Università
Ca'Foscari
Venezia

Digital Marketing and Customer Analysis

Churn Prediction – Report

Kanan Mammadli 888195

Pierfrancesco Montello 887309

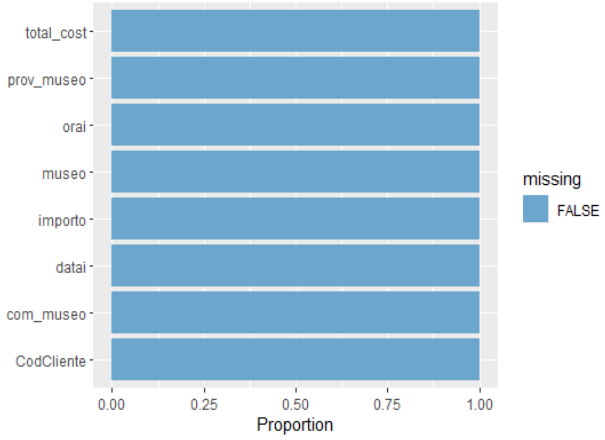
Caterina Tedeschi 883594

0. Introduction

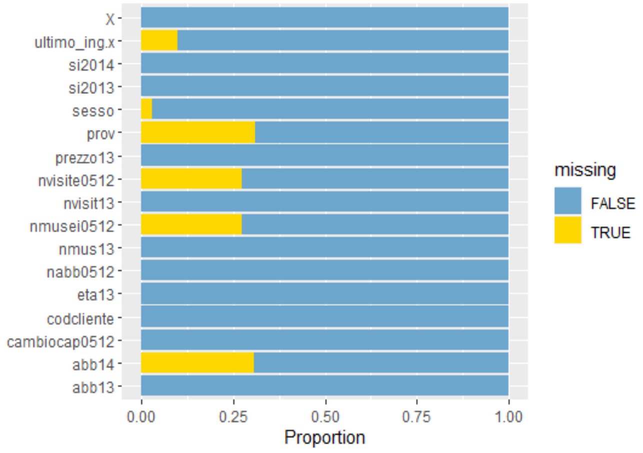
The goal of the project is, after providing an explanatory data analysis on the customers' habits and characteristics and selecting which features have the most impact on the probability to churn, to plan a direct marketing campaign with fixed budget based on statistically-based prediction model for churning probability.

1. Data Preprocessing and Cleaning

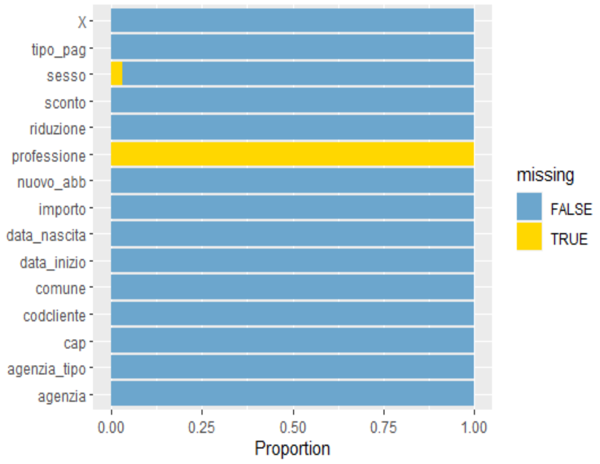
Once we imported our datasets and transformed each variable in the appropriate type of data, we analyzed each dataset to find any missing values.
While the dataset regarding the single visits to each museum (called "use_of_card") was full, we found some issues to solve in the dataset regarding the subscription's renewals ("renewals") details and in the one with the client's characteristics ("cardholders").



Missing value percentage of “use_of_card” dataset



Missing value percentage of “renewal” dataset



Missing value percentage of “cardholders” dataset

In “cardholders” data having 100% missing values for the “profession” column we dropped it while, since we had 3% missing values in “sesso”, we imputed with the mode, assuming that even if we missed the right label to impute in such few cases, we wouldn’t compromise the results of the analysis.

For the “renewal” dataset we did same to column “sesso” and imputed using the mode (if the column referred to a qualitative feature) or the median (if the column referred to a quantitative feature; the median is also more robust statistic compared to the mean in case of outliers), except for the “prov” column in which we replaced the missing values with a new value: “unknown”.

Moreover, in the last two dataset we addressed the issues in “data_nascita” and “eta13” which were presenting unrealistic values like birth dates preceding 1900 or negative ages.

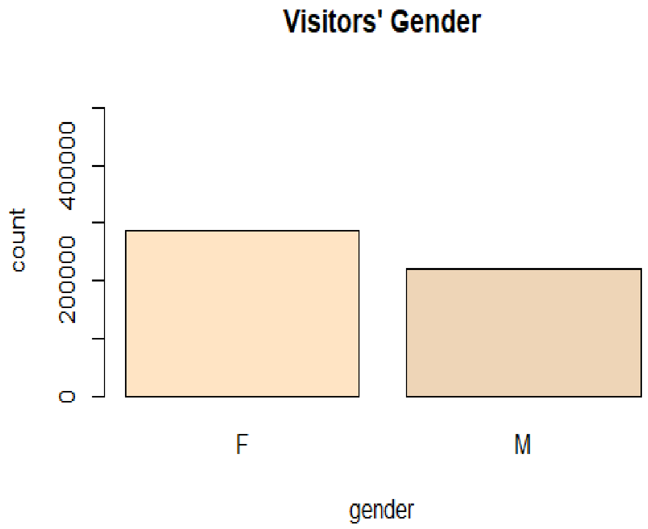
After that, we counted how many different values there were in each column of each dataset and then we merged the three datasets using the client code as key.

After the merge, some rows with a lot of missing values appeared, but since they accounted for less than 1% of the data were dropped, as well as other rows which had logical inconsistencies (e.g., clients that had “renewal subscription” as discount type while never been subscriptors before).

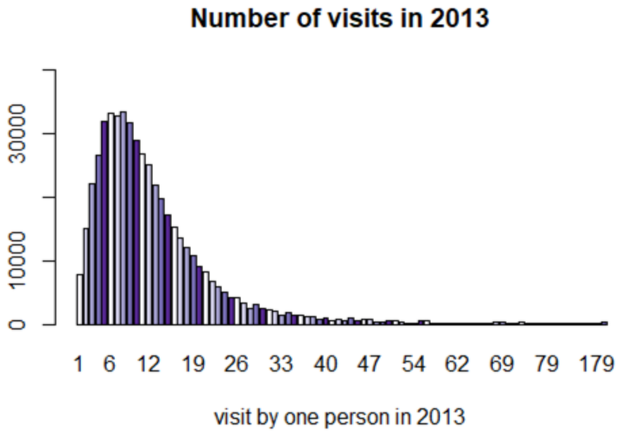
2. Exploratory Data Analysis and Feature Engineering

Before moving on, with the analysis of the variables, the correlations and the impact of those characteristics on the probability of churn with created some other variables that summarize the information about the dates of visits, the times, if the two or more people went together, the total amount that should’ve been spent if the customer didn’t have the card and the distance in days between the subscription and the last visit at a museum.
In particular, the variables with more than 100 different values were dropped because they would have been too heavy on models to compute on and were generally replaced by a summarizing variable (e.g. “cap” and “com_cliente” were replaced by a new province variable which compares the names of each comune with the ones in a file .csv downloaded from Istat website and is filled by the relative province if it’s in Turin, in a province nearby or in a big city; otherwise the variable “prov” takes the value “altro”. Lastly, the coding of the variable that detects churn (“si2014”) was switched (churn from 0 to 1 and viceversa) since it’s easier to identify important variables and compute models that deal with churning probability.
Then we move on to analysis through plots and correlation matrix.

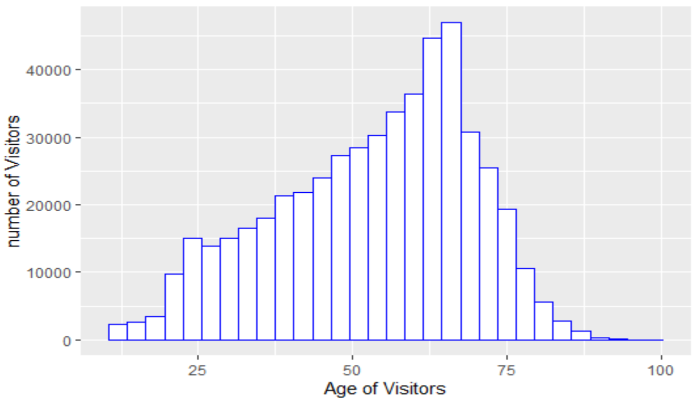
First we want to analyze the distribution of some important variables alone.
We can see that in the whole dataset we have an higher proportion of females compared to males.

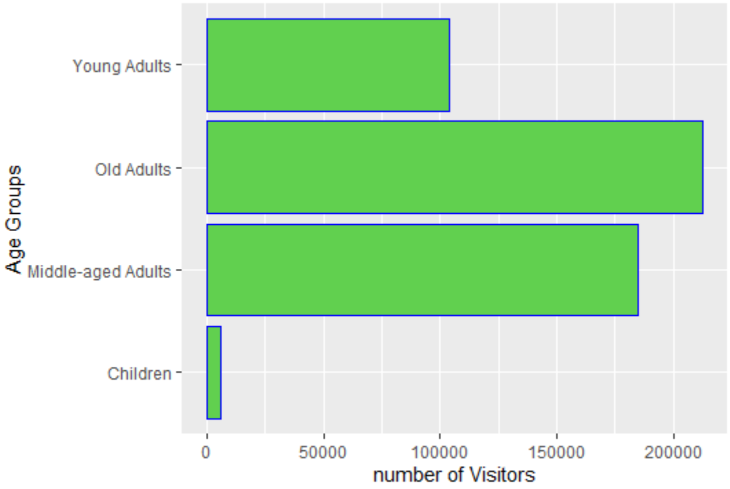


Then we plotted the distribution of the number of visits in 2013 and we can safely affirm that the distribution is highly skewed (we have a mean which is higher than the median) and there a lot of people who make between 1 and 19 visits while as the number of visits increase less and less clients.

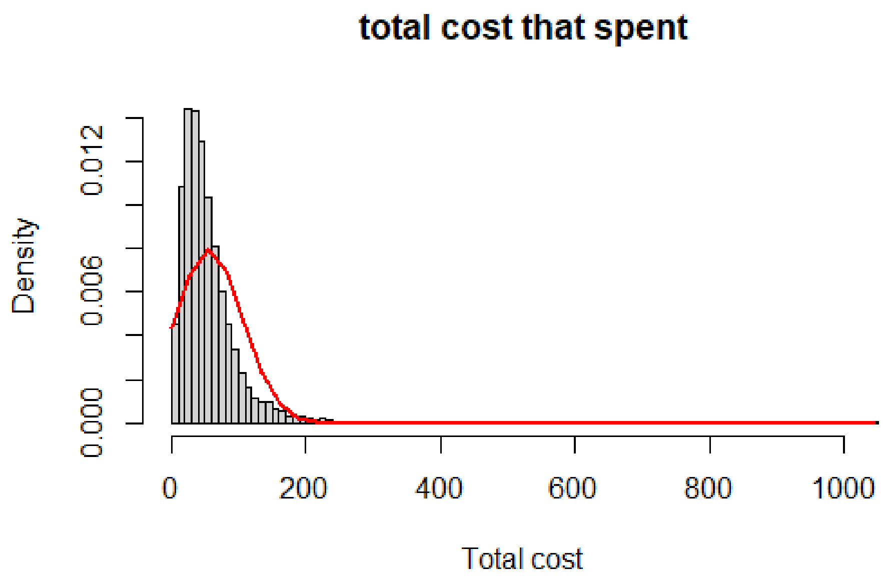


Later, we analyzed the age distribution and we point out the fact that most of the visitors are between 25 and 75 years old, with a mode around 67 years. We can also say confidently that the age category that visits more museums is the one of “old adults” (older than 59 years old) followed closely by “middle-aged adults” (between 40 and 59 years old).



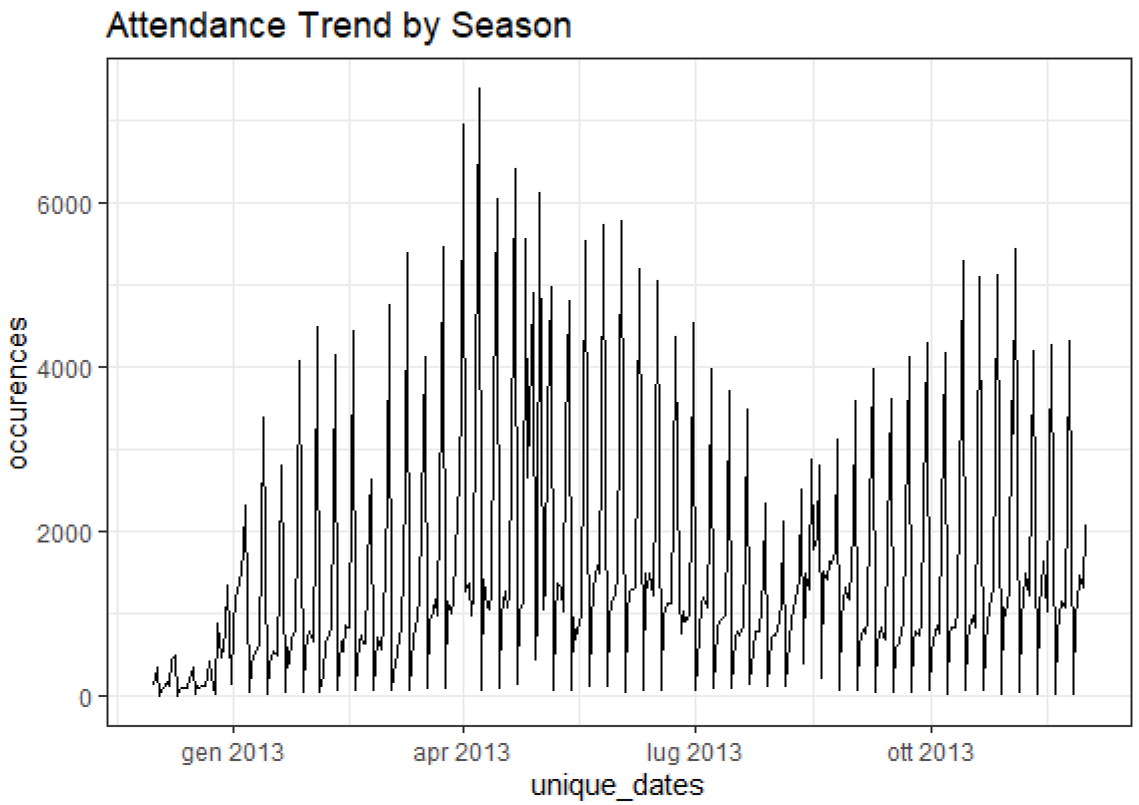


Then we plotted the distribution of the amount of money each consumer should have spent to visit the museums if she didn't have the card. It's evident that few clients are "hardcore visitors" of museums since more than 90% of the customers would have spent less than 150 euros.

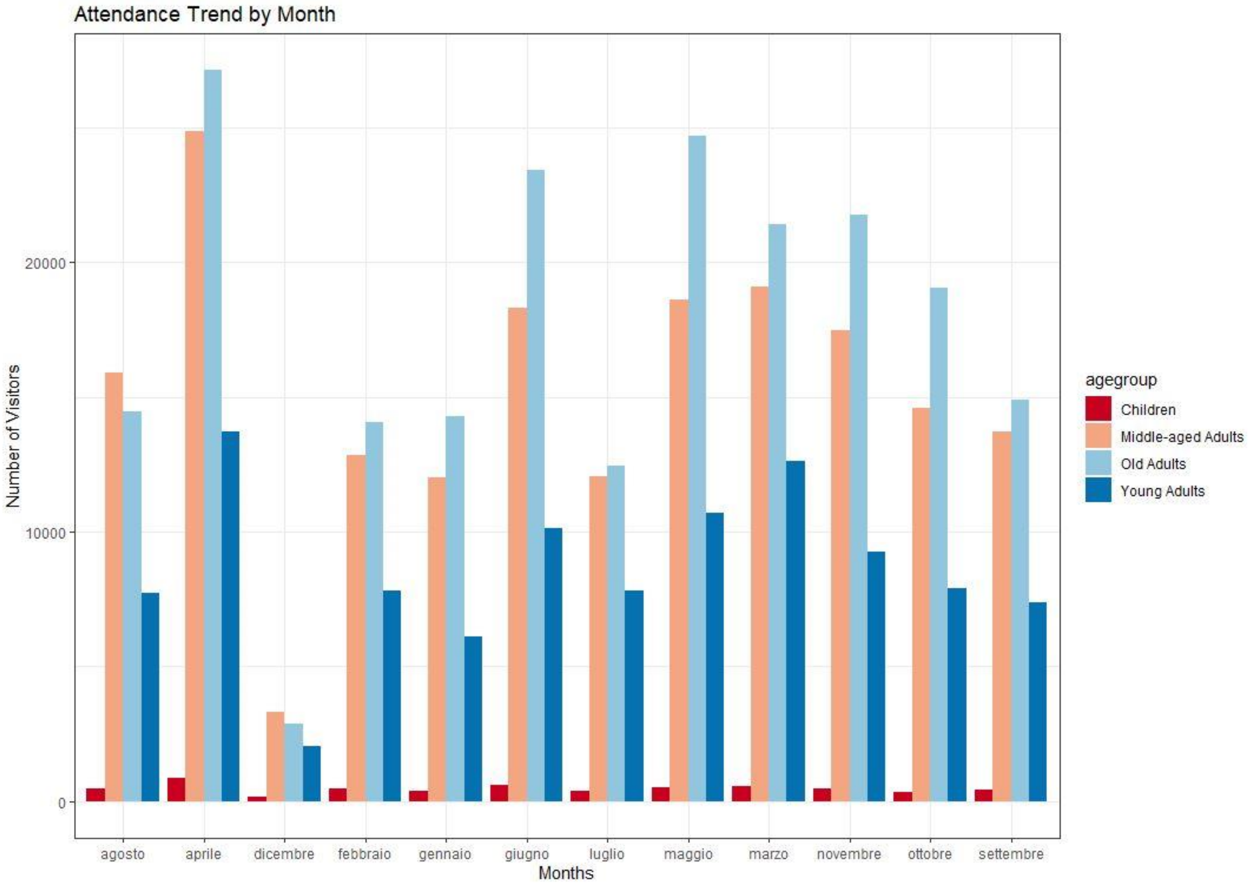


Then, we move to analyze the relationships between variables and the propensity to churn.
To evaluate typical customers' behaviors, the analysis begins with the examination of the attendance trend of clients in different periods of the year from two perspectives: the seasonal and the month ones.

In the first figure, the relationship regards the variable "unique_dates", that results from the calculation of the unique values in the "datai" variable shown in the dataset "use_of_card" arranged by "CodCliente", with the variable "occurrences", that extrapolates the frequency of distinct clients in the same days and compares the original variable "datai" with the "unique_dates" one. The results shows that the majority of customers is concentrated in the spring period, from March to June.

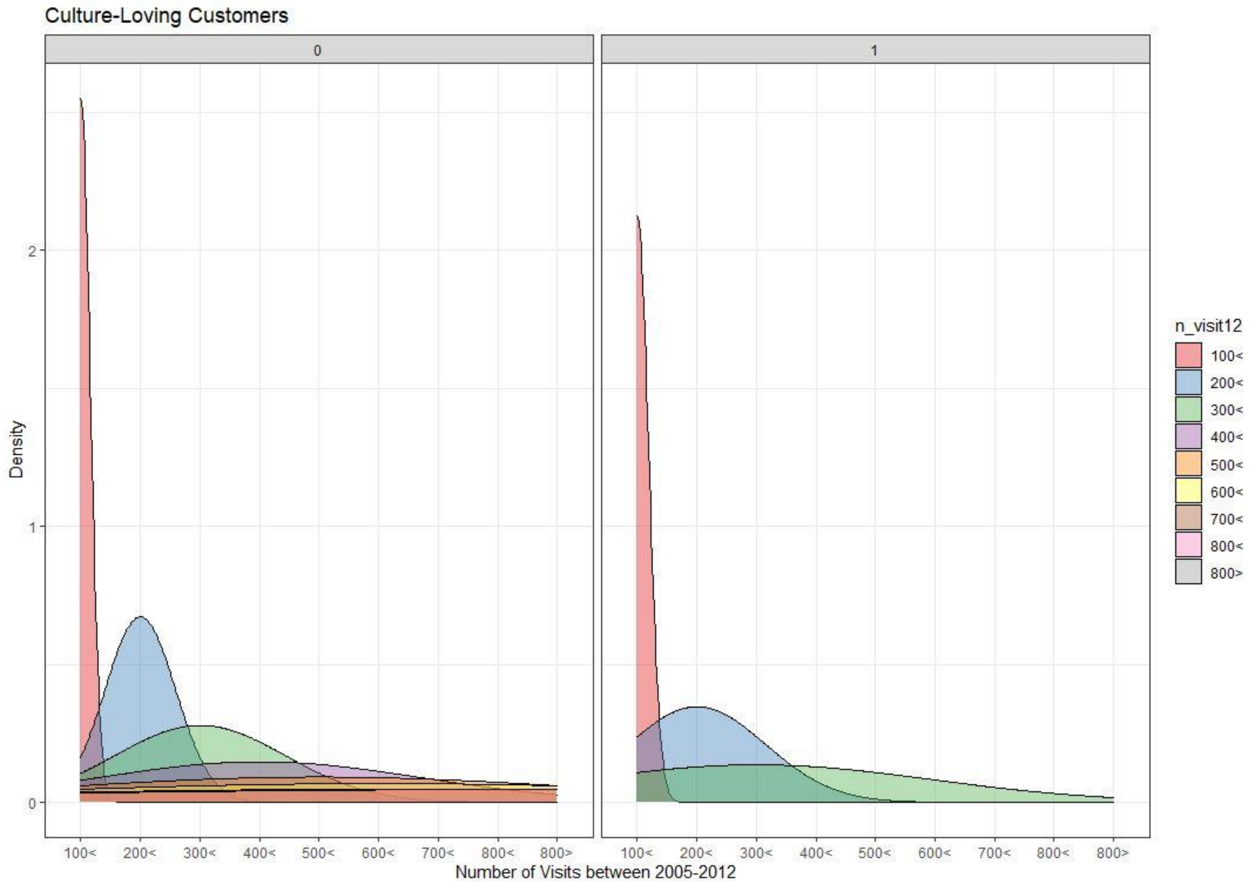


This fundamental output is confirmed in the following graph, where the trend of the number of visitors, divided per age, is displayed in relationship with the diverse months. The variable “agegroup” derives from the distinction of the variable “eta13” from the “full_dataset_for_plots” into four sections: “Children” (0-17 years old), “Young Adults” (18-39 years old), “Middle-Aged Adults” (40-59 years old) and “Old Adults” (from 59 years old on). It is crucial to deduce which are the more popular months and which is the age of clients that are more interested in these cultural activities. In addition, the variable “Months” is the name of the “month_1” column of the dataset “full_dataset_for_plots” based on the values of the column called “data” organized by month through a for loop. The final graph, determined with the combination of the aforementioned variables, indicates that the most attractive months is April, followed by March and June. The presence of children is quite stable in the all periods, whereas the most fundamental one is represented by the “Old Adults” and the “Middle-aged Adults”.

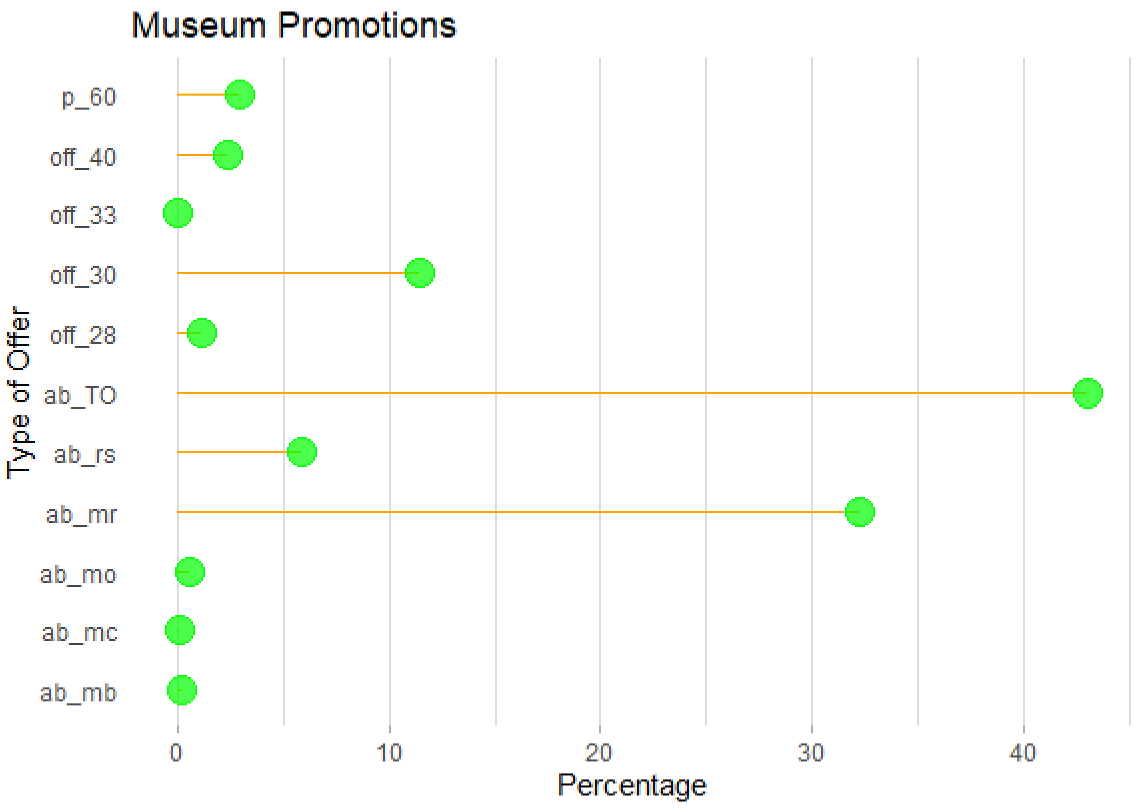


Therefore the museums described in the dataset “full_dataset_for_plots” could launch some seasonal promotions, in particular in the spring period, the most popular one. Moreover, the final result would certainly be more important if some offers had reserved for the diverse age groups, like the children one, to increase their presence.

The variable “nvisite0512” of the dataset “full_dataset_for_plots” is fundamental to detect how many clients are attracted by the museums present in the overall study and also to evaluate their level of interest in culture. In fact, it shows the number of total visits by each client during years from 2005 to 2012. Furthermore, the number of visits are classified into nine groups: from 1 to 100 number of total visits, from 100 to 200 ones until the last group that present a number of total visits more than 800. In the following density plot also another variable is crucial: the “y” one that indicates the level of churning in the audience. The value 0 represents the single client who decided to renew the museum subscription in the passage from the year 2013 to 2014, while the value 1 represents the customers who choose to churn. The final graph shows that the part of the audience who is really interested in the activities promoted by museums, the culture-loving customers, is also the one who renovate the subscription (the left part).

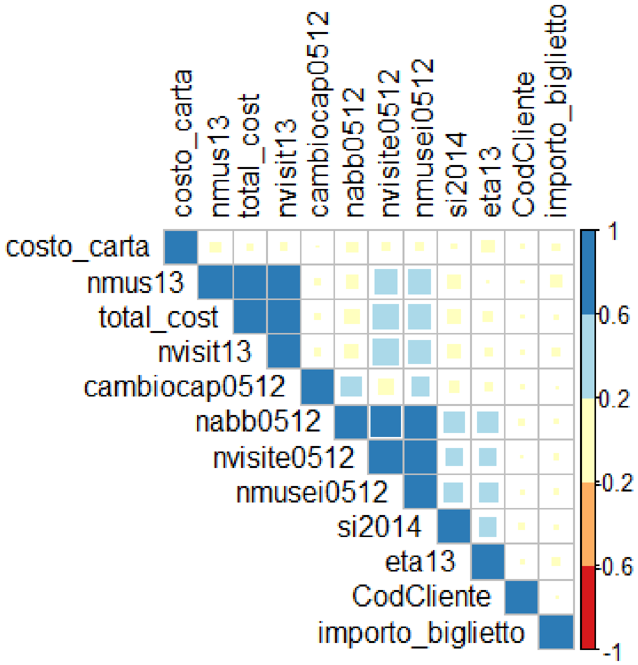


Another significant variable is the “Riduzione” of the dataset “dataset_for_model”. It regards the numerous offers launched by the museums like “Abbonamento musei ridotto”, “Abbonamento musei omaggio” and so on. It is used as the base for the creation of the variable “Offer” that presents the identical values of “Riduzione”, but with the abbreviation of the distinct names to obtain more comprehensible graphs. In particular, the first image reveals the percentage of each offer in the overall dataset. The most diffused is the “ab_TO” (equals to ”Abbonamenti musei Torino”), “ab_mr” (equals to “Abbonamento musei ridotto”) and “off_30” (equals to “Offerta su quantitativo 30€”).



The same variable is plotted in the second figure in relationship with the variable “y”. The result states that clients who decide to confirm the subscription are also the ones who received a higher proportion of promotions (the left graph). Therefore, the museums mentioned in the analysis should introduce more offers to diminish the level of the churning process. In addition, these initiatives could be organized in the spring period, the most popular according to the previous graph “Attendance trend by month” to reach as much people as possible and negative influence the number of churners.

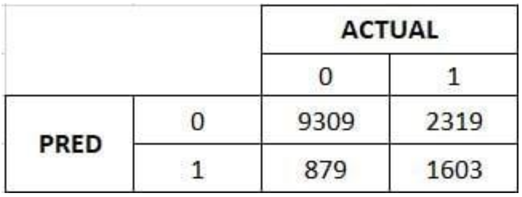
Lastly, plotting the correlation matrix we can see that some important variables are highly correlated and we may consider to drop some in favor of others but since we don’t want to lose any information and think that most of the correlated variables are highly important, we decided to keep them. In any case, if any of those variables would turn out to be not useful to predict churning probability, it’s easy to take it out from the model.



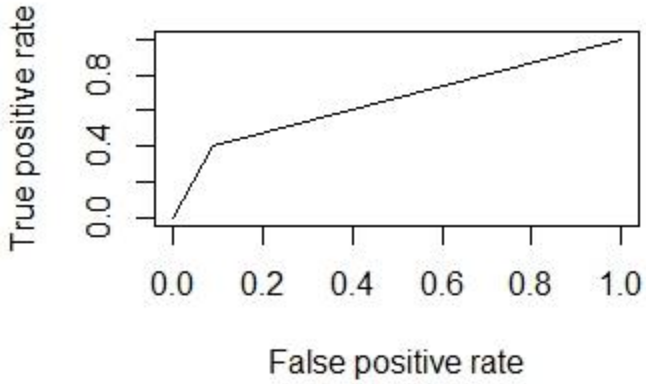
3. Model Development and Testing

Before building any model, as it is good practice in data science field, we split randomly our dataset in two parts: the training set, on which we will train our models, and the test set, on which we will evaluate the models' performances. We decided to keep the usual splitting fraction, 80% data into the training set and 20% into the test set. From that point onwards, we didn't modify the data anymore, even though we had unbalanced classes into our splits (the training had 27.7% of its data as churnerns and the split just 27.8%). That decision was taken after some research (class balancing it's a quite debatable topic) and in finding a specific article (https://win-vector.com/2015/02/27/does-balancing-classes-improve-classifier-performance/) we agreed on taking the route of keeping the imbalance between the classes. Once we were done with all the splitting, we moved to building the models.

The first model we built was a generalized linear model, a logistic one with canonic link function, using all the variables we had, even though from the correlation matrix we could observe some high correlations. We wanted to keep as much information as possible since our goal was to get the highest performing model possible and, if necessary, we could've dropped any non-important variable through model refinement. In fact, over the course of a 7-times-repeated process, we dropped several variables whose parameter were not statistically significant (using an anova test) and could highlight the most important features. The resulting model had **77.3% accuracy**, **40.9% specificity** and **91.4% sensibility**.



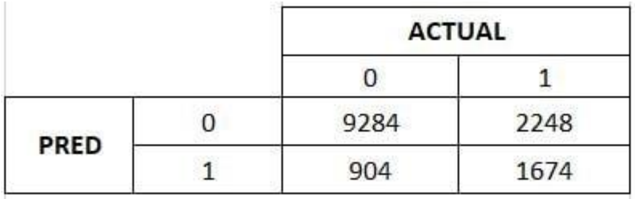
		ACTUAL	
		0	1
PRED	0	9309	2319
	1	879	1603



Quite high values considering that was the first model produced. However, and not surprisingly, the AUC value was still low **0.6612**, meaning either that some improvements could still be made or that maybe we could use other tools to get a better performance.
As a first attempt in improving, we modified a little our model, changing the link function (we tried a probit, a logarithmic and a complementary log-log function), but the results were overall worse in terms of accuracy, AUC or even more importantly, specificity. The specificity, along with relatively high accuracy and AUC, was our main decision-making metric since our main goal was to detect client that are actually churning, considering that a non-detected churn causes an higher loss rather than a non-churn wrongly classified. Assessed that we weren’t on the wrong path to improving performance, we put in our last effort to further refine the glm: we tried both the ridge and LASSO regularization but with poor results.

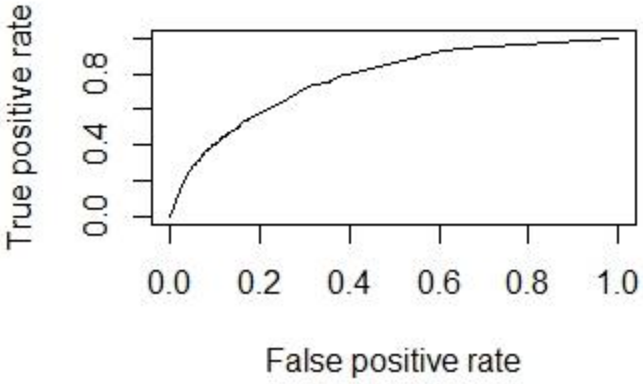
After having dropped any further idea on the logistic model, we moved to a new method: the recursive partitioning. Unluckily, that approach led only to a slight increase in sensibility along with decrease in all our important metrics. Obviously, we kept our glm as the best model.

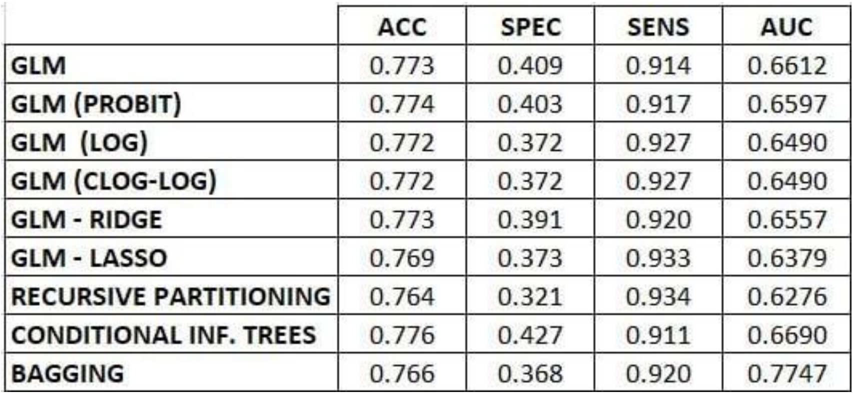
Our next attempt was the method of conditional inference trees which, indeed, had improved performance compared to the glm; **77.6% accuracy, 42.7% specificity** (a highly valuable increase of 1.8% in our main metric) and **91.1% sensibility** (acceptable decrease). Also, the AUC grew to **0.6690**. Comparing the metrics it was obvious to choose the conditional inference trees method as our new best model.



		ACTUAL	
		0	1
PRED	0	9284	2248
	1	904	1674

Lastly, we tried a bagging approach and something awkward happened: while the first three metrics were pretty much the same as the other models and approaches (in fact, lower than the ones of CIT method), the AUC skyrocketed to **0.7747**, in this specific case. That brought us in front a dilemma: "which method is the best to assess our issue?". After careful reasoning we decided to base our following assumptions on the CIT method due to almost 6% higher specificity and the fact that the bagging method has an high AUC because it performs relatively well at different thresholds while the previous methods and models were better only using the usual 0.5 threshold (which was the best and only used) to divide the classes between churners and non-churners.





	ACC	SPEC	SENS	AUC
GLM	0.773	0.409	0.914	0.6612
GLM (PROBIT)	0.774	0.403	0.917	0.6597
GLM (LOG)	0.772	0.372	0.927	0.6490
GLM (CLOG-LOG)	0.772	0.372	0.927	0.6490
GLM - RIDGE	0.773	0.391	0.920	0.6557
GLM - LASSO	0.769	0.373	0.933	0.6379
RECURSIVE PARTITIONING	0.764	0.321	0.934	0.6276
CONDITIONAL INF. TREES	0.776	0.427	0.911	0.6690
BAGGING	0.766	0.368	0.920	0.7747

4. Marketing Campaign Planning

With our CIT method we identified 2578 potential churners among our 14110 clients in the test set. However, to conduct an effective anti-churn marketing campaign that aims to get the most profit (as opposed to get the largest market share, which would imply a different approach), we should try to keep only the clients that generate profit, i. e. the ones that pay for the card subscription more than they make us spend when they visit a museum (every time a client goes to a museum with the card, our company must pay half of the price of the ticket to the museum). Among our 2578 predicted churners only 2411 effectively generate profit, so we're going to focus our attention on them.

Furthermore, we must consider that we pay a cost trying to contact those clients to remind them to renew the subscription. We have two ways in which we can contact those people:

- we can call them by phone spending 1 euro for each call and we have a response rate of 50%;
- we can send them an email spending 0.15 euro each and we have a response rate of 4%.

Our budget is 5000 euros and we assume that once a client responds to the contact she will renew the subscription paying the same amount she did last time.

Taking into account the spending we have, we cannot just cannot contact all the clients that generate profit, but only the ones that have an **Expected Profit** greater than zero assuming that some of them may not answer and that attempt in contacting must be worthwhile (e. g., we mustn't spend 1 euro to keep a client which generates less 1 euro in profit).

Comparing the two contact ways we find that the only one we should use is phone calls: even if we send seven emails to a client (for 1.05 euros of cost) the rate on non-response is still about 75% (assuming the response is a binomial random variable with probability of 4% and that each email is an attempt is unrelated with the others) while with a phone call (and 1 euro of spending) we would get a 50% response rate. Furthermore, we must take into account that we should have a positive expected profit, so we can mathematically summarize everything as:

Exp. Profit (email) >0 : 0.96*(-1) + 0.04*(-1 + client_profit) >0 ➔ client_profit > **3.75 €**
Exp. Profit (call) >0: 0.5*(-1) + 0.5*(-1 + client_profit) > 0 ➔client_profit > **2 €**

With phone call we can contact more people and spend relatively less.

Of our previously identified 2411 profitable clients, 2371 of them generated a profit higher than 2 euros so our campaign will be directed to them. Since, we have € 5000 to spend in our campaign to have an higher probability of people to renew the subscription we can call each client at least twice, in particular we can call the 258 most profitable clients three times and the remaining 2113 twice (calling a client twice the probability response/non-response becomes 25/75% and three times 12.5/87.5%, maintaining the assumption used to compare calls and emails).

The cumulative expected profit obtained from the three-times-contacted clients is **10022.48 €**, while the one obtained by the other contacted clients is **45227.30 €**.
The total expected profit accounts for **55249.78 €.** Multiplying this value by the probability that each client contacted is an actual churner (**~ 64.9%**) and then dividing the result by the campaign budget, you get an expected **ROI** of **717.52%**.