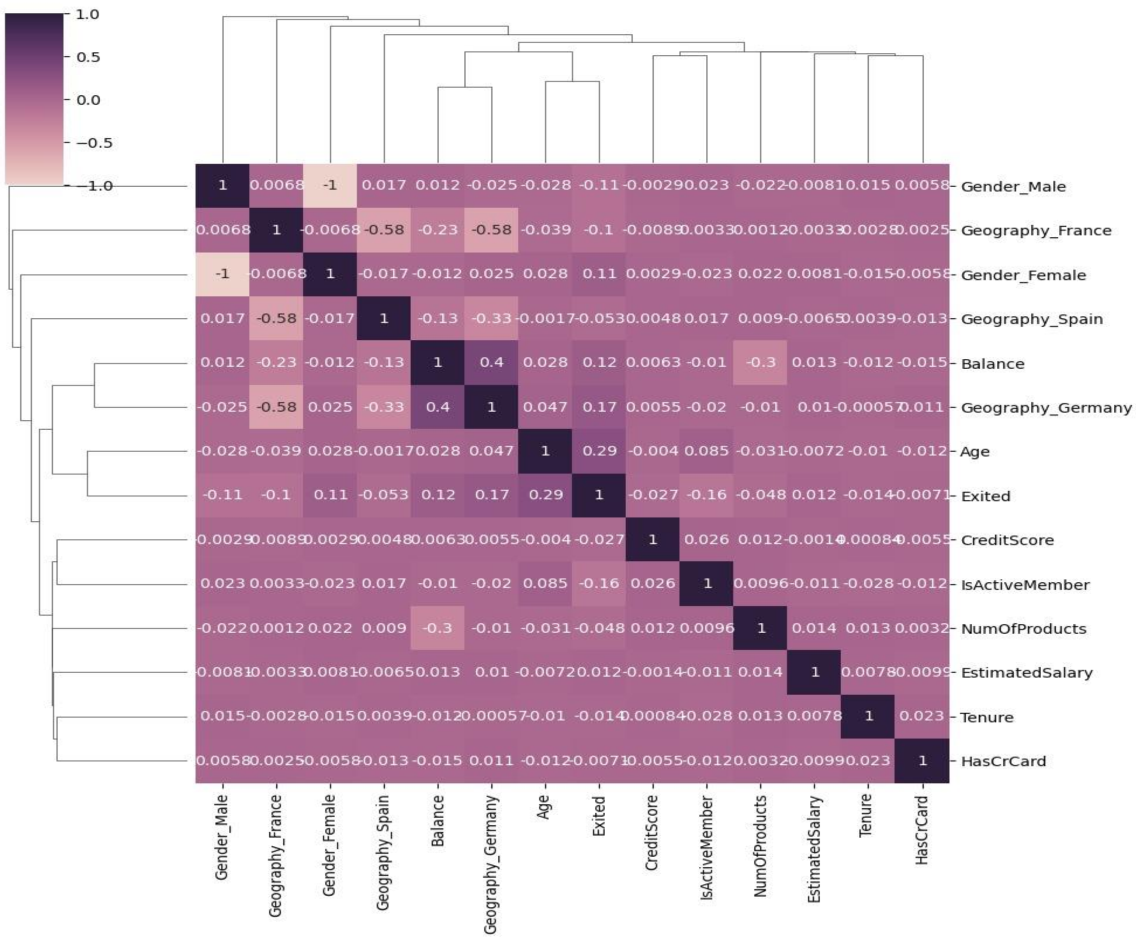

Figure 5: Correlation Plot.

In Figure 5, there is a heatmap for correlations with dendrograms for the features in the dataset. It uses a color gradient scale from dark purple when the correlation value is 1 to light pink when the negative correlation is present. It consists of some important variables like Age, Exited, Geography, Balance, NumOfProducts, CreditScore, IsActiveMember, EstimatedSalary, Gender, Tenure, and HasCrCard. Some of the correlations include a moderate positive correlation between Age and Exited columns, which is 0.29, which means that as the age increases, the customers are more likely to exit whereas the negative correlation between the IsActiveMember and Exited, which is -0.16, implies that the active member is less likely to exit, which is also evident as seen in Figure 3. When it comes to Balance and NumOfProducts, there is a negative correlation, -0.3, which states that customers with higher balance purchased fewer products. The variables clustered together at the top and the left side of the heatmap are automatically grouped from dendrograms, making it easy to recognize.

The bar chart in Figure 6, which is the Feature Importances graph, has been created using ExtraTreesClassifier to display the level of significance of the variables in predicting customer behaviour. The graph shows ten features ranked from left to right according to their decreasing importances, each bar corresponds to their importance scores whereas black vertical lines in each bar correspond to the standard deviation across all trees. Age has the highest feature importance, while HasCrCard, has the least importance score.

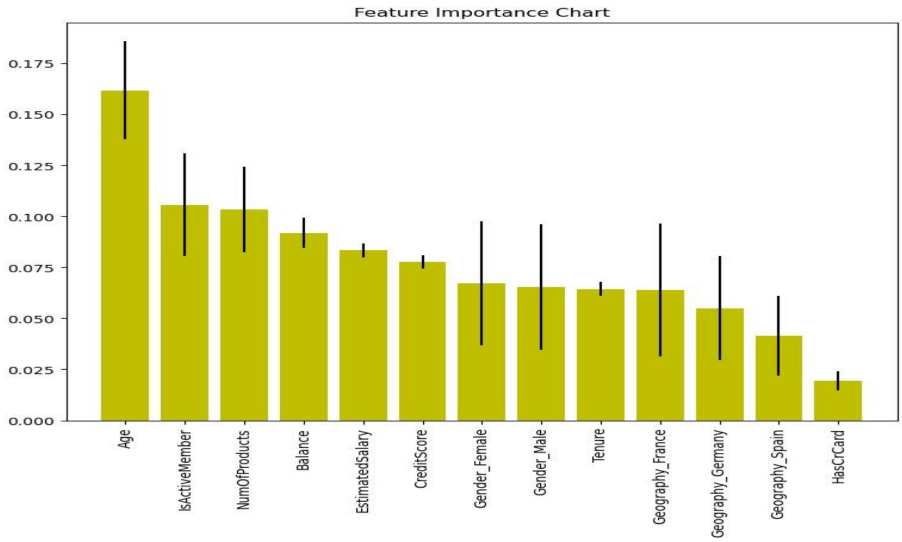

Figure 6: Feature Importance Chart.

3.4 Data Splitting and Modelling

The preprocessed dataset was divided into training and testing subsets using a 90:10 ratio to sustain the class distribution. This approach allowed the models to complete training on most of the data while also preserving the strict subset of the data for the testing. After the model training of Logistic Regression, Adaptive Boosting Classifier and Extreme Gradient Boosting Classifier, to evaluate the constructed model, cross-validation was used to divide the dataset into ten parts which helps to train on the nine parts and then test on the remaining one part. This approach reduces overfitting and calculates model performance, nevertheless, XGBoost yields the best mean CV accuracy, followed by AdaBoost Classifier and then Logistic Regression.

The tuning of hyperparameters was also done on XGBoost using GridSearchCV to find the best combination of parameters including the learning_rate, max_depth and n_estimators. GridSearch was utilized rather than other hyperparameter tuning techniques because of the size of the dataset utilized in this study. The dataset has a substantial number of rows and columns. Therefore, GridSearch provides a reliable result for improving the performance of XGBoost model for a manageable dataset size and features and ensures no optimal settings are missed. This optimization proved to enhance XGBoost's predictability by a good margin. Finally, the Stacking classifier was used to ensemble all the trained models of Logistic Regression, AdaBoost Classifier and the optimized XGBoost Classifier where the meta-classifier, the same optimized XGBoost Classifier will make its own conclusion from the results produced by the other models.

4 Design Specification

Figure 7 depicts the workflow of the suggested system that is followed in this research. This includes the input of the dataset, data preprocessing, feature importance analysis, data splitting and concluding by modelling and evaluation. The different classifiers included in the model training are Logistic Regression, Adaboost Classifier, XGBoost Classifier and finally the Stacking Classifier integrated with the Logistic Regression, Adaboost Classifier and the tuned XGBoost Classifier.

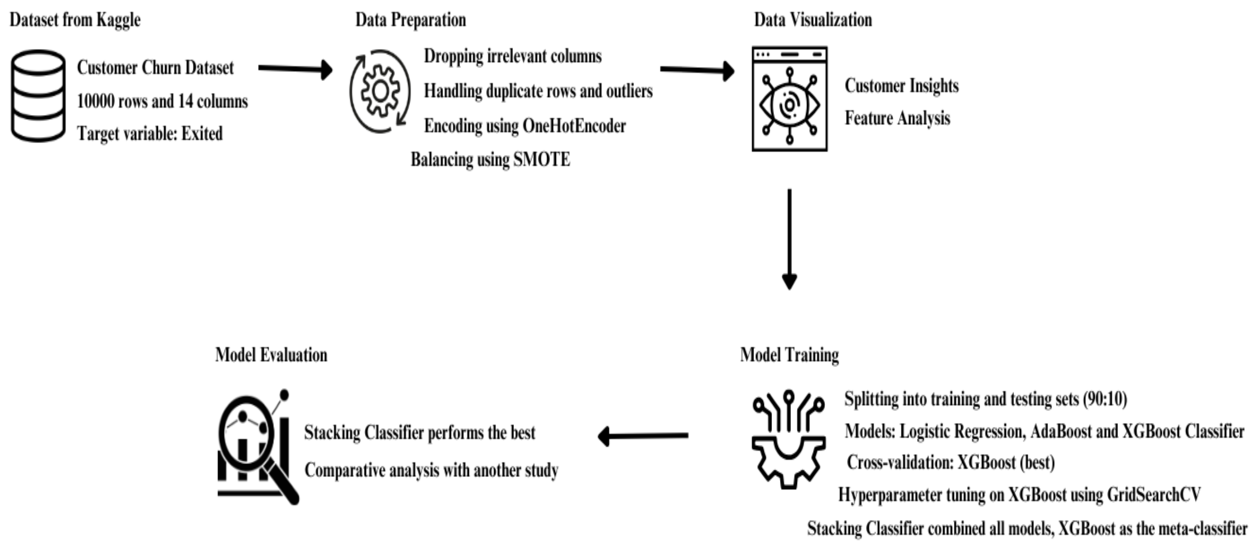

Figure 7: Workflow Diagram of the Proposed System.

The proposed system for the customer churn prediction as shown in Figure 7 consists of a workflow that is systematic, and it initially includes data collection which is sourced from the Kaggle website. The collected data is cleaned and preprocessed by transforming categorical variables using methods like OneHotEncoding and the dataset is balanced using SMOTE technique. In addition to that, Exploratory Data Analysis (EDA) is conducted to find out the characteristics of the dataset such as trends, relations and patterns in the data to gain insights about the characteristics of the dataset. Visualizations and the insights gained from EDA have been useful for determining influential drivers for churn prediction by choosing relevant features and designing a model.

The model training phase initially involved training three machine learning models which are Logistic Regression, Adaptive Boosting Classifier and Extreme Gradient Boosting Classifier. To improve the performance of the prediction, hyperparameter tuning is applied to the XGBoost Classifier using GridSearchCV to search for the best possible values of the parameters such as maximum depth, the learning rate and the number of estimators. The tuned XGBoost yielded better performance and modelling generalization. Finally, a stacking ensemble model is developed using Logistic Regression, AdaBoost Classifier and the tuned XGBoost Classifier as base classifiers with a meta-classifier to merge the forecasts which yielded the best results.

5 Implementation

For model training, sklearn.linear_model provided LogisticRegression, sklearn.ensemble provided the AdaBoostClassifier while the XGBClassifier under xgboost provided the gradient boosting technique. The StackingClassifier from mlxtend was also used to increase the accuracy of that classifier, as it uses the strengths of several classifiers for improved performance.

5.1 Logistic Regression

In Figure 8, the results achieved for the Logistic Regression model for customer churn prediction demonstrated a moderate performance. The confusion matrix as seen in the given figure, uses a heatmap where colors range from light pink to dark purple with the latter representing higher values of predictions. The true positives cell in the lower right corner of the matrix shows the correctly predicted number of customers who left as 534. The false positives count are 255 customers where the model has provided an inference that the customer churn is imminent while they have not churned and the false negatives were 262, where the model has suggested that the customers have not churned when in fact they have.

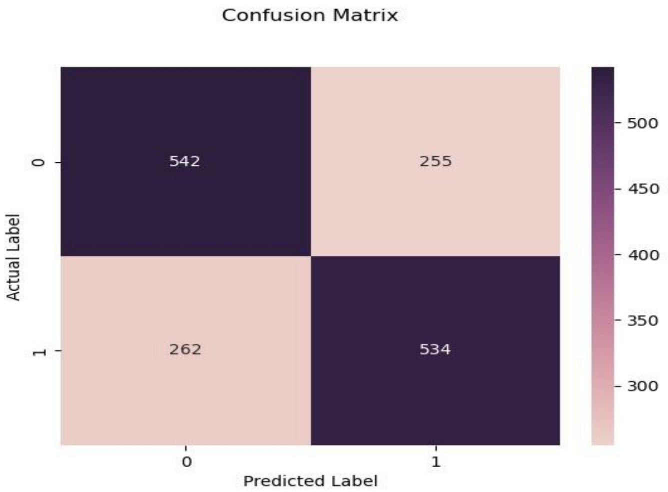

Figure 8: Confusion matrix of Logistic Regression.

5.2 Adaptive Boosting Classifier

Figure 9 given below gives the confusion matrix using the AdaBoost Classifier to analyze the effectiveness of the classifier model in customer churn prediction. The matrix reveals an impressive classification performance where true negatives in the first cell indicates 694 correctly predicted customers who did not churn and the true positives in the bottom-right quadrant show 674 were correctly predicted as leaving. The model's misclassifications are represented in the other cells where the false positives in the upper right quadrant are 103, where it is the count of customers that were predicted to leave but did not and the false negatives in the bottom-left quadrant, 122, where the customers were predicted to stay but left.

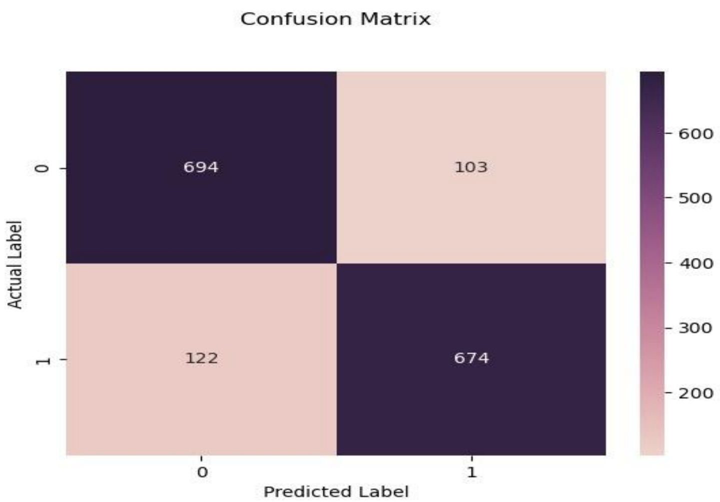

Figure 9: Confusion Matrix of Adaptive Boosting Classifier.

5.3 Extreme Gradient Boosting Classifier

Figure 10 displays the confusion matrix that shows the performance indicators for the XGBoost Classifier in customer churn prediction. The matrix shows an exceptional classification result. As seen in the figure, the true negatives cell on the top left of the confusion matrix indicates 725 correct predictions of customers who did not churn and the true positives cell on the bottom-right of the confusion matrix shows 662 correct predictions of customers who churned. The misclassification cases are lower where the count of false positives are 72 customers who were classified as churned but did not and the false negatives are 134 consumers who were classified as loyal customers but churned.

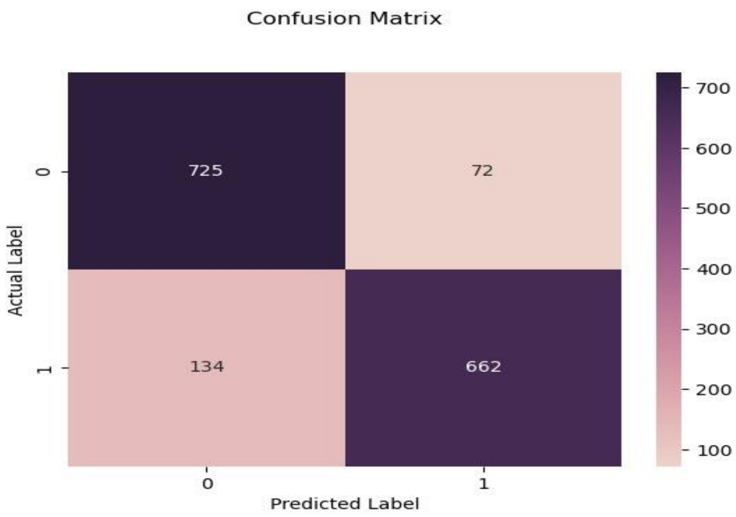

Figure 10: Confusion Matrix of Extreme Gradient Boosting Classifier.

5.4 Cross-Validation for Accuracy Comparison

A cross-validation was performed to analyze the performance of the three machine learning models, Logistic Regression, Adaptive Boosting Classifier and Extreme Gradient Boosting Classifier, based on the accuracy comparison using 10-fold cross-validation. This method ensures that the evaluation of the model is strong because the dataset is divided equally into ten portions and the model is trained on nine portions and then tested with the remaining one portion which is kept iterated. The Logistic Regression had a cross-validation accuracy of 68.74% indicating a moderate accuracy. A much-improved accuracy was seen in AdaBoost Classifier with an accuracy of 86.55%. Finally, the XGBoost Classifier gave the highest cross-validation accuracy of 88.92% which shows a better capacity of generalization in analyzing the customer churn.

5.5 Hyperparameter Tuning for XGBoost Classifier Using Grid Search

When choosing the best model and tuning the parameters for the classification, the focus was given to the best result in the customer churn prediction. Initially, three machine learning algorithms were built which were Logistic Regression, Adaptive Boosting Classifier and Extreme Gradient Boosting Classifier, among which, XGBoost Classifier showed the highest accuracy. For optimization of the XGBoost Classifier, GridSearchCV from sklearn.model_selection was used for tuning the hyperparameters of the classifier. A parameter grid with maximum depth, learning rate, and the number of estimators has been defined to determine the best parameter setting. From the results obtained, it showed that the tuned XGBoost model was able to achieve comparable best accuracy thus performing a better generalization of unseen data after cross-validation.

5.6 Stacking Classifier

Further extending the XGBoost Classifier approach, the parameters obtained from tuning were incorporated into a Stacking Ensemble Model which uses the tuned XGBoost Classifier, Logistic Regression and AdaBoost Classifier as base classifier models. This ensemble model was given a meta-classifier which was the tuned XGBoost Classifier. The stacking approach builds on the achievement of these individual models and thereby minimizes the bias as well as variance.

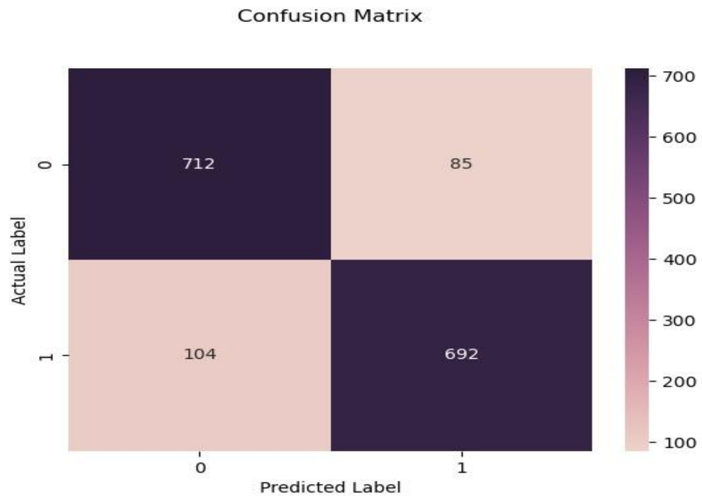

Figure 11: Confusion Matrix of Stacking Classifier.

The performance of the Stacking Classifier as seen in Figure 11 shows the highest performance. The first quadrant of the matrix, true negatives, includes 712 correctly predicted customers who remained loyal while the true positives show an outstanding 692 correct customer exits. Much fewer misclassifications are observed for this model where only 85 customers were falsely predicted to churn when they did not, while 104 customers are predicted to remain loyal when they defected.

6 Evaluation

6.1 Evaluation of Machine Learning Models on Customer Churn Dataset for Customer Churn Prediction

This section provides a detailed experimental evaluation of various machine learning models utilizing the preprocessed customer churn data. It also focuses on the comparison of the Logistic Regression, AdaBoost Classifier, XGBoost Classifier and the Stacking Classifier in determining customer churn. In addition to that, the trade-offs between the measures of accuracy, precision, recall and F1-scores for each model are discussed. Specifically, the Stacking Classifier that uses the tuned XGBoost Classifier, Logistic Regression and

AdaBoost Classifier as base classifiers and XGBoost as a meta-model reveals the highest accuracy and the best predictive capability. The focus remains on how each model enhances the understanding of which customers are at the risk of attrition and how to decrease error rates while offering insight into the efficacy of ensemble approaches into the general prediction mechanism. The performance of each of the models based on accuracy can be seen in Table 1.

Logistic Regression

The Logistic Regression model has obtained an accuracy of around 67% which seems reasonably moderate. The value for precision in predicting the non-churning customers was about 67% while that of the churning customers (class 1) was 68%. The recall or sensitivity was found to be 68% for the non-churned and 67% for churned customers while the f1-score was 68% for class 0 and 67% for class 1. These results suggest that the model has a reasonable forecasting precision but should be improved and it doesn't excel in capturing either of the classes because of its inability to capture more complex patterns in the data.

Adaptive Boosting Classifier

The AdaBoost Classifier was seen to perform better than the previous model, Logistic Regression with an accuracy of around 85%. The precision was 85% for the non-churned customers and 87% for churned customers. Regarding recall, for class 0, it was found to be 87% and for class 1, 85% while the f1-score for the two classes was computed to be 86%. It was therefore helpful in improving the accuracy of predicting customer churns due to its ability to handle the nonlinear relationships in the data, while at the same time increasing the level of precision, making it preferable for use in identifying customer churn risks.

Extreme Gradient Boosting Classifier

The XGBoost Classifier helped to increase predictive capability and increased the accuracy to almost 0.87 by outperforming both the Logistic Regression and the AdaBoost Classifier. Precision values were 0.84 for non-churned customers and 0.90 for churned customers. The classification report reveals strong precision and recall values for both the classes. This performance shows that the model is indeed strong and reliable when it comes to sorting customers for effective churn prediction.

Cross-Validation Results and Hyperparameter Tuning for XGBoost Classifier

The performance of the three machine learning models, Logistic Regression, Adaptive Boosting Classifier and Extreme Gradient Boosting Classifier was assessed using cross-validation by comparing accuracy and it was seen that the XGBoost Classifier had the highest cross-validation accuracy of 88.92% followed by AdaBoost Classifier and Logistic Regression. Thereafter, hyperparameter tuning with GridSearch was utilized to identify the best parameters for the XGBoost Classifier and this optimization has improved the performance of this model, reaching an accuracy score of 0.90.

Stacking Classifier

The Stacking Classifier was built from the base models, namely XGBoost Classifier, Logistic Regression and AdaBoost Classifier with XGBoost Classifier as the meta-classifier model.

For XGBoost Classifier, the best hyperparameters were found through hyperparameter tuning using a GridSearchCV, which improved the model's performance. Stacking was trained and the accuracy of the stacking model was determined to be 88%. This approach offered better precision, recall and F1-scores for both the churned and the non-churned classes. The stacking classifier also showed strong predictive performance and lower error rates than individual classifiers used above. Precision was the highest for this model at 89% for the churned customers and 87% with the non-churned customers. The recall rates for the two classes were 89% for class 0 and 87% for class 1 where the f1-score was 88% for both the classes. This model has combined the strengths of the individual models and enhanced the proportion of predictive accuracy while at the same time balancing between the number of true positives and false positives as well as true negatives and false negatives and hence can be said to have complemented the reduction of both.

Table 1: Performance of the Classification Models in Customer Churn Dataset

Classification Model	**Accuracy (%)**
Logistic Regression Model	67
Adaptive Boosting Classifier Model	85
Extreme Gradient Boosting Classifier	87
Stacking Classifier	88

6.2 Comparative Analysis with an Alternative Dataset

To evaluate the performance of the proposed system, a comparative analysis was conducted with a methodology presented by Awasthi (2023). The study mentioned has used the "Ecommerce Customer Churn Analysis and Prediction Dataset" available on Kaggle, to predict customer churn using advanced machine learning models. To ensure a fair comparison, the same dataset was utilized, and the proposed system in this research was implemented and trained using the same data preprocessing steps and machine learning algorithms on that dataset. The results obtained from this implementation demonstrated superior performance compared to the outcomes reported in the methodology by Awasthi (2023). Models used in our study not only enhanced predictive accuracy but also improved the balance between precision, recall and F1-Scores across all classes. This comparison highlights the robustness and the scalability of the proposed system when applied to a different dataset. The ability to outperform the results on the same dataset of the referenced study underscores the reliability of the ensemble-based approach in accurately identifying churn patterns and minimizing false predictions. The findings validate the proposed system's capability to generalize effectively while delivering better results than previous studies in the domain of customer churn prediction.

Table 2: Performance of Proposed System on Alternative Dataset

Classification Models	**Proposed System (Accuracy)**
Extreme Gradient Boosting Classifier	93%, 97% after cross-validation,

	98% after hyperparameter tuning
Stacking Classifier (XGBoost, Logistic Regression, and AdaBoost as Base Models and XGBoost as the Meta-Classifer)	99%

Table 3: Performance of Referenced System on Alternative Dataset

Classification Models	Referenced System (Accuracy)
Extreme Gradient Boosting Classifier	91%
Stacking Classifier (Decision Trees, Random Forest, K-Nearest Neighbors and Support Vector Machines as Base Models and Logistic Regression as the Meta-Classifer)	98%

The models in this proposed system demonstrated superior performance on the "Ecommerce Customer Churn Analysis and Prediction Dataset" compared to the referenced study (Awasthi, 2023). Specifically, the XGBoost Classifier achieved 93% accuracy, outperforming the 91% reported in the referenced study. After cross-validation, the XGBoost model further improved its performance, reaching 97% accuracy and 98% after hyperparameter tuning, showcasing the robustness and effectiveness of the proposed system. The Stacking Classifier developed in this research was using three base models, namely XGBoost Classifier, Logistic Regression and AdaBoost Classifier, with XGBoost Classifier as the meta-classifier. For the XGBoost Classifier, the best hyperparameters were determined through hyperparameter tuning using GridSearchCV, which enhanced the model's performance. In the referenced study, a Stacking Classifier was built using four base learners which were Decision Trees Classifier, Random Forest Classifiers, K Nearest Neighbors (KNN) and Support Vector Machines (SVM) with Logistic Regression as the meta-learner. The Stacking Classifier achieved 99% accuracy, compared to 98% in the referenced approach. These results highlight the improved predictive capability of the models in accurately identifying customer churn in our proposed system. Refer to Tables 2 and 3 for a detailed comparison of the performance between the proposed system and the referenced system on the alternative dataset used.

6.3 Discussion

The performance of the Stacking Classifier in the utilized dataset and compared dataset is because of its stacking property that enables it to combine the performance of the base models. This approach of fusion minimizes the bias and variance normally inherent in single models as it takes advantage of such models' strengths. The Stacking model's accuracy is due to XGBoost that identifies intricate patterns, AdaBoost that handles nonlinear relationships, as well as the interpretability of the Logistic Regression model. In addition to that, the meta-model, XGBoost, improves the rankings of the outcome of the base models. This combination of different classifiers resulted in a higher level of prediction and Stacking Classifier achieves better results than all the other models evaluated on the used dataset as well as the comparative study. Its strong performance can also be attributed to its lower chances of overfitting, improved performance and better hyperparameters by using cross-validation and GridSearchCV techniques and SMOTE balancing, which is also not seen in the referenced study. This approach not only helped to achieve the maximum level of performance for each model, but also provided evidence for the effectiveness of the ensemble methods to increase the level of accuracy and general reliability of predictions.

7 Conclusion and Future Work

This study was centered on customer behaviour prediction through the utilization of a variety of machine learning predictive models using a good quality global customer churn dataset. In the proposed system, three machine learning models were developed, and it was found that XGBoost Classifier performed the best with an accuracy of 87%. Thereafter, GridSearchCV was used for parameter optimization of the XGBoost model to identify the best hyperparameters. On this basis, a stacking ensemble model that further strengthened the effect of all the three models were created. This ensemble approach enhanced accuracy, and the final accuracy was 88%. It is understood that the stacking model can help identify the customers that are in danger of churn. The proposed system was then compared to the methodology that was described by Awasthi (2023) which showed that the aforementioned models could be outperformed by the proposed system when the system was used in analyzing and predicting the "Ecommerce Customer Churn Analysis and Prediction Dataset". This proposed system yielded 98% accuracy higher than the 91% accuracy for the XGBoost Classifier when using the same dataset. The Stacking Classifier also had an accuracy of 99% and was higher than 98% in the referenced approach.

Even though this study has discovered suitable machine learning classifiers for predicting the likelihood of a customer churn, the following limitations should be taken into consideration for future research. Firstly, it may be argued that the data employed in the analysis is extensive but the scope of customer behaviour under this study is not likely to be exhaustive. An additional collection of samples that present a wider range of customers and their behaviour could improve the models' stability. Furthermore, further research with more advanced ensemble methods, including blending or boosting with a larger number of algorithms, might demonstrate even higher performance. One limitation in this research is that there is no temporal analysis done because of the fluctuation in churn behaviour with time and therefore the inclusion of time series data would potentially increase the accuracy of the model. However, other variables, for example, comments or feedback, a customer interaction with the customer support section might be an important indicators of customer behaviour prediction and can be included in the future study. Therefore, additional features and deep learning could be applied in future studies to provide more details of complex data patterns. Finally, model interpretability has not lost its relevance for business use cases, therefore, additional efforts to increase the transparency of the model prediction can be beneficial to stakeholders seeking to apply retention strategies successfully. These limitations, if worked around, may lead to better churn prediction models that might be of greater practical use.

References

Anudeep, C., Venugopal, R., Aarif, M., Valavan, A.T., Vuyyuru, V.A. and Muthuperumal, S., 2024, June. Predicting Customer Churn in E-commerce Subscription Services using RNN with Attention Mechanisms. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

Awang, M.K., Makhtar, M., Udin, N. and Mansor, N.F., 2021. Improving customer churn classification with ensemble stacking method. *International Journal of Advanced Computer Science and Applications*, *12*(11).

Awasthi, S., 2023. Customer Churn Prediction on E-Commerce Data using Stacking Classifier. *Authorea Preprints*.

Chaudhuri, N., Gupta, G., Vamsi, V. and Bose, I., 2021. On the platform but will they buy? Predicting customers' purchase behavior using deep learning. *Decision Support Systems*, *149*, p.113622.

Haleem, A., Javaid, M., Qadri, M.A., Singh, R.P. and Suman, R., 2022. Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*, *3*, pp.119-132.

Jain, H., Khunteta, A. and Srivastava, S., 2020. Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Computer Science*, *167*, pp.101-112.

Kim, S. and Lee, H., 2022. Customer churn prediction in influencer commerce: An application of decision trees. *Procedia Computer Science*, *199*, pp.1332-1339.

Mariani, M.M. and Wamba, S.F., 2020. Exploring how consumer goods companies innovate in the digital age: The role of big data analytics companies. *Journal of Business Research*, *121*, pp.338-352.

Oladimeji, O.M., Ajiboye, A.R. and Usman-Hamza, F.E., 2023. An optimized stacking ensemble technique for creating prediction model of customer retention pattern in the banking sector. *Gadau Journal of Pure and Allied Sciences*, *2*(1), pp.22-29.

Sani, I. and Febrian, W.D., 2023. Marketing Analysis And Customer Satisfaction On Customer Loyalty Global Competition Era. *Jurnal Perspektif Manajerial Dan Kewirausahaan (JPMK)*, *3*(2), pp.111-120.

Senthan, P., Rathnayaka, R.M.K.T., Kuhaneswaran, B. and Kumara, B.T.G.S., 2021, April. Development of churn prediction model using xgboost-telecommunication industry in sri lanka. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1-7). IEEE.

Sharda Thete, P. and Sharma, D.K., 2023. Customer Behavior in Telecom: A Machine Learning Approach to Predictive Modeling and Analysis of Churn with Logistic Regression. *International Journal of Multidisciplinary Innovation and Research Methodology, ISSN: 2960-2068*, *2*(4), pp.1-5.

Sunarya, P.A., Rahardja, U., Chen, S.C., Lic, Y.M. and Hardini, M., 2024. Deciphering digital social dynamics: A comparative study of logistic regression and random forest in predicting e-commerce customer behavior. *Journal of Applied Data Sciences*, *5*(1), pp.100-113.

Sundararaj, V. and Rejeesh, M.R., 2021. A detailed behavioral analysis on consumer and customer changing behavior with respect to social networking sites. *Journal of Retailing and Consumer Services*, *58*, p.102190.

Usman-Hamza, F.E., Balogun, A.O., Amosa, R.T., Capretz, L.F., Mojeed, H.A., Salihu, S.A., Akintola, A.G. and Mabayoje, M.A., 2024. Sampling-based novel heterogeneous multi-layer stacking ensemble method for telecom customer churn prediction. *Scientific African*, *24*, p.e02223.

Zhang, J.Z. and Chang, C.W., 2021. Consumer dynamics: Theories, methods, and emerging directions. *Journal of the Academy of Marketing Science*, *49*, pp.166-196.