Network Capacity and Machine Learning

A Technical Paper prepared for SCTE/ISBE by

Dr. Claudio Righetti
Chief Scientist & Security
Cablevisión S.A.
Gral. Hornos 690, Buenos Aires, Argentina
Phone: +5411 5530 4468
crighetti@cablevision.com.ar

Emilia Gibellini
Data Scientist
Cablevisión S.A.
egibellini@cablevision.com.ar

Florencia De Arca
Data Scientist
Cablevisión S.A.
fdearca@cablevision.com.ar

Carlos Germán Carreño Romano
Data Scientist
Cablevisión S.A.
caromano@cablevision.com.ar

Mariela Fiorenzo
Data Scientist
Cablevisión S.A.
mafiorenzo@cablevision.com.ar

Gabriel Carro
VP Engineer and R&D
Cablevisión S.A.
gcarro@cablevision.com.ar

Fernando Rodrigo Ochoa
Security Analyst
Cablevisión S.A.
fochoa@cablevision.com.ar

Table of Contents

List of Figures

List of Tables

Abstract

The purpose of this paper is to introduce STEM-ML, an extension of our network-dimensioning tool, which allows us to define the strategy to face the increasing demand, of both our Internet broadband and "Flow", our Internet Protocol Television (IPTV) services. This tool makes use of machine learning techniques to characterize the optical nodes that integrate our network. Based on such characterization, we can define the technologic and commercial strategy for the access network so that Cablevisión (CVA) is able to afford the short and long-term demand.

Until the development of STEM-ML, characterization was made at hub level, and it was based on the average bandwidth per subscriber parameters. With STEM-ML, the analysis is made at optical node level, and monthly consumption, households passed (HHP), and protocol types, among other variables. Moreover, data from "Flow" our IPTV platform is added. The increasing data volume generates the need for introducing machine learning and multivariate analysis techniques.

Content

1. Introduction

We decided to apply machine learning techniques as an extension of our network dimensioning tool, STEM, presented in Cable-Tec Expo '16 [1]. We called this extension STEM-ML and it makes use of algorithms such as Principal Components Analysis (PCA) and Artificial Neural Networks (ANN). The objective of this work is to characterize the nodes that make up our network in order to define the strategies that Cablevisión will use to meet short and long term demand.

In STEM-ML, we carried out different analysis at node level based on variables such as monthly consumption, households passed, traffic per port and downstream channels distributions, among others. We use a huge volume of data from different sources to obtain examples for the training sets used in the algorithms. As the obtained results are needed for a large number of cases and on a regular basis, it is necessary to automate these processes applying machine learning and multivariate analysis techniques.

1.1. Machine Learning Overview

The main applications of Machine Learning technology in telecommunications and in particular in the cable industry are listed below.

While machine learning has been under research and development for decades, we may wonder why it has just now become Strategic Technology and is in peak expectations of the Gartner's Hype Cycle for Emerging Technologies [2].

The reason why it has become strategic is the great processing power available and the thousands of algorithm developers who have improved the performance of that technology. In addition to the large investments made by companies such as IBM (Watson project), Google (TensorFlow project) or Microsoft (Azure).

Machine learning is the subfield of computer science that, according to Arthur Samuel in 1959, gives "computers the ability to learn without being explicitly programmed" [3].

Machine learning technologies can learn from historical data based on it making predictions or making decisions. That is the fundamental difference between any other applications developed from a program's instructions that are ran in a deterministic manner (Figure 1).

Machine learning is based on algorithms that learn from data without relying on rules-based programming (Figure 2).

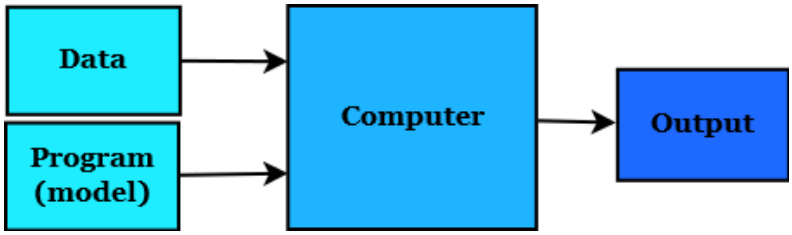

Figure 1 - Traditional Programming

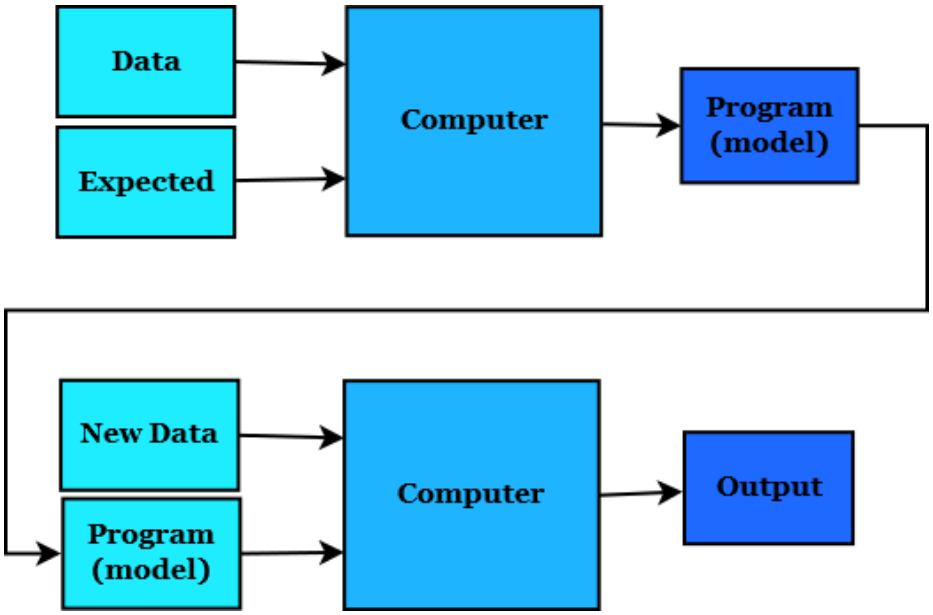

Figure 2 - Machine Learning

At the TM Forum [4] presentations and discussion panels, possible uses and potential applications in the Telecommunications business were presented in the analytics sessions. Among the advantages of the use of Machine Learning, we can mention:

- It allows fast and automatic analysis of large volumes of data that are becoming more and more complex. Getting faster and more accurate results that allow you to make reliable and repeatable decisions.
- Focus on behavioral analysis to detect and predict possible "anomalous" events at an early stage.
- Automate real-time analysis in the orchestration of end-to-end services in a virtualized world.

• Identification and mitigation of security threats in services through predictive analytics and machine learning to detect attacks that escape traditional preventative static defenses.
• Prediction of Churn. Unlike traditional strategies, machine learning allows a multi-class classification of our clients, for example to predict whether they belong to a low, medium or high-risk class.
• Support for automation and management of network orchestration and traceability of end-to-end transactions across the network and OSS / BSS environment.

For the cable industry in particular, Sundaresan et. al. in [5] provides an overview of Machine Learning algorithms, and how their potential applications could be applied:
• Software Defined Networks (SDN) Routing
• Profile Management on DOCSIS 3.1 cable modems [6]
• Proactive Network Maintenance (PNM): for DOCSIS
• HFC´s Network Health KPI

Some applications are being implemented in:
• Internet Traffic Characterization
• Network Traffic Engineering
• Wi-Fi Proactive Network Maintenance (PNM)

There are two main paradigms of machine learning, called supervised and unsupervised.

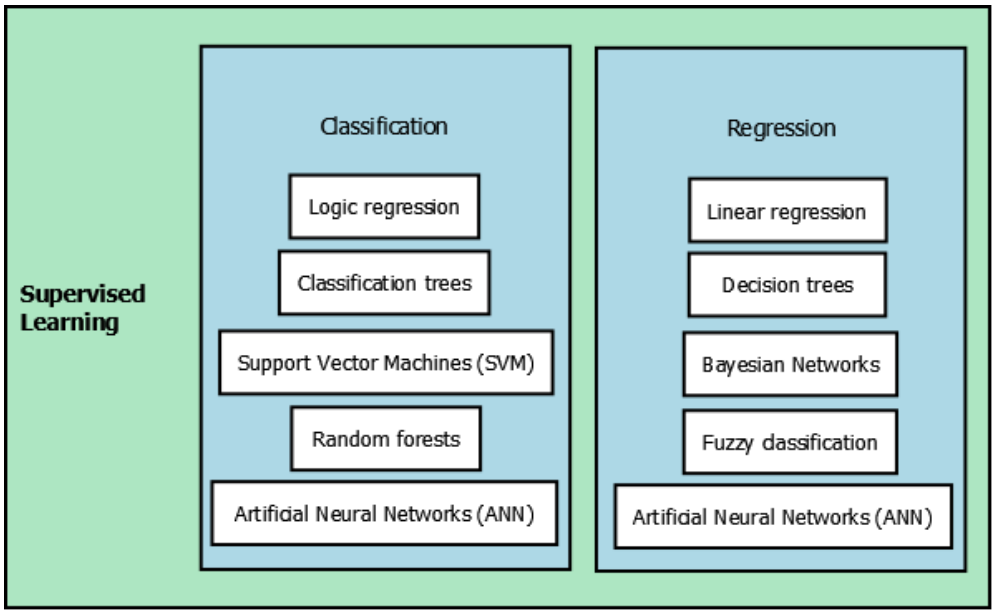

Figure 3 - Most used supervised machine learning techniques.

In supervised learning, your training data consists of some points and a label or target value associated with them. The goal of the algorithms is to figure out some way to estimate that target value. Learning stops when the algorithm achieves an acceptable level of performance.

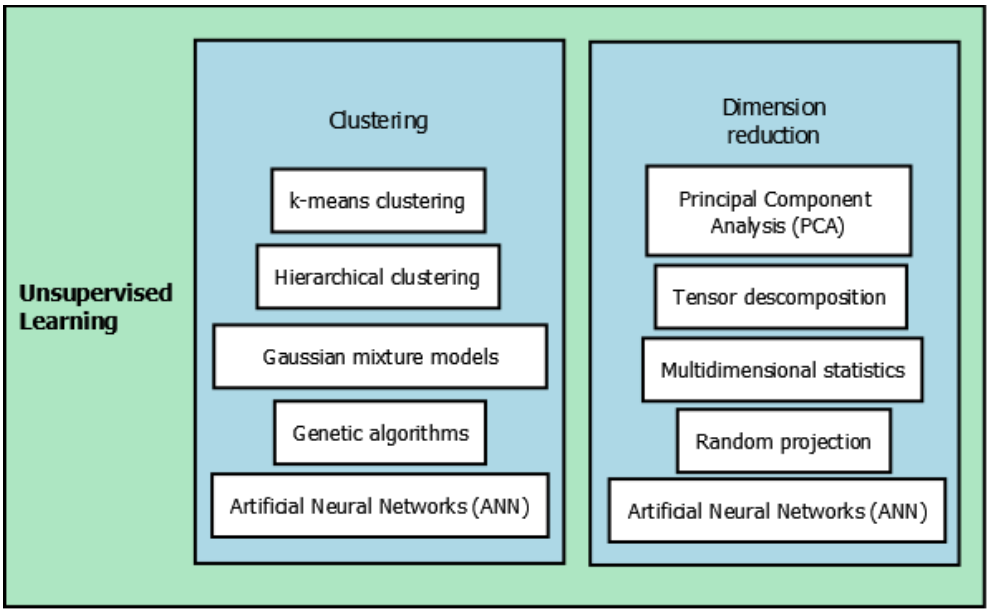

Figure 4 - Most used unsupervised machine learning techniques.

In unsupervised learning, there is just raw data, the output is not known beforehand. Unsupervised algorithms are used for finding hidden underlying structure in the data; there are no correct answers and no teacher.

Supervised learning is somewhat more common in real applications. However, unsupervised learning algorithms are often used as a preprocessing step for extracting meaningful features from a data point, with those features ultimately getting used for supervised learning [7].

1.2. Motivation

In order to define the access network strategy in Cablevisión, last year we carried out a first analysis of optical nodes, in which we characterize them according to their high or low demand, and it was assumed that in all of them the average amount of bandwidth per subscriber during prime time (T_{Avg}) has a 50% Compound Annual Growth Rate (CAGR). Based on this forecast, we suggested that in cases with higher values of this index (T_{Avg}), a new optical node should be installed using architectures N+0 such as Remote MAC-PHY or Fiber Deep, aligned with DOCSIS 3.1 evolution. For those cases with lower values, the suggestion was to enable more QAM channels or to segment the optical node.

At that time, we used historical data about traffic and cable modems volume at HUB and CMTS level. Now we are focused on this same data but at node level so we can make a more detailed and deeper analysis which helps us optimizing the investment plan.

1.3. Datasets Treatment

"Data today is often compared with oil, as in its raw form, its uses are limited. It is through refinement that oil becomes useful as kerosene, gasoline and other goods, and similarly it is through the refinement process of cleansing, validation, de-duplication and ongoing auditing that data can become useful in the kinds of advanced analytics that are starting to shape our world" [8].

Data plays a critical role in the development of smart solutions. Poor data quality puts organizations at risk of making unwise decisions, missing opportunities and undermining the customers' confidence. Rule of thumb: If your human experts struggle to come to conclusions with your existing data, ML will not fix it by itself.

As part of the data treatment phase, some decisions about error treatment based on business guidelines had to be made. Next, an overview of the data processing, main errors found, and ways used to offset them were provided.

As for the terminology used in this paper, a service group (SG) refers to the SCTE definition [9], which is a group of nodes, each node having a number of homes passed (HHP). A headend/hub serves multiple service groups. All nodes in a service group are served by a common switched RF spectrum.

Service group has been borrowed from the video world and has been defined in DOCSIS as the complete set of upstream and downstream channels that can provide service to reach a single subscriber device. Those channels may come from different MAC Domains and even different CMTSs. They could also come from video Edge QAMs.

We add two more concepts: segments and technical zones. Segments are the legs of an optical node that cover a geographical area, they relate to the number of modulators inside the node. Technical zone is the name we use to refer to a specific node/segment combination. This way, we say that a segmented node has an individual modulator for each segment. Therefore, we can say that a node is segmented 1x1, 2x2 or 4x4 depending on the number of downstream and upstream modulators. As an example, a group of technical zones called BON001A, BON001B, BON001C, BON001D refers to the four segments (A, B, C, D) of the node BON001.

Our objective is to estimate traffic at node level (or even segment level in some cases). We will briefly explain how we do this starting from traffic at port level. To do this, it is assumed that the registered traffic at one port is distributed between the nodes or segments in proportion to the CM number in each port.

With this objective in mind, we integrate data that comes from different sources: customer portfolio (cable modem count in each node/segment), traffic in each CMTS port (during prime time periods, Sundays from 18 hrs. to 00 hrs.), and ports/nodes combinations (Figure 5).

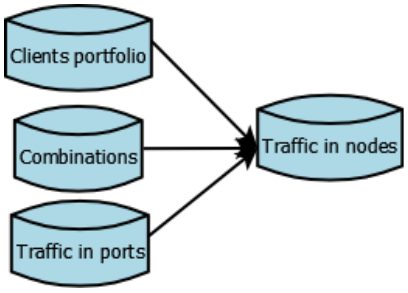

Figure 5 - Dataset treatment simplified scheme.

1.3.1. Portfolio, Combinations and Traffic

First, we take CVA data on subscribers and compute the cable modem count in each zone. After unifying the segments' notation, we obtain a data set which combines each CMTS port with its connected nodes, so that they match the one in the portfolio. We merge both data sets and we obtain Table 1.

Table 1 – Example of the resulting database after merging the nodes and CMTS data.

HUB	CMTS_Name	DS_Port	Node	Segment	Zone	CM_Count
ACC	CMT1.ACC1-BSR64K	12/0	ACC001	AB	ACC001A	163
ACC	CMT1.ACC1-BSR64K	12/0	ACC001	AB	ACC001B	164
…	…	…	…		…	…

Despite the corrections made, there still are some errors.

- The node is already segmented in the portfolio but not in the combinations data set. **Solution:** copy HUB, CMTS and port information to all segments.
- The node is segmented in the combinations data set but not in the portfolio. **Solution:** divide the cable modem count by the number of segments. This is an arbitrary decision, since the cable modem count is not necessarily evenly distributed among the segments.
- It is not known which port corresponds to each segment. **Solution:** sum up the cable modem count and divide it by the amount of registered ports.
- Some zones have more than one port associated. This is not a data set merge error, but probably a segmented node not yet updated. **Solution:** distribute the cable modem count equally among the ports connected to that zone.

The last dataset that is left is the one that contains traffic data. As we said before, we take this data during prime time and then use the maximum traffic per port.

After joining all the data sets, we still have to decide how to distribute the traffic of each port among the zones that it is connected to. At this point, there are three possibilities:

1. The data we have is complete; we have the cable modem count for each zone connected to the port.
2. The data is partially complete; we have the cable modem count for some zones, but not all.
3. None of the zones associated to the port has the data about cable modem count.

For the first case, we obtain the following weight:

$$weight = \frac{cable\ modem\ count\ in\ the\ zone}{cable\ modem\ count\ in\ all\ the\ zones\ connected\ to\ the\ port}$$

Then, each port traffic is distributed using this weight.

For the second case, if one of the segments connected to the port does not have the cable modem count, we consider it a segment with few cable modems, or one that is not yet active. Therefore, we consider this zone has zero cable modems, and the traffic is distributed between the ones that do have this information.

For the third case, another type of estimation is used: the total registered traffic at port level is evenly distributed among all the zones that are connected to it.

Finally, Table 2 shows the errors in the final data set.

Table 2 - Errors in the final database.

Description	%	Cumulative %
No errors.	90.15%	90.15%
Some zones without portfolio information - FIXED.	1.13%	91.28%
No portfolio information - FIXED.	1.31%	92.59%
Ports not associated with any zone.	7.48%	100%

We continue to work towards improving our dataset quality, since it is a common task for data science.

1.4. Variables in this Analysis

This analysis uses traffic data collected every Sunday during prime time, between February 19th and June 4th 2017, for all the ports registered in Cablevisión network.

In particular, one of the variables in these datasets contains the maximum traffic (Kbps) registered in each port. Our approach consists in analyzing two key indicators:

- Average bandwidth traffic per residential subscriber at peak time.
- Ports usage.

The former will provide information about the zones where there is a need for higher bandwidth, and the latter will help us find the optical nodes where ports are operating at almost their full capacity, conditioning the Quality of Service (QoS) and limiting the demand.

To assess the average bandwidth traffic per residential subscriber at peak time, the metric is defined:

$$Average\ BandWidth\ per\ Subscriber\ [Kbps] = \frac{Port\ Traffic}{\#\ Subscribers\ connected} \quad (1)$$

For measuring ports usage, the maximum utilization was defined:

$$Max\ utilization\ [\%] = \frac{Max\ Port\ Traffic}{Port\ capacity} \quad (2)$$

For practical purposes, the maximum utilization metric is also referred to as utilization.

To gather data about how the two key indicators relate to other variables, we also included in our analysis: the count of segments or zones connected to one port, CMTS model, classification of optical nodes according to the region where they are located, 2016 investment plan status, network capacity (1GHz or other), DOCSIS 2.0 and DOCSIS 3.0 cable modems count, HHP, network extension (in Km) and total monthly downstream consumption.

We also included the variable ‘downstream channels available per port’. Depending on the CMTS model, and knowing how many of downstream channels are already being used, we calculate how many more channels the network can support.

2. Exploratory Data Analysis (EDA)

We conducted an extensive exploratory analysis to assess the variables variation ranges, the correlations among them, whether the variables influence the traffic and utilization, and which is the magnitude and trend of that influence to determine its significance in the analysis.

A typical process for data modelling is: Problem → Data → Analysis → Model → Conclusions.

Data visualization is a key aspect of the exploratory phase. A correct visualization should be clear and easy to understand, in order to help any reader detect trends. We used SAS® to obtain a variety of plots and charts, a selection of which are shown in this section.

2.1. Average Bandwidth per Subscriber

Over the surveyed period, there was a slight increase of the average bandwidth per subscriber mean and median. In addition, in Figure 6, the percentile 95 is plotted. This tells us under which values do 95% of the observations lie. The colored area is greater as time passes, so we understand that on every Sunday there are new higher values of the average bandwidth per subscribers, and this means that every Sunday the metric varies among a wider range.

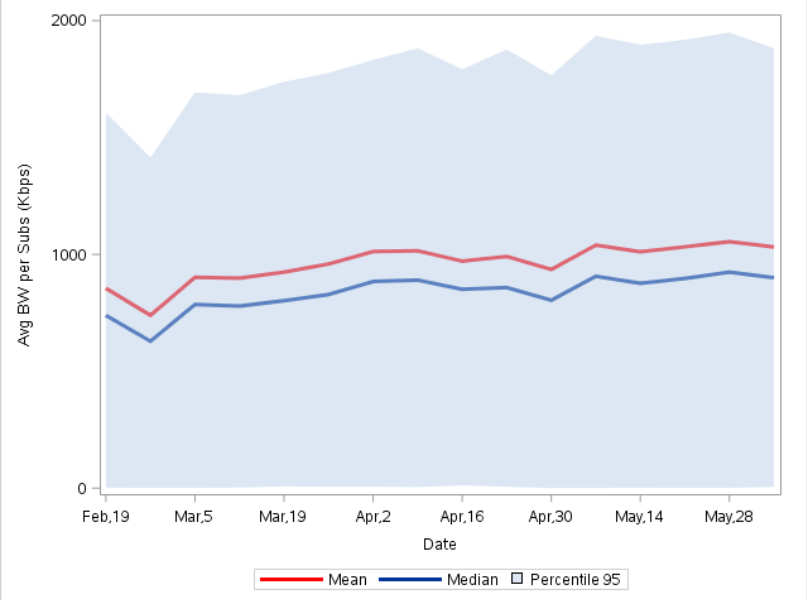

Figure 6 – Evolution of traffic over the surveyed period.

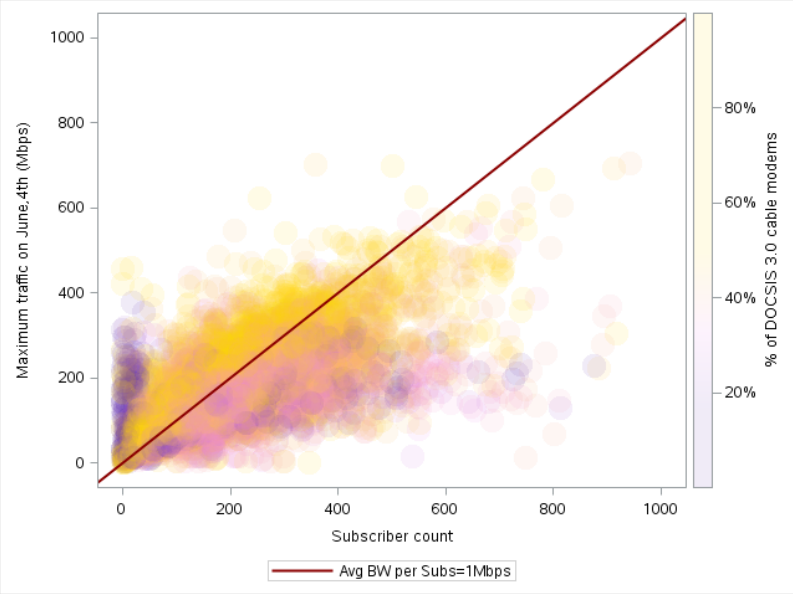

Figure 7 - Traffic at segment level on June, 4th (Mbps) versus subscriber count, colored according to the % of cable modems with DOCSIS 3.0 in each segment.

The diagonal line in Figure 7 divides the segments in the plot. Above it, there are segments in which the average bandwidth per subscriber is higher than 1Mbps. Below, there are the ones for which the metric is lower than 1Mbps. Notice that in most of the zones above the diagonal, the percentage of cable modems with DOCSIS 3.0 is about 60% or higher. On the other hand, the zones below the line tend to have more DOCSIS 2.0 cable modems.

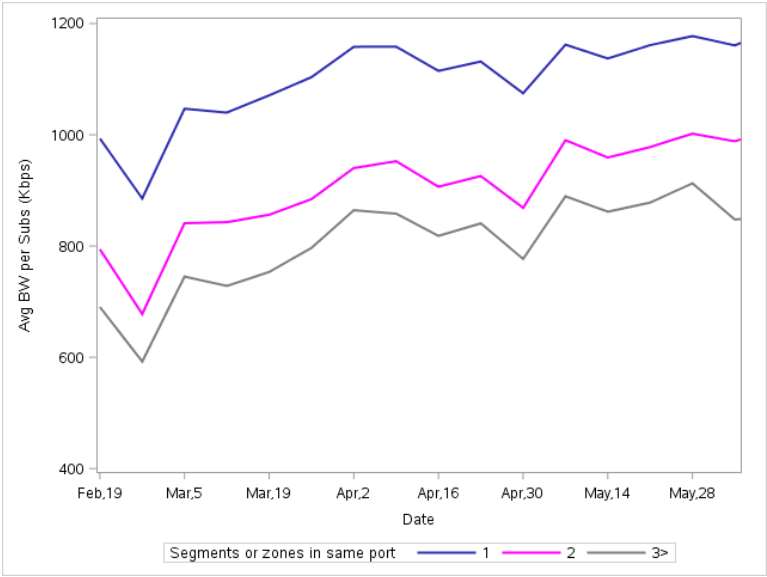

Figure 8 - Evolution of Avg BW per subscriber according to the number of segments in same port.

Figure 8 shows the evolution of the average bandwidth per subscriber mean in the surveyed period. It can be seen that when two or more segments are connected to the same port, the average bandwidth per subscriber drops off. For the cases with two zones connected, the mean is about 20% lower than the same metric for the ones with only one zone. When there are three zones or more, this difference increases another 10%.

Table 3 - Classification of ports according to the number of segments connected. Mean and standard deviation of cable modem count for each type of port.

Segments per port	% of ports	Cable modem count mean	Cable modem count SD
1	65%	208	120
2	30%	322	141
3	5%	472	192

2.2. Correlation between Traffic and Monthly Consumption.

We were interested in studying the correlation between total monthly consumption and the average bandwidth per subscriber. It is well known that the ports with higher traffic are the ones in which there are more cable modems. That relationship actually exists, as is shown in Figure 9, and the correlation between these two variables in May 2017 was 0.85.

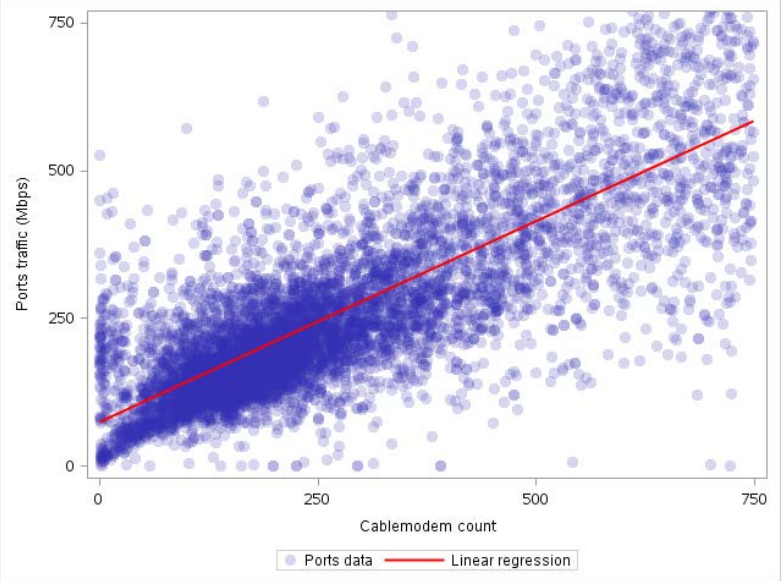

Figure 9 - Traffic versus cable modems count at port level.

On the other hand, monthly consumption at segment level is defined as the sum of all consumption that came from the subscribers in that segment. The correlation between monthly consumption and number of cable modems in May 2017, displayed in Figure 10, was 0.61.

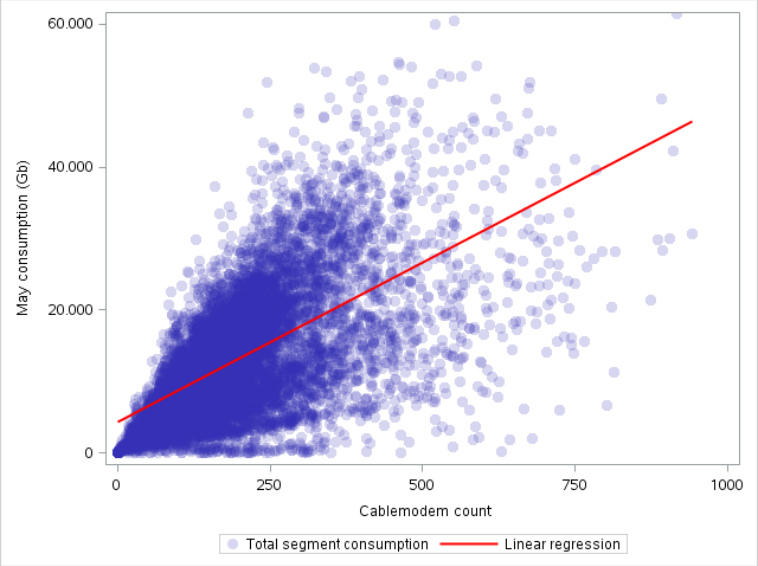

Figure 10 – Monthly consumption versus cable modems count at segment level.

If we plot the ports traffic versus monthly consumption values in the segments connected to each port, it may seem as if they correlate. Nevertheless, we already know both variables are associated to a third variable, which is the number of cable modems in each segment.

In order to assess if there is a correlation between maximum traffic during peak hours and total monthly consumption, we tried to remove the effect of cable modems count on both variables. Therefore, we postulated two linear regression models:

Model 1:

$$Y_i = \alpha_0 + \alpha_1 . X_i + \varepsilon_i \quad (3)$$

Where

- Y_i: Maximum traffic registered in port i
- X_i: Sum of the cable modems count of all the segments that are connected to port i
- α_0,α_1: Fixed coefficients to be estimated
- ε_i: Random error component, $\varepsilon_i \sim N(0, \sigma_1^2)$

Model 2:

$$Z_i = \beta_0 + \beta_1 . X_i + \delta_i \quad (4)$$

Where

- Z_i: Total monthly consumption registered in May 2017 in segment i
- X_i : Cable modems count in segment i
- β_0,β_1: Fixed coefficients to be estimated
- δ_i: Random error component, $\delta_i \sim N(0, \sigma_2^2)$

In both cases, residuals can be calculated. A plot of the residuals from the first model versus residuals from the second will provide information about correlation between traffic and monthly consumption once the effect of the third variable, cable modems count, is removed. This is equivalent to calculating the partial correlation of the first two variables, conditioned by the third. In Figure 11, it can be seen that there is no association between traffic at peak time and monthly consumption, at least when they are measured at port level. In fact, the estimated partial correlation is -0.06.

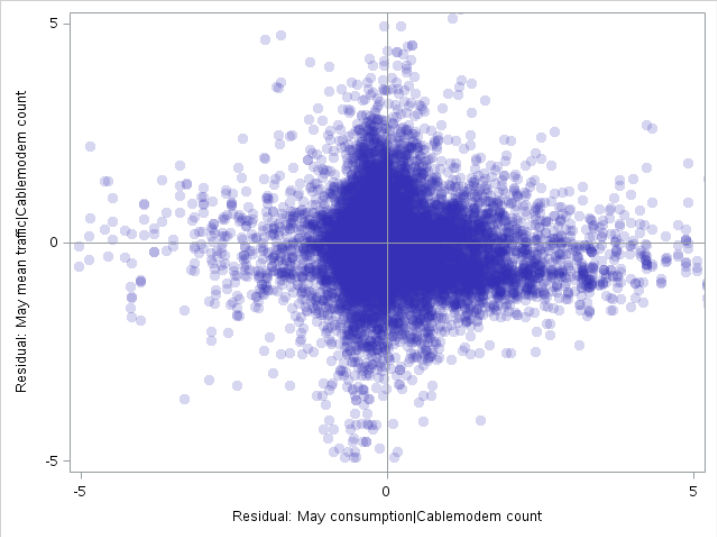

Figure 11 - Residuals from Model 1 versus residuals from Model 2.

We evaluated the same correlation at client level. Considering that for the majority, the expected monthly consumption is about 70 GB. We defined heavy users as the ones who download more than 1000 GB per month. It was found that heavy users also produce high traffic during prime time. We can conclude that these clients do have an impact on the ports traffic.

2.3. Principal Component Analysis

Sometimes data is collected on a large number of variables, so it becomes too large to study and interpret properly; there could be too many pairwise correlations between the variables to consider. That is why it may be useful to use a dimension reduction technique to help simplify the analysis. One of these techniques is Principal Component Analysis (PCA).

PCA takes in a collection of d-dimensional vectors and finds a collection of d "principal component" vectors of length 1, called PC1, PC2… and PCd . This means you can get as many main components as original variables. A point x in the data can be expressed as:

$$x = a_1 \cdot PC1 + a_2 \cdot PC2 + \cdots + a_d \cdot PCd \quad (5)$$

However, the PCi are chosen so that generally a_1 is much larger than the other a_i , a_2 is larger than all a_i except for a_1, and so on [7]. Therefore, there will be a subgroup of components explaining a high percentage of the total dispersion of data usually 2 or 3 components.

Principal Components are the underlying structure in the data. Their interpretation will be based on the weights obtained from the original variables. PCA is a way of identifying patterns in data, and expressing it with fewer variables.

We made a PCA for the ports database. The variables included were:

- Maximum traffic per port for each prime time (Sundays from 18 hrs. to 00 hrs.) from 02-19-2017 to 06-04-2017 (Kbps_port1 – Kbps_port16)
- Count of downstream channels in use per port (Channels_used)
- Count of areas connected to each port (Areas_port)
- Households passed (HHP)
- Residential subscribers per port (Subscribers)
- Traffic Management

Table 4 shows the proportion of the total variance of the data that is explained by each component. The first one explains 77.24% of the total dispersion. For the second one, its 5.7%. This means that the first two principal components explain 82.94% of total variance.

Table 4 - Percentage of the total variance explained by each PC.

Principal Component	% of variation	Cumulative %
1	77.24	77.24
2	5.70	82.94
3	4.30	87.24
4	3.66	90.90
…	…	…
21	0.13	100

Each principal component is just a weighing of the original variables. Therefore, in order to assign them a name and a meaning, we analyze the weights for the other variables. We concluded there are two main PC, which can be explained as follows:

- PC1: It will take higher values for those ports that registered more traffic during the time surveyed. Channels in use, areas, amount of cable modem and HHP per port also have a positive yet lower impact.
- PC2: This component will take higher values as the amount of areas, cable modems and HHP per port increase, as well as traffic management. On the other hand, it takes lower values as the number of downstream channels in use increases. This variable informs about a port's incapacity to provide a good service in highly populated areas.

We decided to compare PC1 and PC2 to understand the information they provide about the ports.

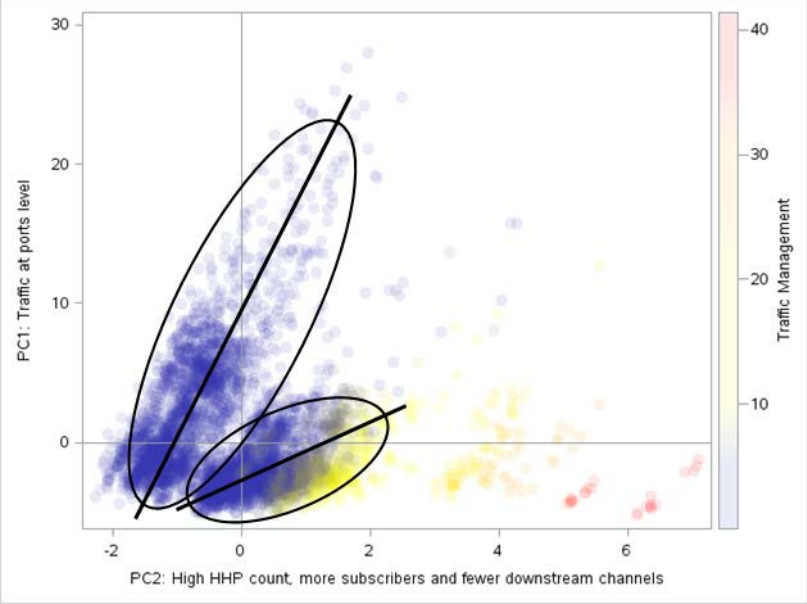

Figure 12 - PC1 vs PC2 and its relation with traffic management.

Figure 12 shows two groups. The first group where ports connect few subscribers and have many downstream channels. An increment in the subscribers count is associated to a huge increment in the traffic at port level. This increment is possible because these ports have a higher capacity due to the large number of downstream channel. This is not the case for the second group, which presents more subscribers and fewer downstream channels. The increase in subscribers count does not seem to have such an impact in the total traffic at port level. There are few cases where traffic is slightly influenced by traffic management.

In the same figure, we can also see a few ports where the Operations team was performing tasks affecting the service and because of that the traffic management has higher impact on the traffic. These ports connect highly populated areas, meaning, they are associated to high cable modems and HHP count, with many areas connected, but still few downstream channels in use.

We conclude the PCA highlighting that there are two groups of ports. One where the aggregation of more subscribers draws a substantial increment in the traffic, and another where the impact of adding subscribers is lower.

2.1. Ports Usage

In Figure 13, we can observe that the percentage of ports with utilization below 80% decreases over the observed period.

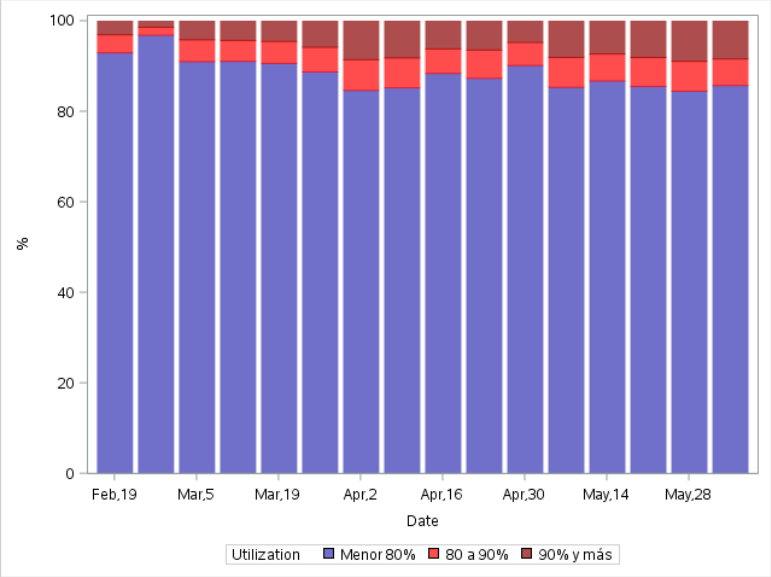

Figure 13 - Evolution of utilization over time.

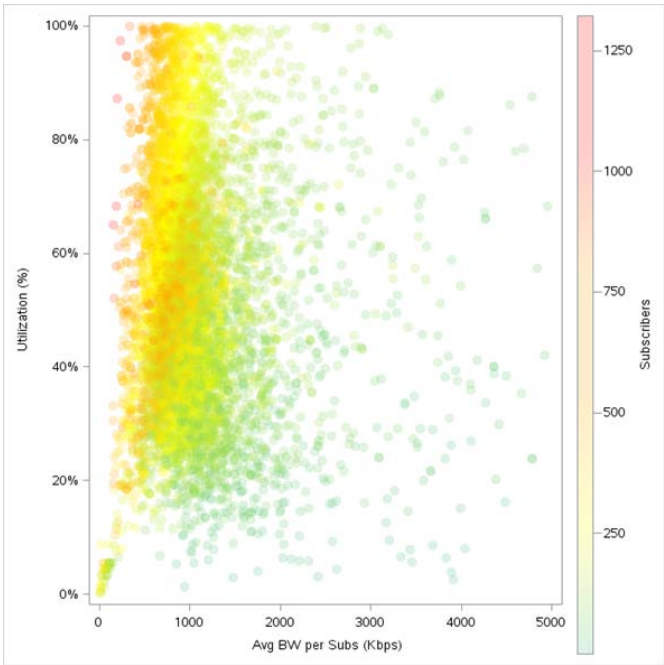

Figure 14 - Avg BW per Subs (Kbps) versus Utilization (%), colored by number of subscribers.

Figure 14 shows that the traffic per subscriber is generally below 2 Mbps. For obvious reasons the more subscribers the lower average bandwidth per subscriber, the fewer subscribers the higher average bandwidth. We can also interpret that, as mentioned in section 2.2, there is no correlation with utilization. If we pay attention to the coloring in the plot, we can see that when there are more subscribers, the ports utilization is also higher.

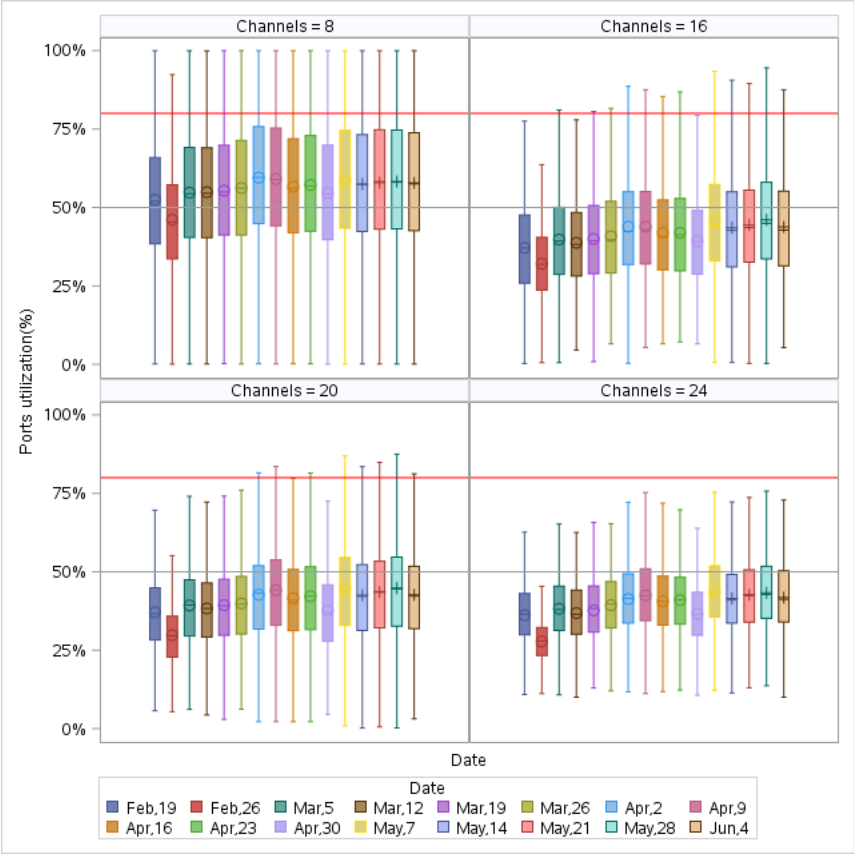

Figure 15 - Utilization distribution according to count of downstream channels being used.

Figure 15 shows the distribution of utilization over time according to the downstream channels distribution. Notice that utilization not only is higher for ports with eight channels, but it also has higher variability. In all cases, utilization is lower on February 26th due to a long weekend effect, as on February 27th and 28th it was national holiday in Argentina.

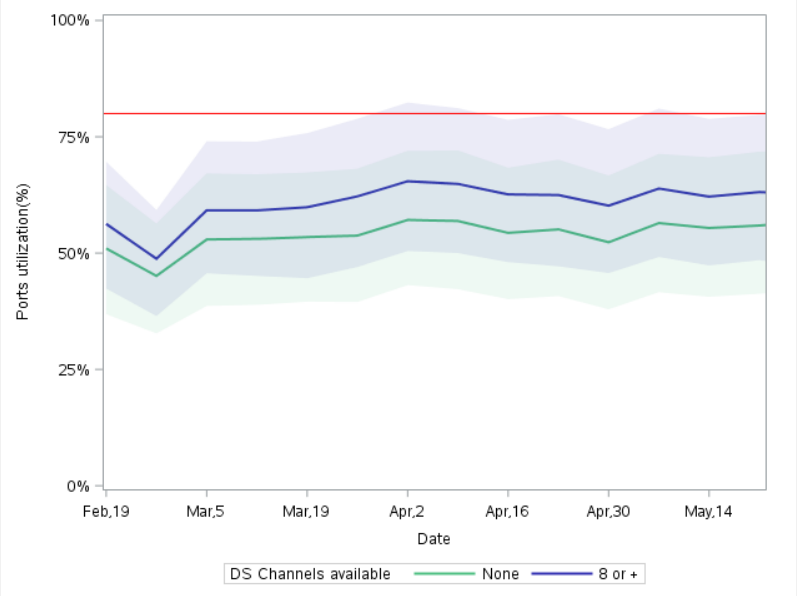

Figure 16 - Utilization in ports with 8 DS channels used, none and 8 or more channels available. The colored area shows the interval between the 25th and 75th percentiles.

We compare the utilization distribution in the ports where there are no channels left to be used against the ones where there is some capacity left. In Figure 16 the blue area shows that half of the ports with eight or more downstream channels available have utilization values that lie between 40% and 80%. Utilization of a quarter of the ports in this group is below 40% and another quarter, is above 80%. On the other hand, the green area shows that percentiles of utilization for the ports with no channels available is generally lower.

3. Construction of an Artificial Neural Network (ANN)

We based our ANN on the following principles [10]:

- Parsimony Principle: the simplest model that fits the data is also the most plausible.
- Sampling Bias Principle: if the data is sampled in a biased way, then learning will produce a similarly biased outcome.
- Data Snooping Principle: if a data set has affected any step of the learning process, its ability to assess the outcome has been compromised.

3.1. Artificial Neural Networks

The inventor of one of the first neurocomputers, Dr. Robert Hecht-Nielsen, provides the simplest definition of an Artificial Neural Network (ANN). He defines a neural network as:

"...a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs" [11].

The original goal of the ANN approach was to solve problems in the same way that a human brain would. They provide a practical method for learning discrete-valued functions from examples [12].

ANNs are typically organized in layers. Layers are composed of a number of interconnected nodes, which contain an "activation function". Patterns are presented to the network via the "input layer", which communicates to one or more "hidden layers" where the actual processing is done via a system of weighted connections. Most ANNs contain some form of "learning rule" which modifies the weights of the connections according to the input patterns that it receives. The hidden layers then link to an "output layer". If the network generates a "good" output, there is no need to adjust the weights. However, if the network generates a "poor" output, then the system adapts, altering the weights in order to improve subsequent results [13].

A network of many neurons can exhibit incredibly rich and intelligent behaviors. Once a neural network is "trained" to a satisfactory level, it can be used as an analytical tool on other data. They are simple to use and effective classifiers.

3.2. Network Access Strategies

In order to increase the capacity of the network and consequently the access speed for DOCSIS 2.0 and DOCSIS 3.0 HFC networks, we decide which of these four strategies to apply:

1. **Chassis upgrade**
   It consists on an upgrade of the firmware in the CMTS, starting with one or 8 QAMs per physical port, we can upgrade to 16, 20, 24 and 32 QAMs on the same port.

2. **Recombination**
   This process is an upgrade inside the HUB or internal plant. During the deployment of an optical node, and based on the capacity of multiple channels per downstream physical port, a common way to size a service group is to split one downstream port into 1:2 RF combiner. Each one goes to an individual transponder and by the optical Tx goes to two different optical nodes. When we make this kind of upgrade, we remove the RF combiner and reconnect each Tx to a new physical port.

3. **Node segmentation**
   This process is an upgrade in the external plant. We open an optical node, plug a new Tx or Rx (or both) module, and reconnect the segments to the new modules. One optical node may be segmented or not depending on how many modules it has. Typical segmentation schemes are 1x1, 1x2, 2x2, 2x4 and 4x4, where the first and second numbers belong to the number of Tx modules and Rx modules, respectively.

4. **Node division**
   Another task we use to increase the capacity of the network is node division and it can follow two different approaches: the first one is to divide the existent node into two new nodes, reassigning the distribution of subscribers among the two nodes. The other one point to a deep fiber approach. The deep fiber approach requires a node n+0 topology. With the existent network it implies that each amplifier will be replaced with a new fiber node reducing the number of HHP per node by the rate of the existing active devices.
   Depending on the chosen strategy, in terms of numbers, it implies that we can jump from a node with 500-1000 HHP to two nodes with 250-500 HHP in the case of node division, or 500-1000/x where x is the number of the actives above the node in the n+m existing topology. For example, if we have n+3, n+4, n+3, n+2 in each segment of a node with 500-1000 HHP, we will increase the number of nodes to at least 4 new deep fiber nodes and the jump goes to 120-250 HHP.

We expect our ANN will classify Cablevisión network nodes into one of these four strategies. The criteria that it will learn is the one given by the expert team.

3.3. Data Sample

We want to get a sample of nodes for the expert team to classify. Then, we use this data to train a neural network to do the same job for all the nodes in the network.

To determine the characteristics of the sample, we classify nodes into three strata: in the first one, we'll have the nodes in which the ports have a mean utilization below 50%; the second stratum contains the ones where mean utilization lies below or is equal to 80%, and in the third one, the nodes for which the mean utilization is above 80%.

The HHP count in each optical node is a determinant factor when it comes to choosing a strategy. Table 5 shows that when utilization is high, there is an increase of households passed mean and standard deviation.

Table 5 – HHP statistics according to the utilization range.

Utilization range	Strata size	HHP Mean	HHP Standard Error
<=50%	2,539	636.10	364.98
50%-80%	1,992	727.73	418.36
>80%	426	804.61	450.16

It is in our interest to have a fair representation of the HHP variable in our sample. Therefore, we will look for a sample size, so that if we had to estimate the mean HHP in each stratum, the estimation would have a certain standard error. As the investment in nodes with higher utilization has a higher priority, we want to be more accurate in these cases. Hence, the desired standard error for the second and third strata will be lower than the desired standard error for the first one.

One way to obtain a sample in a stratified population is to treat each stratum as a "population" and calculate a simple random sampling for each [14]. As the wanted precision varies from one stratum to the other, we decided to calculate separately for each stratum:

$$n_h = \frac{S_h^2}{V_h}$$

Where

- S_h: Squared standard error, which is the same as the variance of stratum h
- V_h: Desired variance for the sample in that population.

The desired standard errors and sample sizes (n_h) are shown in Table 6.

Table 6 - Sample size calculation for each stratum.

Utilization stratum	HHP Desired Standard Error	n_h
<=50%	100	13
50%-80%	80	27
>80%	80	32

If we sum n_h, we obtain a total sample size of 72 nodes, which are randomly selected within each utilization stratum.

3.4. Neural Network Training

It is possible that after training the neural network, which supposedly returns highly accurate classifications, the predicted classifications do not make sense when new data is introduced. This is generally due to overfitting, which means that the neural network adjusts too well to the data used for training but to that dataset only.

In order to prevent overfitting, we split the sample data in three subsets: training, cross-validation and testing. We assign 60% of the cases to the training set, so we use them to estimate the weights in the neural network. Then, another 20% goes to the cross-validation set, which helps us validating the model in terms of the variables and optimization parameters selected. Finally, we use the remaining 20% as testing set. This part of the sample does not participate in the construction of the neural network, it is only used to check whether there is overfitting or not by measuring how well would the network classify 'new observations'.

The first time we tried to train a neural network, using the software Octave [15], we introduced the variables that usually appear in the process of choosing the strategy. Specifically, these variables were: ports' utilization, HHP, cable modem count, average bandwidth per subscriber, downstream channels being used, downstream channels available, count of segments and nodes sharing ports, network capacity and segmentation level.

We soon discovered that the existing correlations between these variables were decreasing the network's accuracy. We used a technique for variable selection known as *stepwise*, to find the optimal combination of input variables. As a result, the first layer of the network contains four inputs: utilization, downstream channels available, count of segments and nodes sharing ports and segmentation level.

At first, accuracy was low and we wanted to improve it. We used the cross-validation set to obtain a learning curve that could help us understanding if we needed some more training observations or to change the selected variables. Figure 17 shows that after 25 observations, even if the sample size increases, the errors in the training set and in the cross-validation set are approximately the same. This indicates that we don't need a bigger sample to train the network. Actually, we need to reflect a more complex structure.

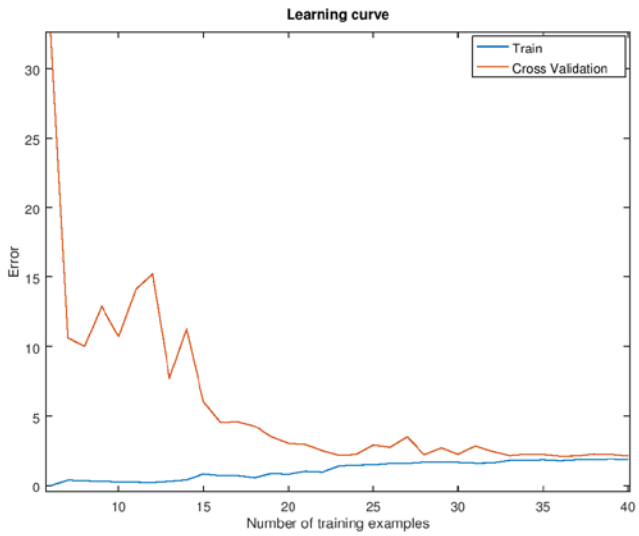

Figure 17 - Learning curve for the ANN based on four selected variables.

We included quadratic terms to search for higher accuracy. Therefore, as Figure 18 shows, the network's first layer contains eight inputs, the four variables mentioned (represented as x) and the same variables at square (x^2). The second layer, also called hidden layer, contains eight data points too ($a^{(1)}$), and finally the output layer contains five classes ($a^{(2)}$), which refer to the four strategies already detailed and the fifth option 'no action needed' for the nodes where there is no need for investment at the time.

The optimal solution for the weights in the network throws an accuracy level of 96% with the training set. After evaluating the classifications through the testing sample, we found that the accuracy is actually around 90%.

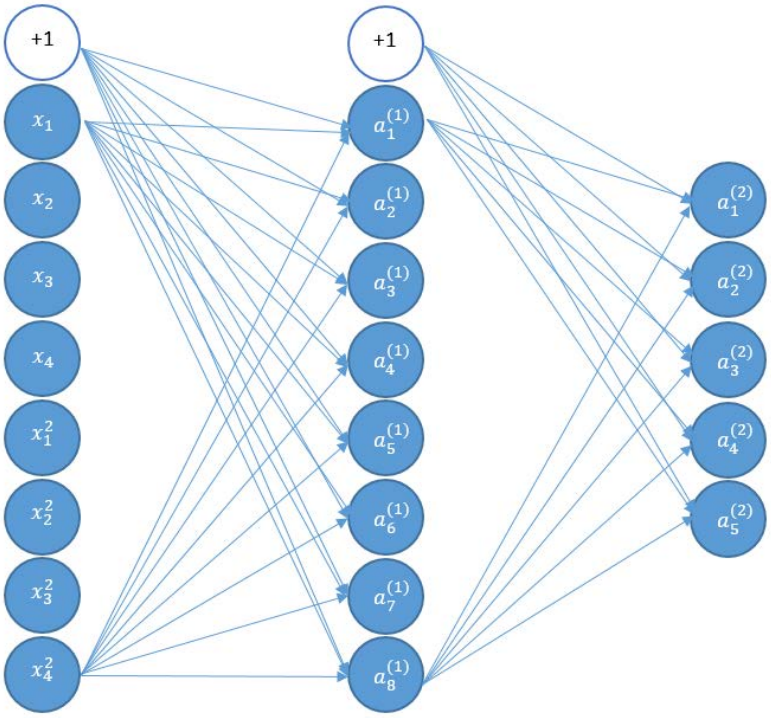

Figure 18 – Our Neural Network scheme.

4. On-going and Future Work

The next step is to incorporate upstream traffic data and improve the ANN accuracy, towards 95% or higher. We will then use it to evaluate the impact of long term actions (in approximately ten years), under different T_{Avg} growth scenarios.

The ANN will be periodically used on all of Cablevisión nodes to guide future investment actions. A Machine Learning tool that involves different scenarios with many variables can be transformed into a very powerful, accurate and scalable resource.

We are using STEM to analyze the impact on a campaign in which we double the speed of access to our customers. It allows us a quick evaluation of the conditions of the nodes and their capacity.

This duplication is complementary, generating a great satisfaction in our customers.

Conclusion

The applications of the technology of Machine Learning have made a very strong advance in the industry of the telecommunications and especially in the Cable industry, providing multiple advantages as detailed above. In addition, it facilitates the operation given the trend of the strong virtualization.

However, we must not forget that the application success is based on the tasks performed by people. In general terms the tasks to be developed to find the learning models are:

- Data: Separate development and validation data. Define instances, classes and attributes.

- Experimentation: Selection of attributes. Performance measures. Cross-validation.

- Validation of the models: Processes intended to verify that models are performing as expected, in line with their design objectives and business uses. It's the most important step in the model building sequence.

In short, applications developed with machine learning technologies are based on human art and science.

Abbreviations

ANN	artificial neural network
Avg	average
BW	bandwidth
CAGR	compound annual growth rate
CM	cable modem
CMTS	cable modem termination system
CVA	Cablevisión S.A.
DOCSIS	data over cable service interface specification
EDA	exploratory data analysis
GHz	giga hertz
HFC	hybrid fiber coaxial
HHP	household passed
IPTV	internet protocol television
Kbps	kilobits per second
Km	kilometers
Mbps	megabits per second
ML	machine learning
OSS/BSS	operation support system/business support system
PCA	principal components analysis
PNM	proactive network maintenance
QAM	quadrature amplitude modulation
QoS	quality of service
SD	standard deviation
SDN	software define network
SG	service group
STEM	Science, Technology, Engineering and Mathematics
Subs	subscribers

Bibliography & References

[1] M. Fiorenzo, C. Righetti, C. Carreño Romano, G. Carro. IP Traffic Analysis: a Tool for Network Dimensioning. SCTE Expo 2016.

[2] Gartner's Hype Cycle for Emerging Technologies, 2016, http://www.gartner.com/newsroom/id/3412017

[3] Arthur Lee Samuel, “Some studies in machine learning using the game of Checkers," IBM Journal of Research and Development 3 (1959).

[4] TM Forum Live , May 15-18 2017, Nice France

[5] Karthik Sundaresan, Nicolas Metts, Greg White, Albert Cabellos-Aparicio, Applications of Machine Learning in Cable Access Networks SPRING TECHNICAL FORUM, CableLabs SCTE NCTA 2016 Spring Technical Forum Proceedings.

[6] Greg White and Karthik Sundaresan, DOCSIS 3.1 Profile Management Application and Algorithms, SCTE NCTA 2016 Spring Technical Forum Proceedings.

[7] “The Data Science Handbook”, Field Cady, 2017, Wiley.

[8] Pitney Bowes (2017), “The Data Differentiator: How Improving Data Quality Improves Business”, Forbes Magazine.

[9] DOCSIS Engineering Professional, SCTE Course Participant Guide, 2014.

[10] Professor Yaser Abu-Mostafa, Caltech's Machine Learning Course - CS 156, Lecture 17 - Three Learning Principles. https://www.youtube.com/watch?v=EZBUDG12Nr0

[11] "Neural Network Primer: Part I" by Maureen Caudill, AI Expert, Feb. 1989.

[12] Machine Learning. Tom Mitchell. 1997. McGraw-Hill.

[13] Daniel Shiffman, The Nature of Code, Chapter 10: Neural Networks, http://natureofcode.com/book/chapter-10-neural-networks/

[14] W.G. Cochran, Sampling Techniques, Chapter 5: stratified random sampling (page 65).

[15] John W. Eaton, David Bateman, Søren Hauberg, Rik Wehbring (2016). GNU Octave version 4.2.0 manual: a high-level interactive language for numerical computations.

