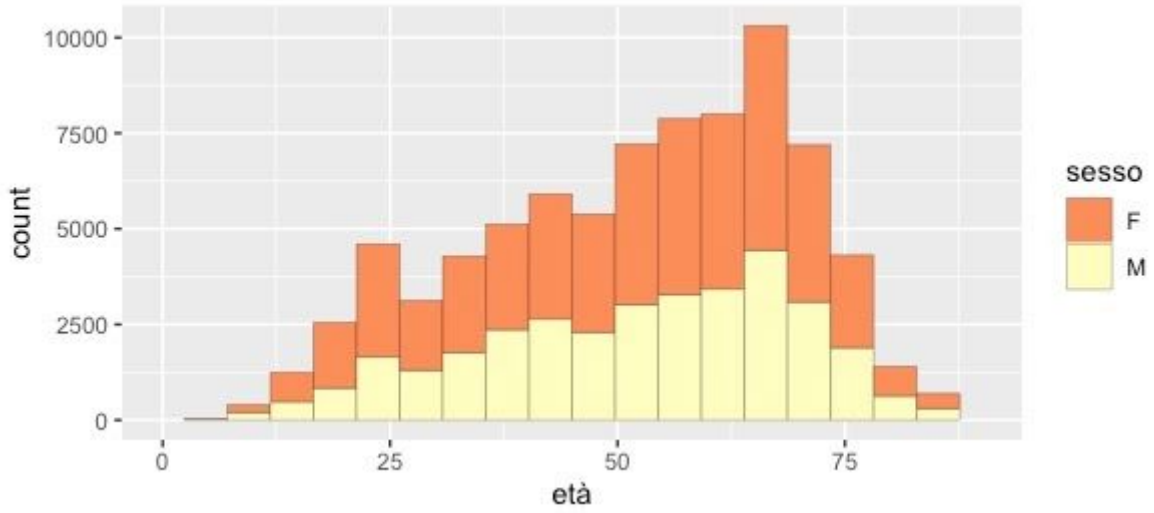
Bassignana Tommaso Matricola 819060

This report's goal is to give intuitive insights about the museum's card customer base and suggest proactive actions to maximize profits.

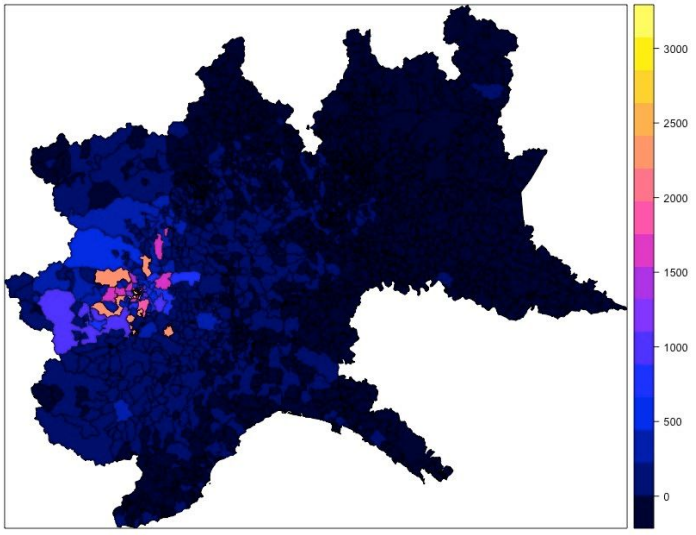
This **first segment** is dedicated to the exploration of the customer base.
Demographic segmentation:

- The members' age range is between 3 and 101 years of age, as in 2014 no limits were imposed. 50% of the total population has an age between 39 and 66 years of age and, on average, they are 51 years old. We can see from the graph below that across all ages, females are slightly more represented than men.

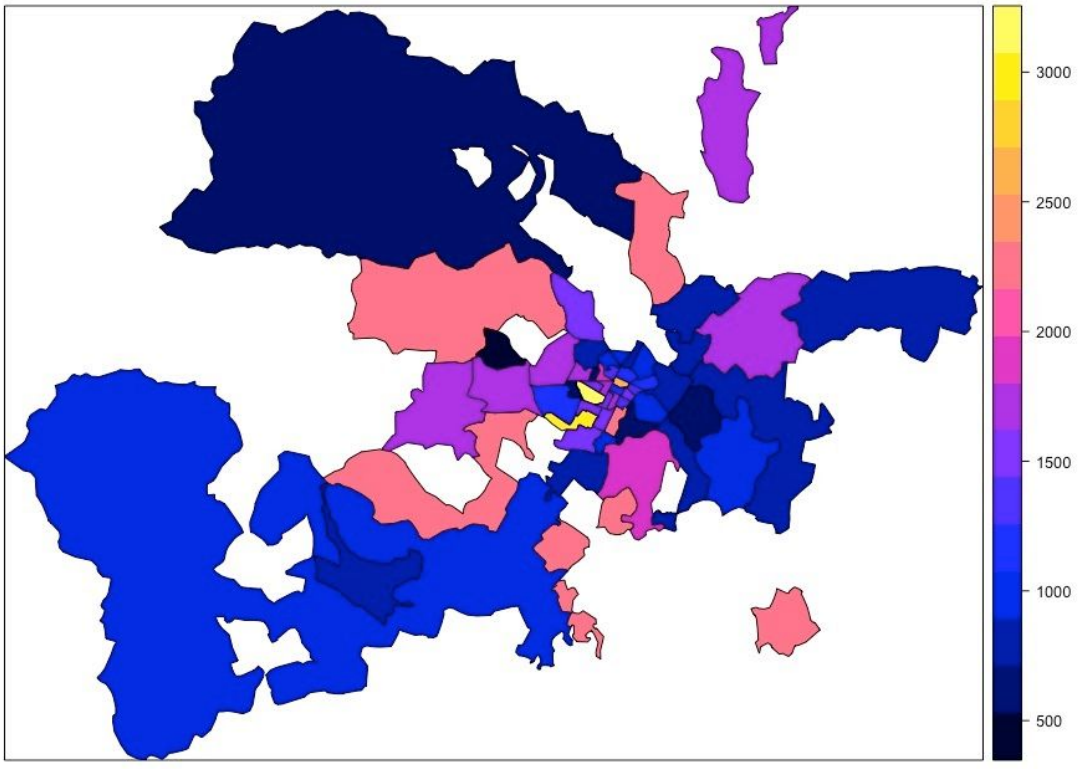


- The female gender, like in the case above, is slightly more represented across all variables, and also in absolute terms.
- The maps below highlight where the majority of the customers live. Colors change based on the number of people residents in the area of a particular CAP, the frequencies are presented in absolute terms i.e. a value of 500 means that in the area delimited by the CAP live 500 people that have purchased the card.

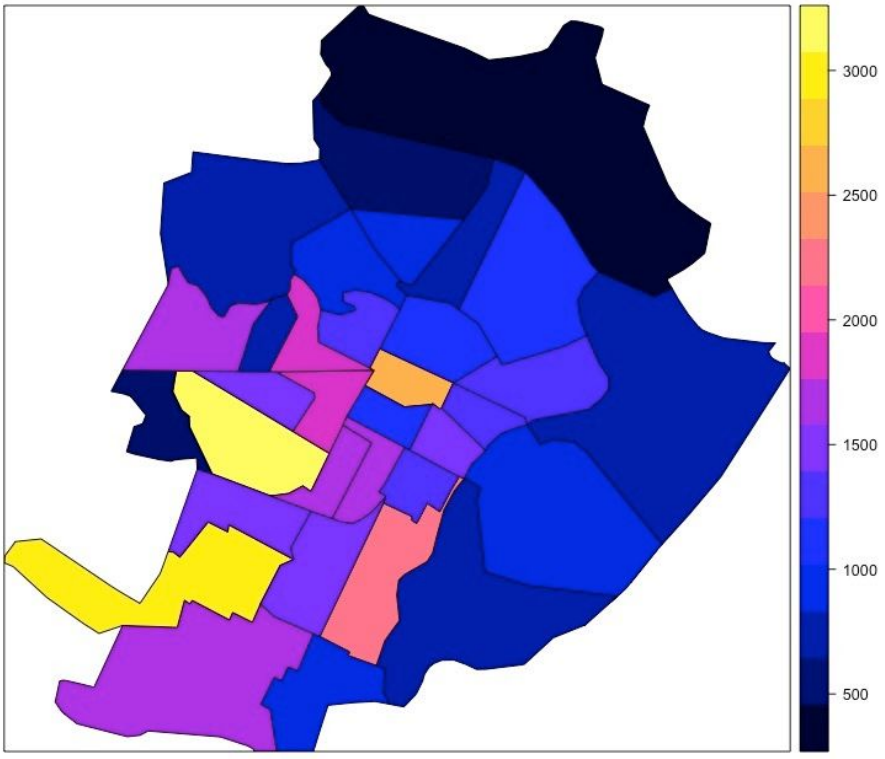
- We can see, in the first map how there are subscribers scattered in a large area around Turin.



- In this second detail, we can see only the regions which have a high(>500) frequency of residents.



- In the last map, we can observe a detail of the city of Turin



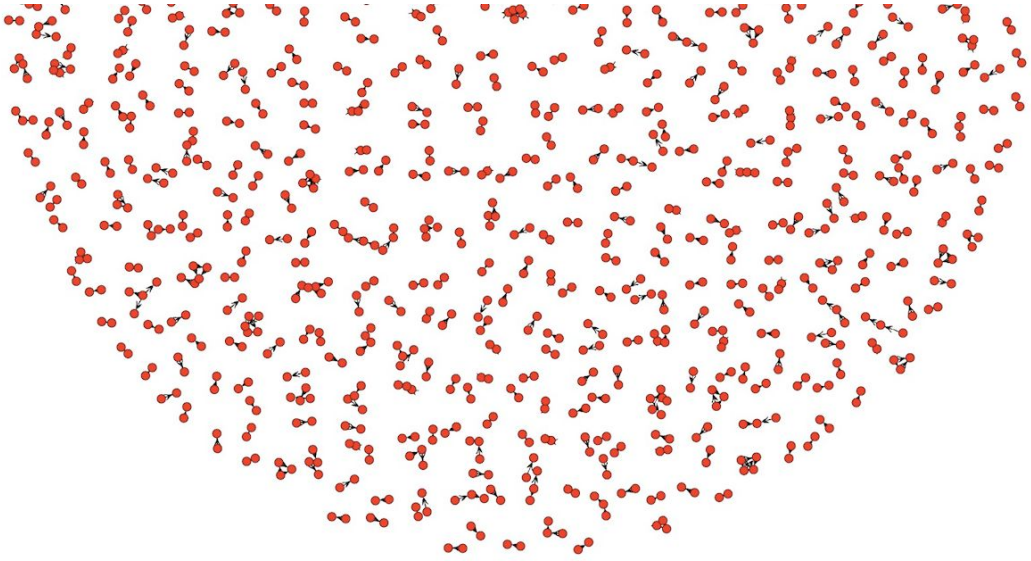
- We can note that despite the areas with the highest frequencies are located inside the city(yellow regions), there are a lot of high-frequency areas dislocated around the city(orange regions in the second map)

Psychographic segmentation:

- The interests more represented in the population are mostly cultural activities, such as visiting museums, going to the theater and purchasing books. The nature of the data forbids us to infer about interests concerning other aspects of life.

Behavioral segmentation: in this section, we explore the usage of the membership cards.

- There are 20582 people, corresponding roughly to the 23% of the total number of members, that go often to the museums with one or more friends. The exact ids of these people can be found in the variable *friends*.

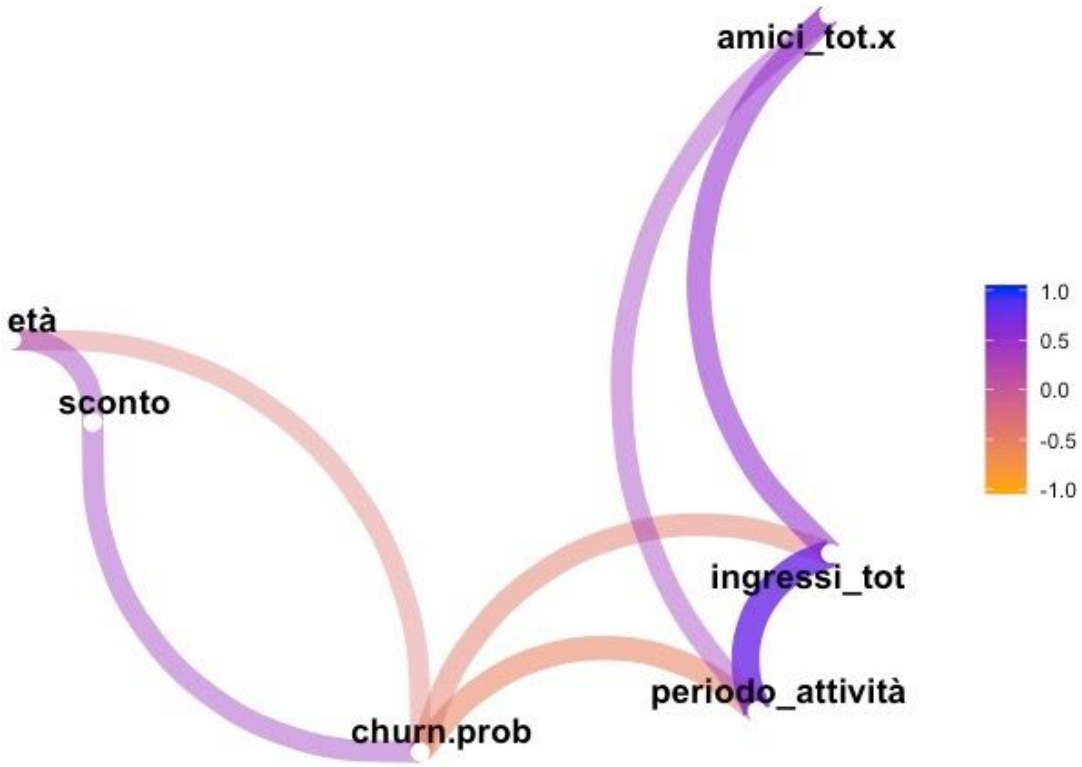


- This network analysis gives us an idea of the groups formed by people that visit together the museums: the more prominent group is the pair, a couple or two friends, but there are also a minority of groups formed by three or more people like shown above.
- Presumably, there are 34465 people that have shared their membership with a person that hasn't had the card. The ids of these people, in conjunction with information about how many times they shared the card, where, when and other characteristics can be found inside the dataset *furbetti*. It is impossible, to the best of my knowledge, to know precisely who they shared the card with.
- 10% of the people who bought the membership never went to a museum.
- 50% of the total of the customers use the card between 2 and 7 times a year, with the mean number of visits being 6.
- Regarding the churn, 30% of the people abandoned the membership. Various pieces of information about these customers can be found in the *churned* dataset.

Occasional segmentation: here we investigate unique circumstances related to the use of the card.

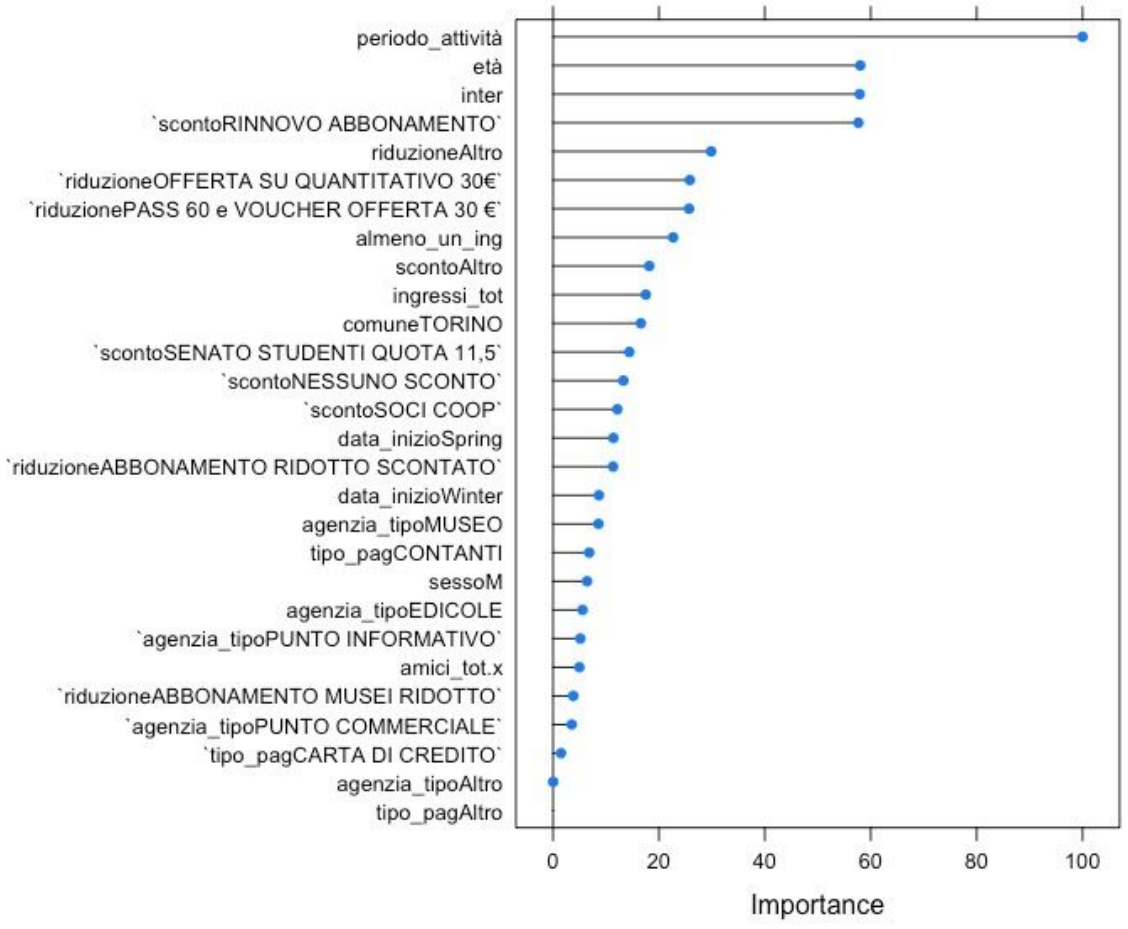
- On average, 64 days elapse between the purchase of the season ticket and the first visit to the museum.
- 55% of the people that used the card at least one time, used it for the last time during the fall season, the 45% of the people used the membership for the last time during summer, spring and winter in equal proportions.
- The use of the card is done, on average, in a window of 172 days.

Correlation between the probability of churn and customer's features



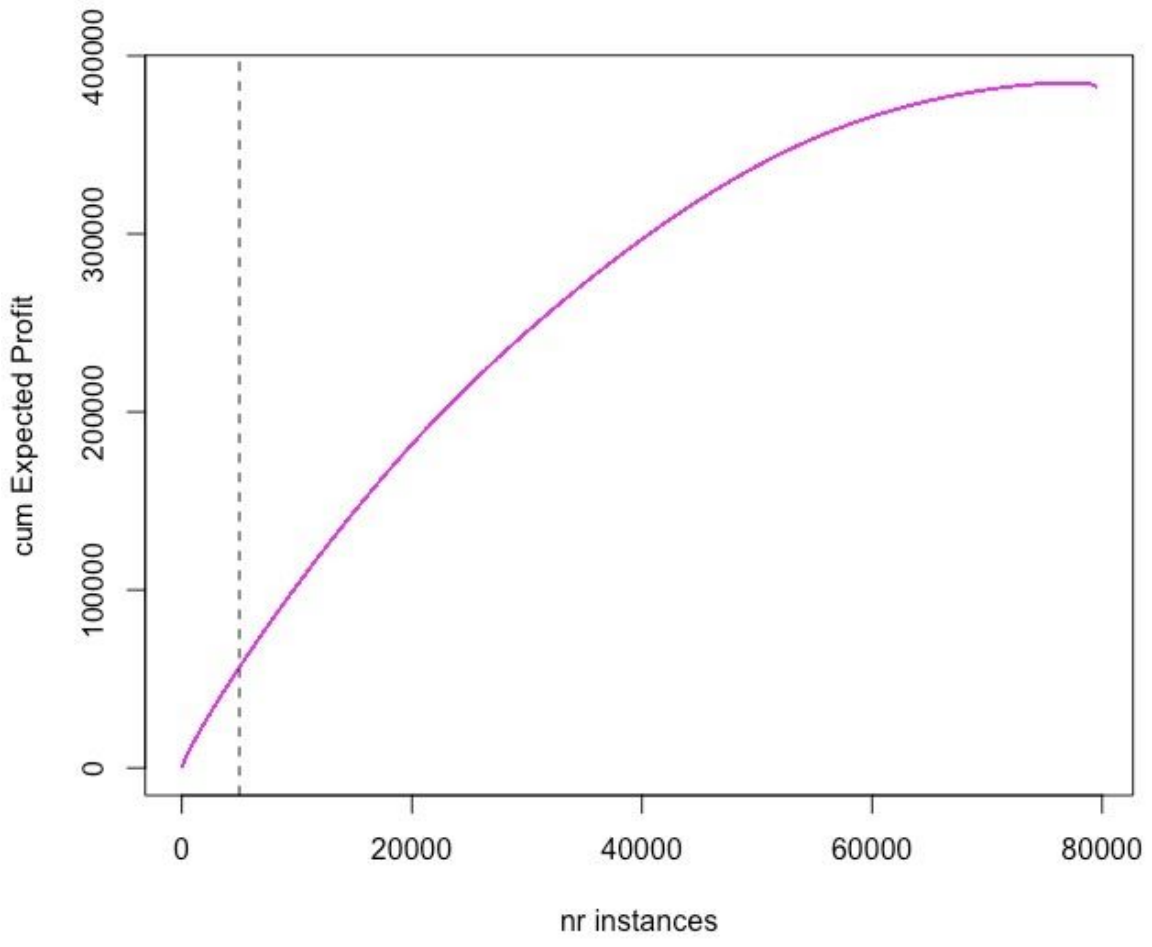
In the figure above are reported the most relevant correlations (or associations in case of qualitative variables) between the probability of churn and all the variables that have a connection with it above a reasonable threshold.
For a better understanding of the relationships between features, I've also included correlations and associations between all the variables connected to the probability of churn.
We can see that a more elevated age, a higher value of periodo_attività and a higher number of ingressi_tot, all diminish the probability of churn, like expected.
Periodo_attività denotes the duration, in days, between the first and last use of the card and ingressi_tot is the total number of visits.
We can also see that the levels of the variable sconto are generally associated positively with the probability of churn and, as expected, the more time a person uses the card the higher the number of visits he or she will do.
Amici_tot.x denotes the number of unique people that the possessor of the id go often to the museum with. This is positively correlated with both ingressi_tot and periodo_attività, a relationship heavily relied upon in formulating the approach in the last part of this report.

The **second segment** of this report is dedicated to exploring what characteristics of the customers have the highest impact on the probability of seeing a customer churn.
Here we present the result derived from a logistic model but other tests, performed with different algorithms, strongly corroborates these findings.
In the plot below, variables are listed in order of relative decreasing importance.



A few things to aid the comprehension of the plot.
The importance measure is relative, meaning that the feature more impactful on the probability of churn will have a score of 100 and, for example, a feature with a value of 50 denotes an impact on the probability of churn of one half compared to the first variable and so on.
Some qualitative feature attributes with low frequency have been collapsed into a unique level, denoted by the suffix “Altro”, placed after the name of the feature, like riduzioneAltro, scontoAltro etc.
This was done because, as can be seen in the graph, there are few variables with high importance but the majority of the impact on the probability to churn is done by a large number of variables, which marginally influences the probability. The attribute Altro was introduced to capture as much information as possible.

The **third segment** of this report is dedicated to providing guidance about maximizing the profits of the Museum's association, which is going to perform a direct marketing campaign. The focus of this analysis is maximizing the profits of the overall marketing campaign rather than focusing on contacting only customers that have a high individual expected profit.
The former approach resulted in higher cumulative profits over the latter.
The best algorithm among a total of six was chosen through the Expected Value Framework; this procedure allow us to evaluate an algorithm performance when there's an asymmetric structure of costs. Note that the expected profit of a customer is weighted by the probability of churn and it is also weighted by the number of people that he/she goes to the museum often with. This is done because of the possible influencing effect that one single customer could have on many others and on their decision to purchase a membership themselves.
The plot below can make some points apparent.



- The fixed price structure of the membership is reflected in a rather steady slope in the cumulative profit curve denoting that no customer will do far better than others in regard to making the Museum's Association some profits.

This highlights that the approach of focusing on a weighted expected profit could be reasonable.

- The vertical dashed line denotes the first 5000 people identified by our model that generates, cumulatively, the highest expected profits possible.
- The people to contact can be found in the dataset *to_contact*. They range from people that have a very high expected profit to people that have a negative profit by themselves but could positively influence others to purchase a membership and everything in between. This denotes, at least, that this approach doesn't focus only on a particular customer type that makes sense only mathematically.
- In those five thousand people to contact, we can find some similarities in terms of characteristics with the full population, but there are some distinctions: there are more people which churn, they have a rather low probability to churn, around 40%, they use the card on average in a window of 100 days.