UNIVERSITY OF PALERMO

Degree course in Statistics and Data Science
Department of Economics, Business and Statistics

Exam – November 2022

Stochastic Networks

Antonino Gagliano
Alessia Granata
Veronica Milazzo
Marco Tarantino

Prof. Antonino Abbruzzo

ACADEMIC YEAR 2022 - 2023

CONTEXT

Term deposits are a major source of income for a bank. A term deposit is a cash investment held at a financial institution. Your money is invested for an agreed rate of interest over a fixed amount of time, or term. The bank has various outreach plans to sell term deposits to their customers such as email marketing, advertisements, telephonic marketing, and digital marketing.

Telephonic marketing campaigns still remain one of the most effective way to reach out to people. However, they require huge investment as large call centers are hired to actually execute these campaigns. Hence, it is crucial to identify the customers most likely to convert beforehand so that they can be specifically targeted via call.

The data is related to direct marketing campaigns (phone calls) of a Portuguese banking institution conducted among May 2008 and November 2010. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed by the customer or not.

The dataset is made up by 45211 observations and 17 variables, these are divided into three groups:

bank client data:
1 - **age** (numeric)
2 - **job** : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
3 - **marital** : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4 - **education** (categorical: "unknown","secondary","primary","tertiary")
5 - **default**: has credit in default? (binary: "yes","no")
7 - **housing**: has housing loan? (binary: "yes","no")
8 - **loan**: has personal loan? (binary: "yes","no")

related with the last contact of the current campaign:
9 - **contact**: contact communication type (categorical: "unknown","telephone","cellular")
10 - **day**: last contact day of the month (numeric)
11 - **month**: last contact month of year (categorical: "jan", "feb", "mar", …, "nov", "dec")
12 - **duration**: last contact duration, in seconds (numeric)

other attributes:
13 - **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14 - **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
15 - **previous**: number of contacts performed before this campaign and for this client (numeric)
16 - **poutcome**: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

Output variable (desired target):
17 - **y** - has the client subscribed a term deposit? (binary: "yes","no")

Missing Attribute Values: None.

The aim of this analysis is to understand which variables influence to open a term deposit and which kind of subject is more likely to open a term deposit with this bank, according to client's information.

DATASET MANAGEMENT

In order to start the analysis, it's necessary to modify the dataset. First of all some variables have been deleted:

- **Contact**, the difference between telephone and cellular isn't crucial for the analysis about a telephonic marketing campaign.
- **Day** and **Month**, starting from these two variables it was created the variable **year**.
- **Pdays**, the information given by this variable is similar to the one given by **previous**.
- **Poutcome**, the meaning of this variable is ambiguous, also the 82% of the values are "unknown".

In addition, all the qualitative variables have been factorized and the quantitative ones divided into categories:

- **Age** (categories: "≤30", "31 – 59", "> 60"), based on the general ageing policy of banks (more concessions for the youngest).
- **Balance** (categories: "negative", "medium-low", "medium-high"), based on the distribution of the data.
- **Duration** ("≤ 3 minutes", "> 3 minutes").
- **Campaign** ("≤ 3 contacts", "> 3 contacts").
- **Previous** ("yes", "no").

At the end the dataset on which the analysis is conducted considers 12 explanatory variables and the response one.

EXPLORATORY ANALYSIS

Considering the target variable, only the 11,7% of the clients opened a term deposit:

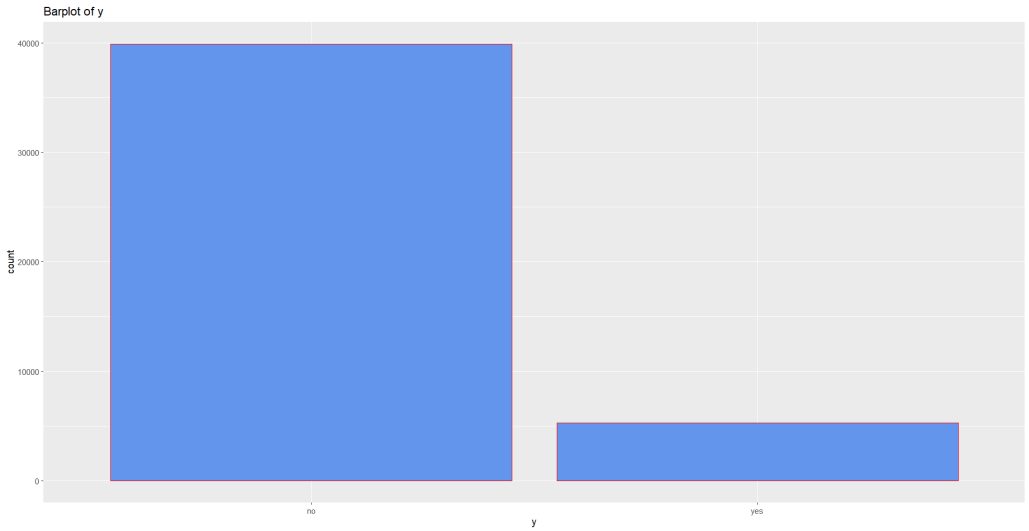

***Figure 1**: Bar plot of the target variable.*

In order to understand the marginal relationship between the response variable and all the explanatory, it is useful to represent all the contingency tables considering one explanatory variable at the time.

In the following graph are reported all the relative frequencies of the explanatory variables, given the target.:

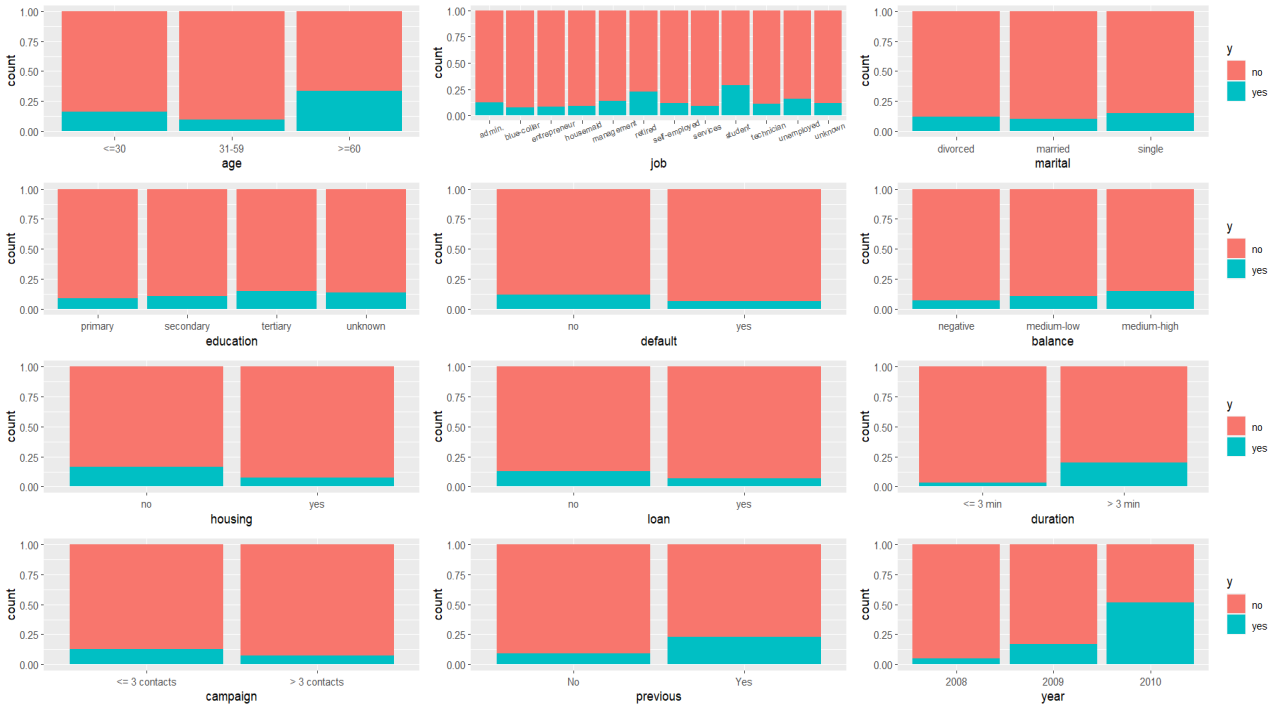

***Figure 2**: Bar plot of the target variable given by the explanatory variables.*

As we can see from the graph, the oldest clients are the most likely to open a term deposit with the bank (34%), followed by the youngest (16%), maybe influenced by the bank concessions. Also, from the marital status it can be seen that the singles open more deposits (15%) than married or divorced people.

Considering the level of education, the higher it is, the more likely is to have a deposit, the same trend for the balance.

As regards the economic variables, it can be seen that people who have housing loan, personal loan or have credit in default, consider less a deposit opening.

Focusing the attention on the contacts during the marketing campaign, people who stay on the phone more than three minutes, or who has been contacted in a previous campaign, are more interested in opening a deposit.

The last consideration has to be done on the year of the contact: as we can see from the table, there is an increasing of deposit opening during the time, in fact, if in the 2008 only the 5% agreed to open a deposit, in the 2010 this value is above the 50%, ten times the value of the 2008, just two years before, probably due to the financial crisis of those years ("bank-run" of 2007-2009).

It can be also important to study the relationships among the exploratory variables to understand better the net of possible links among them.

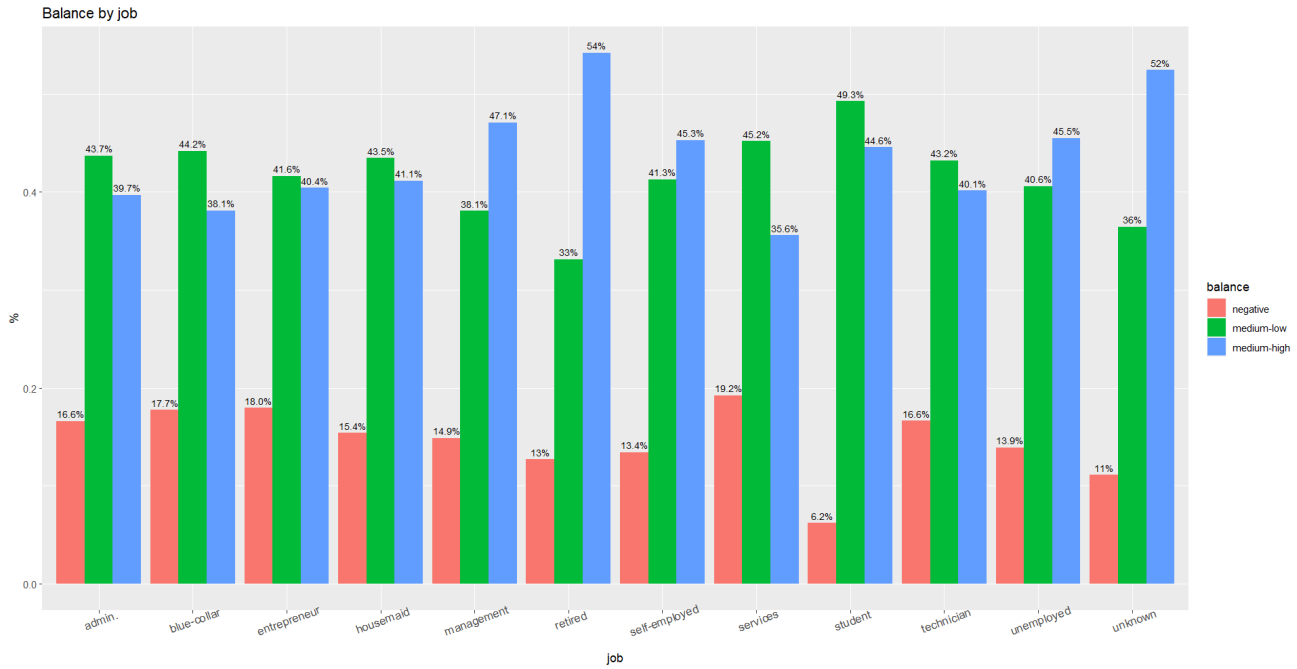

***Figure 3**: Bar plot of the balance by job.*

The first consideration is done between the job of the clients and their balance. As we can see, retired people have the highest percent value of medium-high balance, followed by who didn't declare the job ("unknown"), instead the higher level of negative balance is referred to services job, followed by the blue-collar.

It has to be done a consideration about the students: from the previous table, we have seen that the 25% of them opened a deposit (the highest value observed among all the jobs) and maybe this is due to a very low percent of negative balance (the lower value observed in this table).

Another variable that can be compared with the job is the housing loan: the class with the higher level of housing loan is the blue-collar, followed by people that work in services:

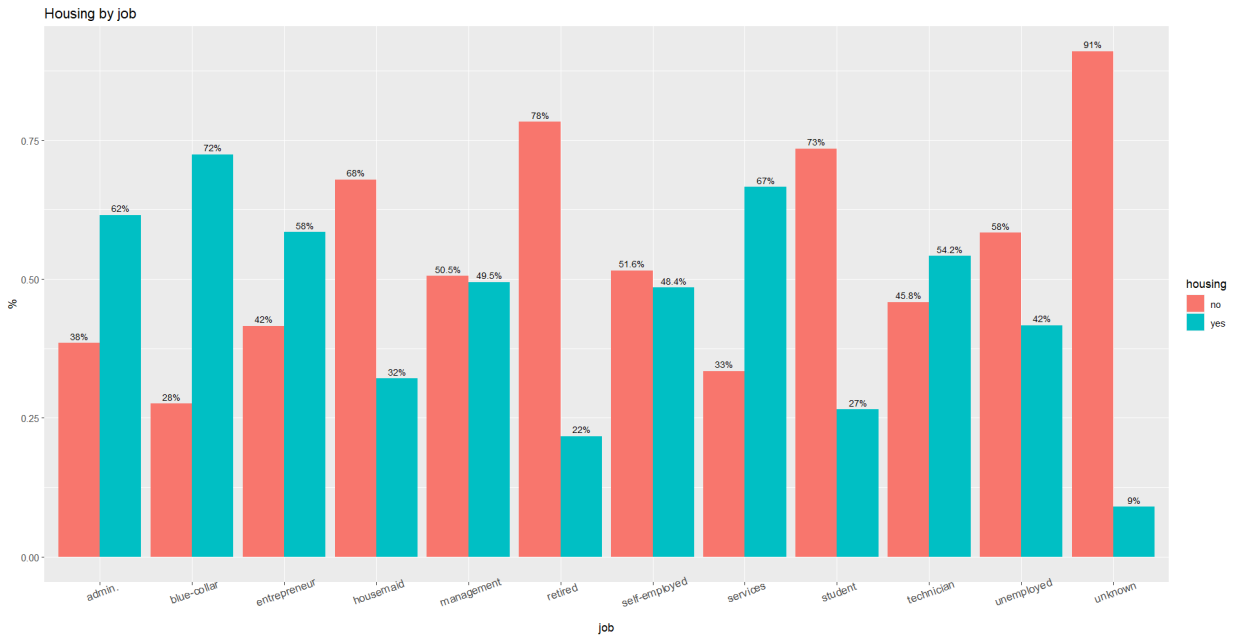

***Figure 4**: Bar plot of the housing by job.*

As we said before, the year can be significant on the target variable because of the financial crisis; it is reasonable to study the relationship between year and balance to identify this phenomenon.

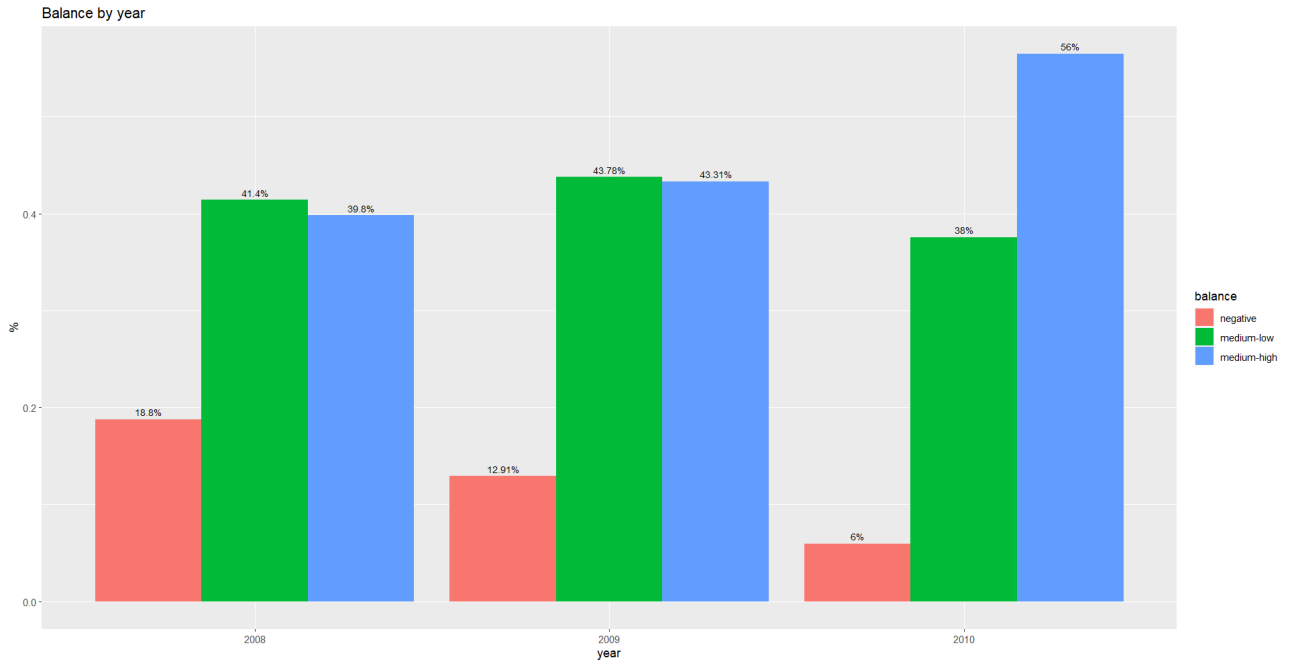

***Figure 5**: Bar plot of the balance by year.*

The graph confirms the supposed trend: if in the 2008 the amount of people with negative balance was 19%, in the 2010 this value is 6% and the amount of people with a medium-high balance is equal to the 56%.

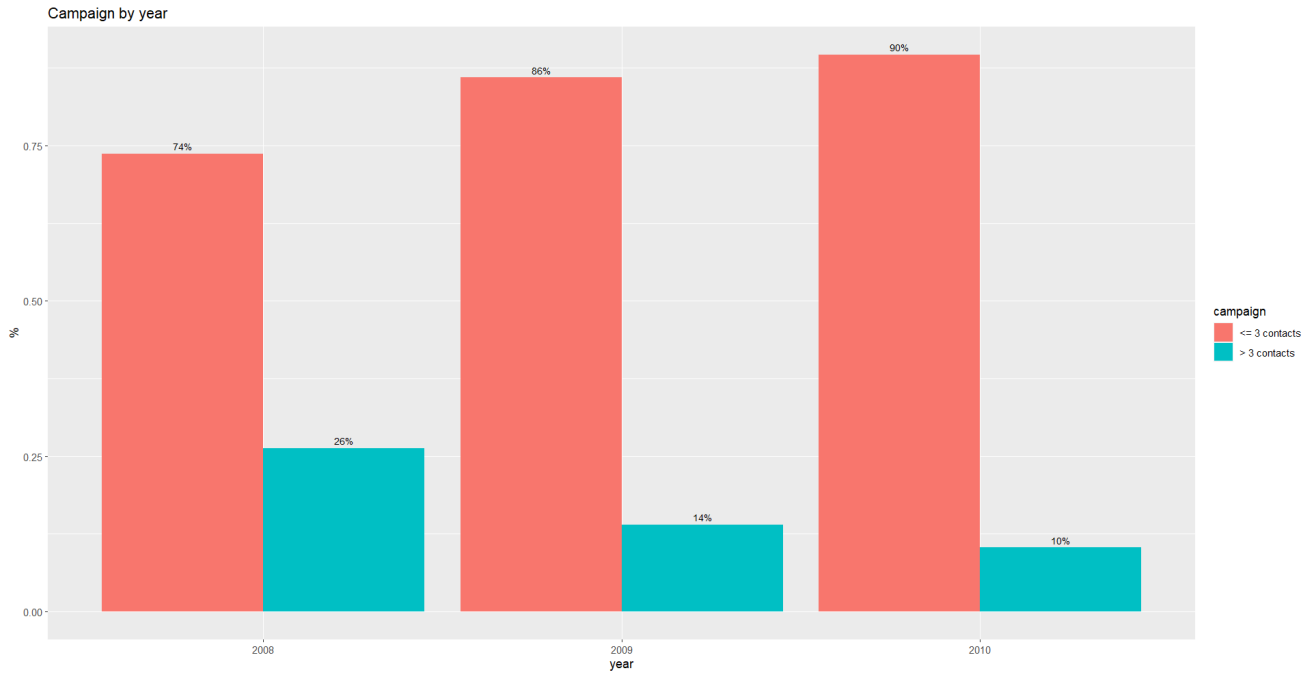

***Figure 6**: Bar plot of the campaign by year.*

Also the relationship between campaign and year suggests the same trend: in 2008 the bank needed income, so it is reasonable that there were more contacts with clients in order to obtain more deposits.

In order to understand why older people open more deposits, other two relationships are investigated: age-housing and age-duration:

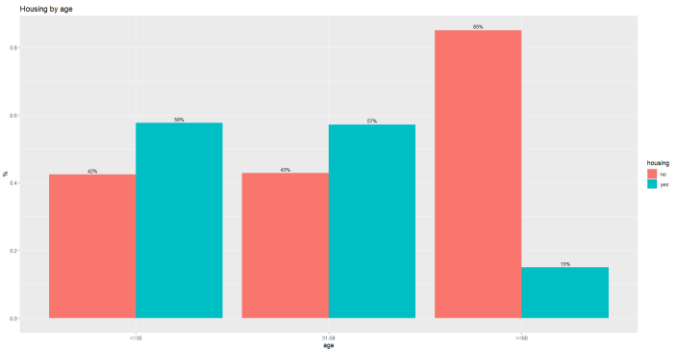

***Figure 7**: Bar plot of the housing by age.*

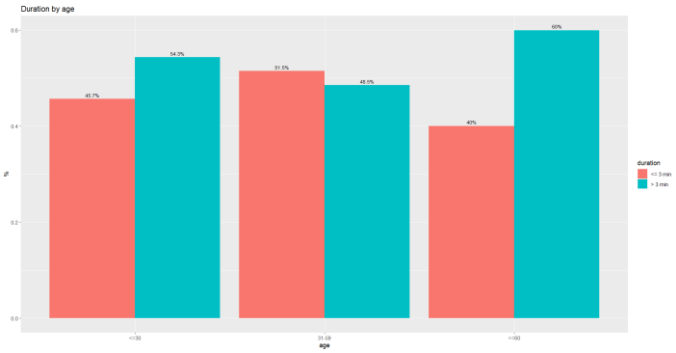

***Figure 8**: Bar plot of the duration by age.*

As we can see old people have less housing loans and also stay longer on the phone than the other two age groups: these were two conditions in favor of opening a term deposit. So housing and duration can explain the relationship between age and the output variable.

The last thing to do in order to understand the association between the variables in the dataset is a corrplot of them, it calculates the association two by two variables (based on V-Kramer):

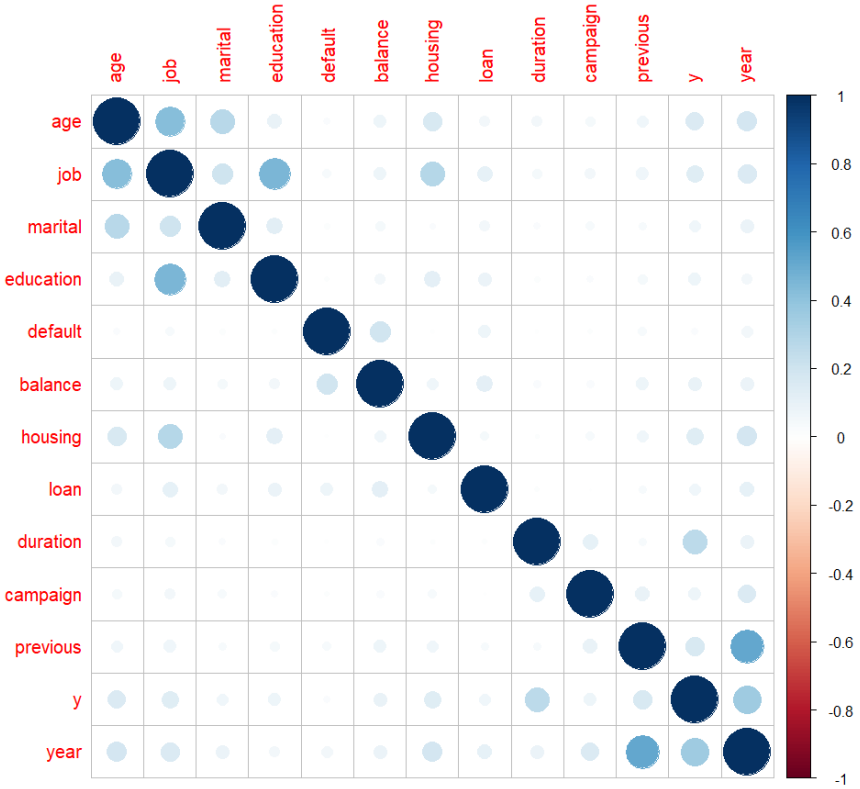

***Figure 9**: Corrplot of the variables.*

As we can see, the strongest associations are education – job, year – previous and age – job.

GRAPHICAL MODEL APPLICATION

A graphical model is a probabilistic model for which a graph expresses the conditional dependence structure between random variables.

This model is useful in order to interpret the relationships among all the variables and to estimate all the probabilities of interest (marginal, joint or conditional).

Before initializing this method, the dataset is divided into two subsets:

- Training set: subset on which we build the graphical model (75% of the observations are included).
- test set: subset on which we test how good the graphical model is in terms of predictions (25% of the observations are included).

The best way to represent the variables of this dataset is a directed acyclic graph (DAG), such that we can interpret the links between variables as cause-effect, but before applying this model we have to make restrictions about the possible links' direction:

It's obvious that variables such "Year" or "Age" cannot depend on the others, so in order to avoid possible links with wrong direction, it's created a blacklist (links that can't be in the graphical model) for these two variables.

Here the representation of the directed acyclic graph:

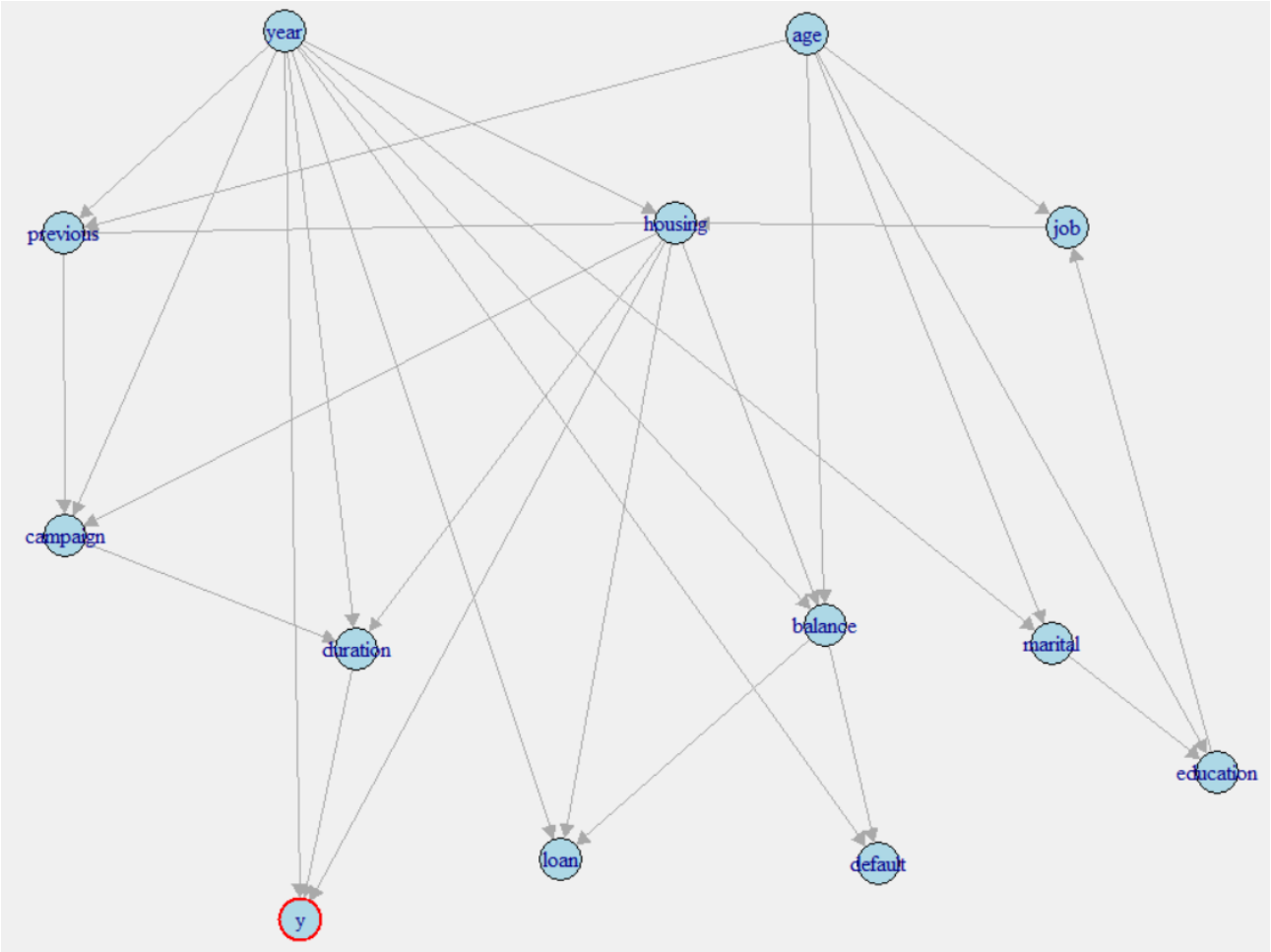

***Figure 10**: Directed Acyclic Graph.*

As we can see from the graph the variables “Year” and “Age” are not influenced by any others (restrictions applied correctly), but they influence a large number of variables, in particular:

- the “Year” affects strongly the marketing campaign of the bank, in fact we can see the links with the variables “Previous”, “Campaign” and “Duration”.

  Also, it’s very important from an economic point of view, in fact, it influences the balance of the clients and also the presence of loans, housing loans and default situations.

  Finally, it can also be noted that the year affects directly the response variable.

- The “Age” instead affects mainly the client’s variables, such “Job”, “Education” and “Marital”. It also influences variables like “Balance” and “Previous”.

As we could expect, the target variable does not affect any other variable, but it’s directly affected by “Year”, “Housing” and “Duration”.

There are other two variables that are at the end of the chain: “Loan” and “Default”. The first one is influenced by “Year”, “Housing” and “Balance”, the second one only by “Year” and “Balance”.

After having underlined the most important relationships in the directed graph, it’s important to build the relative probability tables. The way to do probabilistic inference on the Bayesian Network is the “Junction Tree” method:

The junction tree method is based on a transformation of the Bayesian Network to a junction tree, where each node in this tree is a group or cluster of variables. Probabilistic inference is performed over this new representation, via propagation over the junction tree, the probability of a variable is obtained by marginalization over the “junction”(clique).

Once the junction tree is built, inference is based on probability propagation over the junction tree. Initially the joint probability of each macro node is obtained, and given some evidence, this is propagated to obtain the posterior probability of each junction. The individual probability of each variable is obtained from the joint probability of the appropriate junction by marginalization.

Starting from the Bayesian Network, we can investigate marginal, joint or conditional probabilities by writing different queries, with a particular interest about the two goals of the analysis:

1) Understand which exploratory variables are significant in explaining the target variable.
2) Which are the main clients’ features that increase the probability of opening a term deposit with this bank.

Regarding the first goal, the three marketing variables are investigated, conditioned to the year:

Query 1:

Year = 2008	Duration	
y	<= 3 min	> 3 min
no	0.997	0.895
yes	0.003	0.105

Year = 2009	Duration	
y	<= 3 min	> 3 min
no	0.934	0.732
yes	0.066	0.268

Year = 2010	Duration	
y	<= 3 min	> 3 min
no	0.806	0.328
yes	0.194	0.672

As we can see, if the phone call lasts more than 3 minutes, clients are more inclined to open a deposit. Furthermore, this probability increases during the years, maybe due to an increased trust of the clients after the crisis, in particular if a client stayed in call for long in 2008, the probability of being convinced was 10%, in 2010 it increases at 67%.

Query 2:

marginal	Campaign	
y	<= 3	> 3
no	0.872	0.925
yes	0.128	0.075

Year = 2008	Campaign	
y	<= 3	> 3
no	0.947	0.956
yes	0.053	0.044

Year = 2009	Campaign	
y	<= 3	> 3
no	0.822	0.868
yes	0.178	0.132

Year = 2010	Campaign	
y	<= 3	> 3
no	0.490	0.519
yes	0.510	0.481

Starting from the marginal probability table, it can be deducted that if a client is contacted more than three times, he will be less inclined to open a deposit. If we condition this relationship by the year, we can see how the difference in probability of opening or not a term deposit decreases, but has the same direction.

Query 3:

marginal	Previous	
y	no	yes
no	0.906	0.783
yes	0.094	0.217

Year = 2008	Previous	
y	no	yes
no	0.949	0.949
yes	0.051	0.051

Year = 2009	Previous	
y	no	yes
no	0.818	0.845
yes	0.182	0.155

Year = 2010	Previous	
y	no	yes
no	0.490	0.494
yes	0.510	0.506

Considering the marginal relationship between "Previous" and the response variable, it suggests that if a client was contacted before this marketing campaign, the probability of success is 21,7%, if was not contacted before it's reduced to 9%. But if we condition this relationship by the year, we can see that there is no difference in the probability of success if the client was already called or not. So, the influence of "Previous" on the target is not real (spurious relationship).

It's also important to underline the influence of the economic variable on the target:

Query 4:

Housing = no	Balance		
y	Neg	M-L	M-H
no	0.889	0.841	0.815
yes	0.111	0.159	0.185

Housing = yes	Balance		
y	Neg	M-L	M-H
no	0.926	0.920	0.917
yes	0.074	0.080	0.083

marginal	Balance		
y	Neg	M-L	M-H
no	0.911	0.886	0.869
yes	0.089	0.114	0.131

Is showed a slight growth of subscriptions respect to the balance of the clients, if it's negative the probability of success is 9%, if it's medium-high is equal to the 13%. Conditioning on the "Housing" the relationship between the balance and the response is similar, in general the probabilities of success are lower if the client has a housing loan.

After having analyzed the most significant exploratory variables in the model, now the attention can be focused on what feature a client has to have in order to increase the probability of subscribe a term deposit with the bank.

Query 5:

Job	yes
admin	0.112
blue-collar	0.189
entrepreneur	0.032
housemaid	0.031
management	0.220
retired	0.060
self-employed	0.037
services	0.086
student	0.024
technician	0.168
unemployed	0.032
unknown	0.008

y = yes	Housing	
Job	no	yes
admin	0.102	0.129
blue-collar	0.133	0.277
entrepreneur	0.033	0.031
housemaid	0.040	0.016
management	0.240	0.187
retired	0.085	0.019
self-employed	0.042	0.030
services	0.076	0.103
student	0.031	0.012
technician	0.163	0.175
unemployed	0.040	0.020
unknown	0.012	0.001

From the distribution of the job, conditioned to the opening of a term deposit, it's pointed out that people working as management are the most likely to open it (22%), but if we take note of the presence of a housing loan, then blue-collar become the main target of the bank, in fact the relative probability is 27,7%.

Query 6:

education	yes
primary	0.145
secondary	0.504
tertiary	0.309
unknown	0.042

As regard the level of education of the clients, the bank has to focus the attention on people with a secondary grade.

Query 7:

age	yes
<=30	0.157
31-59	0.798
>= 60	0.044

Considering the age, the model suggests a high probability of success for the class 31-59 (80%). It's important to underline that in the exploratory analysis, we have considered the target conditioned to the age, in this case the table is built on the age, conditioned to a positive response.

For education and age were investigated also conditional relationships, but they hadn't any influence.

Query 8:

marital	yes
divorced	0.108
married	0.572
single	0.319

y = yes	year		
marital	**2008**	**2009**	**2010**
divorced	0.118	0.115	0.085
married	0.633	0.559	0.534
single	0.249	0.326	0.381

The first table points out that married clients are more inclined to open a deposit (57%), followed by the singles (32%). If we condition for the year, we can see that the probability of success increases only for the singles, instead it decreases for divorced and married.

GOODNESS OF THE MODEL

After having interpreted the result of the application of the model, it's important to verify the ability of the model to do the prediction.

For this reason, it is applied on the test set created before:

The aim is to compare the prediction of the target variable in the test set, obtained by the model selected, with the real values of the response observed in the test set.

Here the confusion matrix:

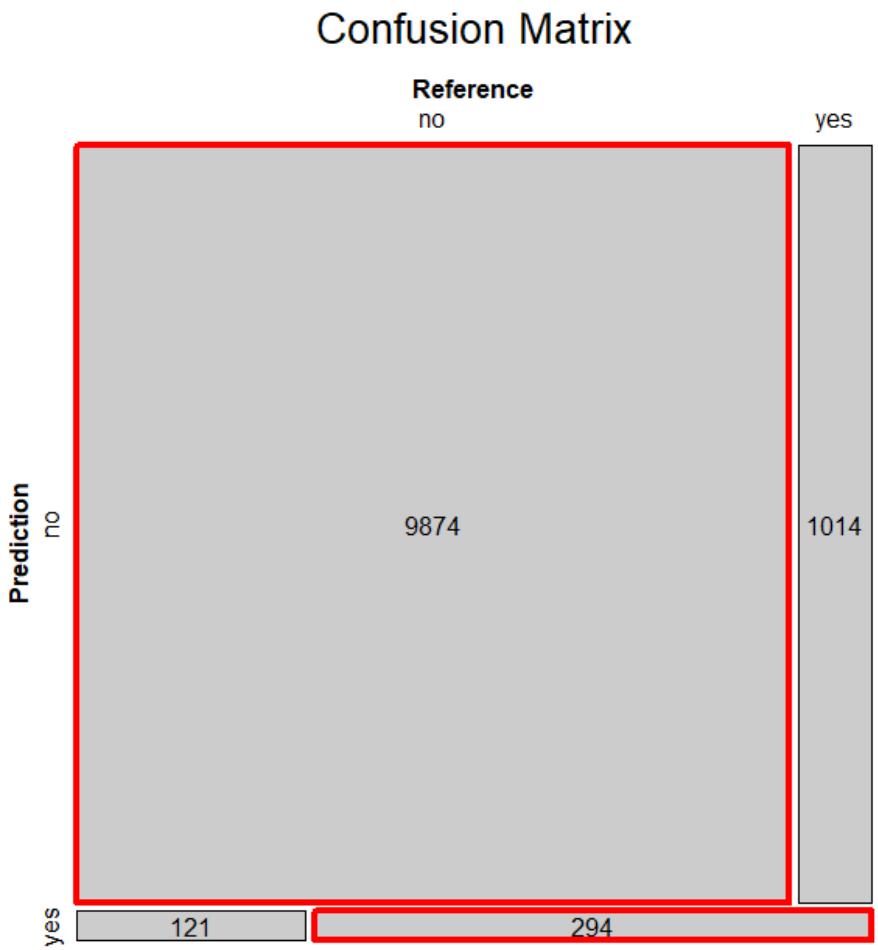

***Figure 11**: Confusion Matrix.*

From the matrix it can be seen that the model predicts correctly the 89% of the data, but our interest in on the modality "Yes" of the target variable, where the prediction is worse: only the 22% of them are predicted right.

Sensitivity = 0.22 (true positive over the true positive and false negative)

Specificity = 0.98 (true negative over the true negative and false positive)

In conclusion, this model is not so useful to predict a positive response of the target because of its imbalance.

It can be useful to explore other method to build a graphical model that solve the imbalance problem. We have considered:

- Over – sampling: the algorithm over-samples the "Yes" modality in order to obtain a balanced test set.
- Under – sampling: the algorithm under-samples the "No" modality in order to obtain a balanced test set.
- Both: creates possibly balanced samples by random over-sampling minority examples, under-sampling majority examples or combination of over- and under-sampling.

Here the comparison of the model selected with these three methods:

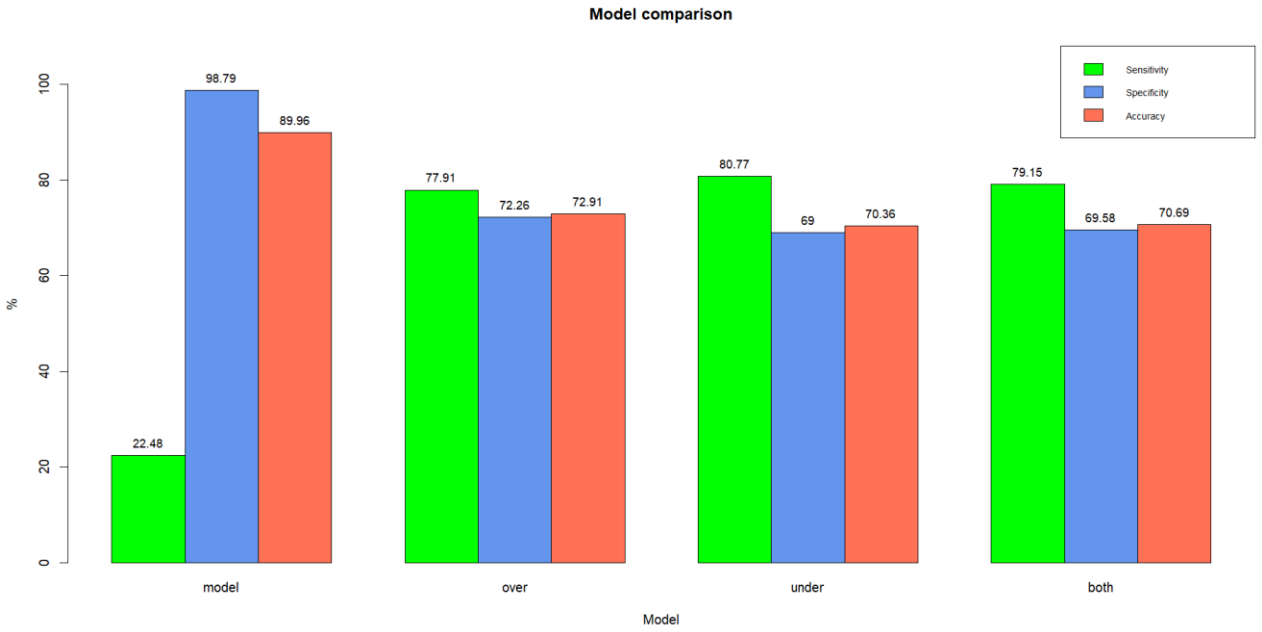

***Figure 12**: Comparison of the different models.*

As we can see, all the three alternatives improve the prediction of target modality "Yes". If we have to choose the model considering our goal (target = "Yes"), then we will choose the under-sampling model, if we consider a trade-off among the three values (sensitivity, specificity and accuracy) may be the best solution can be the model obtained by the "over-sampling" procedure.

CONCLUSION

At the end of this analysis we can say that:

- The opening of a term deposit is influenced by:
  i. The call duration;
  ii. The number of time that the clients are contacted;
  iii. Taking out or not a housing loan;
  iv. The different balance between clients.

However, we have to consider that our conclusions are strongly influenced by the year because of the financial crisis.

- If the bank has to identify the perfect profile to call in order to have a high probability of make a new subscription, considering the results of the model, it will be a client who has the following features:
    i. Married;
    ii. Age between 31-59;
    iii. Level of education equal to the secondary grade;
    iv. belonging to the blue-collar class of work.