Multivariate Analysis
Master in Data Science
Final Project: Predicting Customer Churn
Group 4 (class 11)
January 17th, 2021

Gerard Sanchez……..gerard.sanchez.maltas@estudiantat.upc.edu
Sebastian Paglia…….sebastian.paglia@estudiantat.upc.edu
Emmanuel F. Werr…….emmanuel.werr@estudiantat.upc.edu
Agustina Martinez……agustina.martinez@estudiantat.upc.edu
Meritxell Arbiol…….meritxell.arbiol@estudiantat.upc.edu
Camila Perez……..camila.perez.millar@estudiantat.upc.edu
Gerard Pons……..gerard.pons.recasens@estudiantat.upc.edu

INDEX

1. Problem formulation and data set

Problem: The goal of this project is to explore and address a real-life problem found in banking and many other businesses: customers dropping out of a service and moving to the competition. In particular, our dataset tackles the problem of customers leaving the credit card services from a specific bank.

The objective is to predict the customers who are likely to drop off from an undisclosed bank. This will allow account managers to proactively reach out to the customer and attempt to provide them with a better service. Thus reversing the customers' initial decision to drop off.

This dataset consists of 10,000 customers and contains 16 variables for each one, addressing banking metrics as well as more personal information:

- Numeric features: Customer's age, Number of dependents, Months on bank, Number of products held, Months inactive last year, Number of contacts last year, Credit Limit, Revolving Balance, Average Open-to-Buy amount, Change in transaction amount and in count from Q4 to Q1, Average card utilization ratio, Amount spent in transactions and Number of transactions in the past 12 months.

- Categorical features: Card product (Blue, Silver, Gold, Platinium), Gender (Male, Female), Marital status (Married, Divorced, Unknown, Single), Education Level (Graduate, High-school, Unknown, Uneducated, College, PhD, Post-Graduate), Income Category (Less than $40K, $40K - $60K, Other).

- Target: Attrition flag (Existing Customer (1) and Attrited customer (0))

The dataset was found in Kaggle and can be accessed here: BankChunkers.csv

2. Work Plan and Assignment of Tasks

2.1. Work plan

The workflow of the project will go tightly along with the course theory and lab schedule, as some of the practices and concepts that we are expecting to use will be introduced to us in future lectures. We plan to work on and learn the concepts explained in the labs and add them appropriately to our main project. We will be using the software R, with RStudio, using an online work collaboratory application (i.e GitHub/Google Collab) to enable all of us to keep track of the advances. In the following sections, we will describe a tentative timetable and what each team member is going to work on, as well as the contingency plans.

2.2. Calendar

Selection: Select the dataset that meets the previously stipulated requirements and explore a potential problem reaching out for a solution.	Planned start date: 16/09/2021 Planned end date: 20/09/2021
Pre-process of data: Carry out the first review of the data, detecting outliers, missing values, errors and establish the best decision based on theory.	Planned start date: 20/09/2021 Planned end date: 27/09/2021
Preparing data for analysis: Reach the final model or protocol that fulfills the expected quality, before starting the relevant analysis.	Planned start date: 27/09/2021 Planned end date: 11/10/2021
Analysis: Put into practice a multivariate analysis of the training data set, also choose statistical tools that are efficient to be able to predict solutions to the main problem and the different tests that we could apply. On the other hand, describe the relationship between services offered and customers dissatisfaction and the interpretation of the results.	Planned start date: 11/10/2021 Planned end date: 01/11/2021
Prediction: Choose a model for prediction who is going to get churned so we can provide a proactively solution in order to offer better services and turn customers' decisions in the opposite direction.	Planned start date: 01/11/2021 Planned end date: 09/01/2022

2.3. Assignment of tasks

This is a preliminary assignment based on our current knowledge about the vast majority of the techniques and the workload associated with them. Hence, we are aware that in the future we may need subdivisions of the main tasks, which will be split with the members of the main task.

TASK	MEMBERS
Dataset Selection	ALL
Exploratory analysis, Errors and NAs	Gerard S., Meritxell, Camila and Emmanuel
Outliers	Sebastian, Agustina and Gerard P.
Feature Extraction	Camila, Sebastian and Meritxell
Validation Protocol Election	Emmanuel and Gerard P.
MVA techniques selection	ALL
MVA 1	Agustina, Gerard P. and Gerard S.
MVA 2	Camila, Meritxell, Sebastian and Emmanuel
Prediction Model Election	ALL
Optimal Model Parameters	Camila, Gerard P., Meritxell and Sebastian
Model Performance	Agustina, Gerard S. and Emmanuel

2.4. Analysis of potential risks

During the assignment of tasks, we made sure to assign at least two team members to the same global task. These will enable us to have at least someone with in-depth knowledge of what somebody else is/was doing in case some unexpected problems may arise (from medical issues to lack of work/disagreements). Additionally, we will assign a backup team member to each of the tasks, in order to help with the increased workload that this issue could cause to anyone. The backup assignment is shown in the next table:

TASK	BACK UP MEMBER
Outliers | Optimal Model Parameters	Gerard S.
Exploratory analysis, Errors and NAs	Sebastian
Feature Extraction	Emmanuel
Validation Protocol Election	Agustina
MVA 1	Meritxell
Model Performance	Camila
MVA 2	Gerard P.

Other risks that could occur could be related to an increase of COVID-19 cases that could force us to work from home. If that is the case, it won't cause a large problem, as we will work from an established Google Meet room.

3. Data Preparation

This was the first, and one of the most important steps we did in order to start working properly with the chosen dataset. As a preliminary process, we deleted two features from it because they were calculations previously done with some of the other dataset's features, and were not properly documented. Moreover, the qualitative values were transformed to factors with the appropriate levels.

3.1. Missing Values

A priori, our dataset showed an absence of missing values, but further exploration of it revealed that in fact they were treated as a new category: '*Unknown'*. It must be noted that we found missing values for only three features, and all of them were categorical features: Income Category (11%), Marital Status(7%), and Education Level(15%). As can be seen in *Figure 1*, we found that there were clients having one, two, or all of the categories missing.

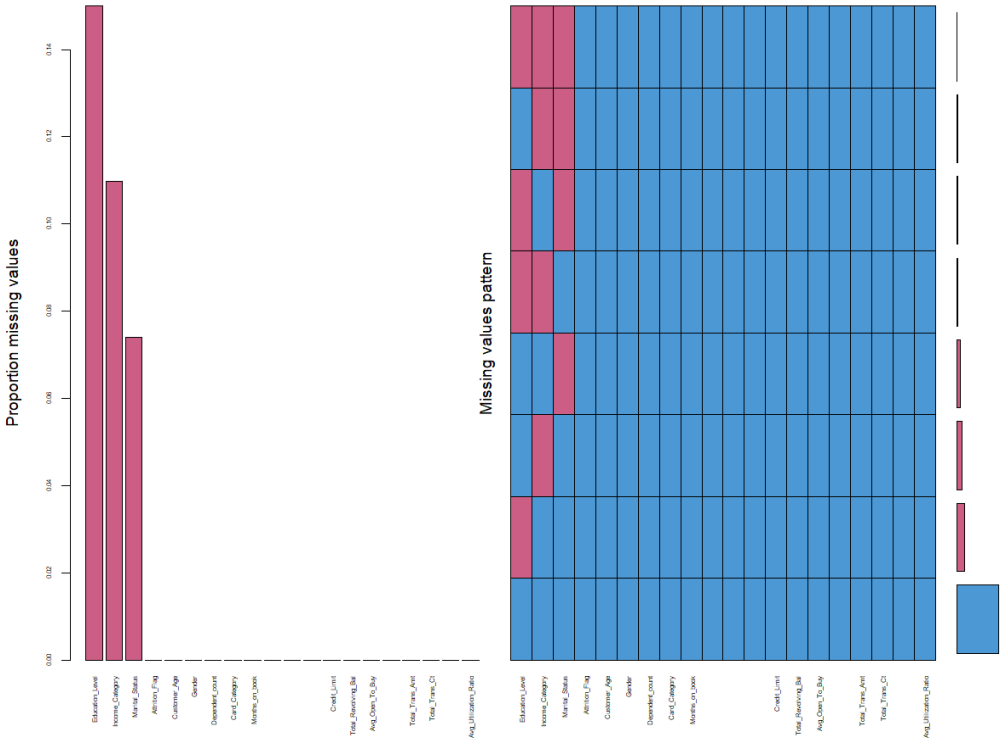


Figure 1

3.2. Logical Imputation

We started by trying to see if we could assess some clear relation of those missing values by exploring our assumptions/intuitions:

- First of all, we looked into a possible relation between **Income Level** and **Education Status**, thinking that people with higher education levels tend to have better-paying jobs. We performed a Chi-squared test on these attributes, which yielded a p-value of p=0.22, indicating that we can not refuse the null hypothesis, which was that features were independent.

- We also tried our second initial guess, which was a relation between **Marital Status** and **Number of Dependents**, assuming that people married have more economic dependents on them than people who are single and have divorced people in between. The p-value obtained was p = 0.02, hence there could be a slight dependency. However, as seen in *Figure 2* and by the fact that we will need to manually establish the rules of imputation, the variations and correlations are so subtle and the number of dependent categories is too large in order to properly identify sensible relations.

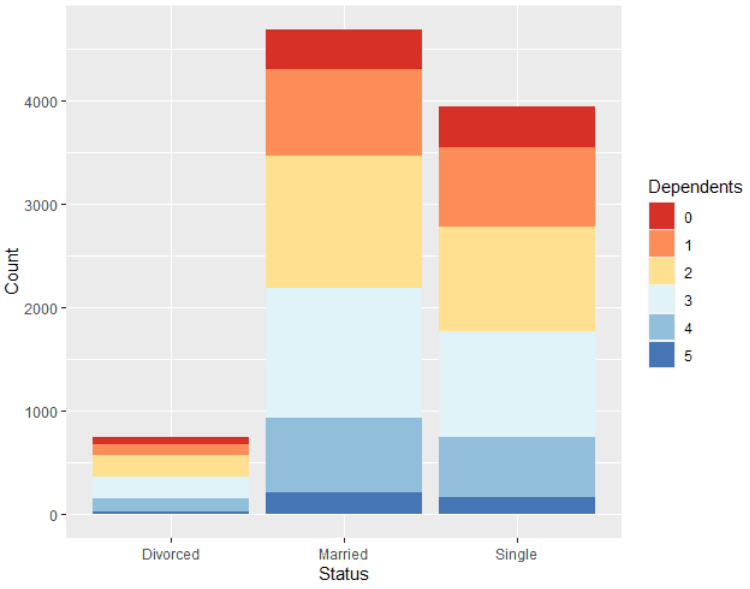


Figure 2

With those unsatisfying results, we discarded the logical imputation.

3.3. Automatic Imputation

We decided to address the imputation with the well-defined methods that R libraries provide us with. Concretely, we decided to approach the problem with two different algorithms: Random Forest and MICE, for which we tried different parameters of the functions to try to find the one that suited most our data. The results are shown in *Figure 3*, where we can see the frequency for each category for the original dataset, and with the different imputations. We can clearly observe that the MICE outperforms Random Forests as it preserves nearly exactly the original dataset's distribution in Education Level and Marital Status. However, it encounters some minor problems with the imputation of the Income Category variable. Nevertheless, we will continue working for now with the imputation mechanism as the frequency variation is in the worst of the scenarios of 2%.

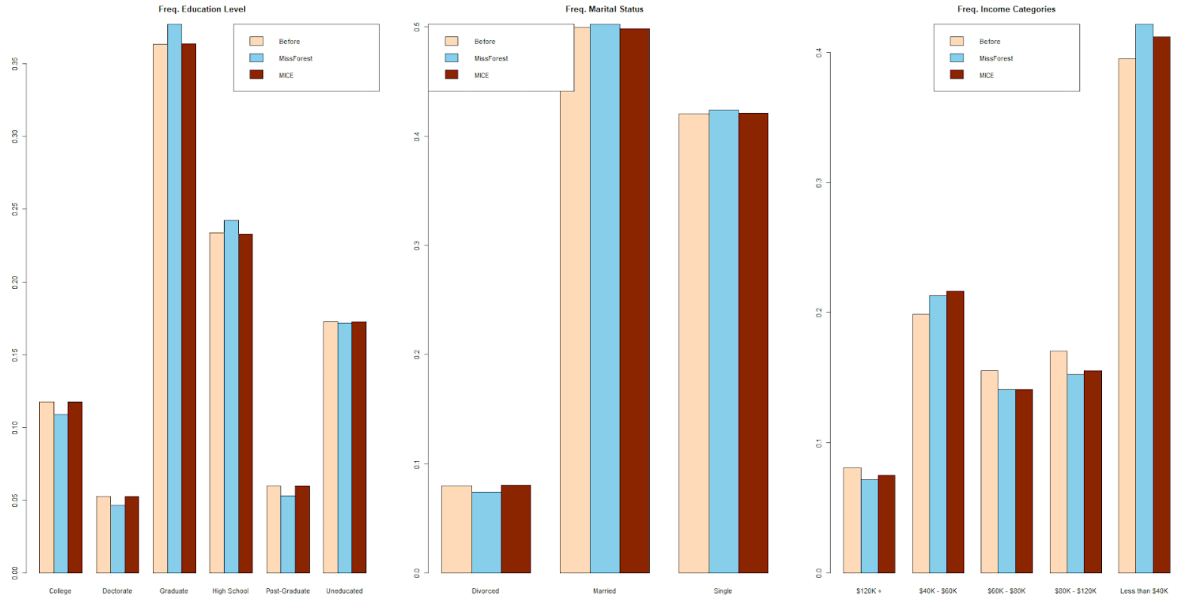


Figure 3

3.4. Outlier Detection

The next important step in data preprocessing is outlier detection. On the one hand, their presence could distort the statistical analysis and prediction models. On the other hand, they can be very informative and help us detect some extra grouping of the data and suggest different approaches to treat distinct parts of the dataset.

As we are dealing with a multivariate dataset, we performed multivariate outlier detection using the Mahalanobis distance. As we can observe in *Figure 4*, which represents the density distribution, there are a few outliers above the cut-off value, which is calculated to represent the 0.99 quantiles. Precisely, there are 173 users in that range that account for a little bit under 2% of entries in the dataset.

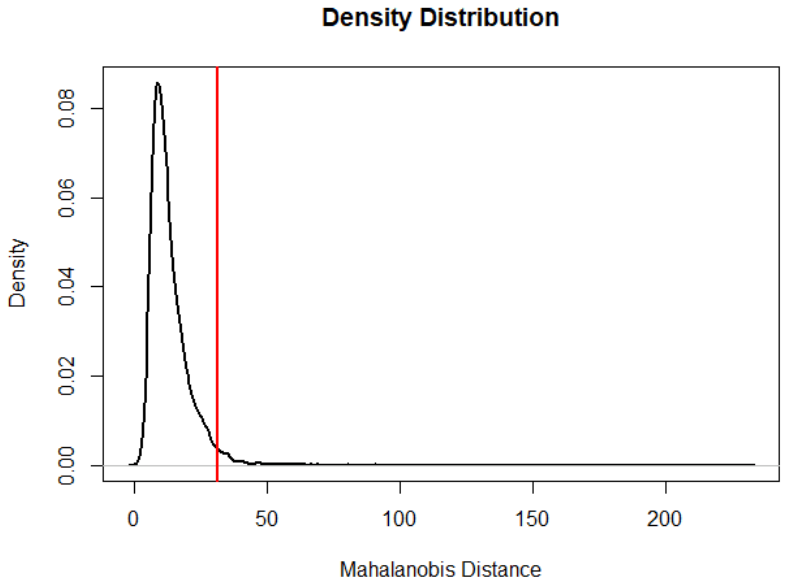


Figure 4

First of all we explored if there was a relationship between the outliers and our predicting task, but we could see that in the whole dataset 19.1% of the clients dropped off, whereas in the outliers the percentage was 17.2%. This made us conclude that there was not any strange behavior in terms of attrition among the outliers.

After that, we attempted to see if there were some outliers that were there because of an error while constructing the dataset in one or more variables that caused their values to be extreme. Hence we looked into outliers in a univariate manner and fortunately, mild outliers were only detected for the variables Credit Limit, Average Open to Buy, and Total Transaction Amount. As seen in *Figure 5*, Total Transaction Amount is the only one of the variates that have extreme outliers. However, we can not observe any out-of-place value that suggests that was due to an error.

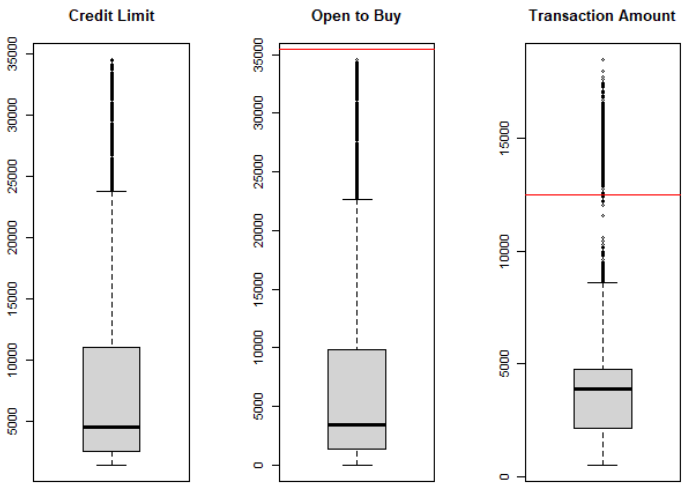


Figure 5

Hence, due to them accounting for only under 2% without the possibility to correct their values, we decided to remove them from the principal dataset and treat them as a separate one, trying to see if we can find an underlying pattern that originates from this new group. The first thing that strikes out is the percentage of men, as in the complete dataset the value is 47.1%, and in the outliers group is around 77.3%. Also, as we can see in *Figure 6*, there is a very significant increase in income compared to the original dataset, especially for the higher income categories (+120K and 80K-120K) and a very substantial decrease in the lower paying. This is also reflected in the Credit Card categories, as the percentage of the Blue one, which is the one with the lowest benefits, dropped 30%, while the more exclusive ones all increased up to 350%. Also, we observed these other relevant metrics:

	Mean Credit Limit ($)	**Mean Open to Buy ($)**
Original	22286	21085
Outliers	8632	7469

So, we describe the outliers as people with high income, mostly men, which tend to allocate more money to possibly big expenditures.

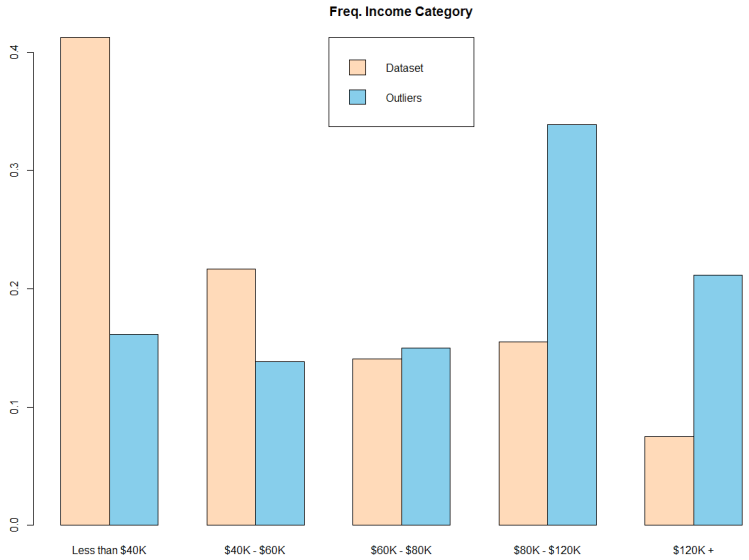


Figure 6

4. Dimension Analysis

As our dataset contains 10 numerical and 5 categorical variables, we decided it was worth exploring both PCA and MCA and trying to draw conclusions from them. However, MFA was also performed to use all the available descriptive variables.

4.1. Principal Component Analysis (PCA)

After applying PCA to our data frame imputed by the MICE method, and pulling apart the categorical features, we observed that nine dimensions from a total of fourteen, are enough to retain above 90% of the total inertia (variation) contained in the data. Taking into account that performing PCA allows us to analyze graphically just two dimensions at a time and that our two first dimensions only account for 30.8% of the total variation we decided to select Dim1 and Dim2 just to check the contribution and correlation among them, knowing that it is insufficient to fully describe the dataset.

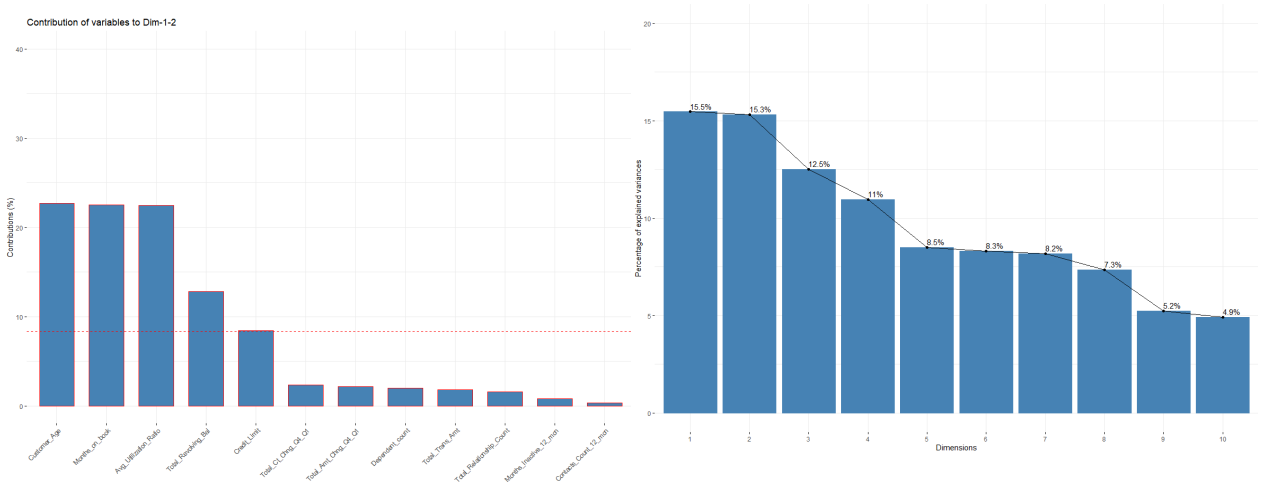


Figure 7

From the above plots that show the quality and the contributions of representation, we can infer that the best two variables for Dim1 are Customer_Age and Months_on_book. Regarding Dim2, we can mention Avg_Utilization_Ratio and Total_Revolving_Bal. Finally, the two features that contribute the most to the definition of both dimensions are Customer_Age, Months_on_book, and Avg_Utilization_Ratio. The variables not mentioned are not perfectly represented by PCA.

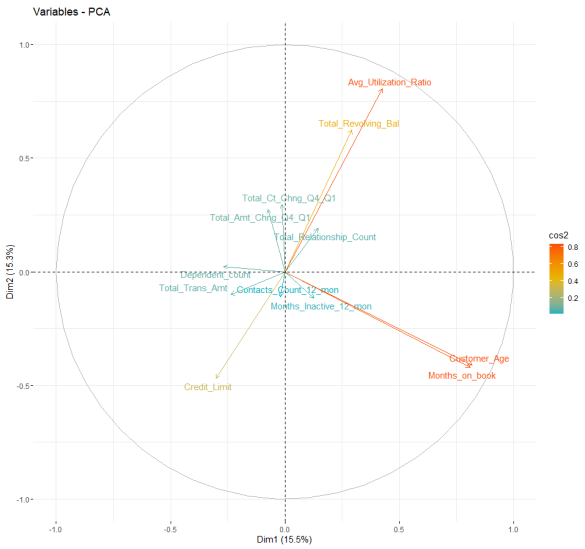


Figure 8

From the correlation circle plot we can mention potential positive correlation variables such as Customer_Age and Months_on_book. As a first approach to detect potential clusters, both features could belong to the same one. Moreover, there are negatively correlated variables, positioned in opposite quadrants, such as Credit_Limit with Avg_Utilization_Ratio - Total_Revolving_Bal.

In the previous graph, we can see that features near the center could be not significant and potentially not useful when creating the future models, for example, Dependent_count, Months_Inactive_12-mon.

4.2. Multiple Correspondence Analysis (MCA)

We performed an MCA analysis with our categorical variables and we encountered some of the same problems we bumped into in PCA. As seen in *Figure 7*, our first dimensions explain a little proportion of the inertia of the dataset, and all the following dimensions account basically for the same level of inertia. Hence, the conclusions drawn in this section should be interpreted with some caution. Moreover, as our dataset consists of 10000 observations, we will not plot the cloud of points of individuals, as it would be confusing.

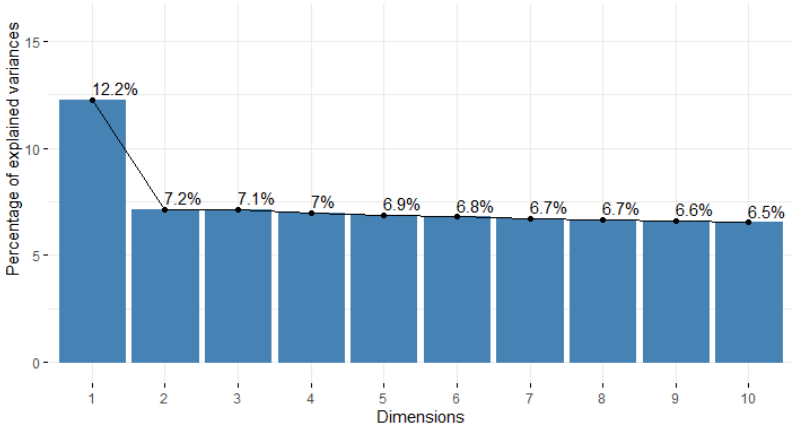


Figure 9

We analyzed the first two dimensions to see which variables (and categories) influenced them the most. First of all, it can be seen in *Figure 8* that the variables Gender and Income Category contribute to the first dimension. Indeed, as Gender is a dichotomous factor it only needs one dimension to be explained. The second dimension has not so strong contributions and is of the Marital Status, Card Category and Income Category. Observing the figure, Gender is completely separated by the first dimension along with different types of income, showing that men are associated with higher salaries than women. Also, we can not say much about Education as it is not well represented by these two dimensions.

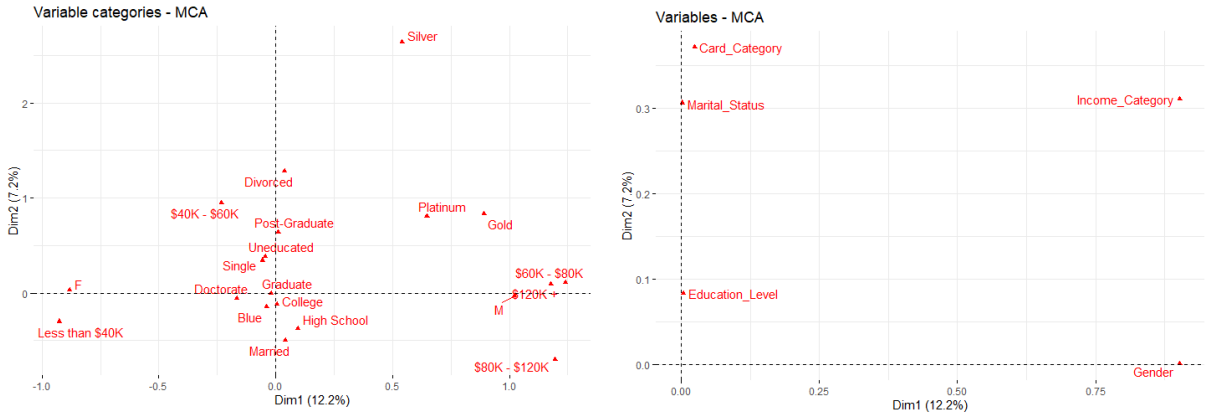


Figure 10

4.3. Multiple Factor Analysis (MFA)

To calculate the multifactor analysis to our data, first, we have excluded the variable Attrition Flag since it is our target, and we have decided to group the variables to calculate the MFA as follows due to the significance that each one has:

- *personal_num*: it groups age and economic dependence, two factors that we believe are closely related (Dependent_count and Customer_Age).
- *personal_cat*: Here we have the personal information (Variables: Gender, Education_Level, and Marital_Status).
- *personal_money*: We group two categorical variables related to the personal money information (Variables: Income_Category and Card_Category).
- *interactions*: we have grouped all the variables related to some kind of interaction with the bank (Variables: Months_on_book, Total_Relationship_Count, Months_Inactive_12_mon, and Contacts_Count_12_mon).
- *movements_ind*: Here we have grouped all the money movements that users perform (Variables: Credit_Limit, Total_Revolving_Bal, Avg_Open_To_Buy, Total_Amt_Chng_Q4_Q1, Total_Trans_Amt, Total_Trans_Ct, and Total_Ct_Chng_Q4_Q1)
- *move_ratio*: We have left one variable alone because it is a ratio, and we do not see that it makes sense to group it with any of the others (variable: Avg_Utilization_Ratio).

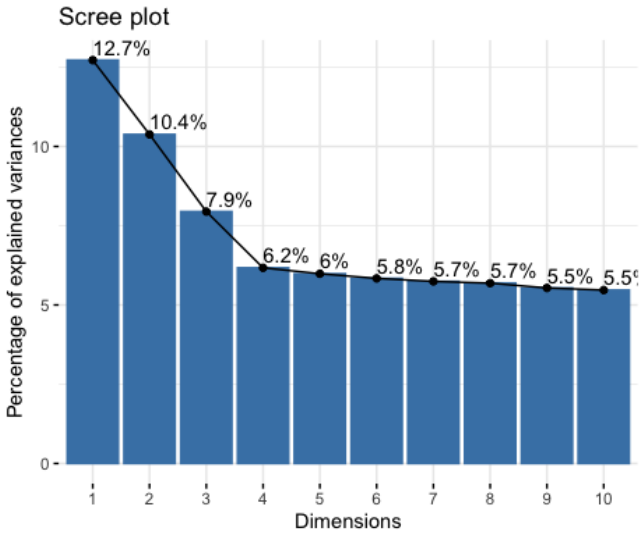


Figure 11

After computing the Multiple Factor Analysis we see that with the first two dimensions we have a total of 23.1% variance of the total of the original database (*Figure 11*). It is not a very large number, but after doing several tests as well as modifying the grouped variables, this is the maximum variance that we have been able to represent in only two dimensions.

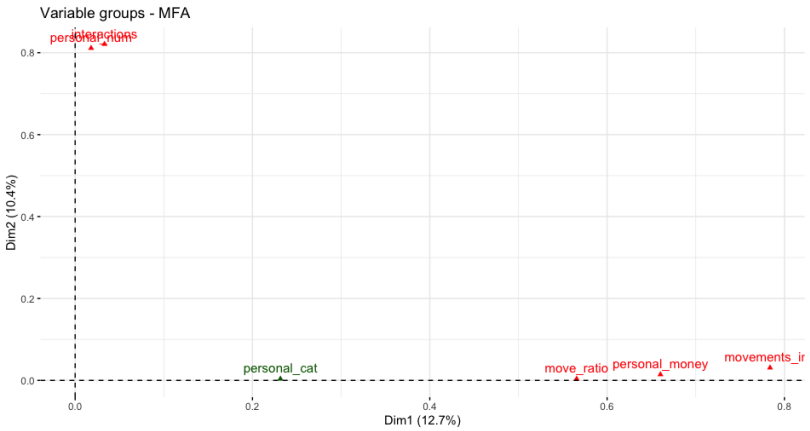


Figure 12

We can see in *Figure 12* that the groups move_ratio, personal_money, and movements_ind compose the first dimension which also means that they are highly correlated, and regarding dimension 2, it is mainly composed by the groups' interactions and personal_money.

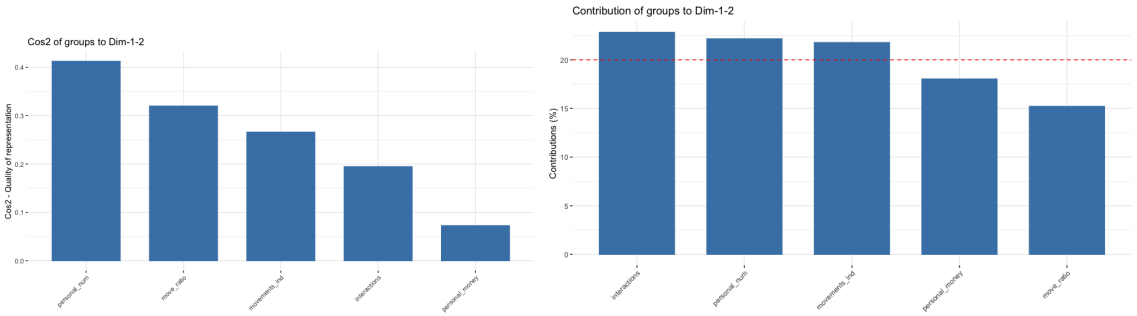


Figure 13

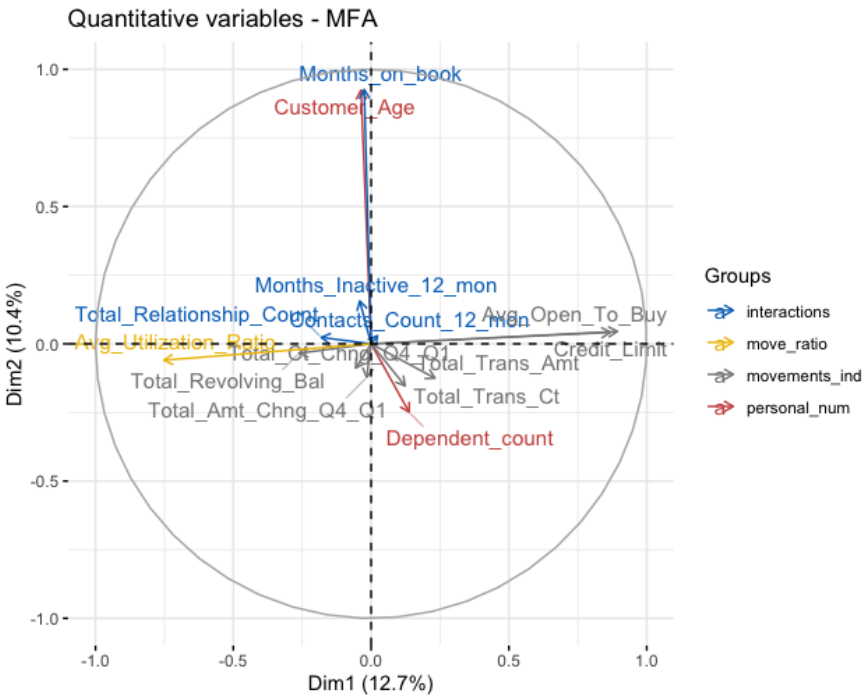


Figure 14

In the image above (*Figure 14*) we can check the correlation between the quantitative variables and the dimensions. One thing, in particular, catches our attention: the correlation between the variables Credit_Limit and Avg_Open_To_Buy because they are one on top of the other. We decided to learn more about these two variables using the cor.test (test for association/Correlation between paired Samples) and found that they do indeed have a correlation of 0.99, and it makes perfect sense since in the end one variable is calculated from the other. We also see that both variables are negatively correlated with Avg_Utilizatin_ratio, so we conclude that we will have to eliminate 2 of these 3 variables before calculating the model, to avoid collinearity effects, keeping only the variable Credit_Limit. We can also observe that the second dimension is due to Month_on_book and Age and that they are correlated, which makes perfect sense as the older customers are, the more likely they are to have been part of a bank for more time.

As a final remark, we have seen that for our dataset we can not obtain very relevant information with any of the unsupervised dimension reduction methods applied. We hope to be able to obtain more meaningful and insightful results in the next methods that we use with our dataset and that it does not have consequences or hinders the future creation of predictive models with our data.

4.4. Linear Discriminant Analysis (LDA)

LDA is a dimensionality reduction technique used as a preprocessing step in machine learning and pattern classification applications. We are aware that LDA is not a flexible technique

because there are many conditions to be satisfied in order to apply it. First of all, we have to check if our variables satisfy the three conditions:

a. All the predictors need to have a Gaussian or Normal Distribution.
   We approached this condition through histograms, qqplots, and Shapiro test, in order to know if we need to apply a transformation or not.

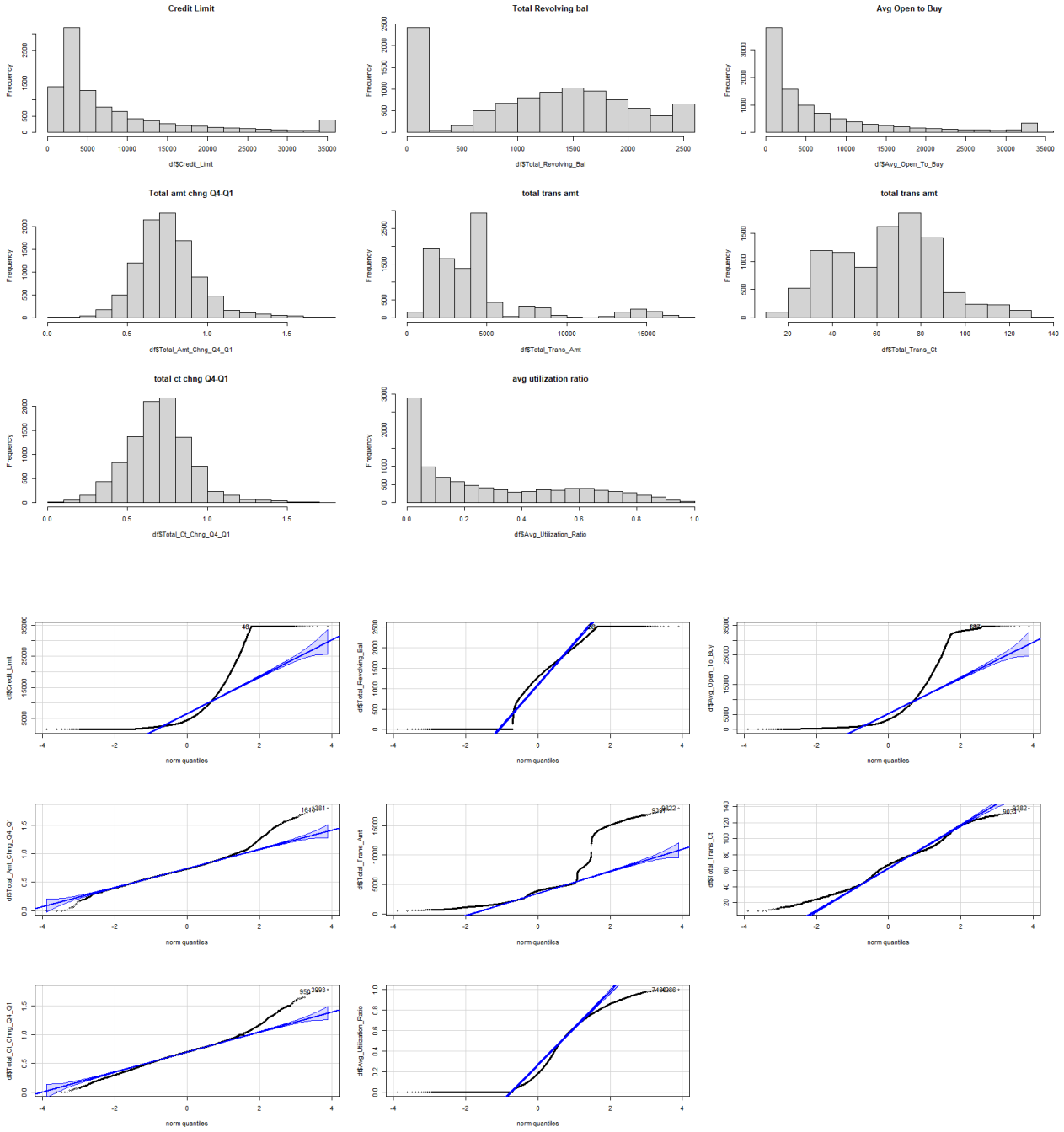
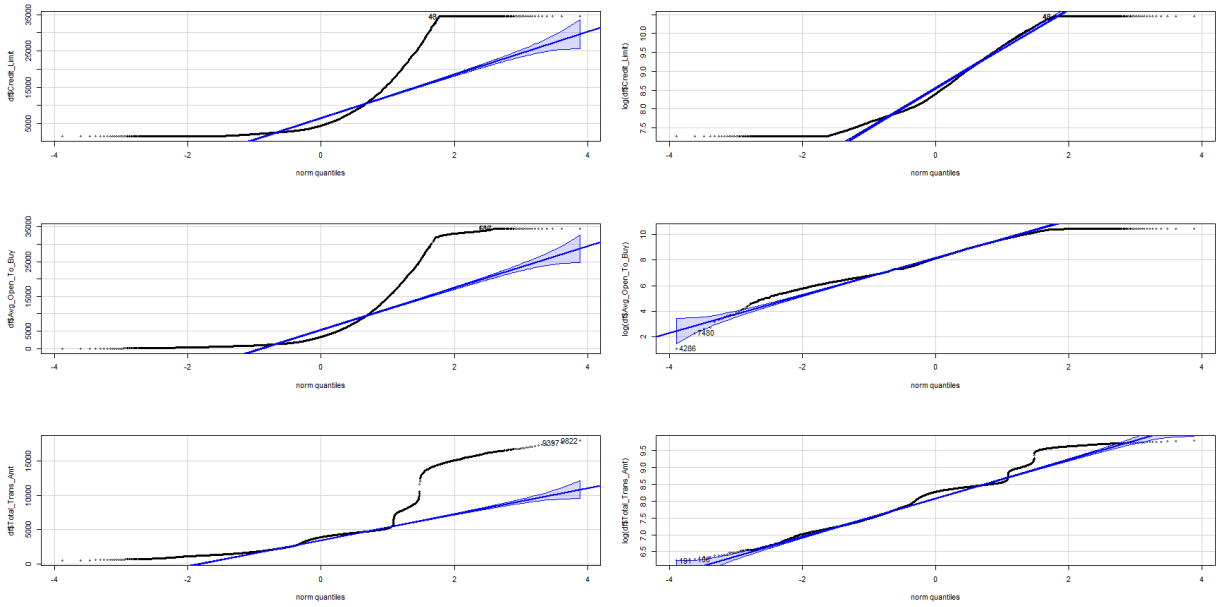


Figure 15

After the plots above, we realized that our variables do not follow a normal distribution, mainly Credit_Limit, Avg_Open_To_Buy, and Total_Trans_Amt. Therefore, we will try to normalize them with a logarithm transformation.

Plots showing before (left side) and after (right side) logarithmic transformation for these variables:



Even after this transformation, with the Shapiro test, the p-values obtained are very close to 0 so we have statistical arguments to reject the null hypothesis of normality.

b. The output needs to have a Multidimensional Gauss.
   In order to check this condition, we used mvn function but we obtained a very low p-value, thus we have statistical arguments to reject the null hypothesis and affirm that our variables don't follow Multivariate normality.

c. The covariance between groups in the data has the same variance.
   Finally, to check the variance between groups we have used the boxM function, and we have again a p-value close to 0, so we reject the null hypothesis. Therefore, there is no Homogeneity of Covariance Matrices.

Even though the conditions are not fulfilled, we continued with the analysis to get a first and initial predictive model.

The first step is to split the data into 80% training and 20% test. Secondly, we applied the LDA function to create our model with Attrition Flag as our binary target and the explanatory numerical variables.

With the following plot, we should be able to discriminate the response variable into its distinct classes but it looks poor because there is a clear overlap between both classes (possibly a consequence of not satisfying the previous conditions).

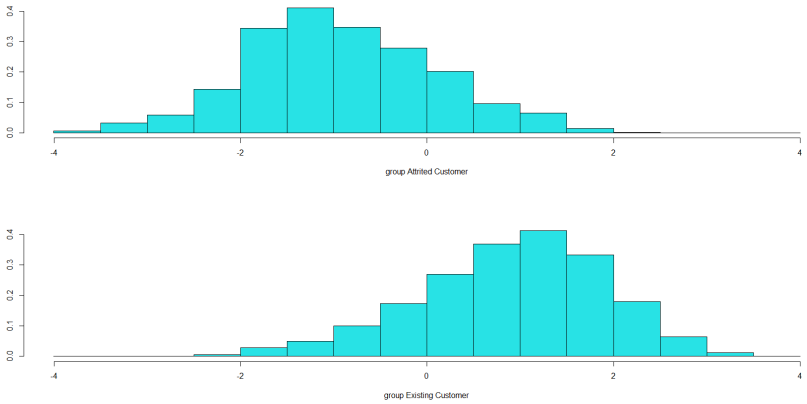


Figure 16

As a third step, we have made the prediction based on our model with the data test and we obtained the following results with an accuracy of 88%:

```
> table(test.transformed$Attrition_Flag, predictions_lda$class, dnn = c("Actual Class", "Predicted Class"))
                  Predicted Class
Actual Class        Attrited Customer Existing Customer
  Attrited Customer               168                148
  Existing Customer                85               1572
> mean(predictions_lda$class==test.transformed$Attrition_Flag) #accuracy
[1] 0.8819057
```

We are going to perform the QDA method too because the sample size is big and it does not assume the equality of group covariance matrices mentioned in our third condition. We have made the prediction based on this model with the test data and we obtained the following results with an accuracy of 87%:

```
> table(test.transformed$Attrition_Flag, predictions_qda$class, dnn = c("Actual Class", "Predicted Class"))
                  Predicted Class
Actual Class        Attrited Customer Existing Customer
  Attrited Customer               202                114
  Existing Customer               125               1532
> mean(predictions_qda$class==test.transformed$Attrition_Flag) #accuracy
[1] 0.8788647
```

Having in mind that we didn't fulfill the three main conditions, our results are not robust. Therefore, we will take them into account as an approach but we won't use it.

5. Clustering

The following step in the data exploration part was regarding the clustering of the observations. The first step needed was determining the best number of clusters into which we wanted to partition the data. We decided to start by assessing first a hierarchical clustering method to get an intuition of what would be a good number of clusters, and then move on to a partitioning algorithm.

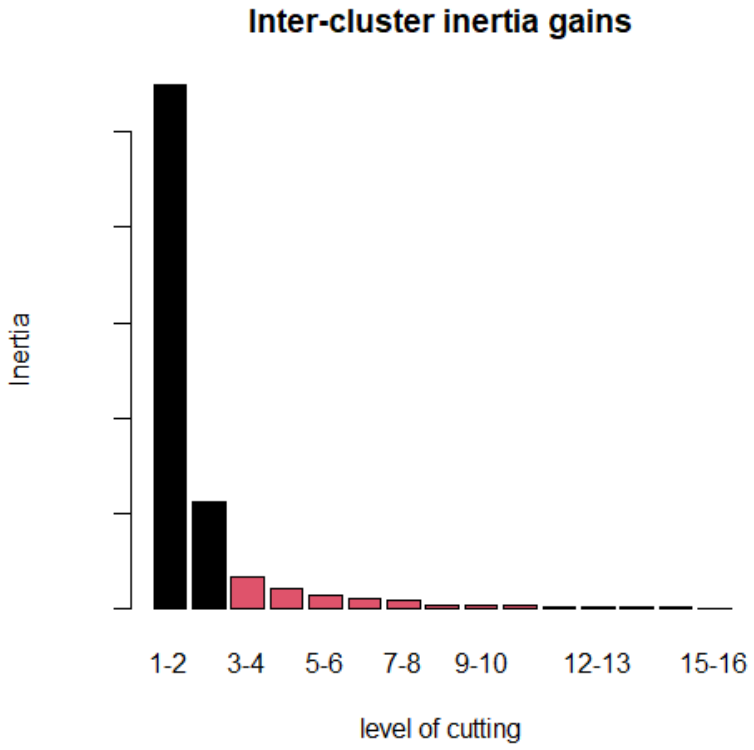
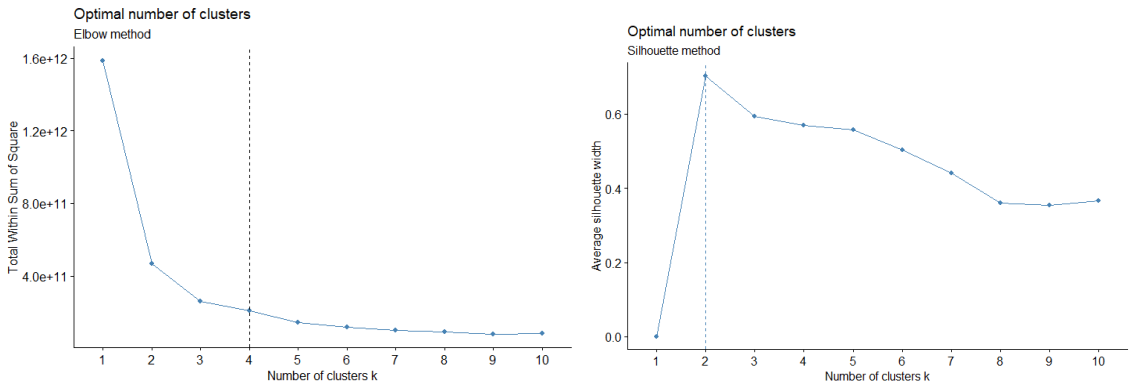


Figure 17

We can see that in this case, it could indicate to work with 3 clusters as the inertia loss can be argued not to be significant further increasing the groups. However, as other indicators can be used to argue the number of clusters, we decided to also use those. First of all, we looked into the within-group sum of squares behavior and using the heuristic elbow method, 4 clusters are suggested. Also, the Silhouette method has also been considered, which in this case suggests selecting only 2 clusters. Additionally, the GAP statistic method has also been assessed, and in this case, the value suggested is 6. The plots of all these figures can be seen below:



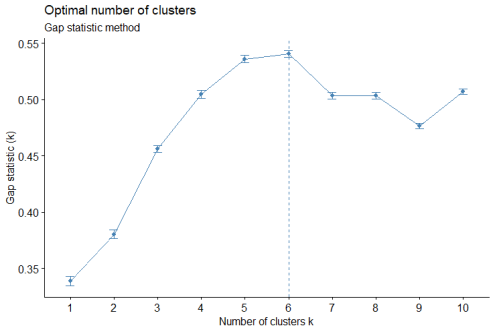


Figure 18

To deal with these discrepancies in the suggested number of clusters, we decided to use multiple indexes using the NbClust library (in our case 23 were used, due to the high computation cost that had some of the algorithms over our dataset) and chose the one that was suggested more times, as can be seen in the figure below:

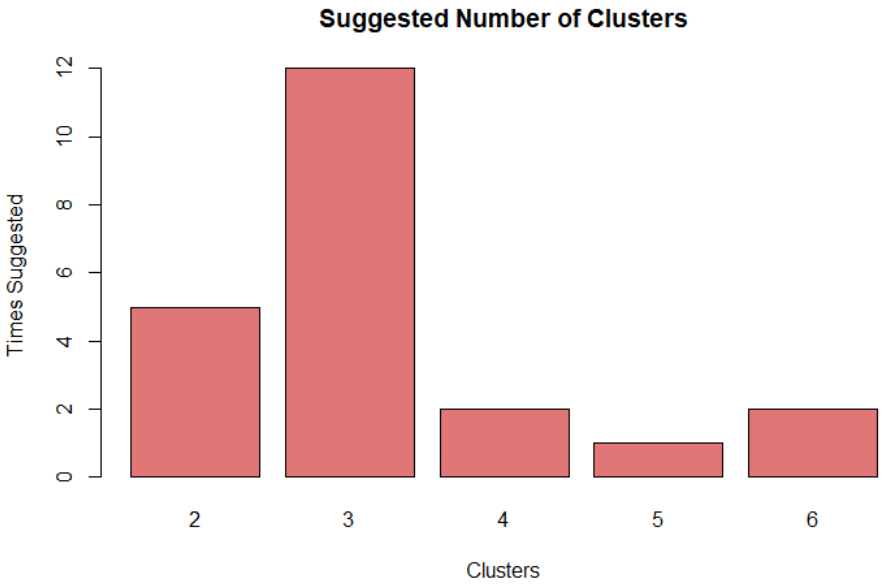


Figure 19

It can clearly be seen that the vast majority of the indices suggest 3 clusters, just as we obtained with the hierarchical approach. In the visualization of the clusters, we can observe that Dimension 1 is the one that helps differentiate between them the most and that Dimension 2, while apparently not being relevant for clusters 1 and 2, seems important in cluster 3, as it takes mainly positive values. Regarding the inertia, separating with the three clusters explains 22.54% of the total Inertia, which is smaller than the two first components of the PCA. Other than that, they follow a simple visual pattern, so in order to gain a better insight into the groups, the profiling of them needs to be addressed.

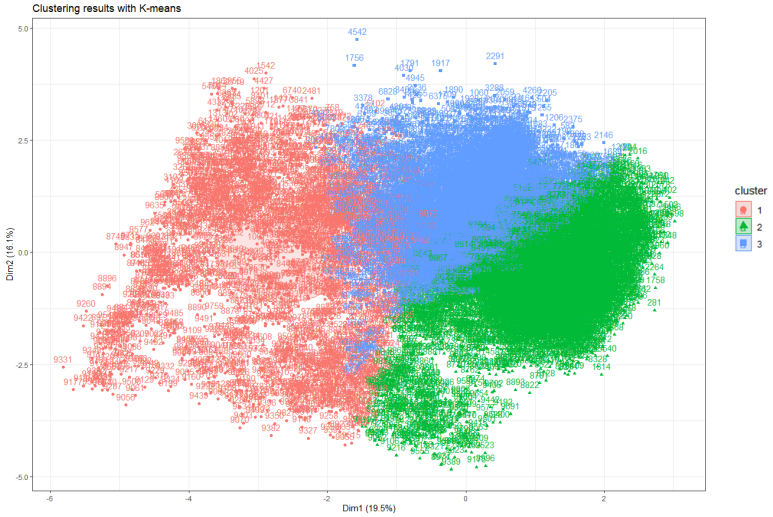


Figure 20

5.1. Characterization of the clusters

As an overall characterization of the clusters, it can be seen that the categorical variables whose proportions are significantly different among clusters are Income/Card Category, Gender, Marital Status, and Attrition Flag. Among the numerical ones, all of them except those not related to banking (age and dependent count) show significant correlations according to the p-value statistic. A more in depth characterization of the groups can be done:

Group 1: It is a group that consists of very high earners, with high credit limits, who are typically male (79.7%) and accounting for the larger proportion of not married individuals. This is a group that uses the credit card very frequently but it has few contacts with the banking services. It has a high proportion of customers that are not attired, so it could mean that they are content with the credit card services.
Group 2: It consists of people with low income and mostly female and married people. This group has obviously low credit limits but more utilization of the credit card than the other two groups: this indicates that people with families, in general, need to spend money on a wide range of products, activities, and services. Additionally, they contact the bank more frequently than the other groups. Additionally, the proportion of attrition is also low.
Group 3: In this group, there is not a significant distinction between genders or marital status, but there is a very clear increase in the number of contacts with the bank. It is a group with an income between the first two ones and with a low utilization ratio. It can be seen that the number of attrited customers is high among this group.

Also, it must the sizes of the different groups must be assessed:

GROUP 1	GROUP 2	GROUP 3
1794	4466	3607

We can observe that the grouping sizes reflect well the reality, as people with low/average salaries are in the largest group and few people represent the high earners.

As our response variable was only added on the characterization step and not in the clustering, the fact that we have found a relationship between the proportions could suggest that we could find good models to predict the intention of customers in the future:

	GROUP 1	GROUP 2	GROUP 3
ATTRITED	10.4%	7.1%	29.8%
EXISTING	89.5%	92.9%	70.2%

6. Classification Trees

In the last part of the project, classification tree models were assessed in order to predict the outcome of whether a customer will leave the bank or not. If the project was a real case scenario, this part would be very valuable, as the outcome of the model (in this case the split of the branches) could give important business information to the bank, enabling them to act on the customers that are likely to change banks (i.e offering better services) so they do stay on the bank.

6.1. Working with an unbalanced dataset

The first challenge we faced is that we worked with an unbalanced dataset. This is not surprising as this is a real dataset one would not expect 50% of the people leaving the bank. Concretely, 16.01% of the observations belong to attrited customers, which is a value that suggests that some balancing techniques need to be done. It must be noted that different approaches had to be taken when balancing the dataset depending on the output of the classification methods. The technique shown below is the one that yields better results, and the models that will be presented hereunder are all with the same balanced dataset to allow comparison:

- The size of the dataset was preserved, with 10000 observations and a ROSE method was used.
- Oversampled the Attrited Customers to 35%.
- Undersampled the Existing Customers to 65%.

It has to be said that it made no sense to balance the dataset to 50/50, as the oversampling of the underrepresented class would lead to overfitting.

Having done the balancing of the dataset the train/test split was performed. In our case it was decided to be done in a 70/30% split, leaving 7000 and 3000 observations respectively.

6.2. Model

In order to model with simple trees, three different modeling packages were used and compared. First of all, the cost penalty (which will impact the size of the trees) was decided using the one standard deviation rule:

1. First, we created a tree with a cost penalty of 0.

2. Then, we identified the tree with the minimum CV relative error and added to that value 1 standard deviation of it.
3. After that, we found the smallest tree whose CV relative error was less than the previously obtained value.
4. Finally, we pruned the initial tree to that CP value.

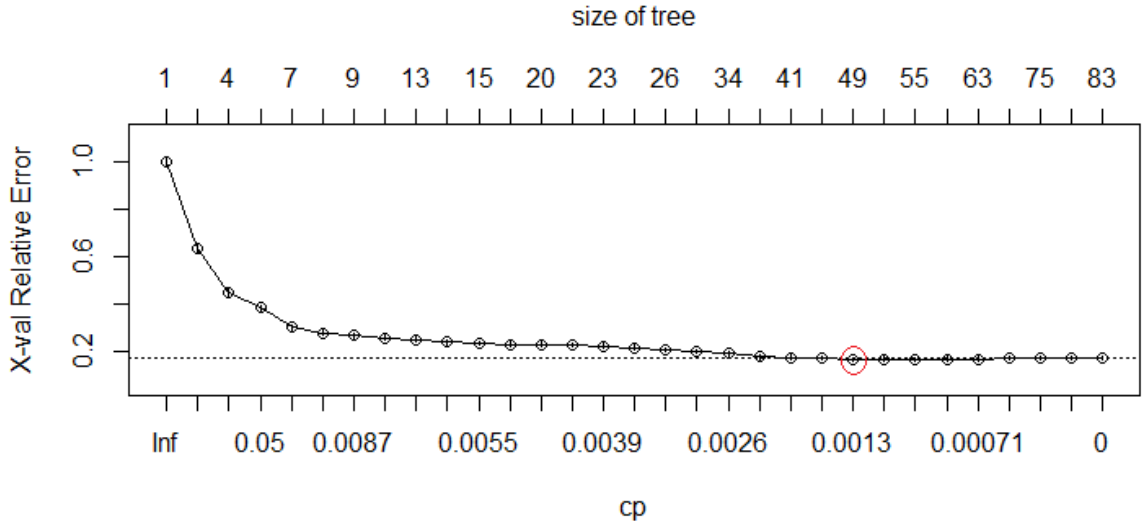


Figure 21

The best resulting tree can be seen in *Figure 22* and as it can seem to be a big tree, overfitting must be addressed to ensure that not many splits have been created and that the model is able to generalize well. The hypothesis of overfitting has been discarded as it can be that the training accuracy and the testing accuracy only differ by 1.5%, which is an acceptable difference.

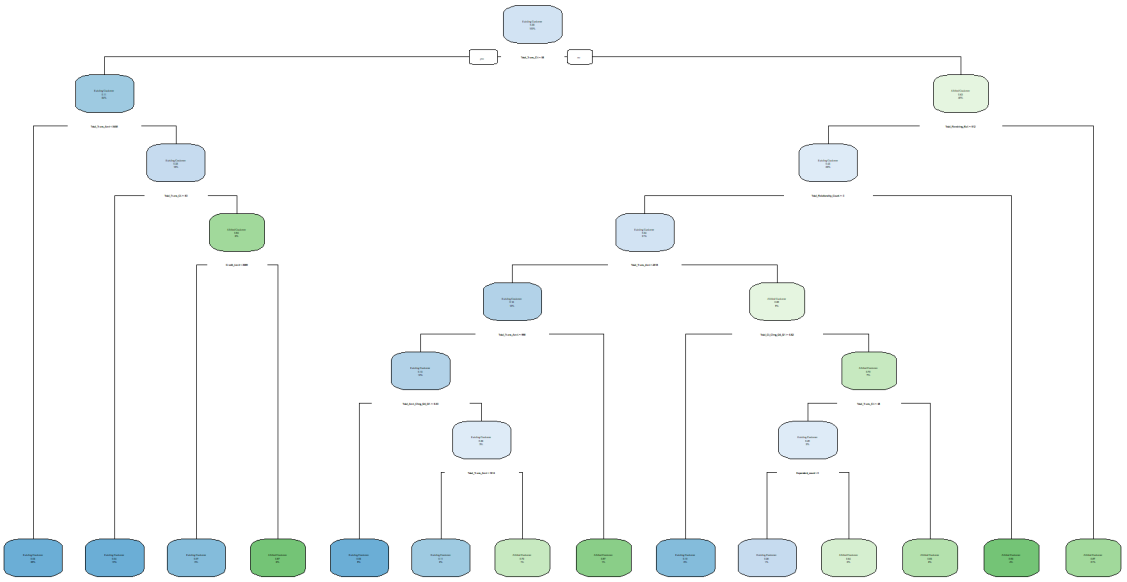


Figure 22 Overview of the best tree

A more detailed view of the results of the best-found tree can be seen in the figure below. One thing to remark is that although all the metrics should give satisfactory results, one should address the problem in a case-sensitive way. In our case, we are dealing with bank attrition, so we expect the models to be used by the bank to identify customers that are likely to leave

the bank and try to change their minds. In that case, one of the metrics we should inspect is `**recall**` as in our scenario, is better to have a false positive than a false negative: is better to for instance offer better conditions to an already happy customer than not interacting with a customer that is unsatisfied.

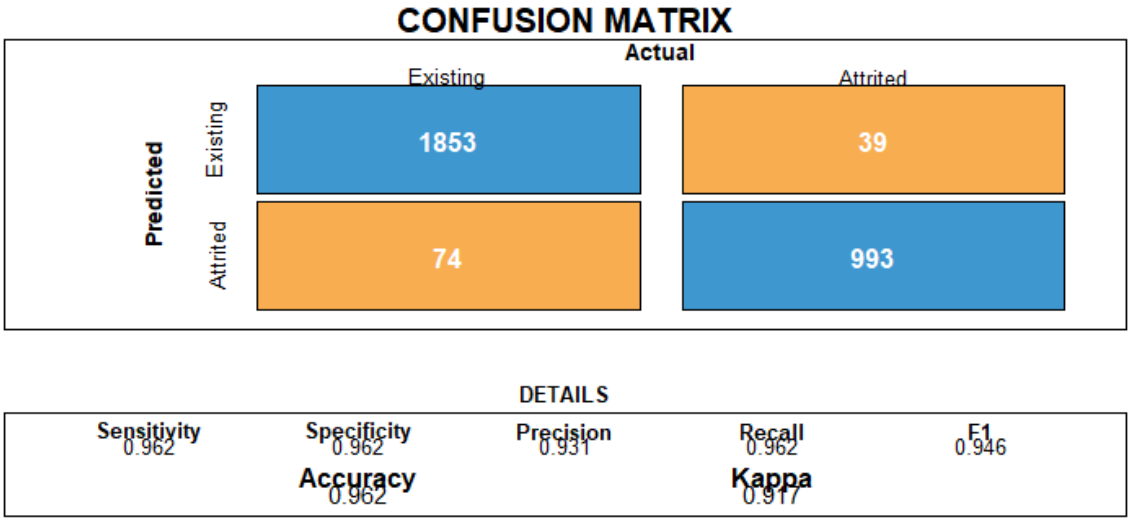


*Figure 23 shows the Confusion Matrix and metrics for the best model found over the **test** dataset.*

One other interesting property of the tree creation that we can assess is the importance that it gave to the different available descriptors (numerical and categorical). In this case, only 11 of the 19 available were used. As can be seen in *Figure 24*, the most important features used are the ones regarding amounts of money and interactions with the bank, and we can not see the personal information of Marital Status, Education, Dependent People, or Salary.

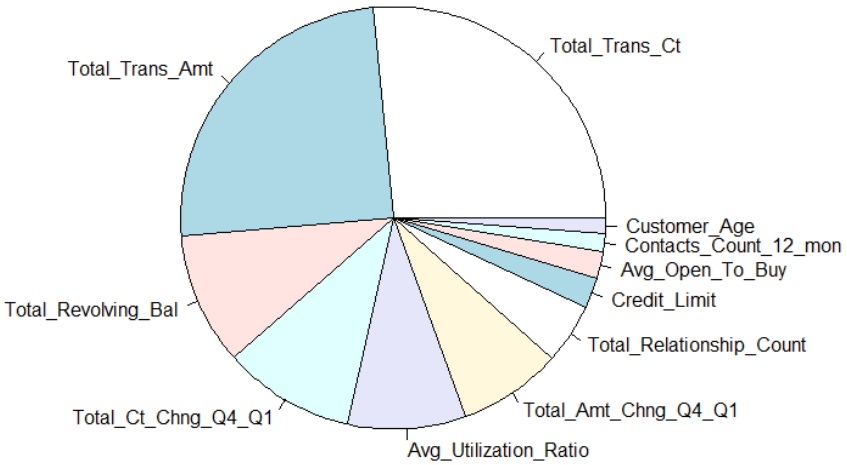


Figure 24 shows the percentage of importance of the descriptors in the construction of the tree.

6.3. Random Forest

To further explore the capabilities of the classification trees, bagging with random forests has also been assessed and compared. In our case, the best results were found by using 50 trees trying 4 variables at each split. If we increased the number of trees, the accuracy did not improve, so we decided to not overcomplicate the modeling process. The overview of the best model found, over the test dataset can be seen below:

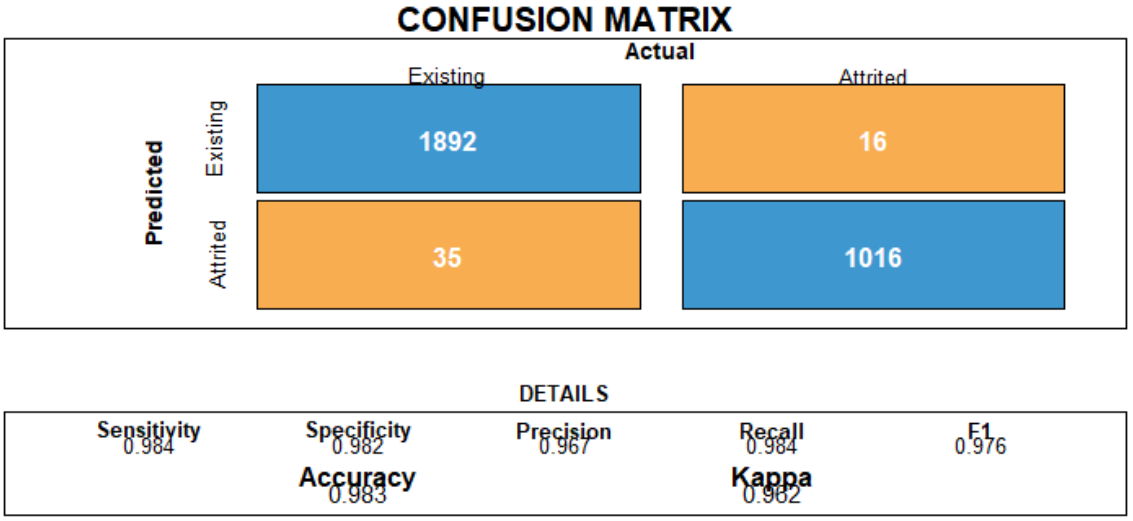


Figure 25 shows an overview of the predictive power of the Random Forest model, over the ***test*** *dataset.*

It can be seen that there has been an accuracy increase of 2% on the test dataset, bringing the accuracy up to 98%, which is a satisfactory result.

Regarding the more important variables, there has not been a change in the ones that are more present in the creation. However, as now in the creation of the trees random samples of the variables and observations are taken, some variables are now present in the creation of some trees but are obviously underrepresented.

6.4. XGBoost

In order to continue exploring the effectiveness of classification trees, gradient boosting with XGBoost was assessed and compared to previous models. XGBoost is a robust ensemble ML algorithm that doesn't require much preprocessing of data, apart from the one-hot encoding of our nominal attributes, and the conversion of our input data frames to matrices. Namely Dmatrix (dense matrix) objects, which are internal data structures used by xgboost that are optimized for both memory efficiency and training speed.

The xgboost model with the highest accuracy had 97.5%. It was fit by using default values for all hyperparameters except "max.depth" = 3, "nround" = 60, and "scale_pos_weight" = 2. However, It was not the best model obtained. Given the unbalanced nature of our dataset, the model mostly did a great job predicting existing customers but not attrited customers. The model shown below in *Figure 26* with 96.3% accuracy does a better job at predicting attrited

customers at the expense of correctly predicting existing customers, while making sure not to over-fit.

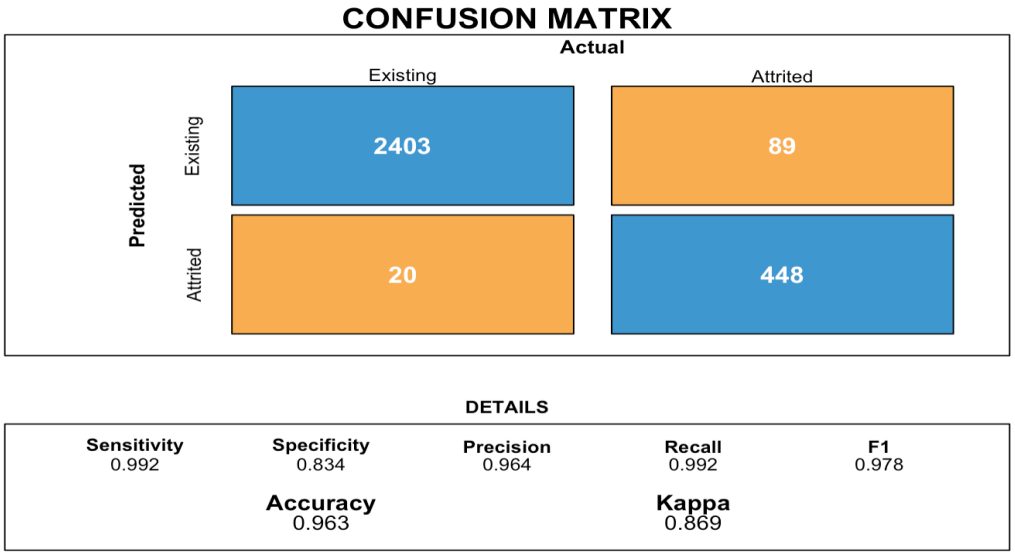


Figure 26 shows the confusion matrix and metrics of the best overall XGBoost model.

This model was obtained by using using default values for all hyperparameters except “max.depth” = 4, “nround” = 50 with “early_stopping_rounds” = 3, and “scale_pos_weight” = 5. Still, it does not outperform the previously computed Random Forest model. However, after using gridsearchCV to tune the xgboost hyperparameters with the aim to maximize the accuracy of predicted attrited customers, we were able to compute a model, referenced below in *Figure 27*, that correctly predicts attrited customers at 98.12% accuracy. This is done at the expense of correctly predicting Existing Customers, whose accuracy has lowered to 94.12%. The model is fit by using default values for all hyperparameters except “max.depth” = 3, “nround” = 70 with “early_stopping_rounds” = 3, “gamma” = 0.25, “lambda” = 10, and “scale_pos_weight” = 15. The resulting tree for this xgboost model is shown below in *Figure 27*. Its confusion matrix is also shown below in *Figure 28*.

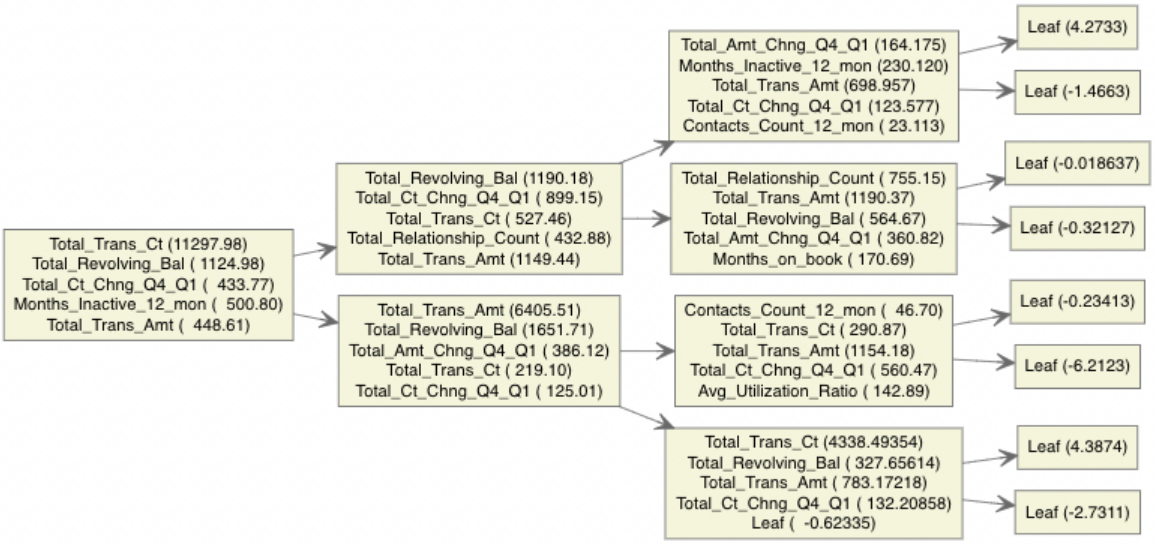


Figure 27 shows the resulting tree of the XGBoost model referenced above

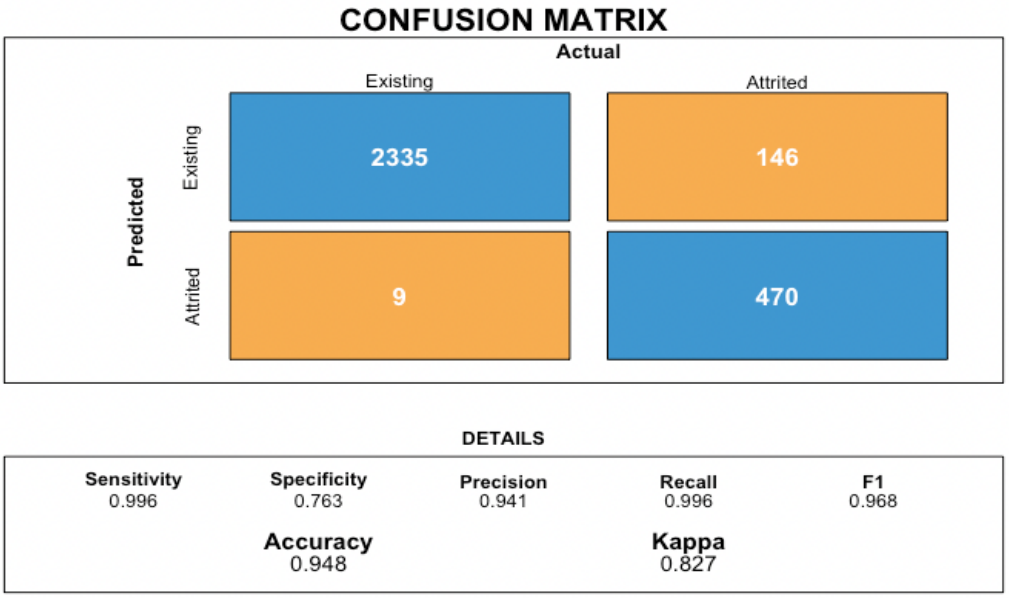


Figure 27 shows the confusion matrix and metrics of the best XGBoost model at correctly predicting attrited customers.

There is a case to be made that correctly predicting attrited customers would be of more value to our organization, even at the expense of incorrectly predicting existing customers, since customers who leave also take their money with them. Knowing which customers are about to leave represents an opportunity to keep the customer's business and continue making money. Because of that, even though the above model has a lower overall accuracy than all previous classification models, it still carries value. It makes more errors in prediction, but they are better errors.

6.5. Interpretation of the results

As commented previously, the rules created by the classification tree can be used to extract knowledge about in which situations the bank should act and which strategies should they implement to reduce customer attrition. Looking at the rules for the best single tree model, some suggestions can be made:

- Be aware when activity levels are low: it has been seen that customers with low levels of activity (transaction _count<45) are the ones with the highest probability to leave the bank. Hence, the bank should try to engage these customers to use their credit cards more.
- Keep track of people who although they do not use their credit cards frequently, go to the bank more than the average customer. This situation may be due to the client being unsatisfied.

It is important to also remark that, as explained in the outlier detection part, a group of 261 observations was considered to be outliers and to belong to a separate group. For these observations, a model was also trained on them. In this case, the obtained accuracy was around 84% for the test dataset, which can be explained by the fact that the number of observations is significantly lower.

7. Association Rules

Our dataset contains a mixture of categorical and numeric variables and needs some preparations before it can be transformed into a transaction subset for association mining.

We performed discretization of numerical variables, grouping categories of the factors created according to the shared characteristics presented among their levels.

Firstly we decided to remove some variables such us: "Dependent_count", “Months_on_book", ”Total_Relationship_Count", ”Months_Inactive_12_mon", "Contacts_Count_12_mon”, "Total_Revolving_Bal”, ”Avg_Open_To_Buy" , ”Total_Amt_Chng_Q4_Q1”, ”Total_Trans_Ct”, “Total_Ct_Chng_Q4_Q1”, ”Avg_Utilization_Ratio", since according to the previous analyses, they didn’t contribute most value and give the greatest explanatory power to the model.

Next, we needed to map the remaining variables (“Customer_Age”, “f.customer_age”, “Gender”, “Education_Level”, “Marital_Status”, “Income_ Category”, “Card_Category”, “Credit_Limit” and “Total_Trans_Amt”) to ordinal attributes by building suitable categories.
For the first one, we divide the attributes Customer_Age into three suitable categories using knowledge about the life cycle of the people: Young adults, Middle-aged adults, and Older adults. For Income_Category, we have created three levels according to low, medium, and high income, and lastly for Total_Trans_Amt according to the number of transactions performed in low, medium, and high corresponding to customers.

To consider which items are important in the data set we can use the ItemFrequencyPlot. To reduce the number of items, we only plot the item frequency for items with support greater than 10%.

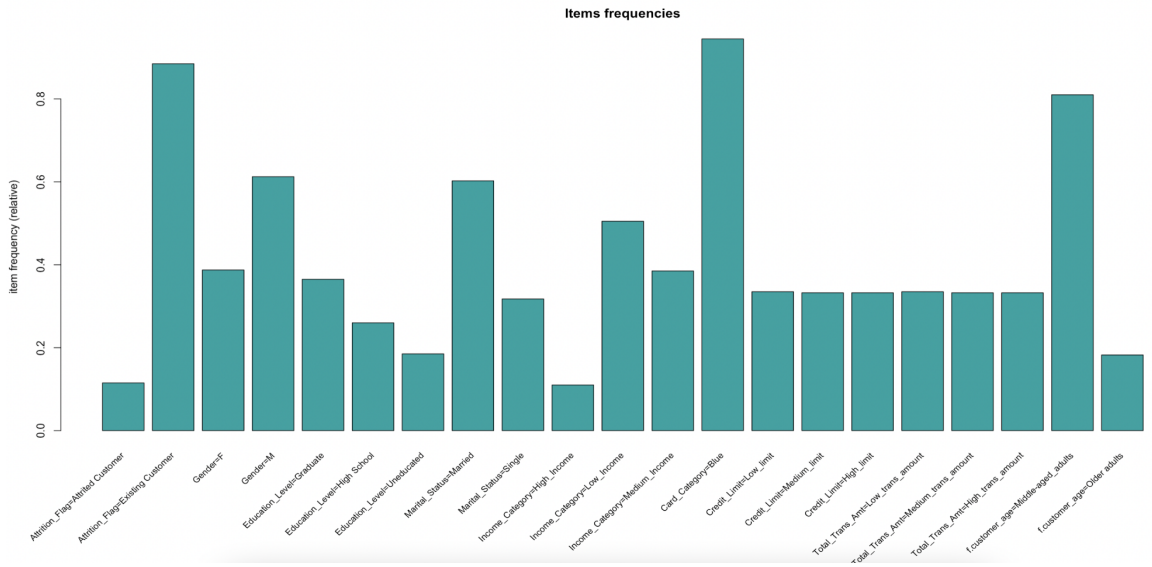


Figure 29

The summary of the transaction dataset gives a rough overview showing the most frequent items, the length distribution of the transactions, and the extended item information which shows which variable and which value were used to apply the association rules into our data.

After that, we used the function apriori in order to find all rules (the default association type for apriori with minimum support of 1% and a confidence of 0.6).

The function prints the used parameters. Apart from the specified minimum support and minimum confidence, all parameters have the default values. It is important to note that with parameter maxlen, the maximum size of mined frequent itemsets, is by default restricted to 5.

The result of the mining algorithm is a set of 28.539 rules. For an overview, through the summary, we can observe the number of rules, the most frequent items contained on the left-hand side and the right-hand side, and their respective length distributions and summary statistics for the quality measures returned by the mining algorithm.

To analyze these rules, for example, a subset can be used to produce separate subsets of rules for each item that resulted from the variable Attrition_Flag in the right-hand side of the rule. At the same time we require that the lift measure exceeds 1.0

7.1. Created rules

For the analysis of the results obtained, we focused on the lift parameter, since lift is a measure for the deviation of the rule from the model of statistical independence of the rule body and rule head. Taking into account the resulting lift value is greater than 1, that indicates that the rule body and the rule head appear more often together than expected; this means that the occurrence of the rule body has a positive effect on the occurrence of the rule head.

Existing customer: From the output above we can observe that the three most redundant rules evidence that exists a direct relation between clients with a Doctorate and the attrition flag, being single, having Medium or High transaction amount tend to remain their bank accounts open.

```
> inspect(head(rules_existing_customer, n = 3, by = "confidence"))
    lhs                                          rhs                                 support confidence coverage     lift count
[1] {Education_Level=Doctorate,
     Marital_Status=Single}                  => {Attrition_Flag=Existing Customer}  0.0100          1   0.0100 1.129944     4
[2] {Education_Level=Doctorate,
     Total_Trans_Amt=Medium_trans_amount} => {Attrition_Flag=Existing Customer}  0.0125          1   0.0125 1.129944     5
[3] {Education_Level=Doctorate,
     Total_Trans_Amt=High_trans_amount}   => {Attrition_Flag=Existing Customer}  0.0150          1   0.0150 1.129944     6
```

Attrited customer: Taking into account that the first three most redundant rules related to the attrited customers show some interactions between Marital Status, Credit Limit, Total Transactions Amount and Customer Age, we can conclude that clients most likely to close their bank accounts are married older adults with a low credit limit.

8. Summary and Conclusions

We worked with an unbalanced dataset with few missing values in three of the attributes, which were imputed trying to maintain the marginal distributions of all of them.

The original data presented a group of multivariate outliers, which were removed from the dataset and treated separately as they had a similar profile.

From performing PCA we obtain a first approach to detect potential clusters according to positive correlation between variables “Customer Age” and “Months on book”, and negative correlation between “Credit Limit”, “Avg Utilization Ratio” and “Total_Revolving_Bal”. The analysis of the first two dimensions also allowed us to detect not significant and potentially not useful features when creating the future models such as “Dependent count” and “Months Inactive 12 month”.
As for the results extracted from the Multi Correspondence Analysis are a bit confusing, as there is not a great representation of variability in the first dimensions. We see that the first dimension represents a variability of 12%, and the following dimensions all represent between 6 and 7% variability. For this reason, this analysis is not useful in our dataset because to reach a variability of 70% (as it would be correct) we would have to take many dimensions.
Regarding the Multi Factor Analysis, it has helped us to see how the variables behave within the dataset, also if there is any correlation with another, as well as which variables have a greater contribution in each dimension, but we have seen that for our dataset we cannot obtain very relevant information with any of the dimension reduction methods applied.

Dealing with Linear Discriminant Analysis (LDA), we have to check first if our data satisfy the three main conditions but we conclude that our dataset does not fulfill those conditions. Therefore, the results obtained won’t be robust. Despite this setback, we reached a predictive model as a first approach, obtaining an accuracy of 88%. As a second attempt, we also applied Quadratic Discriminant Analysis (QDA) taking into account that this method works better with big samples (like ours). However, the final accuracy of the model ended up very similar to the one of LDA (88%).

Classification tree models were assessed in order to predict the outcome. Before training the different models proposed, we found that our data was unbalanced. The technique used after trying different approaches is the ROSE method which oversampled the positive outcome to 35% and under-sampled the negative outcome to 65%. Next, we trained three simple trees from different packages finding the optimal cost penalty. The results were satisfactory as we obtained an accuracy of 0.962 and we didn’t find the overfitting problem as the test and train results are quite similar. Moreover, another metric to take into account was recall, because it is better to have a false positive than a false negative, the bank can offer better conditions to an already happy customer than not interacting with a customer that is unsatisfied. Therefore, a random forest using 50 trees trying 4 variables at each split was used and compared with the best model so far. The result was more satisfactory than before as the accuracy increased up to 98%. All the models trained arrive to the same conclusion that the features regarding amounts of money and interactions of the bank are the most important.
With the trees obtained we can extract new knowledge and information about the customers of the bank. If the customers have a low level of activity and if they go to bank, they are most likely to leave the bank.

According to association rules, we have previously prepared our dataset because it contained both numerical and categorical variables, so we could transform it into transactions, after that we grouped categories of the factors according to the shared characteristics among their levels. ItemFrequencyPlot, was very helpful for inspecting the item frequency distribution for objects and determining which variables were most relevant.

We have also taken into account the lift parameter, it took a value greater than one, it means that the rule body and the rule head appear more often together than expected; so the happening of the rule body has a positive effect on the occurrence of the rule head.

In this way we were able to obtain the most redundant rules of those customers who would keep their products in force and observe who would be the customers most likely to close their accounts at the bank.

As a final conclusion, it can be said that despite the fact that dimension analysis results were far from being promising, the final model created had a very high predictive power. Hence, the usefulness of the Classification Trees, and more concretely Random Forest, could be clearly seen.