An Explainable Machine Learning Framework for Telecom Customer Churn Prediction Using RFM Clustering and LightGBM USING R

An Explainable Machine Learning Framework for Telecom Customer Churn Prediction Using RFM Clustering and LightGBM USING R

Submitted To:

Mr. Ahmed Imran Kabir

Assistant Professor

School of Business & Economics (SoBE)

United International University

Submitted By:

Name: Shahoriar Parvez

ID: 111 201 180

Major: Business Analytics

Registration Trimester: Fall 2025

School Of Business and Economics

United International University

Date of Submission: 11th March, 2026

Letter of Transmittal

March 11, 2026

Ahmed Imran Kabir

Assistant Professor,

School of Business and Administration (SoBE)

United International University (UIU)

United City, Madani Avenue, Badda, Dhaka-1212

Subject: **Submission of Research Project on "An Explainable Machine Learning Framework for Telecom Customer Churn Prediction Using RFM Clustering and LightGBM USING R."**

Dear Sir,

It would be my honor to present my detailed research paper, the Customer Churn Prediction through RFM-based Segmentation and Explainable LightGBM Model in R. The literature is a detailed examination of a customer churn prediction model based on artificial intelligence and machine learning approaches to telecommunications in order to enhance its predictive capabilities, which can be achieved through the demographic background of a customer..

It reports an analysis of our method, including the Customer segment RFM model, the LightGBM algorithm and the SHAP evaluation method with segment + LightGBM with 80.68% accuracy

The report is provided with experiment results, their analysis and presentation using a web-based interface. It is also a detailed description of the system architecture, the procedure of implementation, and the performance appraisal. The work has various contributions of significance to the prediction of churn among customers and strategic modelling. The given approach is already tested on a large amount of data and has shown better results as compared to the current methods.

I would like to thank you that you have read this finest work. I would very much appreciate your feedback or suggestions on the improvement.

Sincerely,

Shahoriar Parvez

ID: 111 201 180

Major: Business Analytics

School of Business and Economics

United International University (UIU)

Certification of Similarity Index

An Explainable Machine Learning Framework for Telecom Customer Churn Prediction Using RFM Clustering and LightGBM USING R

ORIGINALITY REPORT

9%	4%	6%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Tianpei Xu, Ying Ma, Changyu Ao, Min Qu, XiangHong Meng. "A Novel Telecom Customer Churn Analysis System Based on RFM Model and Feature Importance Ranking", Interdisciplinary Journal of Information, Knowledge, and Management, 2023 Publication	4%
2	dspace.uiu.ac.bd Internet Source	2%
3	dspace.uiu.ac.bd:8080 Internet Source	1%
4	scholar.archive.org Internet Source	<1%
5	Submitted to National University of Ireland, Galway Student Paper	<1%
6	Kazim, Hind Tawfiq. "Predicting Customer Churn in E-Commerce.", Rochester Institute of Technology Publication	<1%
7	Submitted to United International University Student Paper	<1%
8	Submitted to Monash University Student Paper	<1%

9	Submitted to University of Strathclyde Student Paper	<1%
10	research-explorer.ista.ac.at Internet Source	<1%
11	www.nature.com Internet Source	<1%
12	dos Santos Costa Oliveira, Ana Raquel. "Exploring and Preventing Churn in Gyms", Universidade do Porto (Portugal), 2025 Publication	<1%
13	ir.amu.ac.in Internet Source	<1%
14	Fuchang Han, Shenghui Liao, Chao Xiong, Haitao Wei, Renzhong Wu, Yingqi Zhang. "Explainable Prediction of Whether The Acetabular Cup Is Placed in The "Safe Zone" from X-ray Images", 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021 Publication	<1%
15	Submitted to Universitat Oberta de Catalunya Student Paper	<1%
16	Yinglin Xia, Jun Sun. "Machine Learning for Microbiome Statistics", CRC Press, 2026 Publication	<1%
17	d-nb.info Internet Source	<1%
18	rdrr.io Internet Source	<1%
19	www.coursehero.com Internet Source	<1%

20	"Intelligent and Fuzzy Systems", Springer Science and Business Media LLC, 2025 Publication	<1%
21	www.frontiersin.org Internet Source	<1%
22	Haque, Ekramul. "Enhancing Network Security: Dynamical Intrusion Detection Systems Leveraging Zero Trust Architecture.", Tennessee State University Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

Declaration of the Student

I would like to state that the present research project entitled Customer Churn Prediction Using RFM-Based Segmentation and an Explainable LightGBM Model in R is a project of my personal work, which will be conducted under the supervision of Ahmed Imran Kabir. The report has been prepared as part of the condition of completion of my Bachelor of Business Administration (BBA) degree course at the School of Business and Economics (SoBE), the United International University.

I also testify that I have not entered this work to any other institution in any direction to receive any degree or academic qualification. I conducted all experiments, analyses, and implementations went through in this report under the appropriate supervision.

I consider the advice and help that I have had throughout the semesters of this research. The programs, information and findings used in this paper are real and have been formulated as per the acceptable research procedures.

Md Shahoriar Parvez
ID: 111 201 180
Major: Business Analytics
School of Business and Economics
United International University (UIU)

Acknowledgement

To commence with, I appreciate the fact that the Almighty has granted me the supremacy, intelligence and patience to complete this research work.

On my part, I wish to offer my warm gratefulness to my supervisor on my research project because of his support, great suggestions, great comments and encouragement in conducting the research.

It was possible due to my research work and my academic peers who provided feedback at various phases of this work, discussed different aspects of this work, and helped to evaluate the system in the course of user studies. They have been able to enlighten us on how we can streamline our process and work on the system. In particular, I would like to get a chance to express my sincere gratitude to my friend, S.M Shahariar Maruf, who motivated me, suggested ideas contributing to the creation of this research, and assisted me in all its stages.

We owe the open-source community the existence and maintenance of the Research Hub, Creator of RStudio, and other computer vision libraries and other powerhouses that are used in parts of this research. My prototype research is accelerated through the presence of pretrained models and datasets.

Lastly, I would like to appreciate my family with support and being patient. It is through their realization that this would be achieved after thousands of hours of research and development.

Thank you all.

Abstract

Customer churn prediction is a very sensitive activity in the telecommunication sector since it is cheaper to retain customers than to acquire them. This paper will suggest a hybrid analytic and machine learning platform that combines Recency-Frequency-Monetary (RFM)-based customer segmentation, K-means clustering, and Light Gradient Boosting Machine (LightGBM) model with SHAP-based explainable AI (XAI) to improve prediction accuracy and model interpretability.

LightGBM algorithm performs both for the original data and the clustering data set. The experiment result shows that the cluster data set improves the prediction accuracy by 0.8068 and F1 by 0.8760, and the original data accuracy is 0.7765 and F1 is 0.8395. And the AUC value cluster data and original data are 0.8484 and 8582, which means the strong predictive capability of the prediction.

SHAP analysis will be employed in order to show the most significant features influencing customer churn. According to the findings, the most important predictors of different categories of customers are the type of contract, tenure, monthly charges, and total charges. Specifically, the month-to-month contracts and increased monthly fees are closely linked with a greater degree of churn risk. The importance of the features also differ within clusters, which highlight the customer behavior distinctions

The findings provide that the customer segmentation RFM model with ML algorithm and explainable AI techniques can provide deep insight into customer behaviour and an important churn prediction model strategy

Keyword:

Customer churn prediction, Customer Segmentation, K-means, WSS, Cluster, LightGBM, SHAP, R Programming, Explanable AI (XAL)

Table Of Contents

Table of Contents

List of Figures

List of Tables

List of Equations

1. INTRODUCTION

1.1 BACKGROUND OF THE STUDY

Another interesting area CRM research among telecom companies to enhance customer retention is customer churn as it is a critical issue in the telecom market, which is highly competitive. Most of the researchers are engaged in creating a telecom customer churn analysis system to identify the cause of customer churn that can be incorporated to enhance the accuracy of prediction.

Telecommunication firms face risky issues which are tied to the churn of customers (new consumers are expensive to maintain in terms of research and analysis, etc) compared to retaining. The termination of service by the consumers is referred to as customer churn. The factor that enables the enabling companies to retain interventions is the churn prediction, especially in competitive sectors like the belated rocketry.

Each customer is unique, in this case, we are referring to the age, location, psychology but the most significant parameter is the purchasing behaviour. Accordingly, each category of customers needs a different product at a different cost. Thus, two groups of customers need different marketing strategy. The first thing that we need to create differently is to stratify like-minded customers into a single segment (Shirole, Salokhe, and Jadhav, 2021). In the context of the telecommunication sector, which is experiencing an escalation of competition resulting in a severe customer churn issue in the companies involved, to ensure the improvement of business further, there is a need to offer not only high-quality technical services but also high-quality personalised services, to avoid churn(Sudharsan & Ganesh, 2022). The cost of obtaining a new customer is estimated to be 5-6 times higher than the cost of obtaining an existing customer(Santharam & Krishnan, 2018).

It is noted that numerous small online shops and new comers to the online retail industry are keen to undertake data mining and consumer-centric marketing within their business, but technically could not know how to do it(Chen, Sain, and Guo, 2012).

Since customers are the greatest asset of any telecommunications firm, there is need to know and treat them well. Customer relationship management (CRM) is also referred to as the framework that regulates the relations between the company and its existing or prospective clients (Santharam and Krishnan, 2018).

1.2 STATEMENT OF THE PROBLEM

Customer churn is an issue that telecommunication companies have been experiencing, and this has led to loss of revenue and high operational costs. Although the range of studies that apply ML algorithms to the prediction of churn is high, several challenges are still observed.

Firstly, the majority of churn prediction models do not consider customer heterogeneity and predictive performance only. The behavioral traits and usage habits of the customers tend to be different, and this can be the one that will dictate whether they will churn or not. The primary step deals with the identification of the important attributes that would differentiate the various market segments, hence market targeting and positioning of products. Companies in competitive markets need to know their customers in order to develop suitable market matchings. However, it is not easily explainable by a large number of customers(Rungruang, Riyapan, Intarasit, Chuarkham, & Muangprathub, 2024).

Second, traditional machine learning models tend to be black box and it is hard to describe the variables that affect the churn predictions. The problem of lack of interpretability may arise whereby an organisation may be challenged in translating model results into business strategies. Segmentation, targets and positioning (STP) marketing is one of the basic marketing principles which are the creation of high level customer value and helps in the growing of products and services(Rungruang et al., 2024). The presence of a booming e-commerce and the World Wide Web (WWW), a powerful tool in the customer relationship management (CRM), referred to as the RFM analysis model, has been utilised to ensure that large businesses earn more money. The CRM system, along with data mining technologies, can automatically predict the future behaviors of customers in order to boost the customer retention rate(Wan, Chen, Qi, Gan, and Tang, 2022).

Third, the number of studies integrating customer segmentation with the use of modern machine learning models and explaining AI practices into one piece of analysis is minimal. This, in turn, implies that there is a need to perform research, which would bring together these approaches to make the predictive ability and interpretability stronger.

The most basic issue in this industry is that the information about the subscribers of telecom companies is often grossly incomplete, and the thickness of the information of a given feature is low. This is an issue that is common among telecommunications companies across the world, rendering the majority of the conventional prediction methods inaccurate. In addition, the existing research tends to ignore the paramount significance of helping companies with the diagnosis of the causes of churn. To address these gaps, the present paper proposes a new and suitable system of customer churn analysis. By combining the data extension method with the feature importance ranking, the system will be geared towards giving the telecom companies a better competitive advantage and more knowledge.

This paper, therefore, addresses these concerns by developing a framework that can integrate the RFM-based segmentation, LightGBM modelling, and SHAP explainability to study customer churn behaviour.

1.5 DEFINITION OF KEY TERMS

The primary goal of the study, however, is to use explainable AI approaches and behavioural segmentation to develop an explainable machine learning model of customer churn prediction.

The customer segmentation methods are not standard and methods employed are wide ranging with the businesses employing the most appropriate method based on the unique characteristics of the respective industry. K-means and RFM models are applicable in the formation of new feature in progressive churn prediction processes. As the model is introduced, it becomes easier to interpret the data and thus learn more with an objective of making more precise predictions (Barus, Nathasya, and Pangaribuan, 2023). Clear, secure and reliable usage of machine learning models is dependent on model explanations. Contrary to popular belief, SHapley Additive exPlanations (SHAP) framework has been considered as the gold standard of local explanations because of its solid theoretical foundation and because it is universal(Mosca, Szigeti, Tragianni, Gallagher, and Groh, 2022).

The following are the main conclusions of the paper:

1. The new and innovative churn analysis design is suggested, the combing of the importance of features rankings, customer churn prediction, and customer segmentation into one system.
2. The K-means clustering algorithm is applied along with the RFM model in developing a data extension method by segmenting customers. This would not only segment the customers, but it would also add new informative features to enhance the dataset.
3. An interpretable predictive modelling approach is implemented, utilising the LightGBM algorithm to achieve high-accuracy churn prediction. The SHAP method is then applied to quantify and rank the importance of individual features, providing transparency and actionable insights into the key drivers of customer churn.

1.5 SCOPE AND LIMITATIONS OF THE STUDY

This study's primary goal is to create an interpretable machine and a better prediction with full accuracy. Nonetheless, the research has a number of weaknesses. First, the sample may not reflect the consumer behaviour in other businesses as it was indicative of a certain telecommunications scenario. Second, alternative methods can give more data but the study uses one machine learning model. Lastly, the study does not take into consideration the aspect of customers' behaviour over time and instead utilises cross-sectional data. Bangladeshi data is not available for analysis.

1.6 ORGANISATION OF REMAINING CHAPTER

This paper will be divided into the following form. In Section 2, the literature review of the key notions, including the clustering techniques, the RFM models, and the analysis of the feature importance are presented. Section 3 gives the information about the dataset that was used in this study and the methodology of this study that is being proposed. In section 4, the discussion and presentation of the experimental results are given. Finally, the paper concludes with the conclusion in Section 5 in which findings and conclusions are summarised.

2. LITERATURE REVIEW

2.1 RELEVANT THEORY

Technological progress is a major determinant to becoming the market leader and being a market player. Meanwhile, the competition and the rules of the game in the telecom industry have already been transformed due to technological advancement. However, the contemporary customers are now more sensitive to differentiated and value-added services, and this has increased switching costs, but consumer loyalty has also increased(Zhang, Moro, and Ramos, 2022).

Subscriber churn has been a significant problem that has been affecting telecom companies in recent years. Some literature (Ma, 2023) states that predictive accuracy is compensated in the form of providing useful customer segmentation and management information. This is a limitation to the utility of most of the research because it is not widely applicable to real-world telecom conditions. This has resulted in an abrupt need for methods that would be capable of hitting a compromise in the accuracy and depth of prediction as well as the cause of churn analysis. This is bridged in this paper not just by improving the performance of the prediction, but also by a systematic analysis of the drivers of churn would be done with an aim of improving the strategic decision-making process.

One of the potential directions that can be taken to enhance the predictive accuracy is clustering techniques. Other researchers have also thought of integrating clustering methods in the prediction pipeline to elicit the underlying customer group processes in order to improve better performances of the model.

2.2 LITERATURE SURVEY

2.2.1 CLUSTERING

Clustering, the concept, has been widely utilised in data mining and machine learning to identify unintuitive patterns in the customer data and enhance the predictive modelling performance. Clustering- This is an unsupervised learning algorithm, which is highly used in data mining to reveal the presence of similarities in giant data sets by grouping similar observations into homogeneous cluster. The main objective of the clustering is to maximize similarity of a cluster and different clusters, which will result in the ability to identify meaning patterns in extremely complex data landscapes. Throughout the recent years clustering methods have become an essential part of customer analytics and, primarily, within the industries that generate a huge volume of behavioural data. K-means algorithm, which is partition based clustering algorithm, is deemed as one of the most widely used due to their efficiency in processing and the capability of scaling-up on large data sets as they possess high capacity (Jain, 2010). Simple K-means algorithm clusters the data into a set number of clusters, by optimising the within-cluster variance which is usually than the squared Euclidean distance between the observation and the centroid of the cluster

The most significant one is how to find the number of the clusters. My best cluster is the one that I am determining with the help of the elbow method though a different method can be suggested

The clustering analysis greatly relies on the number of cluster as it is a very important process. The facilities of the number of clusters are directly related to the quality of the results of the. segmentation. Usually, various measures have been suggested towards. Identify the maximum number of clusters such as the elbow test, silhouette coefficient and information-based measures(Rousseeuw, 1987). In addition to the partition-based schemes, the density-based clustering schemes have also been given a lot of attention since it is able to detect clusters of mixed shapes and correct noise. DBSCAN is a density-based algorithm and is among the most popular, and is used to cluster the points based on the density distribution of the points rather than specifying the number of clusters (Ester et al., 1996). The other density-based algorithm is called

OPTICS and is the extension of the DBSCAN algorithm, which gives a ranking of the data points, which demonstrates the underlying structure of the clustering, and does not require a threshold density (Ankerst, Breunig, Kriegel, and Sander, 1999).(Tsiptsis and Chorianopoulos, 2011)

The clustering methods have been popular in the study of customer behaviour and market segmentation. Another area of application of clustering has been in telecommunication and service industries to identify clusters of customers who have common usage patterns, purchasing behaviour and service preference. This form of segmentation helps organisations to develop certain marketing strategies and manage better the relationship with customers. Previous studies have defined the customer segmentation based on clustering as potentially very effective in enhancing predictive model performance particularly in regard to churn prediction activities, in order to discover latent behavioural pattern in the customer data (Tsiptsis and Chorianopoulos, 2011). Using the results of clustering as inputs to predictive model pipelines, researchers can design additional behavioural characteristics, which can be incorporated into the dataset and allow machine learning models to characterise more complex customer dynamics. Consequently, feature construction by clustering has become an effective feature technology to increase the predictive accuracy of a customer churn prediction system.

2.2.2 RFM MODEL

Recency Frequency Money (RFM) models remains of the best-known behavioural customer analytics as well as relationship management models. The model quantifies the engagement and value of customers through three behavioural indicators namely; recency that is the time interval the customer last interacted or purchased a product, frequency which is the amount of interactions or transactions a customer had with the business over a certain stage and budgetary value which is total outlay of the customer. These are highly concise but powerful variables of customer buying behaviour and are frequently used to describe valuable customers and predict customer future behaviour. RFM models is a simple models that are easy for interpret and thus is still used in modern data-driven marketing and predictive analytics (Black et al., 2023) (Gupta and Kim, 2020).

The latest studies have integrated the RFM analysis and ML and data mining process to improve customers segmentation and predictive modelling. Indicatively, the RFM model has been employed by scholars alongside cluster algorithms such as of the K-means and hierarchical clustering identify distinctive customer groups in relation to their purchasing patterns. These types of segmentation can enable businesses to categorize the customers into segments such as loyal customers, potential churners, and low-value customers and then apply certain marketing strategies and customer retention programs (Putra et al., 2021). Segmentation of customers according to their behavioural attributes enables the companies to manage their marketing resources effectively to improve their performance in their customer relationship management.

RFM variables have also been rampant in predictive modelling frameworks of churn prediction as input variables along with segmentation. Recent research indicates that added behavioural predictors using RFM can significantly enhance predictive power on ML algorithms because they can used to identify patterns of customer engagement that are difficult to see on raw transactional inputs. Using an example, it has been demonstrated that machine learning models trained on RFM features can more effectively distinguish between loyal and potential churners, hence they can better classify them when it comes to churn prediction (Huang et al., 2022).

Besides, the availability of large volumes of data on customers has prompted the researchers to integrate the traditional RFM model with the new machine learning models and ensemble algorithms. Another similar engineered factor in these hybrid techniques of supplementing the original data are the RFM variables, so that the predictive models are able to learn more about behavioral associations in the data. It has been found that this integration has enhanced effectiveness of the churn predictive systems, customer value analysis in telecommunications, e-commerce and financial services industries (Zhang et al., 2023). The RFM model remains, therefore, relevant to modern customer-centric business models as a useful analytical instrument to understand customer behaviour and improve predictive analytics.

2.2.3 FEATURE IMPORTANCE RANKING

The ranking of feature importance is the other approach that is not novel with the goal to render the predictive models more intuitive and observe the most prominent sources of customer behaviour and churn (Wojtas and Chen, 2020). The machine learning models that are used to predict the feature importance include other models, such as Simple boosting algorithms (LightGBM) as well as XGBoost algorithm (M. Chen et al., 2019); there are also regularised linear models (LASSO, Ridge regression, etc.), as well as the Simple boosting algorithms in the category. Such techniques have been found to have the capabilities of large predictive as well as high propensity to solve complex data; the gradient boosting algorithms are one of them. The lightGBM in particular has gained quite some popularity due to the fact that it boasts of reasonable levels of computational efficiency, scalability in addition to the fact that it can handle large scale information with relative accuracy. However, ensemble models are very complex model, which cannot be easily comprehended. To address this limitation, Lundberg and Lee (2017) created one such trendy explainable artificial intelligence (XAI) algorithm termed as SHapley Additive exPlanations (SHAP) algorithm. SHAP is based on the cooperative game theory and the worth provided by an input feature could be identified by means of computing the Shapley values that is the marginal worth of a feature to the model wisdom with all other combination of feature sets fixed. Local and global ones can be explained on the basis of the quantification of the behaviour of the model, of the effect of single variables using SHAP (Guo et al., 2021). Therefore, by integrating SHAP and machine learning algorithms, such as LightGBM, researchers will be in a better position to be grateful to the impacts of the different inputs attributes to the predicted outcomes and, hence, will be handy in informing organisations on the most relevant customer churn drivers and the decision making process.

To conclude, clustering algorithms have been utilized for feature expansion in order to improve prediction accuracy within the customer churn domain The RFM model for it learns the best and making telecom companies trying to get more such quality customers. Xgboost Feature ImportanceRanking Since organizations need to justify the cases of customers churning, the method gives a good overview of how feature

importance is computed from a customer point-of-view. Enabling the entire telecom customer churn measurement and prediction process

3. MATERIALS AND METHODS

3.1 METHODS

Based on three main components, the proposed customer churn analysis system outcomes e.g. customer segmentation, churn prediction and countering the churn factor. The first phase involves the clustering of data. RFM (Recency, Frequency, Monetary) model on top of the from K- means clustering algorithm, and then picking up the perfect clusters by the assistance of the elbow method. Then, it uses features to predict churn. The final. For explainers of the LightGBM approach, we have SHAP (Shapley Additive exPlanations) , and LightGBM explained in order to understand the

the model output and with a custom-built structure of feature importance ordering.

The analysis of this methodology with the data is done in an open-source manner.

telecom customer dataset having 7,043 cases and 21 features. The R programming

Written in a language that can be utilized by the ML algorithm.

3.2 DATASET

The system that I am doing research on is the customer analysis system. I use an open-source customer churn data set that is available in the Kaggle webpage. Reports on one of the telecom companies that serve 7032 customers in California as of the 3 rd quarter of 2019 without phone and internet services. This data might not have been literally reflective of the whole telecommunications industry in the world, but will reflect the vast majority of the common issues and prevailing trends visible in the industry-such as churn of customers and consumer behaviour. It shows the churned, the ones that remained and the ones who just subscribed to the services of the company, among other 21 details of the customers. Majority of the important demographic data can also be located in the data. Churn is the target variable that is divided into two groups such as yes and no with the customer having or having not churned. Table 1 provides a list of the features, their descriptions, and their data types.

Table 1. Dataset Description

Feature	Description	Data type
customerID	Customer ID	Nominal
Gender	Whether the customer is a male or a female	Nominal
SeniorCitizen	Whether the customer is a senior citizen or not	Nominal
Partner	Whether the customer has a partner or not	Nominal
Dependents	Whether the customer has dependents or not	Nominal
Tenure	Number of months the customer has stayed with the company	Numeric
PhoneService	Whether the customer has a phone service or not	Nominal
MultipleLines	Whether the customer has multiple lines or not	Nominal
InternetService	Customer's Internet service provider	Nominal
OnlineSecurity	Whether the customer has online security or not	Nominal
OnlineBackup	Whether the customer has an online backup or not	Nominal
TechSupport	Whether the customer has tech support or not	Nominal
StreamingTV	Whether the customer has streaming TV or not	Nominal
StreamingMovies	Whether the customer has streaming movies or not	Nominal
Contract	The contract term of the customer	Nominal
PaperlessBilling	Whether the customer has paperless billing or not	Nominal
PaymentMethod	The customer's payment method electronic check,mailed check, bank transfer	Numeric
MonthlyCharges	The amount charged to the customer monthly	Numeric
TotalCharges	The total amount charged to the customer	Numeric

3.3 PROPOSED CUSTOMER CHURN ANALYSIS SYSTEM

As Figure 1 demonstrates, the suggested CCAS will comprise customer segmentation, feature construction, customer churn prediction, and detecting the customer churn factor. The system's goal is to identify client churn and its contributing elements. Customers are segmented using K-means algorithm and RFM model, and new customers features are created based for the segmentation outcomes. Churn factors are identified using the LightGBM and SHAP algorithms, whereas churned consumers are predicted using the LightGBM algorithm.

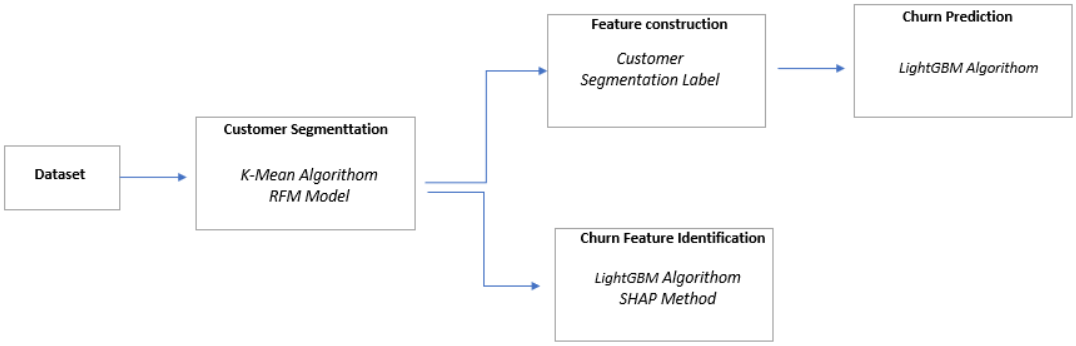

Figure 1. Churn analysis system structure

3.3.1 CUSTOMER SEGMENTATION

Customer segmentation refers to the process of classifying customers basing on common behaviours.. This approach helps businesses anticipate what customers might buy, identify potential new audiences, and create more targeted marketing campaigns. The organisations are aware of the unique needs of the different categories of customers and can thus commit their resources in empowerment. Here, it is achieved through using the RFM model and the K-means algorithm. K-means is one of the most widespread centroid-based clustering algorithms applicable to the unsupervised machine learning(Liu, Feng, Ai, Chen, and Lin, 2026) in grouping the data into K categories. Calculate the distance between each data point and all centroids. Assign each point to the closest centroid (usually using Euclidean distance).

This forms the initial clusters. The characteristics of the K-means algorithm cluster are centres and attributes of the cluster. The analysis would involve both the elbow method and the heuristic method that can estimate the best number of clusters to use in the analysis as it would know when introducing an extra cluster would not add any significant value to the model.

The RFM model remains is the most widely usable techniques for behavioural segmentation in customer analytics and relationship management. It evaluates customer engagement and value based on three key behavioral indicators: recency, which refers to how frequently a customer made a purchase or attraction with the brand; frequency, or how often they transact within a given period; and monetary value, which reflects their total spending. This paper adopts the RFM scoring convention established in prior literature, using upward (↑) and downward (↓) symbols to indicate relative performance. Specifically, if a cluster's Recency, Frequency, or Monetary value exceeds the overall aggregate average, it is marked with an upward arrow (↑); otherwise, it receives a downward arrow (↓). The customer segmentation process is illustrated in Figure 2(Lee et al., 1998).

Step 1: This study presents the value profile of telecom customers in terms of Recency-Frequency-Monetary. RFM model are a model of buyer segmentation that is widely used in profiling puchasing behaviour using three main measures, namely recency, frequency and monetary value.

Customers, is most instances, pay their communication providers monthly in the case of the telecom industry. Therefore, the Recency (R) variable is connected with the habit of the customer in terms of payment within a month. The recency value will be identical in this analysis (value = 1) since the telecom billing process is usually scheduled on a definite monthly cycle and all the customers are charged with a definite monthly charge.

Frequency (F) variable is a measure of how many times a client consumes the services of the company in terms of time. The frequency indicator used in this data is the element Tenure that is the measure of the years a customer is a customer of the telecom provider. The extended time period is a sign of the more loyal customers and the usage of services more often.

Monetary (M) variable entails the sum of money that a customer injects in a company. The attribute in the dataset used in the study to represent the financial cost is the attribute MonthlyCharges since it represents the amount of finance a customer spends on telecommunication services in a month.

Step 2: When this has been constructed, the variables of RFM are then followed by the next step of establishing the right number of customer segments. Since K-means clustering is an unsupervised learning algorithm, a number of clusters or (K) must be typed in prior to the actual process of clustering the data taking place.

K-means will attempt to group the data into k distinct clusters in such way that the customers within a specific cluster are very similar and the customers within one cluster are not similar to their counterparts in another cluster. This is achieved through the minimization of The Within-Cluster Sum of Squares (WSS) which is a measure of squared dissimilarity among the individual data and the centroid of that particular cluster.

Best value of K is obtained by taking the Elbow Method. Under this strategy the values of WSS on different numbers of clusters are checked and it is observed at what point the rate of decrease in WSS is greatly diminishing. This is the stage of the series that is the most appropriate number of clusters of the dataset that is the one that is referred to as the elbow point.

Step 3: Once the optimal size has been reached in the initial steps, it is declared that the K-means clustering algorithm will be used to group customers based on the variables of RFM. The algorithm involves the use of all the customers to cluster the customers but this process is an iterative one where a center point of the cluster is modified until all the customers are pleased with the cluster.

All customers of the data set will be merged into K different clusters with the selected variables of the RFM and the number of clusters. They share similar behavioural and financial characteristics, customers who belong to the same cluster, that enable the telecom customer segmentation in a commendable manner

.Step 4: After having sorted the clusters into clusters, each cluster is examined using the aid of RFM scores to interpret the characteristics of the customer groups. The study, according to the approach proposed by (Ha and Park 1998) makes use of

upward (↑) and downward (↓) indicators to show that the cluster of RFM value is either above or below the mean of a dataset. Where a cluster is found to have a greater R, F or M value than the mean, then it is denoted using the arrow (↑).

The down arrow (↓) is applied in case the value is lower than the total average. It is the analysis of these RFM score patterns that the clusters can be comprehended and

elucidated in accordance with their customer value and behavioural patterns. Such segmentation will facilitate the identification of several groups of telecom customers such as high-value customers, loyal customers, or low-engagement customers that may be applied to carry out more specific customer management and churn-mitigating strategies

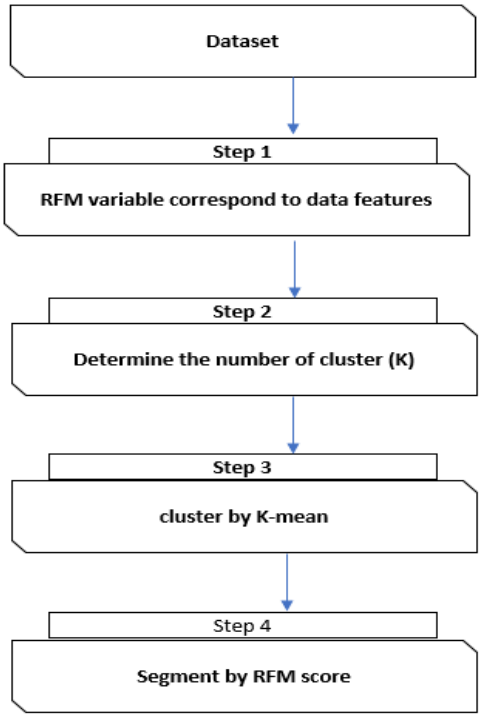

Figure 2: Customer Segmentation Process

3.3.2 FEATURE CONSTRUCTION

Process of constructing features assists in improving performance of predictive models that transform the raw information to significant variables which represent better the customer behaviour. The current work has conducted the feature building to elicit the variables in terms of RFM over the given original telecom data, and will assist in examination of customer segmentation and customer churn forecast. K-means clustering algorithm was subsequently provided with the acquired RFM variables as features taken as input in order to establish the various customer groups which were in essence incorporated into machine learning models to forecast telecom customer churn.

3.3.3 CHURN PREDICTION

After construction of RFM features, K-means clustering algorithm was applied to define discrete customer groups in respect to behavioural and financial profiles. The outcomes were then combined to the initial data as an additional feature. Finally, a LightGBM model was shot on the enriched data set including the RFM variables and the cluster data, to forecast the telecom customer churn.

Customers of the telecommunication services are likely to shift service providers often because of numerous factors therefore causing the high customer churn rates in the industry. The telecom companies have therefore a great stake in customer identification of probable leavers as they can employ certain strategies to ensure that they hold onto their customers and reduce churn. The new machine learning techniques are to be used in order to identify the potential churners successfully. The current study has been performed based on Light Gradient Boosting Machine (LightGBM) algorithm as it is efficient, scaled up and powerful in predicting issues in customer churn prediction classification. LightGBM model is more precise and is more effective to make predictions(Sun, Liu, and Sima, 2020). The prediction using the model also makes it possible to predict the model performance, which can be observed by observing some of the metrics, e.g. accuracy, precision, recall and f1-score (Hossin and Sulaiman, 2015).

3.3.4 CHURN FACTOR IDENTIFICATION

The telecom companies need to determine what influences the churn of its customers in an attempt to formulate effective customer churning strategies. Predicting churn with the use of machine learning models has become popular because fact of the ability of such models to replicate complex correlation amongst customer characteristics. LightGBM, which is a form of gradient boosting has been demonstrated to possess a high predictive performance when used on classification. Explainable artificial intelligence (XAI) techniques have been introduced in order to fulfill the explanation of model outputs(Sinaga, R., and Widianto, S. 2024). The SHAP (SHapley Additive exPlanations) is among the most common interpretability techniques that quantifies the contribution of each feature to the model prediction. The findings indicated that SHAP framework drastically affects the churn issue and marketing department can implement the right strategy based on customer personalisation.

Therefore, in the research, LightGBM is used to predict the customer churn, and SHAP is employed to explain and understand the churn predictors which comprise of the transaction behavior, customer engagement and service usage patterns(Arai, K., et al. 2023).Other papers have also used SHAP to explain and understand the churn predictors which include the transaction behavior, customer engagement and service usage patterns. Use of SHAP visualization process, such as SHAP summary plots and dependence plots. This can be calculated using R programming, though it can be calculated manually

The SHAP value for a feature represents the average marginal contribution of that feature to the prediction across all possible feature combinations. The SHAP value is calculated as follows:

$$\Phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\left(|F| - |S| - 1\right)!}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S\left(x_S\right) \right]$$

Φi represents the SHAP value of the i-th feature.

F denotes the set of all features used in the model.

S represents a subset of features selected from F that does not include feature i.

|S| is the number of features in subset S.

fS(xS) is the prediction of the model when only the features in subset SSS are considered.

fS∪{i}(xS∪{i})represents the prediction of the model when the feature i is added to subset S.

xS denotes the values of the input features contained in subset S.

4. RESULTS

4.1 CUSTOMER SEGMENTATION

I use R Programming for doing this analysis. The elbow method is applied to determine the optimal cluster. WSS shows the cluster distance number. Algorithm Range is 1 to 10. The K value is 4 because the 1 to 4 difference is high. And there is little change from 5 to 10, so the K = 4 is the best cluster for that.

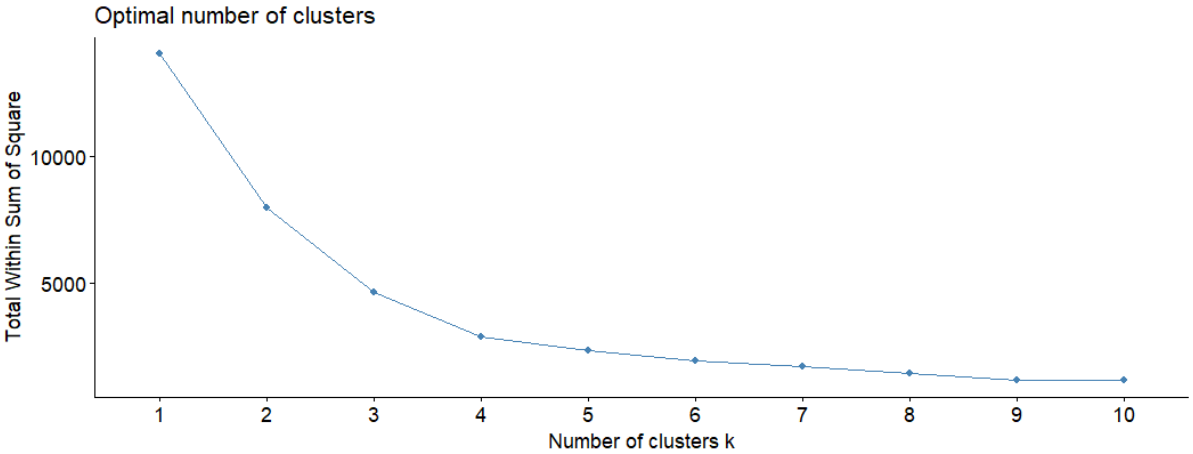

Figure 3. WSS value curve

This data set is separated into the 4 clusters based on the RFM model, where the RFM contain the MonthlyCharges, tenure, and MonthlyPayment. Figure 4 shows the scatteredplot for the 4 clusters, which is devide in 4 groups based on algorithom.

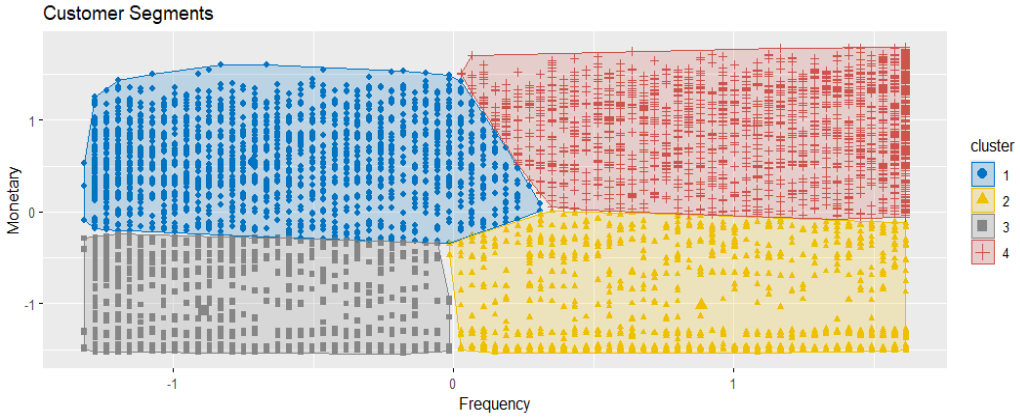

Figure 4. Scattered plot for 4 clusters

The average of the RFM variable is indicated in Table 2. The percentage of 4 clusters, name of the cluster and the RFM score. The F value, which is higher than the mean value represents customer value, which describes the commitment of the company strongly and often remains longer as members. Conversely, customers of lesser F values are not so loyal, and their relations with the organisation are short-lived. Regarding the monetary value (M), individuals with higher-than-average spending will be related to a much higher profitability, and the customers with low spending will be willing to consume discounted products.

Table 2. Customer segmentation results

Data	R(mean)	F(mean)	M(mean)	RFM score	Name of Cluster	Amount (%)
character	numeric	numeric	numeric	character	character	numeric
1	1	14.8	81.2	R-F↓|M↑	HCSTC	31.3
2	1	54.1	34.0	R-F↑|M↓	LCLTC	16.4
3	1	10.5	32.5	R-F↓|M↓	LCSTC	24.6
4	1	58.7	93.0	R-F↑|M↑	HCLTC	27.7
Dataset	1	32.4	64.8			100.0

Here we can see all the customer clusters in our data set, cluster 1, F lower and M higher than average value which is cluster name is High Consuming Short-Term Customers (HCSTC). Cluster 2, F is a larger value and M is less than the average value and so, the name of the cluster becomes Less Consuming Long-Term Customers (LCLTC). Cluster 3, F and M are smaller in comparison with the average value meaning that the name of the cluster denotes Less Consuming Short-Term Customers (LCSTC). Value of F in the Cluster 4 is higher than since it is higher than the average and only M, which means that the name of the cluster is Less Consuming Long-Term Customers (LCLTC). It is here that we learn that the proportion of the value HCSTC is the 31.3 percent that all customer is high consumers, yet are short term, is a cluster of those who like high-quality products and services. The company should evaluate their favor-ites in order to create and sell high quality products. The LCLTC of all customer 16.4% that Low consuming, but not leave Long-term. The firm should suggest appropriate quality products and services to encourage good

consumption and a high expenditure in the long-term. LCSTC percentage of all customer is 24.6% reading Low consuming and they are not out of the short term. The company needs to go and evaluate the nature of this cluster to achieve more involvement and reduce the churn. HCLTC 27.7% of all customers High consuming, they are kept in the Long term. To make the customer loyalty degree larger, the company should maintain a contact with them and offer various promotions.

4.2 FEATURE CONSTRUCTION

After find the customers into four fertile clusters based on their value to the business, we noticed something important: customers within the same group share similar value characteristics, while those in different groups don't. This means we can create a new customer feature that clearly shows these value differences.

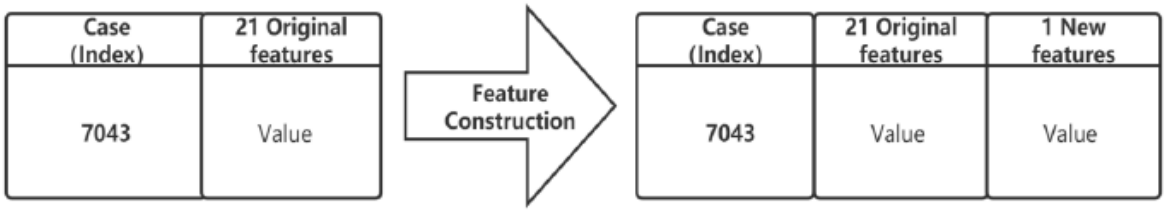

Figure 5. The process of dataset feature construction

Figure 5 illustrates how we transformed the original data by adding this new feature

4.3 CHURN PREDICTION

I operated the LightGBM with the data of the initial information and cluster information and tried to determine the 4 type of matrices to evaluate the performance and impact of the model performance as shown on Table 3. The model performance, as we may

see, is significantly improved and the accuracy is achieved by nearly 3.03%.

Table 3. Comparison of model performance after feature

Dataset	Model	Accuracy	Recall	Precision	F1-score
Orginal dataset	LightGBM	0.7765	0.7825	0.89	0.86
New dataset	LightGBM	0.8068	0.9291	0.9	0.88

Table 4 The validity of the proposed customer churn analysis system prior to the proposed method in this paper is contrasted to the other developed machine learning techniques on the identical data set, 4. As can be seen, the highest accuracy percentage is concentrated in our customer segmentation based churn prediction model.

Work	Model	Accuracy%
(Halibas et al., 2019)	Naive Bayes	73.00%
	Generalized linear model	75.70%
	Logistic regression	75.70%
	Deep learning	74.30%
	Decision tree	70.70%
	Random forest	75.20%
	Gradient boosted tree	79.10%
Nandhini, S., & Chitra, P. (2025)	PCA-Supported LightGBM	80.10%
My method	Customer segmentation +LightGBM	80.68%

Table 4. Comparison with others' work

4.4 CHURN FACTOR IDENTIFICATION

Table 5 display the top 5 feature importance of predicting customer churn based on the SHAP method. The rankings are presented on the original data set and the four RFM-based customer clusters. These findings prove that not all types of customer segments are subject to the same churn drivers, and jumpstart implying that one global churn model can fail to capture the key behavioural dissimilarities among customers.

Table 5. Top important features for the original dataset and 4 clusters

Ranking	Original_Dataset	HCSTC	LCSTC	LCSTC.1	HCSTC.1
1	Contract.Month-to-month	tenure	Contract.Month-to-month	Contract.Month-to-month	tenure
2	tenure	MonthlyCharges	TotalCharges	TotalCharges	MonthlyCharges
3	TechSupport.No	Contract.Two year	tenure	tenure	Contract.Two year
4	OnlineSecurity.No	TotalCharges	MonthlyCharges	MonthlyCharges	TotalCharges
5	MonthlyCharges	Contract.Month-to-month	TechSupport.No	TechSupport.No	Contract.Month-to-month

Contract. Month-to-month is the most important feature in the original dataset to predict churn, and tenure, TechSupport.No, OnlineSecurity.No, and MonthlyCharges are the second, third, and fourth most influential features, respectively. This finding is an indication that subscribers to month-to-month contracts are more likely to churn than subscribers to long-duration contracts. Moreover, the churn tendencies also tend to be higher in the case of customers who have shorter tenure, no technical support, and no online security services.

The ranking of the importance changes significantly when the analysis is conducted on cluster-specific datasets.

HCSTC Cluster

The top churn predictor in the HCSTC cluster is tenure, then MonthlyCharges, and Contract.Two year, TotalCharges and Contract.Month-to-month. This means that the duration of the stay of these customers with the company is the overwhelming factor to churn behavior among the customers in this segment.

LCSTC Cluster

In the case of the LCSTC cluster, the most significant feature is Contract. Month-to-month and then TotalCharges, tenure, MonthlyCharges, and TechSupport.No. This indicates that one of the key contributors to churn in this category of customer is the flexibility of a contract and billing factors.

LCLTC Cluster

The ranking of importance in the LCLTC cluster goes the same way as the LCSTC cluster, with the following features taking the top three positions: Contract. Month-to-month, TotalCharges, and tenure. It means that the type of contract and total expenditure is essential in the process of churn in this segment.

HCLTC Cluster

In the case of the HCLTC cluster, tenure, MonthlyCharges, and Contract.Two years, TotalCharges, and Contract. Month-to-month was the most influential factor. These findings indicate that the main churn determinants of this group are customer lifetime and service cost factors.

On the whole, the findings indicate that the churn drivers of customers differ considerably between clusters. Although the type of contract is the prevailing variable in the entire dataset, other variables like tenure, total charges and monthly charges have a stronger effect among specific groups of customers. Itforms the significance of segment-based churn analysis, whereby companies are able to develop more specific retention measures in accordance with the specifics of every customer group.

5. DISCUSSION

The current research paper suggested a combined customer churn prediction model that involves the use of RFM-based customer segmentation and LightGBM machine learning along with the SHAP (Shapley Additive Explanations) to enhance the accuracy of predictions and interpretability of the model. To begin with, the customers were divided based on the RFM model whereby frequency (tenure) and monetary value (monthly charges) were employed to represent customer engagement and spending behavior. With K-means clustering algorithm, four segments of customers were formed as the dataset has been separated into, which enabled the model to address heterogeneity in customer behavior patterns. RFM is one of the commonly employed customer segmentation methods in marketing analytic that allows distinguishing between customer groups sharing certain behavioral traits and facilitates more specific decision-making (Wedel and Kamakura, 2012; Kumar and Reinartz, 2016).

After the segmentation, LightGBM based churn prediction model was trained. The model realized the accuracy of 0.8068 which depicts high predictive power on churn and non-churn customers. LightGBM is a popular gradient boosting algorithm and is known to be efficient and perform well in classification tasks, especially when large and complicated data are used (Ke et al., 2017). The findings show that the prediction model with the inclusion of behavioral segmentation in the predictive modeling pipeline is capable of capturing the customer patterns better and giving more significant insights than when using the original data only.

In a further attempt to improve the interpretability of the model, SHAP analysis was used to identify variables that had the most significant impact on churn prediction. SHAP is an explainable artificial intelligence (XAI) algorithm designed on cooperative game theory, and that allocates values of contributions to each feature in a predictive model (Lundberg and Lee, 2017). The findings of Table 5 indicate that even churn predictors are heterogeneous in different clusters of customers, as the most significant churn predictors vary according to different groups of customers. The features that were found to be the most significant predictors of churn in the original dataset included month-to-month contract, tenure, absence of technical support, absence of

online security, and increase in monthly charges. This result is in line with other researchers who report that short-term contracts and an increased cost of service significantly contribute to the risk of customer churn (Verbeke et al., 2012; Amin et al., 2019).

The cluster-level SHAP analysis also indicated that there were churn drivers unique to customer segments. As an illustration, in LCSTC cluster, the total charges, internet services type, and tenure features were more significant in churn prediction, thus indicating that the customer in this cluster is more sensitive to the usage of the services and total cost. Conversely, HCLTC cluster revealed that tenure and monthly charges were the two most effective variables, which shows that long-term high-value customers can be churned in case of service costs increasing. In the same manner, the HCSTC cluster had been mainly affected by the type of contract and the total charges, and the LCLTC cluster indicated that tenure and monthly charges were still the determinants of churns. These results confirm the argument that the churn behavior is highly differentiated by customer segment and must be studied at a segmented level instead of applying one global model (Ngai et al., 2009; Ullah et al., 2019).

On the whole, the findings indicate that the combination of RFM segmentation, LightGBM predictive modelling, and SHAP explainability can be used to introduce a significant customer churn analysis framework. The suggested method has not only high predictive validity but also offers clear results on the driving forces of churn in each customer segment. This knowledge can be used to assist telecommunications businesses in developing specific retention plans, better customer relationship management and more efficient marketing resource allocation. In future research, this framework can be expanded by adding more variables of behaviour, more advanced clustering methods, and models of ensemble learning to further augment the churn prediction performance and business decision support.

5.1 SCOPE OF FUTURE STUDY

Future studies may extend this work by exploring deep learning models, incorporating real-time customer behavioural data, and applying alternative clustering techniques for customer segmentation. Additionally, other explainable AI methods could be used

to further enhance model interpretability. Testing the proposed framework on larger datasets and across different industries would also help evaluate its generalizability and practical applicability.

5.2 CONCLUSION

The current study has designed a customer churn prediction model by combining customer segmentation based on RFM and LightGBM machine learning algorithm with SHAP explainability analysis. The K-means algorithm was used to divide customers into four groups, which allowed to represent the differences between customer behaviour guidelines in terms of frequency (tenure) and monetary value (monthly charges). LightGBM model had an accuracy of 0.8068, which is good predictive capability in determining the possible churn customers.

The SHAP analysis also indicated the major churn factors differ in different groups of customers. Variables like month to month contract, tenure, technical support, online security and monthly charges were some of the most powerful predictors of churn in the entire dataset. These are in line with other related research which has shown that short-term contracts and an increase in the cost of services have a strong relationship with the churn probability (Verbeke et al., 2012; Lundberg and Lee, 2017).

On the whole, the findings prove that the integration of customer segmentation and machine learning with methods of explainable AI can enhance the interpretability and efficacy of the churn prediction models. Such a strategy can assist telecommunications firms to gain a better insight into customer behavior and develop specific retention plans to apply to various customer groups.

6. REFERENCES

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record, 28*(2), 49–60.

Barus, O. P., Nathasya, C., & Pangaribuan, J. J. (2023). The Implementation of RFM Analysis to Customer Profiling Using K-Means Clustering. *Mathematical Modelling of Engineering Problems, 10*(1).

Black, R. J., Cross, M., Haile, L. M., Culbreth, G. T., Steinmetz, J. D., Hagins, H., . . . Ong, K. L. (2023). Global, regional, and national burden of rheumatoid arthritis, 1990–2020, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. *The Lancet Rheumatology, 5*(10), e594–e610.

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management, 19*(3), 197–208.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters, 31*(8), 651–666.

Lee, C.-G., Hahn, H., Seo, Y.-M., Min, S. L., Ha, R., Hong, S., . . . Kim, C. S. (1998). Analysis of cache-related preemption delay in fixed-priority preemptive scheduling. *IEEE transactions on computers, 47*(6), 700–713.

Liu, S., Feng, Q., Ai, W., Chen, H., & Lin, B. (2026). Research on K-Means algorithm based on adaptive association rules and its application in commodity segmentation. *Egyptian Informatics Journal, 33*, 100911.

Ma, Y. (2023). ANovel TELECOM CUSTOMER CHURN ANALYSIS SYSTEM BASED ON RFM MODEL AND FEATURE IMPORTANCE RANKING. *Interdisciplinary Journal of Information, 18*, 719–737.

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022). *SHAP-based explanation methods: a review for NLP interpretability.* Paper presented at the Proceedings of the 29th international conference on computational linguistics.

Ong, K. L., Stafford, L. K., McLaughlin, S. A., Boyko, E. J., Vollset, S. E., Smith, A. E., . . . Hagins, H. (2023). Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet, 402*(10397), 203–234.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics, 20*, 53–65.

Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2024). RFM model customer segmentation based on hierarchical approach using FCA. *Expert Systems with Applications, 237*, 121449.

Santharam, A., & Krishnan, S. B. (2018). Survey on customer churn prediction techniques. *International Research Journal of Engineering and Technology, 5*(11), 3.

Shirole, R., Salokhe, L., & Jadhav, S. (2021). Customer segmentation using rfm model and k-means clustering. *Int. J. Sci. Res. Sci. Technol, 8*(3), 591–597.

Sudharsan, R., & Ganesh, E. (2022). A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy. *Connection Science, 34*(1), 1855–1876.

Sun, X., Liu, M., & Sima, Z. (2020). A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance research letters, 32*, 101084.

Tsiptsis, K. K., & Chorianopoulos, A. (2011). *Data mining techniques in CRM: inside customer segmentation*: John Wiley & Sons.

Wan, S., Chen, J., Qi, Z., Gan, W., & Tang, L. (2022). *Fast RFM model for customer segmentation.* Paper presented at the Companion Proceedings of the Web Conference 2022.

Wan, W., Kubendran, R., Schaefer, C., Eryilmaz, S. B., Zhang, W., Wu, D., . . . Gao, B. (2022). A compute-in-memory chip based on resistive random-access memory. *Nature, 608*(7923), 504–512.

Zhang, T., Moro, S., & Ramos, R. F. (2022). A data-driven approach to improve customer churn prediction based on telecom customer segmentation. *Future Internet, 14*(3), 94.

Gupta, S., & Kim, H. (2020). Customer analytics and RFM-based segmentation in modern marketing. Journal of Business Research.

Putra, D., Santoso, H., & Wibowo, A. (2021). Customer segmentation using RFM model and K-means clustering for marketing strategy optimization. Procedia Computer Science.

Huang, C., Li, Y., & Chen, T. (2022). Customer churn prediction using machine learning and behavioral features. Expert Systems with Applications.

Zhang, Y., Wang, J., & Liu, X. (2023). Hybrid machine learning models for customer churn prediction based on behavioral feature engineering. IEEE Access.

Fei, Z., Wang, B., & Li, Y. (2022). Customer churn prediction using random forest and machine learning techniques. Expert Systems with Applications, 190, 116196. https://doi.org/10.1016/j.eswa.2021.116196

Fonti, V., & Belitser, E. (2017). Feature selection using LASSO. VU Amsterdam Research Paper in Business Analytics, 30, 1–25.

Guo, C., Berkhahn, F., & Wang, X. (2021). Interpreting machine learning models using SHAP values: Applications in predictive analytics. IEEE Access, 9, 132531–132542. https://doi.org/10.1109/ACCESS.2021.3115006

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS) (pp. 4765–4774).

Machado, L., Karray, F., & Sousa, J. (2019). LightGBM: A highly efficient gradient boosting decision tree algorithm for predictive analytics. IEEE International Conference on Data Science and Advanced Analytics.

Wojtas, M., & Chen, Y. (2020). Feature importance ranking methods for machine learning models in customer churn prediction. Journal of Business Analytics, 3(2), 89–104. https://doi.org/10.1080/2573234X.2020.1752892

Arai, K., et al. (2023). Churn customer estimation method based on LightGBM for improving sales. International Journal of Advanced Computer Science and Applications.

Sinaga, R., & Widianto, S. (2024). Understanding telecommunication customer churn: Insights from LightGBM predictive modelling and SHAP feature interpretation. ASEAN Marketing Journal.

7. APPENDIX-A

The detailed R code is provided in the R script

Library and Data

```
library(readr)

library(dplyr)

library(ggplot2)

library(factoextra)

library(xgboost)

library(SHAPforxgboost)

library(caret)

library(cluster)

library(pROC)

library(lightgbm)

library(knitr)

data <- read_csv("WA_Fn-UseC_-Telco-Customer-Churn1.csv")
```

RFM Calculation

```
# Compute RFM (adapted for telecom billing)
rfm_complete <-  data_clean %>%
  mutate(
    # Recency: Assume all recent (monthly billing); set to 1 or (max(tenure) - tenure + 1) for variety
    Recency = 1,
    # Frequency: Tenure as usage frequency proxy (months active)
    Frequency = tenure,
    # Monetary: Total lifetime charges
    Monetary = MonthlyCharges
  )
```

WSS and k-means

```
rfm_scaled <- scale(rfm_complete[, c("Frequency", "Monetary")]) # Find optimal cluster


set.seed(201180)

kmeans_result <- kmeans(rfm_scaled, centers = 4, nstart = 25)#nstart how many time analysis cluster

summary(kmeans_result)

rfm_complete$Cluster <- as.factor(kmeans_result$cluster)
```

LightGBM model

```
lgb_model <- lgb.train(
  params = params,
  data = dtrain,
  nrounds = 500,
  valids = list(validation = dvalid),
  early_stopping_rounds = 20,
  verbose = 1
)
```

SHAP analysis

```
dummy_cluster <- dummyVars(Churn_num ~ ., data = cluster_data)


  x_cluster <- predict(dummy_cluster, newdata = cluster_data)
  x_cluster <- as.matrix(x_cluster)


  y_cluster <- cluster_data$Churn_num


  dcluster <- lgb.Dataset(data = x_cluster, label = y_cluster)


  model_cluster <- lgb.train(
    params = params,
    data = dcluster,
    nrounds = 200,
```

```
    verbose = -1
  )

  cluster_results[[i]] <- get_top_shap(model_cluster, x_cluster)

}

top_cluster1 <- cluster_results[[1]]
top_cluster2 <- cluster_results[[2]]
top_cluster3 <- cluster_results[[3]]
top_cluster4 <- cluster_results[[4]]

shap_table <- data.frame(
  Ranking = 1:5,
  Original_Dataset = top_original,
  HCSTC = top_cluster1,
  LCSTC = top_cluster2,
  LCHTC = top_cluster3,
  HCHTC = top_cluster4
)
```