

MGI

Master Degree Program in
Information Management

Churn Prediction in Online Newspaper Subscriptions

Lúcia Madeira Belchior

Project Work

presented as partial requirement for obtaining the Master Degree Program in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

CHURN PREDICTION IN ONLINE NEWSPAPER SUBSCRIPTIONS

By

Lúcia Madeira Belchior

Project Work presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence.

Co-Supervisor: Nuno Miguel da Conceição António

Co-Supervisor: Elizabeth Silva Fernandes

November 2022

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Faro, 2022-11-11

ABSTRACT

In today's connected world, people turn more to the web to be informed of the news. Newspapers in an online environment profit from employing subscription models. Despite that Portugal remains one of the countries with higher levels of trust in news, readers present a low propensity to subscribe. Hence, online newspapers' existing customers are a valuable asset. Therefore, it is in the best interest of such businesses to monitor these customers to identify potential churn signs down the line.

Customer churn prediction models aim to identify customers most prone to attrite, allowing businesses that leverage them to improve their customer retention campaigns' efficiency and reduce costs associated with churn. Two different research approaches, namely prediction power and comprehensibility, have been at the core of churn prediction literature. Businesses need accurate models to target customers' right subset. However, many models are black-boxes and present reduced interpretability. On the other hand, understanding what drives customers to churn can support managers in making better-informed decisions.

This project report presents the development of a plan to tackle churn prediction in a Portuguese newspaper with an online subscription model using Machine Learning methods. The models' performance was evaluated in two experiments. One experiment assessed the performance for all types of subscriptions and another considered only non-recurring subscriptions. The results of the first experiment were tempered by an unplanned marketing campaign that run simultaneously with the experiment on top of the contrasting contexts in which the model was trained and evaluated. On the other hand, the second experiment's results suggest that for non-recurring subscriptions, a phone call from the call centre proved to be an adequate retention measure for probable churning subscribers.

Additionally, models' predictors were analysed and it was found that users with lower fidelity rates and few subscriptions present a higher propensity to cancel their subscriptions. The same occurs with users whose product is annual or longer-lasting. These findings shed light on how to minimize churn and improve reader engagement. Based on the models' results, and predictors' analysis, the newspaper decided to implement a re-engagement newsletter to keep users engaged and prevent future churn.

KEYWORDS

Churn Prediction; Classification; Data Mining; Interpretability; Online Subscriptions; Machine Learning

INDEX

1. Introduction.....	1
1.1. Background.....	1
1.2. PÚBLICO - Comunicação Social, SA.....	2
1.3. Project objectives	2
2. Literature review	4
2.1. Customer Churn.....	4
2.2. Customer Churn Prediction	4
2.3. Churn in the News Media Sector.....	6
2.4. Balancing Techniques	7
2.5. Interpretability in Machine Learning.....	7
3. Methodology	9
3.1. Business Understanding	10
3.2. Data Understanding	10
3.3. Data Preparation	13
3.3.1. Data Cleaning.....	13
3.3.2. Data Transformation	14
3.3.3. Data Reduction	15
3.4. Modelling.....	16
3.5. Evaluation	18
3.5.1. Models Selection	18
3.5.2. Models Evaluation	18
4. Results and Discussion.....	20
4.1. XGBoost Model.....	20
4.2. Model Interpretability	21
4.3. A/B Tests.....	24
4.3.1. Model 1 – Recurrent and Non-recurrent Subscriptions.....	24
4.3.2. Model 2 – Non-recurrent Subscriptions.....	25
5. Conclusion	28
References.....	30

LIST OF FIGURES

Figure 1.1 – International online news payment comparison with Portugal	1
Figure 3.1 – CRISP-DM process model	9
Figure 3.2 – Project framework.....	9
Figure 3.3 – Periods covered by the different kinds of variables exemplified	11
Figure 3.4 – Data cleaning stages.....	14
Figure 3.5 – Daily predictions pipeline	19
Figure 4.1 – Predictive features ranking by SHAP values.....	22
Figure 4.2 – Dependency plot for Days with an active subscription (%)	23
Figure 4.3 – Dependency plot for Days since 1st subscription	23
Figure 4.4 – Dependency plot for Total Days With Visits (30D+60D)	24
Figure 4.5 – Stringency Index 30D ranges for the three datasets.....	25

LIST OF TABLES

Table 3.1 – Raw datasets for each model	10
Table 3.2 – Predictors' meaning and periods which they cover	12
Table 3.3 – Features selected for each model	16
Table 3.4 – Datasets used for model selection	17
Table 3.5 – Parameters considered for grid search cross-validation per model	17
Table 4.1 – Evaluation metrics from models trained with original and balanced data for both modelling contexts	20
Table 4.2 – Evaluation metrics extracted from XGBoost models selected	21
Table 4.3 – Results of the experiment on only non-recurrent subscriptions	26

LIST OF ABBREVIATIONS AND ACRONYMS

AUC	Area Under the receiver operating characteristic Curve
CRM	Customer Relationship Management
DM	Data Mining
DT	Decision Tree
FT	Financial Times
F1	F1 Score
GBM	Gradient Boosting Machine
GDPR	General Data Protection Regulation
LightGBM	Light Gradient Boosting Machine
LR	Logistic Regression
ML	Machine Learning
NN	Neural Network
RF	Random Forest
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting

1. INTRODUCTION

1.1. BACKGROUND

News media traditionally based their business models on a combination of sales and advertising revenue. However, with the economy still recovering from the effects of the 2008 crisis (Arrese, 2016) and the recent Covid-19 pandemic lockdowns, on top of a continuous rise of digital media and a decades-long downwards trend in print newspaper readership, news media have had to re-evaluate their business models (Newman et al., 2021). In order to diminish and offset the obstacles caused by the previously mentioned difficulties, newspaper companies are focusing on digital information business models (Cardoso, Paisana, & Pinto-Martinho, 2020).

Online newspapers' main sources of revenue are advertisements and online subscriptions mainly in the form of paywall models (Pattabhiramaiah, Sriram, & Manchanda, 2018). Nonetheless, the first does not create a long-lasting relationship with the users and can be avoided with the use of ad blockers. Hence, the latter has been adopted by many publishers to increase revenue through subscriptions (Fletcher & Nielsen, 2017). Paywall models, based on content consumption, broadly fall into two categories: “soft” and “hard”, i.e., from some free articles to no article access (Arrese, 2016). Between both, there are the “freemium” paywalls, the most used by the industry, that provide a certain quantity of free articles before inviting the reader to buy a subscription. Moreover, recent studies present new paywall mechanisms based on data science algorithms to define the optimal paywall time in order to increase propensity to subscribe (Davoudi, An, Zihayat, & Edall, 2018).

In fact, from 2019 to 2020, there was a generalised increase in people paying for online news (see Figure 1.1). This trend was also verified in Portugal, reaching 10% of the population affirming to pay for online news, most of them with an ongoing subscription (Cardoso et al., 2020). Nevertheless, Portugal is still characterised by a strongly traditional press industry, where revenue depends highly on the physical distribution of newspapers and, even though digital media subscribers are increasing, the figures are small, which threatens this news ecosystem significantly trusted by the Portuguese. In addition to this, a wide variety of accessible options for readers pose a challenge to the newspaper business on digital platforms.



Figure 1.1 – International online news payment comparison with Portugal
(Cardoso, Paisana, & Pinto-Martinho, 2020)

Thus, subscribers' acquisition and retention are the main focus of publishers (Rußell, Berger, Stich, Hess, & Spann, 2020). As argued by Kotler et al. (2017), across the reader funnel, publishers should map the reader path in order to improve critical touchpoints. Considering the aforementioned environment, digital news media, much like many other businesses, has been increasingly acknowledging their extant database as their most valuable asset (Coussement & Van den Poel, 2008), and engaging in the process of Data Mining (DM) as part of an overall strategy to improve business intelligence, Customer Relationship Management (CRM) and customer churn prevention (Gunnarsson, Walker, Walatka, & Swann, 2007).

1.2. PÚBLICO - COMUNICAÇÃO SOCIAL, SA

Founded in 1990, PÚBLICO - Comunicação Social, SA (hereafter referred to as PÚBLICO), is one of the leading Portuguese newspapers and was the first newspaper in Portugal to have an online edition. With the outburst of the internet, people are turning more to the web and social media to be informed of the news. Over the past years, PÚBLICO has been facing a continuous change in its digital model and has increased its focus on digital platforms, following the developments of international digital media companies (Kueng, 2017). Journalists have been improving their digital skills to offer content prepared for the online platform to increase readers' engagement by increasing, for example, the number of website visits and recirculation on the website. The adoption of a paywall mechanism by the newspaper, alongside exclusive subscriber content and the offer of a large set of discounts and experiences, are some of the approaches that have been encouraging users to convert. The results of these actions are mirrored by their growing subscriber's database, which during the first semester of 2020 more than doubled, compared with the homologous period of the previous year.

With an increasing subscriber database and keeping in mind that attracting new customers is more costly than retaining existing ones (Richter, Yom-Tov, & Slonim, 2010), PÚBLICO knows it can leverage its existing database to implement a prediction model that can help distinguish customers' probable tendency to switch to competitors, thus allowing targeted proactive measures to prevent this from happening.

1.3. PROJECT OBJECTIVES

PÚBLICO's Data Analytics team developed a Logistic Regression (LR) model to identify the main churn factors. Nonetheless, the low number of attributes considered and the low predictive performance induced the need for a model with higher predictive performance, but still, a high degree of interpretability.

This project's first and foremost objective was to develop a churn prediction model, with high predictive performance, to deal with subscriber churn. The application and use of a predictive model can provide analytical support for the organisation to tackle customer attrition proactively, rather than addressing it reactively. To achieve this objective two different models were built to predict churn in different contexts, namely: (1) predicting churn for both automatically and non-automatically renewed subscriptions and (2) predicting churn only for non-automatically renewed subscriptions. The models were sequentially evaluated with two experiments, where daily predictions were performed for a period of three weeks for model 1 and four weeks for model 2.

The interpretation of both models was a complementary goal within this project's development to deliver a valuable solution for the business. Both models' predictors were studied. These predictors' interpretations allowed the understanding of the models and to conclude whether it would be possible to act upon churn drivers.

DM techniques were relied on to tackle the former problem by analysing and preparing the data provided, using it to create several Machine Learning (ML) models and evaluating them to select which one presented the best predictive performance to cope with customer churn. The latter objective was addressed through the analysis of the outputs of the model resourcing to the SHAP values (Lundberg & Lee, 2017) python package.

The remainder of this project report is structured as follows. First, a review of the literature on customer churn and the need for employing predictive models, as well as an overview of studies regarding churn in the media, are present from section 2.1 to 2.3. Balancing techniques for uneven target class distribution and the concept of interpretability in ML with the growing need to address it are outlined in sections 2.4 and 2.5. In section 3, the methodology used is described, and each step of the process is detailed in its subsections. Also, the models' selection and their respective evaluation are present in section 3.5. Section 4 contains the results and, finally, section 5 contains the conclusion of the project.

2. LITERATURE REVIEW

In this section, the meaning and need for churn predictions are addressed, as well as previous studies in customer churn prediction, with a special focus on the media sector. Finally, balancing techniques and the need for interpretation in ML are addressed.

2.1. CUSTOMER CHURN

Customer churn, also identified as customer attrition, turnover or defection, refers to the loss of customers. It is a very well-studied topic in literature and not a novelty for any industry. Studies on the matter have been developed in telecommunication (Coussement, Lessmann, & Verstraeten, 2017; Keramati et al., 2014; Verbeke, Dejaeger, Martens, Hur, & Baesens, 2012), banking (Farquad, Ravi, & Raju, 2014), gambling (Coussement & Bock, 2013), newspapers (Coussement & Van den Poel, 2008) and many other industries (De Caigny, Coussement, & De Bock, 2018).

The literature mention two types of churn: incidental and deliberate (Jayaswal, Prasad, Tomar, & Agarwal, 2016). Incidental churn occurs when customers are forced to end their subscription for some underlying reason (e.g., faulty payments). In contrast, the process of a customer voluntarily deciding to cancel their subscription falls under the deliberate churn definition.

When predicting customer churn, a scoring model estimates the future churn probability for every customer based on their historical data (De Caigny et al., 2018). The use of a predictive model to detect customers' likelihood to terminate their relationship with a company has been proven to be of great importance as a research discipline since acquiring new customers is several times more expensive than retaining existing ones for numerous reasons, namely:

1. A small change in the churn rate can lead to a significant increase in profitability (Coussement & Van den Poel, 2009);
2. Long-term customers become less costly to serve, as the company already has information about them and can understand their needs, as well as be more profitable, because they tend to make more purchases and provide positive referrals through word of mouth (Ganesh, Arnold, & Reynolds, 2000);
3. Customers who churn can persuade other customers within their social network to act alike (Nitzan & Libai, 2011);
4. By losing customers, the need to attract new consumers and costs associated with it increase, whilst profits decrease due to lost opportunities in sales (De Caigny et al., 2018).

2.2. CUSTOMER CHURN PREDICTION

Customer churn prediction has been tackled from two different angles in previous research: predictive performance and comprehensibility (Tianyuan & Moro, 2021). To effectively manage customer churn within a company, it is crucial to build an effective and accurate customer churn model (Coussement & Van den Poel, 2008). Researchers focused on improving customer churn prediction models develop and propose more complex models (De Caigny et al., 2018). Nevertheless, there is no consensus on the performance of churn prediction modelling techniques, given that results differ between industries as well as DM techniques applied (Verbeke, Martens, Mues, & Baesens, 2011). In contrast, researchers want to understand what drives customer churn. Therefore, models that provide actionability are

essential, since researchers with a better understanding of churn can help managers make better-informed decisions in combating churn (De Caigny et al., 2018).

Studies on customer churn prediction in the media sector have been developed (Burez & Van den Poel, 2007; Yuan, Bai, Song, & Zhou, 2017). However, there is a gap in the literature when it comes to digital newspaper subscriptions, as most publishers do not wish to disclose their churn prediction strategies. Other industries, such as the telecommunication sector present a larger range of scientific literature on the topic (Coussement et al., 2017; Keramati et al., 2014).

Coussement and Van den Poel (2008) used Support Vector Machines (SVM) algorithm to predict churn in a print newspaper subscription context. Using grid search with cross-validation on a balanced training dataset, the authors concluded that selecting the model's parameters based on the cross-validated model with the highest Area Under the receiver operating characteristic Curve (AUC), instead of the highest accuracy, resulted in significant improvements in AUC and Top Decile Lift, when the final model was evaluated against a test set with the original churn distribution. Nevertheless, the SVM model is outperformed by a Random Forest (RF) algorithm. Moreover, from the RF feature importance measure, the ten most important predictors were extracted. The subscription's length and the elapsed time since the last renewal are the most important predictors. In addition, client-company interaction data was important for prediction and should be considered. On the other hand, monetary value, frequency of renewals and socio-demographic variables, except for age, were not present in the most important variables.

More recently, MittMedia (Ekholm, 2019) has also addressed the issue of churn prediction by developing three models (Survival Analysis, Gradient Boost Machine (GBM) and Neural Network (NN)) that were supplied with user and behavioural data. The use of several models allowed the authors to compare results and verify models' contradictions. A factor analysis was carried out to perform feature selection and enable insights into which variables were correlated with churn. Analysts found that a high count of notification openings and female users were more correlated to churn, while a higher count of articles read or if the subscriber was a past paper subscriber were more correlated with a lower churn risk. Furthermore, the company concentrates its focus on retaining users who present a likelihood of churning between 60% to 90%, since predictions with churn likelihood below this threshold present low certainty of churning, and above the threshold are represented by a group of users who most likely have already decided to cancel the subscription; therefore, it would take a larger effort to keep them.

The pay-TV sector was also subject to the study of churn prediction (Burez & Van den Poel, 2007). With the use of historical and current subscription data, socio-demographic data and interactions with the customer, the authors studied churn prediction on an imbalanced dataset (15% churners), comparing two LR with a RF algorithm. The results of the study show that for cut-off values lower than 25%, the RF model outperforms other techniques; however, for larger cut-off values, both LR models perform as well as the RF and end up surpassing it.

Bagging and boosting classification techniques to predict churn have been studied in the context of telecommunication subscriptions (Lemmens & Croux, 2006). When compared with LR, results show a significant improvement in accuracy. Both techniques have similar performance, nevertheless, authors consider bagging the most competitive approach since it is conceptually less complicated than stochastic gradient boosting. Furthermore, random oversampling was applied to half of the training

set (51.306 observations) to reach an equal size proportion of churners and non-churners, whereas the other half of the training set kept the original proportions (1.8% churn). Results lead to the conclusion that, in the context of this study, a balancing schema was preferable. Finally, the authors point out the need to correct the bias originated by a balanced training sample and find that selecting a cut-off value different to the central point between classes is substantially better than assigning a weight to each observation in the training set.

Also in the telecommunication sector, hybrid models have been explored to provide better performance. The authors sequentially combined NN with NN and Self-Organising Maps with NN (Tsai & Lu, 2009). The first technique filtered out unrepresentative training data, then, the second used the representative outputs to create the prediction model. Results indicated that the single NN was outperformed by both hybrid models in accuracy and Type I and II errors, being the combination of two NN the best performing model.

2.3. CHURN IN THE NEWS MEDIA SECTOR

Several articles can be found covering interviews and conferences (“Retaining Subscribers: Key Learnings from The Atlantic and MediaNews Group,” 2021; Sonderman & Vargo, 2021) with members of multiple well-known newspapers, addressing retention strategies and churn drivers on a higher level.

The process of retaining customers starts before the users purchase the product, in addition, most subscription-based businesses capture larger churn rates in the early life of a customer (Campbell, 2018). Customers tend to subscribe more easily and keep being subscribers if they are engaged with the product (American Press Institute, 2018), as such, the Financial Times (FT) approach to retention strategies starts from early customer life to the moment of renewal, always considering that each stage of the customer lifecycle can be an opportunity to reduce churn (Financial Times, n.d.). Claiming that creating habits has proven to be more valuable for the FT subscribers' engagement than presenting them articles (Veseling, 2018), the FT tries to materialize those habits from day one. To analyse subscribers' engagement, and anticipate customer churn, the FT considers the last 90 days and examines a Recency – Frequency – Volume (RFV) score for every subscriber stating it provides a good approximation for retention.

On another note, Dagens Nyheter, a Swedish newspaper, has created a research engine that pinpoints more than 200 features that explain churn (Southern, 2019). They identify the lack of frequency in visits as the most common churn driver. In addition, direct debit payment methods and subscribers' seniority are drivers of low churn while monthly invoice reminders are drivers of high churn (SparkBeyond Team, 2021). On the other hand, The Economist favours the time on the website as the single biggest predictor of retention (Suárez, 2019). Other good indicators for churn are the number of premium articles read, consumption across devices, the channel used when a user converted, whether a user hit the paywall or subscribed on a premium article, etc. (Campbell, 2018).

Although metrics that support digital news media companies to identify probable churners may vary, subscribers' engagement metrics are the most commonly used and have proven to be of great importance for many newspapers. Furthermore, most newspapers agree that building product usage habits help in sustaining a longer relationship with customers.

2.4. BALANCING TECHNIQUES

Churn prediction tasks are often characterised by imbalanced datasets, where the churn class is frequently the minority class and of main interest, with churn rates varying between 3% and 25%, depending on the industry (Burez & Van den Poel, 2009). This poses a challenge to the performance of most standard classifiers, which assume balanced class distribution and equal misclassification costs. Data sampling is one of the most common techniques to overcome class imbalance (Burez & Van den Poel, 2009). Its purpose is to eliminate or reduce rarity by altering the distribution of the training data. Under-sampling focuses on eliminating instances from the majority class. Contrarily, over-sampling duplicates minority class examples. However, both techniques have downsides. Under-sampling removes data samples that may be representative of the majority class; thus it can downgrade the performance of a classifier. Over-sampling involves making exact copies of the minority class, which may lead to overfitting the training data.

Advanced sampling techniques, such as Synthetic Minority Over-sampling Technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), overcome both obstacles by using oversampling to create artificial data for the minority class while maintaining the number of instances in the majority class. SMOTE randomly selects an instance from the minority class and finds its k nearest neighbours from the same class. Then, it randomly selects one of those neighbours and creates synthetic data between the two instances. A downside of SMOTE is that it can only handle continuous variables, and using label encoding may lead to unnecessary categories. The SMOTE-NC approach covers this aspect as it can deal with datasets containing both types of data.

Cost-sensitivity learning is another approach to overcoming class imbalance, where each class is provided with a distinct misclassification cost (Sun, Wong, & Kamel, 2009). Making classifiers cost-sensitive by placing a greater penalisation on misclassifying the minority class can help tune models to attain better performances (Burez & Van den Poel, 2009). Boosting techniques are techniques where a group of learners is constructed sequentially, considering the previous models' errors to better predict them in the subsequent models. This technique has proven to be a good method to prevent under classifying the smallest target class (Lemmens & Croux, 2006; Sun et al., 2009).

2.5. INTERPRETABILITY IN MACHINE LEARNING

The need for interpretability in ML models is gaining relevance, particularly in areas where automated decisions are critical (e.g., healthcare, financial markets, criminal justice) (Rudin, 2019). In 2018, the European Parliament adopted the General Data Protection Regulation (GDPR), which entitles individuals affected by automated decision-making the right to obtain meaningful explanations of the logic involved (Goodman & Flaxman, 2017), which further emphasises the need for an understandable model. In ML, interpretability is defined as the capacity to explain or give meaning in understandable terms (Guidotti et al., 2018). Lipton (2018) proposed that the concept of interpretability could be divided into two categories, namely transparency and post-hoc explanations. The former is depicted as some understanding of the mechanism by which the model works, whilst the latter explains predictions without clarifying how a model works.

Neither Decision Trees (DT) nor LR models are intrinsically interpretable since high dimensional models and deep DT could be difficult to understand. Nonetheless, these two algorithms have a superior preference among researchers favouring interpretability (Guidotti et al., 2018). Alternatively, other

models such as NN and SVM have been categorised as black boxes (Farquad et al., 2014; Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015) due to the lack of transparency. Tree-based ensembles are also perceived as black boxes as they are composed by a group of classifiers, which makes the task of comprehending predictions challenging. However, many papers suggest that black boxes can also be explained. For tree-based ensembles, this can be done by computing each feature's relative importance in prediction and analysing Partial Dependency Plots (Lemmens & Croux, 2006), which allows the understanding of which variables are most important in explaining churn and how they affect that decision, respectively.

SHAP values (Lundberg & Lee, 2017) are a novel approach for model interpretation, based on Shapley values from game theory. It assigns each feature an importance value for a determined prediction allowing us to see how much it contributes global and locally to the prediction. TreeSHAP (Lundberg, Erion, & Lee, 2018) is a high-speed derivation of SHAP developed for tree-based algorithms, such as Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM), allowing a visualisation-based approach to interpret these types of black boxes. Local Interpretable Model agnostic Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016) is a method that also supports models' interpretation. It trains local models with permuted data to study changes in the predictions when presented with slight variations to the data. LIME is characterized by high local fidelity, nonetheless, global and local interpretations with SHAP are consistent (Främling, Westberg, Jullum, Madhikermi, & Malhi, 2021) and literature shows a much stronger agreement between human explanations and SHAP than with LIME (Lundberg & Lee, 2017).

3. METHODOLOGY

This project was carried out applying a widely used methodology known as Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000). It defines a data mining project as a cyclic process of six phases, shown in Figure 3.1, where the sequence of phases is non-rigid, and several iterations can be used to allow a final result fine-tuned towards business goals.

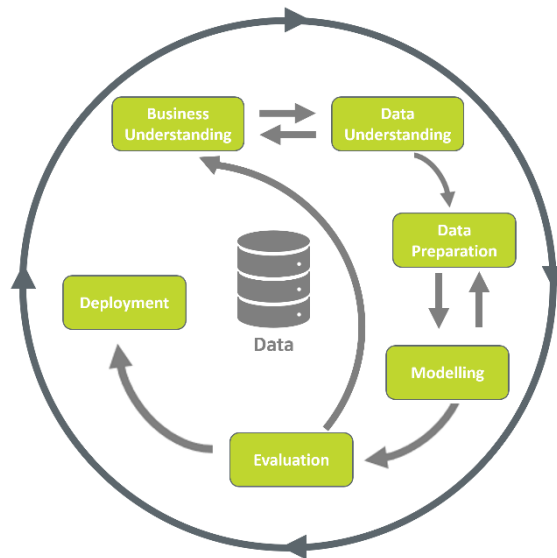


Figure 3.1 – CRISP-DM process model

(adapted from Chapman et al., 2000)

Two experimental setups resulted in the development of two models. One to predict churn for recurring (automatically renewed) and non-recurring (non-automatically renewed) subscriptions (1) and a second one to predict churn only for non-recurring subscriptions (2). The models were sequentially evaluated with daily predictions being subject to A/B tests, in which one group was contacted with retention actions (variation group) and the other was not (control group).

As presented in Figure 3.2, both models' development went through a similar preprocessing phase: data was cleaned, transformed and filtered. The modelling and respective evaluation phases followed the same framework for each setup. The project framework is detailed as follows.

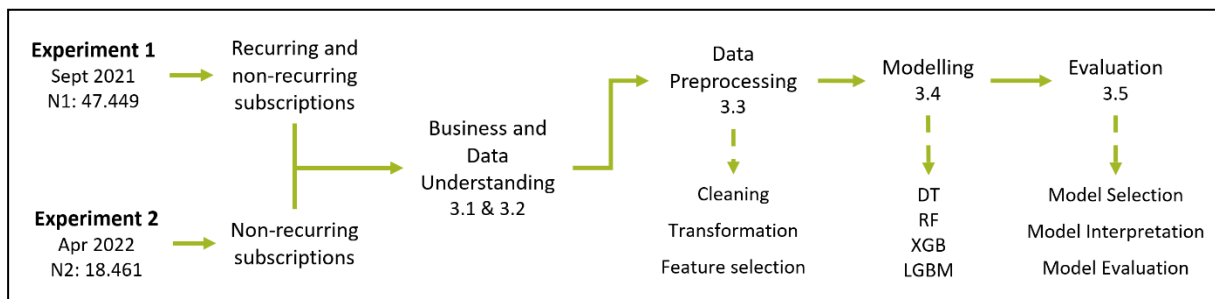


Figure 3.2 – Project framework

3.1. BUSINESS UNDERSTANDING

The initial stage concentrates on understanding the project's objectives from a business perspective and transforming it into a DM problem. In this project, customer retention of PÚBLICO's online subscription was tackled through the task of churn prediction, formalising it as a supervised ML research for binary classification. The target variable was defined as churning (1) if a customer did not renew his or her subscription 30 days after the renewal date; otherwise, the customer was classified as a non-churner (0).

Two sets of customers were studied, one containing all kinds of subscriptions and another containing only non-recurring subscriptions. The decision of performing a second trial considering only non-recurring subscriptions arose from the fact that the churn rate in only non-recurring subscribers is far larger than considering both groups, thus the business deemed it would be of high importance to understand and attempt to mitigate churn within this group of subscribers.

The tool used in this project's development was a Jupyter Notebook, employing python language and relying mainly on the following python packages:

1. pandas: offers multiple data structures and functions which are used for data manipulation and analysis (McKinney, 2010).
2. Scikit-learn: contains a wide range of algorithms for supervised and unsupervised ML, modules for data preprocessing and evaluation of models' performance (Pedregosa et al., 2011).
3. XGBoost: open-source software providing a GBM framework to implement the XGBoost algorithm (Chen & Guestrin, 2016).
4. LightGBM: open-source library, originally developed by Microsoft, used to implement the LightGBM algorithm (Ke et al., 2017).

3.2. DATA UNDERSTANDING

PÚBLICO's team provided a dataset containing 47.449 subscriptions to build the first model and 18.461 for the second model, both with 44 predictive features regarding website and newsletters consumption, subscription product specifications, user-related variables and historical information regarding users' affiliation with the company. The data was collected 30 days after the ending date of a subscription, along with a subscription status identifier. The datasets contained respectively a churn rate of 9,6% and 22,9%, and covered a period of subscriptions ending dates from November 9th, 2020 to June 21st, 2021, and from November 9th, 2020 to January 11th, 2022.

Table 3.1 – Raw datasets for each model

	Model	
	1	2
Subscriptions	47.449	18.461
Churn Rate	9,6%	22,9%
Period	2020-11-09 2021-06-21	2020-11-09 2022-01-11

Different kinds of variables cover different periods, namely: product's usage data is related to the last two months of the subscription, collected for a period of 30 to 0 days and 60 to 30 days before the end of the subscription; historical data on the subscriber concerns the entire period since the subscriber first registered on the website; characteristics of the subscription product represent the entire subscription period. In Figure 3.3, the covered periods by these variables are exemplified with a subscription that started on December 3rd, 2019 and ended on December 3rd, 2020. Usage of the website variables characterize the last subscription month, i.e., the last 30 to 0 days of the subscription, and the month before that, i.e., the last 60 to 30 days before the subscription. In the example, these periods are between November 3rd, 2020 and December 3rd, 2020, and between October 4th, 2020 and November 3rd, 2020, respectively. As for subscription related variables, these characterise the current subscription, therefore, they represent the period from the first day of the current subscription to the last day of that same subscription. Finally, the subscriber's historic with the company would represent the period from the registration on the website, May 20th, 2014, and the last day of the current subscription. Churn would be predicted for January 2nd, 2021, 30 days after the subscription ended.

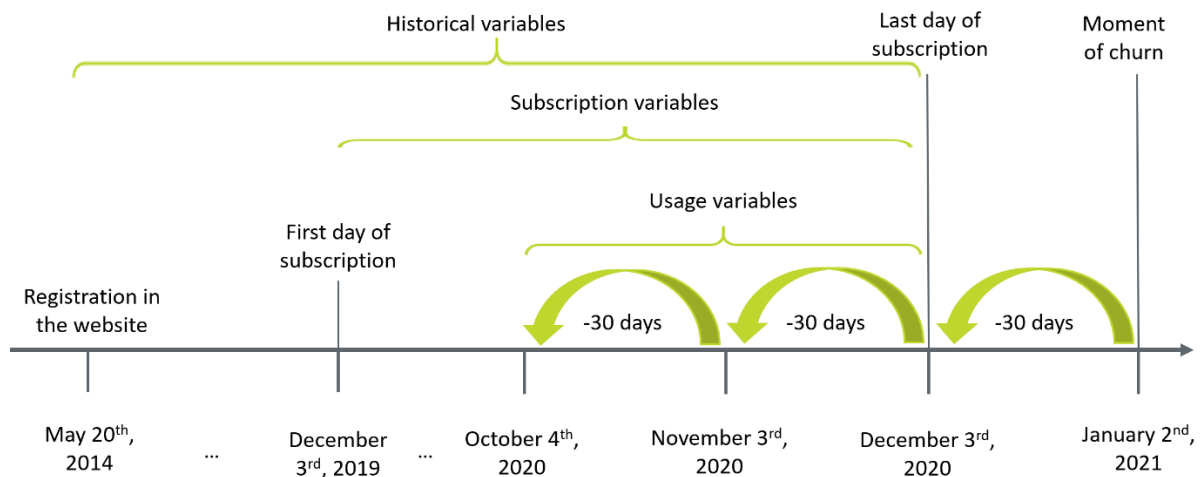


Figure 3.3 – Periods covered by the different kinds of variables exemplified

In addition, with the Covid-19 pandemic, PÚBLICO has seen more interaction with the website. To understand if the stringency of the government measures in response to Covid-19 is related to the customer's decision to keep subscribing to the newspaper, a dataset containing daily scores of the stringency of these measures in Portugal was used for the first model. The average stringency indexes analysed cover the same periods as the usage related variables. This data was retrieved from the Oxford COVID-19 Government Response Stringency index (Petherick et al., 2021) which collects measures taken by governments in response to the pandemic on 23 indicators and scores those measures on a scale of 1 to 100.

Considering this second dataset, 47 predictive features were available for the models' development and are described below in Table 3.2.

Table 3.2 – Predictors' meaning and periods which they cover

	Variable Name	Variable meaning	Period
Usage Variables	N_60_M_RFV_60D, N_30_M_RFV_30D	Maximum RFV value	0D - 30D & 30D - 60D
	N_60_N_dias_visitas_60D, N_30_N_dias_visitas_30D	Number of days with visits	
	N_60_Ultima_Visita_60D, N_30_Ultima_Visita_30D	Days since the last visit	
	N_30_2_Artigos_N30, N_60_2_Artigos_N60	Total articles, which are prone for conversion, read	
	Excl_N60_Exclusivos60, Excl_N30_Exclusivos30	Total exclusive-to-subscriber articles read	
	readability_P2_url_75_60, readability_P1_url_75_30	Total articles in which 75% of the article was read	
	engagement_Share_60D, engagement_Share_30D	Number of shares from the website to social media	
	engagement_COMENT_60D, engagement_COMENT_30D	Number of comments on the website	
	engagement_Guardar_60D, engagement_Guardar_30D	Number of articles saved on the website	
	N_30_S_seccaoPref	Preferred content section	0D - 30D
	NL_30_NL1_b_NL_LIDAS	Number of newsletters read	
	NL_30_NL2_b_NL_cliues	Number of newsletters clicked	
Subscription Characteristics	A1_SUBSCRIPTION_CAMPAIGN	Campaign related with current subscription	Subscription length
	A1_Data_fim	Subscription's end date	
	A1_Data_inicio	Subscription's start date	
	SUBSCRITPION_CHANNEL	Channel through which the subscriber purchased the current subscription	
	A1_SUBSCRIPTION_TYPE	Binary indicating if the subscription is paid or offered	
	SUBSCRIPTION_PRODUCT	Subscription product duration	
	SUBSCRIPTION_PAYMENT_MODE	Payment method used	
	SUBSCRIPTION_ORIGINAL_DAYS	Subscription length in days	
	SUBSCRIPTION_IS_RECURRENT	Binary indicating if the subscription is automatically renewed	
	SUBSCRIPTION_IS_TRIAL	Binary indicating if the price increases in the next subscription	
Historical Variables	SUBSCRIPTION_VALUE_WITH_TAXES	Subscription price	Since the user registered the website
	SUBSCRIPTION_CLASSIFICATION	Customer's relationship with the company at the purchase moment	
	b_assinaturasTodas	Total subscriptions	
	b_Media_Dias	Average subscription length in days	
	b_Max_Dias	Maximum subscription length in days	
	c_Antiguidade	Days since the first subscription	
	e_taxa_fidelidade	e_numerador_T/ c_Antiguidade	
	e_numerador_T	Days with active subscriptions	
Subscriber Socio-demographics	i_product1	Preferred subscription length in days	30D after subscription ends
	USER_create_date	Subscriber registration on the website	
	A2_USER_GENDER	Subscriber gender	
	A2_USER_BIRTHDAY	Subscriber birthday	
Target	A2_USER_JOB	Subscriber job	30D after subscription ends
	A3_SUBSCRIPTION_IS_ACTIVE	Binary indicating if the subscription is active	
Stringency Index	avg_stringency_index_30, avg_stringency_index_60, avg_stringency_index_total	Average stringency index	0D - 30D, 30D - 60D, 0D - 60D

3.3. DATA PREPARATION

Data preprocessing is a critical phase in the data mining process (Antonio, de Almeida, Nunes, Batista, & Ribeiro, 2018), as it is when representational data and its quality are addressed. When there is irrelevant, redundant, unreliable or noisy data present, the process of finding patterns while training the algorithms is more complex and challenging (Pyle, 1999). As such, detecting this data while preprocessing, and properly dealing with it will influence the success of the classification model (Neslin, Gupta, Kamakura, Lu, & Mason, 2006).

Preprocessing data for it to be ingested by algorithms to perform prediction tasks usually involves steps such as data cleaning, data transformation and data reduction (Han, Kamber, & Pei, 2012). Data cleaning requires dealing with missing values, duplicate rows, outliers and tackling incoherencies. Data transformation, on the other hand, encompasses normalising the data scale and feature engineering (Chapman et al., 2000). Data reduction involves shortening the size of the dataset, which can be done with feature selection. The adopted procedures for preparing the data for both models were equal and are described in the following subchapters.

3.3.1. Data Cleaning

When developing a ML model it is often mandatory to deal with missing values, and there are multiple approaches to deal with this task depending on the context in which the data is missing (Pyle, 1999). Deleting rows with missing data might not always be the best approach because there may be a systematic pattern of missing data and other valuable data might be lost during the removal. Another approach is to discard the variable with missing data (Han et al., 2012), which can be a good method when there is a large quantity of missing data and neglecting it does not affect the analysis. Other methods encompass inputting data with a constant value or a random value or extracting a flag variable to indicate when the variable is missing (Han et al., 2012).

Within the provided datasets multiple variables contained missing values. Variables related to the usability of the website (see Table 3.2) had a minimum value of 1, therefore the missing values were filled with a constant value of 0, indicating that there was zero interaction of that kind. Users' birthdays had 1/3 of the data represented by a constant value, therefore these instances were treated as missing values. Also, only ages between 18 and 100 were considered, thus the remainder was treated as missing data. Since a large number of entries were missing, to use this variable the missing data was inputted with the mean of the subscribers' ages on a further step.

Another important step is the removal of duplicate data entries, which can bias the model to focus on predicting those entries better, as well as outlier elimination, as it can lean the model to predicting entries which can be misleading. The datasets contained duplicate subscriptions with distinct information on the users' birthdays and registration dates to the website. Each time a user updated their birthday information on the website, a duplicated entry was showing up on the dataset regarding the same subscription, however, each instance would have a different birthday and registration date. Thus, to prevent the existence of a subset of duplicated data, the most recent alteration on the birthday information and the first registration date were considered and used to correct the subscriptions' data. Furthermore, only subscriptions with the most accurate data from duplicated subscriptions collected on different days and duplicated subscriptions with different information were considered.

In terms of outliers, it was defined that all users with less than 20 days of subscription duration should be removed from the analysis, as in this type of subscription instances, users were offered some days of the subscription until their actual subscription started. Also, subscriptions without visits but with any kind of engagement variable were removed. By keeping unique instances and removing the outliers, the datasets were reduced by 10.8% for the first model and 10.1% for the second, where only 0.01% were identified as outliers on both datasets.

Finally, it is important the data used to build a model can be reliable, therefore, confirming its coherence is an indispensable task. For categorical variables, it was necessary to validate and group smaller categories into one larger group. For example, the subscription's favourite content section variable contained multiple categories which did not exist, and it was decided to convert those into a grouped category indicating that the user did not log in during the month. On the other hand, the subscription product was a variable with more than 70 entries, which for a ML model can become complex to understand. Therefore, as these categories had the subscription duration in common, it was possible to unify them into fewer groups with larger frequencies.

In addition, some subscription durations that did not match the subscription product were rectified, and subscriptions which had a duration of less than 0 days were corrected by computing the difference between the end and start day of the current subscription. Users' registration dates on the website which were after the first subscription was purchased were replaced by the first subscription's purchase date. Furthermore, coherence checks on the subscription's classification and target variable were performed and some inconsistencies were found and rectified.

Missing values	Duplicated subscriptions	Outliers	Incoherencies
- Usability variables - Birthday	- Collected on different dates - Different birthday and/or registration date	< 20 subscription days - With usability but no visits	- Inexistent or faulty categories - Subscription duration - Target variable

Figure 3.4 – Data cleaning stages

3.3.2. Data Transformation

Having available a large number of features that interact with each other allowed the exploration and creation of new variables. For usability variables in which there was available data for the last two months before the subscription ended, the following measures were derived:

1. Sum of activity in both periods (e.g.: $N_{30_2_Artigos_N30} + N_{60_2_Artigos_N60}$);
2. Proportion of activity per period (e.g.: $(N_{30_2_Artigos_N30} - N_{60_2_Artigos_N60}) / (N_{30_2_Artigos_N30} + N_{60_2_Artigos_N60})$), allowing to understand in which period the user had a higher activity;
3. Average activity per day for each period and both periods aggregated (e.g.: $N_{30_2_Artigos_N30} / N_{30_N_dias_visitas_30D}$);
4. Classes of activity level per period using bins formed according to the percentiles of each variable (e.g.: $N_{30_2_Artigos_N30}$ divided in percentiles of 20%) and the number of days in

a month for the number of days with visits, this way reducing the impact of extreme values on the model's performance.

Furthermore, binary variables indicating if the user interacted with the website pages (comment, share or save) were created and the maximum RFV variables were split into three columns, each indicating a value in the RFV variable. For newsletters, the click-through rate and a sum of the clicks and openings were computed, as a measure of engagement and magnitude, along with a binary variable indicating if there was any interaction.

From the birthday information, it was possible to extract the user's age at the end of the subscription. The start and end dates of the subscription resulted in the weekday and the day of the month in which the subscription was purchased/ended, and the registration date to the website resulted in days since the user registered. Variables of days since the first subscription (c_Antiguidade) and days with active subscriptions (e_numerador_T) were converted into years.

Through the subscriber's favourite product, a binary variable indicating if the current product was the preferred product was created. Also, the average price per day on the current subscription was computed.

Finally, categorical variables such as the subscription product variable were unified into three classes comprising monthly, quarterly and semestral, and annual or larger subscriptions. The most read section in the last month variable had some classes with less than 5% of instances, so those classes were merged into one, representing other sections, in order to avoid over fitting these categories to few examples. As most ML algorithms cannot use text to perform predictions, categorical variables must be converted into numerical variables. Therefore, variables such as the payment method, most read section in the last subscription month, subscription classification and subscription product were targeted with label encoding and weight of evidence techniques to deal with categorical data.

Feature engineering provided a wide range of 151 variables.

3.3.3. Data Reduction

Feature selection plays a key role in the model's performance and computational costs. Selecting a small but highly predictive subset of variables reduces collinearity issues and increases model interpretability (Verbeke et al., 2012). Having multiple variables representing the same information can create bias in the model's results.

Firstly, the user variables gender and job were removed, because they contained a large number of missing values (33% and 95%, respectively) and user gender could bias the model into having preferred behaviour towards one gender, thus not being compliant with GDPR. Furthermore, multiple variables which were duplicated were eliminated, as well as variables that during the coherence check process indicated unreliable data.

To perform variable selection, within the 151 available variables, 18 groups of features representing the same information were identified and a base group of variables with one variable per group was created, this way defining a base dataset constituted by features which do not represent the same information.

Then, to select a variable from each group of features representing the same information, the base group of variables was used to compute XGBoost models with the default parameters given by the library, in which all variables in the base group were constantly present in the models, except for the variable that was being analysed, which was sequentially replaced by all other variables that represent the same information. To select one feature within each group, the evaluation metrics SHAP value for feature importance, F1 Score, gain, cover and weight were extracted for the features in analysis, and a rank for each of these metrics was computed. Then an average of the computed ranks was generated, and the best average ranked feature was defined as the selected one within the group.

After selecting a variable from each group, the Pearson correlation between the selected variables was assessed and variables with absolute correlations over 0.75 were either replaced with another variable representing the same information in another format, if available, or the variable with the smallest correlation to churn was removed, if another equivalent was not available. Finally, tests for removing one variable at a time were performed, assessing the model performance, and confirming if removing it improved the F1 Score, then the variable would be removed.

The final training sets contained 16 predictors, for the 1st model, and 11 predictors, for the 2nd model, which are described in Table 3.3. As the algorithms considered for this project have built-in feature selection implementations, maintaining a relatively large number of predictors does not jeopardise the model's performance.

Table 3.3 – Features selected for each model

Features Selected	
Model 1	Model 2
A1_SUBSCRIPTION_ORIGINAL_DAYS	A1_SUBSCRIPTION_PRODUCT
Total_Visits	N_30_N_dias_visitas_30D
Prop_ultima_visita_period	N_30_Ultima_Visita_30D
30_Total_ArtExcRead_class	30_Total_ArtExcRead_class
c_Antiguidade	c_Antiguidade
Age_class	Age_mean
weekday_Data_fim	DayOfMonth_Data_fim
N_30_S_seccaoPref_outra_LE	N_30_S_seccaoPref_outra_woe
A1_SUBSCRIPTION_CLASSIFICATION_LE	A1_SUBSCRIPTION_CLASSIFICATION_woe
e_taxa_fidelidade	e_taxa_fidelidade
bassinaturasTodas	bassinaturasTodas
A1_SUBSCRIPTION_IS_RECURRENT	--
A1_SUBSCRIPTION_PAYMENT_MODE_LE	--
Product_is_FavProduct	--
avg_stringency_index_30	--
NL_30_NL1_b_NL_LIDAS	--

3.4. MODELLING

During the modelling stage, preprocessed datasets were sorted by subscriptions' end date and split into train (75%) and test data (25%) for the first model, and into train (48%), validation (32%) and test

data (25%) for the second model. The churn percentage in the first model was 8% on train and test data, comprising a time period with recurrent and non-recurrent subscriptions ending between November 9th, 2020 and June 21st, 2021. Whereas for the second model, churn was 17% for train data, 25% for validation and 22% for test data, containing only non-recurring subscriptions ending between November 9th, 2020 and January 11th, 2022.

Table 3.4 – Datasets used for model selection

Model	Dataset	Subscriptions	Subscriptions (%)	Churn	Time period
1 Recurrent & Non-recurrent	train	31.741	75%	8%	2020-11-09 to 2021-04-23
	test	10.581	25%	8%	2021-04-24 to 2021-06-21
	total	42.322	100%	8%	2020-11-09 to 2021-06-21
2 Non-recurrent	train	7.615	48%	17%	2020-11-09 to 2021-06-22
	validation	5.078	32%	25%	2021-06-23 to 2021-11-02
	test	3.174	20%	22%	2021-11-02 to 2022-01-11
	total	15.867	100%	21%	2020-11-09 to 2022-01-11

The oversampling technique SMOTE NC was applied to the training datasets to improve predictive performance by balancing the instances in each class. The validation and test datasets maintained the original distribution to assess the models' performance, presenting a natural distribution of churners. The original training data and the oversampled training data were relied on to build models based on state-of-the-art algorithms XGBoost and LightGBM algorithms. RF and DT models were also constructed for performance comparison purposes. For the algorithms developed on top of the original training data distribution, the parameter `class_weight` was set to balanced.

These algorithms have several parameters that can be used to optimise the models' performance and reduce the overfitting effect on training data. To select such parameters, for each combination of algorithm and training dataset, a grid search with 3-fold cross-validation was computed with the parameters defined in Table 3.5. Parameter optimisation was also implemented, with the same framework, for DT and RF algorithms to obtain optimal results for comparison. Using a low number of leaves and a reduced max depth tends to make the algorithm less prone to overfitting.

Table 3.5 – Parameters considered for grid search cross-validation per model

	XGBoost	LightGBM	Random Forest	Decision Tree
learning rate	[0.1, 0.2, 0.5]	[0.01, 0.1, 0.2]		
max_depth	[2, 3, 5, 7]	[-1, 2, 3, 5, 7]	[None, 2, 3, 5, 7]	[None, 2, 3, 5, 7, 8]
n_estimators	[200, 500, 1000]	[200, 500, 1000]	[200, 500, 1000]	
base_score	[0.2, 0.5]			
colsample_bytree	[0.5, 0.8]	[0.5, 0.8]	['auto', 5, 8]	[None, 5, 8]
subsample	[0.3, 0.5, 0.8]	[0.3, 0.5, 0.8]		
gamma	[1, 20, 50, 100]			
min_samples_leaf		[None, 100, 200]	[1, 100, 200]	[1, 100, 200]
num_leaves		[5, 10, 20, 30]		
subsample_for_bin		[500, 1000, 200000]		

3.5. EVALUATION

3.5.1. Models Selection

It is important to evaluate the models' performance on unseen data to ensure it is not overfitted to the training data and, consequently, resulting in poor predictions (Moisen, 2008). Therefore, the test data was used to evaluate the classifiers' capability to predict unseen data.

Several metrics used to evaluate binary classifiers are extracted from a confusion matrix, which summarises the number of correct (true positives and true negatives) and incorrect predictions (false positives and false negatives). Accuracy, precision, recall, specificity and F1 Score are some of these measures (Burez & Van den Poel, 2009; Farquad et al., 2014). However, most churn prediction problems are characterised by imbalanced datasets, and accuracy does not distinguish between classes, therefore, it can be misleading and should be analysed cautiously. Another commonly used evaluation metric in churn prediction problems is the AUC (Coussement et al., 2017; Höppner, Stripling, Baesens, Broucke, & Verdonck, 2017; Verbeke et al., 2012).

K-fold cross validation trains and tests predictions on k datasets, as such, for each dataset it is possible to evaluate models' performance individually and compare it against other datasets to avoid overfitting. For this project, the grid-search employed to tune the models used a 3-fold validation and the averages for multiple evaluation metrics (accuracy, balanced accuracy, precision, recall, F1 Score and AUC) on the 3 train and test sets were extracted. When selecting the best combination of parameters from the grid search, the five highest models for each average evaluation metric on the test datasets were considered for each combination of algorithm and dataset, producing $360 (5 \text{ (top 5)} \times 6 \text{ (evaluation metrics)} \times 4 \text{ (algorithms)} \times 2 \text{ (datasets)})$ different models.

The models selected with the grid-search results were evaluated against the test and against the validation dataset for the first and second models, respectively. Overfitting models were discarded and thresholds for each metric were considered, allowing better filtering of the models. The results of the selected are discussed in chapter 4.1.

3.5.2. Models Evaluation

Once model 1 was selected, PÚBLICO's team suggested implementing a first experiment over a period of three weeks in September 2021, to perform daily churn predictions on recurrent and non-recurrent subscribers who would end their subscription during that period. Model 2 was tested in a second experiment in April 2022, in which daily churn predictions were conducted on non-recurrent subscriptions. For both models, subscribers with a predicted probability to churn over 50% were considered possible churners, and 50% of the predicted churners were targeted with a retention campaign as a means to perform A/B tests.

After meeting with the Analytics and the Subscriptions teams, it was defined that the retention measure of the first experiment would be the offer of one month of free subscription for subscribers who purchased a yearly subscription. As for the second experiment, having already analysed the results of the previous experiment it was decided to target probable churners with an engagement newsletter, containing content about the war occurring in Ukraine, and a phone call, when a phone number was available, reminding subscribers that their subscription was ending.

To employ both experiments, the data was collected from BigQuery to a Google sheet, from where a python script fetched that data and prepared it for model ingestion. For the purpose of an A/B test, the output of the models was equally split into two groups, one for subscribers who were predicted churners and should be impacted by the retention measures, the variation group, and another for predicted churners who should not be impacted by the retention measures, the control group. Having an A/B test allows understanding, on the one hand, if the retention measures performed well in keeping the subscribers active, and, on the other hand, if the model predicted well on subscribers not impacted by the retention measures.

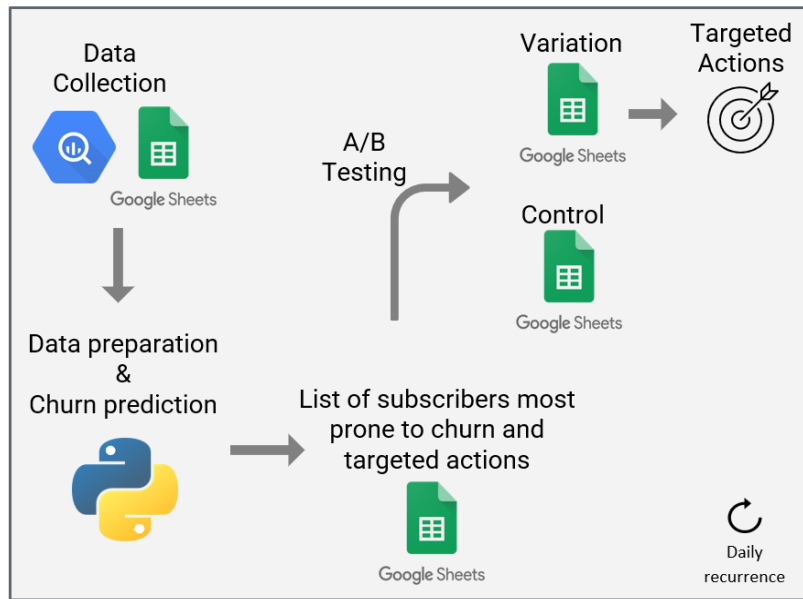


Figure 3.5 – Daily predictions pipeline

4. RESULTS AND DISCUSSION

4.1. XGBOOST MODEL

The filtered models attained with the grid-search for parameter selection resulted in 8 possible modelling options for both modelling contexts. The evaluation metrics of these models are presented in Table 4.1. For recurrent and non-recurrent subscriptions (model 1), DT and RF are models that present the lowest performance considering the F1 Score (F1). While, Light GBM and XGBoost have higher performance, both in train and validation data. However, only XGBoost built with the non-balanced dataset does not overfit the training data. As for non-recurrent subscriptions (model 2), the F1 measure increases on all algorithms and shows a good generalization when predicting unseen data. The 8 models present a similar performance and XGBoost trained with the original dataset has the highest F1 performance. Furthermore, for both modelling contexts, the models trained with balanced datasets show overfit to the train data, not being able to generalize when introduced to new data. Considering the same combination of algorithm and dataset, the non-recurring models capture better patterns within churning subscriptions, which can indicate that analysing individually the behaviour of non-recurring subscribers can provide better insights than analysing all subscribers as one group.

Table 4.1 – Evaluation metrics of models trained with original and balanced data for both modelling contexts

	Algorithm_dataset	Accuracy		Balanced Accuracy		Precision		Recall		F1		AUC		Specificity	
		Train	Validation (Val)	Train	Val	Train	Val	Train	Val	Train	Val	Train	Val	Train	Val
Model 1	DT_original	94,52%	94,58%	82,09%	82,65%	68,34%	64,32%	67,08%	68,52%	67,71%	66,35%	90,56%	91,86%	97,09%	96,78%
	DT_smote	82,45%	94,57%	82,45%	82,54%	95,94%	64,31%	67,77%	68,28%	79,43%	66,24%	91,90%	91,83%	97,13%	96,79%
	RF_original	92,97%	93,28%	91,42%	89,43%	55,55%	54,47%	89,54%	84,87%	68,57%	66,35%	98,00%	96,45%	93,30%	93,99%
	RF_smote	93,31%	92,24%	93,31%	88,92%	92,84%	50,18%	93,85%	84,99%	93,34%	63,10%	98,63%	95,73%	92,76%	92,85%
	LGB_original	99,48%	96,08%	99,71%	87,84%	94,24%	73,38%	100,00%	78,09%	97,03%	75,66%	100,00%	95,22%	99,43%	97,60%
	LGB_smote	94,20%	92,55%	94,20%	88,26%	93,63%	51,42%	94,85%	83,17%	94,24%	63,55%	98,67%	95,64%	93,55%	93,35%
	XGB_original	95,10%	95,15%	91,92%	88,67%	66,04%	65,27%	88,07%	80,99%	75,48%	72,29%	98,38%	96,10%	95,76%	96,35%
	XGB_smote	98,10%	92,52%	98,10%	85,47%	98,16%	51,41%	98,04%	77,12%	98,10%	61,69%	99,81%	93,13%	98,16%	93,83%
Model 2	DT_original	93,60%	91,02%	93,89%	91,03%	75,15%	76,80%	94,32%	91,05%	83,65%	83,32%	98,62%	95,41%	93,45%	91,01%
	DT_smote	88,75%	89,78%	88,75%	86,87%	92,31%	78,20%	84,54%	81,14%	88,26%	79,64%	92,80%	91,97%	92,96%	92,61%
	RF_original	95,10%	92,73%	93,63%	92,35%	82,33%	81,28%	91,37%	91,61%	86,62%	86,13%	98,82%	97,59%	95,89%	93,10%
	RF_smote	88,66%	91,97%	88,66%	89,45%	94,89%	83,16%	81,71%	84,49%	87,81%	83,82%	97,06%	96,22%	95,60%	94,41%
	LGB_original	94,66%	92,44%	93,90%	92,59%	79,75%	79,75%	92,73%	92,89%	85,75%	85,82%	98,85%	97,71%	95,06%	92,29%
	LGB_smote	94,84%	91,79%	94,84%	91,51%	94,54%	78,92%	95,19%	90,97%	94,86%	84,52%	98,94%	97,47%	94,50%	92,06%
	XGB_original	95,64%	93,70%	93,21%	92,62%	85,96%	84,92%	89,48%	90,49%	87,69%	87,62%	98,81%	97,69%	96,93%	94,75%
	XGB_smote	95,31%	92,04%	95,31%	91,71%	94,81%	79,59%	95,87%	91,05%	95,34%	84,94%	99,13%	97,65%	94,76%	92,37%

The selected best performing models were defined based on the F1 measure and AUC evaluation metrics and were both based on the XGBoost algorithm, with the original datasets' distribution. F1 is a metric that combines precision and recall, therefore it is a favoured metric when dealing with class imbalance problems. On the other hand, AUC considers sensitivity and specificity as individual class performance metrics over all possible thresholds. The model built with the complete dataset (recurrent and non-recurrent subscriptions) attained a F1 measure of 76% for the train dataset and

72% for the test data. As for the AUC, the results were 98% and 96%, respectively. When considering the model built on top of non-recurrent subscriptions the evaluation metrics were 88% for train and validation data, and 86% for test data for the F1, and 99%, 98% and 98% for the AUC. The models did not present a large overfit to the training data, when evaluated against the validation and test datasets, and should, therefore, perform as good on new and unseen data. All metrics for the selected models are present in Table 4.2.

Table 4.2 – Evaluation metrics extracted from XGBoost models selected

Model	Dataset	Accuracy	Balanced Accuracy	Precision	Recall	F1	AUC	Specificity
1	Train	95,10%	91,92%	66,04%	88,07%	75,48%	98,38%	95,76%
	Test	95,15%	88,67%	65,27%	80,99%	72,29%	96,10%	96,35%
2	Train	95,64%	93,21%	85,96%	89,48%	87,69%	98,81%	96,93%
	Validation	93,70%	92,62%	84,92%	90,49%	87,62%	97,69%	94,75%
	Test	93,64%	92,57%	82,45%	90,64%	86,35%	98,34%	94,49%

4.2. MODEL INTERPRETABILITY

The selected models were derived from a XGBoost algorithm, so it is possible to extract the importance of the features from the classifiers in terms of their weight, gain and cover. However, it may occur that, for each measure, features have different importance. On the other hand, SHAP values ensure consistent and accurate results, allowing for a more appropriate interpretation of the features' importance (Lundberg & Lee, 2017). Furthermore, Tree SHAP extends the evaluation of features' importance by identifying how much impact each feature has on the model's output for individual instances. Using SHAP's summary plot, features are vertically sorted from most to least important across all data samples, and their impact on the instance's outcome can be analysed horizontally. The colour of each data point denotes the true value of the feature. The SHAP summary plots from both models are displayed in Figure 4.1, where the impact of the features used to train them are exhibited. In both models, the predictors show a similar effect on churn likelihood and share identical importance ranking in churn prediction. The two predictors that have the highest impact on churn prediction are the fidelity rate (percentage of days with an active subscription) and the total amount of subscriptions purchased. Higher values for both features are related to a lower propensity to churn. In addition, for both models, a group of users with a low number of total subscriptions present a large impact on these predictions, indicating a higher propensity to churn. This shows that brand fidelity metrics are a good indicator of churn probability.

Furthermore, the third most important predictor is the current subscription length (in days), in model 1, and the current subscription product (monthly, annual and larger, or in between), in model 2. Although both are on different scales, they represent similar information: subscriptions with one year present a higher propensity to churn, whereas monthly subscriptions are more related to a lower likelihood to churn. This may be related to the higher prices paid for annual subscriptions.

Finally, another variable where the difference in the churn likelihood is highly visible in both models is the total number of days the subscriber visited the website during the last subscription's months/month, where more days with visits are related to a lower likelihood to churn.

Feature Importance Analysis

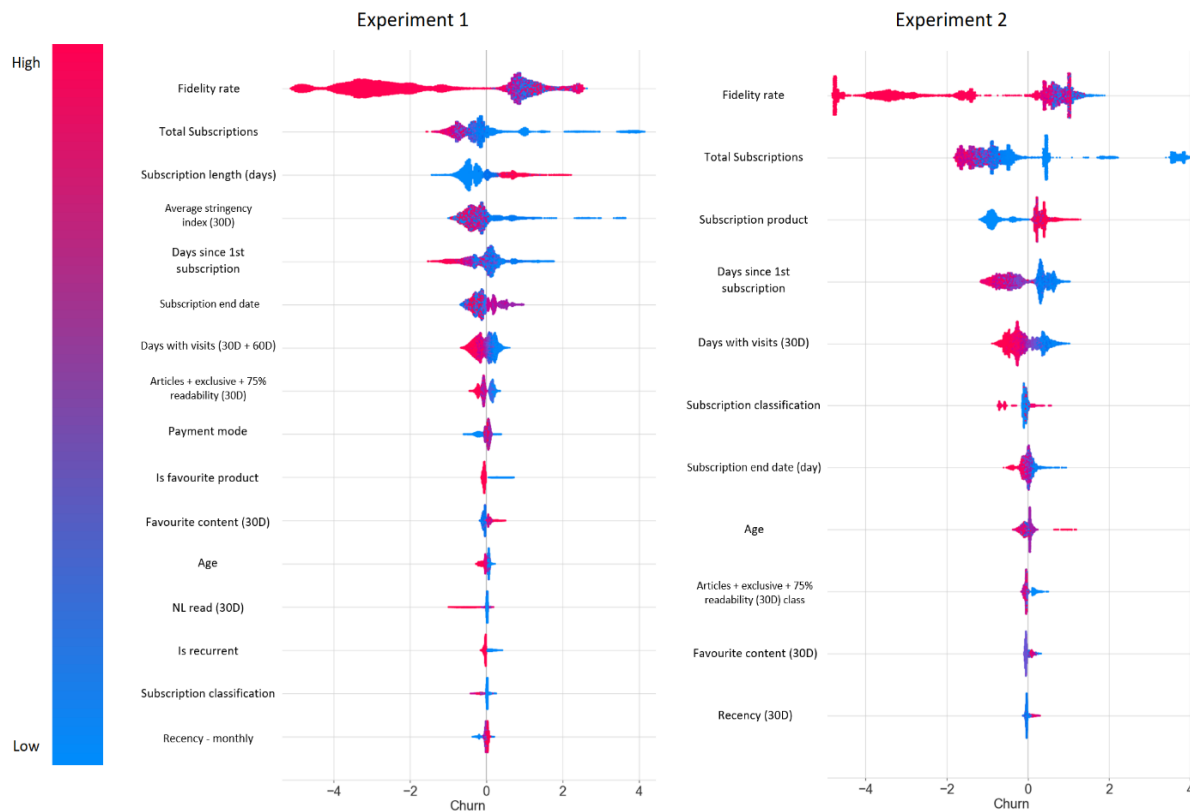


Figure 4.1 – Predictive features ranking by SHAP values

SHAP dependency plots represent the effect of individual features over predictions. Each point on the graph represents a subscription, and the colour scale indicates the values of the variable the dependency plot is being plotted against, being higher values represented in red and lower values represented in blue. Below there are some of the dependency plots for some predictors used to build model 1.

From Figure 4.2, it is visible that subscribers who have almost always had an active subscription since their first subscription (96% of the time) are less likely to churn. In addition, Figure 4.3 shows a downward trend in the propensity to churn as the days since the first subscription increase, having a turnover moment of around 2 and a half years, most visible for a group of subscriptions.

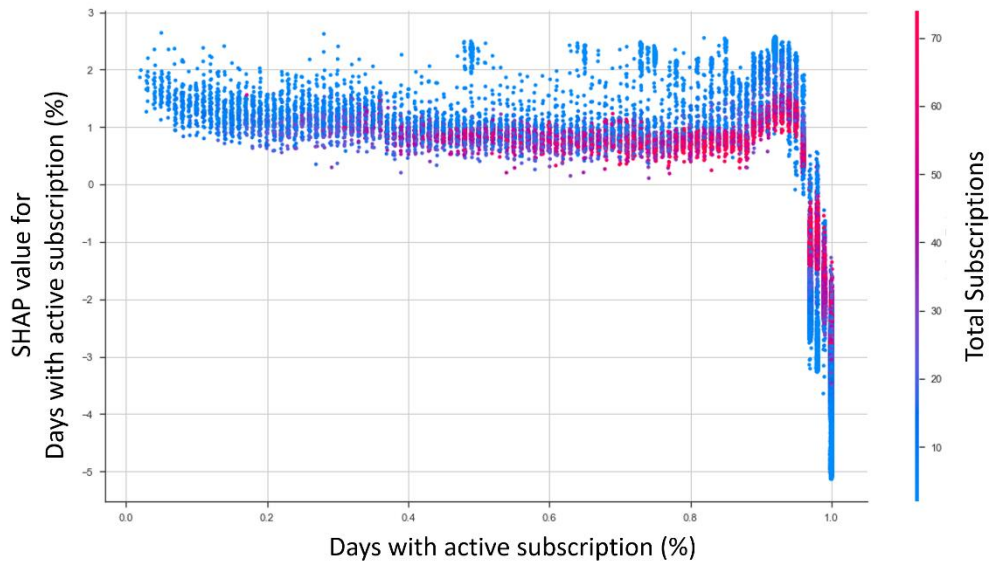


Figure 4.2 – Dependency plot for Days with an active subscription (%)

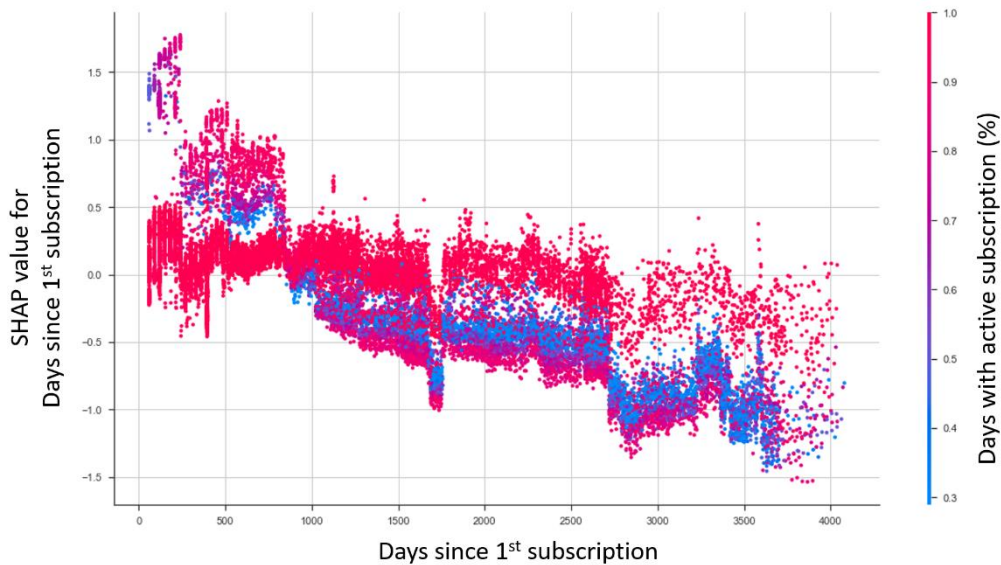


Figure 4.3 – Dependency plot for Days since 1st subscription

Furthermore, focusing on Figure 4.4, the same downward trend is portrayed in the total days with visits during the last two months of the subscription. Here the turnover moment is around 25 days with visits in the last 60 days of the current subscription. Using this information PÚBLICO tested an automated weekly newsletter to re-engage users who had less than 25 days with visits in the last two months of the subscription. With six newsletters sent, from all targeted subscribers who clicked the newsletters, 55% increased the number of active days and 18% maintained it. Furthermore, of all subscribers who received these newsletters, 67% renewed their subscription. However, considering only the ones who opened it, 70% renewed their subscription, whereas, from the ones who did not open it, 66% remained active users.

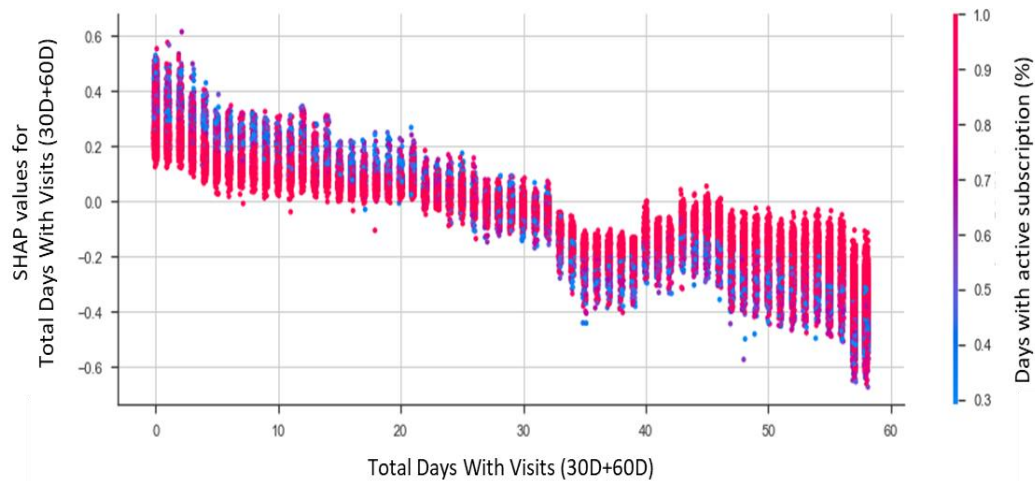


Figure 4.4 – Dependency plot for Total Days With Visits (30D+60D)

4.3. A/B TESTS

4.3.1. Model 1 – Recurrent and Non-recurrent Subscriptions

During the evaluation period of model 1, the model predicted 4.376 subscriptions, from which 1.003 (20%) were considered churners. Furthermore, 501 (50%) predicted churners were impacted by the retention campaign. Model 1 predicted a higher percentage of churn by 10pp in the evaluation group than it predicted when training and testing the model. The results of the experiment were much lower than expected. The F1 considering only the control group and users predicted as non-churners was 14.5%, which indicates that the model predicted poorly, as this group was expected to present results similar to the train and test sets defined during the model development. Moreover, considering all users predicted as non-churners and the variation group the F1 was 16%, which indicates that most of the predictions defined as churners were, in fact, non churners. However, these results do not directly indicate that the retention action was successful, as the results of the control group were much lower than expected.

To explain these results, it is important to understand the business context in which the model was trained and the context in which it was evaluated. One of the reasons for these results was the value range of the variable that represented the average index of the stringency of restrictive measures in Portugal to combat Covid-19, during the last month of the subscription. The existing range during the train and test periods was similar. However, the restrictive measures in Portugal during the experiment period were softer, comprising a range of values never considered by the trained model (Figure 4.5). It is worth reminding that this variable was the fourth most important variable for prediction considered by the model. Therefore, presenting input values never studied by the model downgraded the model's performance for prediction.

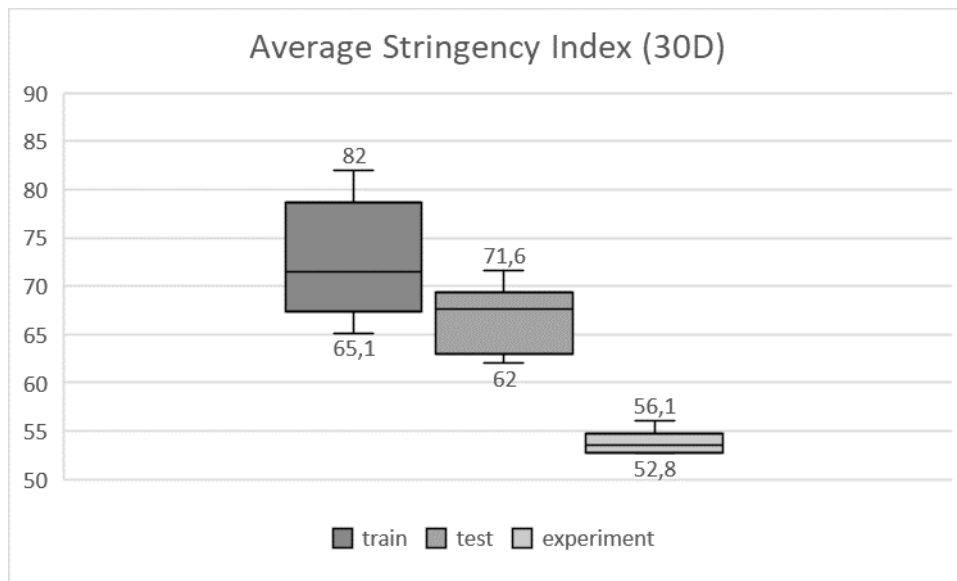


Figure 4.5 – Stringency Index 30D ranges for the three datasets

Secondly, starting from the second week of the experiment, a parallel marketing campaign was activated and visible to all website visitors. This second campaign offered three months of free content, for a yearly subscription. However, this campaign was only available for new users. While the retention campaign pushed during the experiment was offering one month, for a yearly subscription, for existing customers who were likely to churn. Competing against a better offer than the one being pushed and monitored for the experiment resulted in biased results. Customers, most likely, cancelled the current subscription and register with a new email, as new customers, leveraging the most appealing campaign.

Lastly, the model's training was 30 days prior to the churn moment (last subscription day). However, to impact customers with the retention campaign before the end of the current subscription, for the experiment period, predictions were carried out 37 days before the churn moment (seven days before the subscription's end day).

4.3.2. Model 2 – Non-recurrent Subscriptions

Due to the results attained during the 1st experiment it was decided that model 2 not use the average stringency index as a predictor to avoid the lack of representative data in the trained model. Furthermore, the predictions were computed 30 days before the churn moment, following the model's training framework. As a retention measure, the newspaper decided to send a newsletter with content about the war in Ukraine, to show subscribers the quality of the journalism provided by PÚBLICO, as well as a way to re-engage subscribers into using the website again.

Considering only non-recurrent subscriptions during the second experiment, 1.286 subscriptions were predicted, from which a churn rate of 30% was derived, while the real churn rate was 24,6%. Of subscribers predicted as churners, 193 (50%) were targeted with the retention actions, nonetheless, 114 of those subscribers cancelled their subscription 30 days after the last day of the subscription. As for the control group (194 subscribers), 108 subscribers cancelled their subscription. In other words, a higher percentage (59%) of subscribers cancelled when targeted with the retention actions, when compared to subscribers who were not approached by it (56%), nonetheless this is not a significant

difference to confirm that it is better not to contact customers who are ending their subscription. In addition, only 65 customers opened the re-engagement newsletter and, of those, 45% kept their subscription, on the other hand, 39% of the ones who did not open the newsletter renewed their subscription.

The call centre was able to contact 86 subscribers from the variation group to persuade them into renewing their subscriptions, from those, 56 answered the phone from which 43% renewed their subscription. In contrast, only 17% of subscribers who did not answer the phone call renewed their subscription, and 20% of customers without a phone number associated, thus not contacted via a phone call, renewed their subscription. Contacting consumers via a phone call to get them to renew their subscription proved to be significantly more successful when they answered the phone compared to when they did not answer the phone or were not contacted via phone call.

As the predictions in the second experiment comprised only subscriptions which were not automatically renewed, many subscribers did not recall their subscription was ending and thus were not going to renew their subscription. With the call centre phone call these subscribers were actively reminded and requested payment reference numbers to renew their subscriptions. Therefore, comparing the re-engagement newsletter to the phone call, the authors can conclude that less subtle retention actions at the end of the subscription proved to be a more effective action to retain subscribers.

In regards to model performance, analysing the subscribers within the control group and subscribers predicted as non-churners, the F1 metric was 54%. This result was higher than the first experiment however it is not equivalent to the results attained with the train, validation and test datasets.

Table 4.3 – Results of the experiment on only non-recurrent subscriptions

Dataset	accuracy	balanced accuracy	precision	recall	f1	specificity
TRAIN	95,64%	93,21%	85,96%	89,48%	87,69%	96,93%
VALIDATION	93,70%	92,62%	84,92%	90,49%	87,62%	94,75%
TEST	93,64%	92,57%	82,45%	90,64%	86,35%	94,49%
experiment (non-churner & control)	83,44%	72,55%	53,20%	55,67%	54,41%	89,43%

The model was trained during a period when subscribers were looking to get credible information, due to the pandemic, but evaluated during a period when Covid-19 was no longer a novelty and subscribers could no longer have the same interest in purchasing a subscription (Newman, Fletcher, Robertson, Eddy, & Nielsen, 2022). Moreover, literature has shown that the overwhelming amount of news related to the pandemic is triggering news fatigue, which leads them to avoid news consumption (Fitzpatrick, 2022; Mitchell, Oliphant, & Shearer, 2020). When the Ukraine war began, the extensive news coverage of the war brought back the same kind of negative news headlines seen during Covid-19. Reuters Institute reported an accelerated increase in news avoidance, having countries such as Germany increasing news avoidance by 7pp, in two months, compared to the 5pp increase seen in five years from 2017 to 2022 (Eddy & Fletcher, 2022). This kind of news avoidance behaviour might have affected the model performance in prediction, as the decrease in engagement in news readership was also seen in the data provided by PÚBLICO. The average number of days with visits and the total

amount of articles read during the last subscription month have been decreasing over time. Showing that PÚBLICO's subscribers may also be being affected by news fatigue.

5. CONCLUSION

Following a long period of a downwards trend in newspapers' print readership, newspapers have been increasingly adopting digital business models that comprehend online subscriptions of articles. Compared to other online services, the newspaper industry still has low customer adherence. Nevertheless, the quantity of subscribers to online newspaper services has been increasing over the years. During the past two years, a lack of knowledge on top of misinformation on the topic of the Covid-19 pandemic has contributed to an increase in newspaper subscriptions (Newman et al., 2021), as the population searches for reliable sources to be informed.

PÚBLICO increased its subscribers' database and required a method to predict probable churning subscribers to implement preventive measures. To attain this objective, two models were built and tested. One model containing all subscribers and a second model containing only non-recurring subscribers. For the first model, a dataset containing 47.449 subscriptions and 9.6% of churners was used to train and test a XGBoost model, covering a period of subscriptions ending between November 9th, 2020, and June 21st, 2021. For the second model a dataset containing 18.461 non-recurring subscriptions, ending between November 9th, 2020, and January 11th, 2022 was used. Both datasets were provided with 44 predictive features reflecting the website's usability, product specifications, socio-demographic data and user historical data. Information on the stringency of government measures in Portugal was added to the first setup's dataset. Both experiments followed a rigorous and extensive preprocessing process, in which tasks of data cleaning, transformation and reduction were employed to enhance the predictive performance of ML models by removing and reducing noisy data.

Model 1 was evaluated in a three-week experiment. However, the results were not equivalently as good as the results attained in the train and test datasets. The main reason for this was a more appealing marketing campaign pushed simultaneously with the retention measure. The retention measure pushed to probable churning customers offered one free month of subscription if the users purchased a yearly subscription. Whereas the concurring marketing campaign activated during the experiment, available for everyone who went to the website, offered three months of subscription if the users purchased a yearly subscription. Competing against a more appealing campaign translated into poor results in the experiment. Furthermore, the contrasting contexts in which each dataset was evaluated also contributed to this difference in the results. During the experiment, the stringency of the restrictive measures to combat Covid-19 in Portugal, the fourth most predictive variable of model 1, had ranges never introduced to the model while training, thus leading to lower model quality.

The second model was evaluated in a four-week experiment in a more stable context in terms of marketing actions. Subscribers in the variation group were targeted with a re-engagement newsletter with content about the war in Ukraine and contacted via mobile phone whenever possible. A higher percentage of customers in the variation group churned (59%) when compared to the control group (56%). However, among the targeted customers who opened the re-engagement newsletter (33%), 45% kept their subscription, while only 39% of the ones who did not open the newsletter also kept it. Furthermore, successfully contacting probable churning customers via a phone call has proven to be significantly more effective in maintaining them than not contacting them. 43% of customers who were contacted by the call centre renewed their subscription, while only 17% of the ones who did not answer and 20% of customers without a phone number renewed their subscription. Customers who are contacted with a more direct retention action, such as a phone call, right after their subscription

has ended, are more likely to renew their subscription, in a period of 30 days, than customers who do not engage with this contact. Therefore, it would be advisable that PÚBLICO attempts to contact a second time customers who did not answer the phone call, as well as implement a tactic to update their database with customer phone numbers.

In addition, insights on predictors show that users with lower fidelity rates and few subscriptions presented a higher propensity to cancel their subscriptions. Product engagement variables were not found to be the most important variables for churn prediction, nonetheless, they go in line with what was found in blog posts (American Press Institute, 2018; Financial Times, n.d.; Veseling, 2018): higher engagement with the product is related to lower churn. The insights on the number of days with visits in the last two subscription months supported PÚBLICO's decision to further use the re-engagement newsletter tested to impact users who had less than 42% of days with visits to the website.

Nonetheless, literature has shown that the amount of news about the Covid-19 pandemic is triggering news fatigue, thus leading people to bypass news consumption (Fitzpatrick, 2022). In this project, that same reaction is depicted by the importance of the stringency index in model 1, which relates lower stringency values to higher churn probabilities. On top of that, a rapid and sharp increase in news avoidance is being captured in Europe (Eddy & Fletcher, 2022) with the Ukraine war news coverage. This could also explain the lower models' performance in both evaluation periods as they were trained in periods where subscribers were seeking credible information, but evaluated in periods where subscribers no longer felt that need.

This project report fills the gap in the literature related to churn prediction in online news media subscriptions, and it can serve as a basis for future research in this area. Furthermore, the findings described in this project report can support other newspapers, with similar business models, to implement and evaluate their retention strategies. As for next steps, it would be advisable that PÚBLICO studies other indicators of website engagement that could improve prediction performance and provide more understanding of the churn decision, such as preferred access channel, most used device, time on site, interactions with the customer via the call centre and purchase channel. A churn analysis before the end of the subscription should be carried out, to implement retention actions before the subscription ends, to not overwhelm the call centre with phone calls at the end of the subscription and be able to implement automated re-engagement measures. In addition, more retention actions should be tested, as a means to implement distinct and personalised actions for customers with different probabilities to churn.

REFERENCES

- American Press Institute. (2018). Paths to Subscription: Why recent subscribers chose to pay for news. Retrieved from <https://www.americanpressinstitute.org/publications/reports/survey-research/paths-to-subscription/single-page/>
- Antonio, N., de Almeida, A. M., Nunes, L., Batista, F., & Ribeiro, R. (2018). Hotel online reviews: creating a multi-source aggregated index. *International Journal of Contemporary Hospitality Management*, 30(12), 3574–3591. <https://doi.org/10.1108/IJCHM-05-2017-0302>
- Arrese, Á. (2016). From Gratis to Paywalls: A brief history of a retro-innovation in the press's business. *Journalism Studies*, 17(8), 1051–1067. <https://doi.org/10.1080/1461670X.2015.1027788>
- Burez, J., & Van den Poel, D. (2007). CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, 32(2), 277–288. <https://doi.org/10.1016/j.eswa.2005.11.037>
- Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3 PART 1), 4626–4636. <https://doi.org/10.1016/j.eswa.2008.05.027>
- Campbell, C. (2018). *Engaged Readers Don't Churn: Retention lessons for digital subscriptions*. (August). Retrieved from https://datamathstat.files.wordpress.com/2018/09/wan-ifra_reader_revenue-1.pdf
- Cardoso, G., Paisana, M., & Pinto-Martinho, A. (2020). Digital News Report 2020 - Portugal Data. In *OberCom Reuters*.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Rudiger, W. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, (3), 76.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/https://doi.org/10.1613/jair.953>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. New York, NY, USA: ACM. <https://doi.org/10.1145/2939672.2939785>
- Coussement, K., & Bock, K. W. De. (2013). Customer churn prediction in the online gambling industry : The bene fi cial effect of ensemble learning. *Journal of Business Research*, 66(9), 1629–1636. <https://doi.org/10.1016/j.jbusres.2012.12.008>
- Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, 95, 27–36. <https://doi.org/10.1016/j.dss.2016.11.007>
- Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, 34(1), 313–327. <https://doi.org/10.1016/j.eswa.2006.09.038>
- Coussement, K., & Van den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert*

- Systems with Applications*, 36(3 PART 2), 6127–6134.
<https://doi.org/10.1016/j.eswa.2008.07.021>
- Davoudi, H., An, A., Zihayat, M., & Edall, G. (2018). Adaptive Paywall Mechanism for Digital News Media Heidar. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2018)* 205–214, 205–214.
<https://doi.org/10.1145/3219819.3219892>
- De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, 269(2), 760–772. <https://doi.org/10.1016/j.ejor.2018.02.009>
- Eddy, K., & Fletcher, R. (2022). *Perceptions of media coverage of the war in Ukraine*. Retrieved from <https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022/perceptions-media-coverage-war-Ukraine>
- Ekholm, P. (2019). Machine Learning Challenge Winning Solutions. Retrieved from <https://medium.com/mittmedia/we-looked-into-why-our-subscribers-churned-with-machine-learning-d1ac1adf8009>
- Farquad, M. A. H., Ravi, V., & Raju, S. B. (2014). Churn prediction using comprehensible support vector machine: An analytical CRM application. *Applied Soft Computing Journal*, 19, 31–40.
<https://doi.org/10.1016/j.asoc.2014.01.031>
- Financial Times. (n.d.). Lessons on retention from the FT (and beyond). Retrieved from <https://www.ftstrategies.com/en-gb/insights/lessons-on-retention-from-the-ft-and-beyond/>
- Fitzpatrick, N. (2022). No News is Not Good News: The Implications of News Fatigue and News Avoidance in a Pandemic World. *Athens Journal of Mass Media and Communications*, 8(3), 145–160. <https://doi.org/10.30958/ajmmc.8-3-1>
- Fletcher, R., & Nielsen, R. K. (2017). Paying for Online News: A comparative analysis of six countries. *Digital Journalism*, 5(9), 1173–1191. <https://doi.org/10.1080/21670811.2016.1246373>
- Främling, K., Westberg, M., Jullum, M., Madhikermi, M., & Malhi, A. (2021). Comparison of Contextual Importance and Utility with LIME and Shapley Values. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12688 LNAI, 39–54. https://doi.org/10.1007/978-3-030-82017-6_3
- Ganesh, J., Arnold, M. J., & Reynolds, K. E. (2000). Understanding the customer base of service providers: An examination of the differences between switchers and stayers. *Journal of Marketing*, 64(3), 65–87. <https://doi.org/10.1509/jmkg.64.3.65.18028>
- Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision making and a “right to explanation.” *AI Magazine*, 38(3), 50–57. <https://doi.org/10.1609/aimag.v38i3.2741>
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5), 1–45.
- Gunnarsson, C. L., Walker, M. M., Walatka, V., & Swann, K. (2007). Lessons learned: A case study using data mining in the newspaper industry. *Journal of Database Marketing & Customer Strategy Management*, 14(4), 271–280. <https://doi.org/10.1057/palgrave.dbm.3250058>
- Han, J., Kamber, M., & Pei, J. (2012). Data Mining. Concepts and Techniques. 3rd Edition. In *Journal of Chemical Information and Modeling* (Vol. 53).

- Höppner, S., Stripling, E., Baesens, B., Broucke, S. vanden, & Verdonck, T. (2017). Profit driven decision trees for churn prediction. *European Journal of Operational Research*, 284(3), 920–933. <https://doi.org/10.1016/j.ejor.2018.11.072>
- Jayaswal, P., Prasad, B. R., Tomar, D., & Agarwal, S. (2016). An Ensemble Approach for Efficient Churn Prediction in Telecom Industry. *International Journal of Database Theory and Application*, 9(8), 211–232.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154.
- Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing Journal*, 24, 994–1012. <https://doi.org/10.1016/j.asoc.2014.08.041>
- Kotler, P., Kartajaya, H., & Setiawan, I. (2017). *Marketing 4.0: Moving from Traditional to Digital*.
- Kueng, L. (2017). *Going Digital*.
- Lemmens, A., & Croux, C. (2006). Bagging and Boosting Classification Trees to Predict Churn. *Journal of Marketing Research*, 43(2), 276–286.
- Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3), 31–57. <https://doi.org/10.1145/3233231>
- Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). *Consistent Individualized Feature Attribution for Tree Ensembles*. (2). Retrieved from <http://arxiv.org/abs/1802.03888>
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 2017-Decem*(Section 2), 4766–4775.
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). <https://doi.org/10.25080/Majora-92bf1922-00a>
- Mitchell, A., Oliphant, J. B., & Shearer, E. (2020). *About Seven-in-Ten U.S. Adults Say They Need to Take Breaks From COVID-19 News*. Retrieved from <https://www.pewresearch.org/journalism/2020/04/29/about-seven-in-ten-u-s-adults-say-they-need-to-take-breaks-from-covid-19-news/>
- Moisen, G. G. (2008). Classification and regression trees. *Classification and Regression Trees*, (2000), 582–588. <https://doi.org/10.1201/9781315139470>
- Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, 43(2), 204–211. <https://doi.org/10.1509/jmkr.43.2.204>
- Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., & Nielsen, R. K. (2022). *Reuters Institute Digital News Report 2022*. Retrieved from <https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022>
- Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C. T., & Nielsen, R. K. (2021). *Reuters Institute Digital News Report 2021*.
- Nitzan, I., & Libai, B. (2011). Social effects on customer retention. *Journal of Marketing*, 75(6), 24–38.

<https://doi.org/10.1509/jm.10.0209>

- Pattabhiramaiah, A., Sriram, S., & Manchanda, P. (2018). Paywalls: Monetizing Online Content. *SSRN Electronic Journal*, 83(2), 19–36. <https://doi.org/10.2139/ssrn.2916820>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Petherick, A., Kira, B., Cameron-Blake, E., Tatlow, H., Hallas, L., Hale, T., ... Wood, A. (2021). *Variation in government responses to COVID-19*. Retrieved from <https://covidtracker.bsg.ox.ac.uk/stringency-scatter>
- Pyle, D. (1999). Data Preparation for Data Mining. In *Applied Artificial Intelligence* (1st ed., Vol. 17). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Retaining subscribers: Key learnings from The Atlantic and MediaNews Group. (2021). Retrieved from <https://wan-ifra.org/2021/10/digital-media-latam-2021-retention/>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?” Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, 97–101. <https://doi.org/10.18653/v1/n16-3020>
- Richter, Y., Yom-Tov, E., & Slonim, N. (2010). Predicting customer churn in mobile networks through analysis of social groups. *Proceedings of the 2010 SIAM International Conference on Data Mining*, 732–741. Philadelphia, PA: Society for Industrial and Applied Mathematics. <https://doi.org/10.1137/1.9781611972801.64>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Rußell, R., Berger, B., Stich, L., Hess, T., & Spann, M. (2020). Monetizing Online Content: Digital Paywall Design and Configuration. *Business & Information Systems Engineering*, 62(3), 253–260. <https://doi.org/10.1007/s12599-020-00632-5>
- Sonderman, J., & Vargo, G. (2021). 9 subscriber retention strategies. Retrieved from <https://www.americanpressinstitute.org/publications/reports/survey-research/9-subscriber-retention-strategies/>
- Southern, L. (2019). How Swedish newspaper Dagens Nyheter halved subscriber churn in 2 years. Retrieved from <https://digiday.com/media/how-swedish-newspaper-dagens-nyheter-halved-churn-to-8-in-2-years/>
- SparkBeyond Team. (2021). Reducing churn for a European media company. Retrieved from <https://www.sparkbeyond.com/case-study/reducing-churn-for-a-european-media-company>
- Suárez, E. (2019). *How to build a successful subscription news business: Lessons from Britain and Spain*. (February), 61–67. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-02/Eduardo_Suarez_Digital_subscriptions_fellowship_paper.pdf
- Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4), 687–719. <https://doi.org/10.1142/S0218001409007326>

- Tianyuan, Z., & Moro, S. (2021). Research Trends in Customer Churn Prediction: A Data Mining Approach. *Advances in Intelligent Systems and Computing*, 1365 AIST(351), 227–237. https://doi.org/10.1007/978-3-030-72657-7_22
- Tsai, C., & Lu, Y. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems With Applications*, 36(10), 12547–12553. <https://doi.org/10.1016/j.eswa.2009.05.032>
- Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1), 211–229. <https://doi.org/10.1016/j.ejor.2011.09.031>
- Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, 38(3), 2354–2364. <https://doi.org/10.1016/j.eswa.2010.08.023>
- Veseling, B. (2018). The road to 1 million subscribers: FT helps users to build ‘habit.’ Retrieved from <https://wan-ifra.org/2018/04/the-road-to-1-million-subscribers-ft-helps-users-to-build-habit/>
- Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). *Understanding Neural Networks Through Deep Visualization*. Retrieved from <http://arxiv.org/abs/1506.06579>
- Yuan, S., Bai, S., Song, M., & Zhou, Z. (2017). Customer Churn Prediction in the Online New Media Platform: A Case Study on Juzi Entertainment. *2017 International Conference on Platform Technology and Service, PlatCon 2017 - Proceedings*, 0–4. <https://doi.org/10.1109/PlatCon.2017.7883698>