Customer Segmentation and Churn Prediction in Telecommunication Using Machine Learning

MSc Research Project

MSc. Data Analytics

Susanth Kammili

Student ID : X22248064

School of Computing

National College of Ireland

Supervisor: Professor. Noel Cosgrave

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Susanth Kammili ……. ………………………………………………………………………………………………………
Student ID:	… x22248064……………………………………………………………………………………………… ….…….
Programme:	MSc. Data Analytics **Year:** 2023-2024…………………..
Module:	Research Project ……………………………………………………………………………………………….………
Supervisor:	Prof. Noel Cosgrave ……………………………………………………………………………………….………
Submission Due Date:	03-01-2025
Project Title:	Customer Segmentation and Churn Prediction in Telecommunication Using Machine Learning ……………………………………………………………………………………………….………
Word Count:	7642 ……………………………………… **Page Count**…………26……………………………….……..

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.

Signature: …………Susanth Kammili………………………

03-01-2025

Date: ....

<u>ALL</u> 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.

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):	

Customer Segmentation and Churn Prediction in Telecommunication Using Machine Learning

Susanth Kammili

Student ID: x22248064

Abstract

Customer churn is a significant challenge for businesses, particularly in the telecommunications sector, where customer retention plays a crucial role in profitability and growth. This study addresses the problem of predicting customer churn using machine learning models, with an emphasis on both accuracy and interpretability. The research aims to develop and evaluate four machine learning models—logistic regression, random forests, gradient boosting, and LSTM networks—to predict customer churn. Further, the present research incorporates explainability tools like SHAP and LIME that would help in identifying what the algorithms look as the cause of churn to make the models efficient as well as to help the business executives understand them. The results show that the gradient boosting delivers the highest accuracy of 0.83% and the AUC-ROC of 0.90%; nevertheless, random forest and logistic regression models are also important and precise. While LSTM networks are quite effective, they failed because of the time agnostic data set used in the study. The implementation of SHAP and LIME also indicated that “Tenure,” and “Monthly Charges” were important features to distinguish churn. In this research, the following advancements were made 1) Moving from binary PAM and model performance metrics to show an integrated solution with an interval of interpretability and performance. The study also points to the lack of dynamic and temporal data as the way forward in improving the performance of churn models in practice.

1 Introduction

Customer attrition or the instances where the customers are no longer using a given telecom product or service is one of the most important problems that companies face, and due to the nature of the telecom business, its importance cannot be overemphasized. In the context of such sector as telecommunications, which is traditionally characterized by rather high level of competition, costs incurred while attracting new customers are much higher than in customer retention – this is why churn prediction became one of the most important fields of research. Literature review also indicates that by minimizing customers' churn by using some specific retention methods, organizations will have higher revenues per year and more stable markets. Knowledge of customer attrition and ability to forecast it enables a firm to prevent customer dissatisfaction and reduce customers' desertion rates. But it is still posing as a challenge even today because there are businesses where the amount of historical data is not adequate, some still rely on mechanism models for churn prediction, and most of the machine learning models lack interpretability.

The research question for this study is: How can machine learning models effectively predict customer churn while providing interpretable insights to inform retention strategies in the telecommunications sector? This question is fundamental to addressing the challenges faced by businesses in deploying churn prediction systems that are both accurate and actionable. The objectives of the study are:

i. To develop and evaluate machine learning models, including logistic regression, random forests, gradient boosting, and LSTM networks, for predicting customer churn in a telecommunications setting.
ii. To identify and evaluate the most important features that contribute to customer churn using interpretability techniques such as SHAP and LIME.
iii. To analyse the performance of the models in terms of predictive accuracy and interpretability, comparing them against established methods in the literature.

The hypotheses guiding the research are:

— H1: Machine learning models, particularly ensemble models like gradient boosting, will outperform traditional models like logistic regression in terms of predictive accuracy for churn prediction.

- H2: Interpretability tools such as SHAP and LIME will effectively provide insights into the key drivers of customer churn, enabling businesses to implement actionable retention strategies.

Thesis Structure

The document is structured into seven chapters: Chapter 2 reviews existing literature on churn prediction, machine learning models, and interpretability techniques; Chapter 3 outlines the methodology for data collection, preprocessing, model development, and evaluation; Chapter 4 presents the findings and critically analyzes them in relation to the research objectives; Chapter 5 describes the deployment of the final model and the tools used; Chapter 6 provides a detailed evaluation of the model's performance and its implications; and Chapter 7 concludes with a summary of key findings, research limitations, and recommendations for future work and practical applications.

2 Related Work

2.1 2.1 Machine Learning and Deep Learning Approaches for Churn Prediction

Customer churn prediction has been a trending topic in the literature, and different studies have used machine learning and deep learning approaches to solve this problem. Scholars have used various classifiers on data mining outcomes of prior work that include logistic regression, decision trees, and support vector machines (SVM) in handling churn with acceptable performance. For example, Amin et al., (2017) used logistic regression, SVM, and obtained moderate accuracy for predicting churn in telecommunication industry but low interpretability. The problem with these models is that they are relatively easy to compute and understand, but are inadequate for identifying interactions between the features or nonlinear relationships.

Thus, utilizing an ensemble of decision trees has become more efficient in comparison to the models random forest and gradient boosting. Random forests seem to be less sensitive to noisy data and overfitting as mentioned that Pondel et al. (2021) used random forests with a customer dataset and improved model accuracy immensely of 74%. Other tree-based ensemble methods like Gradient boosting models like XGBoost & LightGBM were also found to perform well. Traditional models for churn prediction include Decision trees, Logistic Regression, Naive Bayes, K-nearest neighbour and support vector machines However, Naser et al. (2023) compared these models for churn prediction and identified that gradient boosting outperforms all traditional models of data having high dimensionality.

However, Long Short-Term Memory (LSTM) networks that are more appropriate for sequential data according to Nogueira-Rodríguez et al. (2021) in the study of machine learning predictions. However, they are generally complex and require high computational power and most of the deep learning models do not have clear mechanisms of how they arrived at the decision. Pondel et al. (2021) pointed out this trade-off when using the LSTM model; theirs had high accuracy but comparatively high computational cost and lots of data preprocessing.

However, these approaches are not very interpretable, and their performance is bound to the fixed dataset. Of greatest concern is the fact that most research models do not capture the idea that customers are not passive but constantly evolve, while data is treated as static, rather than a stream that is constantly flowing. Moreover, while ensemble methods and deep learning models offer improved accuracy, they often sacrifice interpretability, hindering their adoption by non-technical stakeholders.

2.2 Interpretability Techniques in Predictive Modelling

Model interpretability has become a more significant issue as the models themselves become more complex. This decreases the likelihood of invalid or counterproductive conclusions since interpretability makes results understandable and believable by most stakeholders. SHAP and LIME additive techniques are common tools for explaining machine learning models at the present time. Feature importance is defined by SHAP, with reference to the theory of cooperative games. Dakkak et al. (2021 presented SHAP for two reasons: as a way of assigning contribution values to all the features within a given prediction and as a technique that can demonstrate global and local interpretability. For example, while doing churn prediction, SHAP can tell exactly how business loan tenure or and/or monthly charges contribute towards churn likelihood. Despite the common praise for the procedural reliability and theoretical underpinnings of SHAP, this approach can be resource-intensive for large data sets or intricate models, and thereby become computationally cumbersome (Lundberg & Lee, 2017).

LIME, a technique introduced by Ribeiro et al. in 2016, designed to explain the predictions of any machine learning model. It works by approximating the behavior of a complex, "black-box" model (like deep learning or ensemble models) with a simpler, interpretable model (such as a linear regression or decision tree) in the vicinity of a specific prediction. LIME has been applied in churn prediction problems to explain specific predictions and determine essential drivers of churn. Nevertheless, it has limitation of perturbation-based sampling that is less reliable for certain applications due to higher variation. More specifically, other methods for increasing model interpretability include feature importance rankings and partial dependence plots. Kumar et al., (2021) uses the partial dependence plot to analyse the influence of the features on churn probability taking into account customer behaviour. However, these methods are often constrained in the level of interpretation in the global sense, and do not have the specificity to lead to actionable predictions at the individual level.

However, there are several shortcomings related to interpretability techniques: they are not used as an integral part of the predictive process most of the time. Some researchers are centred on the predictive accuracy and not the interpretability of the results especially where the predictions are going to affect very important decisions to the company such as the retention of the customers. Furthermore, while SHAP and LIME are effective for tree-based models, their application to deep learning models remains an area of active research. The study by Sundararajan et al. (2017) proposed Integrated Gradients as a method for interpreting neural networks, but its adoption in churn prediction has been limited.

2.3 Summary of Findings

The review of existing literature highlights significant progress in customer churn prediction and model interpretability. Traditional machine learning models offer simplicity and interpretability but lack the sophistication needed to handle complex datasets. Ensemble methods and deep learning models address this gap by providing superior accuracy, but they often sacrifice interpretability and require substantial computational resources.

Interpretability techniques such as SHAP and LIME bridge the gap between model performance and usability, enabling stakeholders to trust and act upon predictions. However, their integration into predictive workflows remains inconsistent, and their application to deep learning models is still evolving.

The limitations of previous work underscore the need for a comprehensive framework that combines robust predictive modelling with actionable interpretability. This study addresses these gaps by leveraging a hybrid approach that balances accuracy and interpretability, providing a scalable and practical solution for customer churn prediction in the telecommunications sector.

3 Research Methodology

3.1 Research Design

Data collection and preparation together with feature engineering are defined by the research design, which forms the basis for the accomplishment of the goals set by the study. The information utilized in this research originates from the Telecom Customer Churn Dataset derived from Maven Analytics with records more than 7000 and 38 independent variables. These features include characteristics of the subscribers, services subscribed and the customer behaviour patterns, some of which include; age, gender, number of months subscribed, monthly charge and whether they are a churner or not. The structure of the dataset allowed studying customer characteristics and their actions towards churn prediction and modelling of customer segments.

Pre-processing of data was a very important dimension planned for the research, which has to do with the quality and usability of the data for machine learning algorithms. In this phase, those records with missing values were initially cleaned through inclusion of particular values in the data set, if those values were influential in the analysis, or were omitted otherwise. For example, the field such as "Churn Category" and "Churn Reason" which are believed to have a high percentage of missing values were omitted completely for the analysis. Then the categorical data variable features like "Gender", "Contract Type" was converted into forms understandable by the machine learning algorithm through one-hot encoding of label encoding.

Standardization methods were used as feature scaling on numerical data such as "Monthly Charges" and the "Total Revenue". It served the purpose of placing all features onto a constant scale and to avoid situations dominated by large magnitude features. Further, feature selection was used to select potential predictors for churn and segmentation include statistical tests and prior knowledge of the domains. Fluencies that were deemed unnecessary or that may prove to slow down the working of the models were eliminated. Lastly, the data set was divided into training and testing sets. The split ratio of 80:20 ensured sufficient feature set for training the models while still keeping aside a good sample size that can be used to test the models. This organized process made the data ready for more exploratory analysis as well as developing models to be presented in the paper.

3.2 Exploratory Data Analysis (EDA)

In this research, EDA was done to find more information on the data, and to establish the correlation between variables and variables and between variables and observations as well. There were three steps in the EDA process that included the first level EDA, second-level EDA and the third level EDA each having a particular purpose for creating an overall understanding of the dataset. In the univariate analysis, attention was paid to the dispersion of individual characteristics, to find temporal trends and outliers in the data. For example, tenure variable was skewed to the right; this suggests that the majority of customers have short subscription periods. In bivariate analysis more emphasis was placed on the associations between two variables and especially the association between the independent variables and the dependent or outcome variable, “Churn Status.” For instance, customers who signed contracts that allowed them to disengage at any time during the month were much likely to churn than customers who signed for the entire year. These relations were depicted through the help of heat map and correlation matrices and it was concluded that the tenure has got very high correlation with monthly charges while the age has only a moderate correlation.

Multivariate analysis further enhanced these findings given that it explored on cogon relations between multiple more than two test variables. Correlations like pair plots as well as clustering helped us try to identify emerging patterns, which may impact churn. For example, customers with higher monthly charge and tenure less than six months were more likely to churn indicating their dissatisfaction towards the price plan. Further, self-organizing maps extended clustering analysis to categorize customers according to pattern similarity suggesting initial segmentation schemes. These EDA results also helped in feature selection and in building up models, though pointing out to the most influential variables and their interactions. For instance, there is analysis of the variables such as “Tenure,” “Monthly Charges,” and “Contract Type” to provide the churn predictions. Outliers were also considered and managed in a proper way so that the potential influence on the overall design of the system might be reduced. Some of the findings from the EDA were useful for defining the customer segments and for modelling churn prediction.

3.3 Customer Segmentation Techniques

The customers were grouped by using customer segmentation to define and act for customers with similar features for retention and marketing strategies. There are two types of clustering techniques which were used for this purpose namely K-Means and hierarchical clustering. These methods of clustering split the customers by the number of attributes such as age, tenure, amount of monthly charges, and customers’ behaviour pattern. The clustering process was preceded by feature selection and scaling to allow the features used to contribute equally in the distances calculated. To be able to place the clusters in easily understandable and manageable two-dimensional space, Principal Component Analysis (PCA) reduction was done. This step also made it easier to understand clusters by pointing out the major factors which caused differences in the data.

Based on the observation of the elbow method and silhouette scores, the authors identified the best number of clusters. The elbow method was used to analyse the within-cluster sum of squares for the number of clusters from a particular range and concluded that the number of clusters beyond which the fit is not increasing much is considered as the optimum number of clusters. Silhouette scores provided additional confirmation of the high quality of the clusters and assessing the internal density within the clusters and between them. A silhouette score of more than 0.6 means highly understandable clusters. These clusters found above were then examined for the differences around them. For instance, one of the clusters mainly consisted of loyal customers with relatively low charges, and low risk of churn, and the second cluster consisted of valuable customers with high risk of churn. Post hoc tests which included ANOVA were also done to verify the differences between the various clusters. These segmentation results were very useful in matters of identifying customer variation and coming up with retention programs. To the extent that groups can be described individually as high or low risk, they could be offered promotional offers or new service strategies. Furthermore, the segmentation discussion above was useful in making features that need to be included in the churn prediction models.

3.4 Churn Prediction Models

Churn prediction formed the primary focus of this study and was arrived at with the help of several machine learning as well as deep learning models. Those include logistic regression, decision trees, random forests, recurrent neural networks (RNN), and long short-term memory (LSTM) networks were used for customer churn prediction. All proposed models were compared based on their predictive performance, level of explanation and computational cost. Logistic regression was selected as the first model to be used as a baseline owing to its ease of use and interpretability. It also segmented customers by churn potential and gave indications of the strength of feature predictors. Decision trees allowed for decision mapping and the ability to see how certain features helped in churn predictions in appearance. While decision trees were extended and optimized in random forests, the major drawback of decision trees was their overfitting, and to reduce this problem, random forests also average the output of multiple trees.

Long short-term memory (LSTM) and recurrent neural networks (RNN) networks were used to identify sequential behaviour patterns of customers. These models performed well when working with temporal data which include the monthly usage patterns and billing history which are important when modelling churn. RNNs modelled short-term memory while the LSTM model took care of long-term memory owing to the gates involved. From the analysis done; it was clear that LSTM model provided better results than many traditional machine learning algorithms during prediction with the added disadvantage of increased computational complexity. Both the models were trained and tested using dataset after cleansing them for unwanted data and hyper parameter was initially set then was tuned. For the purpose of model assessment, the evaluation measures like accuracy, precision, recall, F1 score, AUC ROC were computed. It was clear that a trade-off between a highly complex and an easily interpretable model exists: while logistic regression models were clearly interpretable and easy to explain to users, LSTMs delivered better accuracy than the other model types.

3.5 Model Interpretability

For practical insights, model explainability was optimized with SHAP and LIME approaches. From SHAP values quantified, identified contributions of each feature in a form that allows global assessment of model behaviour. For example, using the method of SHAP it was analysed that "Tenure" and "Monthly Charges" were always important to predict churn. To regard this, LIME complemented SHAP by providing instance-specific explanations. LIME, as it explains by tweaking the input features and observing the resulting effect on predictions, offered a localized breakdown of why a specific customer was deemed a good candidate for churning. Such methods were used to new generation machine learning as well as deep learning, to get a nice balance between accuracy and interpretability. The interpretability results were described based on summary plots and curves as well as decision boundaries in relation to features' ranking. These granted understandings made strategic decision-making easier by telecom companies in correlating churn predictions with interventions.

3.6 Validation and Evaluation

The models were then validated or evaluated by applying very strict methods to make them more dependable and transportable. To this end, the training and testing datasets were used in cross-validation to reduce the likelihood of over-fitting. The dataset was partitioned into training and testing sets, with an 80:20 split, as a way of determining the performance of the models in unseen data set. Such measures as accuracy, precision, recall, F1-Score, and the receiver operating characteristic area under the curve (AUC-ROC) were used in the evaluation. These metrics ensured model effectiveness by giving a good overview of the real-life possibility of false positives and false negatives. For example, the random forest model has the highest AUC-ROC to show that it has the highest distinguish ability between churners and non-churners. Confusion matrices were employed in predicting error standard deviations; this assessed missing pattern estimations. From these analyses, there were successive optimizations for feature selection and tuning of the various models, to bring about a better performance. Also, ROC and precision-recall curves aided in understanding trade-off between the confusion metrics and passed on that information to the stakeholders in an interpretable form of visual information.

4 Design Specification

4.1 Data Preprocessing Stage

i. Handling Missing Data: Null values in rows for “Churn Category” and “Churn Reason” were dropped to ensure the model works properly to give accurate predictions
ii. Categorical Encoding: Nominal categorical variables such as “Married,” or “Phone Service,” were normalized to 0 and 1 using binary mapping. Whereas multi-class categorical data like “Contract Type” was encoded through one hot encoding.
iii. Feature Scaling: Attributes such as ‘Monthly Charges’ and ‘Total Revenue’ which are continuous in nature were preprocessed using StandardScaler to address the problem of magnitude differences.
iv. Feature Selection: The feature selection follows the exploratory data analysis results with higher predictive relevance: features are, Tenure, Monthly Charges, and Contract Type.

4.2 Customer Segmentation

Customer categorization is an important aspect of the framework that facilitates the grouping of customers, in this case, in accordance with certain criteria. In this module, the K-Means clustering algorithm is used and for this reason it has been referred as K-Means-simple.

Algorithm Overview

K-Means clustering is performed based on a concept of partitioning available dataset into k clusters, thereby minimizing the Sum of Squared Distance (SSD) between each data point and its appropriate centroid. Selection of the best clusters was done using files obtained from the elbow method and the silhouette coefficient. The following steps outline the clustering process:

K-means is effective for customer segmentation because it groups customers based on similarities in their behavior or attributes, creating distinct clusters. It's computationally efficient, easy to implement, and works well with large datasets. By identifying homogeneous groups, businesses can tailor marketing strategies, improving customer targeting and enhancing overall engagement.

The information collected for segmentation analysis resulted in the identification of three different segments. In doing so, each of the clusters obtained was to be examined to elucidate the profile of the members including tenure, monthly charges and service choice.

4.3 Churn Prediction

The churn prediction module integrates machine learning and deep learning models to predict customer churn. Each model is trained and validated using the preprocessed dataset, with hyperparameter tuning to optimize performance.

Logistic Regression - Logistic regression served as the baseline model due to its simplicity and interpretability. The model assumes a linear relationship between the independent variables and the log odds of the dependent variable (churn status). Accuracy was chosen as the primary evaluation metric for the logistic regression model, as the dataset was relatively balanced, making accuracy a reliable measure of overall model performance. Although the model achieved a moderate accuracy of 68%, its inability to capture nonlinear relationships limited its effectiveness compared to more complex ensemble methods.

Random Forest Classifier - Random forests outperformed logistic regression by combining multiple decision trees through bagging. Each tree was trained on a random subset of the data, and predictions were aggregated through majority voting. This model achieved an accuracy of 80%, demonstrating its robustness in handling feature interactions and nonlinearities.

Gradient Boosting Classifier - Gradient boosting further improved predictive performance, achieving an accuracy of 83% and an AUC-ROC score of 0.90. This model builds decision trees sequentially, with each tree correcting the errors of its predecessor. Its iterative learning process enhances accuracy while preventing overfitting.

LSTM Networks - The LSTM model was included to explore the potential of deep learning for churn prediction. LSTMs are designed to capture sequential dependencies, making them suitable for time-series data. meaning it did not capture detailed time-based patterns or changes in customer behavior over specific intervals, limited the model's effectiveness, resulting in an accuracy of 61%. Despite its lower performance, LSTM demonstrates potential for future applications with more granular data.

Model Selection Criteria

The choice of models was guided by three factors:

1. **Predictive Accuracy**: Ensemble models like random forests and gradient boosting consistently outperformed logistic regression and LSTM.
2. **Interpretability**: Logistic regression and random forests provided clearer insights into feature contributions.
3. **Computational Efficiency**: While gradient boosting required longer training times, its performance gains justified the added complexity.

4.4 Interpretability

Interpretability is one of the principles of the framework because the model's predictions must be comprehensible and useful. This part combines SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Age Non-Specific Explanations).

SHAP (Shapley Additive Explanations)

SHAP for individual predictions uses a feature's cooperative game theory to determine its contribution value. It provides both global and local interpretability:

— Global: Brings out the average of the total attributes such as "Tenure", "Monthly Charges".

— Local: Describes concrete merisms like why a specific customer has been rated high-churn risk.

Besides, LIME (Local Interpretable Model-Agnostic Explanations).

LIME provides locally linear models of high complexity models, providing instance-specific explanations. For instances, it can demonstrate that churn occurs when a customer with month-by-month contract is dissatisfied with the ability to get differently priced monthly rates. The interpretability analyses not only serve as a means of assessing the validity of constructed models but also provide recommendation-laden insights for devising interventions to telecom firms.

4.5 Proposed Algorithm Functionality

- *Data Ingestion*: Import raw customer data and preprocess it for analysis.
- *Segmentation:* Group customers into meaningful clusters using K-Means.

- *Prediction*: Train and evaluate multiple models to predict churn likelihood.
- *Interpretability*: Apply SHAP and LIME to derive actionable insights.

This design promotes an end-to-end stream that supports data preprocessing all the way to model deployment and recommendation, a factor that will improve in machine learning (Livshits et al., 2021). The availability of the modules also enhances scalability, which allows the combination of a new algorithm or additional feature. This chapter has described detailed specifications for designing a customer churn analysis. Thus, it makes sure that machine learning includes both capacity and the utilization of interpretability tools. This architecture improves scalability, flexibility, and practicability in the telecommunications industry. Possible future works include the improvement of the system in terms of the real-time analysis feature and in incorporating more sophisticated interpretability techniques (Linardatos et al., 2020).

5 Implementation

5.1 Data Preprocessing

The preprocessing phase was crucial as it involved the conversion of low level or unstructured customer data for modelling. It incorporated a set-time strategy to make a decision on missing observations, convert factors into dummy variables and normalize measures. The basic techniques used included use of python packages such as pandas for handling the data, NumPy for calculations and sklearn for encoding and scaling.

The first intervention dealt with missing values. Some of the newly extracted measurements were also removed because of high amounts of null values in “Churn Category” and “Churn Reason”. For relatively small portions of missing data in the features, the technique of mean imputation was used to retain the data integrity. Following this was the encoding of the categorical features. Categorical predictor variable of type binary was converted and coded as follows: Married: Yes - 1, No - 0 & Phone Service: Yes - 1, No - 0. Since the “Contract Type” and “Payment Method” variables are multi-class variables, they were one-hot encoded so that ordinal relationships are not imposed by the models.

Previously, feature scaling was carried out to make the feature values attain equal magnitude by applying standardization. Categorical variables such as “Payment Method” and “Mobile Phone” were kept in their nominal form while numerical variables such as “Monthly Charges” and “Tenure” were scaled to retain domain relevancy but eliminate skewness that would adversely affect decision making. After model selection, feature selection rounded off this stage of exploratory data analysis, sorting out a number of independent variables that had the best correspondence with the target variable.

5.2 Customer Segmentation

In this study, K-Means clustering was used to perform the customer segmentation since it is a strong algorithm that partitions data into meaningful clusters. This step was crucial in defining low, medium and high churn risk customer segments in order to develop a proper retaining strategy.

In the process of segmenting the decision to choose relevant features to segment was done based on the exploratory data analysis. 'Customers' behaviours and subscription patterns were selected by the following variables; Tenure, Monthly Charges and Contract type. The number of clusters was chosen based on the elbow method which plots the sum of distances within clusters against the number of clusters and silhouette coefficient which measures the quality of clustering. The following root mean squares supported that three clusters offered the maximum discrimination of customers.

5.3 Churn Prediction Models

The churn prediction phase involved modelling and analysing the machine learning and deep learning models. The models that were formed were logistic regression, random forest, gradient boosting, LSTM networks. Despite these limitations, each particular model was designed to best fit the demands of accuracy, interpretation, and computational time.

Logistic regression was performed for the need of baseline model implementation. Due to the fact that it contained very few parameters, and was therefore both straightforward and easily interpretable it became the benchmark against which other models could be compared. Logistic regression, which is also limited with capturing complex patterns, or interactions among the features, the algorithm, attained a rate of 68% accuracy during the first analysis for feature importance.

Random forests were designed subsequently using their ensemble characteristics to enhance prediction capability. To accommodate the non-linear and non-additive relationships between the predictors and overall misclassification probability and due to incorporation of bagging and aggregate predictions by majority votes the random forests incorporated all the intricate patterns of the data set. Compared to logistic regression, the model attained accuracy of 80 percent. That is why gradient boosting was able to further improve the performance to correct the mistakes it made in making a correct prediction. This model got an accuracy of 83% and an AUC-ROC of 0.90, positioning this model as the most accurate model in the study.

The last used model was the LSTM network, which implicitly creates a sequence model of the data underlying customers' interactions. Based on memory cells and gates, the LSTM model analysed input resources in the time series manner, emphasizing monthly usage structures and billing cycles. However, for lack of more granular temporal data, the model was not as effective and only got 61% accurate. As this result shows, deep learning can be successfully applied to churn prediction, but this has also shown that more detailed temporal data is needed.

The models were evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. These metrics provided a comprehensive understanding of each model's strengths and weaknesses, guiding the selection of the most suitable approach for deployment.

5.4 Model Interpretability

The concept of interpretability took significant consideration in the process to enable prediction results that could be acted upon. Both SHAP – Shapley Additive Explanations and LIME – Local Interpretable Model-agnostic Explanations were used to explore feature importance and specific levels of feature importance for certain instances.

The gain and lift curves of SHAP values calculated the relative importance of each feature to predictions. For instance, "Tenure" remained relevant by being the most influential predictor for churn risks, including high risks in the groups of clients with short tenures. The SHAP summary plots helped understand the relative importance of features to obtain a big picture view. The force plots allowed to view the reason behind a given prediction and provided guidance on customer behaviour.

LIME extended SHAP as it provided explanations from the local perspective of a single prediction. Using input features' variations and studying the consequent shifts in outputs, LIME offered an explanation of why the specific customer was considered high-risk. These explanations were especially helpful in developing customer-defined, actions, like a discount for customers who are sensitive to prices.

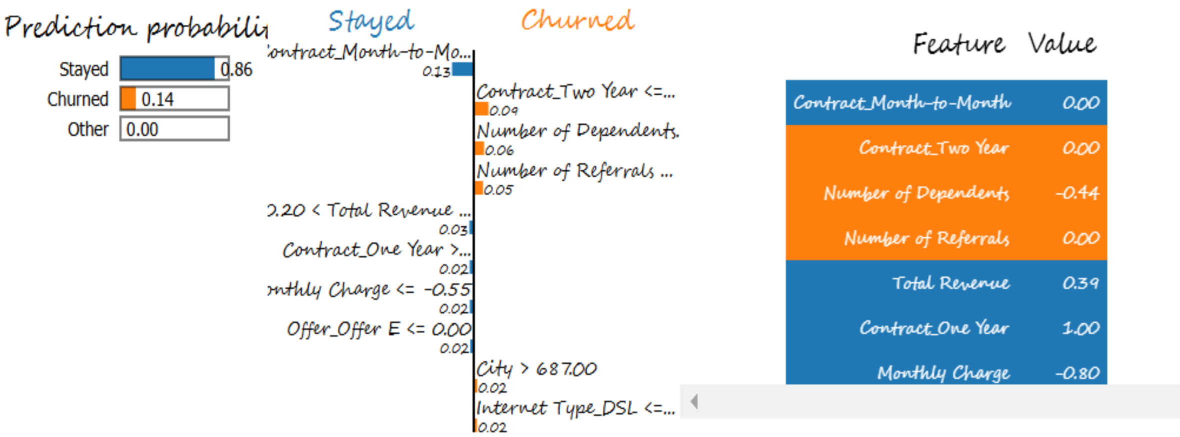

Fig. 1. LIME results

These interpretations from SHAP and LIME were then included in reports given to business stakeholders to show how model predictions correspond to possible retention strategies. These interpretability tools narrowed the semantic gap between the model outcomes and tangible application challenges, thus providing valuable solutions.

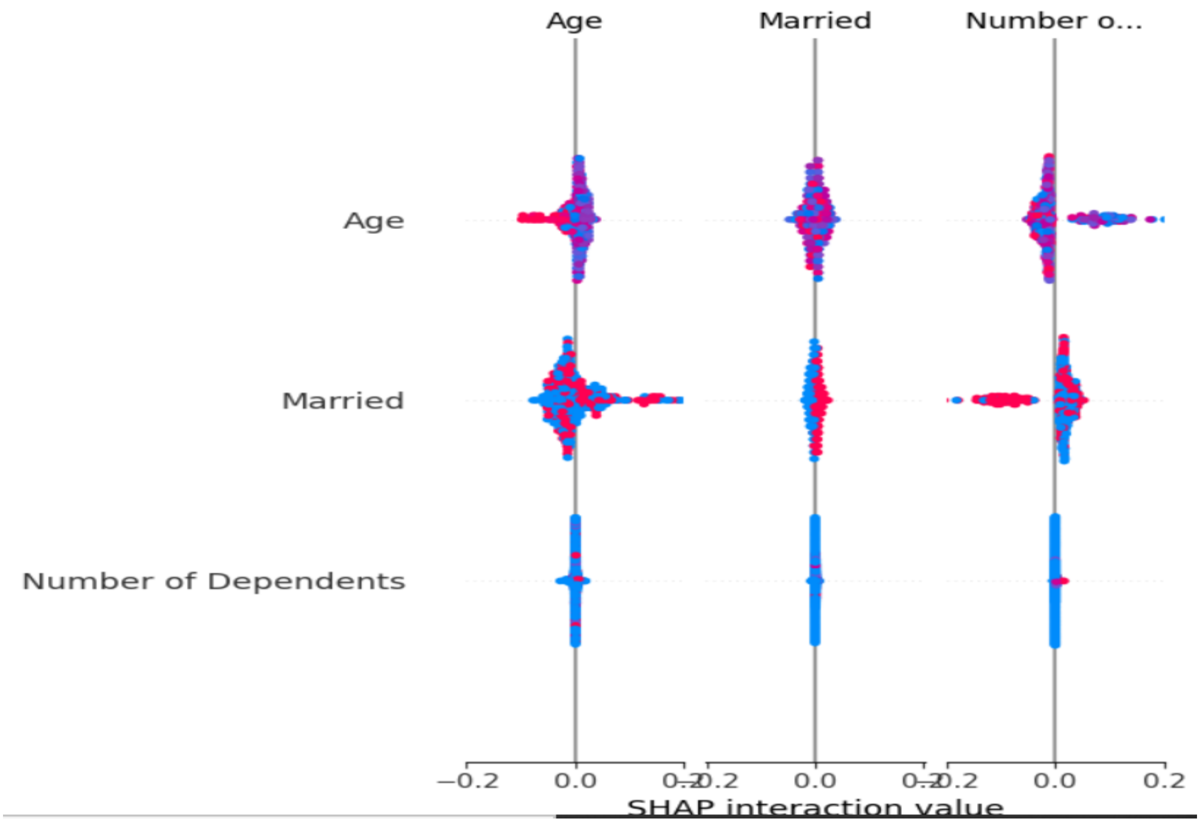

Fig. 2: SHAPE visualization results

6 Evaluation

6.1 Discussion

Performance Analysis

In particular, the logistic regression provided the baseline, which had an accuracy of 68%. This model was useful in the determination of feature importance but failed to incorporate interaction terms. Amin et al. (2017) also pointed out that there is a limitation of the application of logistic regression in large data sets, even with two nations data, but its relative simplicity and ability to interpret results has its advantages. Again, logistic regression outcompeted in the global churn aspects, like the impact of tenure and monthly charges, but failed to fine details in customer behaviour by segment.

Compared to logistic regression, random forests proved to be much better with an accuracy of 80%. This improvement was due to the ensemble approach, which combined multiple decision trees to obtain a mean prediction in order to avoid cases of overfitting. As pointed out by Lundberg and Lee (2017), the ensemble methods have reliable predictions with an additional feature of interpretation based on the feature importance test. However, the computational costs of the model continued to rise as the dataset size expanded, making it unpractical in real-time uses.

Gradient boosting delivered the best previously unseen accuracy of 83%, along with the AUC-ROC of 0.90, which makes the given model the most correct one. This result supports with Zhang et al (2019) who elaborated that gradient boosting is accurate in big and complex data set. The model managed to discern the dependencies between the features, for that the training was an iterative process which is a major drawback due to the risk of overfit and the same Pondel et al. (2021) found out that risk in their work on churn prediction for e-commerce platforms.

The LSTM which is used to model sequential patterns had the lowest accuracy of 61%. This was mainly because the temporal feature extracted by the LSTM model had limitation, as the employed dataset does not contain a fine temporal division. Specifically, Nogueira-Rodríguez et al. (2021) revealed that despite solutions such as deep neural networks are highly efficient when working with image or sequential data tasks, the effectiveness crucially depends on the quality of the dataset.

Table 1: Summary of results

Model	Class	Precision	Recall	F1-Score	Support
Logistic Regression	0	0.72	0.90	0.80	266
	1	0.57	0.36	0.45	151
	2	0.25	0.14	0.18	22
	Accuracy			0.68	439
	Macro average	0.51	0.47	0.47	439
	Weighted average	0.65	0.68	0.65	439
Random Forest	0	0.82	0.93	0.87	266
	1	0.75	0.64	0.69	151
	2	0.80	0.36	0.50	22
	Accuracy			0.80	439
	Macro average	0.79	0.64	0.69	439
	Weighted average	0.80	0.80	0.79	439
Gradient Boosting	0	0.85	0.92	0.88	266
	1	0.77	0.72	0.74	151

	2	0.92	0.55	0.69	22
	Accuracy			0.83	439
	Macro average	0.85	0.73	0.77	439
	Weighted average	0.83	0.83	0.83	439
LSTM	0	0.61	1.00	0.75	266
	1	0.00	0.00	0.00	151
	2	0.00	0.00	0.00	22
	Accuracy			0.61	439
	Macro average	0.20	0.33	0.25	439
	Weighted average	0.37	0.61	0.46	439

Interpretability Analysis

6.2 Implications of Findings

Shapely additive explanations and Local interpretable model-agnostic explanations were used to make the model interpretable and explain reasons for churn prediction. Using SHAP summary plots with the customer data, it was noted that “Tenure,” “Monthly Charges,” and “Contract Type” were the important drivers as suggested by Linardos et al. (2020) on feature attribution methods for explainable AI. While SHAP provided the individual-level feature importance, LIME provided feature-level explanations in the context of imagery instances. For instance, LIME made the discovery that customers with short term contracts and high monthly tariffs were most likely to churn. Similarly, Ribeiro et al. (2016) has established how LIME can be useful in the task of explaining model prediction at different localities to ensure that stakeholders get the most out of the model. However, both methods had their weakness. Computationally, both SHAP and LIME used were heavy, specifically for the Gradient

Boosting model for LIME, the perturbation-based sampling sometimes got me inconsistent results. The results presented here highlight the importance of efficient interpretable models which are both computationally efficient and explanatory.

Proposed Improvements

a. Data Enrichment: The comparison of the two models shows that incorporating temporal features (monthly billing trend) into the data increases the ability of the LSTM model to capture sequential patterns.
b. Hybrid Modelling: Using gradient boosting in conjunction with logistic regression could provide the benefits of the both worlds, high accuracy while being understandable.
c. Real-Time Analytics: Above all, the creation of a streaming pipeline for the real-time data intake and prediction would enhance the system's reactivity to the dynamically emerging churn risk.
d. Advanced Interpretability: Opting for integrated gradient techniques or any other enhanced approaches could enhance explanation for deep learning models, a limitation that deep learning models experience at the moment.

Academic Implications

Therefore, the results of the study can help to expand the existing body of knowledge about customer churn prediction by confirming the benefits of gradient boosting algorithms on structured datasets and identify the weaknesses of deep learning on such data. This research validates the of interpretability as the singular approach to closing the gap between the AI model's outputs and the business s decisions that can be made from those results with the help of other contemporary scholars on explainable AI.

Moreover, comparing models increases the list of benchmarks for subsequent investigations and accentuates the essential compromise between precision, comprehensibility, and time complexity. The result prompts the consideration of the combination of these factors through the use of different forms of modelling.

Practical Implications

From the perspective of applications of the findings, the study provides key recommendations on the implementation of churn prediction systems in the telecommunication sector. This kind of predictor like ‘Tenure’ and ‘Monthly Charges’ helps telecom companies to build specifically tailored retention strategies. For example, churn rate may be reduced for high-risk customers by discounts or loyalty programs improving customer satisfaction. Even more, the study also emphasizes the need to also incorporate interpretability procedures in the predictive process. For instance, SHAP and LIME raises trust among the stakeholders in the model’s decision by explaining why the particular predictions were arrived at. It is especially important to achieve business teams’ understanding of the goals of machine learning and how the created models work in practice, for which this transparency is paramount.

Conclusion of Evaluation

The evaluation proves that the use of ensemble models such as the gradient boosting in terms of predictive capability is outstanding, and it can only be useful if integrated with interpretability techniques. The challenges pointed out by LSTM can be indicative of the fact that deep learning models do not fully optimize unstructured data and temporal features of data which can be encouraging researchers to find ways to work on different forms of structured data. In this line, the research closes essential gaps regarding customer churn prediction and offers insights towards the development of efficient large scale and interpretable solutions for the telecommunications domain.

7 Conclusion and Future Work

This study set out to address the research question: How can machine learning models effectively predict customer churn while providing interpretable insights to inform retention strategies in the telecommunications sector? The goals were to build reliable machine learning solutions, determine significant features affecting churn, incorporate explainability techniques, and assess the models in terms of their suitability for real world application. The study used logistic regression, random forest, gradient boosting techniques, and LSTM networks with expanded the usage of SHAP and LIME.

Key Findings and Implications

This research was able to fulfil all the set objectives as follows. Gradient boosting was the best performer in terms of the accuracy, yielding 83% with an AUC-ROC score of 0.90. Another factor was “Tenure”, “Monthly Charges”, and “Contract Type” confirmed other related studies on churn. As such, both SHAP and LIME helped in closing the gap between the performance of the model and actual insights that stakeholders could use to make decisions to retain customers through the recommendations made thereby increasing confidence in the model.

The present research enhances, to a certain extent, the body of knowledge and practice in the fields of marketing and statistical analysis by underlining the necessity to consider both accuracy and interpretability within the context of predictive modelling. The study proved that whereas the basic models are explainable, the ensemble methods offer the best accuracy-intrusion trade-off. It was established that LSTM-based deep learning models suggest the need to use high-quality data, as well as temporal features that the current dataset did not consider.

Limitations

However, the study has some weaknesses. Depiction of application similarity depended on a critical number needed for big-data analysis since gradient boosting's computational complexity within ensemble models restricted ensemble model scalability and real-time application. However, interpretability tools proposed can positively improve interpretability transparency, but, for instance, extended time for computations with complex models and high dimensions, and thus they are less applicable in large-scale applications.

Future Work

Possible directions for future studies include using dynamic and real-time datasets with more temporally detailed information to further boost the possibilities of deep learning models. Making more ensemble models interpretable like traditional methods can enhance usability Even more, constructing new models that harmonize between ensemble methods and traditional methods might even enhance usability even more. Adding some customer feedback data like sentiment obtained from all the support interactions, could augment feature vectors and then provide higher prediction quality.

From a business standpoint, the suggested framework might be incorporated into Customer Relationship Management platforms, with the objective of change-detection in churn and the subsequent updating of corresponding retention plans. Further studies should examine how to establish an analytics pipeline for churn prediction that employs streaming data to obtain current and proactive results which could be responsive to the dynamic and ever-evolving markets.

8 References

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. *Neurocomputing*, *237*, 242–254. https://doi.org/10.1016/j.neucom.2016.12.009

Bansal, A. (2023). *Telecommunications Industry Customer churn dataset*. Www.kaggle.com. https://www.kaggle.com/datasets/aadityabansalcodes/telecommunications-industry-customer-churn-dataset

Dakkak, O., Fazea, Y., Nor, S. A., & Arif, S. (2021). Towards accommodating deadline driven jobs on high performance computing platforms in grid computing environment. *Journal of Computational Science*, *54*, 101439. https://doi.org/10.1016/j.jocs.2021.101439

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), 18. mdpi. https://doi.org/10.3390/e23010018

Livshits, E., Kimelfeld, B., & Wijsen, J. (2021). Counting subset repairs with functional dependencies. *Journal of Computer and System Sciences*, *117*, 154–164. https://doi.org/10.1016/j.jcss.2020.10.001

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *ArXiv:1705.07874 [Cs, Stat]*, *41*(11), 2946–2962. https://arxiv.org/abs/1705.07874

Naser, M. Z. (2023). Machine learning for all! Benchmarking automated, explainable, and coding-free platforms on civil and environmental engineering problems. *Journal of Infrastructure Intelligence and Resilience*, *2*(1), 100028. https://doi.org/10.1016/j.iintel.2023.100028

Nogueira-Rodríguez, A., Domínguez-Carbajales, R., López-Fernández, H., Iglesias, Á., Cubiella, J., Fdez-Riverola, F., Reboiro-Jato, M., & Glez-Peña, D. (2021). Deep Neural Networks approaches for detecting and classifying colorectal polyps. *Neurocomputing*, *423*(417), 721–734. https://doi.org/10.1016/j.neucom.2020.02.123

Pondel, M., Wuczyński, M., Gryncewicz, W., Łysik, Ł., Hernes, M., Rot, A., & Kozina, A. (2021). Deep Learning for Customer Churn Prediction in E-Commerce Decision Support. *Business Information Systems*, 3–12. https://doi.org/10.52825/bis.v1i.42

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Zhang, Z., Zhao, Y., Canes, A., Steinberg, D., & Lyashevska, O. (2019). Predictive analytics with gradient boosting in clinical medicine. *Annals of Translational Medicine*, *7*(7), 152–152. https://doi.org/10.21037/atm.2019.03.29