Applied Statistics and Big Data: Credit Card Offer

Andrea Rafanelli
mat.4815168

June 5, 2020

Abstract

A bank would like to understand the demographics and other characteristics associated with its customers. This project aims to understand these characteristics in order to respond to clearer and better future marketing strategies. The designed approach allows the bank to control the potentially most important factors to determine how and to whom to target its marketing actions.
In this analysis, the cluster analysis is used to segment customers and create "average customers". The decision tree is used to find out if there is a subset of factors that can be selected as most important to explain the variable, just as a Best Subset Selection is implemented for the same goal. Finally, we will analyze whether it will be possible to build a model to predict the customers who will accept the offer proposed by the bank.
Keywords: *Cluster Analysis, Classification tree, Best Subset Selection, Logisti Regression*

1 Introduction

The analysis is developed in three sections: a section that will concern the data description part (EDA), one that will concern cluster analysis and finally one in which we will try to understand which are the most important predictors in explaining Y and to build a predictive model. The dataset used can be found on the JMP website with the name "Credit card Marketing". This dataset contains 18,000 observations of current bank customers. There are both quantitative and qualitative variables.
Among the numeric variables we can find: 1. Number of non-credit bank accounts; 2. Number of credit cards; 3. Number of houses; 4. Household size; 5. Average balance; 6. Q1, Q2, Q3, Q4 average balance in quarter of a year;
Among the qualitative ones: 1. Offer accepted (Yes, No), that is our target variable; 2. Reward program (Cashback, Points, Air miles); 3. Mailer type (Letter, Postcard); 4. Income level (Low, Medium, High); 5. Overdraft protection (Yes, No); 6. Credit rating (Low, Medium, High); 7. House property (Yes, No);
The two questions we ask ourselves in this analysis are: Why and which bank customers accept credit card offers?; Can we build a model that predicts customer who accept credit card offers?

2 Exploratory Descriptive Analysis

The first analysis we carry out is that which concerns our target variable. Looking at the plot below (figure 1), we understand that the dataset is highly unbalanced. This, in fact, presents only a 6% of "Yes" and a 94% of "No".

Figure 1: Target variable distribution

Subsequently, we try to better understand the characteristics of the customer who accepts the bank's offer.

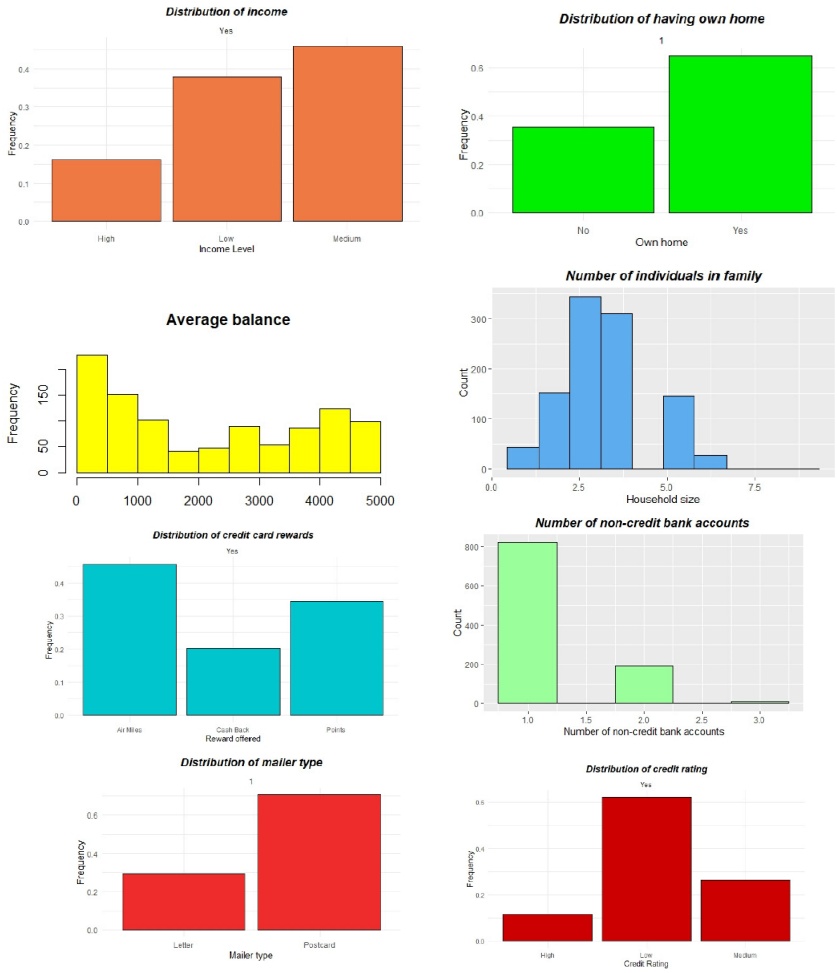

Figure 2: Characteristics of client who accept the offer

Looking at the various plots of figure two, we can find some information about the customer who accepts the offer:

a. This client has a medium-low income;

b. The client's average balance has mostly values below 1000;
c. This client has mostly owns home;
d. The number of individuals within the client's family is 3-4;
e. The type of credit card reward is airmiles and points;
f. The customer who accepts the offer receives mostly this via postcard;
g. The customer has no more than one non-credit card account;
h. The customer has a low credit rating;
All these assumptions will prove to be valuable information for the bank. Moving forward in the analysis, we will try to better understand which of these will be fundamental and to be taken into account in our analysis.

In this section, possible problems within the dataset are also considered.
The first to be considered is that of correlations.

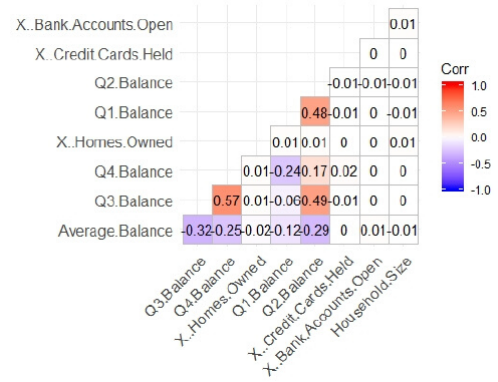

Figure 3: Correlation between numerical predictors

From figure 3, we note the correlations (positive and negative) between the variables Q3-Q4, Q3-Q2, Q2-Q1.
Furthermore, a second problem is identified within the dataset, namely the presence of 24 missing values present in the variables Q1, Q2, Q3, Q4. By developing subsequent analyzes, it is understood that the missing values derive from the same 24 customers.
We have two choices: avoid the 24 observations, avoid the 4 variables.
Considering the presence of correlations and understanding that the "Average balance" variable is a linear combination of the four variables (the average), it is decided to delete the four variables, solving the problem of correlation and missing data.

3 Cluster Analysis

Cluster analysis was performed to try to obtain groups of subjects, in order to create a customer segmentation.
The first thing to consider in this analysis is the presence of numeric and categorical variables. Having both types of variables leads us to use a different type of distance from the usual ones (e.g. Mahalanobis, Euclidian, Minkowski etc.). The distance used, therefore, is

that of Gower. This type of distance allows us to calculate dissimilarity for both continuous and discrete objects.
Once the dissimilarity matrix is obtained, a hierarchical agglomerative method is used to aggregate the observations. The Ward method.
The choice of this method is based on the intention of creating "compact" clusters. In fact, the method minimizes deviance within groups, increasing the between groups deviance. This creates clusters which present within them very similar observations, and very dissimilar to those of other clusters.
Initially the idea of having compact clusters was born from the desire to create a group for customers who do not accept the offer and one for customers who accept the offer.

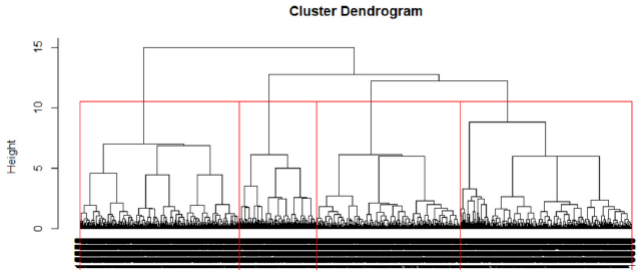

Figure 4: Dendrogram

From the above dendrogram 4 clusters are identified - cutting at the height of 10, where the greatest incremental distance is identified. The four clusters have the following numbers: 1.Cluster: 5165 No, 13 Yes; 2.Cluster: 4685 No, 9 Yes; 3.Cluster: 4633 No, 988 Yes; 4.Cluster: 2494 No, 13 Yes; Thanks to the identification of these four groups, descriptive analyzes are developed that help us identify four average customers:

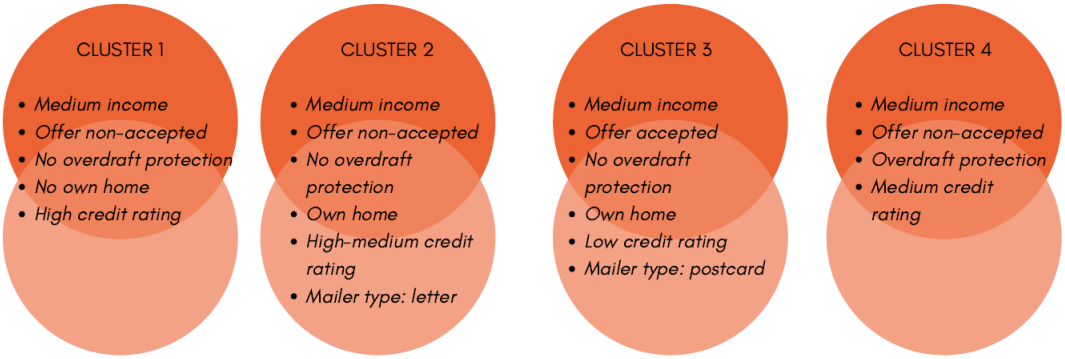

Figure 5: Average clients

The average customer of group 3 appears to be the most important in our analysis. In this group, in fact, there is a prevalence of customers who accept the offer. Analyzing figure 5, the main features are highlighted for this customer: he has a medium income; he has no overdraft protection; he has an own home; he has a low credit rating; he received most the offer with the postcard.
Many of these features had already been highlighted in the section. This analysis therefore

confirmed some of the discriminating characteristics between customers who accept the offer and those who do not accept it. In this analysis they are Credit Rating (low) and Mailer type (postcard).

4 Predictive models

In this section predictive models are implemented. The predictive models adopted will always allow to pursue the first objective, that of understanding the most influential predictors. The first step, in this section, is to carry out a train-test division using a Validation-set approach. The choice of this method instead of the use of a Cross Validation method is due to the attempt to reduce the complexity of calculation, given the large number of observation within the dataset. The dataset is then divided, using 70% as a training set and 30% as a test set. In this process, the same proportion of "Yes" and "No" is also maintained within both, train and test set.
The main problem in this section will be that of the imbalance of the dataset, highlighted in section 2.
When we are going to build a predictive model, if we do not change the distribution of classes within the dataset, the model could focus on the prevailing class ignoring the "rare events" which, in our case, are the class of customers that accept the offer.
Therefore, a resampling method is used to solve the problem of unbalancing the dataset. This method, "both sampling", works by doing an undersampling without replacement of the majority class and an oversampling with replament of the minority class. Within the function, which is located in the ROSE package, the proportion (0.5) of "Yes" and "No" to be obtained is established. The end result is 6031 No and 5953 Yes. This process is obviously applied on the training set. The test set, that will later be used to rent the models, remains unchanged and with the original proportion of Yes and No.

4.1 Classification tree

The first method used is that of the classification tree.
This method has been included among the predictive methods, although it is not recommended in predictive terms but more in terms of interpretability.
In a good decision tree the nodes should be as pure as possible (contain only instances of data belonging to a single class). To "grow" the tree was, for this reason, using the Gini index which measures the purity of the nodes.
Once the tree was built, to reduce its complexity and consequently reduce its variance and make it more interpretable, the "pruning" method was chosen. To select the best subtree, a Cross Valdation was performed and the tree optimization procedure was obtained using the Classification error rate.

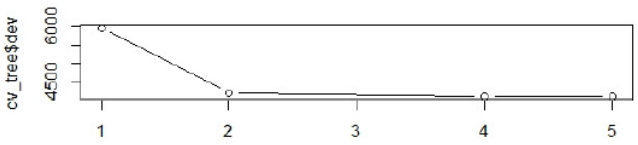

Figure 6: Pruning

As Figure 6 suggests, there are two subtrees we can select, those with the least error pruning, a subtree of 4 or 5. Both have the same result. To reduce the complexity of the model, a k = 4 is chosen.
The final result is represented in figure 7.

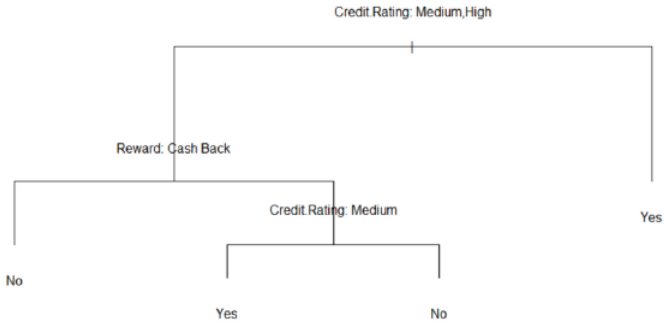

Figure 7: Final tree

Figure 7 shows the importance of two predictors in explaining Y: Credit rating (the most important) and Reward. These two predictors had been highlighted in section 2 as discriminating factors between customers who accept and do not accept the bank's offer. In addition, the Credit rating variable was also present as a discriminating variable within cluster 3 (section 3).
Subsequently, the model is tested in terms of prediction.

Table 1: Confusion matrix

	Predicted No	**Predicted Yes**
No	2618	3047
Yes	55	280

The final model shows an accuracy of 48%, specificity of 46% and sensitivity of 83%. This, therefore, performs very well in Yes's prediction but badly in that of No, moreover the model is not accurate in terms of prediction.
In conclusion, we can derive important information regarding the importance of predictors but we cannot use the model to predict, as the estimates are inaccurate.

4.2 Logistic Regression

To perform the logistic regression, a selection of the most important predictors is made in advance. This is both to pursue the first goal (to understand the most important predic-

tors), and to reduce the number of predictors within the model.
The Best Subset Selection method is then used. The choice of this method in spite of the Stepwise method is justified by the guarantee of finding the best model at the end of the process.
The variables selected by the model are: Credit rating, Mailer type, Reward, Income level, Overdraft protection, Number of houses, Household size, Average balance. We can note that, also in this section, some of the variables that we previously identified as important in explaining our response variable are selected. Here too, in fact, we find Credit Rating, Mailer Type and Reward.
By using these variables, a logistic regression is implemented and the results in terms of model prediction are reported below.

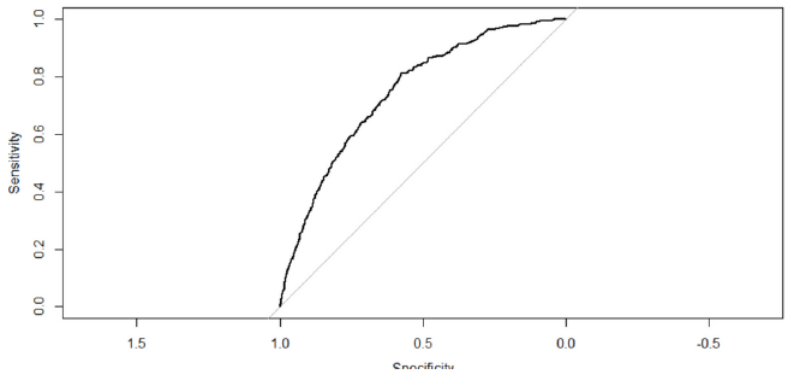

Figure 8: ROC curve

The Roc curve presents an AUC of 75%. This means that the model's predictions are 75% correct.
By better analyzing the specificity, sensitivity and accuracy of the model, the following confusion matrix is implemented:

Table 2: Confusion matrix

	Predicted No	**Predicted Yes**
No	3858	1807
Yes	114	221

The final model shows an accuracy of 68%, specificity of 68% and sensitivity of 66%.
This model shows better results in terms of accuracy and specificity than the classification tree. However, we can ask ourselves what would happen if we changed the threshold. Would the model improve?

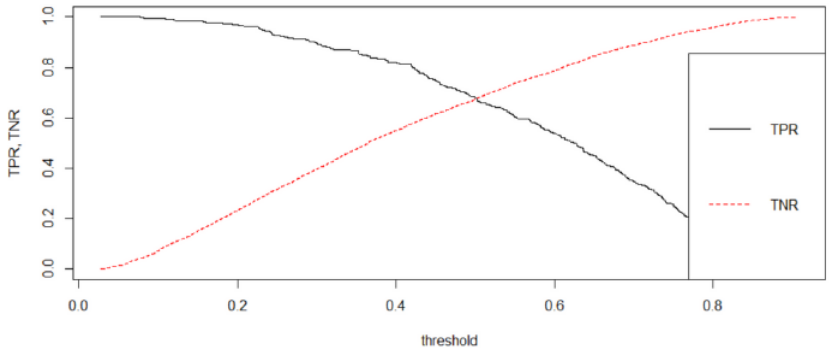

Figure 9: Threshold

From figure 9, we can understand that changing the threshold in this model is a difficult choice. By default, our threshold is equal to 0.5. If we increase it, the specificity of the model will increase but the sensitivity will decrease drastically and vice versa.
following table shows the confusion matrix with a threshold of 0.4.

Table 3: Confusion matrix

	Predicted No	**Predicted Yes**
No	3206	2459
Yes	63	272

The model in this case obtains an accuracy of 58%, a sensitivity of 82% and a specificity of 56%.
We can see how the change in the threshold in our case leads to an overall deterioration of the model.
In conclusion, it is not desirable to change the threshold in our situation. In this section we have identified logistic regression as the best model. In fact, the Classification tree is not optimal in predictive terms.
To improve this section, in the future, it is desirable to try to use other predictive models, such as a Random Forest, which could improve the performance of the Classification Tree.

5 Conclusion

Section 2 underlined the presence of few customers who accept the bank's offer. This leads us to understand that the marketing strategies adopted by the bank are not optimal.
From this analysis we understood what factors to focus on for future marketing strategies. These factors have been highlighted in all sections: EDA, Cluster Analysis, Predictive Models. We can therefore say that Credit Rating, Reward and Mailer type are the most important predictors in explaining Y. What has been understood is that the customers who accept the bank's offers are customers with a low Credit Rating, who receive the offer via postcard and who present credit cards with airmiles and points as rewards.
For the predictive part, we understood that this is difficult to perform, due to the presence

of an unbalanced dataset. We have used a resampling method to overcome this problem but we are not sure that this method is optimal for our data. In the future, this section will be studied in depth, looking for other methods to deal with imbalanced dataset.

References

1. An introduction to statistical learning, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, 2013

2. The R project for statistical computing, www.r-project.org