QUESTION 1: Logistic regression and CLV analysis (60 points)

1.1 Data Explanatory Analysis

There are 31,929 observations in total across 15 variables in the dataset. These observations are spread across 2,000 unique households from the period 2012 to 2017. There were no missing values or duplicates in the dataset.

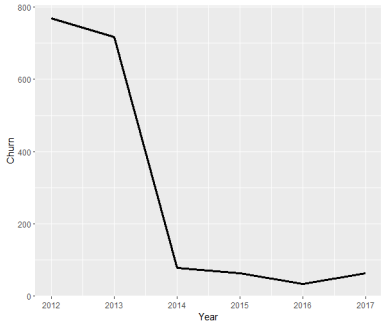

Figure 1 - Evolution of customers lost

From Figure 1, we notice that a substantial amount of the customers terminating their recurring donations did so in 2012 and 2013. From 2013 to 2014, the churn rate drastically decreased, and it is relatively stable in the following years.

Figure 2 reveals the number of unique customers in the dataset, and whether they churned or not. With 1,726 lost customers, a whopping 86.3% of the unique customers obtained in the period 2012-2017, are lost.

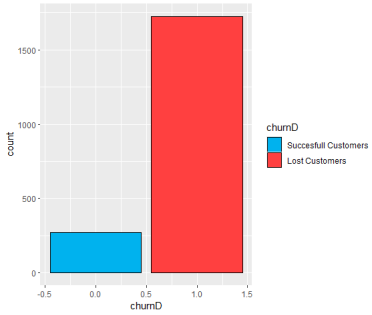

Figure 2 - Unique customers lost

Figure 3 displays how some of the predictor variables are evolving over time. This reveals a fairly even distribution over time for every variable. However, Facebook ad spending seems to be slightly increasing over time. Both TV advertisements and DM´s are often 0, and for TV ad spend, we notice a large abnormal value, but we will say more about that later.

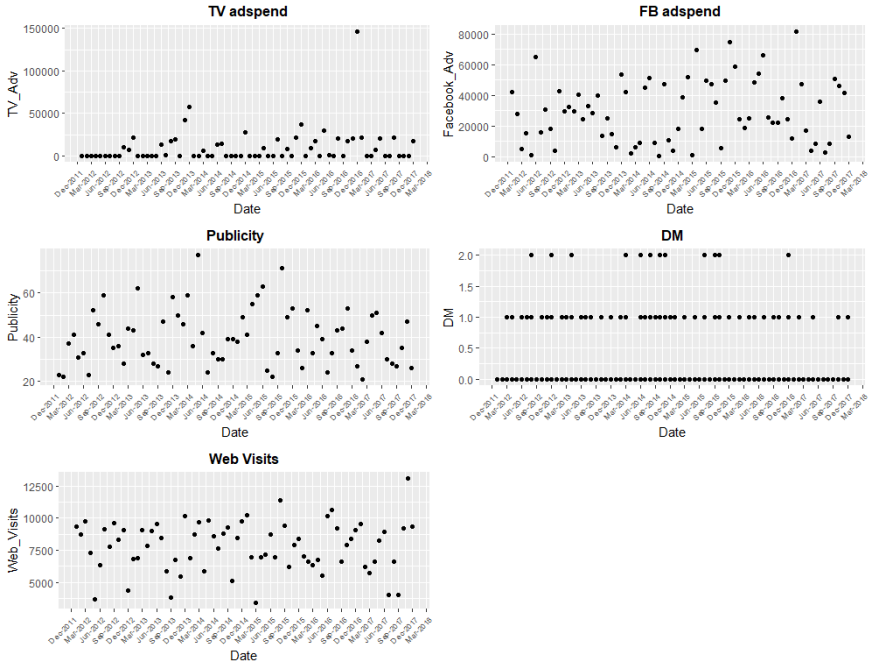

Figure 3 - Evolution of a selection of variables

From Figure 4, we notice that only a few variables are highly correlated, except for RelationshipL and Year (corr = 0.97) and TV_Adv and Month (corr = 0.43), no variables have a stronger correlation than |0.21|. This implies that we don't need to worry about multicollinearity; an issue arises when independent variables that are highly correlated are included in the same regression model. In our model selection, Month and Year are not included, avoiding the possible issue of multicollinearity. It is however interesting to note that costs from TV advertising typically are higher in the later parts of the year.

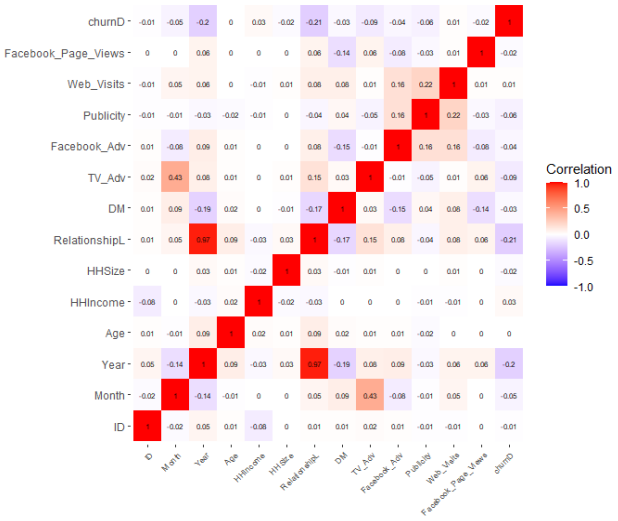

Figure 4 - Correlation heatmap

To identify outliers, we first create a figure displaying the boxplots of each variable. The dots in each corresponding boxplot, that you can see on Figure 5, represent outliers. Using the interquartile range (IQR) method to remove these outliers, the remaining dataset consists of 29,191 observations. Furthermore, removing 2,738 observations lead to excluding 13 households from the dataset. In this process, we removed the previously mentioned abnormally large value for TV Adv.

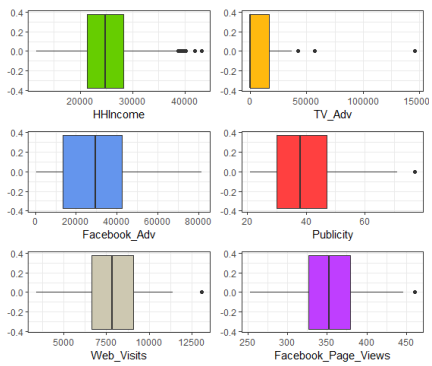

Figure 5 - Outlier detection

Figure 6 reveals the distribution of the total number of churns vs the total number of non-churns in the cleaned dataset. We notice a very big difference between the total number in the two groups. A customer will only churn once, and then "disappear", while non-churning customers appear on the data set month-on-month, year-on-year. After cleaning the data, we are left with 1,679 churned customers. Although this is a large portion of the total number of unique customers (now < 2,000), it is only a fraction of the total number of observations. Only 1,679 of 29,191 observations made an actual churn (churnD = 1), which is equivalent to a proportion of 5,75%. Imbalanced proportions of the dependent variable affect the result must be considered when for instance splitting the data into train- and test sets. Ideally for logistic regression, it would be if the two levels were equally large. We account for these differences by adjusting the threshold, we explain this further down.

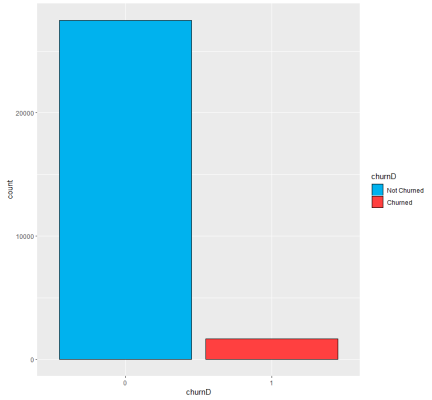

Figure 6 - Total amount of churnD occurring

1.2 Churn Analysis

1.2.a Formulation of hypothesis

In this part, we will highlight the perceived influence of a selection of independent variables (IV´s) on the dependent variable (DV). Furthermore, we will formulate a hypothesis for each IV's and their expected effect (positive, negative, or both) on the probability of a customer churning. The IV´s are: Direct marketing (DM), TV advertising, Publicity, and Facebook advertising.

Table 1 displays the independent variables, the column datatype, the min and max values, and the average value.

Variable	Class	Min-Max	Mean
DM	*Integer*	*0-2*	*0,30*
TV_Adv	*Numeric*	*0-36769,5*	*6902,5*
Facebook_Adv	*Numeric*	*614,4-81416,4*	*29951,4*
Publicity	*Integer*	*21-71*	*39,4*

Table 1 - IVs

DM

We want to highlight that it can be difficult to predict whether the correlation between the DM variable and the response variable will be positive or negative. Indeed, the increase in direct contacts could function as a CRM (customer relationship management) tool and thus reduce the probability of churn. However, the increase in direct contacts could also appear to be spamming and annoying the customer, and thus be associated with higher churn odds.

For instance, Leeflang, Wieringa, Bijmolt and Pauwels explain in their book that Verhoef previously showed that affective commitment can reduce the probability for a customer to churn. Having regular direct contact with someone working for the foundation can be a way for the foundation to create and maintain an emotional connection with the donor, and therefore to reduce the probability of churn (Leeflang et al., 2015). In that case, the coefficient for DM would be negative ($\boldsymbol{H_{A.1.1}}$). 63% of customers give the feeling of not "being helped or appreciated" as a reason for churning. (Pope, 2022)

Furthermore, we learned in lecture 9 of the course that customers appreciate in general when the person contacting them speaks more concretely to them (Le, 2023). According to a study for which 200 service calls were recorded, the customer was more likely to perceive the employee helpful and attentive when the employee adopted a concrete language (Packard and Berger, 2021). It will therefore be interesting to learn more about the way the employee of the Arima Foundation speaks with the donors of the recurring donation schemes, to see how the number of direct contacts can influence the churning ratio.

$\boldsymbol{H_{0.1}}$: *The coefficient for DM is equal to zero, meaning it does not influence the response variable.*

$\boldsymbol{H_{A.1.1}}$: *The coefficient for DM is negative, meaning increasing direct messaging are associated with lowering the log odds for a customer churning.*

$\boldsymbol{H_{A.1.2}}$: *The coefficient for DM is positive, meaning increasing direct messaging are associated with higher log odds for a customer churning.*

TV_Adv

It seems counterintuitive that an increase in ad spending would be associated with higher churn probabilities. The therefor hypothesized negative correlation between television advertising and the response variable indicates that as television advertising expenditures increase, the logarithmic odds of a donor changing his/her mind decreases. In other words, the more the firm spends on television advertising, the less likely it is that donors will leave the recurring donation schemes of Arima Foundation. Television advertising also allows to reach a larger consumer target because it increases visibility. However, several points are to be considered like the quality of the ad but also it is important to notice that TV habits have changed nowadays, and TV is less and less watched since the pandemic as explained by Falon Fatemi in the article of Forbes.

In the Book Database Marketing Analyzing and Managing Customers, Blattberg, Byung-Do and Neslin highlight another interesting point. They point out that Sheehan and Doherty showed that company do not always coordinate the content of their communication between the different means of communication, especially between online and printed. This can lead to a lack of consistency and a blurred understanding of the message for the customer, who is more likely to churn (Blattberg et al., 2009). More than the advertising expenditure spent by the Arima Foundation for on the TV channel, it would be interesting to see if the content of the TV ad is aligned with the content on social media (Facebook) and on the printed communication (publicity). This way, it would be easier to predict the correlation between TV_Adv and the log odds for churning, even if we feel that it is negative ($\boldsymbol{H_{A.2}}$).

$\boldsymbol{H_{0.2}}$: *The coefficient for TV_Adv is equal to zero, meaning it does not influence the response variable.*

$\boldsymbol{H_{A.2}}$: *The coefficient for TV_Adv is not equal to zero. TV advertising and the response variable are negatively correlated, meaning and increase in TV_Adv spend is associated with lower log odds for a customer churning.*

Facebook_Adv

By using Facebook advertising, the Arima Foundation will increase its notoriety and visibility. Hence, we deem it likely that there is a negative correlation between Facebook ad-spend and higher churn probabilities. Moreover, by choosing to advertise on social networks such as Facebook, the brand will reach a different target audience than with other means of communication. But it is always important to consider the quality of a good advertisement. However, assuming that the advertising is effective, we can be fairly confident in our hypothesis regarding the direction of the correlation between ad spending and churn probability.

It is worth mentioning a new element to consider about Facebook ads. Since 2020, a growing number of large companies, like Unilever, have joined the Facebook ad boycott because of the handling of hate speech and misinformation of the social media (Fung, 2020). This trend most likely also reflects a shift in the way Facebook users see the social network. Some are becoming more wary and perhaps less receptive to ads on this social network. Although we assume that

the correlation between the "Facebook_Adv" and the log odds of churning is negative ($\boldsymbol{H_{A.3}}$), we can nuance this relationship with this recent evolution of the perception of Facebook.

$\boldsymbol{H_{0.3}}$: *The coefficient for Facebook_Adv is equal to zero, and hence, does not influence the response variable.*

$\boldsymbol{H_{A.3}}$: *The coefficient for Facebook_Adv is not equal to zero. Facebook_Adv is negatively correlated the response variable, meaning higher observations of Facebook_Adv´s are associated with lower log odds of customer churn.*

Publicity

As for DM, it can be quite difficult to predict whether the correlation between the Publicity variable and the log odds for a customer churning will be positive ($\boldsymbol{H_{A.4.2}}$) or negative ($\boldsymbol{H_{A.4.1}}$). Lke we have seen previously, increasing spending on advertising and its means of communication can have positive effects on brand awareness and visibility. However, it is important to consider that an advertisement must be of good quality to truly have the desired impact and attract the attention of its potential target customer. A bad advertisement can also change the image of the brand and therefore be perceived negatively by a potential future customer. Indeed, the content of the publicity is essential, especially for a nonprofit organization like the Arima Foundation, which truly needs a clear message and a good reputation to keep its regular donors and even attract new ones. Furthermore, the "Publicity" variable represents the number of times the nonprofit had been mentioned on newspapers and similar outlets. We can consider here that it is not necessarily a matter of quantity but rather of quality of the publicity that will encourage the donor to remain faithful to the foundation and to make regular donations.

As we mentioned above, it would be interesting to know whether the content of the printed communication is aligned with the content posted by the Arima Foundation online or broadcast on TV (Blatteberg et al., 2009). This way, we could maybe have a clearer hypothesis on the negativity or positivity of the correlation between "Publicity" and the log odds of churning for a donor.

$\boldsymbol{H_{0.4}}$: *The coefficient of Publicity is equal to zero, meaning it does not influence the response variable.*

$\boldsymbol{H_{A.4.1}}$: *The coefficient of Publicity is negative. Positive publicity will have a negative correlation with the response variable, meaning increased number of times the nonprofit organization has been mentioned in newspapers or similar outlets is associated with lowering the log odds for a customer churning.*

$\boldsymbol{H_{A.4.2}}$: *The coefficient of Publicity is positive. Negative publicity will have a positive correlation with the response variable, meaning increased number of times the nonprofit organization has been mentioned in newspapers or similar outlets are associated with increasing the log odds for a customer churning.*

1.2.b Model Estimation

In this task we create a binomial logistic regression model by using the whole dataset (no splitting between train and test data), with the independent, and dependent variables discussed in 2a). Having a second look at the correlation heatmap for the selected variables reveals no sign of multicollinearity.

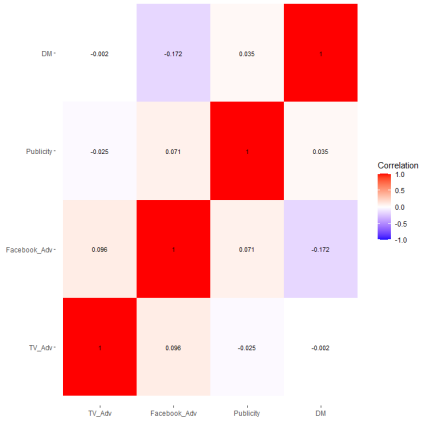

Figure 7 - Heatmap of variable subset

To create a binomial logistic regression model, we use the glm function in R with the relevant response and predictors. We fit this logit model on the entire dataset. When running the summ function from the jtools package in R on the logit model, we notice that the p-value of all estimated coefficients is approximately 0. Since the p-value is lower than 0.01%, we reject the null hypothesis ($H_{0.X}$: coefficients = 0). Hence the coefficient estimates appear statistically significantly different from 0 at any reasonable significance level. Thus, all predictors could potentially be viewed as promising. We do however not confirm our alternative hypothesis ($H_{A.X}$: coefficients $\neq/\leq 0$), we merely reject the null hypothesis. Table 2 below displays the coefficients estimates from this logit model, the 95 % confidence interval to the coefficients, the odds ratio, and the min- and max values of the predictors.

Predictors	**Coefficient Est.**	**95 % Confidence Interval**		**Odds Ratio (OR)**
DM	*-2.802e-01*	*-3.978e-01*	*-1.643e-01*	*0.75565*
TV_Adv	*-7.354e-05*	*-8.175e-05*	*-6.561e-05*	*0.99992*
Facebook_Adv	*-1.051e-05*	*-1.346e-05*	*-7.577e-06*	*0.99998*
Publicity	*-2.287e-02*	*-2.759e-02*	*-1.822e-02*	*0.97738*

Table 2 – Estimated coefficients from model 2b

From Table 2, we notice that all predictors have an odds ratio of less than 1, suggesting negative effects for all predictors. Meaning, the predictors are associated with lower odds of the outcome (churnD = 1). Subsequently, higher expenditures on TV and Facebook ads, more mentions in newspapers, and a higher number of the nonprofit's direct contacts, all are associated with a lower probability for the donors to churn on their payments. However, there are some big differences in the odds ratio (OR) between some of the variables. For the "TV_Adv" and "Facebook", the OR is almost equal to 1. In contrast, the predictor "DM" has an OR of 0.75565 which indicates that increasing the number of direct contacts with the donor by one in that month, is associated with a reduction of the odds of the donor churning by approximately 24.4%. Hence an effective method to make customers stay with the Arima Foundation is to increase their number of contacts with the donors. If we wanted to achieve a similar reduction of the odds of the donor to churn by approximately 24.4% due to increasing the nonprofit's advertising expenditure on Facebook, this would be equivalent to an increase of € 23271,49 in Facebook costs for a given month.

When evaluating the model fit, we measure the Pseudo-R Square (McFadden) which was equal to 0.04803. A better model fit is indicated by a higher McFadden Pseudo-R Square value, as it represents the proportion of the explained variation in the response variable (McFadden, 1974). We also use the LR statistics which had a Chi-square value (X^2) equal to 617.09 and a p-value less than 0.05. Thus, the LR statistics imply that the model is meaningful or that at least one of the regression coefficients is significantly different from 0. This is in alignment with our conclusion in the section above.

Furthermore, to evaluate the model fit, we also use the Akaike Information Criterion (AIC) which was equal to 12241.75. We obtained this by running the AIC() function on the model. In addition, we measure the Bayesian Information Criterion (BIC) which was equal to 12283.16. A smaller AIC and BIC values, suggest a better model fit (Burnham & Anderson, 2004). These metrics will be some of the tools we use to compare the performance between different models in part 1.2c.

1.2.c Model Evaluation

We create three new models, one model where we remove the two variables "DM" and "Publicity" from the original model. In the second and third models, we remove only one variable, "Publicity" and "DM" respectively. We then compare the performance of these three models with each other and finally with the original model from 2.b.

For model validation, we have decided to use a 10-fold Cross-Validation of the entire dataset. The observations were randomly divided into ten non-overlapping groups, where each of these ten groups acts as a validation set and the remainder as a training set (James et al., 2013). We opted for k = 10, as this value is widely adopted in practice because it has been empirically demonstrated to yield test error estimates that suffer neither from excessively high bias nor from very high variance (James et al., 2013). The evaluation metrics of the model fit and the performance of the models, are then estimated by averaging the ten resulting estimates.

With a distribution of total number of not churns (N = 27512) which is significantly higher than the distribution of the total number of churns (N = 1679), then without using any model, we could obtain a total hit rate of 94.25 % (=27512/29191), by assigning everyone to not churn. This total hit rate would be misleading since the organization is interested in predicting donors that churns.

Furthermore, as mentioned previously, the levels of the dependent binomial variable "churnD" are heavily imbalanced. The imbalance of the response variable "churnD" would lead to a narrow range with low predicted probabilities for the donor to churn, and one could experience the model to never predict the donor to churn. When using the range function, we see that the predicted probability often has a maximum level of approximately 13-14 %. With the usage of common threshold value for a binomial logistic regression model, like 0.50, we would therefore never predict the donor to churn. For this specific reason, we adjust the threshold to achieve a

better model validation. When deciding the optimal threshold for each model, we calculate the Area Under the Curve (AUC). Then, we select the threshold corresponding to the point on the ROC curve closest to the top-left corner. We calculate the optimal threshold in 10-fold Cross-Validation for each model instead of setting a fixed threshold for all models to allow for a fair comparison of the models given their optimal specifications.

One tool we used for comparing the models was the likelihood ratio test via the lrtest function in R. In these comparisons, we compared the reduced models with the original full model. We also compared the model consisting of two predictors with the models of three predictors. The null hypothesis of the likelihood ratio test is that the full model is not better than the restricted one. If the test is significant, i.e., with a p-value lower than 5 %, then we can reject the null hypotheses, according to the lecture 3 (Le, 2023). In all our cases, the p-values were approximately equal to 0. Hence, we reject null hypothesis saying that the full model is not better than the restricted one. This also applied when we removed only one variable, either "DM" or "Publicity" from the original model, which provides evidence that the original model with four predictors in part 2.b. performs better when these predictors are included.

Another set of tools we used was to compare the AIC and BIC for the relevant models. The results are displayed below:

IVs	AIC	BIC
TV_Adv, Facebook_Adv	*12371,12*	*12395,97*
TV_Adv, Facebook_Adv, DM	*12337,28*	*12370,4*
TV_Adv, Facebook_Adv, Publicity	*12262,66*	*12295,79*
TV_Adv, Facebook_Adv, Publicity, DM	*12241,75*	*12283,16*

Table 3 - AIC and BIC

From Table 3, we see that the model with four predictors has the best fit, in terms of AIC and BIC.

In addition to the previous methods, we also computed the performance of several other evaluation metrics, including AIC, AUC, Accuracy, Precision, Recall and F1-Score. Table 4 below displays the best model for each of these evaluation metrics.

Metric	Value	IVs
AIC	*12242*	*TV_Adv, Facebook_Adv, Publicity, DM*
AUC	*0,672*	*TV_Adv, Facebook_Adv, Publicity, DM*
Accuracy	*0,731*	*TV_Adv, Facebook_Adv, DM*
Precision	*0,973*	*TV_Adv, Facebook_Adv*
Recall	*0,741*	*TV_Adv, Facebook_Adv, DM*
F1-Score	*0,839*	*TV_Adv, Facebook_Adv, DM*

Table 4 - Evaluation metrics

We notice that the model with four predictors does not necessarily perform best on all metrics. This is however due to the imbalance in the data (the majority of customers who did *not* churned). Predicting 0 for every customer would for instance give a high accuracy and precision as mentioned previously but would not provide any meaningful insight. As 0 is considered as the "positive" outcome, high precision would not mean high model performance regarding the

ability to correctly classify when churnD = 1. The ROC-AUC score provides a more balanced picture of model performance with such an imbalance in the outcome variable in the data. Hence, AUC and AIC are metrics of greater importance for our model evaluation. We thereby conclude with the model from 2.b. being the superior model, disregarding the doubts regarding the importance of "DM" and "Publicity".

1.2.d Optimal Model

To explore what variables might work as good predictors, we divided the original data frame into two subsets, one consisting of observations where churnD = 1, and the other consisting of observations where churnD = 0. Figure 8 displays the distribution of variables in each subset. The variables are considered promising if there is a significant difference in the distributions. The most promising appears to be RelationshipL and advertisement expenditure. Further on, these are used to expand the complexity of the variables.

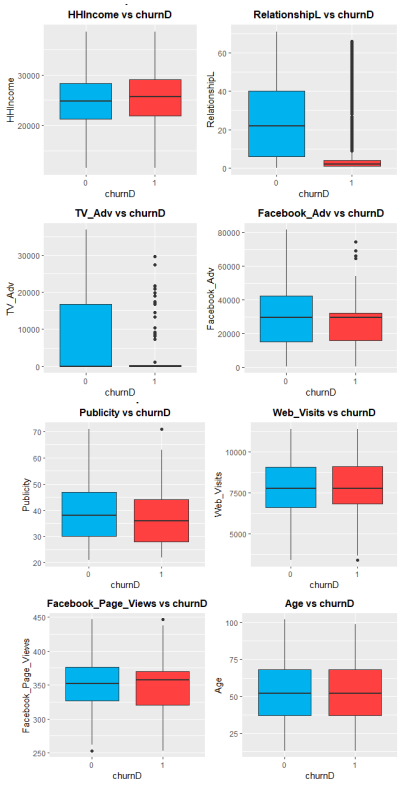

Figure 8 - IV vs DV

Feature Engineering

To improve model performance, we decided to do some feature engineering (changing and creating some new IV´s). The following features was created: *NumberofDonations, ADV_per_visit, RelationshipL_Squared, Age_Group* and *Income_Group.*

NumberofDonations

We found the number of donations per unique customer an interesting feature to include in our model. As previously discussed, the Churn-rate was particularly high in the years 2012 and 2013. This might indicate that customers with fewer donations might be more likely to churn, whilst customer loyalty and perseverance might rise as the number of donations increases. Hence, we expect the variable NumberofDonations and the probability for churn to be negatively correlated, meaning an increase in donations is associated with lower churn probabilities. To calculate this additional variable, we summarized the number of observations per unique ID (representing each customer's donation).

ADV_per_visit

Another interesting aspect is the total cost of advertisement per website visit, effectively representing the *cost-per-click*. This might be able to catch the quality of the advertisement campaigns, meaning a higher cost-per-click would indicate a less effective campaign. Hence, we expect the probability of churn to be lower when cost-per-click is low, meaning a positive correlation with the response variable. The variable is calculated as:

$$ADV\_per\_visit = \frac{FB\_Advertisement + TV\_Adv}{Web\_Visits}.$$

RelationshipL_Squared

RelationshipL is in effect equivalent to the NumberofDonations. However, squaring the variable might reveal a possible non-linear relationship with the response variable. As we will see in the correlation heatmap, the correlation between RelationshipL_Squared and NumberofDonations is not above the acceptable threshold regarding multicollinearity.

Age_Group

Instead of including age as a numerical variable, we wanted to see if there were any systemic differences between age groups. Hence, a factor variable was created, including four age groups: *0-24, 25-44, 45-64,* and *65+*. The expectations regarding age on the possibility of churn are somewhat unclear. Even though older groups typically are wealthier and hence more likely to subscribe to monthly payments, young people appear more idealistic and hence might be more motivated to donate money for a good cause. This is however merely speculation.

Income_Group

As for age groups, we wanted to showcase potential differences between income groups. HHIncome, however, represents the total income of the household, whilst we wanted to look at individual income. Hence, the closest approximation was to divide HHIncome by HHSize. We are aware this is a rough approximation, as there might be large differences within each household that are now being zeroed out. However, three groups of income were constructed: *0-10k, 10k-20k,* and *20k+*. We expect to see lower churn probabilities associated with higher-income groups, as fixed recurring payments would have a relatively smaller impact on their personal economy.

Model selection

For the model selection, we first had to split the data into train and test sets. The train set consist of 80 % of the observations, and the test-set of the remaining 20 %. Stratify samling was used to make sure the disbalance of the response variable (churnD) would not affect model evaluation. Furthermore, a k-fold cross validation with 10-folds was used to validate model performance on unseen data whilst using the train data.

For variable selection, we decided to fit all variables (excluding some that now were baked into interaction terms) to a logistic regression model, whilst using the shrinkage method called LASSO. The selected variables can be seen in Figure 9, along with their correlation. [1] The LASSO method penalizes coefficients, and the *unnecessary* predictors (predictors that poorly predict the response variable), are shrunk to zero. However, the LASSO method did not redeem any predictors to be unnecessary. Hence, we fed all the variables to the logistic regression model and fitted it to the training set with 10-fold cross-validation (later referred to as *the first model*). Evaluation of the model summary revealed only *DM, Publicity, Web_Visits, Facebook_Page_Views, NumberofDonations, ADV_per_visit,* and *RelationshipL_Squared* to be statistically significantly different from zero on >= 95% level, whilst *Male* being the last statistically significant from zero coefficient, on a 90% level. Thereby, we fitted a new model, only using the variables that in the previous model appeared statistically different from zero on a >= 90% level (later referred to as *the second/latter model*).

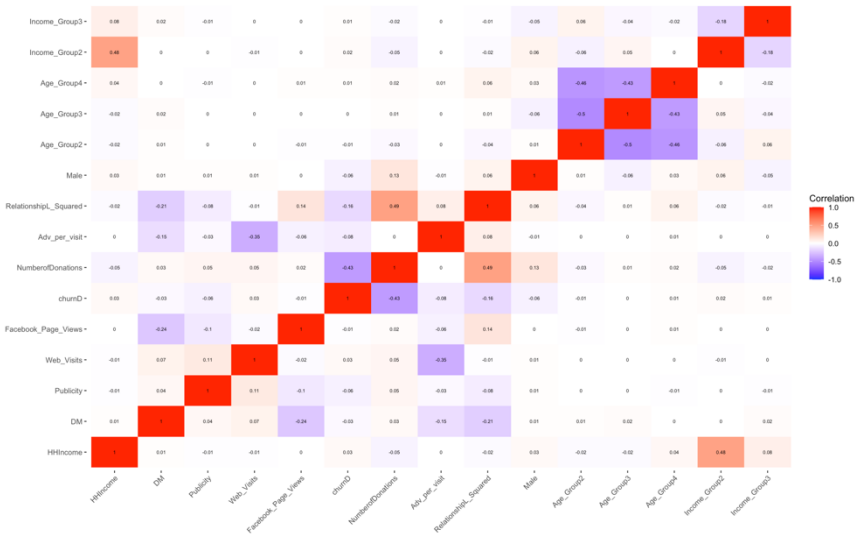

Figure 9 - Correlation heatmap

Comparing the two models, the latter had a slightly lower AIC score *(v1: 5157.1 > v2: 5147.1)*. To further evaluate the model performance, we calculated the area under the curve (ROC-AUC) for the fitted data. The second model gave a marginal increase in the ROC-AUC score *(v1: 0.9585 < v2: 0.9586)*, and an optimal threshold of *0.091* compared to the first model's *0.0884.* Finally, evaluating the two models on completely unseen data (test-data), we see that the second model has a hit-rate (accuracy) of 88.06%, whilst the first model has an accuracy of 87.81%. Meaning, the second model is better at making correct predictions on unseen data. As "positive" class is set to be 0, the specificity reveals how many churns we correctly predict. The first model appears marginally better than the second in this regard *(v1: 0.8657 > v2: 0.8627)* but considering the overall evaluation of model performance on the test and train data, the second model is selected as the preferred.

Fitting the Optimal Model to the entire dataset

Having settled on an optimal model, we fit the model to the entire dataset to be able to do a fair evaluation against the model from 2.b. When using the entire dataset, *Male* no longer is

[1] Note that the Gender column is renamed to Male, taking the value 1 if the customer is a male and 0 if the customer is a female. Meaning the default is female. For income groups Income_Group1 is default, and for age groups Age_Group1 is default.

statistically significant from zero on a 90% level *(p = 0.1530)* and will hence be disregarded in the interpretations of model coefficients. The final model is as follows:

Predictors	Coefficient Est.	P-value	Odds Ratio (OR)
Intercept	*3,2884*	*< 2e-16*	*26,7992*
DM	*0,7630*	*< 2e-16*	*2,1446*
Publicity	*-0,0198*	*< 7,43e-10*	*0,9804*
Web_Visits	*-0,0001*	*0,0145*	*0,9999*
Facebook_Page_Views	*-0,0019*	*0,0113*	*0,9981*
NumberofDonations	*-0,2568*	*< 2e-16*	*0,7735*
Adv_per_visit	*-0,2398*	*< 2e-16*	*0,7868*
RelationshipL_Squared	*0,0031*	*< 2e-16*	*1,0031*
Male	*-0,0900*	*0,1530*	*0,9140*

Table 3 - Estimated coefficients [2]

The intercept being positive indicates that doing no measure is associated with a higher likelihood of churn. The positive relationship between DM and the response variable could be interpreted as direct messaging having more of an annoying spam effect than intended. Contrarily, it could just be that the non-profit usually sends out direct messages to customers they deem likely to churn, without successfully avoiding the churn. Web_Visits, Facebook_Page_Views, NumberofDonations and Adv_per_visit are all associated with lower likelihoods of churn, with NumberofDonations and Adv_per_visits appearing to be the most impactful. Relationship_Squared having an increasing effect on the response variable whilst NumberofDonations having a decreasing effect, it could reveal a possible nonlinear effect regarding the length of customer relationship (and by definition number of donations). It might be that after a significant amount of time has passed (captured by the squared variable getting really big), donors churn likelihood starts to increase again. However, it should be taken into consideration that these interpretations are based on non-causal relationships, meaning they are purely hypothesized by correlations.

Comparing the Optimal Model with the Less Complex Model (from 2.b.)

A direct comparison of key aspects of model fit and performance for each model can be seen in Table 6. We notice the Optimal Model outperforms the model from 2.b. in *every way.*

Metric	Model from 2b)	Optimal Model
AIC	*12 242*	*6 520*
McFadden (Pseudo-R^2)	*4,8%*	*49,4%*
AUC	*0,6723*	*0,9572*
Optimal threshold	*0,0787*	*0,0978*
Hit-Rate	*71,35%*	*88,52%*
Churns correctly predicted	*58,55%*	*88,39%*

Table 6 - Optimal Model VS Model from 2b)

[2] Note that all values are rounded to four decimals.

Lift-, Gain- and ROC-curves are great tools for evaluating classification model performance. The Lift curve focuses mainly on relative improvement compared to a random model, while the Gain curve is emphasizing the cumulative percentage of positive cases the model captures, as a function of the amount of the population targeted. The ROC-curve illustrates the tradeoff between true positive rate (y-axis) and false positive rate (x-axis) for different thresholds, as we learned it in the lab session 3 (Le, 2023). The dotted diagonal line in the plots represented the performance of a null model (random guessing).

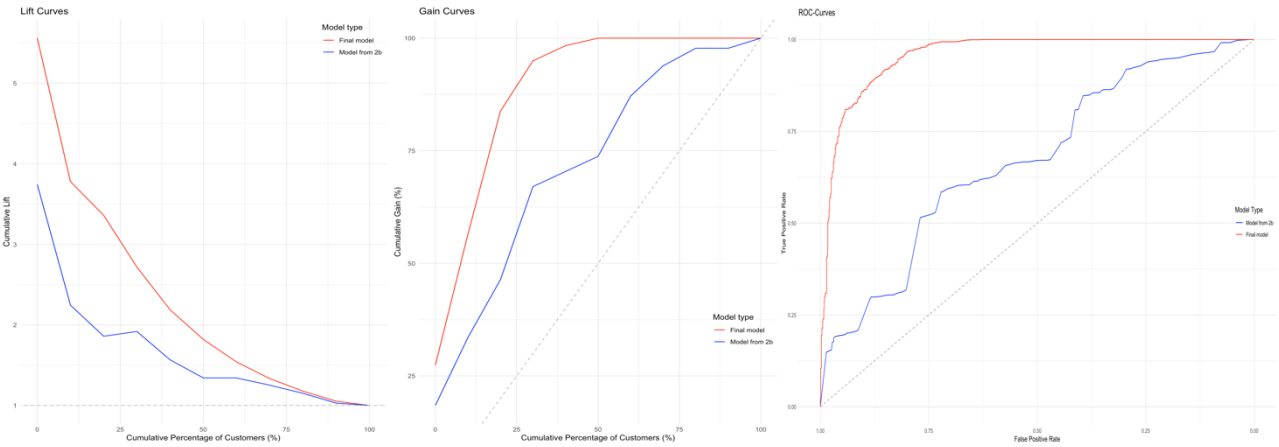

Figure 10 - Lift-, Gain- and ROC-curves

As we can see, all plots reveal the outperformance of the Optimal Model compared to the model from 2.b. The lift-curve is rising more sharply and is maintaining higher above baseline, whilst the gain-curve reveals the Optimal Model to be better at identifying large percentage of positive cases by targeting a relatively small portion of the population. The ROC-curve shows that for the optimal threshold, the Optimal Model is much closer to the ideal value of TPR (1) and FPR (0), than the model from 2.b.

The conclusion from the comparison is that the new model is better at predicting whether a customer will churn or not. With an increased optimal threshold, we are also more confident in our predictions. As the dataset is so skewed regarding churns and non-churns, the hit-rate alone is an insufficient measure. Merely predicting "not-churn" for all customers would provide a high hit-rate. As we see from our optimal model, we are now much better at predicting actual churns, which is what is of the highest importance as churning customers *never* will return. Preventing customers from churning is therefore the number one priority.

1.2.e Managerial implications

In the model from 2.d., we decided to do some feature engineering, slightly expanding the number of predictors, and greatly improving model performance. However, the new model might introduce some necessary changes to current policies. As revealed by the gain curves, the new model is better at capturing positive cases with minimal effort, and hence particularly useful when resources are limited, which is fair to assume would be the case for a non-profit organization. The Optimal Model highlights key aspects for predicting customer churn, and

hence potential areas for improvement. Following are some examples of how these insights can be turned into business actions.

Optimize ad-spend

The relatively large negative correlation between Adv_per_visits and likelihood for churn somewhat counterintuitively indicates that higher CPC (cost per click) correlate with lower likelihood for churn. This does however not make much sense, and it seems that the effect of ad-spend is baked into variables coefficient. Hence, the non-profit should focus on advertisement. But, as they probably have limited resources, they should be seeking to maximize the quality of these advertisement, and maybe link customers to the Facebook page instead of the website, as this appears to have a greater impact on the likelihood for churn.

Lobbying

Increasing publicity appears to have the desired effect on the likelihood for churn. Hence, a strategy for the Foundation could be to increase lobbying for increasing articles written about them etc.

Provide "special care" to freshly obtained customers

Number of donations per customer seems to greatly influence the likelihood of churn, donors with fewer donations being more likely to churn. Hence, the company should focus on providing their new customers with enough care and follow-up to avoid them churning. As we will discuss in the next paragraph, this does not necessarily mean spamming them with DMs. As discussed in our hypotheses above, for example, it might be worthwhile for the Arima Foundation to encourage people in charge of the DM to talk to regular donors in a concrete way (Packard and Berger, 2021). Overall, the key is to create and maintain an emotional connection between the Foundation and the donor (Leeflang et al., 2015).

Do not spam customers with direct messages

In contrast to the model from 2.b, the Optimal Model reveals a positive correlation between DMs and likelihood for churn. It is unclear whether there exists a causal effect, but either way, the positive correlation makes increasing DMs a risk not worth taking.

1.2.f Additional data

If we had the opportunity to go back in time and collect more data we would collect data regarding customer satisfaction, credit score, donation schemes, additional payments, and political orientation. How satisfied a customer is with the subscription process, the non-profit, and overall satisfaction in the first months after subscribing might prove to be valuable for predicting churn, especially considering the negative correlation between the number of donations and churn likelihood. Information regarding the customer's credit score might also come in handy, as customers applying for loans and hence increasing their debt might be "forced" to churn due to financial reasons not explicitly revealed through their household's

income. It would also be beneficial to collect data about the donation schemes for each customer, for instance if the customer made an additional payment in a specific month. By doing this, we would be able to calculate the customer lifetime value (CLV), and potentially locate our "most valuable" customers. Holding on to the most valuable customers would be of greater importance than holding on to fewer valuable customers. Finally, political orientation might reveal a systematic difference between different groups of people, and could possible help predict churn likelihood.

If we were to go even further in the analysis, another aspect may be interesting to look at: it would be the social network analysis. Hælein is quoted by Leeflang, Wieringa, Bijmolt and Pauwels in their book because he has shown, with its churn prediction model, the importance of social interactions in the customer churn decision. Indeed, one donor is more likely to churn if the donor "next to him" or connected to him in the social network previously churned (Leeflang et al. 2015).

1.3 CLV Analysis

1.3 Exploration of the dataset

When exploring the dataset "cf_data", we discovered that the dataset consisted of 59 observations with no missing values or duplicates. We checked for duplicated in all the predictors, and the only predictor with duplicates was the variable "TV_Adv". This variable had 2 outliers, so we were left with 57 observations after pre-processing the data.

From the summary function in R, we find that customer 10569 was a woman in her mid-40s. The household of this donor consisted of 2 people and the total annual income of this household was €36,943 throughout this whole period. Comparing this income with our results from part 1.1, we notice that this income level falls within the upper-end of household income-groups.

1.3.a Manually predict churn probability

To manually calculate the churn probability of the donor in February 2014 based on the binomial logistic regression model we build in parts 2.a and 2.b, we use the following formula:

$$P(Churn_i = 1) = \frac{\mathrm{e}(V_i)}{1 + \mathrm{e}(V_i)},$$
$$where\ V_i = \beta_0 + \beta_1 DM_i + \beta_2 TVAdv_i + \beta_3 FacebookAdv_i + \beta_4 Publicity_i$$

From the binomial logistic regression model in parts 2.a and 2.b, the estimated coefficient size was equivalent to

$$V_i = -1.2102 - 0.2801 DM_i - 0.0000735\, TVAdv_i - 0.0000105 FacebookAdv_i - 0.02287 Publicity_i$$

Then, we created a subset of the dataset "cf_data", where we filtered on the conditions Month and Year to only display the observation of interest. Subsequently, we removed the variables in this subset that was not used for prediction. This gave us the following sizes of the predictors:

$$DM = 1, \quad TVAdv = 0, \quad FacebookAdv = 2140{,}78, \quad Publicity = 46$$

After multiplying the estimated coefficient with the predictor's sizes, we got $V_i = -2{,}566$. Thus, the manually predicted probability for the donor 10569 to churn in February 2014 was:

$$P(Churn_{10569} = 1) = \frac{\mathrm{e}(-2.566)}{1 + \mathrm{e}(-2.566)} = 7.1417\ \%$$

1.3.b Predict churn probability with the predict function in R

When using the *predict* function in R, we rewrite the predict function so the "*newdata*" argument equals the dataset "cf_data". We let the "type" argument unchanged, so it is still equal to "response" since our goal is to calculate the predicted probability with values ranging between 0 and 1. The predicted probability for this donor to churn on its payment in February 2014 with the *predict* function was equal to 7.141 %. This method gave an exact match to the predicted probability we calculated manually.

1.3.c Compute the lifetime value

We start by computing how many times we predict the customer to churn given our logit model in parts 2.a. and 2.b. By using the optimal threshold of 0.07866 for this model, we assign this donor to churn in 14 periods, which is equivalent to 24.56 % of the time. This high probability of churning suggests that it could potentially be a bad decision to keep the customer. We will examine this further by computing the lifetime value of the customer.

In the formula presented in the lecture 2 (Le, 2023) the customer lifetime value can be expressed as follows:

$$CLV_i = \sum_{t=1}^{T} \frac{(Future\ Contribution\ Margin_{it} - Future\ Cost_{it})}{(1 + discount\ rate)^t}$$

Given our logit model in parts 2.a and 2.b and the additional information provided in this task, we can rewrite this formula as:

$$CLV_{10569} = \sum_{t=1}^{T} \frac{\left(Donation\ Scheme_{10569,t} - (DM_{10569,t} + TVAdv_{10569,t} + FacebookAdv_{10569,t} + Special\ Gift)\right)}{(1.006821)^t},$$

where the discount rate in the expression above is equal to the monthly discount rate.

Scenario 1

We are given the information that this customer churns one time and then the Foundation successfully manage to persuade the customer to stay with the organization. In our first prediction of the customer in February 2013, we predict the customer to churn. Therefore, in this scenario, we assume this is the first time where the organization persuades the customer to stay. For this specific reason, we add the one-time cost of the special gift to € 5 in the first period. The next time we predict the customer to churn is at 14 periods later, Mars 2014, we

will therefore calculate the CLV for this customer throughout the period February 2013 to Mars 2014. With a recurring donation scheme at 3 euros per month, we calculate the net CLV for this customer to be equivalent to € 22,55. Since the net CLV is positive (that is, greater than 0), it would be a good decision for the company to keep this customer regardless of our prediction that the customer churns 24,56 % of the time. In these calculations, we added the donation scheme in the period we predicted the customer to churn, since we are assuming that the customer was about to terminate the recurring scheme, but the organization persuade the customer to stay with this special gift. If we increase the donor`s recurring donation scheme to 5 euros per month, it would be even more profitable. In this case, we would get a net CLV of € 47,48.

Scenario 2

In the first dataset "altred_data", the customer 10569 started her recurring donation scheme in December 2012 and terminated her donation in January 2013, indicating that this could be the actual month the company managed to persuade the customer to stay. Since we predicted the customer to churn already in February 2013 in the second dataset "cf_data", it would not be profitable to keep this customer. This holds true for both donation schemes at 3 euros and 5 euros, as the organization would incur the expense of sending a special gift, along with additional costs such as the one-time direct contact with the customer in that particular month and advertising expenditures on Facebook. However, the future contribution margin for the customer would remain at 0 as we assume that the donation is scheduled at the end of a month. Consequently, the net Customer Lifetime Value (CLV) amounts to minus €5,46, irrespective of whether the donation scheme is set at 3 euros or 5 euros.

These two scenarios illustrate how sensitive the computation of the lifetime value of a donor could be, due to relatively high proportions of periods where we assign this customer to churn.

References

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. Journal of Marketing, 1-25.

Blattberg, R. C., Byung-Do, K., & Neslin, S. A. (2009). Database marketing analyzing and managing customers. Springer Science + Business Media LLC, chap 27.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods & Research, 33(2), 261-304.

Eggers, F., & Sattler, H. (2011). Preference measurement with conjoint analysis. Overview of state-of-the-art approaches and recent developments. GfK Marketing Intelligence Review, 36-47.

Fatemi, F. (2022, November 14). How TV viewing habits have changed. Forbes. https://www.forbes.com/sites/falonfatemi/2022/11/14/how-tv-viewing-habits-have-changed/?sh=3028ef684888

Fung, B. (2020). The hard truth about the Facebook ad boycott: Nothing matters but Zuckerberg. CNN Business. https://edition.cnn.com/2020/06/26/tech/facebook-boycott/index.html.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (1st ed.) [PDF]. Springer.

Jockers, M. L. (2014). Text analysis with R for students of literature. Springer. https://doi.org/10.1007/978-3-319-03164-4

Le, N. Q. (2023). Conjoint analysis (part 1). Lecture 5. MBM433 Customer Analytics in a Digital World.

Le, N. Q. (2023). Conjoint analysis (part 2). Lecture 6. MBM433 Customer Analytics in a Digital World.

Le, N. Q. (2023). Customer lifetime value and RFM analysis. Lecture 2. MBM433 Customer Analytics in a Digital World.

Le, N. Q. (2023). Discrete choice and (binomial) logistic regression. Lecture 3. MBM433 Customer Analytics in a Digital World.

Le, N. Q. (2023). Introduction to natural language processing (NLP). Lecture 9. MBM433 Customer Analytics in a Digital World.

Leeflang, P., Wieringa, J. E., Bijmolt, T. H. A., & Pauwels, K. H. (2015). Modeling markets: Analyzing marketing phenomena and improving marketing decision making, chap. 9.

McFadden, D. (1974). *Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), Frontiers in Econometrics (pp. 105-142). Academic Press.*

Packard, G., & Berger, J. (2021). How concrete language shapes customer satisfaction. Journal of Consumer Research, Volume 47, Issue 5, 787-806.

Pope, L. (2022). How to prevent customer churn with retention marketing. G2. https://www.g2.com/articles/customer-retention-marketing.

Steiner, M., & Meißner, M. (2018). A user's guide to the galaxy of conjoint analysis and compositional preference measurement. Marketing (Munich), 3-25.