MUSEUM SUBSCRIPTIONS CHURN PROBLEM

Andrea Cardinali - 911556

Data

The data used for this project consists of information about subscribers to a museum card for the region of Piemonte, Italy.

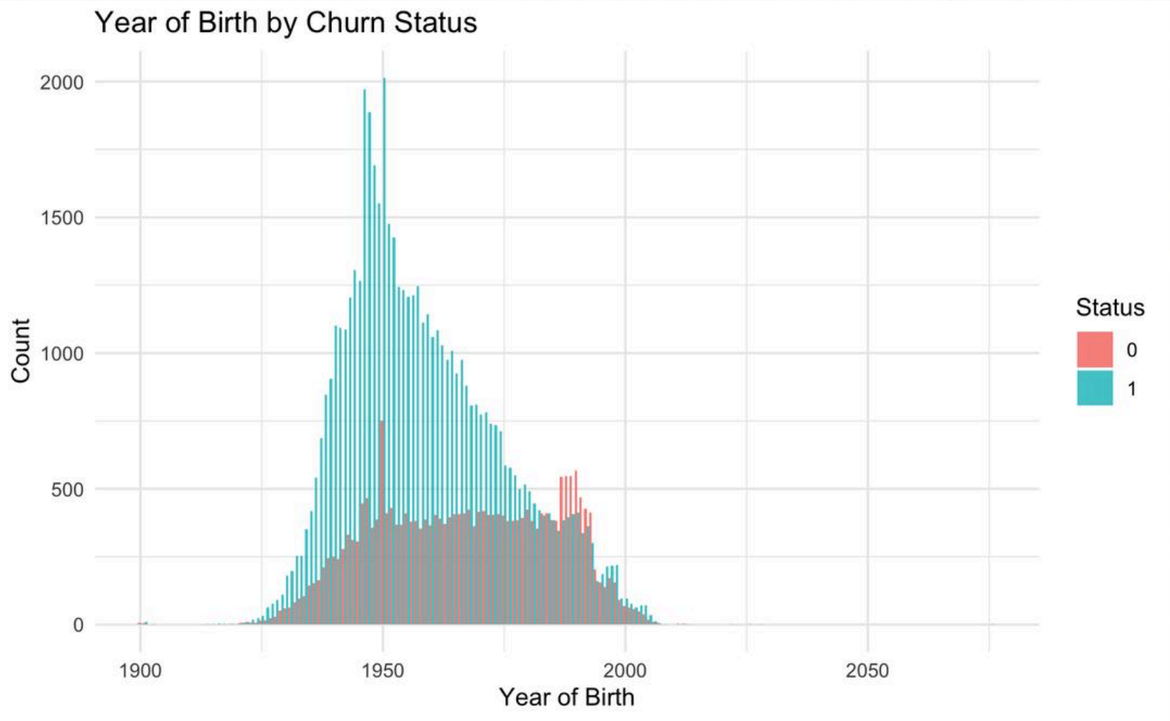
A dataset contains general information about people, like gender, date of birth, their postal code, price paid for the subscription and the discount applied, if any.

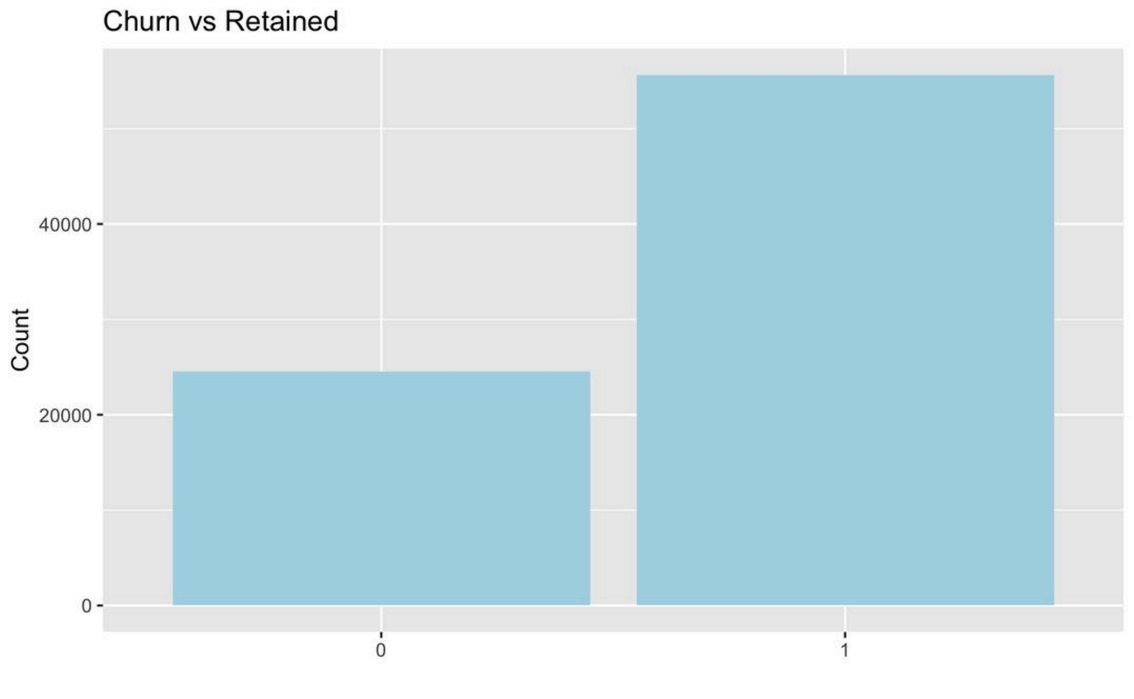
The second and biggest dataset contains each visit made by card holder, like id, museum visited, date and time of the visit, location of the museum and ticket price.

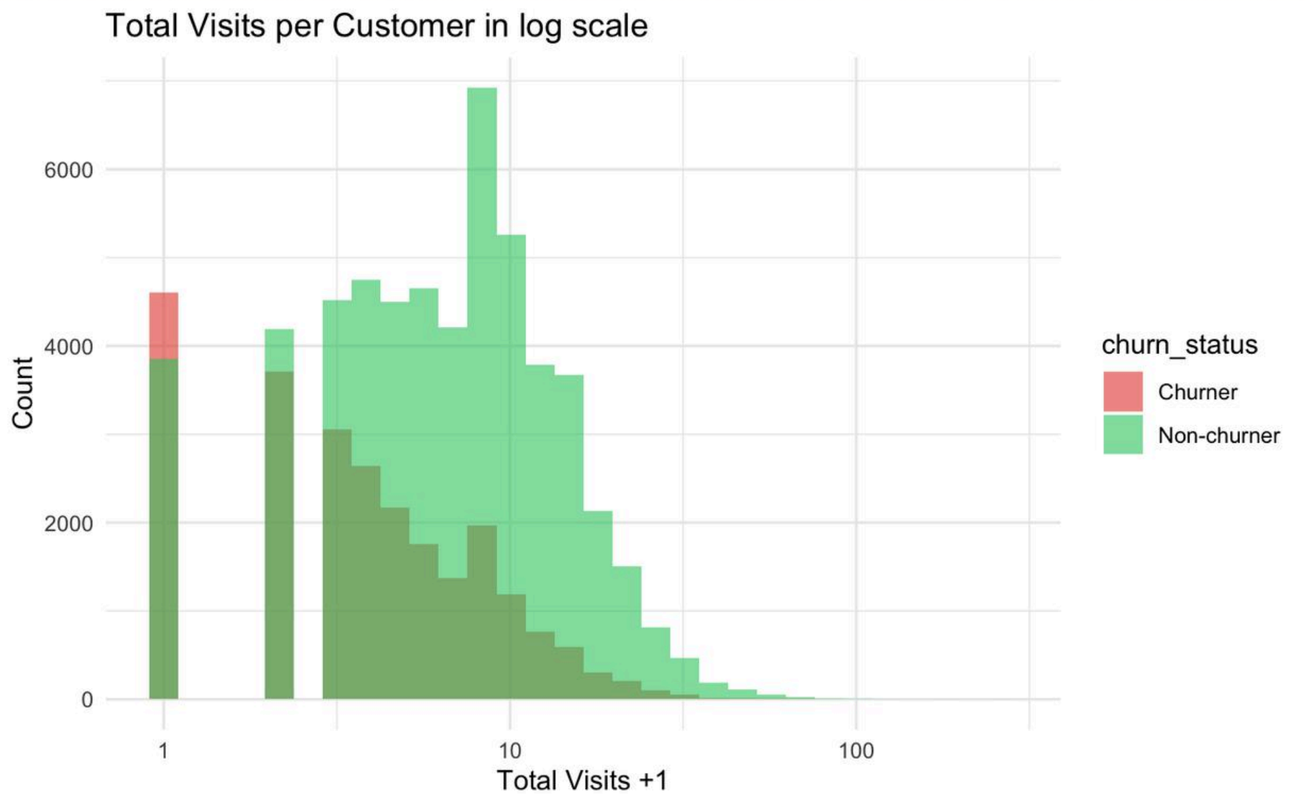
The third and last dataset instead contains the information about customer id, last visit made, whether the customer was also a subscriber in 2013 and if the subscriber churned in 2014 or not.

First thing first it’s important to check how many customers churned for the year 2014. Churners, coded as 1, are a little bit more than 55.000, which is more than double the number of non-churners.

Since it’s not available the age of customers, the date of birth was used instead to compare churners based on the age. We can see that most churners are older people, especially those born around the year 1950. The distribution of non-churners instead is more uniform, without skewness.





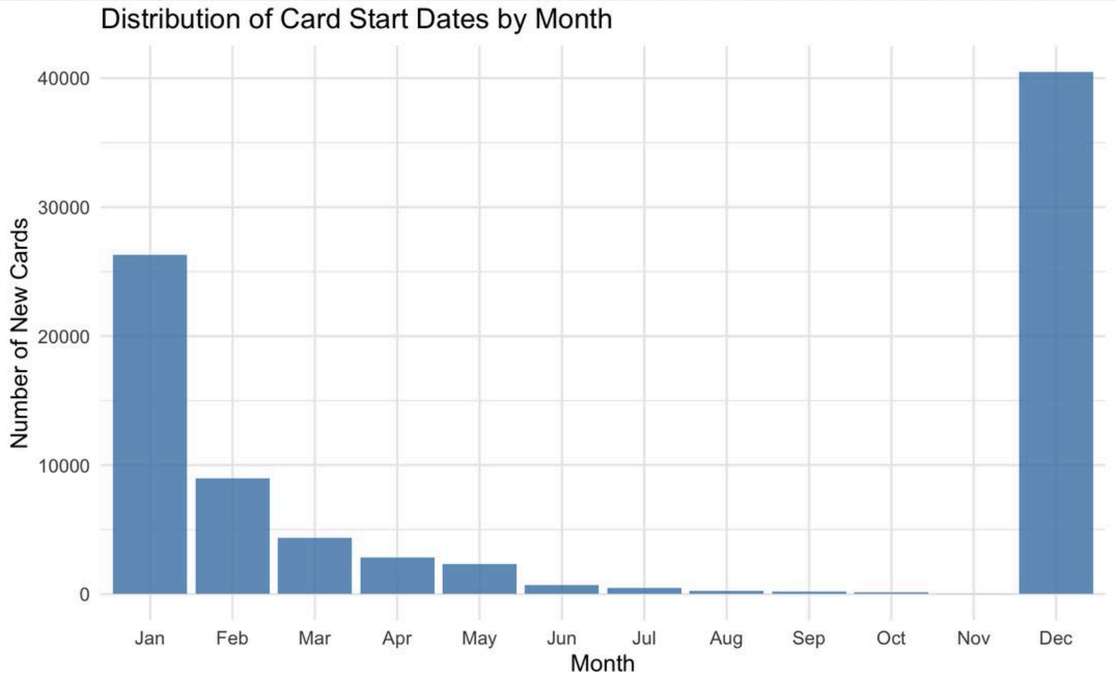


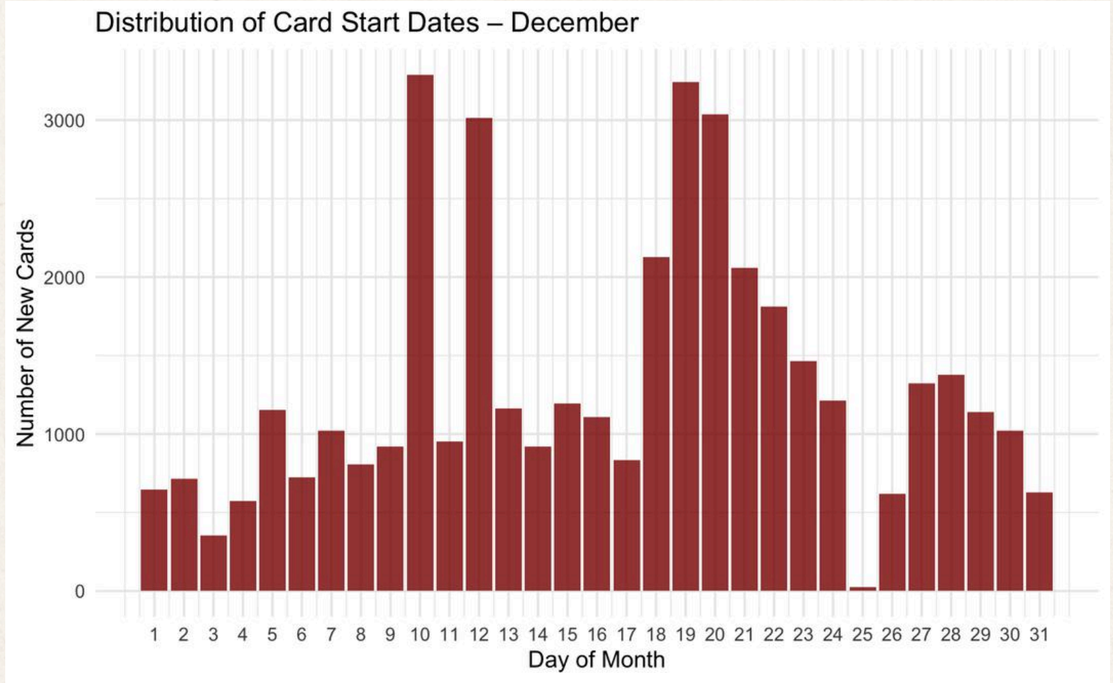
For each customer it was computed the number of total visits made and it was plotted in a logarithmic scale for visualisation.

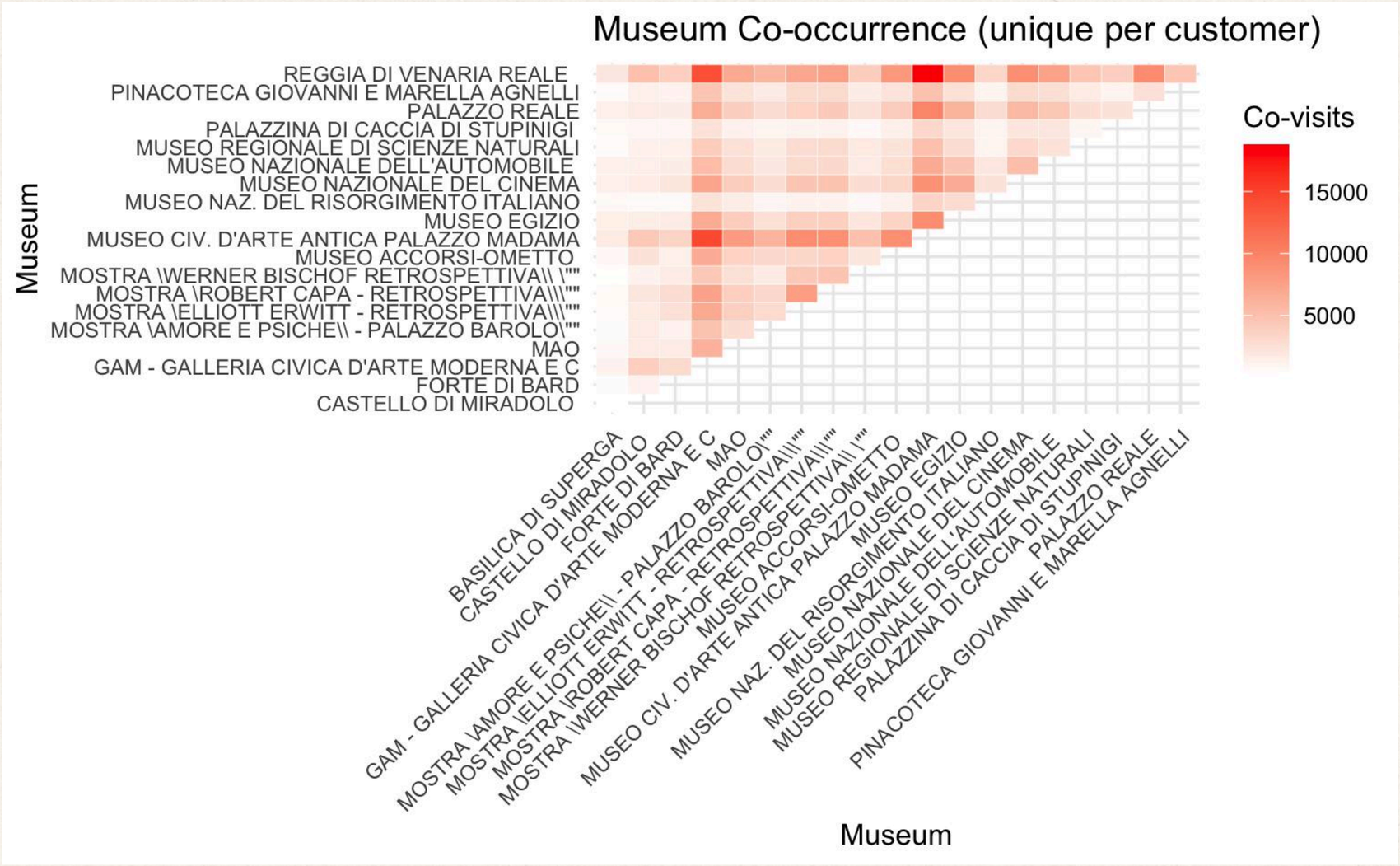
It was considered the variable churn and the results (in the near picture) show that churners are customers that made less visits, in some cases some didn't even visit a single museum.

When considering the possible cause of churning, we can consider the starting date of the museum subscriptions. It's no big surprise that the majority of new subscriptions are made during the winter months of December (especially), January and, to a lesser extent, February.

However, when considering the month of December, most subscriptions are activated or renewed before Christmas.







Taking the 20 most visited museums, a co-occurrence matrix was computed to check whether customers visited two or more museums in the same day.

The occurrences that stand out are the Reggia di Venaria Reale and the GAM, or the Reggia and Palazzo Madama and the Palazzo Madama and GAM.

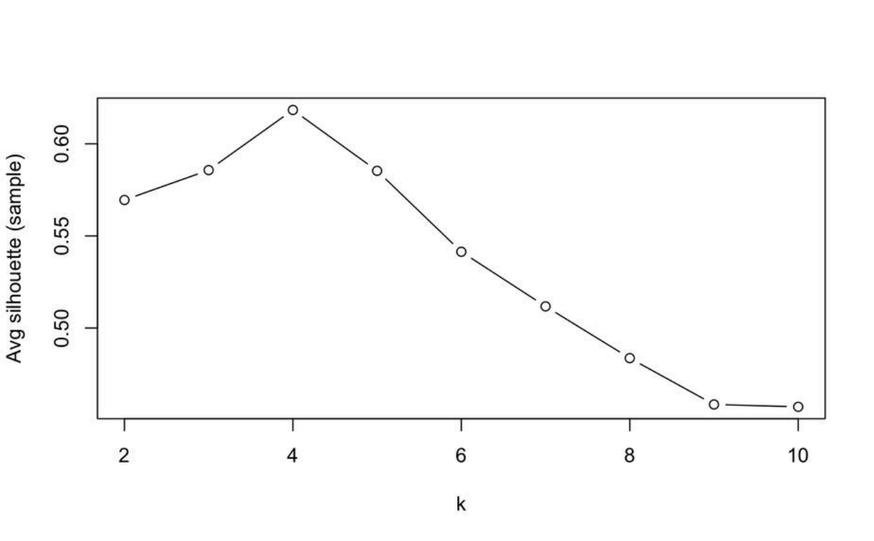
This could be taken into consideration if museums would like to offer combined tickets, which would be cheaper than buying two tickets separately.

Clustering

The data of visits are then clustered in order to be able to check if customers can be clustered according their characteristics such as churn, number of museums visited, tickets price and unique museums visited.

By using the K-means algorithm, a silhouette measure is used to see what is the idea number of cluster to choose for the K-means. From the graph we see that 4 is the best suited number of clusters and each of them has different characteristics.

Cluster 4 has the highest rate of churners (0.49) while cluster 1 and 3 have the highest number of churners (8650 and 9106 respectively), people in this cluster only visited museums 5 times and "spent" 8€ on average. We can see that the second cluster has the fewer churners than any other cluster and is made up of people that visited most museums (almost 22 visits on average in 13 museums) and "spent" 37€ on average.




```
 y
      0     1
1  8650 20686
2   724  5557
3  9106 24010
4  1446  1498
 y
          0         1
1 0.2948596 0.7051404
2 0.1152683 0.8847317
3 0.2749728 0.7250272
4 0.4911685 0.5088315
```

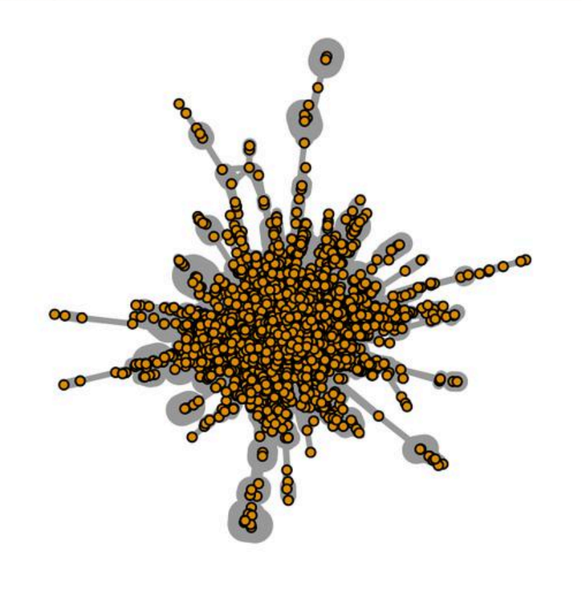
```
$centers
  total_visits unique_museums mean_importo
1     6.021237        4.476582     45.118080
2    21.764846       12.873109     37.080720
3     5.503110        3.973849     28.888332
4     4.754076        3.686141      8.148777
```

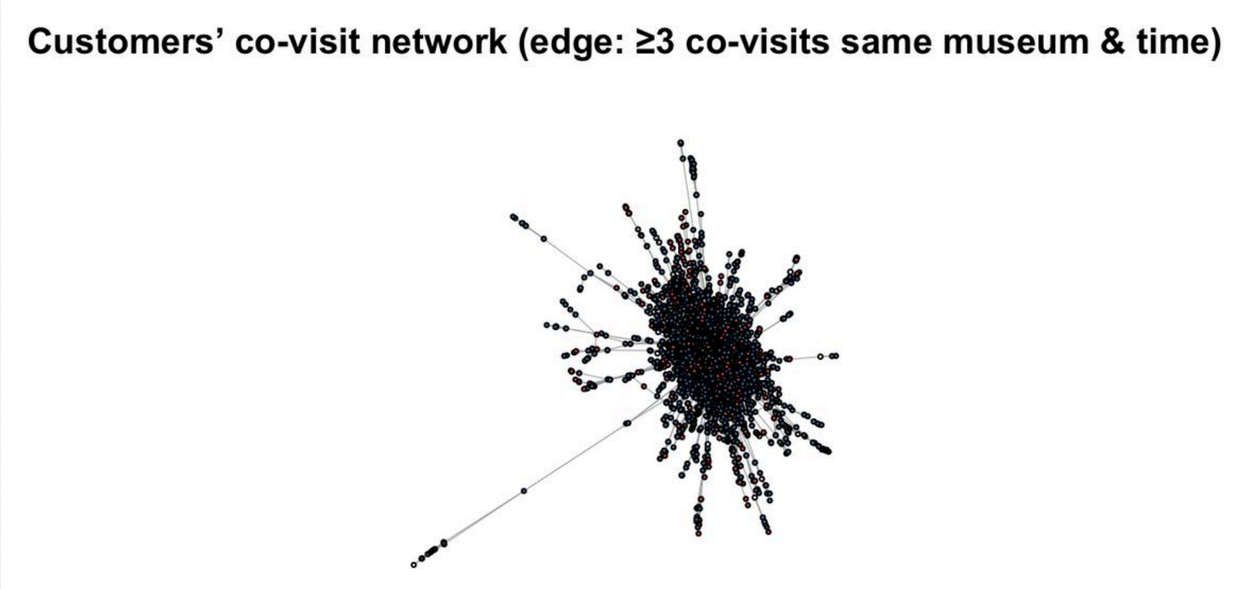
Customers network

From the data at our disposal, it's possible to create a customers' network by considering two or more visits made at the same date and time. By considering unique visits - as the dataset contains duplicates - visitors are matched by the codcliente for each visit. The function creates the left graph, where nodes are the visitors and the edges are the co-visits.

The plot on the right shows visitors with different colour, according to their churn status and the edges are weighted with the pagerank of the nodes.

The pagerank is a measure of the quality of the connection, in which well connected nodes have higher pagerank value.

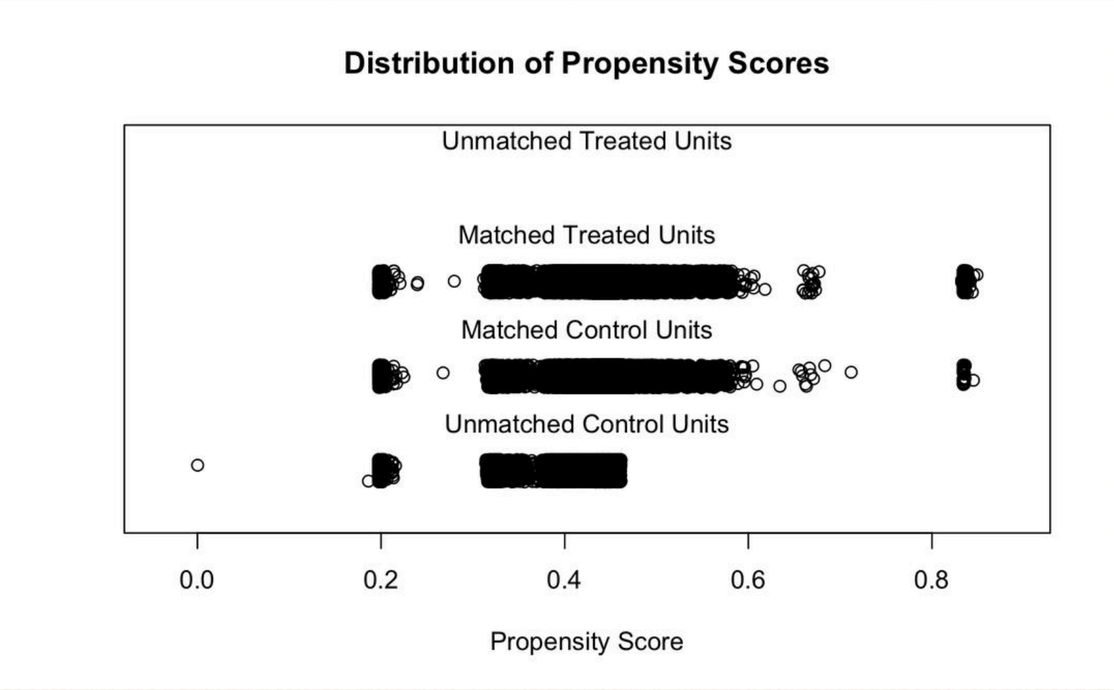




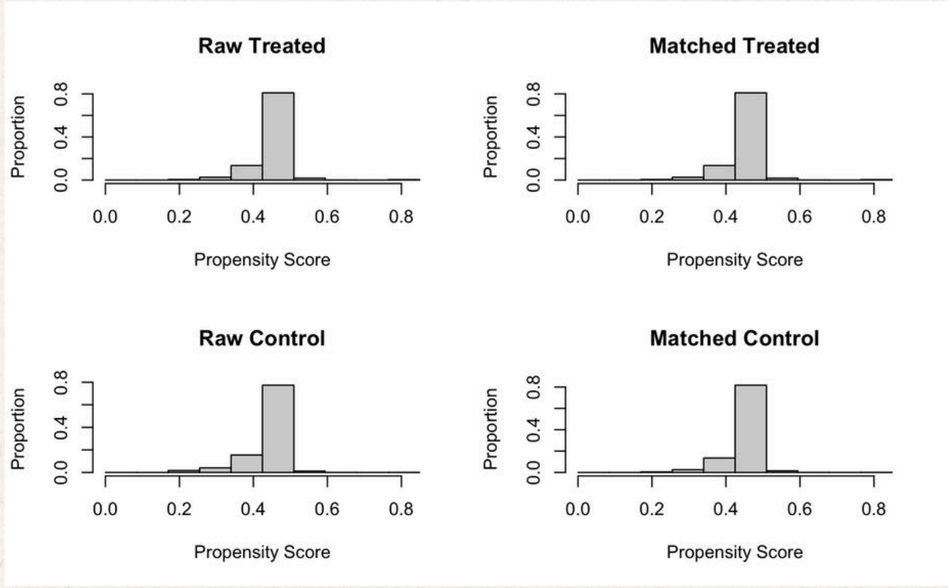
Counterfactual

When considering the impact of gender on churn, it's almost mandatory to match similar units, so to avoid any impact of other variables. So the model tries to match a male with a pretty much similar female from the reference group.

The matching was performed using a subset of variables of price paid for the subscription, the possible discount applied, total visits made, where they got the subscription and year of birth. It was used the nearest method with a logit distance.



From the graph on the left and below we can see that the algorithm is able to obtain quite good results for the matching.



Counterfactual

```
Call:
glm(formula = si2014.y ~ sesso, family = binomial, data = matched_data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.90639    0.01201  75.444  < 2e-16 ***
sessoM      -0.10440    0.01681  -6.209 5.34e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82448  on 67609  degrees of freedom
Residual deviance: 82409  on 67608  degrees of freedom
AIC: 82413

Number of Fisher Scoring iterations: 4
```

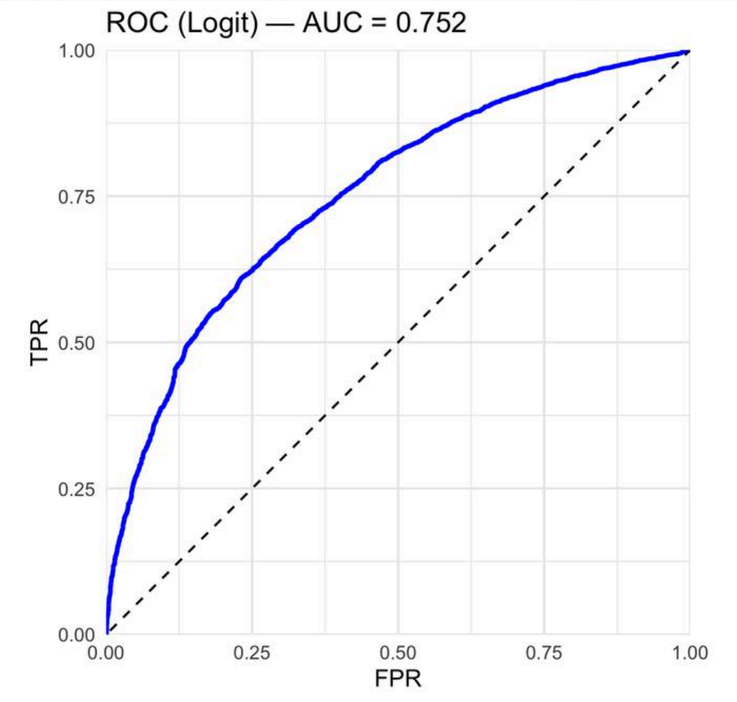
From the matched data, it’s now possible to apply a generalised linear model to estimate the impact of sex on churn. The summary of the model is reported on the left.

The result shows that sex really has an impact on churn, and it’s actually statistically significant. The intercept of the model (0.907) is the log-odds of not churning for the group of female, and the estimate for male sex of -0.104 means that males have a lower log-odds of renewing the subscriptions compared to females.

Prediction models

Before applying any model for predicting churners it's necessary to divide the dataset into 70% train and 30% test sets. After that, the distribution of churners in the train test is checked to be almost equal to that of the original dataset.

The first model applied for prediction is a logit model with covariates such as total visits, sex, discount, date of birth and the network centrality measures like pagerank, degree, strength and closeness.



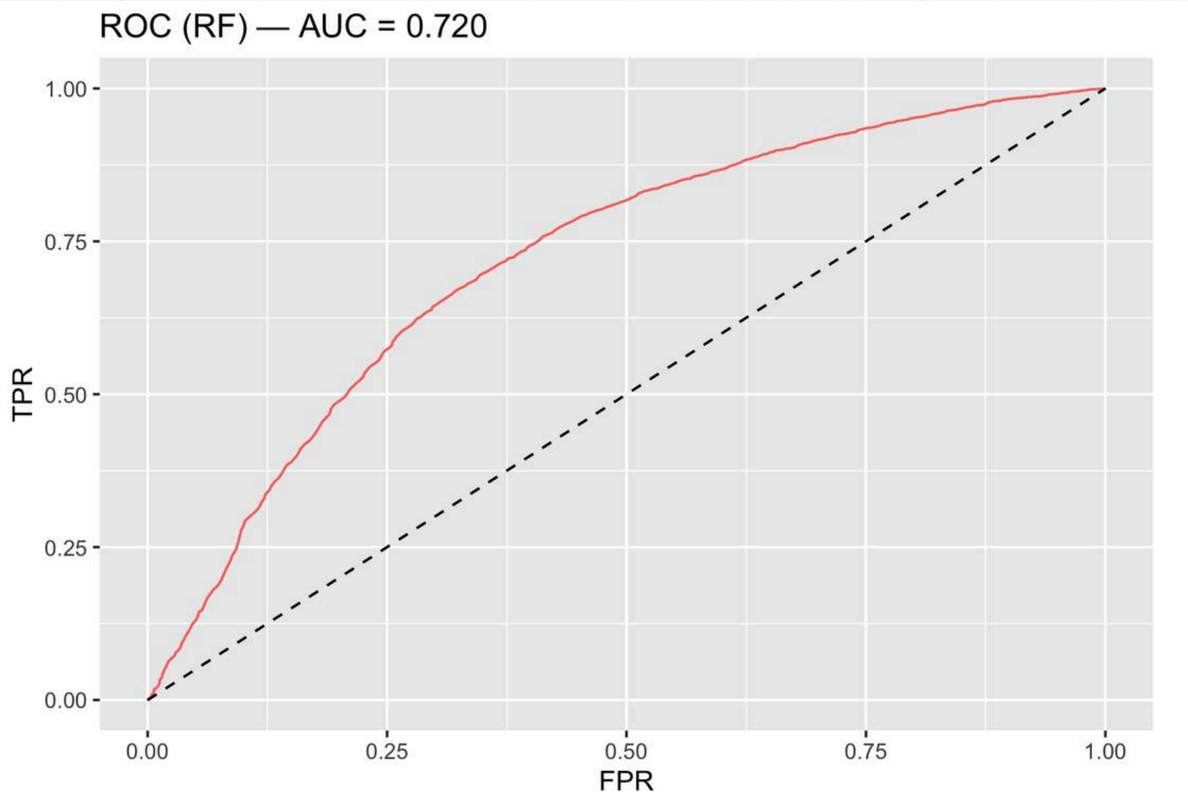
When predicting the new data in the test set, measures such as true positive rate and false positive rate are computed in order to plot the ROC curve of the logit model, and then to compute the AUC.

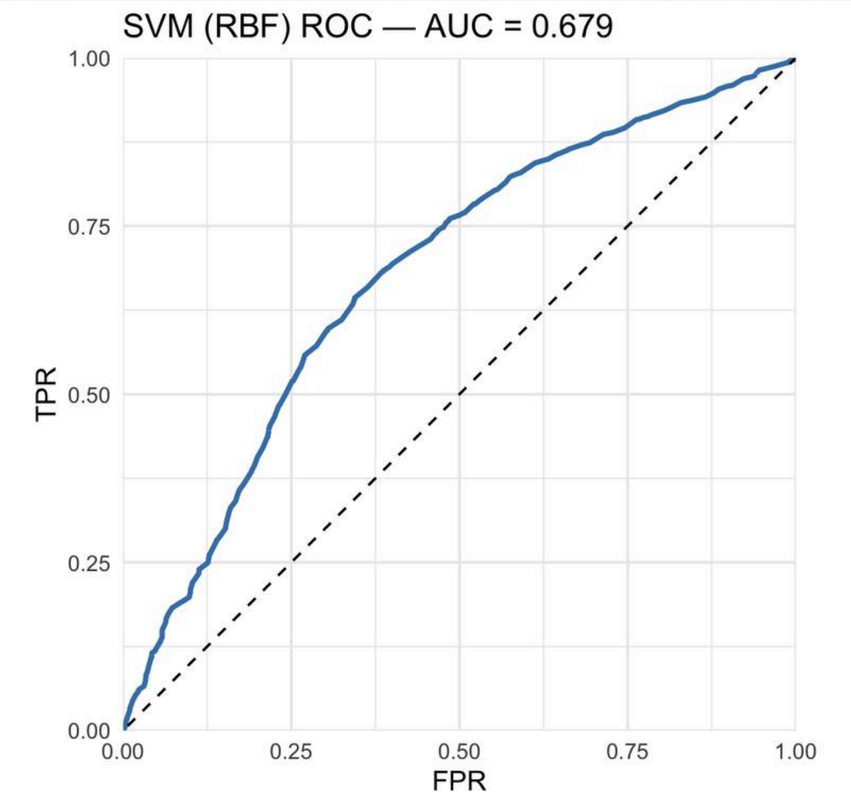
This model reaches an AUC of 0.752 which is a quite good result, meaning that the model is able to predict new data with relative high confidence.

However it would be better to find a model that achieves even better result.

Prediction models

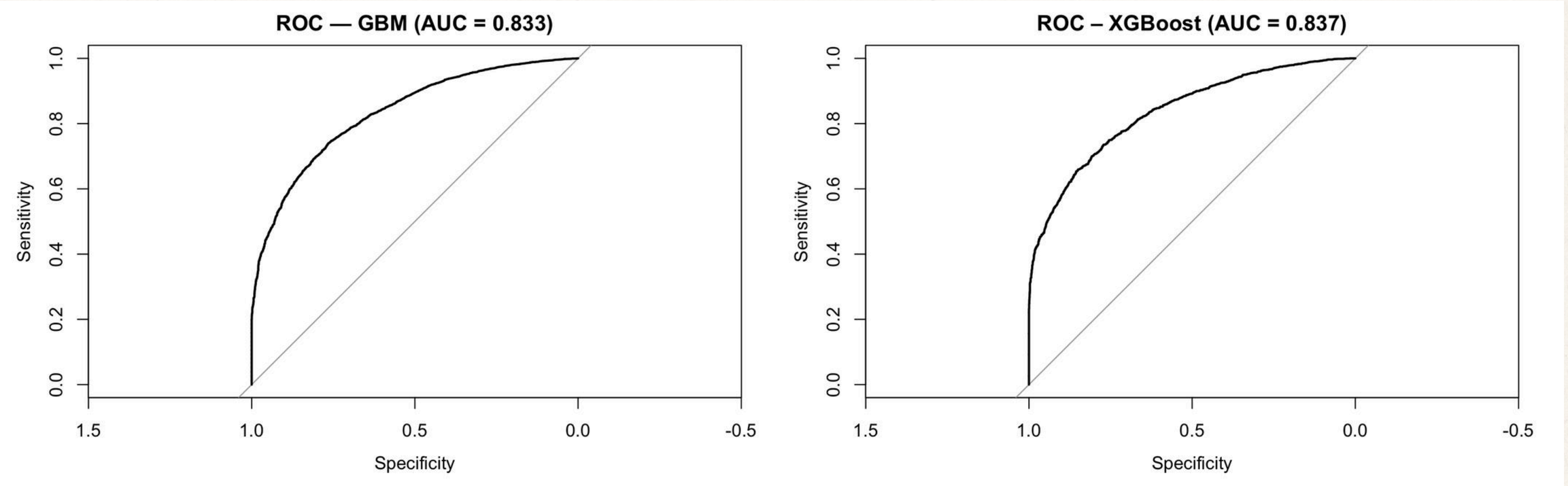
The second and third models applied are a Random Forest with the same covariates as the logit model made up of 500 trees and a SVM with a radial kernel. The. RF model reaches an AUC of 0.72, which is slightly worse to the previous model of 0.752 while the SVM performs poorly, with an AUC of just 0.679.





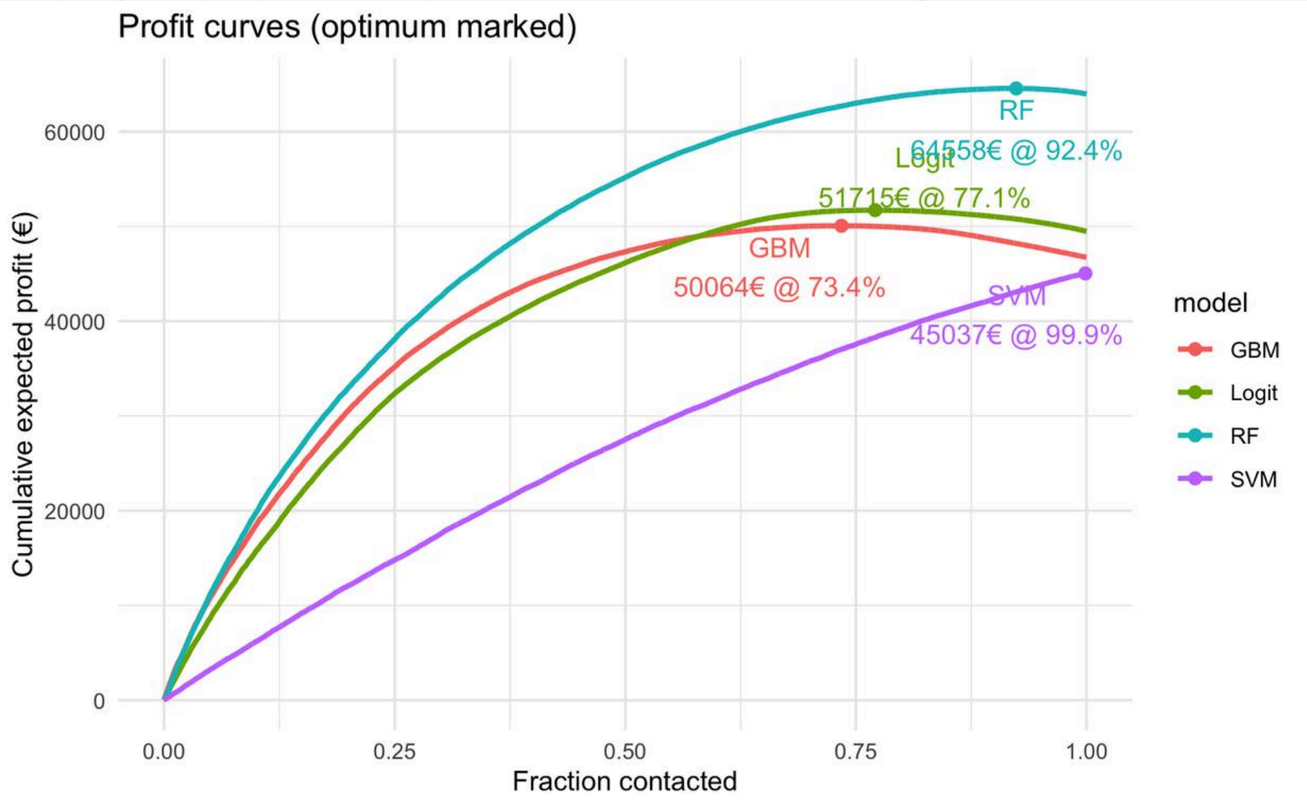
Prediction models

The fourth and fifth models applied are a Gradient Boosting and a XGBoost algorithms. This time the covariates included are much more and include also the date of last visit, date of beginning of the subscription, type of payment and information about the residence. Due to the computational limits, it was necessary to use a sample of the original data. We can see that the results are quite the same (0.833 vs 0.837) so both models are able to correctly predict new data with a high confidence.



Profit curve

After applying different models to the test set, we have to measure how good the models are when taking into account the cost of 2€ for contacting each customer considered as churner by the model and the 10€ discount offered. By contacting customers starting from those with the highest churning probability, the cumulative profits is expected to increase.



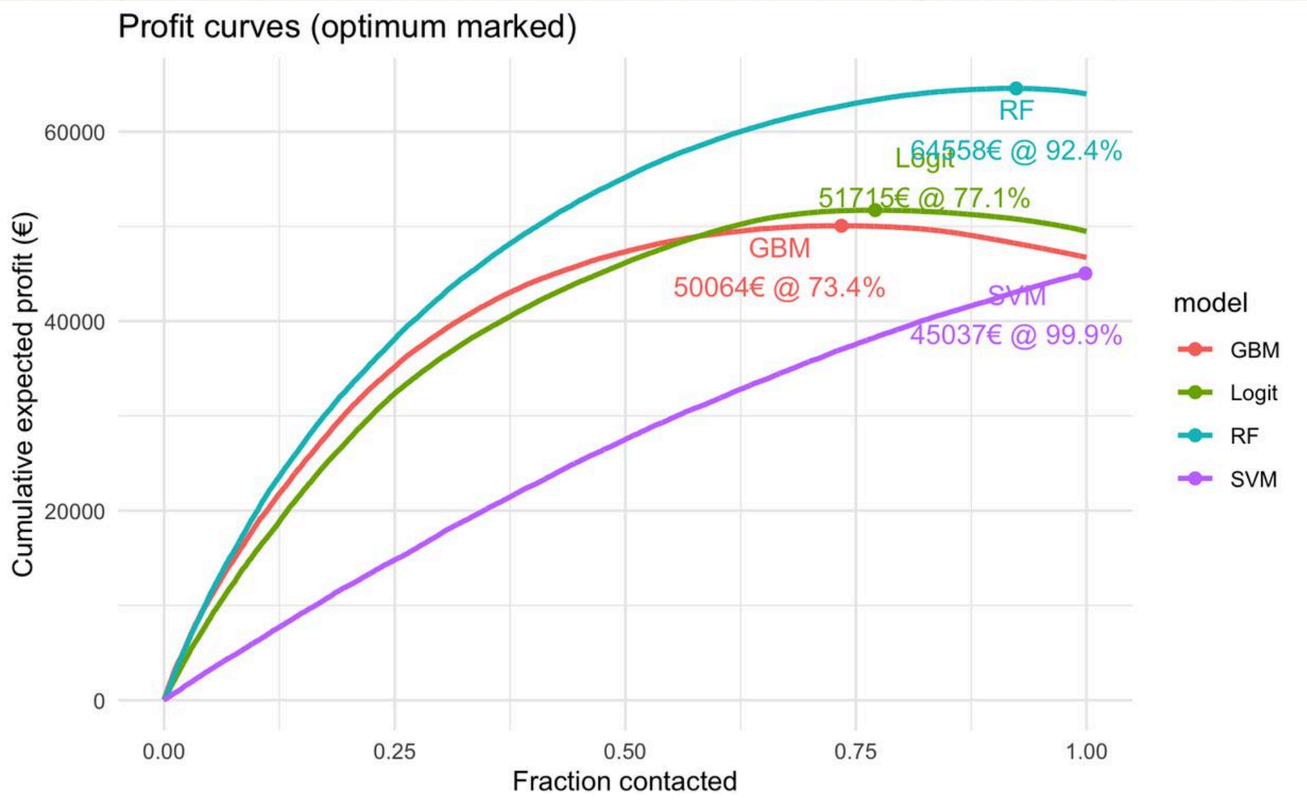
The near figure shows the cumulative profits for the three models of GBM, Logit, RF and SVM because they were applied on the whole test set.

By choosing the GBM and Logit models we can expect to reach the highest profit by contacting 74% and 77% of churners; however, the Logit model is able to better predict true churners and the profit would be circa 52.000€ instead of 50.000€ of the GBM model.

The RF instead is the model that performs best with a profit of just under 65.000€ when reaching 92% of churners.

Profit curve

Such high percentage of customers contacted means that probably the models are not able to discriminate well enough churning and non-churning customers, for example SVM which shows an always increasing curve. Another possibility is that the profit/cost could be profitable even when contacting a big number of customers, like in the case of RF and Logit and GBM, although their ideal fraction is lower.



From these results, we can say that the SVM model underperforms the other models, and while the RF model achieves the highest cumulative profit, we can think that the model is not able to discriminate non-churners.

The other two models, Logit and GBM seem to show a more usual curve, where profits start to decrease when too many customers are contacted. The difference between the two is only about a thousand euros for contacting between 74% and 77% of churners. So in terms of profits, either model will work just fine for maximising profits.