SCTE CABLE-TEC EXPO 23
CELEBRATING 40 YEARS

2023 Fall Technical Forum
SCTE • CABLELABS • NCTA

Machine Learning Model for Customer Claim Prediction in HFC Subscribers

A Technical Paper prepared for SCTE by

Dr. Claudio Righetti
Chief Scientist
Telecom Argentina S.A.
crighetti@teco.com.ar

Matilde Cuenca
Sr Data Scientist Analytics
Telecom Argentina S.A.
crighetti@teco.com.ar

Dr. Diego Martinez Heimann
Tech Scientist Analytics
Telecom Argentina S.A.
crighetti@teco.com.ar

Table of Contents

Title **Page Number**

Abstract 3

Content 3
1. Introduction 3
2. Problem Description and Data Science 4
   2.1. Data sources 5
   2.2. Model 6
3. Prediction Process 8
   3.1. Next Step 9
4. MVP 10
   4.1. Proactive Case Handling 10
5. Model Evaluation – Results and Discussion 11
   5.1. Three-Week Precision 11
   5.2. Perception of Service Degradation 13
   5.3. NPS 14
6. Conclusions and Future Work 15
7. Acknowledgments 15

Abbreviations 16

Bibliography & References 16

List of Figures

Title **Page Number**

Figure 1 - AIOps 5
Figure 2 – ROC curve obtained for a single day prediction in the training phase. 8
Figure 3 – Current data handling tools utilized within the GCP environment. 8
Figure 4 – Proposed configuration for full scalability and automation. 10
Figure 5 – Proactive attention flow designed for the trial MVP. 11
Figure 6 – Three-week model precision calculated for a single week of list emissions. 12
Figure 7 – Model precision measured over the three weeks following the prediction. 13
Figure 8 – Results for the perception of service degradation (P+S) versus not perceived (NP) obtained with the survey. 14

List of Tables

Title **Page Number**

Table 1 – Training dataset for two CM on a two-day lag case 6
Table 2 – Example of daily results for accumulated feature importance results. 9

Abstract

In this technical paper, we propose a machine learning-based approach for predicting customer claims in hybrid fiber coaxial (HFC) subscribers. We are currently finalizing a proof of concept of the model in some areas of our network.

The proposed approach involves collecting data from multiple sources such as network logs, customer service records, hourly collected information from over 3.5 million cable modem (CM) data over cable service interface specification (DOCSIS) and others. We have been able to overcome many technical challenges, two of the most difficult tasks have been dealing with an extremely unbalanced dataset and label noise.

We use several machine learning algorithms, and finally the one selected was extreme gradient boosting (XGBoost) to build the ML model. As a result, we obtain a daily list of customers with high probability to start a claim (95%). From this list, with Customer Experience (CX) and Field Service teams, we can make proactive calls to solve customer problems remotely or to send a technician to their home if necessary.

The proposed approach can help HFC network service providers to proactively identify potential issues in their subscribers' connections and take preventive measures to avoid customer claims that end up in the generation of a technical service ticket. This can lead to improved customer satisfaction, reduced churn rates, and lower operational costs. We measured through surveys how was the experience of our customers and we found that this proactive action has a positive impact on their satisfaction.

Furthermore, the approach can be extended to other types of networks and where predictive maintenance and customer experience management are critical.

Content

1. Introduction

In our networks and services, the artificial intelligence (AI) has the potential to change, the way we operate, and to become the foundation of the transformation that leads to the fourth industrial revolution. But this requires hard work, a long-term commitment, and a deep cultural change.

In the landscape of broadband telecommunications, the DOCSIS protocol has been pivotal for providing high-speed data transfer over existing coaxial cable systems. The widespread adoption of DOCSIS has revolutionized the way internet service providers (ISPs) deliver their services to residential and business customers. However, the complexity of DOCSIS networks, coupled with increasing demand for seamless, high-quality internet service, has posed several challenges for ISPs. One significant area of concern is the predictability of service issues that lead to customer complaints—a critical metric that directly impacts customer satisfaction and churn rates.

To preemptively address the root causes of service issues, proactive network maintenance (PNM) has emerged as an invaluable tool. PNM has been incorporated into DOCSIS since 2005.

Telecom Argentina has been participating in the CableLabs PNM working group since 2013, adopting the best practices proposed. It’s worth noting that for the working group, improving user experience based on PNM data is a constant challenge.

PNM uses advanced diagnostics and analytics to monitor the health and performance of DOCSIS networks. By analyzing metrics such as signal-to-noise ratio (SNR), downstream and upstream power levels, and partial service availability and others, PNM can identify problematic network conditions before they escalate into major outages or other service-affecting issues. However, PNM tools often generate a wealth of data that can be difficult to interpret manually.

In this context, the application of AI techniques offer a promising avenue for leveraging PNM data more effectively. AI can automate the analysis of intricate network behavior, thus providing actionable insights into the future state of network health, including the likelihood of customer complaints.

To create a more holistic view of network health and customer experience, we introduce additional data dimensions such as customer experience index (CEI) HFC, Wi-Fi experience index (WEI) [1] , location, user profile, and modem characteristics (CPU, memory etc.) are increasingly being integrated into the analytical framework. These metrics offer nuanced insights into customer satisfaction and can serve as important predictors for future complaints.

The overarching objective is to shift the paradigm from a reactive customer service model to a proactive one. By leveraging advanced AI algorithms to sift through hourly updates from multiple data sources, our predictive model aims to preemptively identify potential service issues. This proactive approach offers threefold benefits: enhancing customer satisfaction, reducing customer churn, and minimizing operational costs associated with service claims. The paper further details the model's prediction process, its validation techniques, and the successful implementation of a minimum viable product (MVP) to empirically demonstrate its efficacy in proactive customer service.

In the sections that follow, we will delve into the methodologies, technologies, and evaluation metrics that have been instrumental in shaping this ambitious project, highlighting its significance and potential impact on the telecommunications industry.

2. Problem Description and Data Science

In this section, we describe the datasets we use in this MVP. A related work is CableMon [1]. At Telecom Argentina, we employ artificial intelligence for IT operations (AIOps) as our operational framework (Figure 1):

[1] Both the CEI and the WEI are metrics that we have developed at Telecom Argentina. The CEI is an indicator of our customers' experience on the HFC network, while the WEI measures the experience in Wi-Fi usage. It is calculated by taking into account the most important devices in the home, which are defined by their frequency and intensity of use. For these devices, a series of metrics are calculated that show interference, signal strength, the age of the devices, usage of both frequency bands (2.4 GHz and 5 GHz), and packet loss during transmission. These metrics are weighted using a polynomial equation, resulting in a value between 0 and 100, with 100 representing perfect Wi-Fi or HFC network performance.

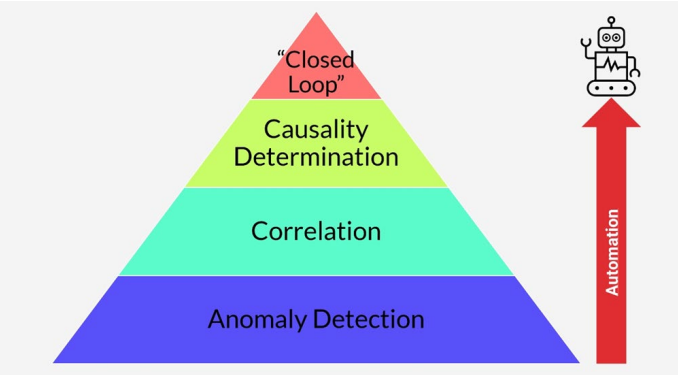

Figure 1 - AIOps

2.1. Data sources

For the training of the predictive model, a collection of anonymized modem-level data is included. The primary source of information considered consists of the proactive network maintenance (PNM) data obtained from the entire spectrum of HFC devices aforementioned. This data is sourced via a universally established standard inherent to the DOCSIS framework. The DOCSIS standard encompasses provisions wherein a cable modem termination system (CMTS) can initiate queries to a DOCSIS-compliant cable modem to retrieve specific performance metrics. These are systematically stored within a local management information base (MIB) in alignment with DOCSIS conventions. Consequently, external processes can harness the simple network management protocol (SNMP) to execute queries on MIBs associated with each CM or CMTS, thereby acquiring comprehensive performance insights [2, 3].

We use this PNM data:

- Timestamp: The time when a PNM data is collected, provided in an hourly basis.
- MAC-ADDRESS: The hashed MAC address of the queried CM.
- SUM_BYTES: Total bytes accumulated in the upstream (US) or downstream (DS) direction.
- AVG_TX_US: Average CM's signal transmission power in the upstream direction.
- AVG_RX: Average received signal power in the CMTS.
- MAC_CER: Maximum package loss percentage.
- MAX_CCER_US: Maximum package correction percentage.
- MIN_SNR: Minimum signal to noise ratio.
- TIMEOUTS: Number of T3 and T4 timeouts the CM has experienced since its last reboot.
- SYSUPTIME: Time since the CM is on.

In addition to the PNM data, we use some other sources of information referred to each individual CM:

- HFC Customer Experience Index (CEI) [2]
- WiFi Experience Index (WEI)
- Location.
- User profile.
- Modem characteristics.

[2] In the CEI, we include additional PNM data.

- Hired service specifications.

Finally, we use as the target of the model, the records of customer tickets generated over the last days which ended in a technical service (TS) being dispatched to the client's location. The relevant fields in each record include the ticket creation time, current status (and closing time if the ticket was resolved), and the distinctive identifier of the customer's account.

2.2. Model

As previously stated, our goal is to predict, on a daily basis, which devices in our HFC network—complying with DOCSIS 3.0 or 3.1 specifications— are likely to generate a customer claim that will escalate to a technical service (TS) assignment within the next few days. Currently, our HFC network comprises approximately 3.5 million devices, which collectively result in a large volume of daily calls to our call center for various reasons. On average, 10% of these interactions escalate to a TS assignment each day. Given the sheer number of devices on our network, the number of such escalations represents less than 0.1% of total daily interactions. This presents us with a highly unbalanced classification problem.

To reduce the volume of information processed, we aggregate the available hourly data into daily values for each individual MAC address. These aggregated daily values include the mean, maximum, minimum, and standard deviation calculated for each of the proactive network maintenance (PNM) data variables.

Given the nature of the available data and the challenges we face, we opted for a technique known as 'Time Series to Supervised Learning.' This approach transforms time-dependent data into a format suitable for conventional supervised learning algorithms. Specifically, it reformulates the time series dataset so that each row represents a unique sample, complete with input features and a target output. This is achieved by generating 'lag features,' which are created based on a predetermined number of previous time steps (or 'lags') to use as predictors for current values.

For each time step, new columns are added to the dataset, each containing the values from prior time steps. This results in a restructured dataset with shifted values, allowing us to apply traditional machine learning algorithms for predictive analysis. This method capitalizes on the inherent temporal dependencies found in time series data, enabling us to employ a wide range of machine learning techniques for forecasting tasks.

An example on how the training dataset looks on a daily basis is shown on Table 1. In this case, a sample representing two CM with a two-day lag is displayed, as if the only predictive feature was SNR.

Table 1 – Training dataset for two CM on a two-day lag case.

ID	date (t-2)	SNR (t-2)	tgt (t-2)	date (t-1)	SNR (t-1)	tgt (t-1)	date (t)	SNR (t)	target
A	20221015	1	0	20221016	2	0	20221017	3	1
A	20221016	2	0	20221017	3	1	20221018	4	1
A	20221017	3	1	20221018	4	1	20221019	5	1
A	20221018	4	1	20221019	5	1	20221020	6	0
B	20221015	7	0	20221016	8	0	20221017	9	0
B	20221016	8	0	20221017	9	0	20221018	10	0
B	20221017	9	0	20221018	10	0	20221019	11	1
B	20221018	10	0	20221019	11	1	20221020	12	1

Features named as previous targets (t-2, t-1) indicate wether the CM had a TS in course on the corresponding date. Then, in the case of the client labeled A, the data shows that a claim was started on

October 17th and remained open until the 19th, when it was closed. For client B, a claim was initiated on October 19th, and it was still open on the last day analized (October 20).

For the prediction dataset we only use one day (file) for each CM, and the target column becomes the model target. In typical conditions, during the active stages of the project (discussed further ahead in the work) we used around 1 million CM for model training, with a time-window of 40 days. In the case of the prediction dataset, we worked with a 300 thousand CM set.The Time Series to Supervised Learning approach has many benefits, some of which include: compatibility with traditional machine learning algorithms, simplification of the feature engineering processes, ease of variable interpretability and feature importance analysis, among others. In our case we selected a time window of five days for the lag duration, and periodically retrained our model coincident with that time.

In terms of model selection, after several tests the choice was made for an XGBoost model, which is known to be particularly well suited for imbalanced classification problems due to several intrinsic characteristics, among which can be mentioned that it is a strong ensemble model that makes use of regularization techniques, especially important in unbalanced datasets, consistent handling of missing values, its robustness to outliers due to the nature of decision trees, etc.

For hyperparameter tuning, we employed a time series cross validation (TSCV) approach featuring a 5-fold split and a 5-day lag gap to offset the temporal dependence introduced by our time series to supervised transformation. This setup was designed to closely emulate the intended predictive application. The optimal hyperparameters we identified are as follows: scale_pos_weight set to 740, **reg_alpha** at 0.004, **n_estimators** at 900**, max_depth** at 4, and **learning_rate** at 0.031.

Traditionally, accuracy has been the default metric for evaluating classification models. However, in cases of class imbalance, relying solely on accuracy can be misleading. The model may achieve high accuracy simply by predicting the majority class, which is generally not the desired outcome in such scenarios. To mitigate this, we selected recall as our refitting parameter. Recall is particularly useful for focusing on the model's ability to correctly identify positive instances, which is often the more critical objective when one class is significantly underrepresented.

Our model's performance was assessed under a typical training scenario, yielding a receiver operating characteristic - area under the curve (ROC-AUC) score of 0.97, as indicated in Figure 2. Additionally, we achieved a precision of 0.11 and an F1-score of 0.13 when using a 90% probability threshold and a one-day prediction window. It's worth noting that the duration of the prediction window emerges as a significant factor in evaluating precision outcomes, a point we will explore further in subsequent discussions.

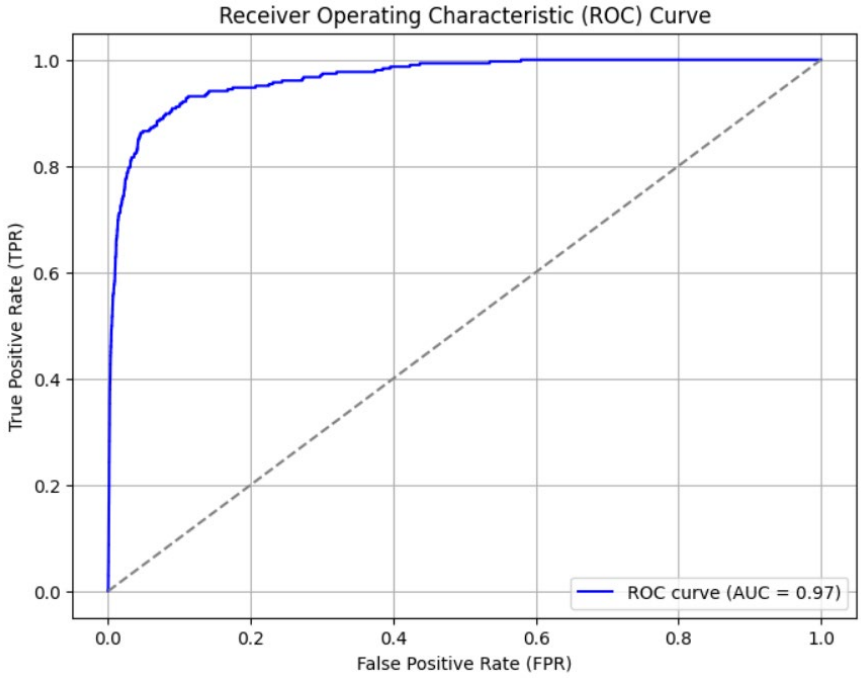

Figure 2 – ROC curve obtained for a single day prediction in the training phase.

3. Prediction Process

This MVP is entirely hosted on the Google Cloud Platform (GCP). As previously mentioned, various data sources—including technical metrics from cable modems, customer information, claims history, and both WiFi and Customer Experience Indexes—are imported into Google Cloud Storage from our on-premise data lake. The data is then processed using Google Cloud Functions and stored in BigQuery for streamlined access. For the analytics component, we leverage Vertex Notebooks to perform daily predictive analyses as well as weekly model training.

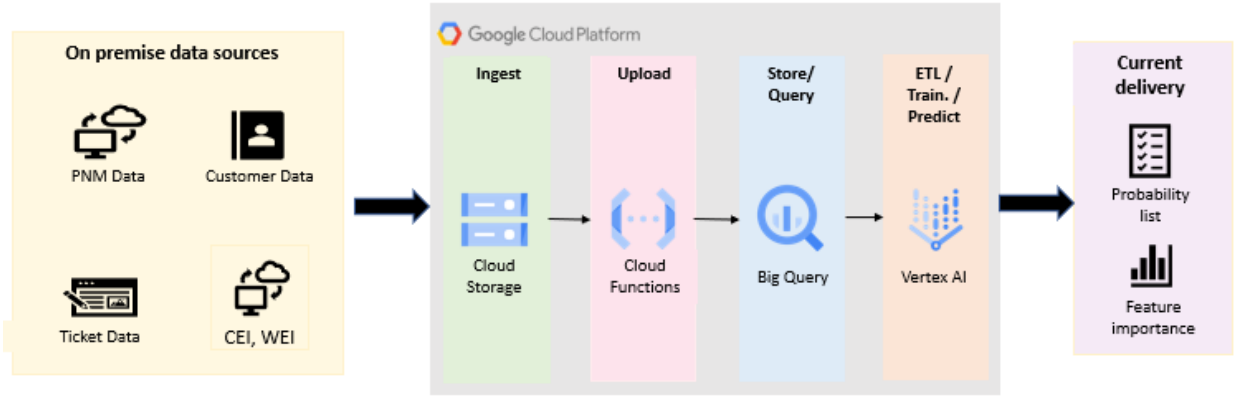

Figure 3 – Current data handling tools utilized within the GCP environment.

Through this configuration, we were able to perform a weekly training over a portion of the entire available data comprising the information of 1M CM for time period of approximately 40 days (of the 3.5M HFC devices available). On the other hand, given the specific characteristics of the experimental procedure (discussed in the following section), the daily predictive process was performed over a set of 300k CM with a lag time window of five days.

Table 2 – Example of daily results for accumulated feature importance results.

Feature	Description	Weight
STDpnm_data.SYSUPTIME_t	SYSUPTIME standard deviation for the predicted day (t)	187
week_day_number_t	Day of the week in a [0-6] range	137
WEI	Wi-fi experience index	127
CEI	Customer experience index	125
MINpnm_data.SYSUPTIME_t	Minimum SYSUPTIME value recorded during the predicted day (t)	125
AVG_CEI_by_node	Mean CEI value for the CM node over past week	104
MINpnm_data.SUM_BYTES_DS_t	Minimum downstream SUM_BYTES for the predicted day (t)	101
STDpnm_data.AVG_RX_US_t	Standard deviation for the upstream AVG_RX value for the predicted day (t)	96
STDpnm_data.SYSUPTIME_t2	SYSUPTIME standard deviation for two days before the predicted day (t2)	90
MINpnm_data.SNR_US_t4	Minimum upstream SNR value for four days before the predicted day (t4)	85

The daily output of the model consisted of the list of customers with the highest calculated probabilities of starting a claim that would eventually end in a TS. The length of the list, or equivalently, the probability threshold was determined by the amount of calls that could be made on a daily basis by the field service teams dedicated to the project. Along with this list, we made use of the Shapley additive explanations (SHAP) library [4] to calculate and add the information of the most relevant features for each prediction individually; this information was further analyzed in order to try to understand the origin of each perceived service inconvenient and develop automated strategies for proactive attention, as will be discussed in the following sections. An example of the results obtained for this feature in an accumulated case for a single prediction day are shown in Table 2. The final index in some feature names (t, t2, t4) indicates whether each variable corresponds to 0 to 4 days prior to the prediction date, according to the time series to supervised learning approach.

3.1. Next Step

To achieve scalability across the entire HFC network, the project is planning a transition from the current extract, transform, load (ETL) stage to a Google Composer-based process. This new stage will leverage the data build tool (DBT) library [5] to handle raw data using standard SQL queries in BigQuery, a shift that is expected to significantly enhance both performance and resource efficiency. Additionally, the existing train/predict operations are being converted into automated workflows through Vertex Pipelines. This automation will streamline the end-to-end process, from model training and prediction to final output. The architecture of this proposed system is illustrated in the figure 4.

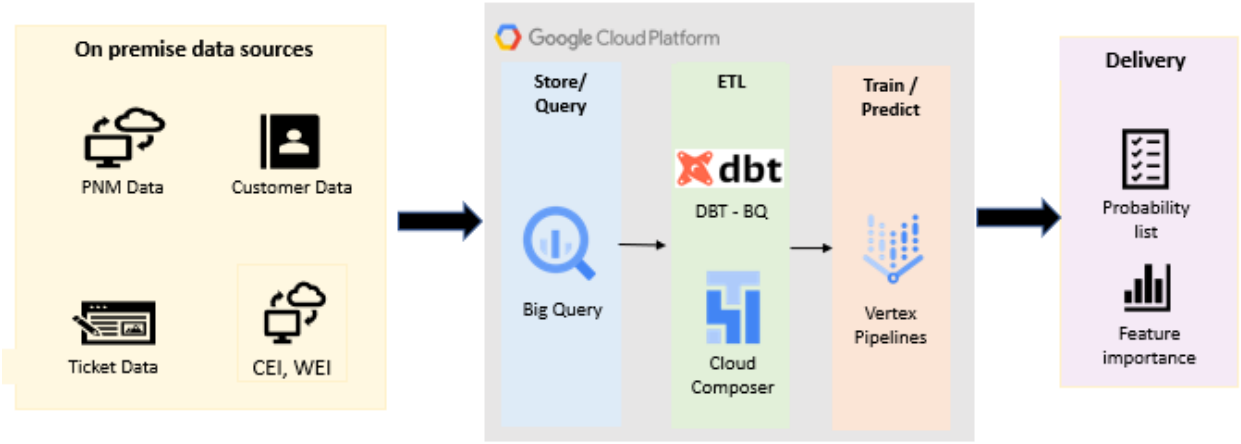

Figure 4 – Proposed configuration for full scalability and automation.

4. MVP

To deepen our understanding and optimization of the MVP process, we developed a proactive attention protocol in collaboration with multiple internal teams, including Quality Management, Customer – Technical Support, Field Service – Selected Technical Bases, Customer Voice, Service, and Field Services. Through iterative refinement of the process, we investigated and identified the most effective procedures, culminating in the development of a final filtering and attention protocol. The key conditions established for the targeted customer segments were:

- Not have an opened or closed claim in the 5 days previous to the listing.
- Not have an ongoing ticket.
- Have a calculated probability higher than 80%
- Not have a massive service interruption reported.
- Have schedule availability from the FS for the following 24/48 hours after the call.

Regarding this last restriction, it was found during the early project stages that if a proactive call is made to a customer which cannot be responded to in a short time period, this tends to be more harmful to the customer perceived experience than not taking action at all.

4.1. Proactive Case Handling

Having established both the daily outcome and the attention process, a trial was performed in a group of selected geographic regions where the service availability was enough to assure dispatch in one or two days after the communication.

For this trial minimum value product (MVP) a subset of approximately 300 thousand CM was established for which the model was weekly trained and daily predictions were made. Following the lessons learned in early stages, a process was implemented for the proactive attention where each customer was contacted according to a specifically designed flow, shown in Figure 5.

First, the prediction list was swept via phone calls made by four Customer Service (CS) representatives. In this step, the average sweep capacity was around of 30% and the effective contact was 30% of that amount. Next the customer was asked if he had perceived a degradation (PD) of the service in the 72 hours prior to the call. For the affirmative answers (excluding the negative ones and the ones that had had a problem but this was solved independently of our intervention), the customer was inquired on whether he was willing to accept the proactive management of the problem.

The customer was then asked if he was available to perform a series of checks and procedures to further assess the service condition, and lastly, a remote diagnosis was performed via automated tools in order to establish if the interaction was due to finalize with the dispatch of a FS or, in certain conditions, the problem was able to be fixed via remote solutions conducted by our CS teams.

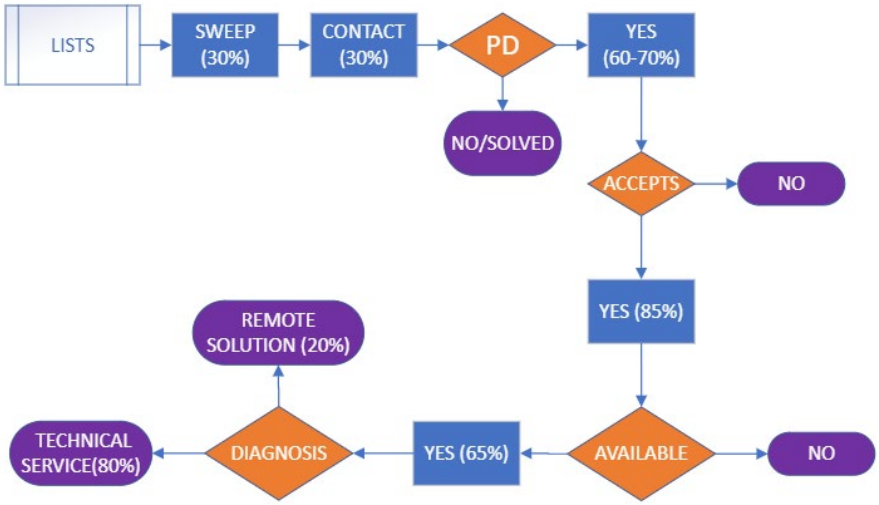

Figure 5 – Proactive attention flow designed for the trial MVP.

With the objective of making the most profit from the customer interaction as possible, a set of ad hoc procedures were developed for the proactive case handling. For example, in cases where the customer was aware of a service difficulty, but he/she was unavailable, or didn't want a home visit, and under certain configurations of the predictive features that were understood to be highly linked to a service dispatch, the technical visit was automatically arranged or the remote solutions available were applied. Some of the key results obtained during this trial are discussed in the following section.

5. Model Evaluation – Results and Discussion

This section focuses on the evaluation of the model's performance and ensuing discussions. To evaluate the MVP, we defined an *ad-hoc precision*.

5.1. Three-Week Precision

During the evaluation phase of our model, we observed that clients frequently did not initiate contact immediately the day after a prediction was made; rather, they often reached out several days or even weeks later. Another noticeable trend was that repeated predictions for a client within a single week increased the likelihood of that client eventually filing a claim, thereby improving the model's predictive accuracy for that particular group. Given these observations and the importance of minimizing unnecessary technician dispatches—which represent a waste of both time and resources—'custom' precision emerges as a

particularly relevant metric for assessing the model's effectiveness. Consequently, precision has become one of our key performance indicators for model evaluation.

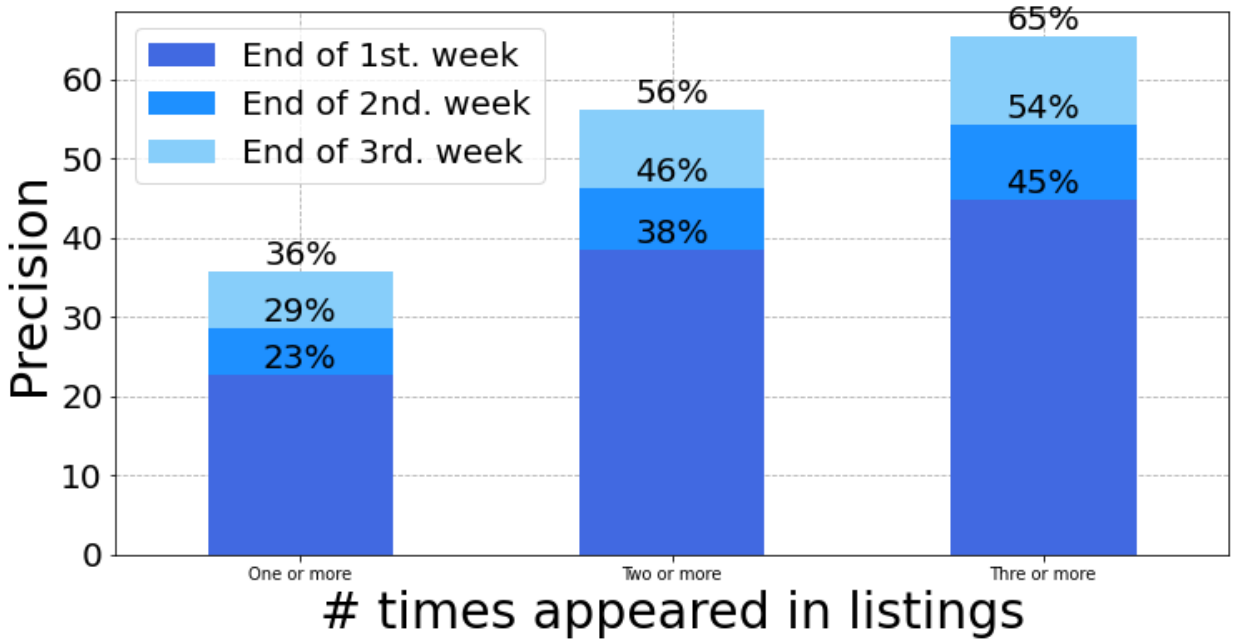

Figure 6 – Three-week model precision calculated for a single week of list emissions.

To account for the inherent delays in customers' claims reporting, our evaluation includes communications that take place up to three weeks after the initial prediction. As a result, a custom precision metric has been developed. Moreover, to mitigate any potential biases arising from proactive outreach initiatives, precision is assessed solely within a control group.

True positive cases are classified as those where clients file a claim subsequent to a prediction. These cases are further categorized based on the time elapsed between the prediction and the actual filing of the claim. We also differentiate among cases based on the frequency of predictions for the same client within a given week—whether they were predicted once, twice, or multiple times within the same week, as shown in Figure 6.

Our data show that while the majority of customer communications occur within a week, a significant proportion take place after one or two weeks. This observed increase in precision among clients who were predicted multiple times aligns with our earlier observations.

To simplify the evaluation metric into a single, unified value, we chose to define our 'custom' precision as the ratio of customers who, having appeared at least twice on our weekly prediction lists, initiated a call leading to a technical Service (TS) assignment within the three weeks following their inclusion on the list. Results for this tailored metric are illustrated in Figure 7. Here, we display the temporal evolution of precision, broken down by the minimum probability thresholds used to filter customers onto the lists. For instance, the curve corresponding to a 0.85 threshold includes only cases where the calculated probability exceeds 85%. It's readily apparent that higher thresholds correspond to increased levels of precision.

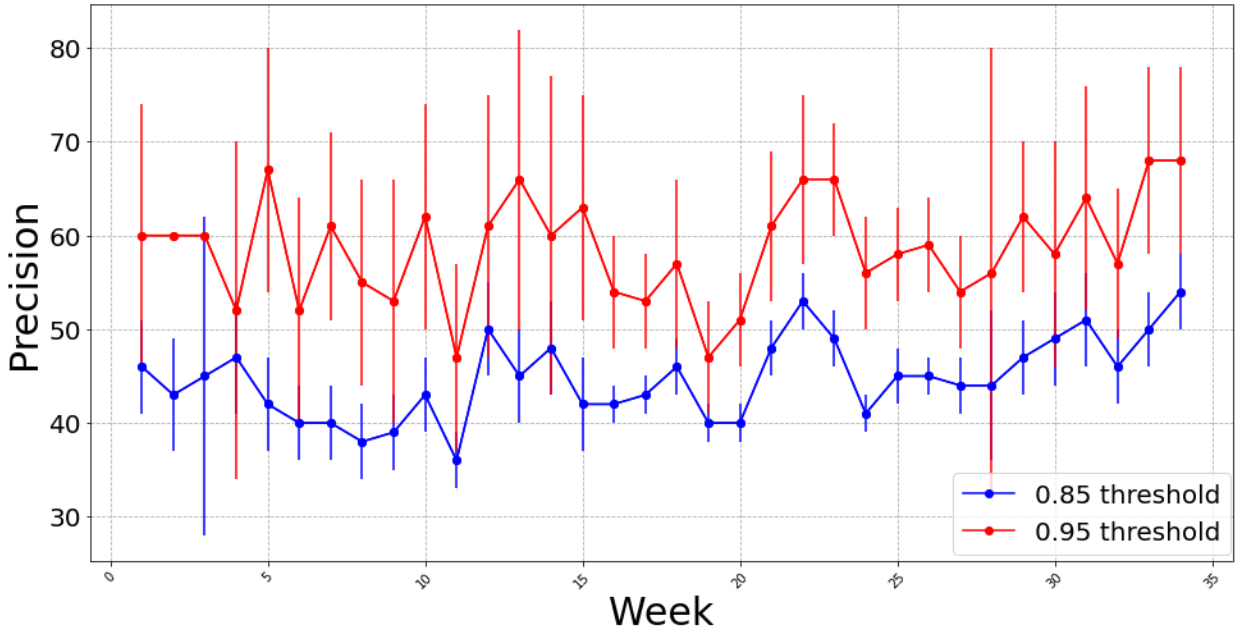

Figure 7 – Model precision measured over the three weeks following the prediction.

The data reveals that, on average, for a threshold of 0.95, our model achieves a precision of approximately 60%. This means that 6 out of 10 predicted customers do indeed file a claim in the weeks following their appearance on the prediction lists. Considering that, as will be further discussed, not every customer initiates a claim, we believe this result underscores the reliability of our model. Additionally, we conducted experiments simulating random customer lists and analyzed the percentage of received calls. We found that for a random sample, the precision falls within the range of 5-10%, which reinforces our initial impression regarding the model's effectiveness.

5.2. Perception of Service Degradation

Given that many clients experience difficulty in establishing contact due to data center congestion, the study augments the precision evaluation of contacted customers by assessing their perception of service malfunctions, indicated as the percentage of perceived malfunctions among those contacted (perceived degradation of service, P), data shown if Figure 8. The value also includes within the perceived curve, those cases where the customer reported to have noticed a malfunction in the service but this was automatically solved in the days passed between the event and the communication (solved cases, S). Finally, the orange curve in the figure represents those cases where the contacted client did not feel that the service had presented any inconvenience in recent times (Not Perceived, NP). It general, it can be seen that the predictive model, after some adjustments made midway in terms of parameter tuning and other refinements, is able to correctly identify between 6 and 7 out of 10clients that are feeling a problem in the service provided.

As previously indicated, the MVP incorporates solicitation of client feedback through surveys, with the aim of validating not only the existence of service-related concerns from their perspective but also their inclination to initiate claims. A comparative analysis between precision and perception underscores that the latter consistently exhibits a marginally higher value than the former. One of the relevant questions within the survey is the inquiry into customers' predisposition to initiate claims following the perception of service degradation. This aspect affords insight into the distinction between precision and the subjective recognition of degradation. It emerges that a considerable number of clients, around 65% of the queried population,

express disinclination towards pursuing claims, citing their awareness of the formidable challenges entailed in establishing effective communication with the data center. This can give us an upper threshold for the precision because even if a customer is perceiving a malfunction, he may not be willing to initiate a claim.

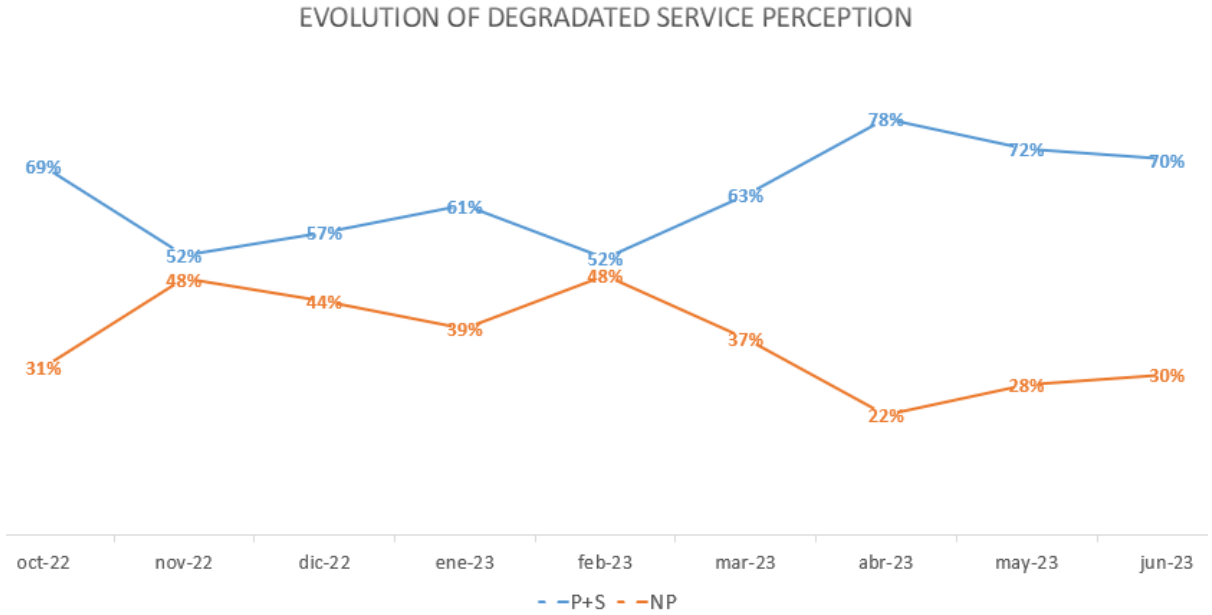

Figure 8 – Results for the perception of service degradation (P+S) versus not perceived (NP) obtained with the survey.

5.3. NPS

Another focal point of our MVP investigation was the impact of proactive customer engagement on customer experience, as quantified through the net promoter score (NPS). We designed a specialized survey to compare the NPS changes between proactive service and traditional customer service channels. This involved crafting precise questionnaires and establishing control groups for a rigorous evaluation.

Our key findings indicate that proactive customer engagement leads to a modest improvement in problem resolution rates when the case is resolved through technical service (TS) — 44% versus 41% for traditional channels. However, the resolution rates jump significantly to 80% when the issue is addressed through one of the various remote solutions evaluated during the MVP phase.

Complementing this is our second observation: the NPS scores generally rise when using proactive engagement channels. Specifically, the NPS value shifts from -9 in reactive channels to 1 in proactive ones. This improvement is even more pronounced in cases resolved remotely, which have an NPS score of 20, compared to an NPS score of -2 for cases requiring a physical visit.

6. Conclusions and Future Work

Our predictive model successfully generates a daily list of customers at high risk of filing claims, each accompanied by pertinent variables. This information empowers our Customer Service and Field Service teams to proactively mitigate potential issues. Through remote problem resolution and targeted technician deployment, we have made strides in boosting customer satisfaction. This is corroborated by our survey results and NPS metrics, which show a marked improvement in customer satisfaction, particularly when issues are resolved remotely.

The MVP stage has provided compelling evidence of the model's effectiveness in preemptively identifying service disruptions and elevating customer satisfaction levels. Moreover, the model's scalability is promising, especially with the automation of the ETL process. The Future Work are:

Algorithm Refinement: There is room for fine-tuning the predictive algorithm to enhance its precision. This includes the integration of additional data sources to augment the model's accuracy.

Root Cause Analysis: Preliminary observations indicate that the model's feature importance calculations could serve as a valuable tool for root cause analysis in early diagnostic procedures, especially when combined with Field Service (FS) reports.

Remote Procedures: We've initiated research into other remote interventions, such as conditional CM resets. These shows promise for improving network maintenance, but more comprehensive studies are needed for conclusive results.

Automated Customer Outreach: A promising avenue for future exploration is the automation of customer interactions through digital platforms. This could exponentially increase the volume of data collected and thereby enhance the statistical significance of our findings.

7. Acknowledgments

We would like to express our sincere gratitude to all those who have contributed to the success of this MVP: Myriam Ramirez, Marcos Avalo, Silvana Regules Silva, Sebastian Seijas, Claudio Miguel Ambrogio, Matias Miguel Parisi, Federico Gomez, Alicia Frassia, Paola A Fusari, Mariela Fiorenzo, Martin Juiz, Leonel Gonzalez Biot and Bruno Pallotta for their tireless efforts, collaboration and commitment to innovation.

Abbreviations

CMTS	cable modem termination system
HFC	hybrid fiber coaxial
FN	fiber node
RF	radio frequency
ISP	internet service provider
DOCSIS	data over cable service interface specification
PNM	proactive network maintenance
CM	cable modem
SNMP	simple network management protocol
MIB	management information base
CEI	customer experience index
WEI	Wifi experience index
TS	technical service
FS	field service
ETL	extract, transform and load
MVP	minimum viable product
CS	customer service
NPS	net promoter score
SHAP	Shapley additive explanations
AIOps	artificial intelligence for IT operations
XGBoost	extreme gradient boosting

Bibliography & References

[1] *CableMon: Improving the reliability of cable broadband networks via proactive network maintenance*, J. Hu et al.; Proceedings of the 17th USENIX Symposium on Networked Systems Design and Implementation (NSDI ´20).

[2] *DOCSIS CableLabs;* Best Practices and Guidelines, PNM Best Practices: HFC Networks (DOCSIS 3.0); Technical report, CM-GL-PNMP-V03-160725, 2016.

[3] *Management Information Base for Data Over Cable Service Interface Specification (DOCSIS)*; W Sawyer; Cable Modem Termination Systems for Subscriber Management. RFC 4036, 2005.

[4] Available at https://shap.readthedocs.io/en/latest/#

[5] Available at https://www.getdbt.com/

