Vavuniya Journal of Business Management
Vol 6 Iss. 2, December, 2023

Faculty of Business Studies
University of Vavuniya
Sri Lanka

Research article

Enhancing Customer Retention with Data-Driven Strategies: Machine Learning Insights from Sri Lanka's Telecom Industry

Himali L. P.

Department of Economics and Statistics, Faculty of Social Sciences and Languages

Sabaragamuwa University of Sri Lanka

himali@ssl.sab.ac.lk

Abeysekara J. D.

Department of Economics and Statistics, Faculty of Social Sciences and Languages

Sabaragamuwa University of Sri Lanka.

Abstract

Customer retention is a crucial factor in ensuring a firm's long-term sustainability, particularly in highly competitive industries like telecommunications. A stable customer base mitigates risks associated with demand fluctuations and enhances business resilience. Research consistently shows that retaining existing customers is more cost-effective than acquiring new ones, a principle widely recognized across industries. This study explores customer retention prediction among university students in Sri Lanka's telecommunications sector using machine learning regression algorithms. Various predictive models were evaluated to determine the most effective approach for forecasting retention, with the Decision Tree Regressor emerging as the top-performing model. Key factors influencing customer loyalty were also analysed, revealing trust as the most significant determinant, followed by perceived pricing and service provider choice. These insights offer telecom providers a data-driven foundation for improving retention strategies, particularly for university students. By leveraging predictive analytics, businesses can anticipate customer behaviour more effectively and implement targeted interventions to foster loyalty, ensuring sustainable growth in a competitive market.

Keywords: Customer Retention, Predictive Analytics, Perceived Price, Service Provider, Customer Loyalty.
JEL codes: C51, C52, C63
Introduction

Customer retention, distinct from churn, focuses on fostering long-term relationships with existing customers, thereby reducing costs and enhancing profitability. This distinction is particularly relevant in the telecommunications industry, where evolving technology and intense competition demand innovative approaches. Telecom providers increasingly leverage AI-powered customer support and loyalty apps to personalize the customer experience and ensure satisfaction.

The COVID-19 pandemic has underscored the critical role of telecommunications, with increased reliance on digital communication for remote work, education, and social interactions. This shift has heightened customer expectations and led to more frequent switching of service providers. University students, a rapidly growing and influential demographic within this context, represent a unique customer base. Their digital dependency presents opportunities for telecom providers to design tailored retention strategies, such as data-centric plans and education-specific offers.

Globally, the telecommunications market was valued at USD 1657.7 billion in 2020, reflecting its economic significance. In Sri Lanka, the industry has transitioned from a state-controlled utility to a dynamic private sector, contributing significantly to the country's economy. However, despite its importance, research on customer retention particularly within the telecommunications sector remains limited compared to studies on churn. This is especially true for university students, a segment whose retention determinants and behaviours are not well understood.

Existing research on customer retention in telecommunications often focuses on traditional approaches, such as customer loyalty metrics or secondary data analysis, while neglecting advanced predictive modelling techniques. Furthermore, these studies rarely evaluate both the determinants of retention and the accuracy of predictive models. This gap is particularly pronounced in Sri Lanka’s telecommunications sector, where university students form a vital consumer group with distinct needs and expectations.

This study seeks to address these gaps by exploring the determinants of customer retention among university students in Sri Lanka. It employs a combination of machine learning and deep learning algorithms, leveraging both supervised and unsupervised methods to enhance predictive accuracy. By analysing key retention drivers such as trust, perceived pricing, service provider choice, and additional factors like network quality and service satisfaction, the research aims to provide a comprehensive understanding of customer loyalty within this dynamic demographic. The findings are expected to offer actionable insights for telecommunications providers, enabling them to design targeted strategies to improve retention and foster sustainable growth.

The findings of this study hold significant practical implications. By understanding the factors influencing retention and leveraging predictive analytics, telecommunications providers can implement targeted strategies to reduce churn rates and maximize profitability.

Focusing on university students offers valuable insights into the preferences and behaviours of younger consumers, enabling businesses to better address their needs and foster long-term loyalty.

In the context of a rapidly evolving telecommunications industry, this research contributes to bridging the knowledge gap on customer retention. It provides actionable insights that can help telecom providers adapt to changing customer expectations and achieve sustainable growth in an increasingly competitive market.

In the context of an evolving telecommunications industry, this research contributes to bridging the knowledge gap in customer retention. By providing actionable insights, it helps telecom providers adapt to shifting customer expectations and achieve sustainable growth in an increasingly competitive market.

The reminder of the paper structured as follows: Section 2 reports a the review of literature and the theoretical background of the study. Section 3 provides the methodology that shows how authors conducted this research. Section 4 interprets the empirical results and its interpretation. Finally, the discussion and conclusion are provided in Section 5.

Review of Literature

Customer retention, as opposed to churn management, focuses on building long-term connections with current customers in order to minimize expenses and increase profitability. In the telecommunications sector, which is marked by fast technical breakthroughs and severe rivalry, new techniques are critical for client retention. The COVID-19 epidemic has emphasized the importance of telecommunications, with an increase in digital communication for remote employment, education, and social activities. This trend has raised client expectations and increased the likelihood of switching service providers. University students, as a technologically savvy and important group, give telecom carriers with unique chances to build personalized retention tactics, such as data-centric plans and education-specific incentives. Despite the economic significance of the worldwide telecommunications sector, which is expected to be worth USD 1,657.7 billion in 2020, research on customer retention is minimal, notably in Sri Lanka and among university students.

Customer retention, defined as a company's ability to maintain long-term connections with its customers and ensure their continued loyalty, is critical for increasing profitability and lowering operational expenses (Silva & Yapa, 2009). While customer acquisition is frequently stressed in the early stages of a firm to increase its client base, retention techniques are critical for cultivating loyalty and engagement in order to retain growth. This is especially important in the telecommunications industry, which is marked by severe rivalry and low switching barriers for customers (Jayantha & Geetha, 2014).

Customer retention concepts include loyalty, contentment, trust, and service excellence. Recent research has highlighted customer satisfaction as a key predictor of the inclination to transfer service providers, with elements such as service quality and customer connections having a substantial impact on retention results (Jayantha & Geetha, 2014). For example, research in the Sri Lankan mobile telecommunications market has shown that service quality and customer satisfaction are critical for customer retention (Silva & Yapa, 2009).

Empirical studies have highlighted the multidimensional character of customer retention, which is impacted by demographic, behavioral, and service-related variables.

Demographic characteristics such as age, gender, and income have been investigated, however the results on their influence on retention are varied (Jayantha and Geetha, 2014). Behavioral elements such as perceived pricing justice, network quality, and length of client connections (tenure) have emerged as strong predictors of retention. In price-sensitive businesses such as telecoms, perceived financial sacrifice is highly linked to customer loyalty and retention (Silva & Yapa, 2009).

Service quality is an important predictor of retention. Service quality is often assessed using metrics such as responsiveness, tangibility, empathy, and network dependability (Jayantha and Geetha, 2014). In the telecommunications industry, network quality—which includes dependability and accessibility—is very important, with studies proving a favorable association with retention.

Predictive analytics has been more popular in retention research, with approaches such as machine learning algorithms used to assess factors and anticipate results (Ahmad, Jafar, & Aljoumaa, 2019). Customer churn and retention have been predicted using techniques such as decision trees, random forests, and gradient boosting machines. For example, Ahmad et al. (2019) constructed a churn prediction model utilizing machine learning techniques on a big data platform and achieved an Area Under Curve (AUC) score of 93.3%.

Despite the rising database of research, there are still shortcomings in the integration of determinant analysis and predictive modeling, particularly in the telecommunications sector. Few studies have focused on specific groups, such as university students, who have unique obstacles and possibilities due to their individual behavioral patterns and requirements. University students, a fast growing consumer category, rely significantly on telecommunications services for education, social contact, and leisure, making them an important demographic for retention efforts (Jayantha and Geetha, 2014).

Addressing these shortcomings requires an integrated approach that brings together retention determinant analysis with modern predictive modeling approaches, with a focus on university students in places such as Sri Lanka. Future study can uncover significant elements impacting retention and evaluate prediction model performance by combining classical machine learning and deep learning methods (Ahmad et al., 2019). Telecom providers may gain actionable information by examining determinants such as trust, perceived cost, network quality, and service provider selection.

The findings of such investigations have important practical implications. By leveraging predictive analytics and understanding the particular needs of university students, telecoms companies may develop focused retention strategies, lower churn rates, and increase profitability. This field of inquiry advances understanding in client retention while providing useful methods for sustaining long-term success in an increasingly competitive industry.

Methodology

This chapter outlines the framework employed to ensure the research's reliability and validity. It provides an overview of the research design, data collection, processing techniques, and the tools used, all structured to align with the study's objectives.

Research Design

Research design serves as the blueprint for the methods and processes used in a study, ensuring that data collection and analysis effectively address the research question while minimizing bias. Research designs are commonly classified into qualitative and quantitative approaches or further categorized into descriptive, explanatory, exploratory, and evaluation research designs.

This study adopts a quantitative research design, focusing on measuring and analyzing numerical data to forecast customer retention outcomes. Specifically, it follows an explanatory design, investigating patterns, trends, and classifications in the data to derive informed predictions.

Research Approach

A mixed-method approach integrating both deductive and inductive reasoning is employed. Deductive methods are used to test existing theories, while inductive methods identify patterns and trends, facilitating the development and testing of new hypotheses.

Data

The study relies on primary data collected specifically for this research, as secondary data may not align with its objectives. Data is categorized into qualitative and quantitative forms based on collection and measurement methods. While most data collected is qualitative in nature, it is quantified using a five-point Likert scale to enable numerical analysis.

Sample Design

The research focuses on predicting customer retention within Sri Lanka's telecommunications sector, with a particular emphasis on university students. The target population consists of 135,599 university students enrolled in government universities, as recorded in the Sri Lanka University Statistics (2020). A sample of 384 students was selected, maintaining a 5% margin of error and a 95% confidence interval.

A two-stage probability sampling approach combining cluster sampling and simple random sampling was employed. Students were first grouped based on their universities, from which two clusters—Sabaragamuwa University and the University of Sri Jayewardenepura—were randomly selected. Within these clusters, 384 students were chosen through simple random sampling.

Data Collection Methods and Techniques

Accurate data collection and analysis are essential for constructing a reliable predictive model. To achieve this, a self-administered questionnaire was developed using Google Forms to facilitate easy distribution and data collection. The questionnaire consisted of two sections. The first section gathered basic respondent information, such as gender, university, and telecommunication services used. The second section focused on independent and dependent variables, primarily measured using a five-point Likert scale. Multiple questions covering the same variable were analyzed using factor analysis to consolidate responses while preserving data integrity. A data privacy note was included to ensure respondent confidentiality.

To validate the clarity and relevance of the questionnaire, a pilot survey involving 20 students from both selected universities was conducted. This step helped refine ambiguous questions and ensured alignment with research objectives.

Data Analysis

The collected data will undergo exploratory data analysis (EDA) before developing a machine learning model to forecast customer retention. Given that the dataset is labeled, the study focuses on supervised learning models. Since the outcome variable, customer retention prediction, is continuous, the analysis will employ regression algorithms to model relationships and predict retention rates. Additionally, ensemble methods, which combine multiple models for improved accuracy, will be utilized. Python and relevant libraries will be used for data processing, model training, and evaluation. This approach ensures that the study effectively identifies key factors influencing customer retention while leveraging advanced predictive analytics to generate actionable insights for the telecommunications industry.

Data Preprocessing

Data preprocessing is a critical step in preparing raw data for analysis. This process involves cleaning, transforming, and reducing data to ensure accuracy and usability in predictive modeling.

Data Cleaning:

In the data cleaning phase, missing values will be addressed using different techniques based on the type of data. If the dataset follows a normal distribution, missing values will be replaced with the mean. If the data is skewed, the median will be used instead. For categorical data, the most frequently occurring value will be used for imputation. In cases where decision tree or regression models are applied, missing values can be predicted based on existing patterns. Noisy or inconsistent data will be handled using binning, regression, and clustering methods to smooth variations and maintain dataset integrity. Data integration will be carried out to combine information from multiple sources while ensuring format consistency and resolving discrepancies.

Data Transformation

Data transformation will modify the structure and scale of variables. Smoothing techniques will be applied to reduce noise and highlight significant trends. Aggregation will be used to summarize data for better pattern recognition. Discretization will convert continuous variables into categorical intervals to simplify analysis. Normalization will scale values between 0 and 1 (or -1 to 1), while standardization will be applied to convert non-normally distributed data into a standard normal distribution. Feature encoding will be used to convert categorical data into numerical values. One-hot encoding will create binary vectors, target mean encoding will replace categories with the mean of the target variable, and frequency encoding will assign numerical values based on category occurrence. A chi-square test will be performed to assess the relationship between independent variables and customer retention.

The Pearson chi-square statistic and a significance threshold of 0.05 will determine whether variables have a meaningful association.

Data reduction

Data reduction techniques will help minimize the dataset size while preserving essential information. Dimensionality reduction methods, such as Principal Component Analysis (PCA) and Factor Analysis, will merge highly correlated variables into a smaller set of components. Numerosity reduction will use model-based storage techniques, such as regression functions, to reduce dataset complexity. Data compression, in either a lossy or lossless manner, will further optimize storage while maintaining critical features. Factor analysis will be specifically used to reduce the dimensionality of service-related data from questionnaires, such as customer satisfaction, service quality, and network reliability.

Exploratory Data Analysis (EDA)

After preprocessing, exploratory data analysis will be performed to identify data distribution patterns and relationships. This step will involve examining individual variables, assessing their distributions, and detecting potential outliers. Univariate analysis will focus on single variables, using visual tools such as pie charts, bar charts, histograms, and box plots to interpret data trends. Bivariate analysis will explore relationships between multiple variables using scatter plots, heatmaps, and stacked bar charts to identify correlations and dependencies.

Train-Test Split

To ensure robust model performance, the dataset will be split into two subsets: a training set, comprising 80% of the data, and a test set, containing the remaining 20%. The training set will be used to develop predictive models, while the test set will evaluate their accuracy on unseen data.

Model Selection

Supervised learning models will be explored to identify the most suitable predictive algorithm for customer retention. Linear regression will be applied to estimate continuous variables by defining relationships between dependent and independent variables. Logistic regression will be used for classification tasks, distinguishing between retained and churned customers. Decision trees will be utilized to segment data into hierarchical structures for both classification and regression tasks. Support Vector Machines (SVM) will classify data points by finding an optimal hyperplane that maximizes the margin between different categories. Naïve Bayes, based on Bayes' Theorem, will be employed for classification, assuming independence among features.

Ensemble methods will also be considered to improve predictive accuracy. Random Forest, which aggregates multiple decision trees, will enhance reliability by reducing overfitting. Gradient Boosting Machine (GBM) will iteratively refine weak learners, focusing on previous errors to improve prediction accuracy. XG Boost, a high-performance variant of gradient boosting, will be used for its computational efficiency and ability to handle large-scale datasets effectively.

Model Training

The model training phase will involve learning optimal weights and biases using labeled data. The objective is to minimize the loss function, which represents the difference between predicted and actual values. Empirical risk minimization techniques will be applied to ensure that the model generalizes well to new data.

Hyperparameter Tuning

Hyperparameter tuning will optimize the model's performance by testing different parameter combinations. A range of configurations will be evaluated to identify the best-performing settings, ensuring the highest predictive accuracy.

Model Evaluation

Model evaluation will be conducted using appropriate performance metrics based on the type of predictive task. Classification models will be assessed using accuracy, precision, recall, and F1-score. A confusion matrix will provide insights into true positive, false positive, true negative, and false negative classifications. Log-loss will be used to measure classification uncertainty, while the area under the ROC curve (AUC) will assess the model’s ability to differentiate between classes.

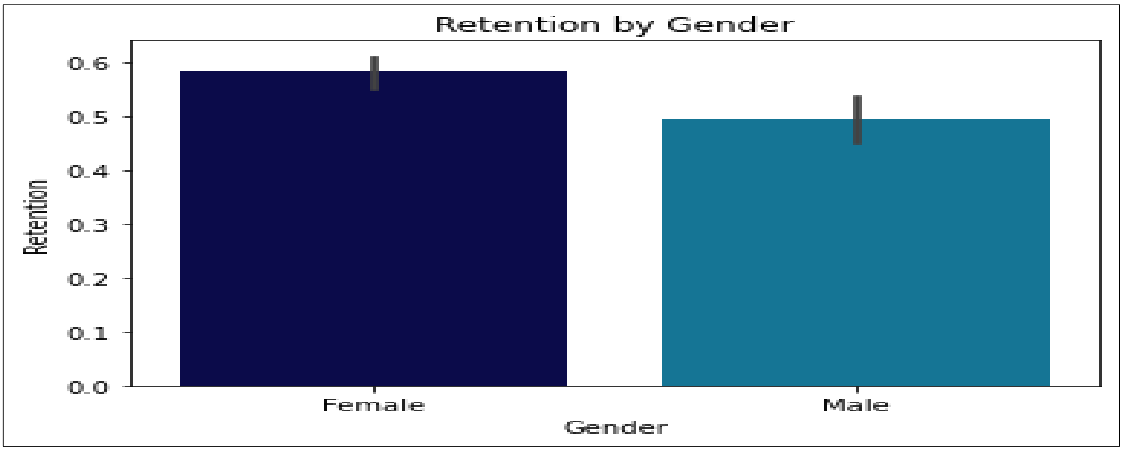
Regression models

Regression models will be evaluated using R-squared and adjusted R-squared values to measure the proportion of variance explained by independent variables. Mean Squared Error (MSE) will be used to quantify average squared prediction errors, while Root Mean Squared Error (RMSE) will provide a standard deviation measure of residuals. Mean Absolute Error (MAE) will be considered to assess the average absolute differences between predicted and actual values.

By comparing models based on these evaluation metrics, the most effective algorithm will be selected to support customer retention strategies in Sri Lanka’s telecommunications industry.

Results and Data Analysis

The original data set consists of 36 variables and 389 records including both qualitative and quantitative type variables. This part includes the exploratory analysis of data mostly represented using graphs. For this, the researcher used Python Programming Language with



the libraries Matplotlib and Seaborn and Jupyter Notebook as the tool to perform the analysis.

Figure 1: Effect of Gender on Retention

Considering the potential influence of gender on customer retention, the data suggests that females exhibit a higher retention rate compared to males. However, it is important to interpret this finding with caution, as the observed difference in retention rates may be a relative measure specific to this dataset. The gender-based disparity in retention rates could be influenced by factors unique to the sample, and further analysis is necessary to determine whether this trend holds in broader contexts or is merely an artefact of the data.

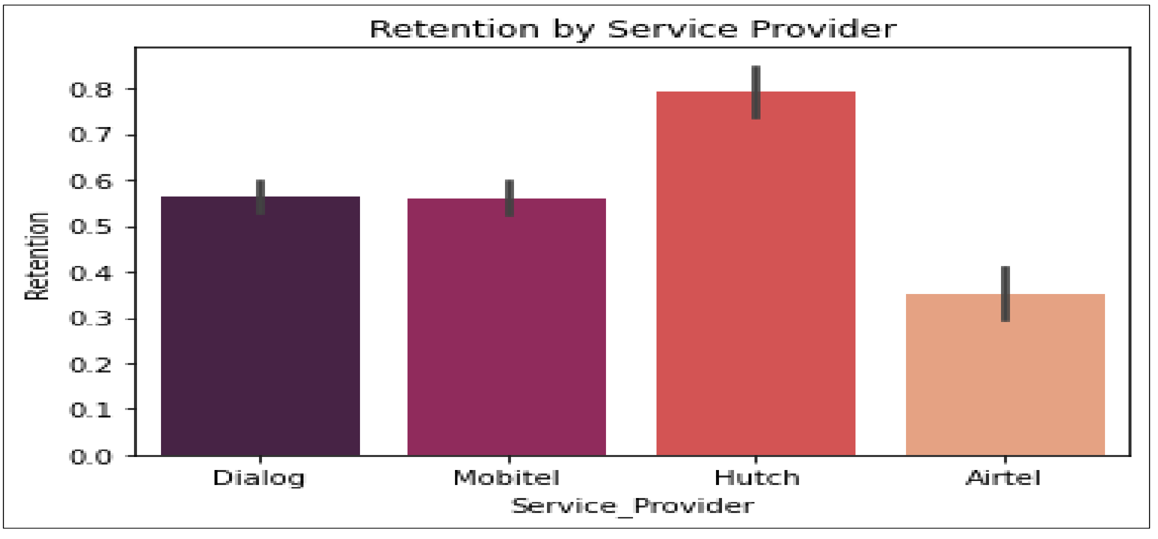


Figure 2: Retention on Service Provider

When examiningfig the relationship between client retention and service provider, Hutch appears to demonstrate a higher retention rate compared to other providers. Although

Hutch has a smaller customer base than some of its competitors, the data suggests that Hutch has effectively managed to foster customer satisfaction and loyalty. This higher retention rate may indicate that Hutch has implemented strategies or features that enhance customer satisfaction, despite the smaller customer base. Further investigation would be necessary to identify the specific factors contributing to this improved retention.

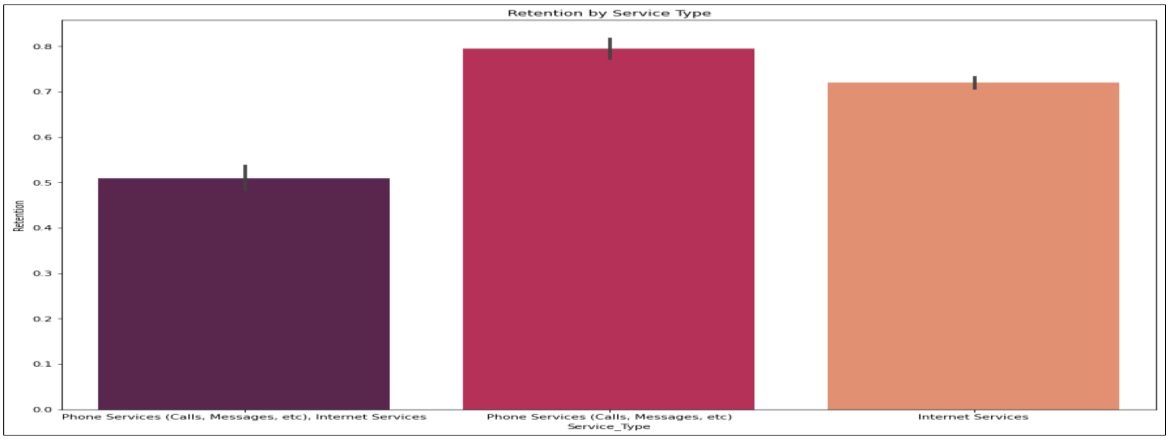


Figure 3: Retention by Service Type

The data indicates that customers who utilize only phone services from telecommunications providers exhibit higher retention rates compared to those who use both phone and internet services. This observation suggests the possibility of inconsistencies or issues in the quality of internet services provided by these companies. It is plausible that certain aspects of Internet service offerings may contribute to reduced customer satisfaction and, consequently, lower retention among customers who use both services. Further analysis is required to explore potential service quality discrepancies and their impact on customer retention.

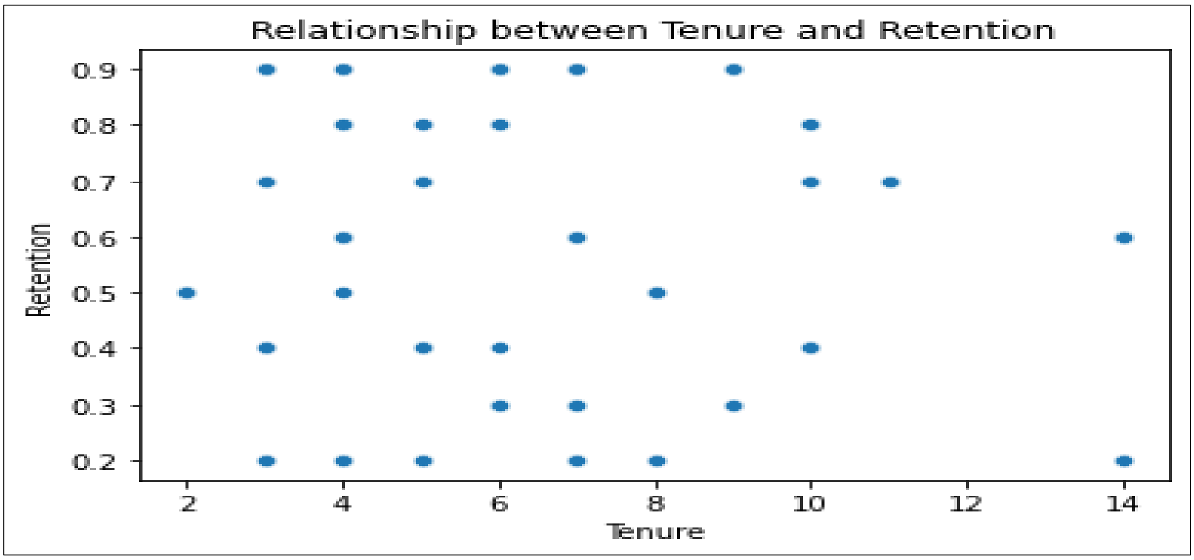


Figure 4.: Relationship between Tenure and Retention

The scatter plot presented above illustrates the relationship between tenure and customer retention. Based on the visual analysis, there appears to be no significant correlation between these two variables. If a relationship exists, it is likely to be of minimal strength, suggesting that tenure may not be a substantial factor influencing customer retention in this dataset. Consequently, any potential relationship may be considered negligible and not warrant further consideration in the analysis.

A chi-square test was conducted for each variable to assess whether a statistically significant relationship exists between these variables and customer retention. The variables considered in the analysis are treated as independent variables for the study, which include both the aforementioned factors and additional variables relevant to customer retention. These variables encompass factors identified in prior research, such as customer satisfaction, network quality, service quality, company image, and trust, among others. The data for these variables were collected using a Likert scale. A summary of the findings is presented below.

Table 1: Summary of Chi-square Test

Independent Variable	P - value	Accept/Reject
University	0.000	Accept
Gender	0.000	Accept
Service Provider	0.000	Accept
Service Type	0.000	Accept
Tenure	0.000	Accept
Subtype	0.000	Accept
Var1	0.000	Accept
Var2	0.000	Accept
Var3	0.000	Accept
Var4	0.000	Accept
Var5	0.000	Accept
Var6	0.000	Accept
Var7	0.000	Accept
Var8	0.000	Accept
Var9	0.000	Accept
Var10	0.000	Accept
Var11	0.000	Accept
Var12	0.000	Accept
Var13	0.000	Accept
Var14	0.000	Accept
Var15	0.000	Accept
Var16	0.000	Accept
Var17	0.000	Accept
Var18	0.000	Accept
Var19	0.000	Accept

Var20	0.000	Accept
Var21	0.000	Accept
Var22	0.000	Accept
Var23	0.000	Accept
Var24	0.000	Accept
Var25	0.000	Accept
Var28	0.000	Accept
Var29	0.000	Accept

Based on the results of the chi-square test, all the variables selected by the researcher to predict customer retention appear to exhibit a statistically significant relationship with customer retention. This suggests that each of the chosen variables plays a meaningful role in influencing customer retention in the context of this study.

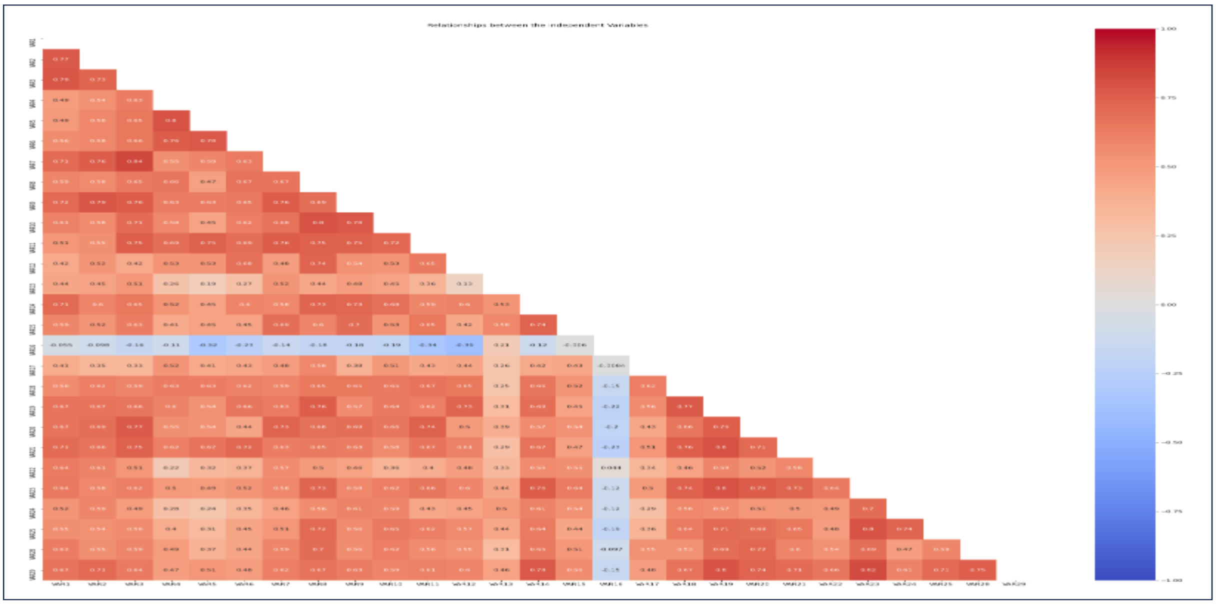


Figure 5: Relationships between the Independent Variables

A data set has been collected based on variables identified in prior research. The objective is to evaluate potential relationships between these variables to determine whether factor analysis is warranted to optimize the results produced by the model. This analysis is conducted independently of the demographic factors, which were identified by the researcher to characterize the sample units and define the independent variables for the study.

Preliminary observations, as visualized in the map above, indicate that some variables exhibit strong correlations. This is noteworthy because relationships among independent variables may introduce multicollinearity, which can adversely affect the performance and interpretability of the fitted model. The outcomes derived from this analysis are presented below.

Table 2: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.870
Bartlett's Test of Sphericity	Approx. Chi-Square	13746.436
	df	351
	Sig.	.000

Before conducting the factor analysis, preliminary tests were performed to assess the suitability of the data for this procedure. The results of these tests are as follows:

- Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy: The KMO value was 0.870, indicating that the sample size is adequate for factor analysis.
- Bartlett's Test of Sphericity: This test yielded a significant p-value of 0.000, which is below the threshold of 0.05 at a 95% confidence level. This result confirms that the variables exhibit sufficient correlation to justify performing factor analysis.

Based on an eigenvalue analysis, it was determined that the variables should be grouped into five factors. The classification of these factors is as follows:

Factor 1 – Corporate Image and Customer Experience

The characteristics most strongly associated with this factor, as indicated by the highest loading values, pertain to the organization's reputation among customers and competitors. Additionally, they encompass aspects of the customer experience, including how the organization interacts with customers and the customers' perceptions of these interactions.

Factor 2 – Perceived Price

This factor encompasses the attributes of perceived price, reflecting customers' evaluation of the fairness of the company's pricing and the value derived from the money spent on the services provided.

Factor 3 – Customer Satisfaction and Network Quality

This factor comprises attributes related to customer satisfaction, including overall satisfaction, satisfaction with the services received compared to expected levels, and satisfaction relative to the perception of an ideal service provider. Additionally, it incorporates characteristics associated with network quality, providing a comprehensive assessment of the quality of both phone and internet services.

Factor 4 – Commitment and Service Quality

This factor encompasses attributes related to commitment, including both affective commitment, reflecting emotional attachment, and calculative commitment, representing economic considerations. It also includes characteristics associated with service quality, such as the overall quality of services provided by the company to its customers.

Factor 5 – Trust

This factor exclusively addresses trust-related characteristics, including the extent to which customers trust the company and situations where customers may feel uncertain about trusting the company.

Based on these identified factors, the researcher selected them as independent variables to proceed with the study, alongside demographic variables such as University, Service Provider, and Service Type. Collectively, these factors account for 78.945% of the variability in the dependent variable, customer retention.

To ensure consistency and mitigate issues related to varying data ranges or variances, the data was standardized before analysis. This step aligns with the prior standardization of the identified factors during the factor analysis.

A summary of the models fitted to the data and the accuracy scores obtained for each model is presented below.

Table 3: Summary of the Models Tested

Model	R^2 value	MAE	MSE	RMSE
Linear Regression	0.5353	0.1372	0.02652	0.1629
Decision Tree Regression	1.0	7.828495686459438e-17	1.1298789007071783e-32	1.0629576194313574e-16
Support Vector Machine Regression	0.8636	0.0818	0.0078	0.0882
Random Forest Regression	0.9929	0.0052	0.0004	0.0201
XGBoost Regression	0.9999	0.0003	3.01818365117548 15e-07	0.0005
Gradient Boost Regression	0.9893	0.0169	0.0006	0.0248

The researcher concluded that the best-fitted model is the one with the lowest Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), as these metrics minimize the influence of errors. Additionally, the model with the highest R^2 value was identified as optimal. Based on these criteria, the Decision Tree Regressor was determined to be the most suitable model for predicting customer retention among university students in Sri Lanka within the telecommunications industry.

To assess whether the model exhibited overfitting, the researcher evaluated its performance using the highest accuracy score of 100%. This was tested by comparing the training and test scores to identify any discrepancies that might indicate overfitting. The results are as follows:

- Test Score: 1.0 (100%)
- Training Score: 1.0 (100%)

The identical performance of the training and test scores suggests that the model does not suffer from overfitting and generalizes well to unseen data.

The researcher concluded that the selected model does not exhibit overfitting issues, as both the training and test scores demonstrate equivalent performance. This indicates that the model generalizes effectively to unseen data. Consequently, the **Decision Tree Regressor** was chosen as the final model for predicting customer retention in the telecommunications industry among university students in Sri Lanka. The corresponding plot is presented below.

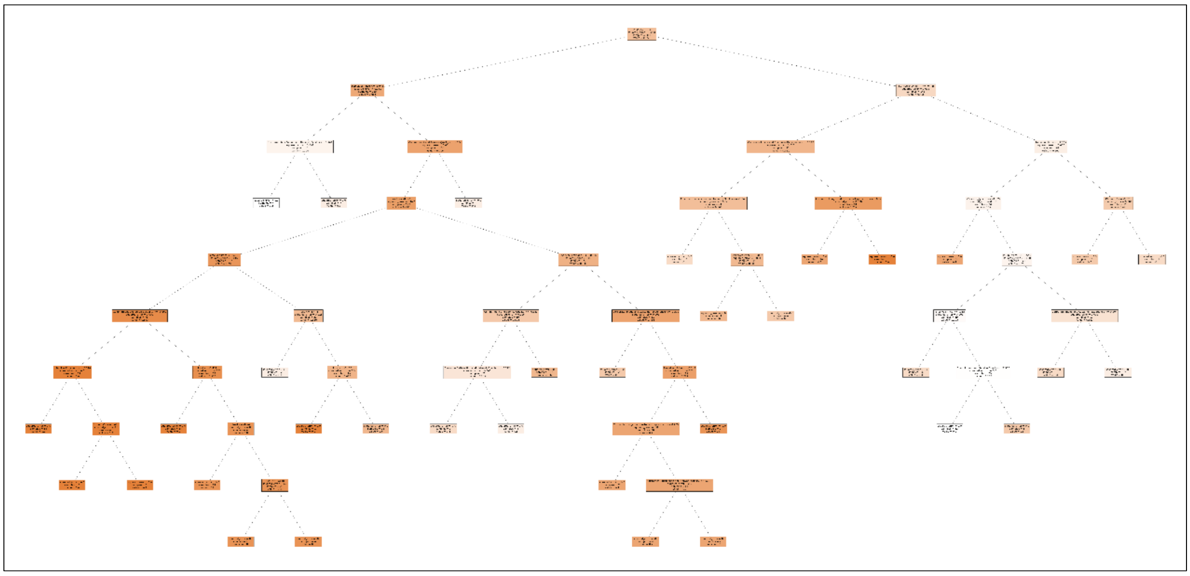


Figure 6: **Decision Tree for the Model Developed**

Unlike traditional statistical methods, this approach does not include a mechanism for testing the statistical significance of each variable using p-values. Instead, it offers a valuable feature that evaluates and displays the importance of each variable to the model. This allows for a comparative analysis of feature contributions, highlighting the most and least influential variables in predicting the dependent variable, customer retention.

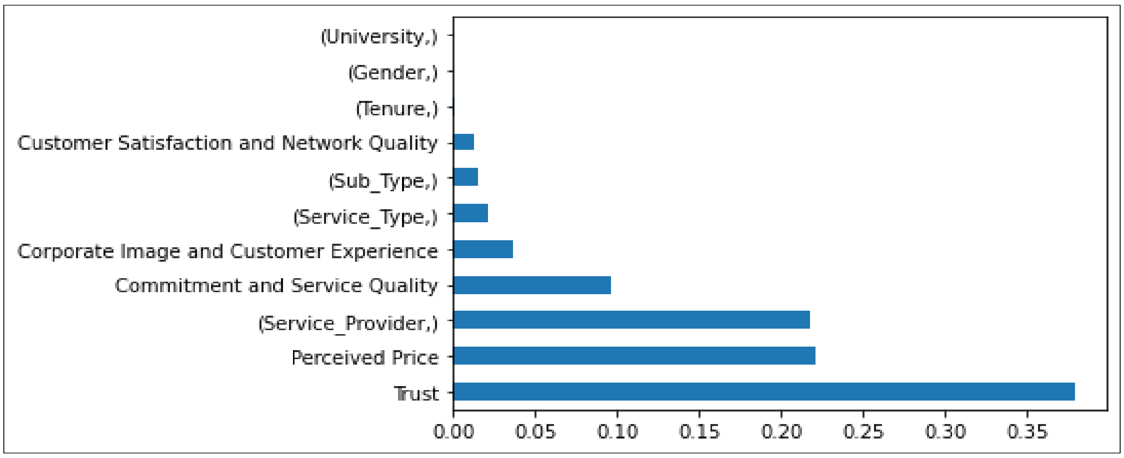


Figure 7: Feature Importance

Based on the analysis of the above graph, it can be concluded that the most significant variable influencing customer retention is Trust. In contrast, the variables University, Tenure, and Gender appear to have minimal impact on customer retention, as evidenced by their low predictive contribution in the Decision Tree Regressor model. This suggests that Trust is the primary factor driving customer retention, while demographic factors such as University affiliation, Tenure, and Gender exhibit negligible effects.

Discussion and Conclusion

In this study, an exploratory analysis was performed to identify the important determinants influencing customer retention in the telecom business. The investigation revealed numerous important factors impacting retention, including gender, institution affiliation, service provider, service kind, subscription plan, and trust. The Decision Tree Regressor model was used to find the most important aspects influencing retention, with trust appearing as the most critical determinant. These findings are consistent with and build on previous research in the subject of customer retention, which has repeatedly identified trust and service quality as key determinants driving customer loyalty.

Several customer relationship management (CRM) and customer loyalty studies have found that trust plays an important role in retention. For example, Morgan and Hunt's (1994) pioneering work on relationship marketing underlined the need of trust in developing long-term connections between businesses and their consumers. Similarly, recent telecommunications studies have emphasized the importance of trust as a predictor of customer satisfaction and retention (Kassim & Abdullah, 2010). Consistent with previous findings, the current study discovered that trust was the most critical factor influencing customer retention, implying that telecom businesses should emphasize trust-building initiatives to create loyalty.

The current study's trust-building strategies, such as regular communication, upholding ethical standards, and avoiding harmful policies, are consistent with previous research, which has found that transparent and ethical practices help cultivate trust and, as a result, increase customer loyalty (Sirdeshmukh, Singh, & Sabol, 2002).

Another notable finding from the present study is the influence of perceived price on customer retention. Similar to the findings in this study, previous research has shown that the perception of price plays a vital role in customer satisfaction and retention in the telecommunications sector (O'Malley & Tynan, 2000). The concept of perceived price refers to how customers evaluate the value they receive from a service relative to its cost. While this study did not suggest direct price reductions as a strategy, it recommended offering additional benefits or features within existing packages to enhance perceived value. This approach aligns with the findings of Taylor and Neslin (2005), who suggested that customers in competitive industries, like telecommunications, are more likely to stay loyal if they perceive that the service offers greater value compared to its price, even without direct discounts.

This approach is consistent with the findings of Taylor and Neslin (2005), who proposed that consumers in competitive industries, such as telecoms, are more likely to remain loyal if they believe the service provides higher value than its price, even in the absence of direct discounts.The current study's findings on the association between service type and customer retention contribute to a growing body of research investigating the role of service quality in customer retention. The survey discovered that customers who chose phone-only services had greater retention rates than those who subscribed to both internet and phone services, indicating possible concerns with internet service offerings. This conclusion is consistent with previous research that have highlighted the crucial significance of service quality in customer retention, particularly in telecoms (Hunt & Morgan, 1995). According to Zeithaml, Berry, and Parasuraman (1996), quality of service is a significant predictor of customer satisfaction, which impacts retention. The current study's figuring that customers using Hutch, despite their smaller customer base, had higher retention rates indicates that Hutch may be providing superior service quality, which is consistent with Berry and Parasuraman's (1991) research, which concluded that service providers with higher quality offerings tend to have higher customer retention rates, even if they serve a smaller percentage of the consumer base.

Gender disparities in client retention, with female consumers showing greater rates, are consistent with earlier research on gender and loyalty. For example, Petersen and Welch (2004) discovered that female consumers have higher loyalty and are more likely to stick with a service provider over time, especially if they believe the service is of good quality and meets their demands. The current study's prevalence of female clients may explain some of the observed retention trends. However, as noted in the report, greater retention rates among female consumers might be attributed to demographic imbalances, as females made up the majority of respondents. Nonetheless, the study's findings are consistent with larger studies indicating that demographic characteristics such as gender might impact customer retention behavior in the telecommunications business. The study also discovered that tenure had no significant link with customer retention, which contradicts earlier research indicating that longer-tenured consumers are more likely to demonstrate stronger loyalty and retention. For example, Boulding, Kalra, and Staelin (1999) discovered that consumers with longer tenure have a greater grasp of the service and are more likely to remain loyal. However, the current study's lack of a significant relationship between tenure and retention could be attributed to a more complex interplay of factors, such as the increasing commoditization of telecom services or the role of external factors such as competition and service quality in influencing customer behavior.

The lack of relationship between university affiliation and retention in the model also warrants discussion. While some studies have found that demographic factors, including educational background, can influence customer preferences and retention (Arora & Verma, 2016), the present study did not find university affiliation to be a significant factor. This could be attributed to the fact that university affiliation is not necessarily linked to customer loyalty or satisfaction in a meaningful way, particularly when compared to more direct factors like service quality and trust. It also suggests that customer loyalty in the telecommunications industry may be more influenced by service-related factors than by demographic characteristics.

One of the most important discoveries of this study is the lack of overfitting, despite the Decision Tree Regressor model's high accuracy score. The model's dependability was demonstrated by the equal performance of the training and test datasets, indicating that the model generalizes effectively to new data. This finding is consistent with previous study on machine learning in customer retention, which revealed decision tree models to give accurate and interpretable insights into factors influencing loyalty (Madhavi and Srinivas, 2013).

However, experts warn against depending too much on a single model, particularly when it achieves very high accuracy, since this may signal a danger of overfitting. In this context, the current study's rigorous validation and equal performance of training and test datasets address these kinds of concerns.

Finally, the findings of this study confirm and extend previous studies on customer retention and loyalty in the telecommunications business. Trust, perceived pricing, service quality, and demographic characteristics such as gender have been identified as major determinants of customer retention, consistent with the findings of various research on relationship marketing and customer loyalty. The findings-based recommendations, such as trust-building techniques, increasing perceived value, and focusing on service quality, offer practical insights for telecom firms looking to enhance customer retention. Furthermore, the study's use of machine learning techniques, such as the Decision Tree Regressor, to estimate customer retention demonstrates the increasing importance of data-driven approaches in analyzing and foreseeing consumer behavior in the telecommunications business.

****

Ms. L. P. Himali is a Senior Lecturer in Statistics in the Department of Economics and Statistics at Sabaragamuwa University of Sri Lanka. She holds a Bachelor of Arts (Hons) in Statistics from Sabaragamuwa University, which she completed in 2008 with second-class honors. In 2014, she earned a Master of Science in Applied Statistics from the University of Colombo, further strengthening her interdisciplinary approach to teaching and research.Since joining the Department of Economics and Statistics at Sabaragamuwa University in 2009, Ms. Himali has been actively involved in undergraduate education within the Faculty of Social Sciences and Languages. Her research interests encompass machine learning techniques, statistical methodologies for analyzing binary outcomes, and quantitative analyses of unemployment trends, causes, and consequences. She also specializes in time-to-event data analysis with applications in healthcare, labor economics, and risk assessment, employing advanced statistical and computational techniques. Ms. Himali has published her research in numerous local and international journals and has presented her work at prestigious conferences and symposia worldwide. In addition to her academic contributions, she serves on the editorial boards and review panels of several journals and international conferences, further demonstrating her commitment to advancing statistical research and practice.

Ms. J. D. Abeysekara holds a Bachelor of Arts (Hons) in Statistics from Department of Economics and Statistics, Sabaragamuwa University of Sri Lanka. Her research interests encompass machine learning techniques, Binary data analysis and Inferential statistics.

****

References

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning and social network analysis in big data platform. *Journal of Big Data, 6*(1), 28. https://doi.org/10.1186/s40537-019-0191-6

Arora, A., & Verma, A. (2016). The role of demographic factors in customer satisfaction and retention in telecommunications. *International Journal of Business and Management Studies, 5*(3), 102-110.

Berry, L. L., & Parasuraman, A. (1991). Marketing services: Competing through quality. *Free Press*.

Boulding, W., Kalra, A., & Staelin, R. (1999). The customer experience and customer retention: A multidimensional perspective. *Journal of Marketing, 63*(4), 75-89. https://doi.org/10.1509/jmkg.63.4.75.18298

Hunt, S. D., & Morgan, R. M. (1995). The comparative advantage theory of competition. *Journal of Marketing, 59*(2), 1-15. https://doi.org/10.1177/002224299505900201

Jayantha, D., & Geetha, T. (2014). Determinants of customer retention with special reference to mobile telecommunication industry in Sri Lanka. *International Conference on Business Management*.

Kassim, N. M., & Abdullah, M. S. (2010). The influence of service quality on customer satisfaction: A study of the telecommunication industry in Malaysia. *International Journal of Business and Management, 5*(5), 61-69. https://doi.org/10.5539/ijbm.v5n5p61

Madhavi, S., & Srinivas, P. V. (2013). Customer retention using data mining techniques in telecommunications. *International Journal of Computer Science and Mobile Computing, 2*(6), 482-489.

Morgan, R. M., & Hunt, S. D. (1994). The commitment-trust theory of relationship marketing. *Journal of Marketing, 58*(3), 20-38.

O'Malley, L., & Tynan, C. (2000). Relationship marketing in consumer services. *Journal of Marketing Management, 16*(5-6), 503-520.

Petersen, J. A., & Welch, L. S. (2004). Gender differences in loyalty and retention behavior in telecommunications. *Journal of Marketing Theory and Practice, 12*(1), 10-17.

Sirdeshmukh, D., Singh, J., & Sabol, B. (2002). Consumer trust, value, and loyalty in relational exchanges. *Journal of Marketing, 66*(1), 15-37.

Silva, K. A., & Yapa, S. T. W. S. (2009). Customer retention: With special reference to telecommunication industry in Sri Lanka. *International Conference on Business and Information*.

Sri Lanka University Statistics 2020. (n.d.). *University Grants Commission - Sri Lanka.*

Taylor, E. R., & Neslin, S. A. (2005). The role of loyalty in customer retention: A comparison of telecommunications service providers. *Journal of the Academy of Marketing Science, 33*(1), 101-110. https://doi.org/10.1177/0092070304265293

Zeithaml, V. A., Berry, L. L., & Parasuraman, A. (1996). The behavioral consequences of service quality. *Journal of Marketing, 60*(2), 31-46.