Predicting credit card owners' churn

Yijiang Fan, Filippo Michelis, Tommaso Perniola, Luisa Wall

April 2021

Abstract

This work presents the results obtained by applying different statistical learning's methods to a notorious problem in machine learning and statistical learning: attrition of banks' clients. To do so we have implemented different technique trying to assess two of the major problems that arise in this situations: unbalanced datasets and lack of observations.

1 Introduction

For our analysis, we used a dataset (retrieved from Kaggle.com) that contains information about credit card owners. The information available in the dataset regards demographic and socioeconomic characteristics of the costumers (i.e. education level, income, gender, age, marital status..), and information about their relationship with the bank and the credit card services it offers (i.e. credit card category, n° of months of inactivity, credit limit on the credit card, average card utilisation ratio, total transaction amount, etc.) The aim of the analysis is to predict which of the bank clients included in the dataset will attrite. Attrition happens when costumers leave the bank's credit card services. In the banking sector, using costumer information to predict attrition is useful to proactively target the customers, in order to provide them better services and persuade them to stay (Verbeke, 2011). In the context of costumer relationship management, preventing churn is considered important, since attracting new costumers is less profitable than preventing current costumers to abandon the company (De Caigny, 2018). In the service industry, establishing a long term relationship with clients, allows the firm to focus on their requirements and needs; in this process, identifying a "valuable costumer base", which consists in identifying the costumers that bring more value to the firm, is crucial (Ganesh, 2000). On this basis, after pre-processing our dataset, we tried to identify the best prediction algorithm for credit card owners' churn.

In the first part are presented the methods used to conduct our analysis. The second part presents the results obtained and offer an interpretation. The third part summarise our analysis underlining the most important results and offering a perspective on further methodologies applicable in further researches.

2 Methods

2.1 Data and preprocessing

In our dataset, we had 10127 observations and 19 variables, 14 continuous and five categorical. We decided to transform 2 of the categorical variables, *Education_Level*, and *Income_Category*, into continuous ones, since they could be more informative. The variable *Income_Category* took on values like *"50.000$ to 70.000$"*. To transform it into a continuous variable we simply took the mid point of the class it belonged to. The variable *Education_Level* took on instead values like *"Uneducated, High School, Graduate, ..."* so we have replaced each category with the corresponding education years.

We had missing values for 3 variables, *Education_Level*, *Marital_Status* and *Income_Category*, so we decided to impute these missing value with k-Nearest Neighbours. We choose 10 as value for k. We choose this technique to avoid the bias that could derive from removing missing value, since the fact that a customer did not share that information could be particularly informative in view of predicting attrition. Not giving that information could indeed be correlated with other variables. Anyway, even when we remove the missing values, the proportion of attrition is the same.

As we can see in Figure 1, most of the variables are poorly correlated with each other.

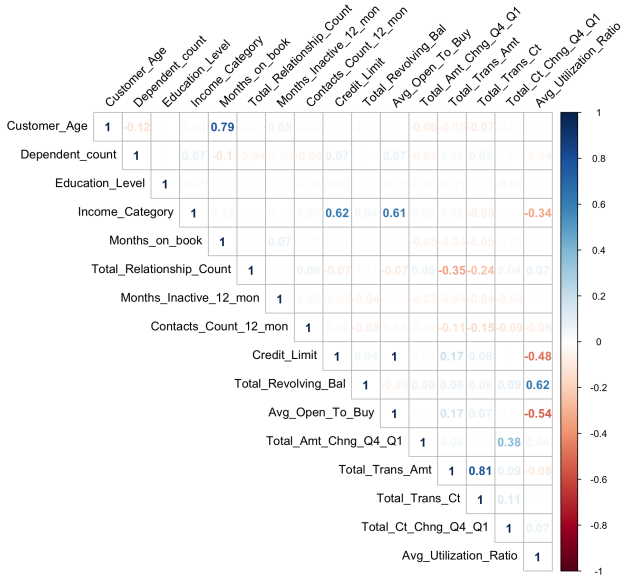

Figure 1: Correlation matrix

The dataset obtained after these first steps of preprocessing still presented an issue. The dataset is unbalanced, in fact Non attrited customers (flagged as Y=0) are the majority, with 8500 observations, compared to 1627 observations of Attrited customers. This imbalance is problematic because the model tends to focus on the prevalent class and to ignore the rare events (Menardi, 2014), the estimation of the classification and the model accuracy computed on the unbalanced dataset is shown in the section Results for comparison. In order to overcome this problem we tried different approaches and we created three training datasets on which we trained the algorithms.

In dataset 1 we got a balanced sample from the original dataset, creating a training sample with 800 random selected Attrited customers and 800 random selected Non attrited customers. The remaining observations (7700 Non attrited customers and 827 Attrited customers) were included in the test sample. With this approach we created a balanced training set, which leads to a better model estimation, and we tested the model on a unbalanced test set which reproduces the distribution of attrited and non attried customers in real life. The training dataset obtained with this approach is smaller in dimension compared o the test set, leading to a higher bias.

The alternative approach is the oversampling method, which is duplicating the observations from the minority class (Menardi, 2014). Before performing this preprocessing method we divided the original dataset in traing set(85% of the observations) and test set. On the training set we performed a simple oversampling, by adding five times the observations from the minority class, in order to obtain a new training set (dataset 2) with 7225 Non attrited customers and 6915 Attrited customers. A more sophisticated method is to add a random error to the observations duplicated (multiplied five times in our case), for each variable we chose a error term with a variance equal to 1/10 of the variance of the variable itself. The training set created with this method is indicated as dataset 3 in the following discussion. It is noticeable that adding a random error gives more variance to the data, but it does not take into account the correlation among the variables. The negative effect of this simplification is expected to be relatively small, since the variables are poorly correlated.

2.2 Statistical methodology

To better understand our dataset, we performed two unsupervised learning techniques, Principal Component Analysis and clustering. Before that, we scaled our data since the variables are measured in different units. PCA allows us to derive fewer representative variables that collectively explain most of the variability in our original dataset. Since, as we said, variables are poorly correlated with each other, we do not expect great results in applying this technique. We use

clustering instead to try to identify homogeneous subgroups in observations. In particular, through K-means clustering we try to partition our dataset into k different non-overlapping clusters. At the basis of K-means clustering, there is the idea the within-cluster variation, expressed as squared Euclidean distance, should be as small as possible. To evaluate the clusters obtained we used silhouette widths, that is a measure of how similar an object is to its cluster compared to other clusters.

In order to perform predictive analysis on the observations of our data set, we used different classification techniques. A model of logistic regression was used to predict the qualitative responses "Non costumer attrition" (encoded as Y=0) and "Customer attrition" (encoded as Y=1). The model is based on the logistic function, that gives outputs between 0 and 1 for all values of the 19 predictors (X variables). According to the maximum likelihood method, the beta coefficients for the predictors are estimated, based on observations in the training data, so that the predicted probability of attrition for each individual corresponds as closely as possible to the individual's attrition status (0 or 1); in mathematical terms, the beta estimates are chosen to maximise the likelihood function.

We performed also a non parametric classification of the clients, using the method of K nearest neighbours (KNN). The method identifies the K points in the training data that are closest to a point x_0, then it estimates the probability that given $X = x_0$, Y belongs to class j as the fraction of points in the neighbourhood whose response values equal j. Finally, KNN classifies the test observation x_0 to the class with the largest probability (James, G., 2013).

We discarded the methods of Linear discriminant analysis and Quadratic discriminant analysis, since a critical number of important variables are not normal.

To evaluate the performance of the classification models described above, a 10-fold cross validation approach was implemented. This method consists of randomly dividing the set of observations into 10 folds of equal size. The first fold is treated as a validation set and the estimated classification models are fit on the k-1 folds. The error rate (which corresponds to the number of missclassified observations) is computed on the observations in the fold treated as a validation set. This procedure is repeated 10 times, each time using a different group of observations as a test set. The confusion matrix computed to estimate the specificity (rate of responses Y=0 correctly predicted), the sensitivity (rate of responses Y=1 correctly predicted) and the accuracy (rate of correctly classified observations) of the logistic regression, performed on unbalanced data set, is based on the results of the 10-fold cross validation. The confusion matrices displayed for the subsequently presented classification models (logistic regression on balanced data sets, KNN classification, ridge and lasso regression models) are estimated by testing the models, which are

trained on the balanced data sets, on the unbalanced data set. This approach is chosen, since the aim of the analysis is performing predictive analysis on the original data group of observations, which is naturally unbalanced. In performing KNN classification, the cross validation method is used to tune the parameter K.

Finally, we tried to improve our model through shrinkage methods. To do so we estimated Ridge and Lasso. In both cases we estimated the tuning parameter λ with cross validation, finding the λ which minimised the error rate: we first choose a grid of tuning parameters, than, due to cross validation, we compute the error associated to each of them and, lastly, we select the one with minimum cross validation error. The models are then refitted with the selected value of the tuning parameter. Performing lasso was the only computationally feasible way to operate a feature selection, given the the number of variables and the information being dispersed among a high number of them, as enlighten by the correlation matrix and the PCA.

3 Results and Discussion

The following images represent the results of PCA.

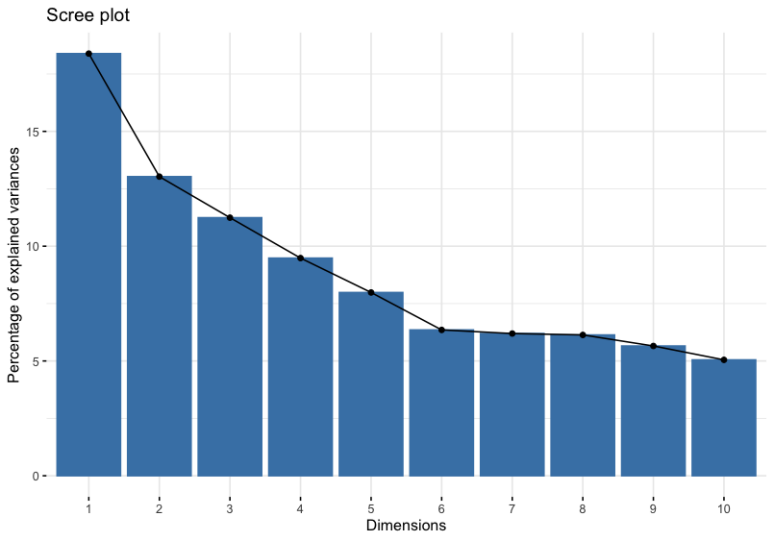

Figure 2: Percentage of variance explained by the principal components

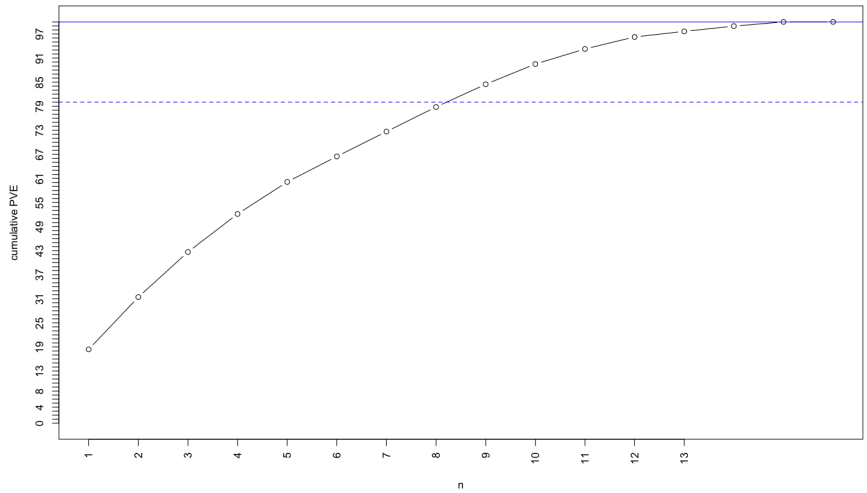

Figure 3: Cumulative percentage of variance explained by the principal components

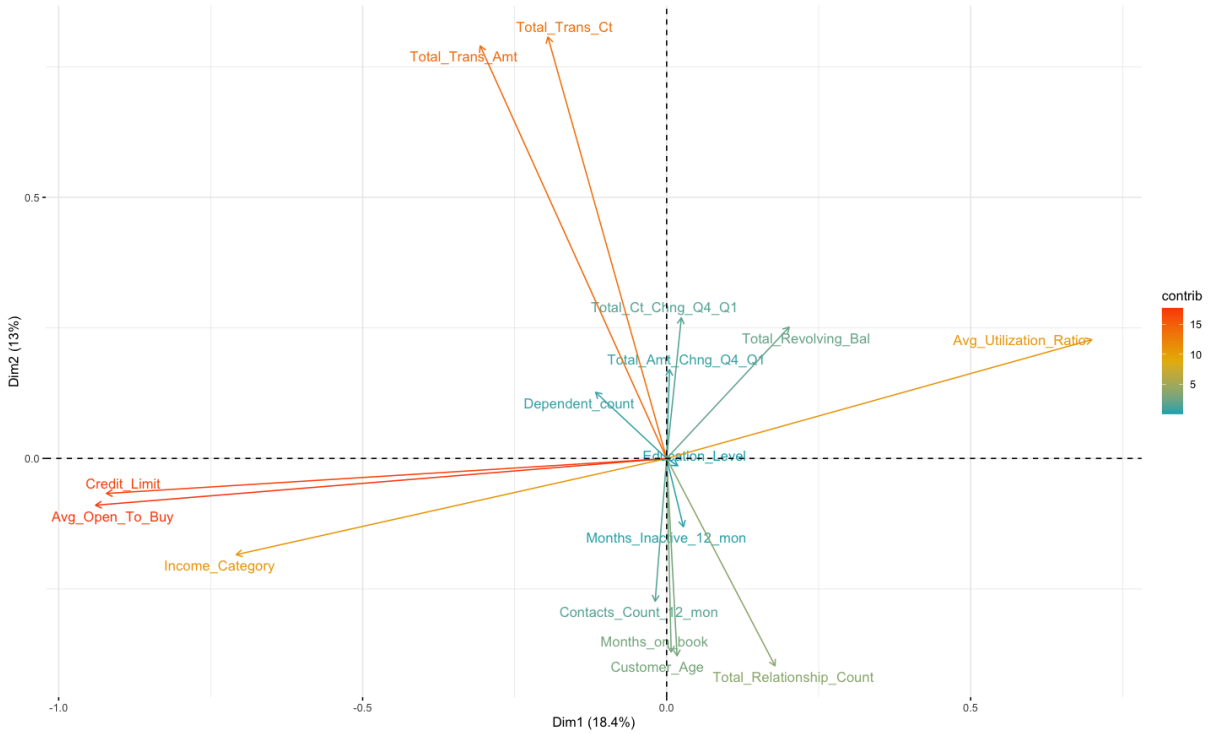

Figure 4: First two principal component loading vectors

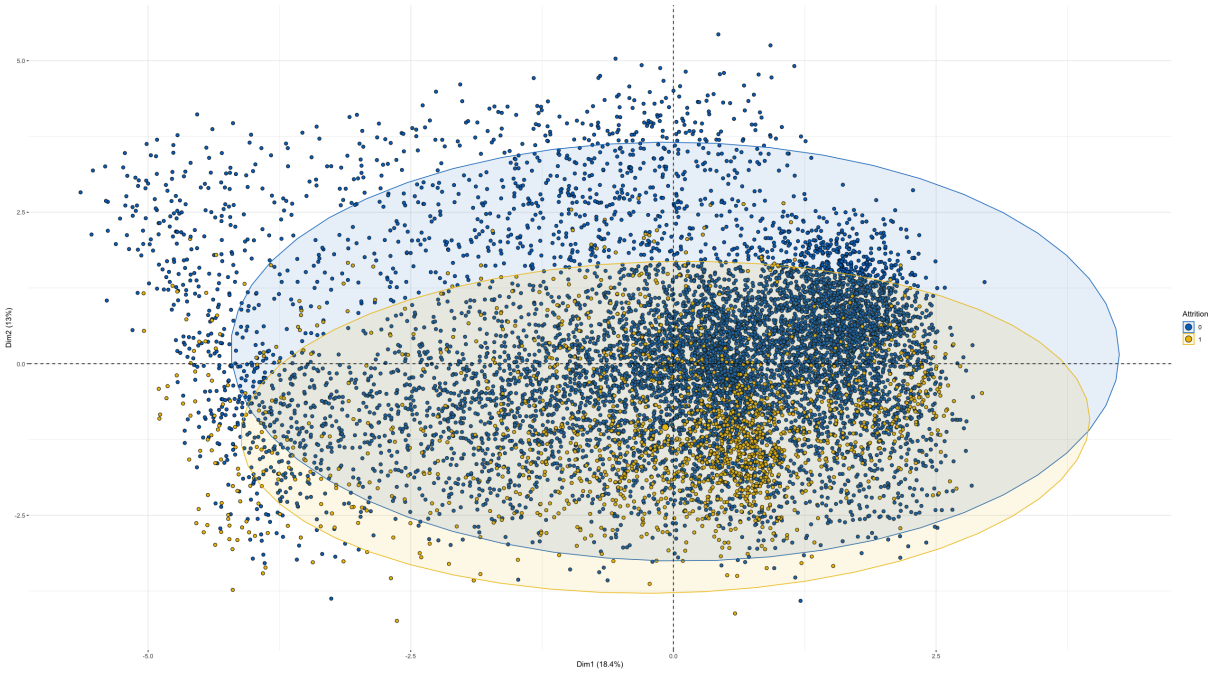

Figure 5: Costumers who committed attrition in yellow vs. costumers who did not in blue

In Figure 2 and 3 we can see the percentage of variance explained by the principal components. As we said, variables are poorly correlated with each other and this could explain the low percentage of variance explained by the first two principal component. In Figure 4 the first two principal component loading vectors are depicted. Finally, we tried to see if the customers who attrited were graphically identifiable. In Figure 5 the blue points represent indeed the customers who continued to use their credit card while the yellows ones are those who defected. We have also drawn a 95% confidence interval ellipse. As we can see, we are not able to individuate two definite clusters. Nevertheless, we tried to apply K-means clustering. In Figure 6, indicating the total within sum of squares of each cluster, we can see that no cluster number drastically improve the total within sum of squares. This could imply that there are no well-defined clusters. This seems to be confirmed by Figures 8 and 9 where are represented the clusters obtained by setting k = 2 (once again with the idea of trying to distinguish the customers who attrited from those who did not) and k = 3 (as suggested by the Hartigan Index and by the average silhouette method, shown Figure 7). In Figure 13 (in Appendix) are instead represented the average silhouette width of both clusters, which are very low. We also tried to perform Hierarchical clustering but without better results. Furthermore, once again, through clustering, we are unable to distinguish customers who attrited from those who did not. It is probably due to the fact that our data are sparse.

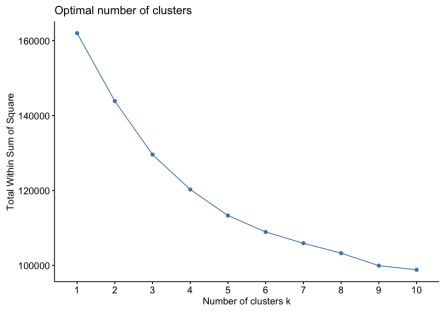

Figure 6: Optimal number of clusters

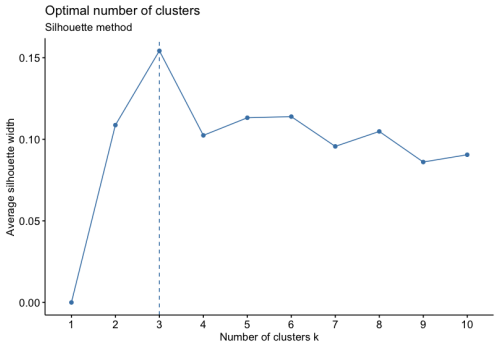

Figure 7: Optimal number of clusters, silhouette method

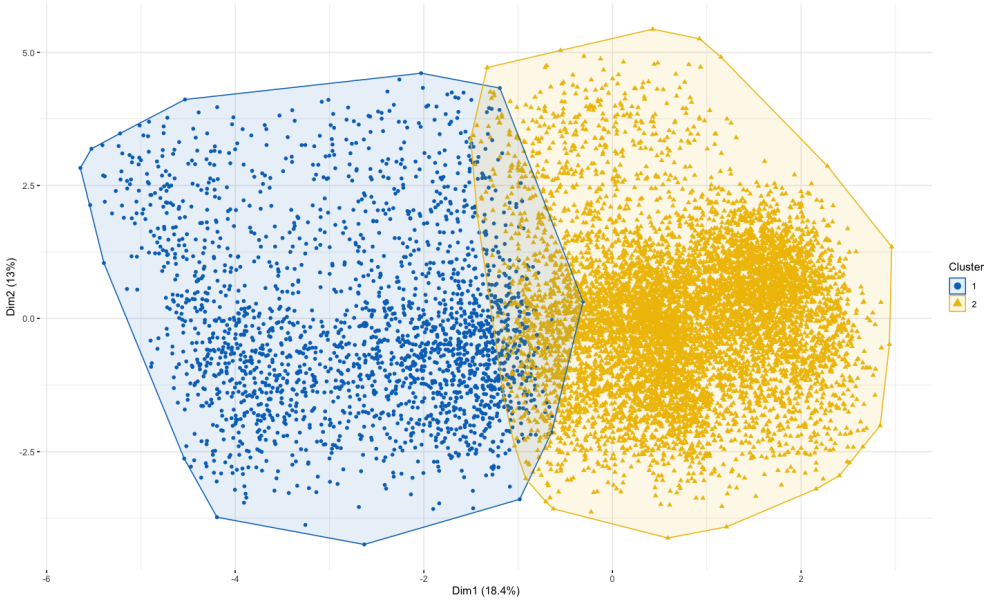

Figure 8: K-means clustering, k = 2

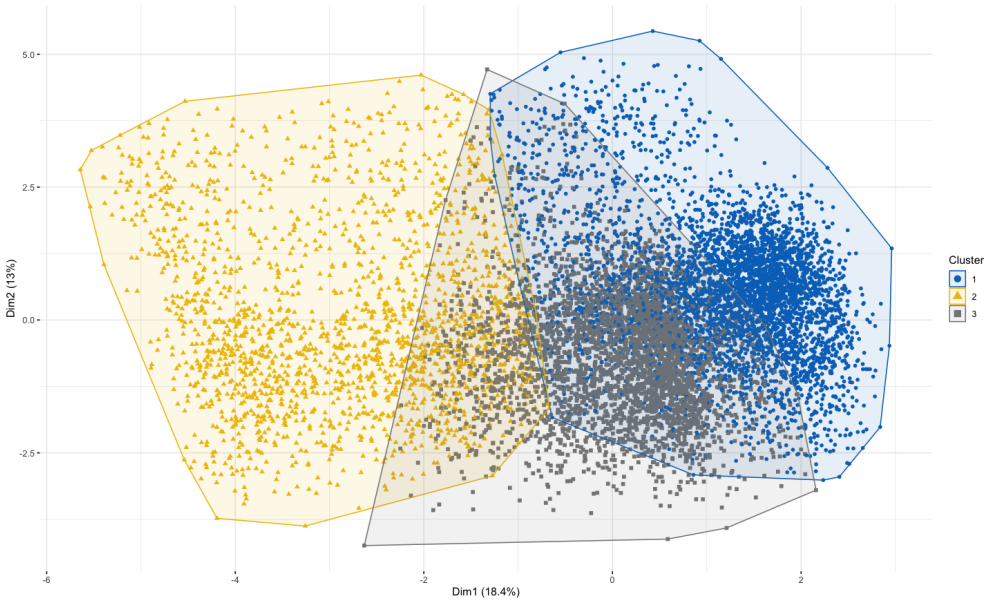

Figure 9: K-means clustering, k = 3

We are introducing the results of the classification models, estimated on different versions of the original dataset, starting from the logistic regression model. For this model we also display (in the section Appendix) the coefficients of the regression, computed on dataset 2, since the coefficient computed on other datasets are very similar and inference is not the aim of our work. In order to predict the attrition of customers, first we applied logit model to the unbalanced dataset. The accuracy results are shown in the 10 fold confusion matrix in table 1.

Table 1: Cross validated (10 fold) confusion matrix, for the logistic regression estimated on the original unbalanced dataset

	Reference	
Prediction	0	1
0	8211	679
1	289	948

While the overall accuracy and the specificity of the logistic regression model are reasonably good (respectively, 0.9037 and 0.966), the sensitivity is less good (0.5826675). The sensitivity, which indicates the rate of positive responses (Y=1, "Costumer attrition") correctly predicted, improves when the logistic regression is fitted on the three differently balanced datasets. Table 2 displays the confusion matrix obtained, by confronting the predictions of the logistic regression model trained on dataset 1 with the observations in the test sample (unbalanced dataset).

Table 2: Confusion matrix for logistic regression model (train: balanced dataset 1; test: unbalanced test sample)

	Reference	
Prediction	0	1
0	6447	121
1	1253	706

The accuracy for the model trained on dataset 1 (0.8502) slightly decreased, reasonably due to the reduced dimension of dataset 1, in comparison to the original unbalanced dataset. The specificity for the model is 0.8343; the sensitivity (0.8519) significantly improved. Table 3 and 4 show the confusion matrices computed in order to test the performance of the logistic regression models, respectively fitted on dataset 2 and dataset 3.

The values of accuracy and sensitivity are, respectively, 0.8578 and 0.8689, for the model trained on dataset 2, and 0.83648 and 0.86885, for the model trained on dataset 3. These accuracy and sensitivity results confirm the hypothesis that the classification model trained on an unbalanced dataset, where the positive responses are under sampled, is less powerful in predicting the

Table 3: Confusion matrix for logistic regression model (train: balanced dataset 2; test: unbalanced test sample)

	Reference	
Prediction	0	1
0	1091	32
1	184	212

Table 4: Confusion matrix for logistic regression model (train: balanced dataset 3; test: unbalanced test sample)

	Reference	
Prediction	0	1
0	1089	32
1	186	212

positive responses, in comparison to the same model trained on balanced datasets.

Assessing the accuracy of the KNN classification on test data we obtain reasonably good accuracy using all of the our balanced datasets, from an accuracy of 0.77 on dataset 1 (Table 5), to 0.87 and 0.84 on dataset 2 (Table 6) and 3 (Table 7). On the other hand, compared to the sensitivity computed previously on logit classification, the sensitivity of the KNN classification method is lower. In fact the latter presents a sensitivity of 0.78 on dataset 1, 0.51 on dataset 2 and 0.21 on dataset 3, compared to 0.85, 0.84 and 0.87 of the first method.

It is noticeable the decrease in sensitivity when performing KNN classification with dataset 2 and 3. This phenomenon is explained by the fact that the two training datasets are built with oversampling, therefore in the neighbourhood of a random data point there are multiple copies (5 in this case) of the point itself, if the point has Y=1, while in the neighbourhood of points with Y=0 there are no copies; this fact is distortive for the training of the classification model and it is reflected in the low sensitivity of the model.

Table 5: Confusion matrix for knn model (train: balanced dataset 1)

	Reference	
Prediction	0	1
0	5956	183
1	1744	644

Table 6: Confusion matrix for knn model (train: balanced dataset 2)

	Reference	
Prediction	0	1
0	1200	119
1	75	125

Table 7: Confusion matrix: Confusion matrix for knn model (train: balanced dataset 3)

	Reference	
Prediction	0	1
0	1227	193
1	48	51

The abovementioned values are computed with the optimal k value, tuned with a ten fold cross validation, the results of the tuning (with the optimal k highlighted) are shown in the subsequent figures:

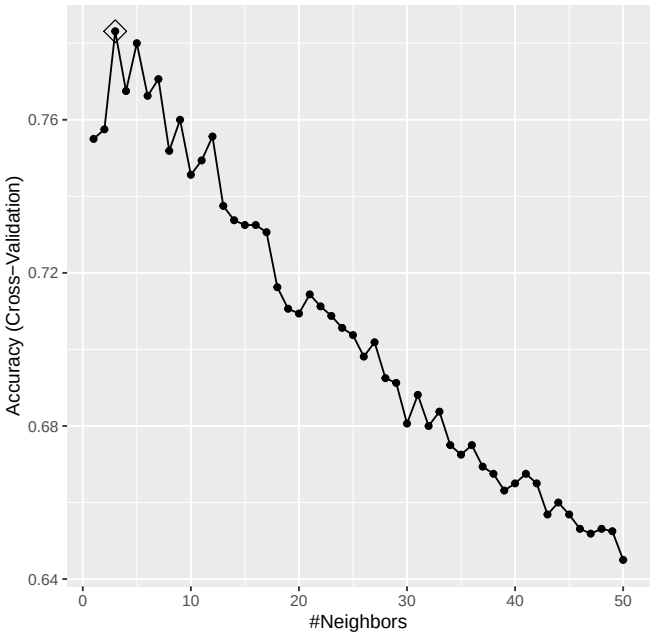

Figure 10: Tuning k (knn on dataset 1)

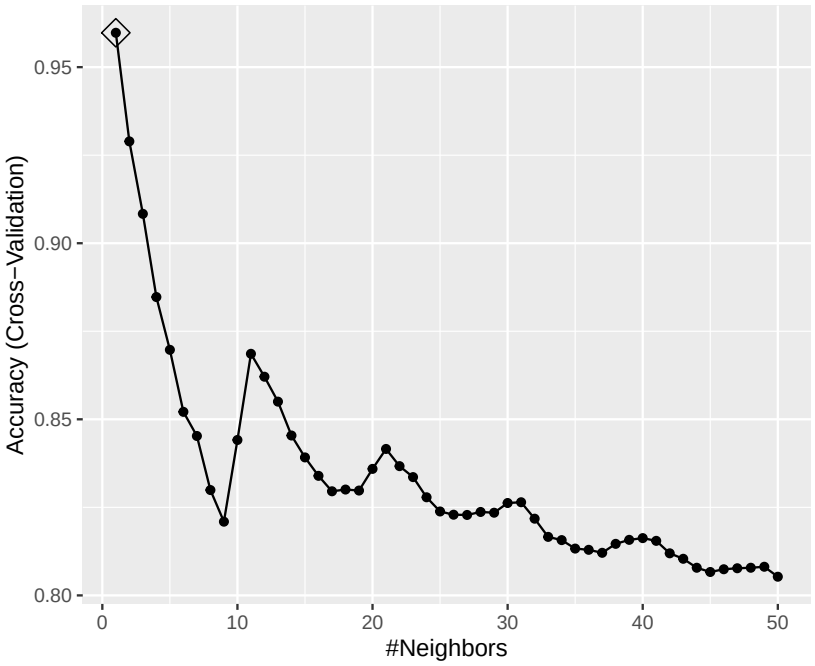

Figure 11: Tuning k (knn on dataset 2)

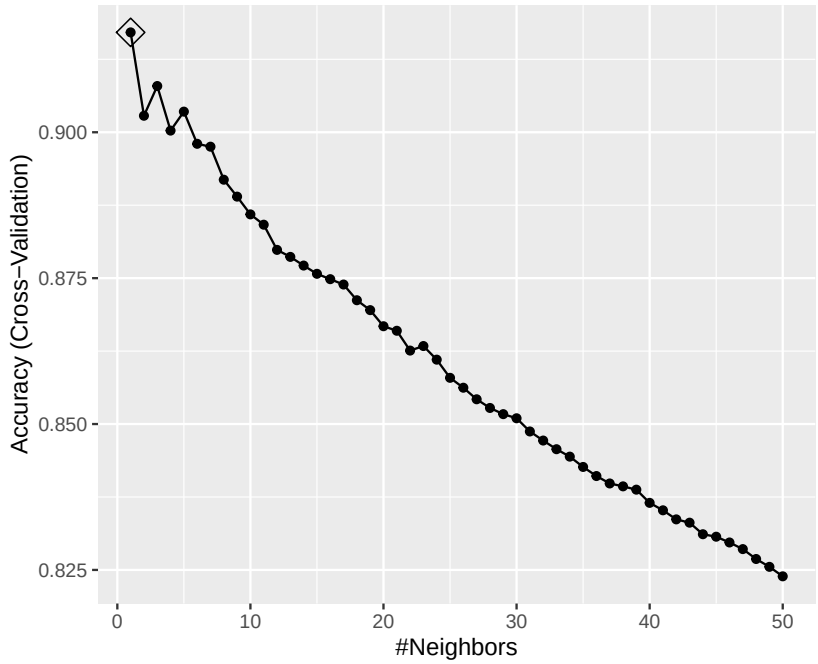

Figure 12: Tuning k (knn on dataset 2)

Performing ridge classification and lasso we obtained similar results in terms of accuracy and sensitivity, however none of them seem to improve over the multinomial logit's predictions: the accuracy is increased at the cost of a worst sensitivity. This results are displayed in Tables from 14 to 16. The lasso forces two of the coefficients estimates to zero. However, the model so obtained is not notably different from the one obtained by using all the covariates. The fact that the lasso led to a very similar subset of the variable can be interpreted as a demonstration of the sparsity of the dataset, which were already suggested by the PCA's results and the covariance matrix of the variables. As a result, the models do not lead to better predictions when compared to the standard logit's results. This can be seen when comparing the confusion matrices to the ones obtained for the other models (as the previous models they both have been estimated and cross-validated on different training sets):

Table 8: Confusion matrix for ridge regression model (train: balanced dataset 1)

	Reference	
Prediction	0	1
0	7011	223
1	689	604

Table 9: Confusion matrix for lasso regression model (train: balanced dataset 1)

	Reference	
Prediction	0	1
0	6804	170
1	896	657

Table 10: Confusion matrix for ridge regression model (train: balanced dataset 2)

	Reference	
Prediction	0	1
0	1191	62
1	84	182

Table 11: Confusion matrix for lasso regression model (train: balanced dataset 2)

	Reference	
Prediction	0	1
0	1151	51
1	124	193

Table 12: Confusion matrix for ridge regression model (train: balanced dataset 3)

	Reference	
Prediction	0	1
0	1191	62
1	84	182

Table 13: Confusion matrix for lasso regression model (train: balanced dataset 3)

	Reference	
Prediction	0	1
0	1152	51
1	123	193

Notably, the models do not lead to better predictions when compared to the standard logit's results.

Table 14: Models computed on dataset 1

	Logit	KNN	Lasso	Ridge
Accuracy	0.83886	0.77401	0.87499	0.89304
Sensitivity	0.85369	0.77872	0.79443	0.73035

Table 15: Models computed on dataset 2

	Logit	KNN	Lasso	Ridge
Accuracy	0.85780	0.87228	0.88479	0.90388
Sensitivity	0.86885	0.51230	0.79098	0.74590

Table 16: Models computed on dataset 3

	Logit	KNN	Lasso	Ridge
Accuracy	0.85648	0.84134	0.88545	0.93465
Sensitivity	0.86885	0.20901	0.79098	0.74590

4 Conclusion and Further research

We assessed the problem of an unbalanced dataset exploiting different techniques from the literature. Firstly, we tried to balance our training set by sampling so that the numbers of observation exhibiting the two values were equivalent. This unfortunately led to a small training set due to the lack of positive response's observations. To overcome this disadvantage, we tried deriving an artificial dataset from the original one. That has been done by adding noise to a balanced sample of observations. Notably, we did not assess the problem of correlation among variables while adding the noise, as we drew each observation as if the value corresponding to each variable were drawn from a gaussian distribution, centred in the real value of the observation and with standard error equal to the ten percent of the standard error of the variable. This might of course lead to biased estimations, however, given the lack of correlation among the variables, we managed to obtain consistent results even in the noised dataset. Hence, we overcame both the problem an unbalance dataset and the problem of a small training set.

As discussed, further attempts to better the predictive ability of the model through subset selection and shrinkage have proven inconclusive. The lasso and ridge classification were not able to increase the accuracy and the sensitivity. That could be interpreted as a result of the correlation among variables: the information needed for the classification laid in almost every variable of the dataset, so it was not possible to estimate a subset of variables which could give lower error rates. The fact that the lasso led to a very similar subset of the variable can be interpreted as a demonstration of the sparsity of the dataset, which were already suggested by the PCA's results and the covariance matrix of the variables. As a result, the models do not lead to better predictions when compared to the standard logit's results.

In further researches alternative methods for balancing can be used, we can generate an artificial dataset with Synthetic Minority Oversampling Technique (SMOTE) method, randomly choosing points that lie on the line connecting the rare observation to one of its nearest neighbours in the feature space; or with Random Over Sampling Examples(ROSE), a smoothed bootstrap approach (Menardi, 2014).

In order to improve the oversampling method already used, in the process of adding a random error to the duplicated observations, we can choose different values for the standard deviation of the random error, in order to tune this parameter.

5 Appendix

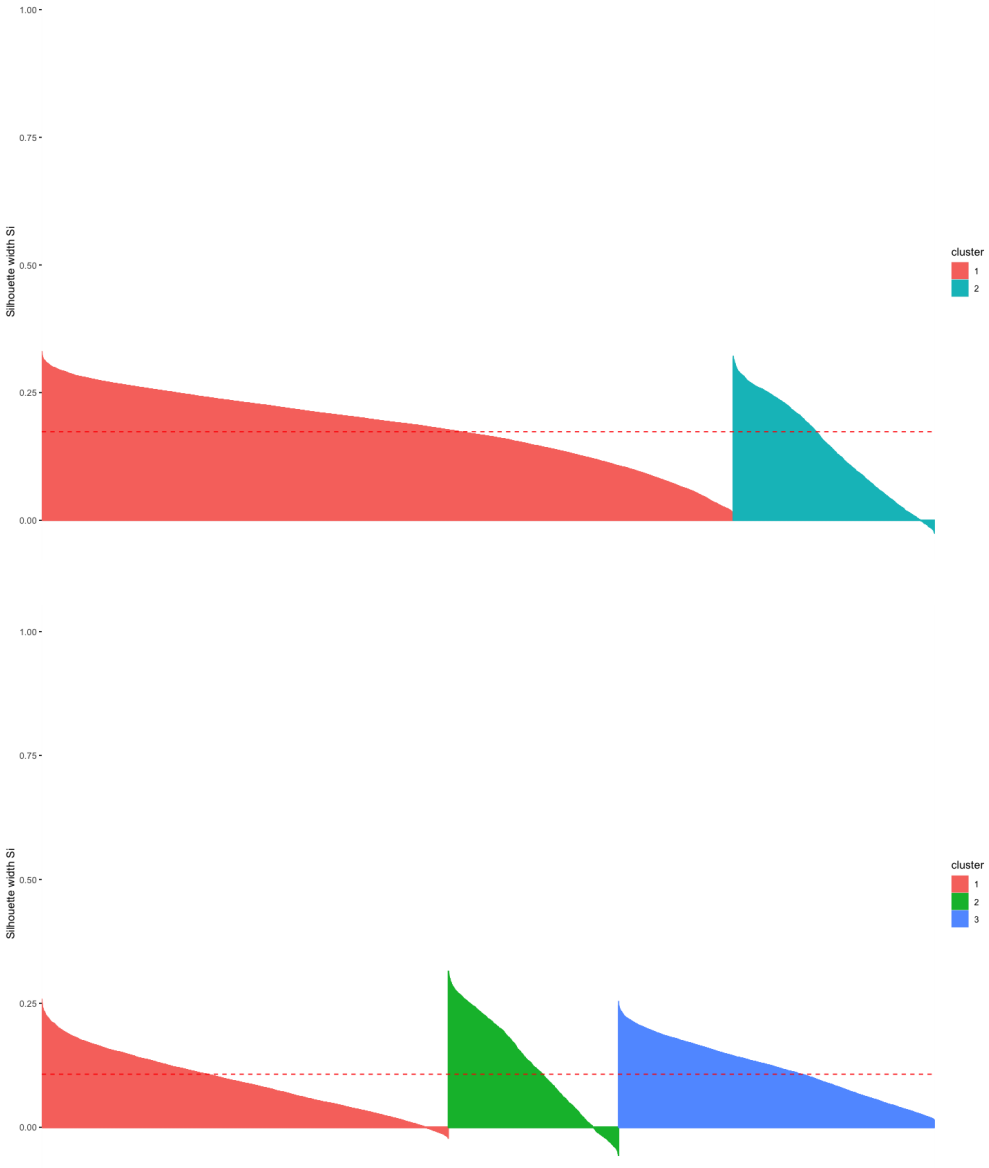

Figure 13: Silhouette width, k = 2 and k = 3

Table 17: Logit regression coefficients (dataset 2)

	Dependent variable:
Customer_Age	−0.001
	(0.005)
GenderM	−0.721***
	(0.079)
Dependent_count	0.127***
	(0.021)
Education_Level	−0.001
	(0.005)
Marital_StatusMarried	−0.434***
	(0.092)
Marital_StatusSingle	0.081
	(0.093)
Income_Category	0.00000**
	(0.00000)
Card_CategoryGold	1.262***
	(0.242)
Card_CategoryPlatinum	0.668*
	(0.385)
Card_CategorySilver	0.330**
	(0.136)
Months_on_book	−0.011**
	(0.005)
Total_Relationship_Count	−0.345***
	(0.017)
Months_Inactive_12_mon	0.509***
	(0.028)
Contacts_Count_12_mon	0.518***
	(0.024)
Credit_Limit	−0.00001**
	(0.00000)
Total_Revolving_Bal	−0.001***
	(0.00005)
Total_Amt_Chng_Q4_Q1	−0.774***
	(0.127)
Total_Trans_Amt	0.001***
	(0.00002)
Total_Trans_Ct	−0.130***
	(0.003)
Total_Ct_Chng_Q4_Q1	−2.540***
	(0.127)
Avg_Utilization_Ratio	−0.239
	(0.162)
Constant	7.344***
	(0.288)

Note: *p<0.1; **p<0.05; ***p<0.01

References

De Caigny, A., Coussement K. De Bock K. W. 2018. "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees." *European Journal of Operational Research* (269(2)).

Ganesh, J., Arnold M. J. Reynolds K. E. 2000. "Understanding the customer base of service providers: an examination of the differences between switchers and stayers." *Journal of marketing* (64(3)):65–87.

James, G., Witten, D., Hastie, T. Tibshirani, R. 2013. *An introduction to statistical learning.* New York: springer.

Menardi, G., Torelli N. 2014. "Training and assessing classification rules with imbalanced data." *Data mining and knowledge discovery* (28(1)).

Verbeke, W., Martens D. Mues C. Baesens B. 2011. "Building comprehensible customer churn prediction models with advanced rule induction techniques." *Expert systems with applications* (38(3)):2354–2364.