livepanel

Processing Surveys from Lookalike Audiences

Table of Contents

Summary

This paper explores the improvements in performance, speed and costs that the introduction of methodologies for passive respondent information processing, Natural Language Processing (NLP) and Machine Learning (ML) could bring about to the processing of surveys in lookalike audiences in public opinion surveys using cell phones through the LivePanel platform.

Aims

- To enable the disaggregation of information (Drill-Down) while maintaining significant sample sizes in each segment and even new combinations in the field work.
- To facilitate the collection of valuable information in environments that are hostile to surveys, whether due to lack of interest in the subject, availability of qualified users, representativeness of the sample with respect to the population pyramid, etc.
- To optimize the availability in turnaround times and costs to the users of the information, while maintaining a reasonable level of accuracy.

Introduction

The mobile world is steadily expanding. Latin America is no exception: according to GSMA reports (2019), 67% of the Latin American population is subscribed to cellular service, and 80% of them has mobile Internet service. This percentage equals approximately 337 million people. This scenario is the domain of LivePanel, a technology company that allows its customers to take surveys on cell phones by leveraging the full potential of these devices, with full coverage in Latin America and the Caribbean. While nowadays the results of surveys are used to feed predictive models[1], the use of AI tools to contribute to the improvement of field work is currently uncharted terrain.

With this paper we will show the improvements in performance, speed and cost reduction that can be generated by the adoption of Artificial Intelligence methods, specifically with respect to Machine Learning and Natural Language Processing (NLP) methods. To this end, two case studies will be developed based on surveys to individuals, and a comparison will be made between them and the results that can be generated by surveys processed by Machine Learning (ML) techniques at different levels of response in lookalike audiences.

In order to provide a framework for evaluating the intended results, the following premises are considered to be valid for the Public Opinion and Markets industry in Latin America at the time of writing:

1. **Details**: The requirements of sample size and composition for the requested studies become evident with respect to creating segments (*Drill-Down*) that offer value while maintaining the significance of the information provided. It is relevant to have the least sample deviation at the aggregate level, but it is also relevant to be able to find richness in the required segments.

[1] Deltell, L., Claes, F., & Osteso, J. M. (2013). Predicción de tendencia política por Twitter: Elecciones Andaluzas 2012. *Ámbitos. Revista internacional de comunicación*, (22).

2. **Pandemic**: The emergence of the COVID-19 crisis has had serious implications for the traditional work of field data collection through face-to-face interviewers, a situation that already showed compromises in terms of security of the field worker and respondent, access to homes, compliance with quotas, etc.

3. **The speed and cost do matter**: Users of field information place value on a survey system that does not present significant quality differences but that manages to deliver partial or total results in a fraction of the time and at a fraction of the cost.

Methodology

The work process consists of a six-step process that starts with preparing the audiences and ends with developing comparative metrics. **Step 1** involves the extraction of a similar audience from our base of total panel members in similar number and distribution of age, gender and country to those intended to be emulated through the procedure, taking into account that none of the users had been included in the original group, nor have they registered on the platform after the completion of the study.

Through *Business Intelligence* tools, a comparative procedure was carried out to establish which questions within the LivePanel database, excluding those of the questionnaire to be worked on, were answered by both the respondents in the field and the audience selected for processing. At this point it is important to clarify that many of these are historical users of the platform, so there are questionnaires that cover much of both samples (especially those linked to the determination of Socioeconomic Level).

Important data of passive nature was also set aside, namely:

- Device Type
- Installed applications in a subset of 100 most installed applications
- Total number of installed applications (Android)
- Total number of applications installed by category (Android)

It should be clarified that all Machine Learning operations include Age Group and Gender as training and prediction variables.

This first step culminates in obtaining a large number of features to work on (on average, 250 variables are usually generated in this first instance). In order to improve the quality of the dataset and enhance the efficiency of the Machine Learning process, **step 2** begins, wherein an automated selection of both historical questions and passive information attributes is carried out using Natural Language Processing (NLP) techniques that can be automated during the management of the process.

After a selection of cases for the execution of each comparison (20% to 90% of field cases, in 10% increments) **step 3** generates two files, training and prediction respectively, containing the cases and a set of variables to which the questions of the questionnaire are added.

The model training performed in **step 4** by an automated procedure in a LivePanel API consists of the evaluation of different Machine Learning methods, accompanied by its respective hyperparameter tuning. To ensure the predictive effectiveness of the model, the k-fold-Cross Validation method was used. The models considered are the following: Tree and Ensembles of Decision Trees (Random Forest and Gradient Boosting) and Logistic Regression. The reason for choosing this variety of ML algorithms lies in the fact that we sought to use models that are linked as much as possible to supervised methods, since they allow the generation of explanations at the level of the characteristics used, something that is usually not possible with unsupervised methods. However, when it comes to determining what each of these methods is used for, they can be grouped into two main groups: classification problems or regression problems. The former are those in which the variable to be predicted (or estimated) is of a categorical type; the regression problems, on the other hand, focus their study on continuous phenomena.

Each of the methods used will be described briefly below:

- The **Decision Tree** is comprised of a sequential series of logical constructions, which derive in a series of possible results (forming something akin to the branches of a tree). Although this method is used mainly for classification problems, it can also work within regression problems (in these cases, they are called "regression trees").

- The **Ensembles** consist of a set of machine learning models. Even though this class can consist of any machine learning model, for these cases we will use ensembles made up entirely of decision trees, because they are much faster to build. Both types are used: the *Random Forest* type, which consists of forming a large number of trees, from which a training subset is selected based on random samples, and *Gradient Boosting*, which is similar in essence with the one exception that each tree tries to learn from the previous one.

- **Logistic Regressions** allow us to estimate the probability that a variable belongs to a categorical variable based on a quantitative one. Its main application consists of binary classifications.

However, this definition is not restrictive. The different methods corresponding to the use of sets of decision trees can also perform regressions with good results.

That said, for each field an **optimal model** trained in a sequential way will be generated. These are chosen through a combination of evaluation statistics among which the **F Statistic** was selected, which combines both Precision analysis (positive predictions that actually are true) and *Recall* analysis (positive predictions over total positive predictions) as well as the latter independently. The choice of statistics is intended **to consolidate the efficiency of the training** of fields with imbalanced samples.

Step 5 performs the sequential prediction of the data sets also in an automated fashion, and for **step 6** two comparative activities were performed:

- First, calculating the average of the absolute deviations for each of the possible answers with respect to the field result. For example: for question number 1, if the results were 75% for option 1, 15% for option 2 and 10% for option 3 for the **total** number of cases in the field, if the prediction were 73% for option 1, 16% for option 2 and 11% for option 3, the overall deviation result would be (2+1+1)/3=**0.75%**

- Secondly, and pursuant to the aforementioned objective, i.e., *to enable the disaggregation of information (Drill-Down) while maintaining significant sample sizes in each segment and even new combinations in the field work,* an analysis similar to the previous point was carried out but with **three levels of disaggregation**, namely: Country, Age and Gender for a total of **72 additional combinations** for each possible answer, for a total of 1,744 segments/response (combination of Country, Age, Gender and Possible answer)

- Imputation of missing values: for field cases, there will be segments that will not show observations in responses. The literature provides information on existing methods for the

imputation of missing values from surveys, such as the work of Rosati (2017), which approaches imputation through Lasso regression ensembles. However, the following criterion will be used: given the existence of missing values (as long as the original field contains responses), a 50% deviation will be imputed. The reason for this is that, if there is a value other than 0 in 100% of responses, and the field work could not predict those answers, technically it would be an infinite deviation (there are infinitesimal zeros in any non-nil number). In order not to penalize the field so much in these cases, it is reduced to only 50%.

The analysis will also take into account considerations such as the time saved by the use of ML models, as well as the hiring of advertising for the collection of observations.

The procedure can be summarized in the following image:

Source: Developed in-house

Case studies

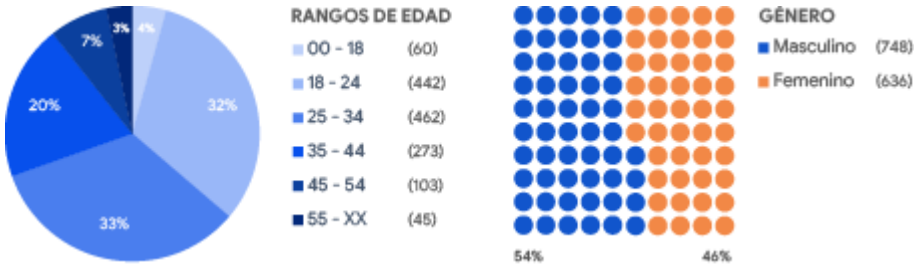
Two cases were developed based on field surveys conducted by LivePanel between 2019 and 2020.

Case 1: 5G Latin America

The first study surveys the impact of 5G in Latin America for the consulting firm SMC+[2], renowned specialists in *Public Policy* in the region, and was actually featured in various presentations in Argentina, Mexico and the United States.

The study was carried out on a total of 2,103 individuals in 17 Latin American countries, with the "free fall" mode and without recruitment reinforcement (online advertising). To narrow down the analysis, a set of 6 countries with the largest sample size and age group representativeness (the sample was limited to people over 16 years old) was selected.

COUNTRY	CASES
Argentina	299
Colombia	202
Ecuador	180
Mexico	395
Peru	148
Venezuela	161



Source: Developed in-house

The questionnaire consisted of 12 questions, but for the purposes of the study only the ones with simple answers were selected, since, given the format consulted, the questions were not cross-related—something which, on the other hand, strongly conditions the following responses and significantly improves the performance measured—in order to preserve the objectivity of the study. The questions were as follows:

1. Thinking about your home, do all people who use cell phones use smartphones?
2. The following questions are of a personal nature. Do you have a mobile data plan?
3. Does your phone connect to 4G mobile networks?

[2] https://www.linkedin.com/feed/update/urn:li:activity:6658011041429012481/

4. What are your expectations of 5G mobile technology?
5. What do you think 5G is going to be more useful for?
6. When do you expect 5G to arrive in your country?
7. Your cell phone spending today…
8. Are you happy with your mobile carrier's service?
9. Would you be willing to pay more to have more services and things you can do with your cell phone?

Preparation of the model

During the development of **step 1**, 56,484 potential cases were found for answering the questionnaire, which satisfied the conditions that they were different from those who had answered the questionnaire and had registered prior to the study. The necessary cases were then filtered according to quantity and lacking distribution of age, gender and country to complete 100% of the base.

Then, by matching the questions users have in common, a total of 250 possible questions were found for use, including active data sources (explicit survey responses plus age group, country and gender) and passive data (which the user provides at the time of accepting the service). The latter includes the following:

- Device Type
- Installed applications in a subset of 100 most installed applications
- Total number of installed applications (Android)
- Total number of applications installed by category (Android)

As previously mentioned, this first step produces about **250 variables** applicable to model training and prediction. After this, **step 2** will start, wherein this number of questions will be reduced by means of Natural Language Processing (NLP) or Regularization (Lasso) techniques.

Upon completion of this step, in this case the following 32 fields were selected, which we will call **Key Training Variables (KTV)**:

Questions

What cell phone plan do you have?

Do you use the same search engine on all your devices (cell phone, laptop, PC, etc.)?

Have you attended any courses or seminars in the last 3 months?

Are you currently in employment?

If you had to define the place where your home is located, ¿would you say…? (check the option you consider most relevant)

If you had to define the type of home you live in, ¿would you say it is…?

And with regards to the people you live with, how many people in TOTAL would you say live in your home (including yourself)? It doesn't matter how you are related

Do you consider that the city you live in is…?

Have you made any hotel or travel package reservations online in the last 3 months?

Passive Data

Applications
Facebook
Instagram
Netflix
Snapchat
Spotify
Twitter
Uber
Waze

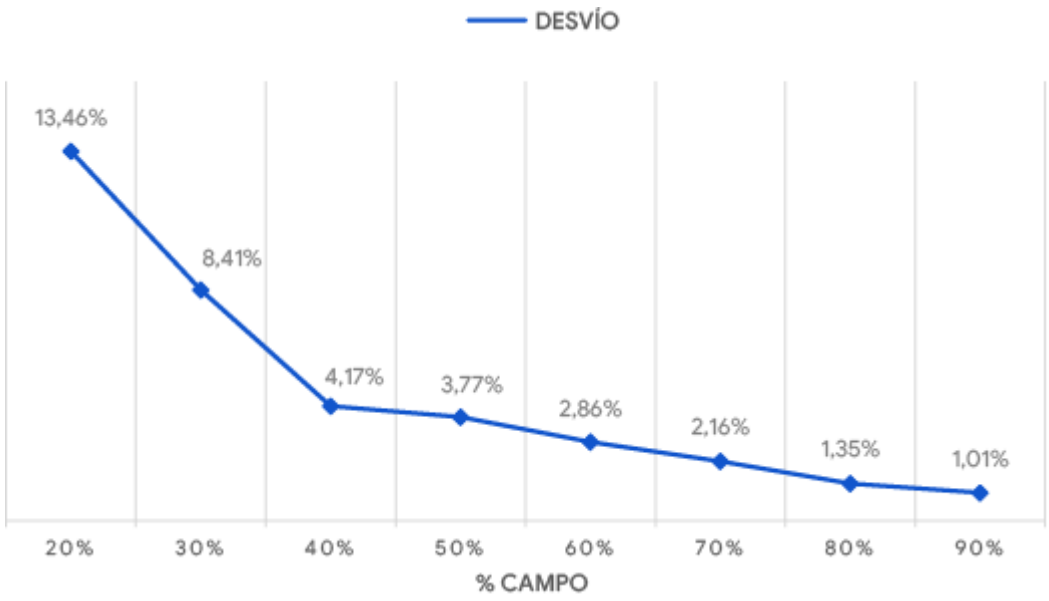
Categories - Other (Quantity of Apps)
COMMUNICATION
ENTERTAINMENT
FINANCE
LIFESTYLE
MUSIC AND AUDIO
PHOTOGRAPHY
PRODUCTIVITY
SHOPPING
SOCIAL
TOOLS
com

Once the different percentages of the sample used to make the comparisons have been created, in **step 3** two files, training and prediction respectively, for each of the cases and the set of variables previously selected, to which the nine questions of the questionnaire are added.

In **step 4** a total of 250 models were trained, of which the model with the highest F and Recall statistics was chosen for each of the fields to be trained, followed by the prediction corresponding to **step 5**.

For **Step 6**, comparisons are made:

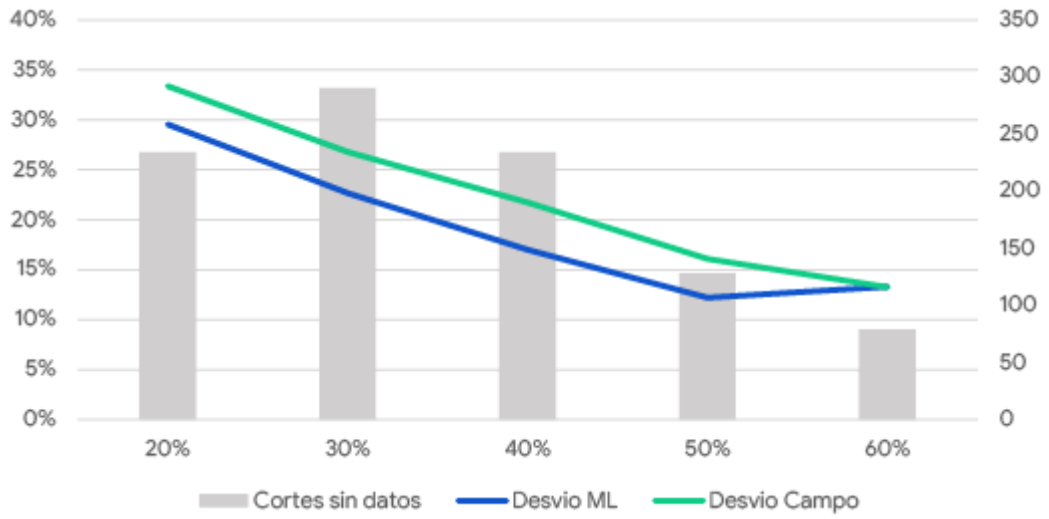
Average of Absolute Deviations (according to sample percentage when training models)



Source: Developed in-house

It can be observed that the deviation decreases abruptly in cases where the survey starts being processed in similar audiences from 20% and 40% of the field, and then decreases with a lower incline to 1%. It can be observed that the value of the deviations from 40% is considerably reduced. However, in order to consider the performance of the method, it will also be analyzed in relative terms.

Deviation averages allowing for Drill-Down

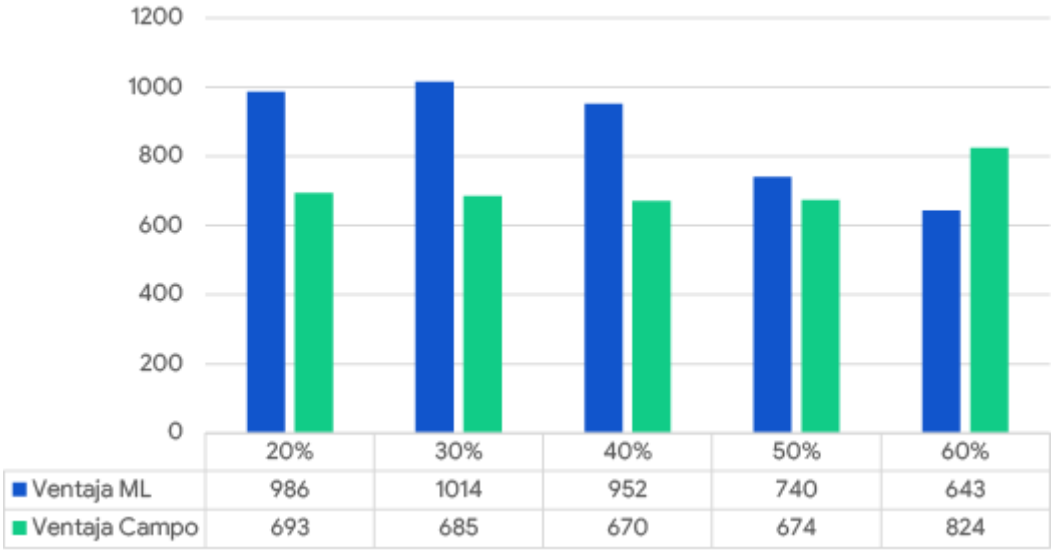


Source: Developed in-house

This graph is the most revealing regarding the impact of the methodology: unlike other procedures that show severe limitations when interpreting or weighting partial results to be comprehensive of a population, the use of a solid methodology that considers both responses and pre-existing information from the panel member allows not only to reproduce responses with greater precision in the scarcest sections of the field, but also brings to light a very important number of segments that do not show responses in the partial field, for a maximum of 16% in the case of the field segment at 30%. In fact, the methodology is capable of **creating value** and this value creation, by definition, can be improved over time with the accumulation of questions of similar nature, greater amount of passive information, improvements in the recurrent training of models, etc.

When comparing the number of segments where this processing shows less deviation with respect to the total field sample, it can be observed that, within the *"sweet spot"* of the contribution of the methodology (20%-50% of field observations), the number of observations (out of 1,744 in total) where this processing has a lower deviation with respect to the total field than the partial sample is also notable. In this case, the segment that presented a difference of more than 1% between the two was defined as advantageous, for example: If the deviation with respect to the field was 2.3% and to the partial segment was 3.6%, the model is considered advantageous; if instead these were 1.3% and 2%, they would be considered "tied".

By incorporating the three levels of Drill-Down and focusing on the segments with less participation by the field sample, where the intervention of the technique is more powerful, the results are as follows:

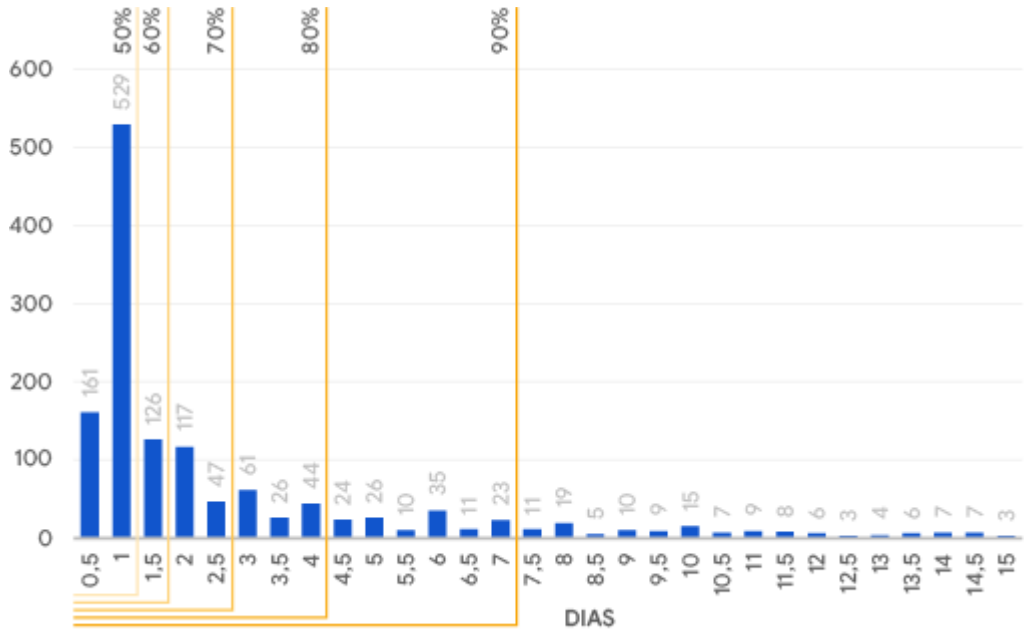


	20%	30%	40%	50%	60%
Ventaja ML	986	1014	952	740	643
Ventaja Campo	693	685	670	674	824

Source: Developed in-house

This graph also highlights the methodology's ability to create value in small samples, which are the most frequent in instances of high-value, low-participation countries and audiences in the collection mechanisms.

Finally, an important point to report is that, in this case, the development of the field project, given that it was a non-paid *partnership*, involved the use of active panel members over the total panel who could opt-in to this low incidence survey in the score/rewards received, so the time distribution of the field cases adopted the following form:



Source: Developed in-house

Where the first 50% of responses was obtained in the first 24 hours in the field, 60% in 36 hours, 70% in 60 hours and so on. Although online advertising modifies this kind of arrangement in other types of surveys, it usually takes on a similar format as a function of the decreasing rates that implies the saturation of audiences in the case of digital marketing efforts, reason why it is an excellent parable of analysis for other cases. In all cases, the comparison with respect to the field results was made by isolating the cases in strictly **chronological order.**

Specific Conclusions

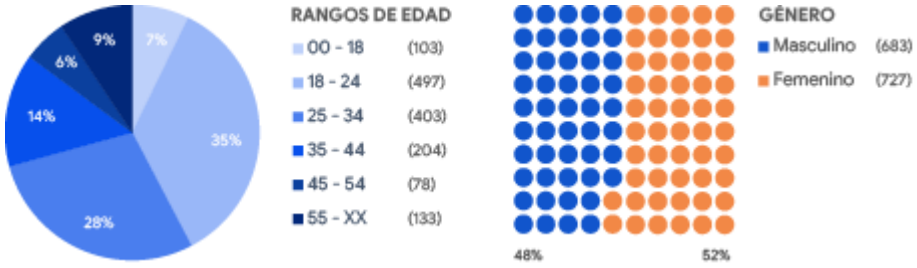
The analysis of 5G's case led to the conclusion that it is possible, using Machine Learning methods, to obtain a premature result to the final delivery that adds value and with a lower level of deviation compared to field values at three levels of Drill-Down. Using a concrete example and relating the four indicators analyzed previously, in a period of less than one day of releasing the survey into the field, we would have an absolute deviation of about 4%, a relative deviation at three levels of disaggregation 6.45% lower than that of the field at that time. In summary, the method represents a tool that makes it possible to save a substantial amount of resources without the need to compromise on the quality of the final results, or better yet, to offer a much more precise detail of the analysis in record time and with limited costs.

Case 2: Fast-Food Restaurants

The second study is an analysis of the fast-food industry in Central America for an important market player. The content of some of the questions has been modified in order to safeguard the confidentiality of the study.

The study was carried out on a total of 1,412 individuals in 3 countries of Central America, with the "free fall" mode and with recruitment reinforcement (online advertising).

COUNTRY	CASES
Costa Rica	464
Panama	369
Dominican Republic	579



Source: Developed in-house

The questionnaire consisted of 26 questions. The questions were as follows:

1 to 5 - Using a scale from Very Healthy to Not Healthy at All, how would you rate this food item? *Pizza* *Sandwich* *Burger* *Hot dog* *Fried chicken*

6 to 10 When you choose a fast-food restaurant to go to or to order from, how important is it to you that the restaurant is competitively priced? I like the type of food/to indulge myself? has food that satisfies me/makes me feel full? has healthy choices? is fast? is reliable and has good quality products?

11 to 18 - And thinking specifically about when you choose the drink to accompany your fast-food order, how important is it to you that this drink accompanies the food well? that I love how it tastes? that it is healthy? that it has no calories? that it is natural? that it has good nutritional value? that it is sugar free?

19 to 26 - Which would you say is the drink that best accompanies this meal? *Sandwich* *Wrap* *Pizza**Salad* And if you had to choose a second option, which would you say is the other drink that best accompanies this meal? *Sandwich* *Wrap* *Pizza**Salad*

Preparation of lookalike audiences

During the development of **step 1**, 19,023 potential cases were found for answering the questionnaire, all of which satisfied the conditions that they were different from those who had answered the questionnaire and had registered prior to the study. The necessary cases were then filtered according to quantity and lacking distribution of age, gender and country to complete 100% of the base.

Next, by matching the active and passive data that are common to continuous users from a filter of the same kind as the previous case, the following 14 KTVs were defined:

Questions

When did you purchase or start using this smartphone?

Do you eat breakfast every day?

Do you do any physical activity in the morning? (gym, walking, yoga, sports in general)

Have you made any hotel or travel package reservations online in the last 3 months?

Are you a member of a sports club?

Have you attended any courses or seminars in the last 3 months?

Was the last course/seminar you attended related to your professional career?

How many times a week do you usually eat at restaurants or fast-food joints?

Do you usually cook the meals in your home?

Do you like cooking?

How knowledgeable do you consider yourself to be in the kitchen?

Would you be interested in learning how to cook better?

How many times a week do you usually order food delivery?

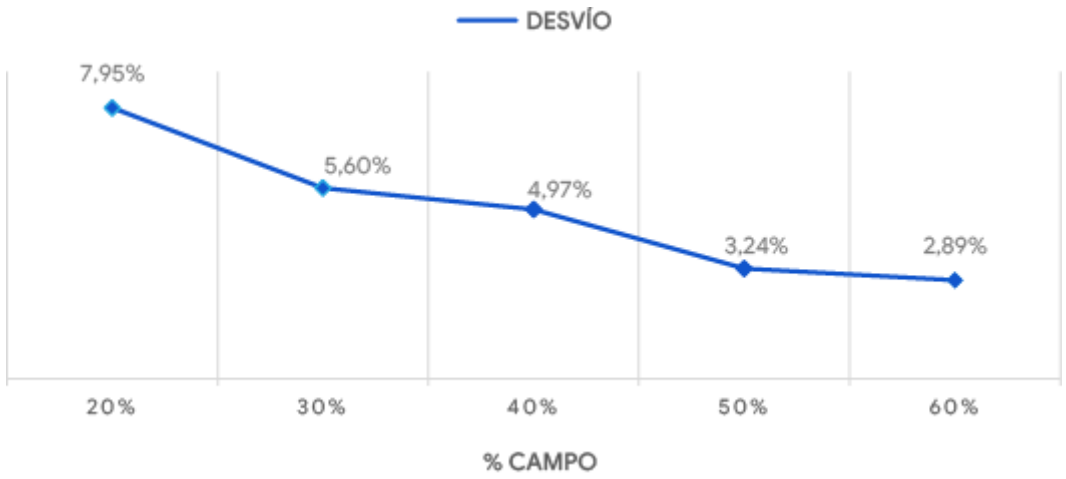
Passive data

Phone model
Applications installed in the FOOD category

After adding these questions to those of the questionnaire and having created the training and prediction datasets for each percentage (step 3), in step 4 a total of 448 models were trained, of which the model with the highest F and Recall was chosen for each of the fields to be trained, followed by the processing corresponding to step 5.

For **Step 6**, comparisons are made:

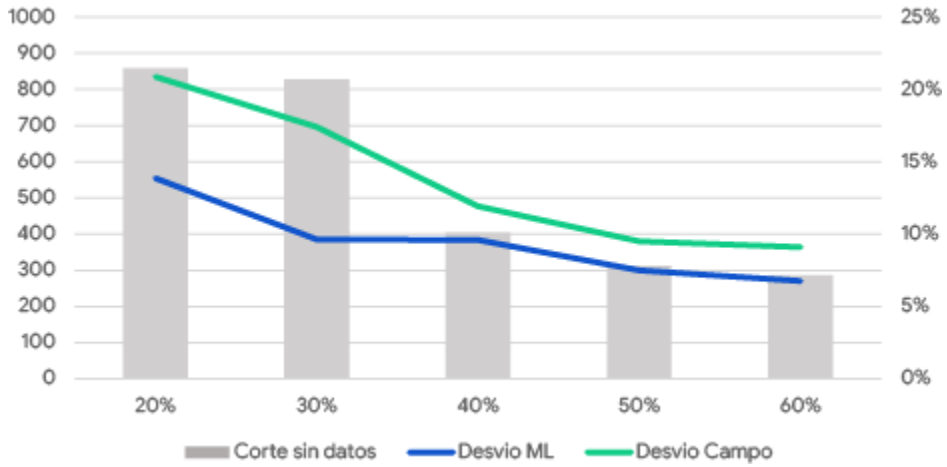
Average of Absolute Deviations (according to the percentage shown when training models)



Source: Developed in-house

A deviation of less than 8% can already be observed here from the first field generated. Moreover, we will analyze, like in the previous case, the comparison of the deviations (both of the field and of the prediction) inherent to 3 levels of Drill-Down.

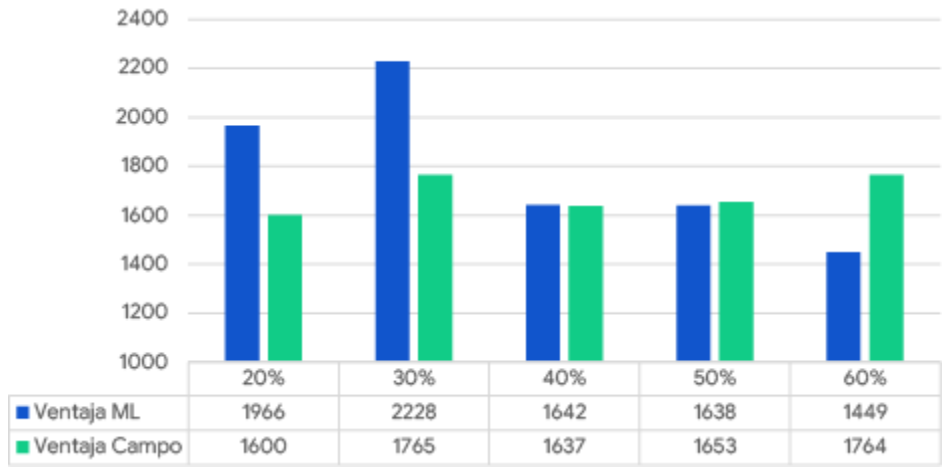
Deviation averages allowing for Drill-Down



Source: Developed in-house

The graph highlights what is shown in the previous graph: for 20% of the sample, the methodology applied by Machine Learning not only shows a deviation 7 percentage points lower (almost 34% less), but added value in 858 segments that were not covered on the field. Although the gap between the deviations tends to decrease, the sample deviation with this processing is always lower, and given the creation of non-existent segments and the time savings this implies, makes this an extremely efficient method in terms of resource savings.

By incorporating the three levels of Drill-Down, the fields where this processing has an advantage (in terms of less deviation) will be shown, like those of the field itself:

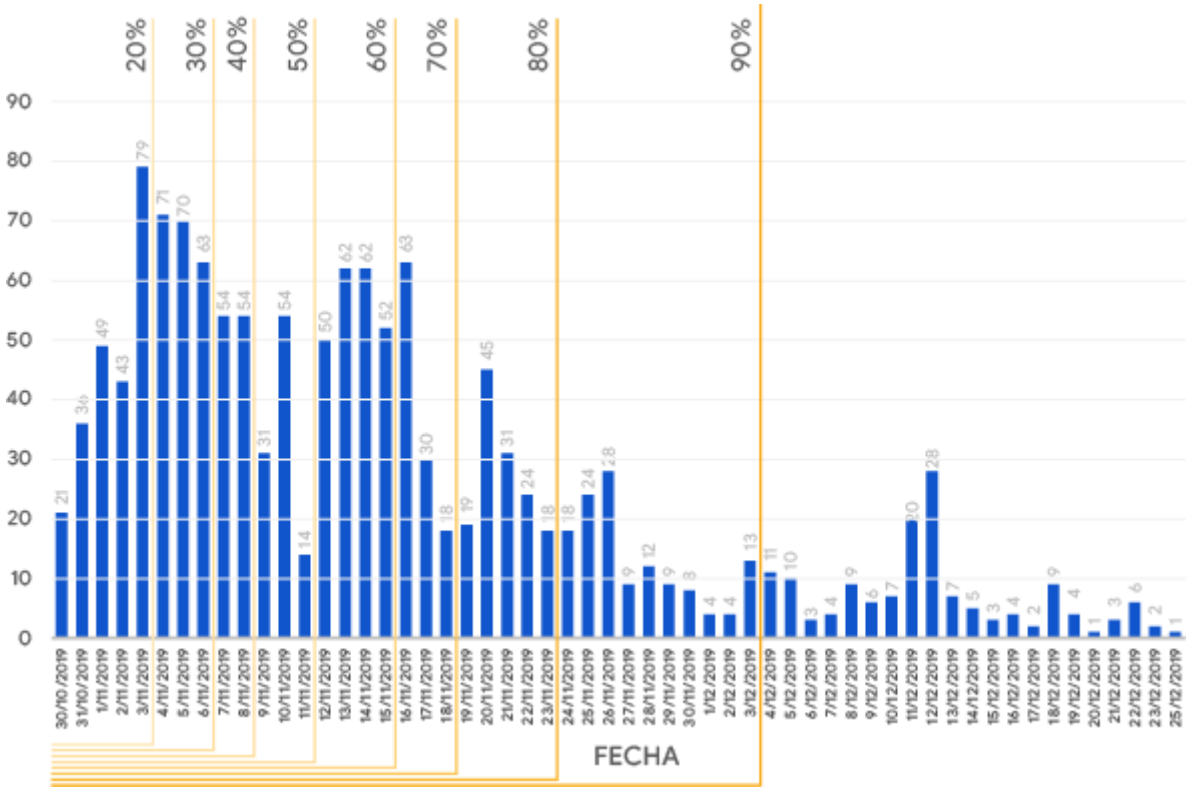


	20%	30%	40%	50%	60%
Ventaja ML	1966	2228	1642	1638	1449
Ventaja Campo	1600	1765	1637	1653	1764

Source: Developed in-house

As in the previous case, the suitability of the methodology for creating value in minority groups either generated by the Drill-Down levels or by the scarce (or sometimes late) participation in the collection mechanisms can be seen.

Finally, with respect to a paid *partnership,* it is necessary to analyze how demanding in terms of time it would be. It is important to consider that, in addition to saving time, having the ability to process missing cases in similar audiences lowers the cost of recruiting panel members. The time distribution can be seen here:



Source: Developed in-house

Here, with paid advertising, the distribution of responses is more uniform than in case 1 in terms of time, although it calls for a particular analysis. The study lasted 57 days, but by the end of the eighth day, with 30% of the total responses collected, value can already be generated at a considerably low level of deviation (less than 10%).

Specific Conclusions

For the Fast-Food Restaurants analysis, it was validated that it is possible to achieve good results processing surveys in lookalike audiences via Machine Learning methods, saving not only time, but also money in advertising for recruiting panel members. Given a specific case, at 40% it is possible to obtain a result with an absolute deviation of 4.97%, a relative deviation considering three levels of

Drill-Down of 9.59% (19.5% less than that of the field) 10 days after releasing the survey, having agreed on an initial period of approximately 60 days.

Final Conclusions

Throughout this paper, by exposing a new methodology for the analysis and processing of surveys presented in two case studies, it is demonstrated that, via the interaction with Artificial Intelligence methodologies, partial and total results can be obtained, and even extend the scope of a rigorously executed field work in an extremely cost-effective manner with useful results in a third of the field work executed, significantly impacting the availability of data in terms of both speed and costs.
We expect that this paper provides precedent on the benefits of using methods such as Machine Learning and NLP in the optimization of resources when planning the methodology and execution of a field study in a rational, and above all, reliable way.

Bibliography

Deltell, L., Claes, F., & Osteso, J. M. (2013). Predicción de tendencia política por Twitter: Elecciones Andaluzas 2012. *Ámbitos. Revista internacional de comunicación*, (22). *Retrieved from* *https://institucionales.us.es/ambitos/prediccion-de-tendencia-politica-por-twitter-elecciones-andaluzas-2012/*

CONICET: Instituto de Ciencias e Ingeniería de la Computación (ICIC) (2019). Minería de Datos: optimización de uso de herramienta de Machine Learning. *Retrieved from* *https://www.livepanel.co/paper/Livepanel%20-%20Mineria%20de%20Datos%20-%20Informe.pdf*

Intelligence, G. S. M. A. (2019). La Economía móvil en América Latina y el Caribe. *Retrieved from* *https://www.gsma.com/mobileeconomy/wp-content/uploads/2020/03/GSMA_MobileEconomy2020_LATAM_Esp.pdf* *[Fecha de consulta: julio 2020]*.

Rosati, G. F. (2017). Construcción de un modelo de imputación para variables de ingreso con valores perdidos a partir de ensamble learning. Aplicación en la Encuesta Permanente de Hogares (EPH). SaberEs, 9(1). *Retrieved from* *https://www.saberes.fcecon.unr.edu.ar/index.php/revista/article/view/132/389*