Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
MASTERS IN DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK
INTERNSHIP REPORT

IMPROVING FIXED LINE CUSTOMER SEGMENTATION WITH FACTOR ANALYSIS

LUKE GORMAN

MARCH - 2024

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
MASTERS IN DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK
INTERNSHIP REPORT

IMPROVING FIXED LINE CUSTOMER SEGMENTATION WITH FACTOR ANALYSIS

LUKE GORMAN

SUPERVISION:
JOAO ALFONSO BASTOS
STEPHEN SMITH

MARCH - 2024

This page is intentionally left blank.

GLOSSARY

ACS – Advanced Consulting Services

ISPs – Internet Service Providers

FTTH – Fiber to The Home

FWA – Fixed Wireless Access

STC – Saudi Telecom

CEM – Customer Experience Management

CRM – Customer Relationship Management

NPS – Net Promoter Score

CLTV – Customer Life-Time Value

MMO – Massive Multiplayer Online

PCA – Principal Component Analysis

KPI – Key Performance Indicator

CEI – Customer Experience Index

QoS – Quality of Service

ARPU – Average Revenue Per User

OPEX – Operating Expenditure

ABSTRACT

Customer segmentation is vital for companies to cluster their customer base into unique groups that allow for a deeper understanding of such a base. This can provide the companies with ample opportunity to grow revenue and increase customer satisfaction. Nokia's Advanced Consulting Services (ACS) Team is utilising Factor Analysis to build unique 'Digital Personas' for Internet Service Providers (ISPs) with the aims to increase revenue, reduce churn and improve customer experience among other KPIs. This paper highlights the methodology of this clustering, results from a feasibility study and shows how Factor Analysis can be used to segment customer bases and provide recommendations and business intelligence for companies in the telecommunications industry. The methods below produced ten unique digital personas based on data collected from Saudi Telecom (STC). Recommendations such as Fibre to The Home (FTTH) were presented to the client as well as dashboards for their consumption.

KEYWORDS: Customer Segmentation, Factor Analysis, Business Intelligence, Consulting Services.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1. INTRODUCTION

Originally founded in Finland, Nokia (NOK) entered the telecommunications industry in the latter parts of the 20th century. In 1987 the company introduced its first handheld mobile phone and by the early 2000s dominated the mobile phone market. The rise of the smartphone and Nokia's inability to adapt to this shift in technology lead to their market share in the mobile phone space being captured by Apple and Samsung. In 2014, Nokia sold its mobile phone business to Microsoft while retaining its network infrastructure, mapping, and technology patents. This would shift Nokia's priority to the development of 5G technology, network infrastructure and more recently, cloud and network services. This has allowed Nokia to form major partnerships with telecommunications companies worldwide.

The telecoms industry is vast, and Nokia has amassed a wealth of industry experience and domain knowledge which allows its Advanced Consulting Services (ACS) team to offer a full suite of telecommunications consulting services. Many branches have sprouted from this arm of Nokia and include Data Science and Analytics Consulting, as well as Cybersecurity Consulting and Cloud Transformation Consulting to name but a few. With over 15 years consulting experience and over 200 field-proven use cases this places Nokia's ACS teams in prime position to leverage advanced analytics to solve customer problems.

The telecommunications industry is composed of telecommunications companies and internet service providers that make communication and connectivity to the internet seamless across the globe. The global telecommunications market was valued at 1.8T USD in 2022 [1]. Internet service providers (ISPs) are the interface between the consumers and access to technologies such as WiFi, 5G and internet. These companies offer a variety of services to customers whether it be fixed line services to homes or mobile data (5G) to customers mobile phones. Across the globe there has been a significant increase in home traffic on consumers' fixed wireless access (FWA). This has been a downstream effect of the 2019 Covid pandemic. A sharp rise in remote working along with lockdowns saw a record increase in customers demand for greater fixed line capabilities.

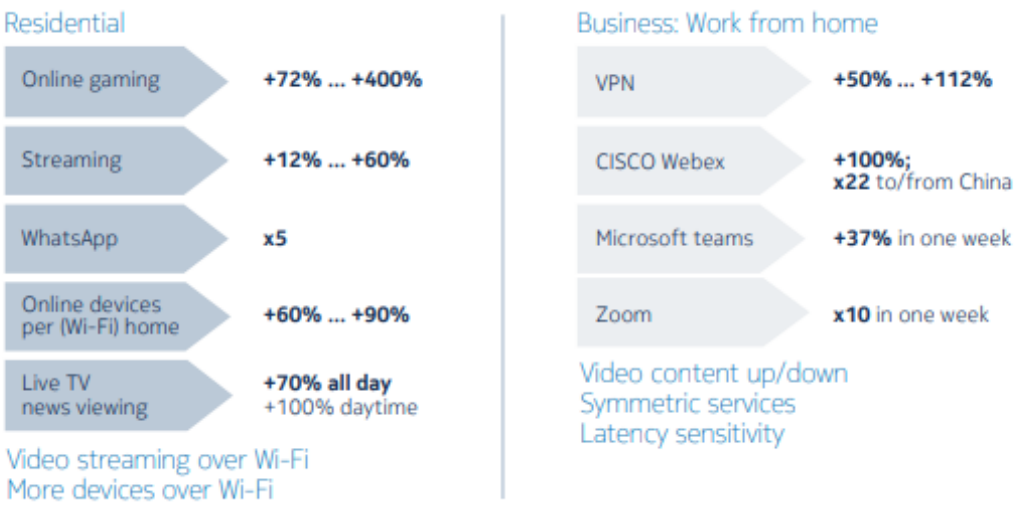

Figure 1 - Increases in digital behaviour activities during Covid-19.

As of 2024 the pandemic has eased and lockdowns have ceased, but demand for remote and flexible work remains. With competition fierce in this industry many internet service providers must adapt and offer better service to their customers with more personalised offerings. A by-product of this fierce competition in telecoms is the amount of churn telco providers may experience from their customer base. Churn in the telecoms industry simply refers to the decision of a customer to opt out of your offering and choose a competitor which naturally will have a negative effect on profits [2]. Customer segmentation offers many benefits to a telecommunications company including personalised offerings, increased customer satisfaction and reduced churn.

1.1 Nokia's ACS

We will now discuss the use case conducted by Nokia's Advanced Consulting Services (ACS) to build 'digital personas' as a means of customer segmentation based on fixed line usage. This use case was initially trialled with Saudi Telecom (STC) who are the largest telecom company in the middle east. Initial results and recommendations made were appreciated by the various STC teams that consumed this business

intelligence. Consequently, Nokia is continuing the development of this use case and ACS is building a service it can offer to other clients.

The service provider wanted to enrich marketing segmentation for customers on fixed line. The segments for fixed customers would be used to improve business results for both marketing and customer experience teams and grow revenues by identifying upsell opportunities. The Nokia Cloud and Network Services Advanced Consulting Services team proposes a solution for ISPs to better segment their customers in the hope that the providers can grow revenue and increase customer satisfaction. By using customer experience management (CEM) data, customer relationship management (CRM) data, service usage and technology usage, the ACS team has successfully identified unique digital personas (customer segments) and created recommendations for ISPs to achieve their financial targets as well as improve on customer related KPIs such as improving customer lifetime value (CLTV), increase net promoter score (NPS) and reduce churn.

Key requirements:

Four key requirements were identified with the customer for the proposed digital personas to be of value:

1. Segmentation must cover 100% of the customer base.
2. Each segment must have a clearly defined profile resulting in what would be called their ‘Digital Persona’.
3. Recommendations must be provided for each persona to address their customer experience and how marketing should approach them.
4. Clearly demonstrate how recommendations can lead to business value.

1.2 Factor Analysis

Factor analysis is a statistical method used to identify commonalities and interpret the underlying factors explaining covariance between variables [3]. The assumption here is that the observed variables are influenced by common factors that are not present in the dataset but can be inferred by latent factors. This makes factor analysis a suitable technique for when there are plausible reasons to believe there are unobserved latent factors. This is true especially in customer segmentation when there are always unobserved social and psychological forces driving consumer behaviour. In our case, users may display elements of correlation between fixed usage. For example, massive multiplayer online (MMO) gaming and online education. Previous studies in 2004 used latent factor analysis to discover a set of latent factors $C = \{c_1, c_2, \dots, c_k\}$ to 'explain' underlying relationships of user's web usage patterns [5]. Moreover, a recent study utilized Factor Analysis to build a regression model to predict churn in telecommunications [2]. For these reasons we have decided to use the factor_analyzer algorithm to segment customers in unique digital personas.

The factor analysis model is expressed as following, if you have p variables $X_1, X_2 \dots . X_p$ measured on n customers, then variable i can be written as a linear combination of m uncorrelated factors $F_1, F_2, \dots . F_m$ where $m < p$.

$$X_1 = ai_1F_1 + ai_1F_2 + ai_1F_m + ei$$

$$X_2 = ai_2F_1 + ai_2F_2 + ai_2F_m + ei$$

……

$$X_p = ai_sF_1 + ai_sF_2 + ai_sF_m + ei$$

Where ai_s are the factor loadings for variable i and ei is the error term or part of variable Xi that cannot be explained by the factors [4]

Factor loadings can be viewed as the strength of the relationship between the latent variables (unobserved variables) and the variables in the dataset (observed variables). They are between $0 < ai < 1$ with a higher loading representing a greater correlation or relationship between the observed variable and the factor. Examples of factor loadings can be seen in figure 2. Each outer box represents an observed variable in our dataset while the inner circles represent a latent factor. Calculating the factor loadings can be achieved in a number of ways, it typically involves a method to obtain initial loadings followed by a rotation step. There are two common methods for obtaining initial loadings. The first is the principal component method, that uses the same method that is used to carry out PCA [4]. Principle axis factoring is another method which aims to find the lowest number of factors which can capture all the variability in the original variables associated with the identified factors. Both methods tend to give similar results when the correlation between variables are high, and the number of variables is also high [5].

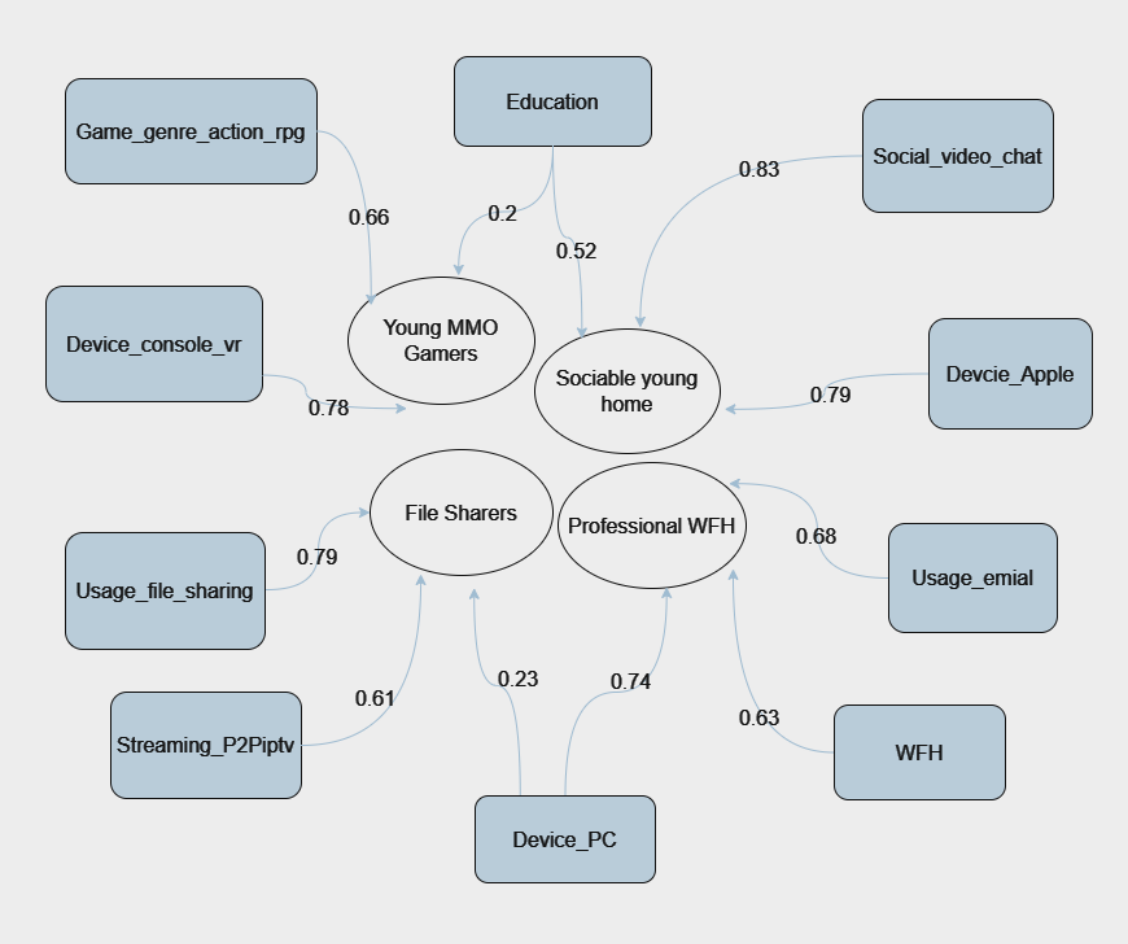

Figure 2 - Sample diagram to visualise the factor loadings.

Factor rotation involves transforming the provisional factors to find new factors that are easier to interpret. If there are groups of variables ('clusters') - i.e. subgroups that are strongly interrelated - then the rotation is done to try and make such variables in the subgroup score as highly as possible for one factor while ensuring the loadings are as low as possible on other factors. Put plainly, the objective of rotation is to ensure all variables have a high loading for only one factor. The rotation method that we use below is the varimax method. Factor rotation can be either orthogonal or oblique. Orthogonal rotation, like varimax, rotates such that the new factors are uncorrelated. After rotation it is desirable that the factor loadings are close to zero or very different from zero [5]. A near zero ai_s means that X_i is not strongly related to F_j.

The last step in factor analysis involves the calculation of the factor scores. These are the values of the rotated factors $F_1 *, F_2 *, F_3 * \ldots\ldots F_m *$ for each n individual for which the data is gathered. The number of factors (m) is selected by the analyst. This can be done upon a request from a client of the use case or with the help of domain experts. It can however be determined independent of human intervention when the principal component method is used. With p variables, there will be the same number of principal components. One may choose m such that the same number of p components account for a particular percentage of the variability (75%). A second method includes plotting the eigenvalues against the number of factors in a scree plot and using an elbow method approach.

In the next section we will describe the methodology for collecting, processing, and clustering the data with the factor analyser algorithm. We will first talk briefly about the deployment of the use case in a virtual environment. The discussion will follow after we have shown the results of the clustering algorithm. It is here where we talk about the possible recommendations and insights that's clients may avail of with this use case.

1.3 Previous Work

Unsupervised clustering algorithms have been used in customer segmentation with classic algorithms such k-means and DBSCAN used as an effective way to segment data in and reveal hidden structure in data. In the telecommunications industry these algorithms can be used to segment customers [9], anomaly detection [8], churn prediction [2] and study customer loyalty in networks [10] .

In the customer loyalty case, factor analysis was used to study service quality as this has a direct correlation with customer loyalty. Surveys were conducted to gather data measuring service quality KPIs such as delivery time and accuracy of billing. From 16 variables associated service quality the team found that service quality can be classified under the four main factors of competence, assurance, tangibility, and reliability [10].

For churn predictions one study used factor analysis in mobile networks to extract factors from different customer tags such as consumption and billing to then feed a regression model for churn prediction [2].

Neither example of factor analysis use above used latent factor analysis for segmenting customer based on their fixed line usage as we have described below. While one study used PCA combined with K-means for telecom customer segmentation [9] this data is not comparable to our fixed line usage analysing customer behaviour based on their internet usage. Therefore we see this application of factor analysis in the telecoms industry as a novel approach to segment customers based on their internet usage behaviour and patterns.

2. METHODOLOGY

2.1 Deployment

The digital persona use case that utilised factor analysis was built first in house and tested on a Linux based virtual lab (vLab). Data was ingested from vendors in the form of CEM, CRM and usage and loaded onto a jupyter notebook launched from the vLab.

From here we performed the clustering using the factor_analyzer algorithm. Results from the clustering were pushed to mariaDB and stored in a results table. The tables were then connected in the vLab to Tableau for subsequent visualisation and dashboard to be consumed by the users. This also allows for scheduled running with cron jobs that update dashboards monthly.

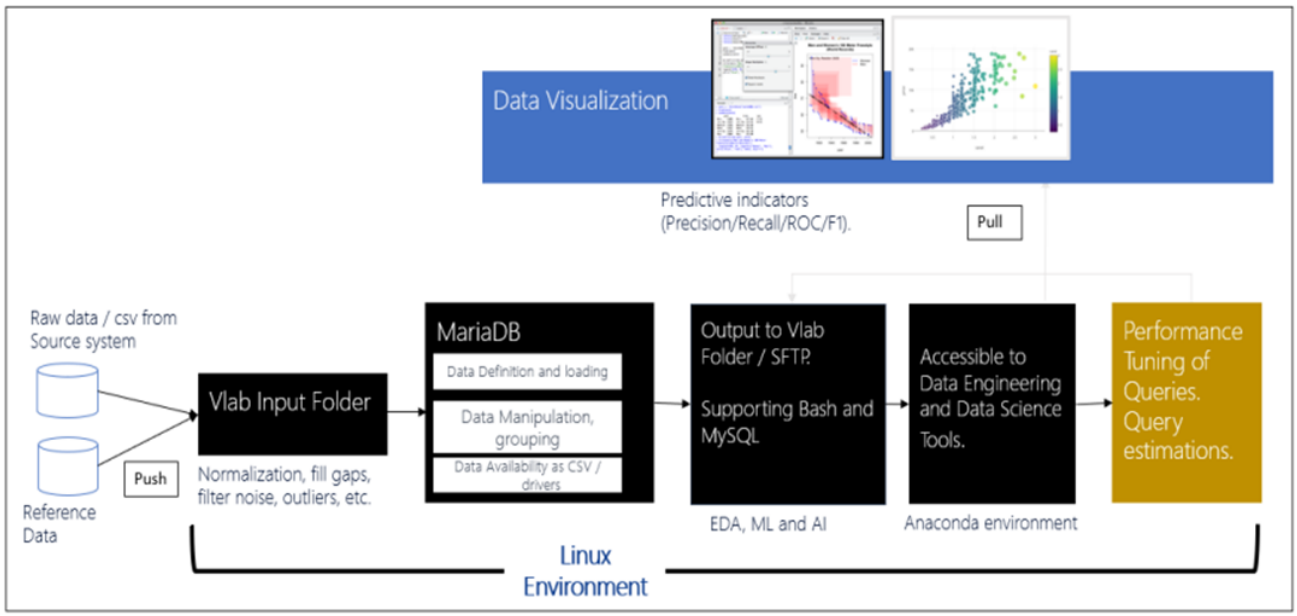

Figure 3 - vLab architecture.

This architecture offers scope for developing the use case on private cloud and public cloud builds, which may be advantageous depending on the customers needs and requirements. Example of public clouds that may support this use case include Azure and Google Cloud Platform, for which Nokia has entered strategic partnerships with.

2.2 Data selection:

Internet usage data was first manually tagged and categorised before loading into the jupyter notebook. A list of 2091 application names was identified as the most used

across the whole subscriber base. These included websites and apps such as 8BallPool, Youtube, CartoonHD, Cricketgateway etc. From this list, websites were grouped by similarity into 34 categories. For example, Udemy, Coursea, NoonAcademy and ClassDojo all comprise the educational category. Traffic volume was aggregated over 90 days and a final dataset called ML_POC_APP_VOL was created. This dataset contained the aggregated volume of each subscriber across all 34 categories of internet usage. A second dataset was utilised which was mainly categorical of both CRM and CEM data. This data was not encoded for clustering but rather for enrichment of insights and analysis to be made from the business intelligence gained from customer segmentation.

2.3 Processing:

Usage data was loaded into jupyter along with the categorical data and bitrate data. With the usage data a total volume and latency were also included. As the data was aggregated over 90 days a new column called volume was created to show the average volume of the subscriber over the 90 days. This column helped to remove outliers as 'data_abusers' > 260 (260Gb per day) and 'data_inactive' volume < .1 (100Mb per day) were dropped from the dataset. Removal of these outliers left us with a final dataset of 1.3m customers with their total volume for each internet usage category, total volume, latency, and average daily volume.

2.4 Preprocessing:

Preprocessing is a necessary step for scaling down the variables and reducing the compute for our clustering. For this we normalised each value in the dataset by dividing it by the sum of internet traffic volume for each user. This significantly scaled our data down. A final transformation step was added by raising the data to the power of 0.1 to scale our data which helps stabilise variance.

2.5 Model selection:

As mentioned above the primary objective of the use case was to segment users into unique digital personas. To decide the number of factors and (the number of personas to segment our customer base into) we first used PCA as a common factor extraction method similar to work done by Bai et al. [5]. As mentioned above this is only an initial

step in choosing m factors. Further analysis was done by calculating the eigenvalues. A scree plot was then built to further decide on the number of factors. As seen in Figure 4, This plots the number of factors against the eigenvalues.

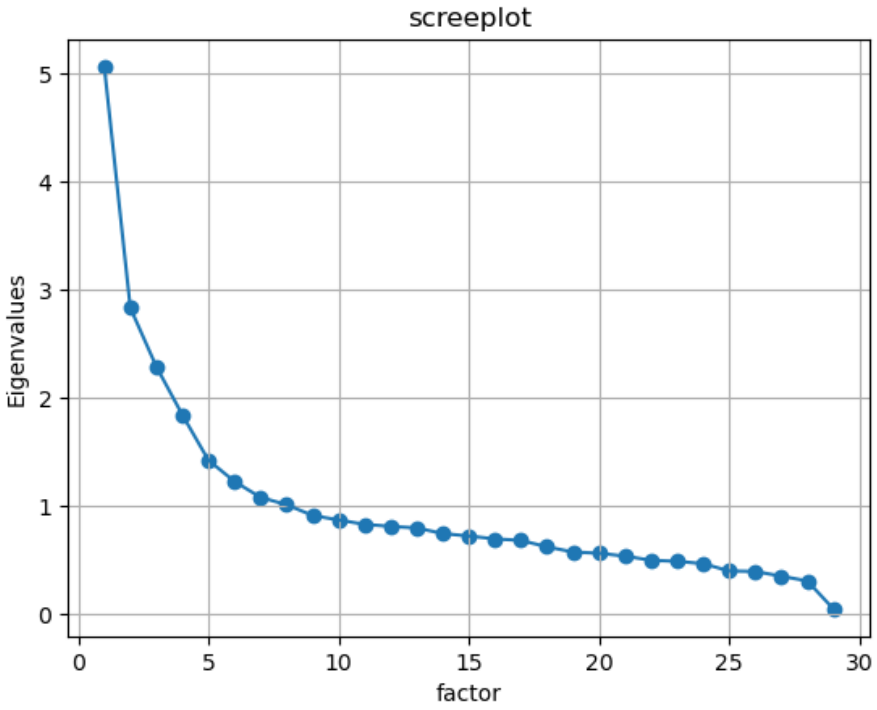

Figure 4 - Scree plot.

3. RESULTS

To measure the factorability of our data we used the Kaiser-Meyer-Olkin (KMO) test. KMO is a test conducted to examine the strength of the partial correlation (how the factors explain each other) between the variables. KMO values closer to 1 are ideal where in general KMO values <0.6 are considered inadequate [6][7]. Our data showed a KMO value of 0.78. Once this was decided the next step was decide on the number of factors for our analysis.

The initial PCA method provided us with insight that ten components explain 74% of the variance in our data and this method aligns with work done by Zhang et al. in their telecoms churn analysis [2]. Combined with the results from the scree plot we chose ten latent factors as our number of factors. After fitting the algorithm to the transformed data, we obtained the factor loadings. This matrix of loadings was converted to a dataframe as seen below in table 1. In the below table we can see the column field is populated with values 0-9 which represent the 10 latent factors in our model. The row values (0-28) contain the internet usage categories.

TABLE I. FACTOR LOADINGS AFTER ROTATION

	0	1	2	3	4	5	6	7	8	9
0	-0.07	0.14	0.28	-0.02	0.66	0.07	-0.05	-0.07	0.01	-0.13
1	0.27	-0.06	-0.05	-0.03	0.34	-0.14	0.09	-0.03	0.57	0.02
2	0.06	0.79	-0.12	0.04	0.08	0.01	0.15	-0.07	-0.17	0.08
3	0.22	-0.19	0.26	0.10	-0.09	0.11	0.04	0.00	0.69	-0.04
4	-0.15	-0.09	-0.20	-0.06	0.78	-0.12	-0.02	0.06	-0.04	0.15
5	-0.08	0.08	0.03	0.02	0.14	0.74	0.07	0.23	0.01	0.09
6	0.29	0.04	0.36	-0.01	-0.02	-0.03	0.03	-0.09	0.63	-0.11
7	0.03	0.04	-0.04	0.97	-0.03	0.02	0.15	0.02	0.03	0.03
8	0.03	0.12	-0.01	0.04	-0.02	0.14	0.00	0.17	0.47	0.41
9	0.57	0.07	-0.12	0.04	0.00	0.06	0.11	0.01	0.14	0.13

10	-0.17	0.83	-0.03	0.05	-0.03	-0.01	0.06	-0.01	0.08	0.04
11	0.13	-0.14	0.77	-0.11	-0.17	-0.01	0.03	-0.02	0.17	0.07
12	0.37	0.02	0.34	-0.03	0.10	0.36	-0.04	-0.12	0.32	-0.28
13	0.15	0.47	0.10	0.01	-0.17	0.06	0.08	0.12	0.47	0.14
14	0.19	0.29	-0.21	-0.01	0.34	0.28	0.05	0.18	0.00	-0.31
15	0.29	-0.02	-0.16	0.04	-0.12	0.63	0.10	-0.04	-0.03	0.27
16	-0.29	0.52	0.16	0.01	0.20	0.20	-0.08	0.14	0.29	-0.04
17	0.22	0.02	0.01	0.00	-0.15	0.68	0.12	0.07	0.28	-0.02
18	0.31	0.18	0.05	0.07	0.06	0.28	0.09	0.00	0.05	0.63
19	0.24	-0.04	0.00	0.08	0.05	0.16	0.82	0.05	0.04	0.13
20	-0.05	0.30	0.01	0.06	-0.11	0.08	0.78	0.11	0.14	-0.06
21	0.06	0.06	-0.04	0.98	-0.04	0.02	-0.02	0.01	0.03	0.04
22	-0.01	-0.06	-0.80	-0.02	-0.15	0.06	0.02	-0.11	-0.18	-0.01
23	0.28	-0.03	0.08	0.03	-0.05	0.03	0.06	0.61	0.11	0.21
24	0.74	-0.03	0.07	0.01	-0.02	0.06	0.00	0.11	0.09	-0.03
25	0.74	-0.18	0.13	0.01	-0.08	0.08	0.03	0.10	0.19	-0.01
26	0.07	0.04	-0.02	0.01	0.05	0.18	0.08	0.79	0.03	-0.11
27	0.07	0.18	0.10	0.01	-0.03	0.23	0.10	0.19	0.58	0.09
28	0.61	-0.08	0.16	0.05	-0.12	0.14	0.07	0.16	0.19	0.18

The table above has been explained below in figure 5 where we have shown the factor loadings of each internet usage category as they relate to each latent factor after rotation on a heatmap. On the bottom of the graph is the cluster groups labelled 0-9. These clusters were then labelled in the preceding step using domain knowledge.

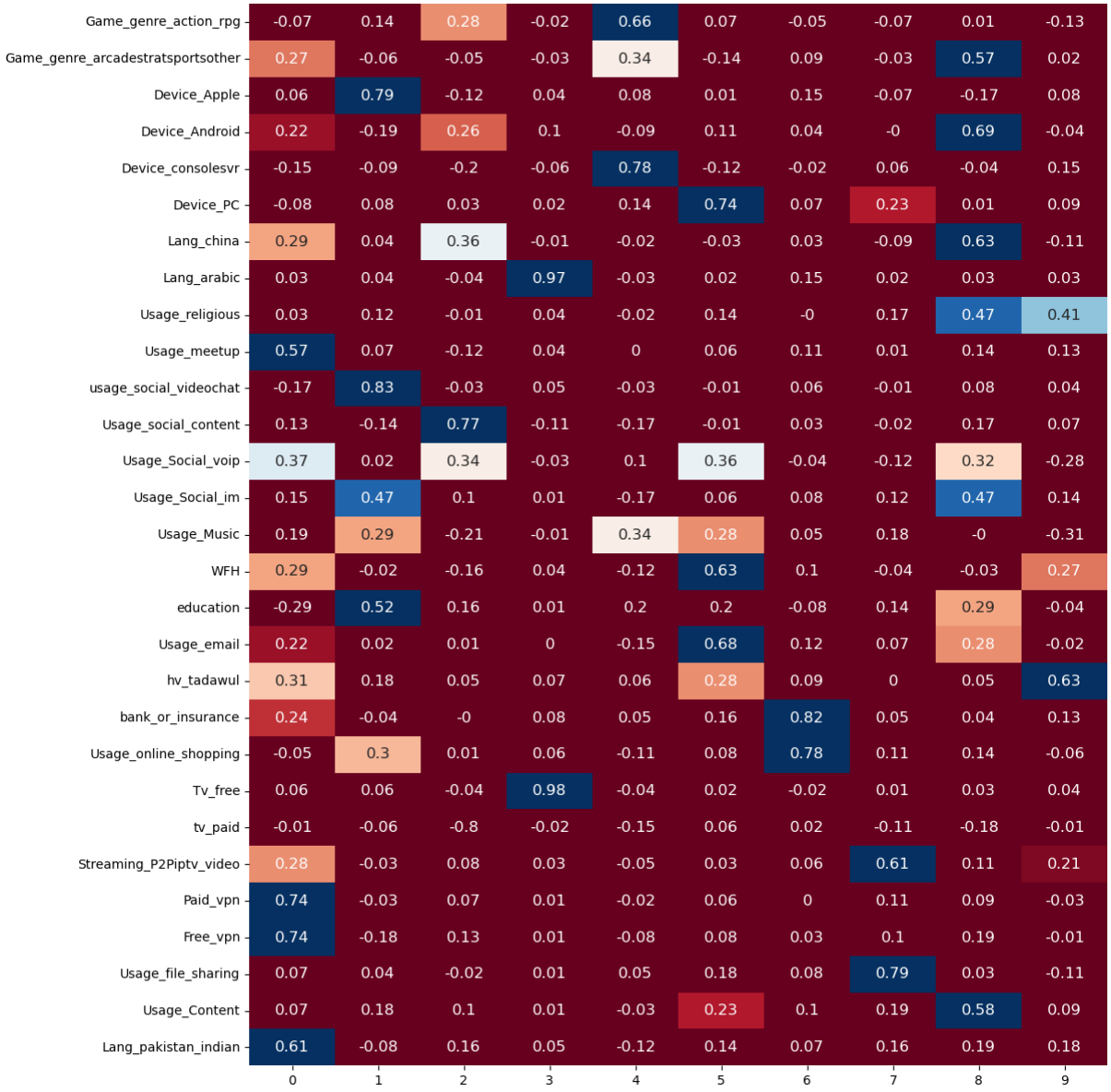

Figure 5 - Heatmap visualising the factor loadings of each variable on the latent factors.

.

The subsequent labelling of the clusters was done in collaboration with domain experts within the NOKIA team. Data was extracted from the above heatmap and plotted as radar charts for another visualisation method. Figure 6 shows such radar chart used to identify clusters. The green line represents the loading value for each variable while the orange is the global average. Here it is observed the cluster 1 (Young MMO

Gamers) observed a higher-than-average usage for categories like Action/MMO games and Console compared to the average.

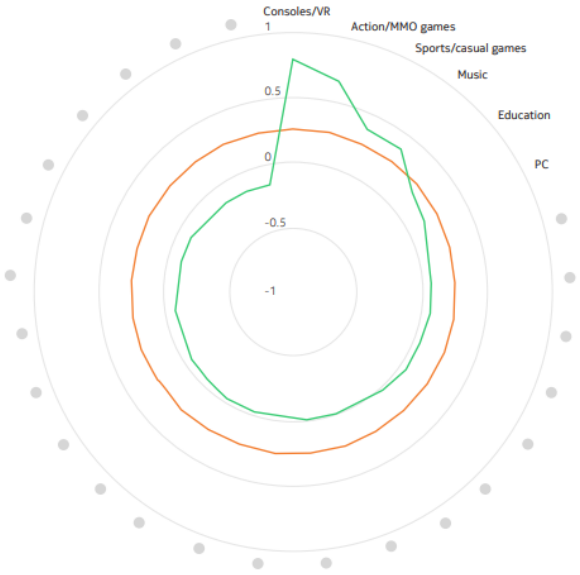

Figure 6 - Example of a Radar Chart for Young MMO Gamer Cluster.

Below shows a summary table of the clusters along with their percentage of the total customer base.

TABLE II. CLUSTER REPRESENTATION OF CUSTOMER BASE.

Cluster_ID	Cluster Name	% of base
0	Professional WFH	11.3
1	Young Console MMO Gamers	11.4
2	File-Sharers	8.7
3	Binge-Watching Arabs	11.3
4	Online Shoppers	11.1
5	Young Lower Budget conservative Family	9.4
6	Sociable Young Home	8.5
7	Social Media Addicts	9.3
8	Conservative Investor	10.2
9	Household with single Expat	9.0

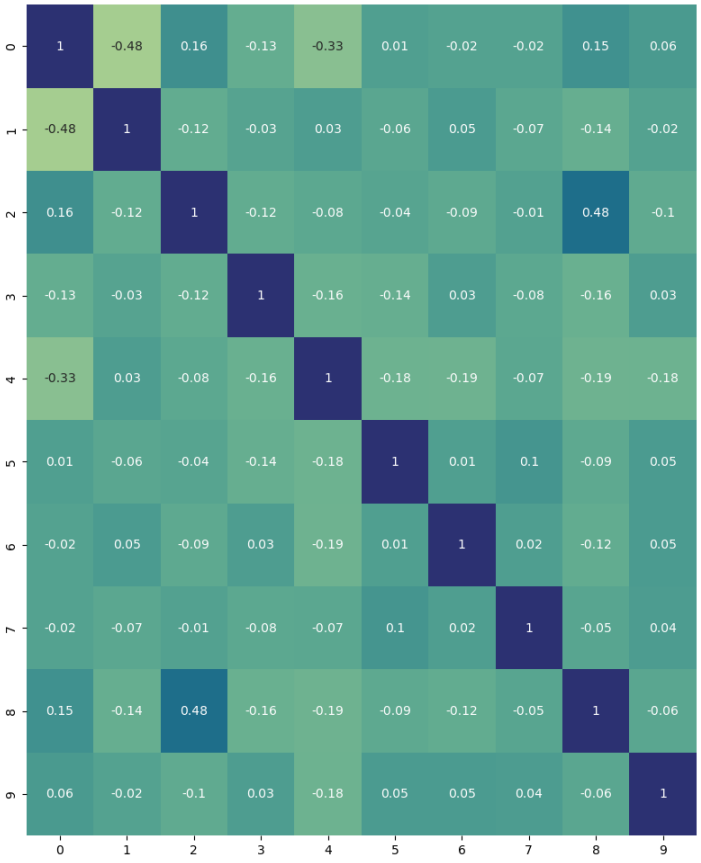

Figure 7 - Correlations heatmap between latent factors.

The above heatmap is the person's correlation between each of our latent factors. Factor analysis requires that the final factors be independent of one another and thus should not exhibit any correlation [3].

4. DISCUSSION

4.1 Tableau dashboards

Building tableau dashboards is a way for customers of this use case to consume the information about the digital personas and use the recommendations to grow revenue and increase customer experience. Multiple BI tools were considered but Tableau was chosen as it integrated with the MariaDB we used in our testing and experimenting. Dashboards were first designed as wireframes in the development stage and then implemented with actual data after the clustering results were obtained.

The dashboards were used to analyse the clusters and gain insights across three different stakeholder teams: operations, marketing, and customer experience (CeX). Subsequent recommendations could be made specific to each cluster across these three departments with the ultimate goals to grow revenue, reduce customer churn and improve customer experience. Operations dashboards were built for teams in the operations department and includes insights that can be used by field technicians to target clusters that may be experiencing high traffic volumes and higher latency. The goal here is to provide the operations team with a dashboard such that they can allocate resources to handle higher volumes in a proactive manner and plan adequately. The Customer experience Index (CEI) map is also present on this dashboard so operations teams can become aware of which regions of Saudi Arabia are having a negative experience and further diagnostics can be ran as they relate to latency or any of the other KPIs available to CEI. CEI is an index created by NOKIA which includes many KPIs concerning customer experience. Latency and packet loss are some of the KPIs that contribute to this index. A higher latency can contribute negatively to customers' experience as this represents the delay in receiving data from the network. Mapping the average latency that each cluster experiences can allow for proactive care for service providers to take the necessary steps to reduce this delay. A sample of this dashboard is shown in figure 8. Top left is the CEI of the cluster across different regions of Saudi Arabia. Top right graph shows the average latency of the cluster against their average volume over a 24-hour period. Bottom left displays the upstream and downstream bitrates for the previous 24 hours.

The middle bar chart shows the average latency for this cluster by sub-technology used and bottom right is showing the average latency over 24 hours with a reference line for poor latency as a KPI and recommendation.

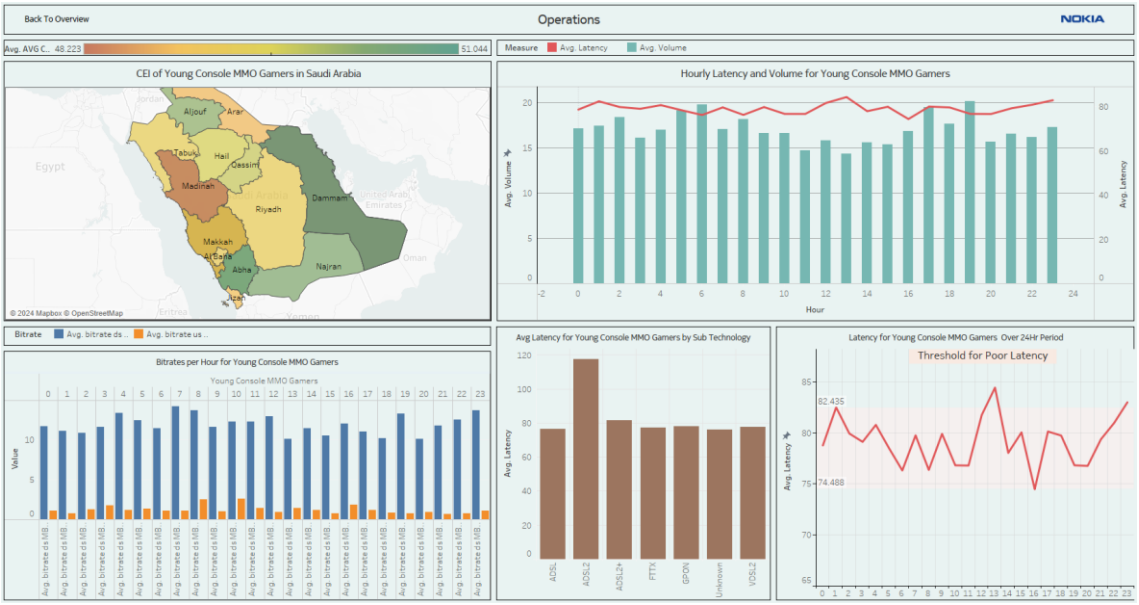

Figure 8 - Sample of operations dashboard for Young MMO Gamer Cluster in Saudi Arabia.

The Marketing dashboard was used a means to reveal potential upsell opportunities to marketing teams but also for marketing to better understand its customer base. Present within the marketing dashboard is the Fibre to the Home (FTTH) quadrant which shows fibre penetration and demand by cluster. Moreover, the percentages of rate plans offered by STC are shown for each cluster. In figure 9 we can see the percentage of plan for cluster five. The dark blue line is the percentage of cluster using this plan, while the grey line represents the average across whole subscriber base.

The customer experience dashboards used the customer experience index (CEI) data collected from CEM and CRM databases. Complaints and CEI data by cluster allowed consumers to map the change in complaints each month by cluster. Recommendations here could be an initiative-taking step in targeting clusters which experienced a higher-than-average complaints and this could further be mapped to the region and technology used by each customer.

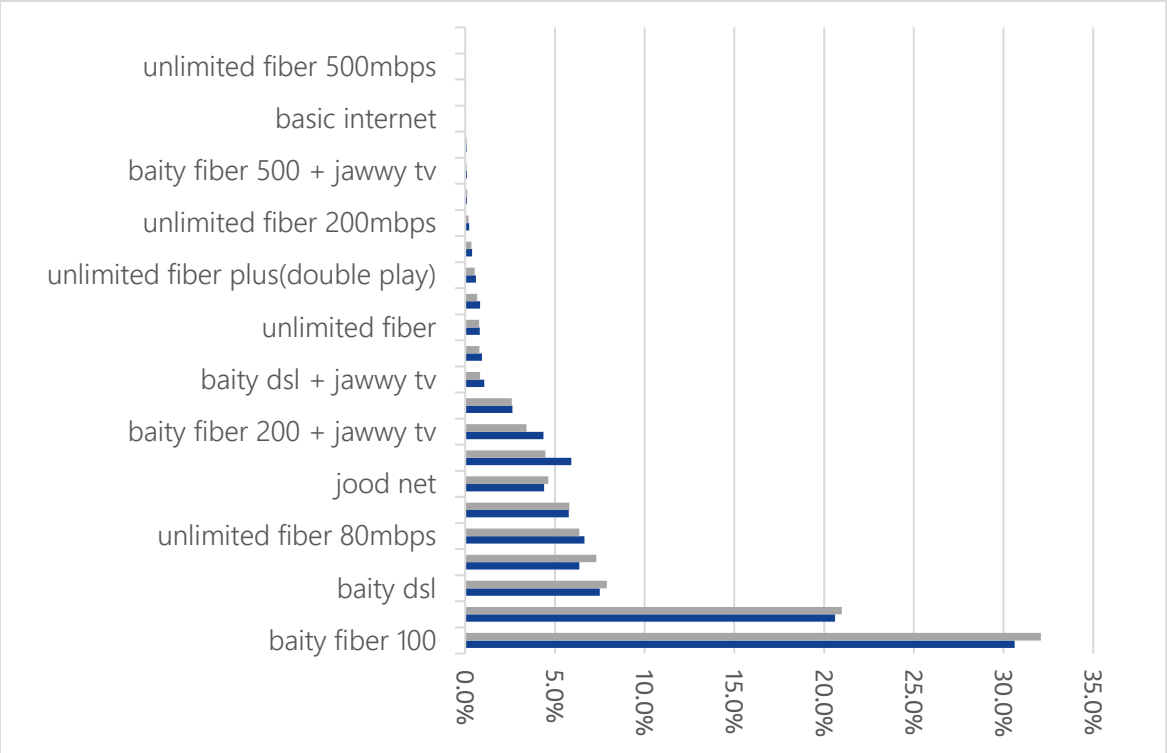

Figure 9 - Distribution of rate plans across customer base for cluster five.

4.2 Analysis and Recommendations

The key challenge of this use case was to discover the relevant patterns in subscriber behaviour and segment this behaviour into unique digital personas. The goal as stated previously was to enhance customer experience, grow revenue and improve on many of the KPIs relevant to the telecom industry (CEI, churn and ARPU etc.) The value gained for customers of such a product include an automated means to segment customers that may be updated as they require, advanced analytics, complex event processing with real time triggering logic and recommendation engines.

Early recommendations for STC have been documented here. For example, marketing could avail of customised offers for Online Shoppers using Qitaf points (loyalty points). Another marketing example includes offering upgraded packages to Sociable Young Homes to satisfy their low latency requirement. From an operations standpoint File Sharers exhibit a high transfer volume and adopting a proactive approach to meet the volume and speed demands of this cluster could reduce churn. Professional WFH

account for 9.5% of the customer base. This group on average uses remote connection tools and higher access to email and professional apps like Teams and Slack. Retaining Quality of Service (QoS) at the network level and checking WiFi quality of this cluster is critical for enhancing their experience.

Fibre to the home (FTTH) is a sub technology offered by STC. The different sub technologies that service providers offer to their subscribers can have a variety of effects on their experience such as speed and quality of service. FTTH offers superior speeds as it transmits data through fibre optic lines compared to traditional technologies such as ADSL (copper) [1]. FTTH is significantly under-represented in the STC customer base and thus offers an opportunity for STC to identify key customers who may benefit from switching to FTTH.

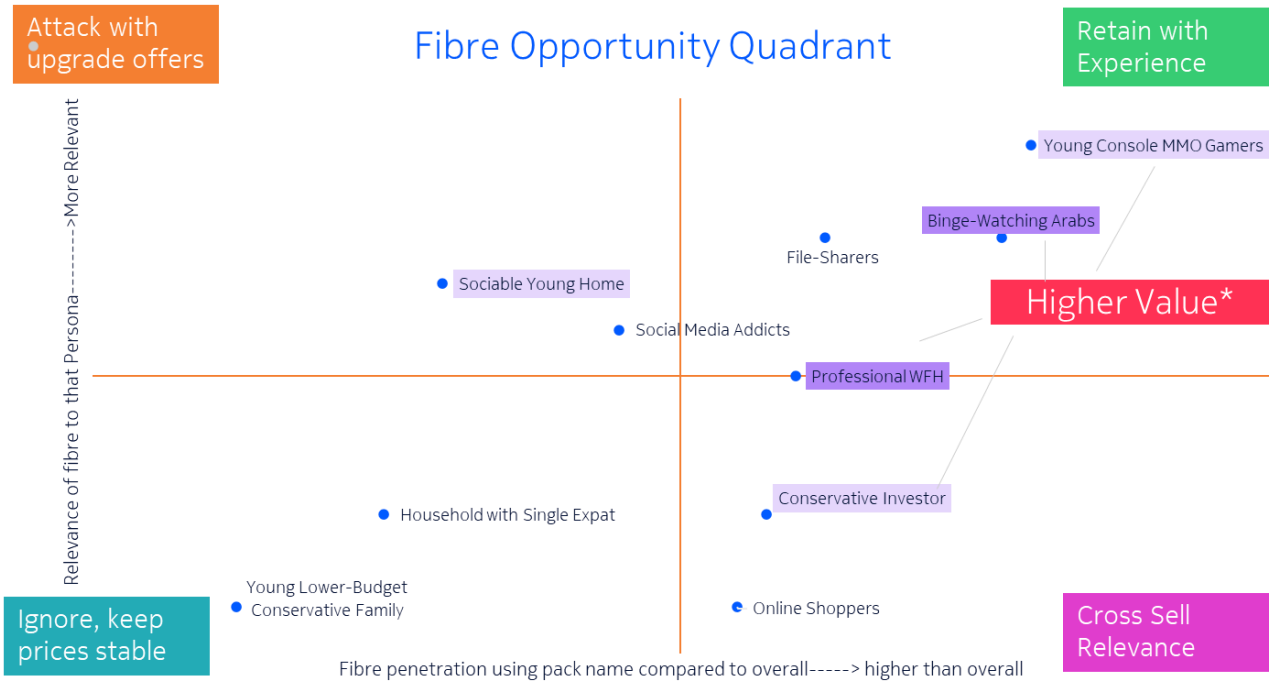

Figure 10 - Fiber To the Home opportunity quadrant.

The above opportunity quadrant highlights key considerations and places each of our clusters into a different quadrant. The Y axis represents the fibre relevance, quantifying how relevant fibre optic speeds are to the given cluster, of course higher speeds are more relevant for Online Console Gamers than Online Shoppers. Along the x axis is the fibre penetration compared to the average, this quantifies the penetration of FTTH in

the cluster and can be used to identify clusters where an upsell opportunity may arise. For digital personas that have high penetration and speeds are highly relevant, a proactive measure can be made to ensure this speed and quality is maintained. Below is a table sampling some of the recommendations that can be made pertaining to the FTTH quadrant.

TABLE III. FTTH QUADRANT RECOMMENDATIONS

Digital Persona	Recommendation	Quadrant
Young MMO Gamers	Retain experience to reduce potential for Churn.	High Penetration, High Demand for Speed
Professional WFH	Offer Cross Sell packages that benefit from Fiber to avoid downgrading.	High Penetration, Low demand for Speed
Sociable Young Home	Offer FTTH packages to satisfy speed demands	Low Penetration, High demand for Speed
Household With Single Expat	Keep Prices stable	Low Penetration, Low Demand

The Nokia ACS team in collaborating with the client for this particular use case exposed previously hidden FTTH upsell opportunities to sociable young homes and social media addicted digital personas. Such a finding would not only help revenue with increased FTTH sales, but also save OPEX by not wasting any time in irrelevant marketing strategies on digital personas for who FTTH is neither relevant nor penetration neither high. By segmenting customers into digital personas consumers of this use case may also monitor churn per persona. This allows for teams to identify clusters that are higher than average churners and it can be a starting point for proactive care. It is accepted that acquisition of a new customer is more costly than retention [2], so having data available regarding churn is valuable. Analysis from STC shows that Single Expats churn 12% more than the average of customer base. Churn percentages may then also be enriched from CRM data. Working with STC, Nokia identified five key churn reasons by grouping together 73 distinct reasons submitted into CRM.

TABLE IV. CHURN REASONS BY PERSONA.

Persona	Service Quality	Charges & Financial	Leaving Country	Staff	Under utilization
Professional WFH	9.34%	**32.34%**	**20.07%**	0.69%	37.96%
Young Console MMO Gamers	**10.86%**	36.41%	**12.45%**	0.51%	40.01%
File-Sharers	9.29%	36.13%	15.96%	**0.27%**	39.38%
Binge-Watching Arabs	9.85%	35.25%	**12.56%**	0.61%	38.56%
Online Shoppers	**10.51%**	35.49%	17.57%	0.70%	38.06%
Young Lower-Budget Family	**8.89%**	**38.74%**	15.83%	0.49%	**42.16%**
Sociable Young Home	**10.62%**	**39.23%**	**12.06%**	0.41%	**42.78%**
Social Media Addicts	9.21%	38.10%	14.53%	**0.87%**	**42.56%**
Conservative Investor	9.74%	36.56%	**11.74%**	0.42%	41.40%
Household with Single Expat	9.92%	**39.00%**	**23.90%**	0.47%	39.56%

The above table provides valuable insight and can be used to empower data driven decisions. It shows the percentages of churn by cluster where red represents highest cluster churning by that reason The main takeaways from this tell us that Charges and Under Utilization constitute the main reasons for churn and that is across all clusters. Households with Single Expats churn mainly due to financial decisions and leaving the country. Recommendations from this insight include proactive care for service quality sensitive clusters. This will be relevant to Young MMO Gamers, Online Shoppers and Sociable Young Homes. At a deeper level, one may analyse the root cause for poor service quality in these clusters such as connectivity scores or WiFi retransmission rates.

5. CONCLUSION

The results from this use case have been appreciated by NOKIA's advanced consulting service branch and for this, has been approved for further development into a minimum viable product that will be offered to other ISPs as a service. One limitation to our data is that we are clustering customers based on the most prominent internet usage in their households. From our data we cannot decipher how many users are in the home and how different users might benefit from different rate plans. For example, a household might have a family of 6 but the children's gaming habits dominate the bandwidth. This may be enriched with more CRM data and could be explored in further analysis.

We know that factor analysis is used to reduce the attribute space and uncover the latent structure of a set of variables [8]. We feel the use of factor analysis is an appropriate clustering method to explain the unobserved communalities in our data. Because internet usage is inherently personal, and 'human' one can assume there are some psychological and psychosocial factors influencing one's usage.

Much like how one study used factor analysis in churn predictions [2] we feel like there is opportunity to use the digital personas we have created here as a variable in future churn prediction models. We hope this can be explored further by Nokia and can become a unique selling point when marketing this use case to other clients.

The FTTH quadrant was one of three principal areas targeted with the results from building the digital personas. Other areas include ARPU which at the request of STC have remained confidential. Churn analysis shows how different clusters value the service provided by ISPs, and this opens many opportunities to target such clusters with personalised marketing campaigns and tailored offers, saving both the marketing teams resources and time. With the insights gained from building digital personas, internet service providers can remain competitive in this increasingly challenging business environment.

REFERENCES

[1] Digital Persona Whitepaper. *Nokia internal document* (2023)

[2] Tianyuan, Z., Moro, S., & Ramos, R. F. (2022). A Data-Driven approach to improve customer churn prediction based on telecom customer segmentation. *Future Internet*, 14(3), 94. https://doi.org/10.3390/fi14030094

[3] RA. Johnson and DW. Wichern, Applied Multivariate Statistical Analysis, vol. 4, Englewood Cliffs, NJ: Prentice Hall, (1992)

[4] Manly, B. F. J. (2004). Multivariate statistical methods. *Chapman and Hall/CRC eBooks*. https://doi.org/10.1201/b16974

[5] Zhou, Y., Jin, X., & Mobasher, B. (2004). A Recommendation Model Based on Latent Principal Factors in Web Navigation Data. *In WebDyn@ WWW*, 52–61. Retrieved from http://ceur-ws.org/Vol-703/paper6.pdf

[6] Bai, A., Hira, S., & Deshpande, P. S. (2015). An application of factor analysis in the evaluation of country economic rank. *Procedia Computer Science*, 54, 311–317. https://doi.org/10.1016/j.procs.2015.06.036

[7] Factor_analyzer documentation [https://factor-analyzer.readthedocs.io/en/latest/factor_analyzer.html#module-factor_analyzer.factor_analyzer]

[8] Wu, N., & Zhang, J. (2004). Factor analysis based anomaly detection. *IEEE Systems, Man and Cybernetics Society Information Assurance Workshop*. https://doi.org/10.1109/smcsia.2003.1232408

[9] Alkhayrat, M., Aljnidi, M., & Aljoumaa, K. (2020). A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-0286-0

[10] Khatibi, A. A., Ismail, H., & Thyagarajan, V. (2002). What drives customer loyalty: An analysis from the telecommunications industry. *Journal of Targeting, Measurement and Analysis for Marketing*, *11*(1), 34–44. https://doi.org/10.1057/palgrave.jt.5740065