

Customer Churn Prediction in the Telecom Industry Via a Clustering-Random Forest-XGBoost Approach: Discovering the Role of Delta in PCA Loadings

Chien-Sing Lee[1,2(✉)] and Sharini Sivadas Nanda[1]

[1] Faculty of Engineering and Technology, Sunway University, Subang Jaya, Malaysia
chiensingl@sunway.edu.my
[2] IEEE, New York City, USA
csleester@gmail.com

Abstract. This study is framed within Technology Acceptance Models 1, 2, 3, in identifying key features which influence churn in the telecommunications industry. Data is obtained from 225 Malaysian survey respondents via random sampling. Mining adapts from Bach et al.'s 3-stage approach. For Phase 1, *k*-means, *k*-medoids and hierarchical agglomerative clustering are utilized to determine the most suitable clustering technique to segment customers. Subsequently, bar graphs form more detailed preliminary inferences and cluster profiling. In Phase 2a, features selected via correlation analysis are complemented with principal component analysis. In Phase 2b, non-linear relationships are assessed using Random Forest and XGBoost. The top 15 identified key features are ranked and compared. Outcomes are compared with findings in Phase I and Phase 2a. Findings highlight the viability of a combination of these techniques, depending on the research objectives, the nature of the dataset, the objectives/ nature of the respective algorithms and the domain investigated. Correlation (covariance matrix) and PCA loadings in Phyton reveal different polarities. Moreover, the delta value of PCA loadings are consistent across all clusters. Chi squared and *p* in CHAID performs comparably with RF and XGBoost. A very humble analogy to quantum states due to polarity, robustness in correlational feature selection reflecting information gain and communication's parity bits to ensure data integrity in interconnected systems due to covariance-PCA loadings' delta as identifiers, may be useful for STEM education, depending on model fit, and to finance, e.g. Buy Now Pay Later.

Keywords: Customer Churn Prediction · Telecom · clustering · PCA loadings · delta · Random Forest · XGBoost · Information Science

1 Introduction

The telecommunications industry is a service-oriented industry that significantly contributes to national income especially in developing countries, such as Malaysia. The Information and Communications Technology (ICT) industry has contributed 23.2%

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16759, pp. 289–307, 2027.
https://doi.org/10.1007/978-3-032-30518-3_19

to Malaysia's GDP in 2021. This is projected to rise to 25.5% by 2025 [1]. Similarly, Mordor Intelligence's Industry Report [2] estimates that the 2025 Malaysian telecom market is valued at USD 9.45 billion. However, it is expected to reach a value of USD 11.54 billion by 2030 at a compound annual growth rate (CAGR) of 4.07%.

Despite the substantial business growth of this industry in recent years, telecommunication firms are in an extremely competitive market because many regions have multiple telecommunications providers offering similar services. This makes it easy for customers to switch providers, if they are dissatisfied with the speed, reliability and affordability of the telecommunication services [3]. Customer Churn, where existing customers choose to unsubscribe from their current telecommunication service provider over a given period [4, 5] is thus a significant challenge.

1.1 Problem Statement

[6] point out that the telecommunications industry has the largest annual churn rate of any major industry (over 30%). Thus, companies often turn to 3 main revenue strategies, i.e., attracting new customers, upselling to the existing customers and increasing customer retention duration [3]. These strategies include aggressive pricing strategies, discounts, promotions, and bundle services to differentiate themselves and attract customers. However, if the cost to acquire new customers is high and the newly acquired customers churn too quickly, this will lead to poor returns on acquisition efforts. Furthermore, customer retention programs are sometimes ineffective, because telecommunication firms sometimes offer a one-size-fits-all retention program to entice them back into resubscribing the service. This fails to address individual needs. Hence, personalized customer retention approaches for churn prediction are critical.

1.2 Research Objectives

Research in this field typically focuses on three key areas, i.e., churn prediction modelling, customer segmentation and model explainability [4, 6, 7]. For churn prediction modelling, we need to ensure that the predictive model adequately considers all relevant factors and corresponding relationships to derive improved churn prediction accuracies and generate rules to aid data-driven decision making.

Given the data collected during this study, the research objectives are to:

a) analyse what the *k*-means, *k*-medoids and hierarchical clustering algorithms will reveal about the clusters;
b) evaluate which features for each cluster are more significant based on correlation;
c) identify which churn prediction features for each cluster are more significant based on Random Forest (RF) and XGBoost;
d) test and compare feature selection performance via correlation with PCA loadings vs. RF/XGBoost's
e) gauge potential for chi-squared automatic interaction detector (CHAID) to enhance RF/XGBoost on model explainability
f) investigate role of delta between PCA loadings
g) possibilities of extending to STEM education

1.3 Research Questions

Our research questions are:

RQ1: what will *k*-means, *k*-medoids and HAC reveal about the dataset's and each cluster's distribution?

RQ2: which features for each cluster are more significant based on correlation analyses and PCA loadings?

RQ3: which features for each cluster are more significant based on RF and XGBoost?

RQ4: how comparable is the Chi-Squared Automatic Interaction Detector (CHAID) with RF, XGBoost?

RQ5: can PCA loadings' delta be analogized for STEM education?

1.4 Research Hypotheses

From the problem statements, hypotheses corresponding to the research objectives are:

- Hypothesis 1: *k*-means will perform better than *k*-medoids for the dataset since it is expected to be imbalanced.
- Hypothesis 2: Tenure and contract type are important predictors not only through intra-cluster correlation analysis but also via principal component analysis (PCA).
- Hypothesis 3: Tenure and contract type will be key predictors across all clusters in RF and XGBoost.
- Hypothesis 4: CHAID's feature selection is comparable to RF/XGBoost, and its correlation complemented by *p* provides a leaner solution
- H5: There are analogical possibilities to quantum computing, which may spark interest, but it needs to be scoped and validated as per scientific inquiry

2 Literature Review

2.1 Significant Factors which Influence Churn

[8, 9] have identified Service Usage Amount, Internet Type, Customer Service Interactions, and Promotional Discounts as the key churn drivers [10, 11]. have pointed to high Monthly Invoices and Billing inaccuracies as large contributors to churn. Contract Type and Tenure also indicate high influence, as month-to-month subscribers and customers with shorter tenure are more likely to churn [12, 13]. Tenure, Contract Type, Monthly Charges, Billing Inaccuracies, Internet Type, Service Usage Amount, Service Type, Customer Service Interactions, Promotional Discounts, Total Services Subscribed are the strongest churn predictors in the telco industry.

These are mainly financial [14] in nature. Hence, an area of interest is the Technology Acceptance Model (TAM) 1 by [15], TAM 2 on social factors [16] and TAM 3 on human factors by [17]. TAM 1, 2, 3 are illustrated in Fig. 2a below.

2.2 Mining Techniques

In terms of mining, clustering, and tree-based machine learning models for clustering/customer segmentation are of interest to this study. These will enable cluster profiling. Next, correlation analysis with Principal Component Analysis (PCA) loadings are reviewed. These enable feature selection, specific to each segment.

2.2.1 Clustering

k-means is considered to be one of the more well-known unsuper-vised clustering techniques, mainly due to its simplicity. However, *k*-medoids and hierarchical agglomerative clustering (HAC) may be more useful when the dataset is more balanced. The precision, recall, and F-measure performance of *k*-means, *k*-medoids and HAC are compared and evaluated using the Silhouette Score (SS) and Davies-Bouldin (DB) Index to determine cluster quality [18–21].

k-means and k-medoids involve initialization, assignment and update. *k*-means splits the dataset into *k* clusters while minimizing the within-cluster sum of squares, which is the total sum of squared distances between each point and its assigned cluster centroid (initialization). The distance between points and centroids are calculated as the Euclidean distance. The goal of *k*-means is to reduce this distance. In the assignment step, each point is assigned to the nearest centroid. With each iteration, the distance to the centroid is updated. The iteration ends when it reaches convergence. *k*-medoids uses minimal average dissimilarity to other data points within a cluster, instead of centroids. Hence, medoids are more robust to outliers (extreme values).

HAC takes a bottom-up approach. Similar clusters are merged based on linkage type, e.g. Ward (minimizes variance within clusters), Average (average distance between points) or Complete (maximum distance between points). This process continues until all points have been agglomerated. To visualize the best number of clusters, a dendrogram is created and the optimal cut-off threshold chosen.

2.2.2 Correlation/Covariance and PCA Loadings

Pearson correlation can enable identification of important churn drivers in the high-risk churn cluster [7, 11]. Correlation, however, has its own limitations as it is not able to capture nonlinear relationships [6]. According to [22], an r of 0.1–0.29 implies weak correlation, 0.3–0.5 moderate, and > 0.5 strong. In Phyton, when PCA is applied on raw data, covariance matrix is used. Chi squared is not used to compute PCA loadings. PC1 loadings refer to that of the variable's, whereas PC2 loadings refer to that of churn.

2.2.3 Tree-Based Machine Learning Models

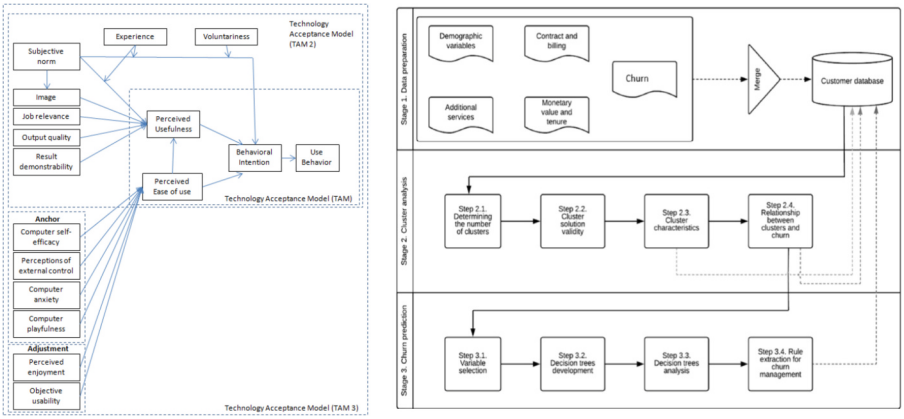
Decision trees build on information gain or its inverse, entropy. They can manage small tabular datasets effectively and accurately, while using less computational resources compared to deep learning models [23–25]. Rule-extraction provides explainability/identification of optimal paths for personalizing customer retention strategies [26].

For Random Forest, the tree splits until the stop criteria such as max_depth and min_samples_leaf are met. The final RF prediction is obtained by aggregating the prediction of all trees by majority voting. XGBoost optimizes loss function through gradient boosting, i.e., XGBoost builds each new tree by correcting the residual errors from the previous prediction. Chi-Squared Automatic Interaction Detector (CHAID) [27–29] applies chi square and p values.

3 Methodology

3.1 Research Design

The Technology Acceptance Models (TAM) 1, 2, 3 (Fig. 1a) are used to frame the research as the authors regard churn as a rejection of the technology [28].'s 3-stage model is presented in Fig. 1b.



(a) **(b)**

Fig. 1. (a) TAM 1, 2, 3 (b) [28]'s model

Ten hypotheses (H1–H10) are developed based on the significant features identified from literature reviewed in Sect. 2 and mapped to TAM 1, 2, 3 in Table 1. The mapped model (Table 1) is used to develop the survey questions and to develop customer retention strategies.

Table 1. Hypotheses, TAM constructs and explanations

Features/Factors	Mapped TAM Construct(s)	Hypotheses
H1: Promos / Discounts (Q18) [8]	Subjective Norm	If family, friends and colleagues enjoy the promotions or discounts provided, then this subjective norm in terms of social influence increase the user's perceived usefulness of the telecom service.

(*continued*)

Table 1. (*continued*)

Features/Factors	Mapped TAM Construct(s)	Hypotheses
H2: Tenure (Q9) [12, 13]	Result Demonstrability	If tenure is longer, then the benefits and reliability of the service become more evident over time (Result Demonstrability), which in turn increases the perceived usefulness of the telecom service.
H3: Monthly Charges (Q12) [10, 11]		If monthly charges are higher, then the benefits and reliability of the service become more evident over time (Result Demonstrability). This increases perceived usefulness.
H4: Usage (Q15) [8]		If usage is higher, then the benefits and reliability of the service become more evident over time (Result Demonstrability), which in turn increases the perceived usefulness of the telecom service.
H5: Internet Type (Q14) (Ribeiro et al., 2023–4) [8, 9]	Output Quality	If the customer is subscribed to a higher-performing Internet type e.g. fibre optics rather than DSL, then the output quality increases with faster speeds and greater reliability, leading to better user experience, and increased perceived ease of use of the service.
H6: Billing Issues (Q13) (Saleh & Saha, 2023; Mouli et al., 2024) [10, 11]		If billing issues are frequent or unresolved, then output quality as well as customer trust decreases, which in turn decreases the perceived ease of use of the telecom service.

(*continued*)

Table 1. (*continued*)

Features/Factors	Mapped TAM Construct(s)	Hypotheses
H7: Contract Type (Q11) (Chong et al., 2023; Zheng, 2025) [12, 13]	Objective Usability	If the contract type is more flexible (month-to-month), then objective usability is higher based on measurable features such as contract length and increases the perceived ease of use of the service.
H8: Customer Service Interactions (Q17) (Ribeiro et al., 2023) [8]	Computer Self-Efficiency	If there are frequent help requests, then Computer Self-Efficiency is low, and decreases the perceived ease of use of the telecom service.
H9: Service Types/ Bundles (Q16) (Ribeiro et al., 2024) [9]	Perceived Enjoyment	If the service types or bundles are more varied and customizable, then perceived enjoyment increases, which in turn increases the perceived ease of use of the telecom service.
H10: Promotions / Discounts (Q18) (Ribeiro et al., 2023) [8]		If promotions or discounts are offered and are perceived as suitable or relevant to the needs of the customer, then perceived enjoyment increases, and increases the perceived ease of use of the service as the service feels more accessible and rewarding to engage with.

3.2 Survey Design/Data Collection

The Malaysian survey dataset is developed based on the reviewed literature and TAM 1, 2, 3. The Google survey form is distributed via social media from May–June 2025. Hence, sampling is random. The researchers have obtained consent from the survey participants via the Google Form in line with Malaysia's Personal Data Protection Act (PDPA). The collected dataset consists of 225 instances.

3.3 Mining

Random Forest and XGBoost are applied to pre-process feature selection before feeding data to CHAID. In Fig. 1b's step 3.3, the CHAID algorithm [28, 29] serves as the final technique as CHAID is sensitive to irrelevant or noisy features. Moreover, CHAID's

decision rules in Step 3.4 can be used to design interventions. For example, if the rule from Node 6 (Fig. 2) flags new Fiber optics Internet customers whose Tenure is 1 month or less, as having a probability of churning at 88.17%, then the system can recommend customer retention strategies such as providing them with proactive welcome call programs, early discounts, or assign priority customer service support for their first month to reduce dissatisfaction. The telecommunication firm can also investigate whether service issues or price sensitivity is the true cause for customers churning early specifically for Fiber optic users.

Adapting from [28]'s research design (Fig. 1b), in the same domain, Fig. 3 shows this study's 3-stage research design. The differences are a) the dataset, i.e., [28]'s is in Croatia and ours, in Malaysia (Asia), b) the addition of RF and XGB prior to CHAID in our study.

```
/* Node 13 */ (IF tenure <= 1) AND (InternetService = "Fiber optic") AND (MultipleLines = "Yes" AND
MultipleLines = "No phone service") THEN Prediction = 'Yes'; Probability = 51.45%
/* Node 14 */ IF ((tenure <= 1) AND (InternetService = "Fiber optic") AND (MultipleLines = "Yes" OR MultipleLines
= "No phone service") THEN Prediction = 'Yes'; Probability = 78.00%
/* Node 6 */ IF (tenure <= 1) AND (InternetService = "Fiber optic") THEN Prediction = 'Yes'; Probability = 88.17%
/* Node 15 */ IF (tenure <= 5) AND (InternetService = "Fiber optic") AND (MonthlyCharges <= 55.25) THEN
Prediction = 'Yes'; Probability = 53.50%
/* Node 16 */ IF (tenure > 1 AND tenure <= 5) AND (InternetService= "Fiber optic") AND (MonthlyCharges >
55.25) THEN Prediction = 'No'; Probability = 71.93%
/* Node 8 */ IF (tenure > 1 AND tenure <= 5) AND (InternetService = "Fiber optic") THEN Prediction = 'Yes';
Probability = 64.89%
```

Fig. 2. Example of CHAID's rules from [28]

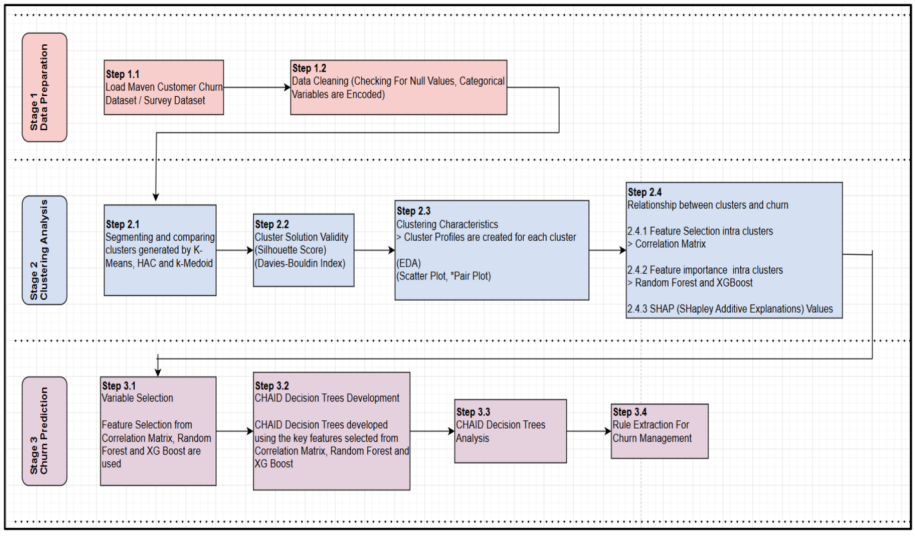


Fig. 3. This study's 3-stage research design adapted from [28]

4 Results

RQ 1: What will *k*-means, *k*-medoids and HAC reveal about the dataset's Malaysian customer segments?

Hypothesis 1: They will reveal slight fuzzy overlaps, due to almost common offerings among providers, differentiated by pricing, capacity/capability of support and provider's incentives.

4.1 Clustering Evaluation

Referring to Table 2, HAC's SS score performs better than *k*-means and *k*-medoids at $k = 2$. The most optimal DB index is also HAC's $k = 2$ at 2.0300. Hence, $k = 2$ has the best cluster quality. The *n* for Cluster 0 is 101, Cluster 1 114 and Cluster 2 10. Hence, only Clusters 0 and 1 are considered. Cluster 2's sample size is too small.

Table 2. SS and DB index of *k*-means, *k*-medoids and HAC across $k = 2$ to $k = 6$

k	SS	DB Index	*k*	SS	DB Index	*k*	SS	DB Index
	k-means			*k*-medoids			HAC	
2	0.0814	3.2377	**2**	**0.0735**	**3.5409**	**2**	**0.2574**	**2.0300**
3	**0.0863**	**2.6510**	3	0.0554	3.5292	3	0.0554	3.3164
4	0.0854	2.6955	4	0.0294	3.5390	4	0.0542	2.7643
5	0.0701	2.8252	5	0.0412	3.1541	5	0.0582	2.7112
6	0.0575	2.8294	6	0.0455	2.9229	6	0.0587	2.6686

4.2 Cluster Profiling

The clusters are profiled based on demographics, service usage, contracts and behaviour. This is to identify the unique aspects of each cluster.

4.2.1 C0 Cluster Profiling

From Table 3, C0's churn rate is 24%, and C1's likelihood of churning is 54%. A possible reason may be the plan type. 84% of C0 customers have family plans, whereas 75% of C1 customers are on individual plans. It is easier for the latter to switch whenever the need arises.

Family plans correlate with 78% of C0's more than 2 years Tenure, and 62% with 2 years Contract. These are binding commitments. Family plans also correlate with most (74%) of C0 customers using fibre optic Internet Type, 44% have Average Monthly GB of 50–100GB and 34% 100-200GB. However, only 51% of C0 customers have Multiple Lines. 68% of C0 customers use mobile data plans.

Due to the Average Monthly GB usage (Table 3), the monthly charges is Gaussian, mainly between RM 101–150 and RM 151–200 (Table 4). Moreover, they are likely to be

mostly well-educated, as most (91%) have no billing issues, and likely to be tech-savvy, as 93% do not need technical support.

Table 3. Key characteristics of C0, and C1 for the Malaysian survey dataset (n = 225)

Features	C0	C1	Plan Type	C0	C1	Contract	C0	C1
n	101	114	Individual	16%	75%	monthly	23%	70%
Predicted churn	24%	54%	Family	84%	25%	1-year	15%	6%
						2-year	62%	24%
Tenure	**C0**	**C1**	**Avg. Monthly Use**	**C0**	**C1**	**Internet type**	**C0**	**C1**
<6 months	0%	1%	< 50 GB	8%	40%	Cable	11%	7%
6–12 months	21%	41%	50–100 GB	44%	40%	DSL	9%	14%
1–2 years	1%	13%	100–200 GB	34%	12%	Fibre	74%	56%
>2 years	78%	45%	Over 200 GB	15%	7%	None	6%	23%

Table 4. Monthly Charge, Multiple Lines, Type of Line, Billing Issues and Tech Support for C0

Monthly Charge	C0	C1	Billing Issues	C0	C1	Tech. Support	C0	C1	Many Lines	C0	C1
< RM 50	2%	24%	No	91%	76%	No	93%	98%	No	49%	77%
RM 50–100	11%	39%	Once	8%	10%	Yes	7%	2%	Yes	51%	23%
RM 101–150	23%	23%	Many	1%	14%						
RM 151–200	22%	11%									
RM 201–250	13%	1%				**Mobile**	**C0**	**C1**	**Land**	**C0**	**C1**
RM 251–300	13%	0%				No	32%	19%	No	78%	93%
>RM 300	17%	2%				Yes	68%	81%	Yes	22%	7%

In terms of promotions, 60% have not received any, possibly because they are mostly satisfied customers (Table 5). Another 30% have received and are still receiving, indicating that some in this group appreciate the discounts more, possibly due to family income. Another 12% however, have not used the discounts received and have allowed the discounts to expire. This implies that possibly 72% of C0 are from M40 or T20.

For this M40 or T20 group, from a cross-selling/market basket analysis (simpler form of [30]'s perspective, the incentive may be in the form of cloud backup, as 97% in C0 have not utilized this and families would need this especially for school-going children's homework. Those in the 18–24 age group will usually use their campus' wifi. For those in the workforce, extending personal cloud backup would be useful. As they get into more senior positions, such personal cloud backup may remain attractive.

Another incentive for those aged 45–54 (23%), and 55–64 (23%), may be beauty/travel-related promotions for female customers (55%). Male customers (45%) may want sports/travel-related promos. Retirees (>65) may want cloud backup to store pictures from travels/gatherings.

Table 5. Possible bases for customer acquisition/retention

Age	C0	C1	Gender	C0	C1	Promo. discount received	C0	C1
18–24	5%	8%	Female	55%	63%	No	60%	54%
25–34	9%	32%	Male	45%	37%	Yes, but expired	12%	16%
35–44	13%	29%				Yes, its ongoing	30%	30%
45–54	23%	12%						
55–64	23%	11%				**Cloud backup**	**C0**	**C1**
> = 65	28%	7%				No	97%	96%
						Yes	3%	4%

4.2.2 C1 Cluster Profiling

For C1, 75% are on individual plan, with 41% having 6–12 months tenure and 45% on more than 2 years' tenure (Table 3). This may imply that the 41% group are renting premises on a yearly basis, whereas the 45% group are renting/staying in premises on a more permanent basis. Based on contract, 70% are on month-to-month, whereas 24% are on 2-year contract. Hence, possibly, 29% out of the 45% are renting their place of stay. More than half (56%) however, have fibre optic Internet Type. This indicates possible use for streaming sports/movies. These are seasonal/on demand. As such, the Average Monthly GB usage is less than 50GB (40%) and for 50–100 GB (40%).

Being on the move explains C1's 81% mobile usage and the lesser charges (less than RM50-RM150) for 86% of C1 (Table 4). Most (76%) compared to C0's 91%, have no Billing Issues, while 14% (compared to C0's 1%) have requested for Technical Support multiple times. Hence, some are not as well-educated or tech-savvy as C0.

In terms of promotions (Table 5), 16% of the promotional discounts have expired, possibly because they may have shifted. Hence, the promotion should enable them to utilize discounts regardless of where they have shifted to. Moreover, most (63%) of C1 are ladies. Hence, from a market basket analysis perspective, they may like discounts for shoes or clothes and groceries and for male customers (37%), travel/clothes.

4.3 Feature Importance Analysis Part I: Correlation (Covariance) Analysis and Principal Component Analysis (PCA)

RQ 2: For each cluster, which features are more significant based on correlation analyses with PCA loadings?

Hypothesis 2: Tenure and Contract Type will be important correlation analysis-PCA predictors.

This section presents the findings of the correlation analysis and Principal Component Analysis (PCA) to guide feature selection for predictive modelling in Random Forest and XGBoost. PC1 represents the feature and PC2 churn. The degree of variance which they represent, implies how much influence PC1 will have on PC2.

In the following Tables, green highlights represents negative correlations (not likely to churn), orange positive correlations (likely to churn). Higher PCA loadings imply higher variance, and thus higher significance.

4.3.1 Correlation (Based on Covariance Matrix) Analysis with PCA Loadings for Cluster 0 (Mostly Customers with Family).

Features which are negatively correlated with Churn (not likely to churn) in descending order (Table 6) are Tenure (strongly negatively correlated), while Average Monthly GB Usage, and Broadband (Internet) are weakly negatively correlated with Churn. Tenure and Number of Referrals are weakly negative, while Unlimited Data is weakly positive (customers may be swayed).

Table 6. Top 10 features' negative vs. positive correlations with Churn and PCA loadings for Cluster 0 (mostly older adults and families)

Feature	Corr.	PC1	PC2	delta	Feature	Corr.	PC1	PC2	delta
Tenure	**-0.885**	**0.100**	**-0.215**	**-0.315**	Broadband	0.239	0.251	0.012	-0.239
Average Monthly GB Usage	-0.277	0.356	0.012	-0.344	Gender	0.141	0.098	-0.040	-0.138
No. of Referrals	-0.116	0.082	-0.255	-0.337	Contract	0.074	0.216	-0.278	**-0.494**
Unlimited Data	-0.086	0.156	0.206	0.050	Monthly Charge	0.068	0.404	-0.201	**-0.605**
Mobile	-0.084	-0.028	0.162	0.190	Age	0.062	0.157	-0.293	**-0.450**

Features which do contribute to the likelihood of Churn are weakly correlated i.e., Broadband (Internet), followed by Gender, Contract, Monthly Charge and Age. However, Age, Contract, and Monthly Charge's PC2 loadings are weakly negative and higher. From Sect. 4.2's findings, bundled offers, e.g. providing discounts for Unlimited Data may increase the Number of Referrals, especially via social media.

4.3.2 Correlation (Based on Covariance Matrix) Analysis with PCA Loadings for Cluster 1 (C1)

The strongest correlation coefficient for C1 is Tenure (Table 7). Contract and Promotional Discounts are weakly and negatively correlated with Churn. The highest positive correlation is Multiple Lines, followed by Cloud Backup. Moreover, though Plan Type's correlation is positive and weakly correlated, it's PC-PC2 loading's delta is the highest, at -0.620. Hence, if the Plan Type fits their lifestyle needs, this can offset other features' influence on the likelihood to Churn. Acquisition/retention strategies can include discounts for multiple lines and cloud backup.

Table 7. Top 10 features most correlated (based on covariance) with Churn and their PCA loadings – Cluster 1 (mostly single and middle/lower income individuals)

Feature	Corr.	PC1	PC2	delta	Feature	Corr.	PC1	PC2	delta
Tenure	**-0.818**	**0.100**	**-0.215**	**-0.315**	Billing Issues	0.163	-0.057	0.202	0.259
Contract	-0.215	0.216	-0.278	**-0.494**	TV / Streaming	0.161	0.286	0.223	-0.063
Promotional Disc. Rec.	-0.178	0.071	0.045	-0.026	Broadband (Internet)	0.158	0.251	0.012	-0.239
Multiple Lines	0.216	0.205	-0.198	**-0.403**	Plan Type	0.155	0.285	-0.335	**-0.620**
Cloud Backup	0.176	0.285	0.379	0.094	Unlimited Data	0.147	0.156	0.206	0.050

4.4 Feature Importance Analysis Part II: Random Forest and XGBoost

RQ 3: Which features for each cluster are more significant based on RF, XGB?

Hypothesis 3: Tenure and Contract type will be significant across all RF, XGB clusters.

Table 8 presents the performance metrics of both models by cluster, followed by the top features contributing to churn prediction in each cluster for both datasets. For this dataset, RF and XGB exhibit similar performance scores for Cluster 0. This implies that the cluster is rather Gaussian and homogeneous. However, RF slightly outperforms XGB for cluster 1, especially for recall (0.106), F1 score (0.062), accuracy (0.059). This implies that this cluster is skewed, possibly due to 63% being female.

Table 8. Model performance metrics by cluster

Cluster 0	RF	XGB	Churn	Cluster 1	RF	XGB	Churn
Accuracy	0.935	0.935		Accuracy	0.912	0.853	
Precision	1.000	1.000		Precision	0.944	0.938	
Recall	0.750	0.750		Recall	0.895	0.789	
F1 Score	0.857	0.857	24%	F1 Score	0.919	0.857	54%

RF = Random Forest, XGB = XGBoost, CR = predicted Churn rate

4.5 Summary of Findings

All 10 hypotheses are accepted as these features are found among the top 16 predictors of churn based on the average rankings by RF and XGB across clusters. Cluster 0 (lowest churn) comprise of customers who prefer bundles of home services and Cluster 1 (highest churn) consists mainly of customers who rely on mobile services with light use of broadband, TV, landline, cloud backup services. From the above, a summarized profile of Clusters 0, 1 and 2 is presented in Table 9 below.

Table 9. Cluster-profiling for Cluster 0, Cluster 1

Features	Cluster 0 (n = 101)	Cluster 1 (n = 114)
Churn status	Low churn rate (CR 24%)	Higher churn rate (CR 54%)
Age	74% adults >35	Gaussian, mainly 25–44
Tenure	78% >2 years	6–12 months (41%), >2 years (45%)
Contract	62% 2-year contracts	70% monthly
Plan Type	84% use family plans	75% use individual plans
Internet Type	74% fibre optic	56% fibre optic, 23% data plan
Multiple Lines	49% single, 51% multiple lines	77% use single line
Avg. Monthly GB Usage	Gaussian 44% 50–100GB, 34% 100-200GB	Skewed left 40% 50–100GB, 40% 100-200GB
Payment Method	46% online banking, 29% credit card	41% online banking, 36% mobile apps
Monthly Charge	Gaussian. 45% RM 101–200	Skewed left. 86% (< RM50–150)
Billing Issues	91% no billing issues	76% no issues, 14% many issues
Promo. Disc. Received	60% no promotion, 30% on-going, 12% expired	54% no promotion, 30% ongoing, 16% expired
Cust. Service Contact Freq.	Contact customer service occasionally	Contact customer service occasionally

5 Comparative Tree-Based Findings: RF, XGB, CHAID

RQ 4: How comparable is CHAID's feature selection to RF/XGBoost?

Hypothesis 4: CHAID's correlation (covariance) complemented by p feature selection is comparable to RF/XGBoost's, for both homogeneous and heterogeneous datasets

RF utilizes majority voting across parallel decision trees. Though rankings differ between RF and XGB, Table 10 and 11 confirm that RF, and XGB concur with regards to each cluster's top 5 and that CHAID is more sensitive towards skewed clusters.

Table 10. RF-XGB ranking across clusters

	Cluster 0		Cluster 1	
Features	**RF Ranking**	**XGB Ranking**	**RF Ranking**	**XGB Ranking**
Tenure	1	1	1	1
Monthly Charge	6	5	5	2
Age	5	3	4	8
Payment Method	3	11	2	6
Number of Referrals	4	9	3	3
Multiple Lines	11	6	9	4
Avg Monthly GB Usage	2	10	7	10
Promotional Discounts Received	10	7	11	9
Contract	9	14	6	5
Gender	7	4	17	13
Broadband (Internet)	8	2	10	15
Plan Type	17	12	14	14
Unlimited Data	12	8	15	11
Billing Issues	14	13	13	17
Internet Type	16	16	8	7
Customer Service Contact Frequency	13	22	12	12
TV / Streaming	15	17	20	18
Mobile	18	15	16	16
Landline	19	18	18	21
Cloud Backup	22	20	19	20
Security Services	21	19	21	19
Technical Support	20	21	22	22

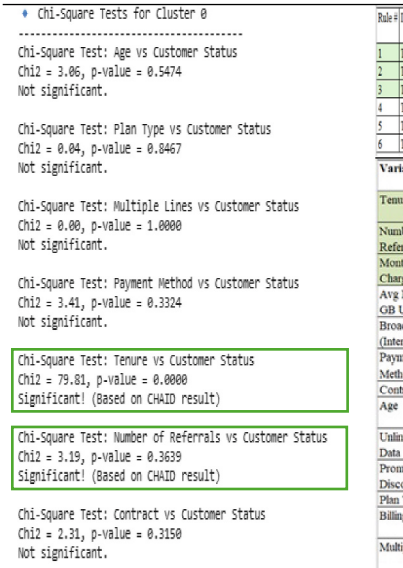
Table 11. XGBoost's top 5 and CHAID's significant vs. RF's top 5 across clusters

RF		**XGBoost vs. RF**		**CHAID vs. RF**	
C0	**C1**	**C0**	**C1**	**C0 (2/3 similar)**	**C1 (3/3 similar)**
Tenure	Tenure	Tenure	Tenure	Tenure	Tenure
Avg. Monthly GB Use	Payment Method	Broadband	Payment Method	No. of Referrals	Payment Method
Payment Method	No. of Referrals	Age	No. of Referrals	Monthly Charge	Monthly Charge
No. of Referrals	Age	Gender	Multiple lines		
Age	Monthly Charge	Monthly Charge	Contract		

Moreover, CHAID's *p*-value is moderated by chi square. For example, for C0, the *chi2* and *p*-value for Payment Method are 3.41 and 0.3324 respectively. CHAID considers it as not significant (Table 9 left). However, for C1, for Payment Method, the *chi2* is 0.96 and *p* value 0.8107 respectively. CHAID regards it as significant. Examples of CHAID-generated rules are presented in Table 12 right.

In summary, the quality of outcomes depends on the quality of the model. RF performs better for more Gaussian and homogeneous datasets, while XGBoost and CHAID are more suited for skewed, imbalanced clusters. Covariance-PCA loadings and its delta may be more stable. We are not certain as the data size is only 225. Further experiments are needed to confirm.

Table 12. Sample CHAID-identified significant features (left) and CHAID rules (right)



```
♦ Chi-Square Tests for Cluster 0
----------------------------------------
Chi-Square Test: Age vs Customer Status
Chi2 = 3.06, p-value = 0.5474
Not significant.

Chi-Square Test: Plan Type vs Customer Status
Chi2 = 0.04, p-value = 0.8467
Not significant.

Chi-Square Test: Multiple Lines vs Customer Status
Chi2 = 0.00, p-value = 1.0000
Not significant.

Chi-Square Test: Payment Method vs Customer Status
Chi2 = 3.41, p-value = 0.3324
Not significant.

Chi-Square Test: Tenure vs Customer Status
Chi2 = 79.81, p-value = 0.0000
Significant! (Based on CHAID result)

Chi-Square Test: Number of Referrals vs Customer Status
Chi2 = 3.19, p-value = 0.3639
Significant! (Based on CHAID result)

Chi-Square Test: Contract vs Customer Status
Chi2 = 2.31, p-value = 0.3150
Not significant.
```

Rule #	IF Conditions	THEN Churn Probability (%)
1	Tenure ∈ {'1–2 years', '6–12 months'} AND Number of Referrals = 0	100.00%
2	Tenure ∈ {'1–2 years', '6–12 months'} AND Number of Referrals = 1	100.00%
3	Tenure ∈ {'1–2 years', '6–12 months'} AND Number of Referrals = 2	50.00%
4	Tenure = 'More than 2 years' AND Monthly Charge ∈ {'Less than RM50', 'RM 251–300', 'RM 151–200', 'RM 101–150'}	0.00%
5	Tenure = 'More than 2 years' AND Monthly Charge = 'Over RM 300'	12.50%
6	Tenure = 'More than 2 years' AND Monthly Charge = 'RM 50–100'	0.00%

Variable	Global p-value 2 (full)	CHAID Local p-value	CHAID Split Used?	Comments
Tenure	0.0040	4.14E-19	Yes	Root split; highly significant both globally and locally
Number of Referrals	0.0043	0.0149	Yes	Second-level split under short-tenure group; locally significant
Monthly Charge	0.0052	0.0797	Yes (borderline)	Used in long-tenure branch; not significant locally ($p > 0.05$)
Avg Monthly GB Usage	0.0259	Not used	No	Globally modest p-value, but not selected by CHAID for any split
Broadband (Internet)	0.0336	Not used	No	Not selected despite moderate global p-value
Payment Method	0.3324	Not used	No	Not significant globally or in CHAID
Contract	0.3150	Not used	No	Not used in this tree
Age	0.5474	Not used	No	High global p-value, not chosen for splitting
Unlimited Data	0.5791	Not used	No	Not selected
Promotional Discounts	0.4652	Not used	No	Not used despite being a model variable
Plan Type	0.8467	Not used	No	Not significant and not used
Billing Issues	0.8523	Not used	No	Highest global p-value; completely excluded
Multiple Lines	1.0000	Not used	No	Not significant and not split

6 Application to STEM Education

RQ5: Can PCA loadings' delta be analogized for STEM education?

From decades of research in quantum information science, [31] points out that there are three primary branches to quantum information science, i.e., quantum computing, quantum communication, and quantum sensing and measurement [32].'s study on quantum correlations and quantum key distribution highlight this as well. This study has proposed covariance-PCA loadings' deltas as identifier/parity bit in metaverses). There may be very humble analogies [33] to quantum states (polarity), robustness (feature selection) and parity bits (covariance-PCA loadings' delta's consistency across clusters). This would extend the role of PCA in [34], covariance in [35].

7 Contributions and Conclusion

For interconnected systems, data size is huge and require lean, accurate discovery of feature selections and adaptivity to dynamic contexts and changes. We have applied *k*-means, *k*-medoids and HAC to determine the optimal number of clusters. Subsequently, correlation (covariance) analysis with PCA loadings are applied to determine significant features. Findings reveal the strengths of RF, XGBoost, CHAID.

Our very humble contributions are:

a) confirmed [6]'s findings on a large dataset that RF exhibits better performance compared to XGBoost for Gaussian datasets,
b) confirmed CHAID's chi square (covariance)-*p* viable applications to finance, similar to [29], but in the future for Buy Now Pay Later explainability,
c) proposed simplifying very complex concepts for STEM education via [33]'s analogical thinking (e.g. RQ5/Sect. 6). However, the scope-validation measures need to be identified according to progressive scientific inquiry.

In conclusion, Information Science grounds the OSI layers. This paper has built on Information Science but at the application layer. Evolutionary learning are promising. These are however, beyond our capability/ resources.

Acknowledgement. This study was initially Sharini's Masters in Data Science capstone project from March-June 2025, supervised by Lee. Lee prepared the paper, added and structured analyses/ interpretations for RQ5, contributions b), c), inspired by Prof. GL Long's CTIS 2025 quantum computing keynote, after Sharini graduated June 2025. We have obtained informed consent and there is no conflict of interest. Lee thanks Prof. Hean-Teik Chuah, Prof. Hong-Tat Ewe, Prof. Yashwant Prasad Singh for past STEM modeling, OSI, and AI foundations (in Multimedia University), Prof. Ashok K. Goel (one of Lee's past Fulbright Visiting Scholar mentors) on analogy, design, creativity, and Prof. GL Long for not rejecting business model-quantum computing concept similarities.

References

1. CEO Insights Asia Team: Malaysia targets High-Tech status by 2030, aims to lead in ICT. CEO Insights Asia (2024)
2. Mordor Intelligence: Telecommunication Industry in Malaysia Size & Share Analysis - Growth Trends & Forecasts (2025–2030). (2024). https://www.mordor intelligence.com
3. Poudel, S.S., Pokharel, S., Timilsina, M.: Explaining customer churn prediction in telecom industry using tabular machine learning models. Mach. Learn. Appl. **17**, 100567 (2024). https://doi.org/10.1016/j.mlwa.2024.100567
4. Maw, M., Haw, S.C., Ho, C.K.: Customer churn prediction in telecommunication: an analysis on issues, techniques and future trends. In: Conference on Sustainable Utilization and Development in Engineering and Technologies, pp. 148–153. IEEE Press, Malaysia (2019). https://doi.org/10.1109/CSUDET47057.2019.9214592
5. Wagh, S.K., Andhale, A.A., Wagh, K.S., Pansare, J.R., Ambadekar, S.P., Gawande, S.: Customer churn prediction in telecom sector using machine learning techniques. Results Control Optim. **14**, 100342 (2023). https://doi.org/10.1016/j.rico.2023.100342
6. Chang, V., Hall, K., Xu, Q., Amao, F., Ganatra, M., Benson, V.: Prediction of customer churn behavior in the telecommunication industry using machine learning models. Algorithms **17**(6), 231 (2024). https://doi.org/10.3390/a17060231
7. Fatima, G., Khan, S., Aadil, F., Kim, D.H., Atteia, G., Alabdulhafith, M.: An autonomous mixed data oversampling method for AIOT-based churn recognition and personalized recommendations using behavioral segmentation. PeerJ Comput. Sci. **9**, e1756 (2024). https://doi.org/10.7717/peerj-cs.1756
8. Ribeiro, H., Barbosa, B., Moreira, A.C., Rodrigues, R.G.: Determinants of churn in telecommunication services: a systematic literature review. Manag. Rev. Q. **74**(3), 1327–1364 (2023). https://doi.org/10.1007/s11301-023-00335-7

9. Ribeiro, H., Barbosa, B., Moreira, A.C., Rodrigues, R.: Customer experience, loyalty, and churn in bundled telecommunications services. SAGE Open **14**(2) (2024). https://doi.org/10.1177/21582440241245191
10. Saleh, S., Saha, S.: Customer retention and churn prediction in the telecommunication industry: a case study on a Danish university. SN Appl. Sci. **5**, 173 (2023). https://doi.org/10.1007/s42452-023-05389-6
11. Mouli, K.C., et al.: An analysis on classification models for customer churn prediction. Cogent Eng. **11**(1) (2024). https://doi.org/10.1080/23311916.2024.2378877
12. Chong, A.Y.W., Khaw, K.W., Yeong, W.C., Chuah, W.X.: Customer churn prediction of telecom company using machine learning algorithms. J. Soft Comput. Data Min. **4**(2), 1–22 (2023). https://doi.org/10.30880/jscdm.2023.04.02.001
13. Zheng, P.: The impact of contract type and monthly charges on customer churn: comparative analysis across age groups. Adv. Econ. Manag. Polit. Sci. **153**(1), 22–27 (2025). https://doi.org/10.54254/2754-1169/2024.19469
14. Kaplan, R.S., Norton, D.P.: The balanced scorecard-measures that drive performance. Harv. Bus. Rev. (1992)
15. Davis, F.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. **13**(3), 319–340 (1989). https://doi.org/10.2307/249008
16. Venkatesh, V., Davis, F.D.: A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag. Sci. **46**, 186–204 (2000)
17. Venkatesh, V., Bala, H.: Technology acceptance model 3 and a research agenda on interventions. Decis. Sci. **39**(2), 273–315 (2008)
18. Vieri, J.K., Munandar, T.A., Srisulistiowati, D.B.: Exclusive clustering technique for customer segmentation in national telecommunications companies. Int. J. Inf. Technol. Comput. Sci. Appl. **1**(1), 51–57 (2023)
19. Bilal, S.F., Almazroi, A.A., Bashir, S., Khan, F.H., Almazroi, A.A.: An ensemble-based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry. PeerJ Comput. Sci. **8**, e854 (2022)
20. Liu, R., et al.: An intelligent hybrid scheme for customer churn prediction integrating clustering and classification algorithms. Appl. Sci. **12**(18), 9355 (2022). https://doi.org/10.3390/app12189355
21. Bagam, N.N.: machine learning models for customer segmentation in telecom. J. Sustain. Solutions **1, 4**, 101–115 (2024). https://doi.org/10.36676/j.sust.sol.v1.i4.42
22. Cohen, J., Cohen, P., West, S.G., Aiken, L.S.: Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Routledge (2013)
23. Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on typical tabular data? Adv. Neural Inf. Proces. Syst. **35**, 507–520 (2022). https://doi.org/10.48550/arXiv.2207.08815
24. Borisov, V., Leemann, T., Sebler, K., Haug, J., Pawelczyk, M., Kasneci, G.: Deep neural networks and tabular data: a survey. IEEE Trans. Neural Networks Learn. Syst. **35**(6), 7499–7519 (2022). https://doi.org/10.1109/tnnls.2022.3229161
25. Rabbani, S.B., Medri, I.V., Samad, M.D.: Attention versus contrastive learning of tabular data: a data-centric benchmarking. Int. J. Data Sci. Anal. **20**, 3069–3091 (2025). https://doi.org/10.48550/arXiv.2401.04266
26. Barsotti, A., et al.: A decade of churn prediction techniques in the TelCo domain: a survey. SN Comput. Sci. **5**, 404 (2024). https://doi.org/10.1007/s42979-024-02722-726
27. Kass, G.V.: An exploratory technique for investigating large quantities of categorical data. Appl. Stat. **29**(2), 119–127 (1980). https://doi.org/10.2307/2986296
28. Bach, M.P., Pivar, J., Jaković, B.: Churn Management in Telecommunications: hybrid approach using cluster analysis and decision trees. J. Risk Financ. Manag. **14**(11), 544 (2021). https://doi.org/10.3390/jrfm14110544

29. Yang, Y., Yi, F., Deng, C., Sun, G.: Performance analysis of the CHAID algorithm for accuracy. Mathematics **11**(11), 2558 (2023). https://doi.org/10.3390/math11112558
30. Rudas, I.J., Pap, E., Fodor, J.: Information aggregation in intelligent systems: an application-oriented approach. Knowl. - Based Syst. **38**, 3–13 (2013). https://doi.org/10.1016/j.knosys.2012.07.02
31. Long, G.L.: Progress in quantum information. National Sci. Rev. **12**, nwaf289 (2025). https://doi.org/10.1093/nsr/nwaf289
32. Franz, T.: Quantum Correlations and Quantum Key Distribution. Universität Hannover, Von der Fakultät fürMathematik und Physik der Gottfried Wilhelm Leibniz (2013)
33. Goel, A.: K.: design, analogy and creativity. IEEE Expert **12, 3**, 62–70 (1997). https://doi.org/10.1109/64.590078
34. Lee, C.S., Singh, Y.P.: Student modelling using principal component analysis of SOM clusters. In: IEEE International Conference on Advanced Learning Technologies, pp. 480–484, Joensuu, Finland (2004). https://doi.org/10.1109/ICALT.2004.1357461
35. Lim, A.H.L., Lee, C.S.: Integrated model-driven business evaluation methodology for strategic planning. Int. J. Bus. Inf. Syst. **3**(4), 333–355 (2008). https://doi.org/10.1504/IJBIS.2008.018037