IADIS International Journal on WWW/Internet
Vol. 21, No. 1, pp. 59-73
ISSN: 1645-7641

DETECTING POOR-QUALITY WEB TRAFFIC IN VIDEO AD MARKETING CAMPAIGNS BASED ON ANALYTICAL METHODS

Marta López and Fernando Perales
JOT INTERNET MEDIA España SL
General Ramírez de Madrid 8, 28020 Madrid, Spain

ABSTRACT

In digital marketing, performance data can be used to determine which variables affect the quality of web traffic generated. There exist different ways of collecting data, which enable the data enrichment and give us a lever to generate new ways of looking at and dealing with information. And yet, a large amount of data, the daily variation and the different levels of granularity require dedicated data preparation. Our data needs prior preparation and cleaning to be able to develop the multivariate and correlation analysis through Cramer's V-diagram. Our objective is to analyze and predict the traffic quality score, a parameter that is not only used by companies to verify the suitability of the audience concerned but is also directly related to the bandwidth awareness and profitability of video advertising campaigns. Analytical identification of the variables that most affect traffic quality as well as the quality prediction can be used to optimize both marketing campaign structure and bidding strategy.

KEYWORDS

Video AD Marketing, Quality Prediction, Data Preparation, Multi-Variant Analysis, Digital Marketing, Web Traffic

1. INTRODUCTION

During the last few years, video marketing has been proven to be a highly efficient medium to engage the targeted audience at reasonable prices with good performance indicators (clicks and return on investment). These aspects motivate the global video ad investment, with a forecast of continuous growth of 12% till 2025, reaching $148 B. This huge market means that an enormous amount of web traffic, representing potential clients, is redirected to the company's websites. However, based on the information reported by the video ad platforms dealing with JOT video campaigns (YouTube), between 20-30% of this traffic comes from inauthentic users,

mainly bots working through virtual protocol networks (VPNs) and programmed to spend the companies´ marketing budget by clicking in the ads with no post-conversion, reducing marketing campaigns impact and decreasing the return of investment.

Due to this fact, in the market there exist several companies offering inauthentic traffic detection services (Click Cease, Click Guard, Traffic Guard, Traffic Cop and PPC Protect). However, none of them have access to historical data of the campaign statistics so they need a long learning process to be able to define and implement reliable and robust traffic filters and traffic blockers. In addition, some of them work on IP addresses which adds a GDPR-related challenge to the development.

Considering this market status, the access to an already existing database describing the video ad campaign web traffic (YouTube) of more than 12 websites during 6 months enables the development of unique web traffic quality analysis, identifying which are the most relevant parameters conditioning the web traffic quality and the development of dedicated traffic prediction service. This disposal also ensures that algorithms are trained, tested, and validated with real data, reaching the required insights level of quality and accuracy to be used in production conditions for decision-making. The dataset generated compiles a set of 50 variables (combining performance indicators and descriptors as attributes).

At the business level, by the multi-variant analysis for the identification of low-quality traffic sources, it is possible: (i) to increase the trust in video ad marketing as a reliable tool for business promotion, ensuring that all the users reaching companies' landing pages are real potential clients, so the marketing budget is fully spent in high-quality audiences; and, (ii) the direct correlation of the video content to the user audience increase the user experience confidence, so the companies will get more and more interactions and visits through this media content.

Considering all the previous aspects, the goal of the paper is to answer the following questions:

- What are the main variables impacting the quality of the traffic?
- Is there any correlation among the variables used to describe the impact of the marketing campaigns?
- How can traffic quality be predicted based on the analysis of historical data?

The paper is organized as follows. Section II presents the technical background related to multi-variant analysis and its relation to quality prediction. In Section III, the required work to generate the dataset is presented, covering data collection, key existing indicators and ETLs (Extraction, Transformation and Loading) needed for data merging. Section IV the technical analysis is presented and discussed. Section V details the development of the predictive models for traffic quality score, including data transformations and classifications. Finally, Section VI highlights key conclusions and open issues for the future.

2. BACKGROUND

The easy and wide access to Digital Marketing (DM) services has allowed the massive and periodic data collection of variables and indicators describing the performance of marketing campaigns. This has pushed the marketing teams to approach data science as the only solution to process and analyze the vast amount of data available. Due to the large number of variables involved, one of the first goals is to understand their importance and correlation depending on the goals of the company, which allow the identification of the target parameter.

Some references are describing the different approaches for multivariant analysis (Black, & Babin, 2019). and (Everitt, & Hothorn, 2011) and specifically applied to marketing (Saura, 2021). They all state that in almost all cases data preparation represents a significant bottleneck for marketing data processing, even conditioning the final quality of the results. For that reason, in our case the final data set used in the analysis has been specifically prepared, enabling both temporal analysis and correlations depending on the target variables: traffic quality (discrete variable) and conversion difference (continuous variable obtained by the difference in the reporting values for the same day with 7 days delay).

The final goal of this work is to generate a decision support system enabling the detection of inauthentic traffic sources based on real performance data analysis and predictive services, so the confidence and trust in the initial data must be maximized. For the generation of the predictive service, the temporal series of the variables must be considered. In (Neusser, 2016) and (Hyndman, & Khandakar, 2008) the two conditions to develop this kind of service are defined:

1. availability of data about the past
2. the assumption that past patterns will be replicated in the future

These two conditions are fully satisfied in our case, as data sets are collected daily from the past month's video marketing campaigns, also society interests show periodic patterns depending on the seasonality and traffic quality can be described based on campaign indicators like the category and the keywords, which are 100% monitored.

The different stages of the process to develop the multivariant analysis are as follows (Arroyo Barrigüete, Fabra Florit, & Redondo Palomo, 2023) and (Salgado, F. B. P. L. , s. f.)

- *Objectives of the analysis*: This is the basis on which the technical work must be guided. The objective of the analysis is to determine whether there are correlations between the different variables that cannot be seen at first glance or by expert visualization and evaluate their influence on our target variables. The final goal is to understand how the quality of traffic to our ads can be improved to optimize the campaign results.
- *Selection of data*: As mentioned earlier, this is a crucial step. The right choice and quality of data determine to a large extent the satisfaction of the conclusions. It must be determined the time range studied, the sample size and the explanatory variables to be observed. Starting from the assumption of independence, it is analyzed whether this is fulfilled.
- *Selection of the target variables*: Representing the variables for which the analysis with show how the rest of the parameters affect. It can be one or several depending on the results obtained. In addition, data has been processed differently depending on its type. There are two types of target variables: (i) Quantitative variable (Composed of numerical values) and (ii) Qualitative variable (Composed of factors).
- *Data cleaning*: Before any type of analysis, the data must be cleaned to ensure that they are optimal. Among the cleaning tasks, a preliminary analysis is carried out, plotting the variables to see their distribution as well as their main statistics. This has allowed me to decide how to transform the variables to select their correct typology. This analysis also provides information about the distributions of the different variables and their possible statistical transformations to create a predictive model. At the same time, the data cleaning also manages the outliers and missing values. Once this initial treatment is developed, the variables will be ready for use in a single dataset.
- *Analysis*: A treatment of correlations has been carried out, both between individual variables that are relevant to our study and between all the inputs. It has been carried

out thanks to the correlation matrix (Cario, & Nelson, 1997). In this way, it has been possible to analyze the interdependence and avoid multi-collinearity in future predictions. In this case, it is a very useful step to see relationships that cannot have been seen otherwise. It has been also seen how the inputs affect the order of importance of the target variables using Cramer's V-plot (Isea, Ojeda, Fernandez, Gutierrez., & Salazar, s. f.).

- *Conclusions*: This is the final step of our analysis. Based on all the data processed, infer the results and whether the hypothesis set out at the beginning has been fulfilled

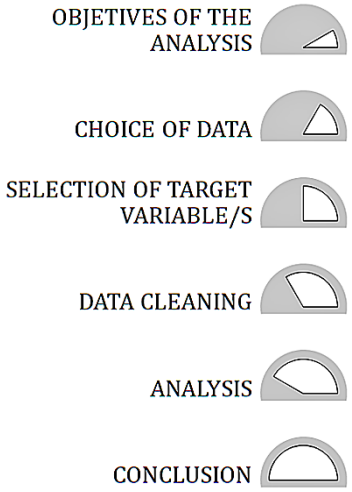

Figure 1. Steps of the data analysis

Massive web traffic generation affects different sectors, although the automation of digital marketing campaigns both in video and search platforms is of great relevance in digital marketing. There are several technical analyses and investigations carried out (Afandi, Bukhari, Khan, Maqsood, & Khan, 2022), (Stevanovic, & Pedersen, 2015) and (Efthimion, Payne, & Proferes, 2018) as an example of the detection of bots and botnets using different techniques such as classification models to avoid fraudulent traffic and using different data sources and platforms like Twitter) with the objective of understanding and avoiding fraudulent traffic. In this work, we aim to go one step further in analyzing fifty KPIs describing the performance of the digital campaign with more than two thousand registers and data with a temporal perspective. This represents an opportunity to discover new ways of improving the reliability of JOT marketing campaigns and ensure the quality of the ads.

3. DATA MARK GENERATION

The most sensible part of any analytical project is the generation of the final dataset. The data included always impacts the results and insights strongly. Therefore, it must be as complete as

possible. Also, the knowledge of the context and quality of the inputs support the decision of choosing the best database to use.

In this case, several data sources have been used:

• *Google Ads*: Providing historical information about the impact statistics and investment of the campaigns, showing the cost information used to measure the profitability and the performance of ads. To do the data crossing with the other databases the ad id and date are used as they are unique identifiers.

• *Partner report*: Showing the economic results (revenue) together with the results of each advertisement once they have passed the corresponding traffic quality filters. This report also includes information on traffic quality per ad.

• *Intern information*: Containing internal information of the accounts, like identifiers enabling the data merging between different tables. It is essential to the crossing because it relates the ad id to the source tag (provided by the partner and includes all the landing pages of the same client) and accounts, campaigns, and ad information.

• *Partner traffic quality report*: Containing a critical part of our analysis, because shows the traffic quality of our ads grouped by a tag that allows distinct the different landings. This report is different from the others. In this case, the data is grouped by source tag. For that reason, the report includes the information by country too, so it can be crossed with the rest of the data sources.

Both the definition of the problem and the knowledge of the project goals at the technical and business level represent adequate guidance for the analysis and predictions. If the data does not feed the predictive model with relevant inputs, the results will be surely poor and with low accuracy.

Data download has been automated through an API and collected in our SQL database. Data access has been customized through SQL queries directly connected to excel, where four tabs corresponding to the four data sources are generated. Within the excel framework, data has been transformed and merged using power query, which allows all the data updated and joined.

The specific technical details of each data source are the following:

- The final dataset (Figure 2) resulting from the data crosses. It includes 50 variables and allows us to begin with the multi-variant analysis.

```
## tibble [237,409 x 50] (S3: tbl_df/tbl/data.frame)
##  $ Fecha               : POSIXct[1:237409], format: "2021-12-01" "2021-12-01" ...
##  $ Dia_sem             : chr [1:237409] "miércoles" "miércoles" "miércoles" "miércoles" ...
##  $ Dia                 : num [1:237409] 1 1 1 1 1 1 1 1 1 1 ...
##  $ Mes                 : num [1:237409] 12 12 12 12 12 12 12 12 12 12 ...
##  $ Año                 : num [1:237409] 2021 2021 2021 2021 2021 ...
##  $ Cuenta_id           : chr [1:237409] "1238797147" "1238797147" "1423963431" "1423963431" ...
##  $ Dispositivo_revenue : chr [1:237409] "Desktop" "Desktop" "Desktop" "Desktop" ...
##  $ Source_tag          : chr [1:237409] "cbs_d2s_xmlb_2236_answersite_gdn2" "cbs_d2s_xmlb_2236_answer
site_gdn2" "cbs_d2s_xmlb_2236_helpwire_gdn" "cbs_d2s_xmlb_2236_helpwire_gdn" ...
##  $ Source_tag_numero   : num [1:237409] 2 2 1 1 1 1 1 1 1 1 ...
##  $ Destino_codigo      : chr [1:237409] "ANS" "ANS" "HEL" "HEL" ...
##  $ Pais                : chr [1:237409] "USA" "USA" "USA" "USA" ...
##  $ Pais_revenue        : chr [1:237409] "USA" "USA" "CAN" "USA" ...
##  $ Categoria           : chr [1:237409] "Insurance" "Insurance" "VideoMeeting" "VideoMeeting" ...
##  $ Keyword             : chr [1:237409] "insurify insurance quotes" "botox for tension headaches cove
red by insurance" "video meeting system" "video meeting system" ...
##  $ Coste_i             : num [1:237409] 0 0.629 1.979 1.979 1.197 ...
##  $ Coste_f             : num [1:237409] 0 0 0 0 0 ...
##  $ Cost_diff           : num [1:237409] 0 1 1 1 1 ...
##  $ Revenue             : num [1:237409] 2.202 0.479 0.607 7.126 0 ...
##  $ Profit_i            : num [1:237409] 2.2 -0.15 -1.37 5.15 -1.2 ...
##  $ Profit_f            : num [1:237409] 2.202 0.479 0.607 7.126 0 ...
##  $ ROAS_i              : num [1:237409] 2.202 -0.239 -0.693 2.601 -1 ...
##  $ ROAS_f              : num [1:237409] 2.202 0.479 0.607 7.126 0 ...
##  $ CPC_i               : num [1:237409] 0 0.629 1.979 0.165 0.599 ...
##  $ CPC_f               : num [1:237409] 0 0 0 0 0 ...
##  $ RPC                 : num [1:237409] 1.101 0.479 0.607 3.563 0 ...
##  $ CPA_JOT_i           : num [1:237409] 0 0.629 1.979 0.99 0 ...
##  $ CPA_JOT_f           : num [1:237409] 0 0 0 0 0 ...
##  $ CPA_G_i             : num [1:237409] 0 0 0 0 0 0 0 0 0 0 ...
##  $ CPA_G_f             : num [1:237409] 0 0 0 0 0 0 0 0 0 0 ...
##  $ CR_JOT_i            : num [1:237409] 0 1 0.5 1 0 0.5 0.2 1 0 0 ...
##  $ CR_JOT_f            : num [1:237409] 0 0 1 2 0 0 0.25 1.25 0 0 ...
##  $ CR_G_i              : num [1:237409] 0 0 0 0 0 0 0 0 0 0 ...
##  $ CR_G_f              : num [1:237409] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Impresiones_i       : num [1:237409] 0 1 4 4 2 2 8 8 2 0 ...
##  $ Impresiones_f       : num [1:237409] 0 0 0 0 0 0 6 6 1 0 ...
##  $ Impressions_revenue : num [1:237409] 1 1 1 12 2 3 1 9 8 1 ...
##  $ Clicks_i            : num [1:237409] 0 1 2 2 2 2 5 5 0 0 ...
##  $ Clicks_f            : num [1:237409] 0 0 1 1 0 0 4 4 0 0 ...
##  $ Clicks_revenue      : num [1:237409] 1 1 1 12 2 3 1 9 8 1 ...
##  $ Conversiones_i      : num [1:237409] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Conversiones_f      : num [1:237409] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Conversiones_revenue: num [1:237409] 2 1 1 2 0 1 1 5 1 1 ...
##  $ Conv_diff_i         : num [1:237409] 2 1 1 2 0 1 1 5 1 1 ...
##  $ Conv_diff_f         : num [1:237409] 2 1 1 2 0 1 1 5 1 1 ...
##  $ Conv_diff_i_%       : num [1:237409] 2 1 1 2 0 1 1 5 1 1 ...
##  $ Conv_diff_f _%      : num [1:237409] 2 1 1 2 0 1 1 5 1 1 ...
##  $ Vistas_video_i      : num [1:237409] 0 1 5 5 2 2 6 6 1 0 ...
##  $ Vistas_video_f      : num [1:237409] 0 0 1 1 0 0 4 4 1 0 ...
##  $ TQ_TypeTag          : num [1:237409] 0 0 2.5 2.5 0 2 3 3 0.5 0 ...
##  $ TQ_SourceTag        : num [1:237409] 2 2 3 3 3 3 3 3 3 3 ...
```

Figure 2. Set of variables included in the dataset combining numeric and categorical variables

To develop the most complete analysis possible, all the potential variables to which we have access have been included. They describe: (i) the financial performance (Cost, revenue, Cost Per Click, Cost Per Acquisition, Revenue Per Click), (ii) the date (day of the week, month, year, date), (iii) the marketing account specifications (id, device, country, code, category, keyword, destiny, source tag) and (iv) performance indicators (clicks, impressions, conversion rate, conversions, traffic quality). These values are collected two times (initial and final), before and after Google traffic invalidation, and their differences are also analyzed.

Thanks to the disposal of this high-quality data, it is possible to carry out a wide catalog of plots for visual inspection, like temporal representations, histograms, data exploratory for error detection and outliers, and so on. This work is part of the data preparation rather than the analysis for predictive modeling so no further details are included. The first benefit of doing this type of representation is that the type of variable is quite clear to appreciate, and it gives us feedback for the next steps. For example, some factor-type variables (categorical) are wrongly coded as a character. Others, such as the day of the week, month, or year, can be changed to factor due to the low number of categories.

4. MULTIVARIANT ANALYSIS AND DISCUSSION

As mentioned previously, this analysis aims to understand what variables affect the traffic quality in video ad marketing campaigns. To do so, two target variables have been selected:

(i) *Conversion difference*, It's the difference in conversions between the cost and revenue report, which explains the invalidation of our ads. We will consider it as the indicator of BOTs, the quantitative variable,

(ii) *TQ (traffic quality)* at source tag level, which gives us information about what value on a scale from 0 to 10 our ads have, where 10 is the highest value.

In general, our ads have a conversion difference above one hundred percent (More conversions in revenue than cost) and a quality score equal to or lower than 3. This makes sense, the existence of bots translates into a high conversion difference and low-quality score. It's because we used these two target variables to investigate more about it. The rest of the variables are considered explanatory or input variables. Therefore, the first action implemented has been to correct the wrong data typologies. This allows us to plot the distribution of the variables where more information was obtained:

(i) Many of the numerical variables show a right tail. It implies that they are skewed variables, with most of the data clustered in the first values of the variable. It would be interesting to analyze whether a transformation to normalize the data will be useful. The most common using transformations are logarithms, exponentials, and roots, but exists models like h2o that do it automatically.

(ii) Another advantage is that potential groupings of the categorical variables are identified to make them more efficient, and it helps us to improve our conclusions.

Given the fact that the data is classified by dates, it was convenient to consider whether there were temporal patterns. Not relevant at all in input variables, but yes in target variables. The invalidation of our ads decreased over time while at the same time, the traffic quality increased. Probably because of Google's actions to improve the quality of the algorithm traffic, avoiding fraudulent traffic.

Once implemented the initial inspection of the data, several transformations are carried out for the data cleaning, modifying the types of variables, and transforming those that can be grouped. In our case, the traffic quality values have been transformed from numerical to categorical (**Error! Reference source not found.**), grouping them into a factor type with four categories (lowest, low, medium, and high), so that it is easier to analyze them.

Table 1. Distribution of ad quality classification

Quality Classification	% of ads	TQ number range
Lowest	3.05	0
Low	86.93	1-3
Medium	7.39	4-7
High	0	8-10

The next step was the analysis of outliers (either very large or very small). The percentage of outliers per variable should be calculated to determine their potential impact and relevance in the data set. Based on the theory, if the result is less than or equal to 10% of the data, the outliers could be transformed into missing values, as they may be considered to have a low incidence. Almost all the variables were below this threshold, so it is considered that they are not impacting the results and can be transformed into missing values. In any case, before that, we will create a new variable to analyze the impact of the outliers, it will be called "Prop_missing" and it contains the provision of outliers.

Finally, missing data were treated, so that if they exceed 50% of the data, the variable should be deleted as it would not yield any information. It is usual to impute missing data by the mean, median or both randomly, as they are then replaced by data that follow the same pattern. In our case the incidence was less than 10%, so we imputed randomly, eliminating any missing values.

Once our database was treated and free of significant outliers and missing values, the complete dataset was analyzed using Cramer's V charts (**Error! Reference source not found.** and **Error! Reference source not found.**). Cramer's V is a measure of the strength of association between two nominal variables based on chi-square, and Cramer's V chart returns in descending order the most significant variables about the target variable. To increase the robustness and reliability of the results and as a quality control, two random variables have been added, so that any variable behind them is pure randomness and can be not relevant at all in the analysis.

The random variables allow us to know what variables have random behavior. Its last position implies a good sign of a good choice of variables because all of them are relevant to our target variables. Excepting one for the conversion difference, the day of the week has a similar behavior as the random variables. In this case, we could remove it.

To answer the first main question, and ignoring the intrinsic relationships of digital marketing, what more relevant in this case, is that the category of the keyword used has quite a lot of weight on this adjustment in conversions. This means that there are keyword categories that show more invalidation than others. When analyzing the value assigned by the partner to the quality of the traffic (*Figure 4*), the category also plays an important role. This time the importance is higher, being the second. It is also remarkable that the day of the month also influences whether the traffic is better or worse considered. Based on business experience, it is also known that depending on the country the TQ is higher or lower. However, due to the differences in measurement between Google Ads and the partners' monetization feed, there are usually traffic leaks to other countries than the ones targeted by the campaigns. This leakage also has some influence on traffic quality. The day is also important because Mondays usually have more records than other days, and it used to be the best day in revenue terms.

5. PREDICTIVE MODELS FOR TRAFFIC QUALITY SCORE

To avoid fraudulent and invalid traffic, different predictive models have been tested, developed, and validated. Thanks to the analysis of the conversion difference shown in the multivariant analysis the presence of BOTs has become evident. Our objective is to improve video ad web traffic quality score over there, including explanatory variables that company marketing managers and technicians could use to predict and avoid it.

5.1 Categorical Transformations

First, categorical variables need to be transformed because Deep & Machine Learning algorithms only understand numbers and cannot understand text in the first place (without considering semantic methods). However, the dataset is maintained before this transformation to try h2o models that already implement this process automatically. The goal is to test if it works better with our transformation or the automatic transformation. For their transformation we have used two methods:

(i) **One-hot encoding**: This has been used in the categorical variables with less than 4 values. One hot encoding can be defined as the essential process of converting the categorical data variables to be provided to machine and deep learning algorithms which improve predictions and classification accuracy of a model. This type of encoding creates a new binary feature for each possible category and assigns a value of 1 to the feature of each sample that corresponds to its original category.

(ii) **Frequency-based coding**: This last coding is based on the only directly measurable property of categorical variables: the frequency of occurrence of each category. Its application is as simple as replacing each categorical value with its frequency.

As a result, a new dataset has been generated to be used by the predictive models.

5.2 Classification Models for Traffic Quality Score

As seen before, the traffic quality score is a measure that the clients (publishers) give to the ads forming the marketing campaigns. The score is from 0 to 10, being 3 the minimum to maintain the campaign active. After analyzing the distribution of the variable in the previous steps, it was decided to process it, so a value of 0 is dedicated for traffic quality values lower than three (insufficient values) and 1 for acceptable and good values (from 4 to 10). It is a dichotomous variable. Looking at the distribution (Figure 3 and **Error! Reference source not found.**), it can be considered as a skewed variable with the most values concentrated in the first values of the variable, the fraudulent traffic values.

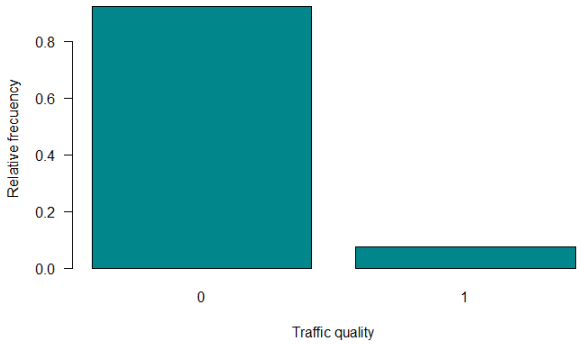

Figure 3. Distribution of ad quality classification

Table 2. Traffic quality score distribution

Traffic quality	Frequency
Bad traffic quality (Less than 3 scores)	92.41%
Reasonable traffic quality (Above 3)	7.59%

The model's objective is to identify and predict bad traffic before launching campaigns. In that way, the marketing team can reduce the percentage of traffic quality scores lower than three and give stability to the company's business model.

The target variable is categorical, it is because classification models have been developed, in the initial tests, three different perspectives were tried:

- Modelling by variable selection
- Classification models
- H2O models, trying different models using the main machine learning algorithms.

5.2.1 Modeling by Variable Selection

The technique used consists of removing variables from an LM model (linear model adapted to the categorical variable) to see the influence of each one of them. Different models were tested, from which can be concluded that all the variables contribute to the predictive model to a greater or lesser extent, but there are two that have a high VIF (Variance Inflation Factor). A high VIF is a sign of the correlation between them and the best practice to do is to remove one of them. These variables are the destination and the source_tag, which makes sense because each source_tag is assigned to a destination.

5.2.2 Classification Models

In this case, the classification models used were random forest, ranger (a fast implementation of random forests), ranger grid, ADA (Stochastic Boosting), and GBM (Generalized Boosted Regression), providing the results shown in Table 3:

Table 3. Predictive capacity of traffic quality classification models

Predictive model	Predictive capacity
Ranger	97.04%
Ranger Grid (300 trees)	97.03%
Random forest	74.58%
ADA	99.26%
GBM	78.28%

Numbers suggest that the predictive capacity of the models is quite high, and therefore very promising. The next step was the search for overfitting since the results appeared to be suspiciously very favorable. After analyzing the differences between train and test and the evolution of model error, it was concluded that overfitting is not present. In all models, the difference in predictive capacity between train and testing was less than 10%. Therefore, this low error in most models can be attributed to a target variable biased at 92.4% to a value, in this case, fraudulent traffic. The fact that the variable is highly biased means that the error is most likely to be biased as well, so in most of the predictive models the error is going to be very small.

5.2.3 H2o Models

The h2o models are very useful because they perform automatic transformations that help to improve the predictive capacity of our models. Thanks to the configuration of each model, also it can be also use techniques to adjust them to our specific data set. In this case, the models tested are the GBM model which follows the principle of gradient boosting, and AutoML which chooses the best model for our data.

Although h2o performs an automatic selection of variables and transformations of categorical variables, the model with our transformations yields slightly better results after testing (Table 4 and corresponding ROC analysis in Figure 4).

Table 4. Results of the transformation test for categorical variables

Predictive model	Predictive capacity
GBM initial dataset	96.15%
GM transformed dataset	96.42%

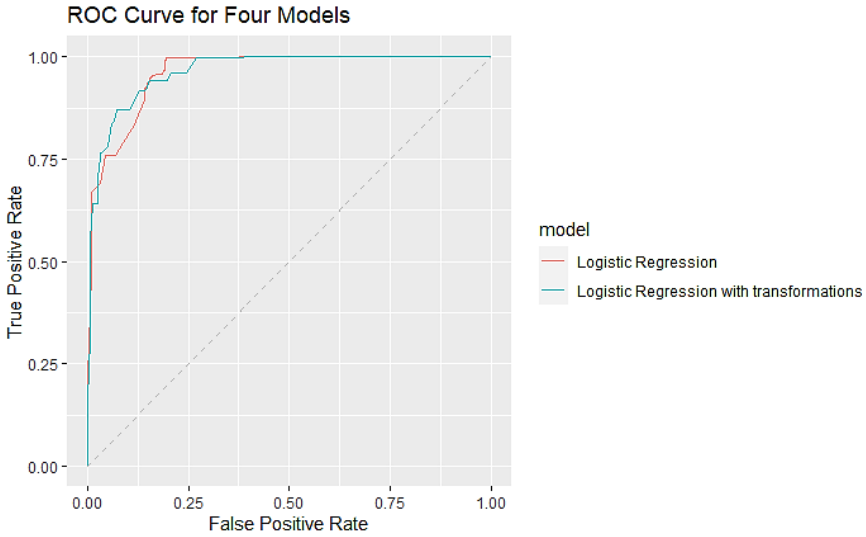

Figure 4. ROC Curve of the transformation test for categorical variables

Testing the AutoML model, the best model is the GBM_grid with an accuracy of 99.57% (Results Figure 5), slightly higher than the one we generated.

	model_id	auc	logloss	aucpr	mean_per_class_error	rmse	mse
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	GBM_grid_1_AutoML_1_20221205_140355_model_7	0.9956638	0.06502728	0.9533016	0.06064093	0.1275520	0.01626952
2	GBM_grid_1_AutoML_1_20221205_140355_model_4	0.9956610	0.05112022	0.9534650	0.07091814	0.1196946	0.01432679
3	GBM_3_AutoML_1_20221205_140355	0.9955800	0.04824214	0.9525534	0.06284388	0.1185562	0.01405558
4	GBM_grid_1_AutoML_1_20221205_140355_model_25	0.9955525	0.04946775	0.9528403	0.07799860	0.1184904	0.01403997
5	GBM_2_AutoML_1_20221205_140355	0.9955379	0.04983817	0.9521646	0.05302649	0.1200587	0.01441408
6	GBM_5_AutoML_1_20221205_140355	0.9954907	0.05489051	0.9515694	0.04694806	0.1224947	0.01500495
7	GBM_grid_1_AutoML_1_20221205_140355_model_8	0.9954805	0.05919361	0.9515380	0.05227445	0.1244279	0.01548231
8	GBM_4_AutoML_1_20221205_140355	0.9954654	0.06101951	0.9516943	0.05398789	0.1244755	0.01549416
9	GBM_grid_1_AutoML_1_20221205_140355_model_29	0.9954651	0.05572043	0.9514443	0.07317341	0.1211226	0.01467069
10	GBM_grid_1_AutoML_1_20221205_140355_model_28	0.9954328	0.05425814	0.9511943	0.04746058	0.1205846	0.01454066
11	DRF_1_AutoML_1_20221205_140355	0.9952417	0.05763474	0.9490834	0.09725592	0.1267876	0.01607509
12	GBM_1_AutoML_1_20221205_140355	0.9952286	0.05531764	0.9491859	0.04221140	0.1234224	0.01523308
13	GBM_grid_1_AutoML_1_20221205_140355_model_23	0.9950784	0.05620922	0.9473501	0.03024981	0.1239601	0.01536611
14	GBM_grid_1_AutoML_1_20221205_140355_model_30	0.9947488	0.05524952	0.9450884	0.04788208	0.1259602	0.01586598
15	GBM_grid_1_AutoML_1_20221205_140355_model_26	0.9947467	0.05727077	0.9446459	0.04119345	0.1264945	0.01600086
16	GBM_grid_1_AutoML_1_20221205_140355_model_24	0.9947366	0.05771723	0.9456777	0.07911996	0.1274824	0.01625177
17	GBM_grid_1_AutoML_1_20221205_140355_model_2	0.9944104	0.06799875	0.9428290	0.07134848	0.1339613	0.01794564
18	GBM_grid_1_AutoML_1_20221205_140355_model_10	0.9942804	0.06941584	0.9409508	0.06737892	0.1364986	0.01863187
19	GBM_grid_1_AutoML_1_20221205_140355_model_9	0.9941581	0.05800696	0.9384995	0.02780209	0.1313292	0.01724735
20	GBM_grid_1_AutoML_1_20221205_140355_model_12	0.9940237	0.06178330	0.9380882	0.02492746	0.1330745	0.01770883
21	GBM_grid_1_AutoML_1_20221205_140355_model_1	0.9923623	0.06880502	0.9283638	0.10492325	0.1380824	0.01906676
22	GBM_grid_1_AutoML_1_20221205_140355_model_11	0.9898317	0.08045541	0.9075433	0.10012553	0.1497901	0.02243708
23	GBM_grid_1_AutoML_1_20221205_140355_model_3	0.9843872	0.10160488	0.8643001	0.12723634	0.1706763	0.02913041
24	XRT_1_AutoML_1_20221205_140355	0.9717462	0.13929682	0.8100281	0.16897776	0.1946486	0.03788806
25	GLM_1_AutoML_1_20221205_140355	0.9588784	0.12560444	0.6822605	0.16973342	0.1905924	0.03632545

Figure 5. Results of AutoML model

Again, the checks show that there is no overfitting so we will attribute the good results to the bias in the target variable. Cross-validation (Figure 6) in turn shows similar AUC (Area under the ROC Curve) results between the different validation columns, in our case we chose five rows.

	mean <chr>	sd <chr>	cv_1_valid <chr>	cv_2_valid <chr>	cv_3_valid <chr>	cv_4_valid <chr>	cv_5_valid <chr>
accuracy	0.975805	0.001552	0.977642	0.976850	0.975229	0.973609	0.975697
auc	0.995641	0.000153	0.995714	0.995847	0.995579	0.995435	0.995629
err	0.024195	0.001552	0.022358	0.023150	0.024770	0.026391	0.024303
err_count	672.000000	43.092922	621.000000	643.000000	688.000000	733.000000	675.000000
f0point5	0.824865	0.017306	0.848349	0.833621	0.822122	0.802139	0.818096
f1	0.849905	0.004324	0.848721	0.857270	0.848791	0.845782	0.848959
f2	0.877066	0.016873	0.849093	0.882299	0.877249	0.894446	0.882243
lift_top_group	13.101583	0.293901	13.542174	12.942684	12.906598	12.852846	13.263610
logloss	0.054500	0.000787	0.053647	0.054190	0.055053	0.055568	0.054042
max_per_class_error	0.103499	0.029389	0.150658	0.100186	0.102695	0.069875	0.094078
mcc	0.838525	0.004152	0.836650	0.845824	0.836795	0.835593	0.837761
mean_per_class_accuracy	0.939412	0.012842	0.918606	0.941557	0.939539	0.953701	0.943654
mean_per_class_error	0.060588	0.012842	0.081393	0.058443	0.060460	0.046299	0.056346
mse	0.014726	0.000314	0.014369	0.014571	0.014975	0.015128	0.014585
pr_auc	0.953187	0.001503	0.952706	0.955776	0.952919	0.951859	0.952676
precision	0.809225	0.026759	0.848101	0.818567	0.805254	0.775463	0.798737
r2	0.791190	0.002554	0.789905	0.795622	0.790486	0.789158	0.790778
recall	0.896501	0.029389	0.849342	0.899814	0.897305	0.930125	0.905922
rmse	0.121343	0.001294	0.119869	0.120710	0.122374	0.122996	0.120768
specificity	0.982322	0.003820	0.987871	0.983300	0.981774	0.977278	0.981386

Figure 6. Cross-validation results of the winner of AutoML model

5.3 Chosen Model: AutoML GBM Grid

The model selected to predict the traffic quality score is the model generated by AutoML. This model has been launched with a grid, to find the most beneficial configuration, showing the following results (Table 5):

Table 5. Results of the predictive model GBM Grid (21 trees) with AutoML

AUC	LogLoss	PR AUC	Mean_per_class_error	RMSE	MSE
0.9953	0.0650	0.9533	0.0606	0.1276	0.0163

The predictive capability is very good, results that are explained by the fact higher than 92% of the traffic is categorized as insufficient or of poor quality. At the same time, this makes the probability of failure and error very low. Looking at the confusion matrix, it is seen that the error is higher when classifying good traffic, but the overall summary is still very good.

Table 6. Confusion matrix of the predictive model GBM Grid (21 trees) with AutoML

Traffic quality score	0	1	Error
0	126562	1708	0.013316
1	1253	9351	0.118163
Totals	127815	11059	0.021321

This work dealt with an unbalanced dataset, so it is required the calculation of the F1 score. This parameter is a harmonic average between precision and recall so it considers both false negatives and false positives.

Table 7. Results of F1 score of the predictive model GBM Grid (21 trees) with AutoML

Precision	Recall	F1 Score
0.882	0.846	0.863

The model's ability to both capture positive cases and be accurate with the cases it does capture is 0.86 (as shown in Table 7). In general, when the F1 score is between 0.8 and 0.9 it can be considered that the results are valid. The same conclusion can be seen if the results of the cross-validations (Figure 6) are checked, with the same results in the different validation stages.

6. CONCLUSION

Thanks to the multivariant analysis of the video ad marketing data, some unexpected relationships and variables that impact the conversions and TQ are deduced that would have been impossible to identify otherwise. This type of analysis has supported the basis for further knowledge and research into the predictive patterns resulting from it.

To answer the main second question about correlations, most of the explanatory variables in the predictive model were calculated from other variables used as a basis. This means that the variables are correlated with each other. This has been demonstrated by generating plots of influences and the correlation matrix. That is, exists intrinsic relations in the video marketing business model. On the other hand, thanks to the correlation matrix between the 50 variables, a low to medium correlation between them was detected, with the highest correlation being between initial clicks and initial cost with an 80% correlation. It is normal as in this business model each click is paid independently.

Once the key parameters impacting the final quality of the traffic have been identified, the next steps were focused on the development of the algorithms enabling the prediction of the video-ad marketing traffic quality, the main third question. Some objectives of the algorithms were: (i) The reduction of inauthentic users to increase the trust in this digital media as a channel to improve industry investment in video marketing, (ii) Generate a new approach to optimizing the impact of the campaigns, reducing the dependencies in third parties, (iii) Development of video marketing campaigns decision support based on real user interactions, not being affected by the actions carried out by malicious bots and (iv) deploy a new way to engage the real targeted audience and promote competitiveness in digital media content. All of them consider the existence of multicollinearity and that there may be external factors that cannot be handled or unseen before. Therefore, it is important to have up-to-date information, digital marketing evolves day by day and marketing managers must react accordingly.

After the realization of the different predictive models, with high accurate results. The overfitting was analyzed and realized how important it is to use a balanced dataset. The existence of bias in the target variable greatly affected the error and therefore the results. Thanks to the large volume of data, the dataset can be updated to include more positive traffic quality values resulting from the latest campaigns.

Currently goal is to work only with campaigns that are not associated with fraudulent traffic. Obviously, given that 92% of campaign traffic is currently invalid, the team will suffer a significant reduction in the campaign volume which would affect profitability in the early stages. However, it is a very positive aspect for the company as we can consolidate our

marketing campaigns, giving stability to the business model and favoring controlled and secure growth. On the other hand, by enriching the data set with a more balanced data, additional adjustment of the predictive models will be required to achieve an expected prediction accuracy.

ACKNOWLEDGEMENT

This work received partial funding from the European Commission Horizon 2020 AI4Media project (grant number 951911).

REFERENCES

Black, W., & Babin, B. J. (2019). *Multivariate data analysis: Its approach, evolution, and impact. In The Great Facilitator: Reflections on the Contributions of Joseph F. Hair, Jr. to Marketing and Business Research* (pp. 121-130). Cham: Springer International Publishing.

Everitt, B., & Hothorn, T. (2011). *An introduction to applied multivariate analysis with R. Springer Science & Business Media*.

Saura, J. R. (2021). *Using data sciences in digital marketing: Framework, methods, and performance metrics*. Journal of Innovation & Knowledge, 6(2), 92-102.

Neusser, K. (2016). *Time series econometrics*. Springer.

Hyndman, R. J., & Khandakar, Y. (2008). *Automatic time series forecasting: the forecast package for R. Journal of statistical software*, 27, 1-22.

Arroyo Barrigüete, J. L., Fabra Florit, M. E., & Redondo Palomo, R. (2023). *Análisis multivariante.*

Salgado C., F. B. P. L. &. (s. f.). *2 Tipos y estructuras de datos en R | Introducción a R y SIG*. https://bookdown.org/chescosalgado/intro_r/tipos-y-estructuras-de-datos-en-r.html

Cario, M. C., & Nelson, B. L. (1997). *Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix (pp. 1-19)*. Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Isea, R., Ojeda, V., Fernandez, J., Gutierrez, A., & Salazar, V. (s. f.). *COEFICIENTE V DE CRAMER (V).* https://mariafatimadossantosestadistica1.files.wordpress.com/2018/06/coeficientes-v-de-cramer-y-c-de-pearson.pdf

Afandi, W., Bukhari, S. M. A. H., Khan, M. U., Maqsood, T., & Khan, S. U. (2022). *Fingerprinting Technique for YouTube Videos Identification in Network Traffic*. IEEE Access, 10, 76731-76741.

Stevanovic, M., & Pedersen, J. M. (2015, June). *An analysis of network traffic classification for botnet detection*. In 2015 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA) (pp. 1-8). IEEE.

Efthimion, P. G., Payne, S., & Proferes, N. (2018). *Supervised machine learning bot detection techniques to identify social twitter bots*. SMU Data Science Review, 1(2), 5.

Shevtsov, A., Tzagkarakis, C., Antonakaki, D., & Ioannidis, S. (2022, May). *Identification of Twitter Bots Based on an Explainable Machine Learning Framework: The US 2020 Elections Case Stud*y. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 16, pp. 956-967).