MAA

Mestrado em Métodos Analíticos Avançados

Master Program in Advanced Analytics

Enhanced Web Analytics for Health Insurance

Piero Maggi

Internship Report presented as the partial requirement for obtaining a Master's degree in Data Science and Advanced Analytics

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

Enhanced Web Analytics for Health Insurance

Piero Maggi

Internship Report presented as the partial requirement for obtaining a Master's degree in Data Science and Advanced Analytics

Advisor: Leonardo Vanneschi

External Advisors: Paula Santos, Jéssica Raquel Zaqueu

February 2020

DEDICATION

To my family which made all of this possible.

ACKNOWLEDGEMENTS

To Andrè Rufino and Paula Santos, thank you for giving me the opportunity to work with you and for making sure that this internship was fulfilled with precious learning experiences. I also want to thank the whole host company and my team for the help and support provided along these nine months.

To Jéssica Zaqueu, what a ride this has been. Thank you for the support you always gave me, as a colleague but most importantly as a friend. Through the many challenges that together we have been facing you taught me how to stop, zoom out and see the problems from a different perspective.

To Leonardo Vanneschi, thank you for giving me the opportunity to join this Master and for being my supervisor.

To my mates Andreas and Dalì, with whom I shared this experience with.

ABSTRACT

Nowadays companies need invest and improve on data solution implementation within most of the business workflows and processes, in order to differentiate the offer and stay ahead of their competitors. It's becoming more and more important to take data driven decisions to boost profitability and improve the overall customer experience. In this way, strategies are defined not anymore on common beliefs and assumptions, but on contextualized and trustful insights.

This reports describes the work that has been made during a 9-month internship, in order to provide the business with a new and improved solution for enhancing the web analytics tasks and supporting the improve of the online user digital experience. User-level data related to the website activity has been extracted at the highest granularity level. Afterwards, raw data have been cleaned and stored in an Analytical Base Table with which an initial data exploration has been made. After giving initial insights to the digital team, a predictive model has been developed in order to predict the probability of the users to buy the insurance product online. Finally, based on the initial data exploration and the model's results, a set of recommendations has been built and provided to the digital department for their implementation in order to make the website more engaging and dynamic.

KEYWORDS

Web Analytics, Google Analytics, Supervised Learning, Decision Tree, Logistic regression, Random Forest, User Behavior,

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SEM	Search Engine Marketing
SEO	Search Engine Optimization
ABT	Analytical Base Table
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
GATC	Google Analytics Tracking Code
CRISP-DM	Cross-industry standard process for data mining
ELT	Extract, load and transform
VPN	Virtual Private Network
CRM	Customer Relationship Management

1. INTRODUCTION

This project has been developed for a health insurance company that has been operating in Portugal for more than twenty years. With its wide range of products and services offer, the business can be considered one of the top players in the Portuguese market. Main stakeholders for this project were the Analytics Department, designed as the project owner, and the Digital department as the one that benefits the most by the project outcomes.

The Analytics department has as a mission pursuing excellence in the Analytics areas, with the research and implementation of new algorithms and sophisticated machine learning techniques. It acts as an enabler for the other business units by delivering data driven insights to better guide the planning of their strategic decisions.

The Analytics team is composed by three different profiles:

- Solution Architects, who designs the solutions integrated with the stakeholders' requirements and supervises its implementation;
- Data Engineers, who builds the data architectures in order to support the work of the other areas;
- Data Scientists, who extracts patterns from data by using complex algorithms, predictive models, clustering within the most advanced machine learning techniques.

The Digital Department is the business owner of the website and it is responsible for the following tasks:

- Designing and implementing the structure of the website within its content, online marketing campaigns, Search Engine Optimization (SEO) and Search Engine Marketing (SEM) techniques;
- Monitoring the activity of the users on the website in order to better define the previous ones;
- Designing and constantly adapting the website user's experience in order to boost its engagement and increase profitability through online sale of the insurance policy.

The topic area of development for this project was the web analytics, seen as an ensemble of technologies and methodologies used to collect and analyze data of a website activity. This report is structured in two main parts: within the first part, a technical overview will be given to the reader about machine learning and web analytics concepts in order to ease the comprehension of the second part. The second part is then focused on the development of the project, with deep focus on the objective and achievements of each phase.

1.2 Problem Statement and Process Description

The internship object of this report has been made during a period time of nine months. Its main objective was to provide a solution in order to address the following business question:

"How can we increase our online conversion rate by making our website more relevant, taking into account how the user interacts with the contents of the website itself?"

The conversion rate of a website is a metric used by business which sell their products online in order to evaluate the profitability of that channel. This metrics is calculated with the equation below:

$$conversion\ rate = \left(\frac{total\ conversions}{total\ simulations}\right) * 100$$

In this particular business context, a simulation is the sequence of completed steps that the user is required to perform in order to simulate the price of a health insurance policy on the website. A user can be then considered as converted if he or she finalized the process of buying of the insurance policy online.

The main idea of this project was to provide the business with a new enhanced analytics solution which would enable the digital department to take proactive actions and improve the experience of the user with the website. With having a brand new improved experience on the website, the business aimed at boosting the number of online conversions and increase the profitability of the digital channel sales. In addition to this, the solution would be also used to provide resourceful insights to the other functional areas of the business impacted by the website activity.

The analytical solution proposed to the business was a set of three main deliverables expected to be developed and delivered at the conclusion of the project:

- A data architecture able to extract web activity data at the highest granularity level, load into a data lake, transform it and make it available in a table for further ad-hoc analytical tasks;
- A predictive model which would predict the probability for an online user to buy the insurance on the website;
- A set of recommendations regarding the architecture and functioning of the website to be implemented by the digital department, in order to have quick improvements in the user experience.

In order to achieve this, a deep business and web analytics knowledge domain analysis needed to be conducted first to better contextualize and address the project within its scopes and objectives. After that, focus was put on the data requirements: it was necessary to identify the data source of the website traffic to properly configure and implement the data extraction and loading workflow. Once this is in place, an Analytical Base Table (ABT) needed to be built in order to further proceed with the next phases of the project. In particular, a deep initial data exploration was necessary to start giving the first insights to the business and identify potential challenges and threats for the business.

Several tools and APIs needed to be tested in order to identify the one which could allow the writer to get information at the highest granularity level. Once the final table was obtained, supervised learning techniques needed to be applied in order to develop the predictive model which would output the probabilities for each customer to buy the insurance on the website. Logistic regression, decision tree and random forest were the algorithms chosen for the scopes of this project.

Before this project was developed, the web analytics activities were performed by the digital department by using the Google Analytics platform: the disadvantage of this approach was that the analysis were conducted at high-level and based on aggregated information. It was necessary then to introduce a new approach that would enable the business to conduct deeper analysis based on user level data, in order to generate more specific insights and put in action targeted strategies.

As it will be better discussed over the next chapters of the document, within the main challenges of this project there was the retrieving of the user-level data, level of granularity required for the scope of this project.

Considering the complexity of the business problem and the many challenges that have been faced over the duration of the internship, the solution has been developed as a proof of concept with space of improvements after an initial implementation within the business processes. In the last part of report are described several immediate actions that could be taken in order to already start improving the website performance, other than a brief description of the next steps for the solution implementation.

1.2 MOTIVATION

In a scenario where companies need to differentiate and emerge from a pool of competitors, it is becoming more and more essential to invest and improve on data solution implementation within most of the business workflows and processes. Taking data driven decisions can only have a positive impact on both profitability and customer experience as the strategies are defined not anymore on assumptions and common beliefs, but on trustful and contextualized insights.

Traditional web analytics techniques that are usually implemented in businesses who aim at digitalizing their ways of working, are becoming outdated and ineffective: these activities are usually performed manually and they are highly time consuming. This highlights the fact that the human brain is not able to work with increasingly larger amounts of complex data that the activity of the

users on the website generates. The insights generated result to be too simple and too dependent from the analyst, along with predictions which are made in a very rough level.

It becomes very important for the business to invest in such approaches as it allows to have a clear view of who the customers actually are and how do they behave in determinate contexts. Indeed, segment profiling and ad-hoc defined strategies allow the business to leverage on an improved customer experience to increase retention rates within their portfolios and more in general the conversion rate.

In this particular case study, the business wanted to invest for the implementation of an advanced analytics solution combining web analytics and data science techniques, with the objective of better understanding the patterns within the users' activity data on the website. Once completely implemented, the solution would be used by the owning department to provide ad hoc insights and recommendations to the other functional units of the business who might need to improve their processes by using information of online website activity, having as a main point of contact the digital department.

Within these business units are worth mentioning along with the digital departments, the products innovation and marketing. The spectrum, though, is not limited only to the mentioned ones as the business was planning to digitalize most of the processes which are still not covered by the website, aiming at increasing the data driven component of the business culture.

The immediate implementation of the solution is aimed at improving the experience of the customer on the website in order to positively reflect this on the conversion rate. The main stakeholder for this project, as previously mentioned, would be the digital department that would benefit from having more detailed insights about the browsing details of the users to improve the structure of the website and better define the digital strategies.

Also impacted from this project is the marketing department which will use the advanced knowledge of the users behavior on the website to better plan targeted actions and improve customer retention and acquisition. Products innovation department is also a stakeholder for this project because the advanced analytics solution would benefit the definition of the new products launching by deep diving in the website audience composition and behavior.

2. LITERATURE REVIEW

2.1 MACHINE LEARNING

The origin of this field can be traced back to the years between 1763 and 1950, when important names as such Thomas Bayes, Adrien-Marie Legendre and Alan Turing started to create the bases of machine learning as we know it today.

Over the past years machine learning has been defined as the ability of machines to learn without being explicitly and preventively programmed. The first one creating the term "machine learning", though, was Arthur Lee Samuel in 1959, even though the most accredited definition considered by the scientific community is the one of Tom Michael Mitchell who says:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

In easier terms, machine learning allows computers to learn from experience: there is learning when the performances of the program improve after each iteration, which can be intended as a process with the final objective of completing a task or a sequence of steps.

Machine learning is strictly correlated to pattern recognition and to the computational learning theory and it studies the building of algorithms which can learn from a sample of data and inductively build a predictive model based on them. It is applied in those fields of computer science where designing and implementing specific algorithms it is not feasible: the most popular use cases, between all, are spam detection for the emails, customer churn and frauds detection.

The problem to be solved is usually defined as such. Given a set of pairs:

$$D = \{(x1, y1), (x2, y2), .., (xn, yn)\}$$

Find a function F such that:

$$For\ each\ i = 1,2, \dots, n \quad F(xi) = yi$$

Where xi is the input and yi is the target value.

2.2 Supervised Learning

Supervised learning is that part of machine learning in which a model is built starting from labelled training data, with the main objective of making predictions on future unseen data. Supervised means indeed that in the sample dataset we have instances for which the output is already known because previously labelled.

With this type of learning, the objective is to find the most accurate mapping function so that when the algorithm is fed with new input data, it's able to predict the output: this is done following an iterative approach, according to which every time the algorithm outputs a prediction, it is adjusted or given feedback until when a good performance is reached.

As previously discussed, the data on which the model is trained includes a set of examples with an input and a desired output: after enough iterations, the algorithm will be able to distinguish and categorize unlabeled data.

Regression and classification are categorized under the same umbrella of supervised machine learning. Both share the same concept of utilizing known datasets (referred to as training datasets) to make predictions.

Supervised learning methods splits up in two different branches: classification and regression. If the desired prediction output is numerical or continuous, then we have a regression problem while whenever the desired output is categorical or discrete, we are in a classification problem.

For this project, the first business problem to solve was a classification problem and therefore the following models, which will be furtherly illustrated in the next few chapters, are:

- Logistic regression
- Decision tree
- Random Forest

2.3 Classification Algorithms

As previously highlighted, classification is a branch of supervised learning where the main objective is to predict the labelled output class of new upcoming instances based on past observations.

2.3.1 Logistic Regression

The logistic regression, despite the name, is a classification algorithm which is often applied to solve supervised learning problems. Is a non-linear regression model which is used when the dependent variable is dichotomous, and the objective of the model is to identify the probability of an observation can be identified as one class rather than the other.

The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

The formula of the function is:

$$1 / (1 + e\text{^} - value)$$

Where e is the base of the natural logarithm and value is the actual numerical value that needs to be transformed.

In order for the probabilities to be calculated, the following formula is applied:

$$P(y = 1) = 1/(1 + e - (b0, b1x1, \dots, bnxn))$$

The coefficient beta 0, beta 1, beta K as shown in the fig above are selected to maximize the likelihood of predicting a high probability for observations belonging to class 1 and predicting a low probability for observations actually belonging to class 0.

2.3.2 Decision Tree

Decision trees are non-parametric supervised learning algorithms which can be implemented in both classification and regression problems. The way the algorithm works is by learning from the data, approximating a sine curve with a group of if-then-else decision rules.

The splitting rules get more complex as deeper the tree grows and therefore the model becomes fitter. Decision tree algorithm divides the data set in always smaller subsets and at the same time an associated decision tree is built. As an outcome, the algorithm outputs a tree structure where each internal node of the tree represents a variable, every branch going to the child node represents a possible value for that property and the leaves represent a predicted value for the target variable.

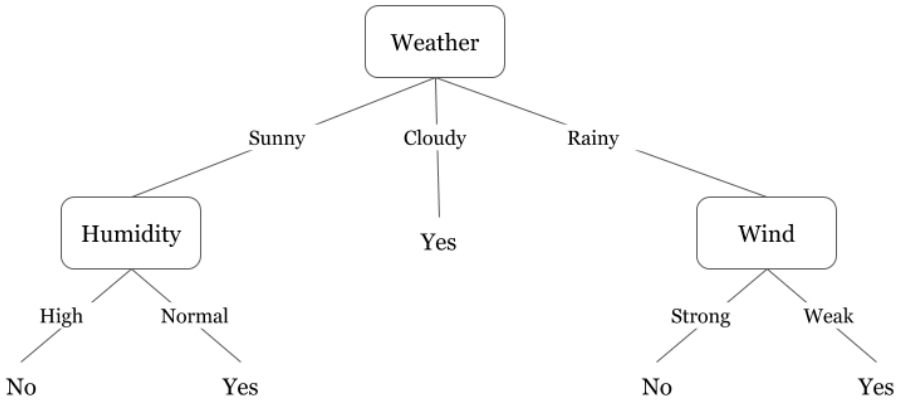

Figure 1. Example of a decision tree

Best practices for this algorithm include the definition of a halting or pruning criterion, with the main objective to control the maximum depth. Increasing the depth of a tree, indeed, does not affect directly the performance of the model; instead, when the decision tree is excessively grown and therefore designed to perfectly fit the all the samples in the training set, there could be a situation of overfitting.

2.3.3 Random Forest

Random forest is a supervised learning method which consist of an ensemble of many decision trees. All the decision trees give as an output a class prediction and the most recurrent one becomes the model's prediction.

Each tree of the random forest is grown as follows:

- At first, N cases are sampled based on the number of cases in the training set. This is done by replacement of the original data. The obtained set of N instances will be the training set for the tree;
- Considering M as the input variables, a subset m<<M is identified such that at each node, m variables are randomly selected and used for splitting. The value of m remains constant during the whole forest growth.
- The prediction of each tree is averaged in order to obtain the final prediction of the model

By using a traditional bagging with decision-trees, the resulting trees might be very correlated because the same features will tend to be used multiple times to split the samples.

The restriction of the features number during the splitting, allows to decrease the correlation between the decision trees in the ensemble; not only, having smaller sets of features at the split of each node the tree can be learnt much faster and then learn more trees in the same amount of given time.

2.4 IMBALANCED DATASET

In a classification problem scenario, the dataset is considered imbalanced when it contains many more instances of certain classes than other. In this case, the classification rules that predict small classes tend to be rare, ignored or undiscovered: for this reason, the samples of the test set which belong to the small classed are more likely to be misclassified rather than the ones who belong to the most prevalent classes.

The imbalance of the dataset in classification problems causes struggles for the classification algorithms as their accuracy is highly impacted by the inequal distribution in the dependent variable. The performance of the existing classifier get biased towards the majority class: since the algorithms are accuracy driven, their main objective is to minimize the overall error and the contribution of the minority class in this is very small. Also, machine learning algorithms assume that the datasets have balanced class distribution and that the errors obtained from the different classes have the same cost.

There are four main methods to solve the imbalance problem in datasets for classification problems:

1. Undersampling: this method reduce the number of instances from the majority class in order to balance the dataset. Usually is the best practice when the dataset is big enough and reducing the number of the training samples has a contained information loss and good impact on latency and storage issues;

2. Oversampling: in this case, the instances of the minority class are replicated in order to increase the frequency of the minority class in the dataset. The biggest advantage of this approach is that there is no information loss;

3. Synthetic oversampling: it is itself an oversampling technique with the main difference that instead of replicating instances of the minority class, it generates artificial data;

4. Cost Sensitive Learning: this technique is not aimed at balancing the dataset itself, but it rather evaluates and highlight the cost of the misclassification of the observations.

For this project, considering the dimensions of the dataset, it is been chosen to apply oversampling for the minority class in order to fix the imbalance. The ROSE package in R has many function which help performing these tasks and in particular the *ovun.sample* one has been used.

2.5 METRICS

The task of choosing performance metrics to evaluate the implemented algorithms could be critical when working with imbalanced datasets. The greatest part of the classification algorithms calculate the accuracy based on the proportion of instances correctly classified. As the minority class has a relative impact on the overall accuracy, in the case of imbalance data the results could be high deceiving.

In order to overcome the above stated issue, the following metrics have been used to assess the performance of the predictive models and choose the winning one:

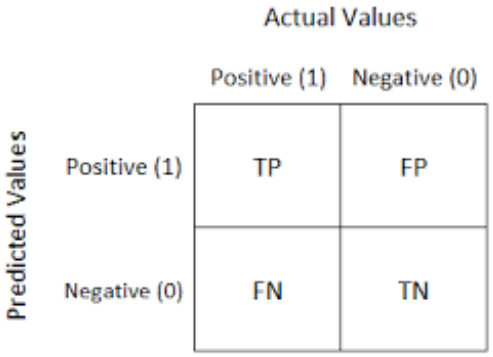

Figure 2. Confusion matrix

A true positive is an outcome where the model correctly predicts the positive class. Similarly, a true negative is an outcome where the model correctly predicts the negative class.
A false positive is an outcome where the model incorrectly predicts the positive class. And a false negative is an outcome where the model incorrectly predicts the negative class.

Metric	Formula
Accuracy	$(TP + TN)/(TP + TN + FP + FN)$
Precision	$TP / (TP + FP)$
Recall	$TP/(TP + FN)$
Kappa	$K = ((p0 - pe))/((1 - pe))$
F1 score	$F1 = \frac{precision * recall}{precision + recall}$
Balanced accuracy	$BACC = (TP/P + TN/N)/2$

Table 1. Metrics used for model assessment

2.6 WEB ANALYTICS

This section is aimed to giving the reader a general context regarding the concepts strictly related to web analytics and its glossary of terms that are usually mentioned when referring to web analytics techniques and concepts.

Web analytics is widely referred as an ensemble of technologies and methodologies that allows us to collect, measure, analyze and provide reports about websites and web applications usage data (Burby & Brown, 2007).

The factors that contributed to the development of this practice within companies of all sizes and across all the sectors are the possibility of measuring most of the actions performed by the website users, the increasing contribution of online activities in companies' profits and the increasing availability of web analytics tool and options (Arson, 2012).

Web analytics can be also seen as an important component of Web Marketing, considering its crucial contribution in calculating the return on investment (ROI) of all the web marketing actions that a business puts in place in order to determine its digital competition in the market. Its role of analysis quantitative and qualitative data of website usage is really crucial when it comes to improving the digital user experience on the website. (Chardonneau, 2011).

Within all the tools available on the market for website activity tracking, Google Analytics is the most used with an estimated market share of 85,4%. The tool developed by Google allows to track different statistics such as the duration of the session, the source of the traffic, the number of pages visited, the pages that are visited the most, the geographical localization and many others.

This tool was the one that was implemented in the business where this project was developed: in the next chapter a focus will be put on some more technical terms which will be often mentioned through the document when going through the details of process.

2.6.1 Technical Terminology

As previously mentioned, Google Analytics is an online tracking tool which is implemented alongside with the website by adding a "page tag". It is a tracking code of Google (GATC, Google Analytics Tracking Code) which is a block of JavaScript code that the user who wants to track its website activity, implements in all the pages of its web domain. This code interacts directly with the Google server, allowing to collect the browsing data of the users which after are elaborated, developed and shown within the platform.

The GACT, other than transferring the data to Google's server, it sets the first party cookie in every visitor's computer in order to memorize anonymous information such as the type of visitor (if it's new or returning), the duration of the visit and the type of the source. In this way, the tool can collect information regarding the user behavior on the website in order to provide the owner of the website with some insights regarding its performance.

Keeping this in mind, Google Analytics uses sampling and aggregating data techniques in order to make reports faster and efficient for the final user, sometimes by compromising the integrity of the information. The result of this approach is that Google Analytics allows the exportation of data from the user interface with the lowest granularity level of information: data can be extracted per group of users based on age, geographical information, source of the user visits between other dimensions.

Although this type of information could be very useful while looking at the big picture of what's going on between the users and the website, if some more deep techniques need to be applied to better understand the users and their preferences while browsing, this type of data could lead to drawing inaccurate conclusions.

For the scope of this project, for machine learning techniques to be applied, a higher level of granularity was needed as the focus needed to be put on the individual user rather than its belonging group: techniques used to reach such results will be better illustrated in the upcoming chapters.

Moving to a more practical definition of the terms related to web analytics, the table below summarizes the main terms web analytics related that will be recurrently used throughout this report.

Bounce Rate	Rate of visitors that leave the website after having viewed only one page	**Paid Search**	Users that landed on the website by clicking on an ad
Conversion	The point at which an activity or response to a call to action fulfills the desired outcome	**Returning Visitor**	a visitor who can be identified with multiple visits, through cookies or authentication
Hit	Also called a page hit, the retrieval of any item (image, page) from a web server	**Session**	A record of a single visitor browsing a website during a given time period. This can include multiple screen or pageviews, events, or ecommerce transactions. Sessions end at midnight on the day a session was initiated or after 30 minutes of inactivity
New visitor	Visitors who have reached a site for the first time. This is important in comparison with return visitors as an indication of loyalty and site value	**Cookie**	A text file placed on a visitor's computer while browsing a website. Cookies are used to track returning visitors.

Organic Search	Describes search that generates results that are not paid advertisements	**Landing Page**	The page intended to identify the beginning of the user experience resulting from a defined marketing effort. In other words, a landing page is a standalone web page that has been designed for a single objective
Channel Grouping	Rule based grouping of the website traffic source	**Goal**	Activity completed by the user which bring value to its online visit

Table 2. Technical terms web analytics

3. METHODOLOGY

3.1 PROJECT DESCRIPTION

As has been previously highlighted, the project was developed for the Analytics department, in collaboration with the Digital Department. The phases and the tasks below listed have been planned accordingly with both departments and regular follow up meetings were held in order to sync between the activities of the two departments and avoid overlapping of activities.

The project was planned to be developed in five main phases on a time frame of nine months. The structure of the project has been designed by following the guidelines of the framework CRISP-DM (Cross-industry standard process for data mining), with the addition of the data extraction phase which was not included. Below a brief overview of the phases and the tasks performed in each one of them will be given: all of the phases will be then further developed within the next chapters.

Figure 3. Crisp-DM framework

Business Understanding:

- Overview of the business and its organizational structure;
- Understanding of insurance and web analytics related concepts;
- Project contextualization;
- Definition of the project objectives and scopes;
- Definition of departments highly impacted by the project's outcome;
- Literature review.

Data Extraction and Cleaning:

- Data source identification;
- Testing of different APIs for data extraction;
- Research for data extraction tool;
- Data extraction implementation;
- Data wrangling and ABT table construction;
- Data cleaning.

Data Exploration:

- Further deep dive into web analytics concepts and website structure;
- Initial data exploration;
- Building an initial set of insights to provide business for immediate actions.

Modelling:

- Selection of models to be implemented;
- Feature selection;
- Implementation of algorithms selected;
- Models assessment and winner selection.

Results and Solution Implementation:

- Interpretation of model results;
- Validation of results with stakeholders;
- Building a set of recommendations based on first insights and model results;
- Proposal for project implementation.

The table below, instead, provides an indication of the planned duration for each of the phases at the beginning of the project and the actual duration. As it can be noticed, the delays experienced in the data extraction and cleaning and data exploration phases had inevitably negative effects on the timing for the other phases.

Phase	Planned duration/ Actual duration
Business Understanding	6 weeks / 6 weeks
Data extraction and cleaning	9 weeks / 16 weeks
Data exploration	5 weeks / 7 weeks
Modelling	9 weeks / 3 weeks
Result interpretation and implementation	6 weeks / 4 weeks

Table 3. Project phases, planned vs actual duration

The whole project from data wrangling and cleaning to modelling has been developed in R programming language, by using the following R packages:

- *dyplyr*;
- *readr;*
- *sqldf;*
- *tidyr;*
- *naniar;*
- *caret;*
- *RandomForest;*
- *Rpart;*
- *ROSE.*

The data extraction part, as it will be better explained in the following chapters, was performed by Scitylana. Scitylana is a tool that allows to extract raw data from Google Analytics, avoiding sampling and making information available to the highest level of granularity. Data were then stored on Google BigQuery.

3.2 Business Understanding and Scope Definitions

The first phase of the project consisted in exploring the business within its operational areas, product families and products portfolio with the objective of shedding a light on the functional units which were going to be impacted the most by the final outcomes.

At first, an overall analysis has been conducted based on the informative material that had been made available by the company, alongside with the public information which could be found on the website. This has been an important step in order to understand the business contextualization and position in the market, other than focusing on its organization within the business units, how and with which products was able to reach every market segment. For the scopes of the project, it was really important understanding how the business operates and through which channels sells its products. The digital channel turned out to be the one with the lowest percentage of insurance policies sold, fact that reinforced the importance of this project's objectives: an improved user experience would inevitably reflect in an increase of the profitability for this sales channel.

Focus has been put on how the digital department had conducted its analytical activity so far and key points of improvements have been spotted during this phase of research. The way the activity on the website has been monitored up to the moment the project started, was through the Google Analytics online platform. This provides high level visualization on the website audience composition, but it doesn't allow to drill down to a user level information. Also it makes it impossible to make cross analysis through several features as the platform is not flexible to perform such tasks.

After this step, individual interviews have been made to the subject matter experts of each area in order to get a more operational understanding of the underlying processes from whom was daily involved in the core activities characterizing that area. Furthermore, this step was crucial for

gathering insights on how the project could help to ease their day to day activities and detecting potential challenges that could be faced during the project development.

Within the important considerations that were taken from this first phase there was the fact that the project needed to be developed and implemented in synergy with the Digital department, also identified as the main stakeholder, in order to enhance their work and avoid useless overlapping of activities. Furthermore, from all the meetings held with the area responsible, it emerged a unified need of making the most out of the online users' behavior before they would become their customers in order to increase proactivity on the business side and ease the acquisition. Proactive actions that have been mentioned during the meetings were targeted advertisements, enhanced pricing procedures and tailored new products launching strategies.

Finally, key takeaway was that the digital channel was the way that the business had to acquire customers at the lowest cost, therefore it was important to improve it in order to leverage its performance and contribute to increase profitability.

Regarding the potential risks and challenges of the project mentioned by the stakeholders there were the difficulties in accessing the web data at a hit level, poor footprint of the customer linked to its online behavior and the low level at which the web development framework of the company was.

At the end of the meetings with the responsible of each area, it was then defined that the main stakeholders of the project could be identified in the following departments, along with the potential impact that the project might have on their activities:

- **Digital Department**: identified as the main stakeholder for this project, would benefit from an advanced analytical solution in order to better identify patterns within the website activity. This would have a positive impact on the definition of its digital strategy and make the website more dynamic and effective. The possibility of zooming in the activity of the single user, makes possible to promptly intervene and adapt the website providing contents according to some of the users features;
- **Marketing Department**: the ability that this project would give in better understanding the footprint left by the single user while visiting the website, will have an impact on the definition of remarketing actions and tailored advertisements or promotions for the users;
- **Products Innovation Department**: highly impacted by the project, this department would see its activities strongly enhanced by the insights provided from this solution when considering the launching of new products. Getting deeply to know the audience of the website would trigger the creation of specific strategies both in a market research phase when designing a new product but also to enhance its advertisement in conjunction with the marketing department.
- **Pricing Department**: slightly impacted by the outcomes of this solution, it would surely benefit from an improved conversion rate of the website when defining new pricing strategies

Another crucial part of this phase was defining the scope and objectives in order to better plan and schedule the required activities for the project to be developed. Based on the analysis previously conducted and the results of the meetings held, the following scopes and objectives have been defined.

Objectives:

- Providing the business with a brand new integrated advanced analytics solution in order to have a better view of the activity of the user on the website. The whole solution includes all the phases from source identification and data retrieval to a generation of quick insights;
- Identifying key points of improvement for an improved website user experience. On a long term this is expected to have positive impacts on the website's online conversion rate.

Scopes:

- Identifying a data source which could provide information regarding the online user activity under the highest granularity level;
- Implementing a whole ELT (Extract, Load and Transform) process in order to extract data and make them available for transformation;
- Generating quick insights according to ad-hoc requests from the different units
- Developing a predictive model in order to predict which users are more likely to buy a policy on the website (conversion);
- Propose a set of recommendation based on an enhanced advanced analytics activity on the data extracted at a user level and the results of the model.

3.2 DATA EXTRACTION

The purpose of this phase was to identify the source of the data needed for the project and verify that the level of granularity for the information provided was the correct one in order for the required techniques to be afterwards implemented.

After initial analysis, it's been noticed that the data regarding the activity of the user on the website was not available at the highest granularity level. Several tests gave as a result that even though Google Analytics makes available the data stored on its server through its APIs, what could be pulled was aggregated data, making it impossible to have the single user footprint of the activity on the website. Data are aggregated on the Google server by traffic source, geographic localization, page visited between other segments that can be defined on the platform.

In order to reach this conclusion different attempts were made:

- API calls with different parameters have been made to the “Google Core Reporting API” through Python and R;
- Extracting data from the reports directly on the Google Analytics platform;
- Enabling the Google Analytics connector on PowerBI.

None of these approaches led to a solution for the problem, not only because the output obtained was not aligned with the required one in terms of data granularity, but also because this tool limited the extraction to only seventeen attributes. Indeed, for the purpose of the project, data at the exposed granularity level could not be used in order to implement the machine learning techniques required to accomplish the objectives defined: it was then in this phase that the very first challenge of the project was faced.

It was necessary to find an external tool that could give the writer the possibility to obtain the data in the format that was required by the scopes of the project: after a brief market research, the tool that was selected for solving the issue was Scitylana. Scitylana is a software that captures online usage data of the website and daily dumps a text file in a selected data storage. The data extracted was purely raw, meaning that the text files contained a record per each interaction that the user had with the website.

The main drawback of this approach was that it was not possible to obtain historical data as it was made available only from the date in which the software was implemented. This led inevitably to a delay of this phase, together with the fact that it was necessary to wait some time to gather a reasonable amount of data to work with.
At the end of this phase, the data was gathered in sixty text files which have been then aggregated in a single table with the characteristics showed in the figure below:

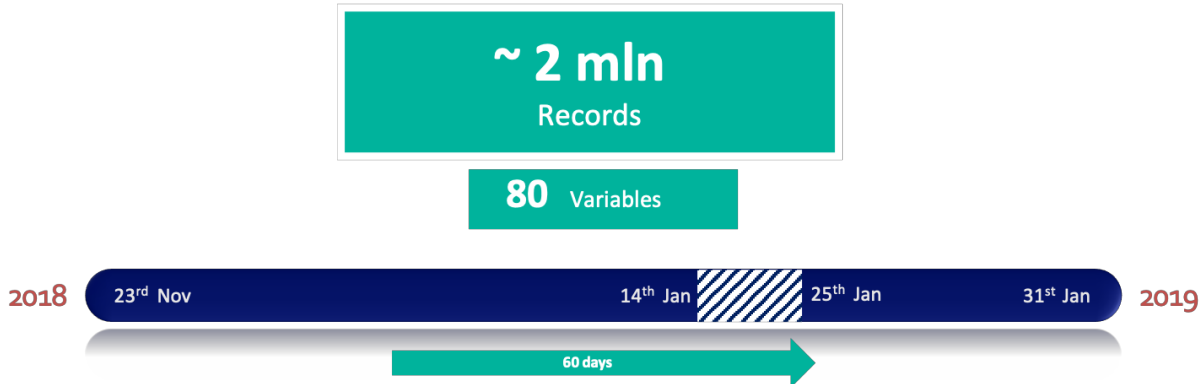

Figure 4. Scitylana Raw Table

The final result consisted in an aggregated table with the below shown characteristics. The time period of the data extraction went from the 23/11/2018 until the 31/01/2019, with an interruption of nine days because of an outage in the collection process. The data was collected during this limited time period because a trial license has been used for the development of this project.

3.4 Data Wrangling and Cleaning

For analytical and modelling purposes, it was necessary to summarize the source raw table in order to have unique records per user. Therefore, in this phase, several aggregation methods have been applied according to business requirements in order to have as an output an ABT which could be used as a starting point for the upcoming model's phases.

During this process, several variables have been dropped as they had more than 70% of missing values either because of technical disruptions that occurred during the extraction or because the information was missing in the source. Also, some variables have been dropped as the information which they contained were considered redundant.

In some cases it was possible recovering the errors or the information missing in the source by computing the values based on other variables, for example:

- It was found out that a goal was not marked as completed in the source, even though the user had been visiting that specific page and completed the action for which the goal completion is triggered. As a reference, in several cases the user had completed the simulation of a policy price, but the correspondent attribute was not valued;
- The variable that captured the count sessions was for some users not valued: in this case, the missing values have been replaced with the count of the distinct session_id for that user.

In order not to avoid any loss of information, all the above mentioned processes have been developed also comparing the final result with the information showed in the Google Analytics dashboard for the time period considered. Point of attention of this task was to make sure that the information contained in the final ABT matched the one showed in Google Analytics.

Feature engineering tasks have also been performed in order to furtherly feed the model with diverse information by extrapolating new variables that are not present in the source, but that could help improve the performance of the models themselves.
The final output of this phase was an ABT of 49 variables and ~200 000 records, containing information regarding the following macro features:

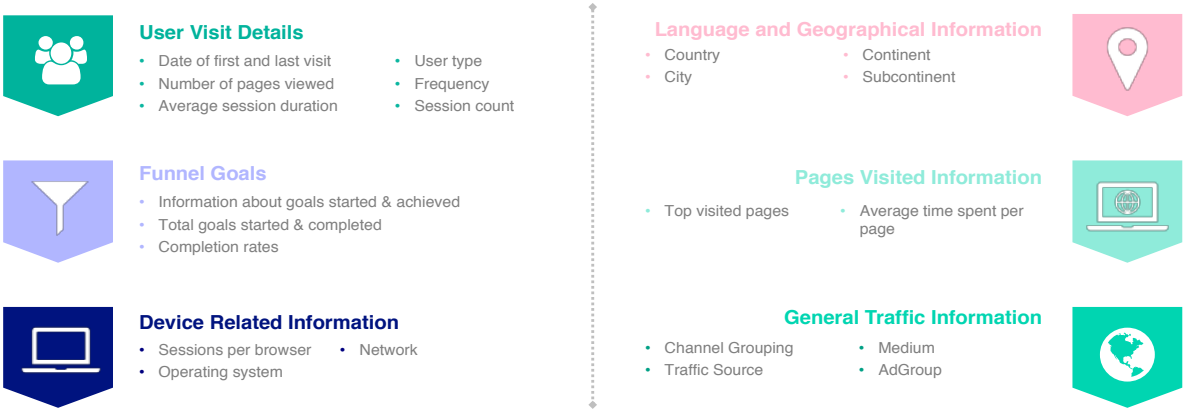

Figure 5. Macro features description

3.5 DATA EXPLORATION

Once the ABT has been finalized, the next step of the project consisted in performing an initial data exploration. This was to find patterns within the data which could help better defining the next steps within the modelling phase. It was also during this phase of the project that a deeper dive has been made in the web analytics terminology and its contextualization within the related business framework:

- It's been highlighted by the digital department that a user is considered converted when it completes the subscription process on the website, even if it didn't proceed with the payment of the policy. This implies that it cannot be fully certain that those users actually became customers;

- The only way possible to link an online user to a customer who logged in the private area, was through an ID assigned by the company. The last two digits of this number identify if the person logged in is the policy holder or one of its dependents. For privacy purposes, though, the digits identifying the position of the customer were encrypted on source side, making it hard to understand if who was browsing the website;

- The location of the user is given by the geographic localization of their IPs, hence this information could be biased on the moment in which the user might be using a VPN (Virtual Private Network);

- In some cases, a single user_id might have two or more than one company assigned ID: this can be given by the fact that different users logged in and visited the website from the same pc and browser. Indeed it needs to be reminded that Google Analytics is able to recognize if those users have been already on the website by the cookies installed in the cache of the machine.

The constraints clarified above have been taken into consideration at the final stages of the project, both for the evaluation of the models' performance and during the formulation of the best actions to be implemented in order to achieve the business goals. They were also main focus points when conducting the initial exploratory data analysis which results will be discussed below.

Next step of this phase was an initial data exploration which has been performed on the ABT, in order to immediately highlight key issues and furtherly guide deeper analysis towards the right direction. The ABT contained information regarding ~200 000 online users, including converted and not converted ones.

In the graphs below it can be observed the languages of the users which converted and the ones of who didn't: the languages showed are the ones of which users' browsers are set on, which could differ from the users' actual language. In both groups, there is a prevalence of European Portuguese as expected since the company mainly operates in the Portuguese market.

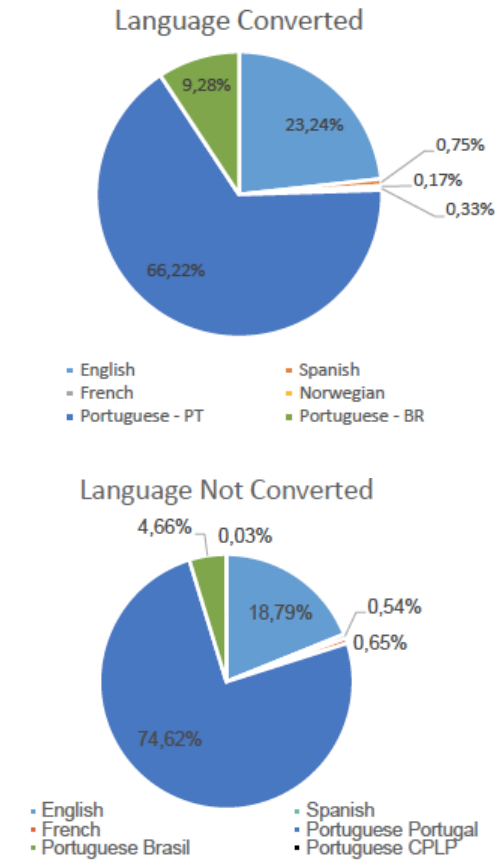

Figure 6. Users description per language

As expected, users converted more from Europe (96,26%): it's worth mentioning though that even for a small percentage of users converted also from the Americas and Africa. This part of users can be potentially a representation of Portuguese living overseas. Not converted users visited the website also from Asia and Oceania: these visits could have been done by Portuguese customers abroad or potential customers that are planning to move to Portugal.

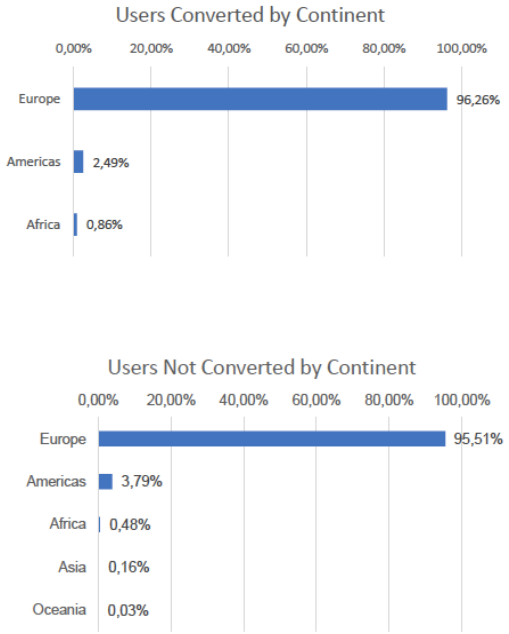

Figure 7. Difference between groups by continent

Even though the business mainly operates in Portugal, the website has been visited in the period of time analysed from 115 different countries. For both converted and not converted users, Portugal along with the Netherlands and the US have been the countries with the highest number of visits.

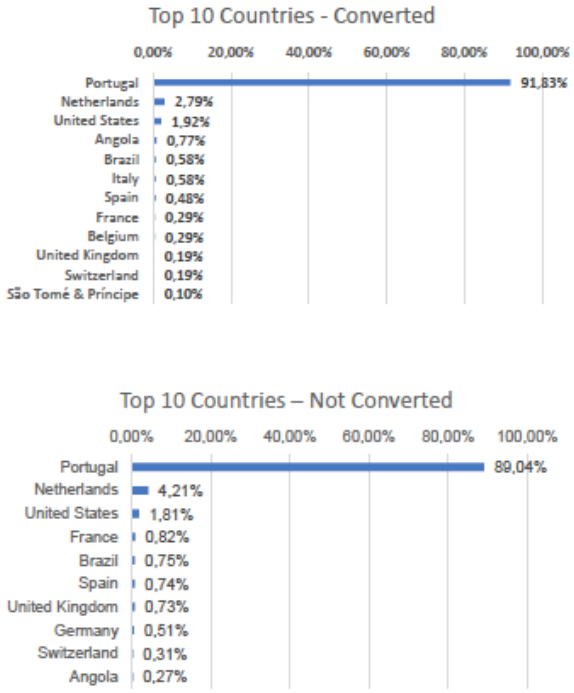

Figure 8. Difference between group by country

In the following charts it can be noticed thar a considerable percentage of users converted from a mobile device: this highlights the fact that a portion of the users do not mind using a device with smaller resolution in order to complete a process as important as purchasing an health insurance. Comparing the two pie charts, it can be assumed that a big part of the users checked first the website from their phones and after converted from a desktop device.

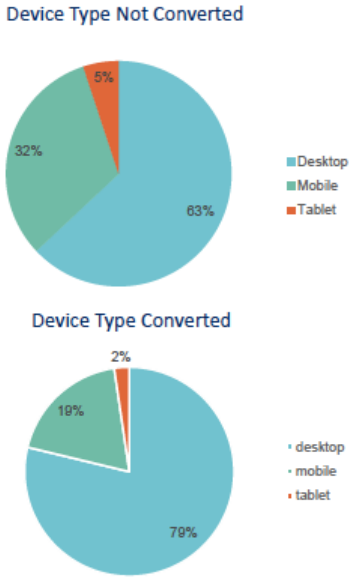

Figure 9. Difference between group by device

As it can be seen in the charts below, the users tend to buy the insurance on the website during the middle of the week, while the weekend and the first days of the week are characterized by a very low conversion rate. The opposite would be expected as it is when the biggest part of the users theoretically has more free time to allocate this kind important tasks.

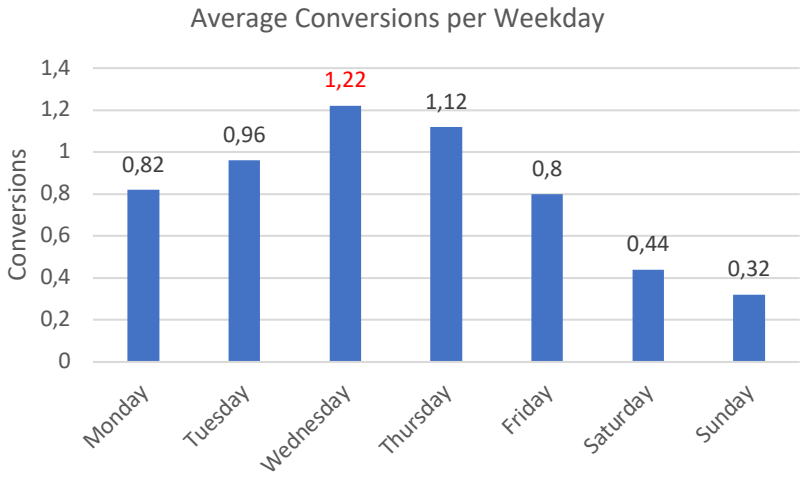

Figure 10. Average conversions per weekday

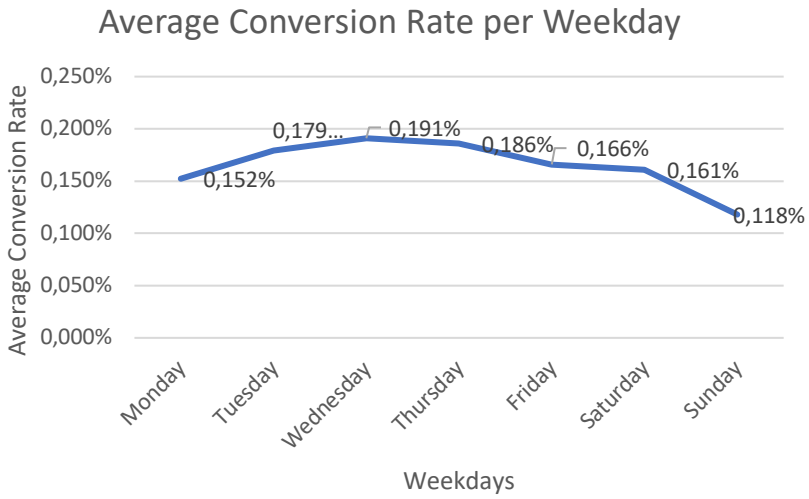

Figure 11. Average conversion rate per weekday

The results of this initial data exploration have been important not only for a better understanding of the data in preparation for the upcoming phases, but also to give some meaningful insights to the digital team. This has been crucial for them in order to take some recommended actions and redesign some of the ways the traffic was canalized through the website.

3.6 Data Preparation

After the initial data exploration, for modelling purposes the ABT has been filtered in order for everything that happened from the moment of the conversion onward to be discarded. As the objective of the model was to predict whether or not that user would convert, there was no need of

keeping information regarding what the user did after buying the policy on the website as this might influence the model's results.

The ABT was composed by categorical variables with a considerable number of levels: this could have been a problem for some models when it comes to performance. As a reference, the *randomForest* package in R doesn't allow variables with more than 50 levels to be passed to the algorithm. For this reason it was necessary to reduce the levels for those variables and this has been achieved by following two approaches:

- When it was possible to contextualize the variable meanings within the business, the levels have been combined into similar groups by using domain and business experience. This has been performed to variables related to pages of the website visited, geography, languages and general traffic information;
- When it wasn't possible to follow a business logic and a domain knowledge is not available, the levels have been combined by looking at the frequency in the observation and at the response rate of each level. This criteria has been applied to all the variables related to device utilization and funnel goals.

The *preprocess* function of the caret R package has been used to center and scale the numerical variables in preparation for the next phases.

Another task that has been performed during this phase was balancing the dataset. As previously mentioned, the dataset was highly unbalanced with 98% for the negative class and 2% for the positive class: in order to correct this unbalance, the minority class has been overbalanced to achieve a 80/20 distribution. This has been possible by using the ROSE R package which allows to random oversampling the minority class.

3.7 MODELLING

In this phase of the project, the algorithms have been selected and fed with the data that have been prepared along the previous phases. As it has been highlighted in the previous chapters, the data extraction and wrangling process took the majority of the time allocated for the project, leaving the rest of the phases with a considerable time constraints. For this reason, only the baseline model for each of the selected algorithm has been implemented as a proof of concept. Another reason behind of this choice was the latency and business context, other than the need of making the models as simple as possible for easier interpretation and faster interpretation.

The three algorithms that have been implemented in this phase were:

- Decision Tree;
- Logistic Regression;
- Random Forest.

While for the random forest no feature selection is needed as per the nature of the algorithm itself, for the rest of the models implemented the features have been selected by using the *importance* of the *randomForest* R package. This function returns the list of the features with the relative importance: the top 10^{th} percentile has been selected as input variables for Decision Tree and Logistic Regression.

Considering the problem, to prevent overfitting a K-fold cross validation has been performed while implementing DecisionTree and LogisticRegression. This has been possible by using the function included in the R package *caret,* which is the one that has been used in order to train these two models.

Accuracy, Precision, Recall, Kappa, F1 and Balanced Accuracy have been used in this case study to assess the performance of the implemented models on validation set. Since the data are highly unbalanced, considering the accuracy as a reference metric was not enough to evaluate the models and therefore the balanced accuracy has been introduced.

Model	Accuracy	Precision	Recall	Kappa	F1	Balanced Accuracy
RandomForest	0.99	0.78	0.62	0.69	0.68	0.80
DecisionTree	0.92	0.13	0.75	0.20	0.22	0.84
LogisticRegression	0.97	0.18	0.29	0.21	0.23	0.64

Figure 12. Metrics per algorithm implemented

Considering the metrics in the table above obtained when running the model on the test set, the selected winning model was Random forest as it was the one with higher values for most of the metrics.

3.8 RESULTS INTERPRETATION

The probabilities output of the winning model have been reconciled with the dataset to perform some results interpretation. The graph below shows that errors have a negative impact on users' experience: the more error they get while browsing the website, the lower is the probability for them to convert. Included in this category there can be found general website backend errors and wrong actions performed by the user during its browsing like wrong password or wrong dates inserted.

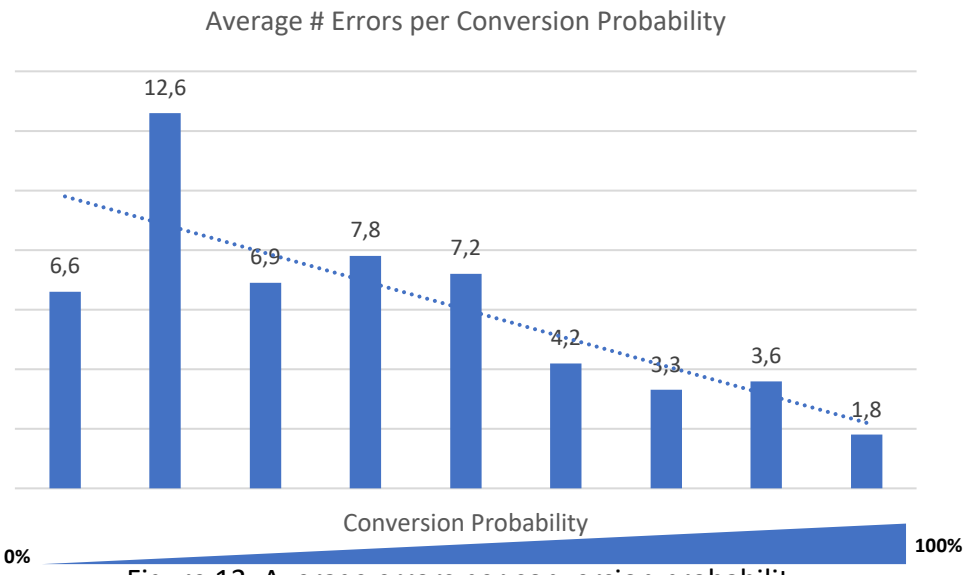

Figure 13. Average errors per conversion probability

The graph below shows how users with lower probabilities of converting have less visits through the Paid Search channel compared to the ones with an high probability: actions could be taken to redirect resources for the Paid Search Campaigns from the second group to the first one in order to increase the conversion rate. As a reminder, channel groupings are ruled-based grouping of the website's traffic sources: Paid Search channel includes traffic coming through the ads posted on the search engines (Google, Bing), Organic channel includes traffic coming through the search engines non paid and Direct is the way the user arrives on the website by directly typing the website address in the bar of the browser.

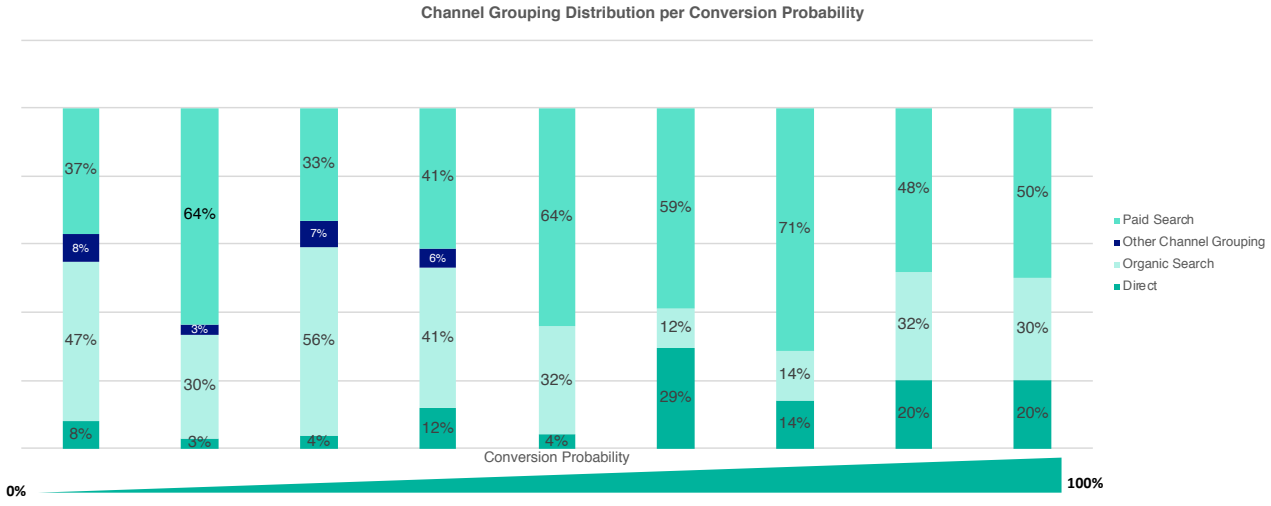

Figure 14. Channel grouping distribution per conversion probability

When it comes to analyzing the differences in visited pages across probability groups, it turned out that users with lower probabilities of converting tend to browse more the website. This could imply that the users do not have enough information regarding general insurance terms, other than the product itself. Indeed, those users before converting tend to visit the pages in the Customer Support category, which are basically the pages where customers and perspective customers ask questions. A

considerable percentage of these users, moreover, asked to be contacted by the contact center in order to clarify their doubts.

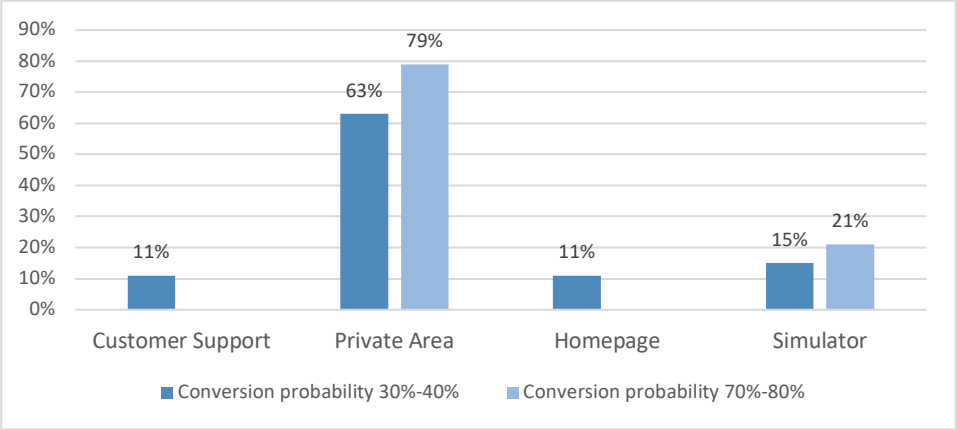

Figure 15. Percentage of users per conversion probability level per page visited

4. SET OF RECOMMENDATIONS

The last deliverable of the project, as mentioned earlier during the project contextualization, was a set of recommendations to provide to the digital team in order for them to start taking proactive actions and slowly drive the website performance to an increase in the online policies conversion rate. This set of recommendations has been built based on the results of the enhanced advanced analytical tasks performed on the website data and on the model's output, and it was aimed to both help enhancing the website optimization and furtherly develop the predictive modelling part.

First of all, a remark has been made on why getting user level data is important in this type of scenarios: having the complete footprint of the users that visited the website allows a better understanding of its activities and highlighted features, other than taking the appropriate proactive actions for acquisition and retention between others.
During the advanced analytics phase, the following scenarios have been spotted and deeply investigated:

- Some users were shown as they had converted multiple times, but by crossing the information with the company internal database, it has been found out that only one policy was successfully payed and therefore issued. This highlights the fact that relying on the aggregated information available on the Google Analytics platform, might lead sometimes to wrong metrics calculations;
- It was also by crossing web traffic data with the internal information that has been spotted a small percentage of users that bought the policy only to get a promotional voucher and cancelled straight after one month. This could be a flag for the marketing department which could think to plan and implement differently their campaign in order to avoid this loss of profit;
- As mentioned during the data cleaning phase, several conversion and simulations were not shown in the Google Analytics platform as they were mistracked by the backend system. Again, this on a large scale might lead to not have the exact and clear picture of simulations and conversion.

Coming to the recommendations, the first part concerns more the further developments that could be done in order to improve the prediction of the users' behavior and implement the correlated proactive actions. More precisely:

- By taking into account the output probabilities of the model, it would be possible to restrict the target audience and optimize costs and resources allocation to leverage conversions on those users that are more likely to buy the insurance policy on the website;
- When more data are available, it would be possible to perform some time series forecasting on the website activity in terms of sessions, visits and purchases, in order to identify the time when the online campaigns would be more effective. Also, predicting the time in which the website will receive the highest number of requests in terms of users' access, it would allow the digital team to provide the website with the needed resources in order to improve in efficiency;

- Having a way to cross online data with the internal information, would allow this solution to be integrated with other data science projects currently up and running in the company in order to furtherly clarify the position of the customers.

The second part of this set of recommendations concerns more the improvements that could be done to the analytics and website optimization tasks in order to get the most out of the information that is pulled regarding the users' interactions with the website. In details:

- It was proposed to ask the email of the customer at the act of the simulation of the insurance policy's price. This could be implemented as a new step in the simulation process on the website and validated with an A/B test in order to evaluate the impact on the user experience. Having the email of the customer at the act of the simulation would allow to have an additional way of tracking the user, other than improve the acquisition and follow-up campaigns.
- Since a small percentage of users shown different location while browsing the website during the time frame considered, another recommendation is to track the localization change through all the users' visits to the website. This could be implemented by using an external tool to the one that it's currently used to track the website and it could be useful to have this information in the moment when the users has been visiting the website from abroad and then starts visiting it from Portugal. This could be flag saying that a user has been reading about the insurance before moving to Portugal and it might be interested in buying one;
- One of the main critical points raised during the data exploration was that main users don't convert or struggle with the conversion process because they're not familiar with the general insurance terms or with the products themselves. The recommendation here is to implement pop up banners on the website whenever the user is spending more time than average on the pages part of the conversion funnel. This could prevent the user leaving the website because of confusion, resulting in a missed conversion.

5. IMPLEMENTATION OF THE SOLUTION

For this project, a trial version of the Scitylana tool has been used in order to develop a proof of concept. In order to proceed with the data extraction, a business version of the software needs to be purchased by the business.

It is important during the implementation to keep the same settings that have been used to set up the extraction process for this project, in order to keep continuity in the data structure and consistency in the code implemented.

Once the account for the business version is set up, the external tool will start tracking and extracting daily data regarding the activity of the users audience on the website and dump a .txt file on a cloud storage. The storage account needs to be identified within the market offer and set up as a project dedicated data lake (Amazon Web Services, Microsoft Azure or Google Cloud Platform).
The dumping of the file will serve as a trigger for the whole process described along the report to be run and have refreshed daily probabilities per each user.

6. CONCLUSIONS AND FUTURE WORKS

In the current scenario, it is essential for a business to get to know its current and potential customers in order to tailor targeted action and improve retention and conversion rates.

The main scope of this project was to deliver a solution able to implement advanced analytics techniques, by enhancing the existing analytical framework of the website monitoring. This would have positive impact on the objective of increasing the online conversion rate by improving the online user experience. Also, this new analytical approach is supposed to support decisions within the other departments of the company who might benefit from having more precise insights regarding the activity of the website.

During the implementation of this project, the key aspect was to identify a way of retrieving data under the highest granularity level possible, in order for machine learning techniques and more ad-hoc and deep analysis to be conducted on the behavior of the user on the website.

After testing several options, an external tool that works as a plugin for Google Analytics has been selected and the data extraction has been implemented. The data extracted was then appended every day into a raw table. At the end of the extraction period, an ABT table has been built in order to be able to perform an initial data exploration. Three baseline models have then been implemented in order to predict the probability for users to buy an insurance policy on the website. The last deliverable for this project was then a set of recommendations for the digital department in order to implement some changes for an immediate improvement in the user experience.

A key challenge that has been faced during the development of this project was the availability of data: the tools that the business had weren't able to capture information regarding the single user footprint on the website. Moreover, once the external tool has been implemented, it was found out that historical data could not have been accessed. This has negatively impacted the project in two ways: first, it was necessary to wait a considerable amount of time in order to get enough data to model with, which resulted in a long delay. Also, the poorness in the volume of the data didn't leave too much space in experimenting and implementing further techniques rather than the ones descripted in the document.

Another critical limitation that this project has spotted is the lack of a CRM (Customer relationship management) system implemented within the company: this makes it harder to keep track of the interaction that the business has been having with the customers so far and leaves out important details which could somehow impact the results obtained.

It is hard to quantify in monetary and performance terms the benefits that the outcome of this project brought to the company: the proof of concept nature of this solution and the many challenges faced during the development of the project, make it hard to evaluate the general impact that the solution provided had on the business operations. Also, the recommendations given as a deliverable of this project were complex and their implementations within the website required resources in terms of timing and labor that go beyond the duration of this project.

However, with the development of this project the business has been provided with a new tool that will drastically reduce the time that was beforehand spent in manual analysis and provide the business with more precise data driven insights.

Once fed with fresh new data, the predictive model developed can give updated probabilities of conversion for each user: these probabilities can be used, along with other attributes extracted during this project, to better define advertisement actions. For example, if the probability of conversion for a user is higher than a fixed threshold and the user has been visiting determined pages of the website, tailored advertisement campaigns could be sent out to that customer.

Furthermore, the possibility of crossing the internal data with the website activity information for the users that are already customers, makes it possible to enrich the clustering techniques already implemented within the business to improve the customer segmentation.

Within the lessons learned during this internship, it needs to be mentioned that the writer got the opportunity to learn how to get the maximum out of the limited resources that sometimes we find ourselves to work with. Also very important was learning how to work with several business stakeholders and communicate the results obtained by the tasks performed: it's crucial to understand the business perspectives and link them with the insights coming out from the data.

7. BIBLIOGRAPHY

A Survey on Decision Tree Algorithms of Classification in Data Mining. (2016). *International Journal of Science and Research (IJSR)*. https://doi.org/10.21275/v5i4.nov162954

Armando Vieira (2016), Predicting online user behaviour using deep learning algorithms, Redzebra Analytics
Arson B., (2012), Web Analytics: Méthode pour l'analyse web, Pearson, Paris

Bekavac, I., & Garbin Praničević, D. (2015). Web analytics tools and web metrics tools: An overview and comparative analysis. *Croatian Operational Research Review*. https://doi.org/10.17535/crorr.2015.0029

Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer
Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., & Cutler, A. (2012). Breiman and Cutler's random forests for classification and regression. *Package "RandomForest."*

Chardonneau, R. (2011) Google Analytics: Améliorer le trafic de votre site pour améliorer ses performances, ENI edition

Coursaris, C. K., Van Osch, W., López-Nicolás, C., Molina-Castillo, F. J., & Rapp, N. (2013). Driving website performance using web analytics: A case study. *19th Americas Conference on Information Systems, AMCIS 2013 - Hyperconnected World: Anything, Anywhere, Anytime*.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In *Ensemble Machine Learning: Methods and Applications*. https://doi.org/10.1007/9781441993267_5

Digital Analytics Association. (2008). Web Analytics Definitions.

Google Analytics reports, metrics and dimensions, https://support.google.com/analytics/answer/9143382

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*. https://doi.org/10.1126/science.aaa8415

Michal Brys (2017) , Using Google Analytics with R, michalbrys.com

Mitchell, T. M. (1999). Machine Learning and Data Mining. *Communications of the ACM*. https://doi.org/10.1145/319382.319388

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of Machine Learning (Adaptive Computation and Machine Learning series). In *The MIT Press*. https://doi.org/10.1007/978-3-642-34106-9_15

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research*. https://doi.org/10.1080/00220670209598786