Università di Catania

Credit card analysis

Statistical learning report

Prepared By:
Emanuele Fiorito
Paolo Vivera
Rosario Licciardello

Summary

0. Introduction to the report

The aim of this report is to perform a statistical analysis, using R, on the dataset "Credit card", uploaded by Rohit Udageri and available on www.kaggle.com. The training data set is composed of 1,238 observations on people and 17 variables, with one more added to the test data.

First of all, here's a quick overview of the variables involved in the analysis.

- **Gender** → Categorical dichotomic variable which identifies the person's gender.
- **Car Owner** → Categorical dichotomic variable which states if the person owns a car or not.
- **Property Owner** → Categorical dichotomic variable which states if the person owns a property or not.
- **Children** → Quantitative discrete variable which counts the number of children for each person.
- **Annual Income** → Quantitative continuous variable which identifies the person's annual income.
- **Type of Income** → Categorical nominal variable which identifies the type of income associated with each person.
- **Education** → Categorical ordinal variable which gives information about the person's educational level.
- **Marital Status** → Categorical nominal variable which tells information about the person's civil state.
- **Housing Type** → Categorical nominal variables which show the person's living conditions.
- **Birthday Count** → Quantitative discrete variable which counts the person's age (in terms of days passed from the birth date of the person).
- **Employed Days** → Quantitative discrete variable which counts the number of days passed since the employment day (basically the working time).
- **Mobile Phone** → Categorical dichotomic variable which states if the person owns a mobile phone or not.
- **Work Phone** → Categorical dichotomic variable which states if the person owns a work phone or not.
- **Phone** → Categorical dichotomic variable which states if the person owns a home phone or not.
- **Email ID** → Categorical dichotomic variable which states if the person owns an email ID or not.
- **Type Occupation** → Categorical nominal variable which identifies the person's occupation.
- **Family Members** → Numerical discrete variable which counts the number of family members in each person's family.

The additional variable added to the test set is:

- **Acceptance** → Binary variable which express, for each person, the result of acceptance of a credit card (0) or not (1). Already present in the training set, the response variable is predicted in the test set.

The basic idea of the analysis is to build a predictive model on the test set based on the information contained in the training set, capable of foreseeing the possible relationship between credit card acceptance and the other factors, which will be considered as the predictors of the model.

1. Exploratory Data Analysis (EDA)

The exploratory analysis concerns a preliminary overview of the data: the main idea is to get a general understanding of the data set before performing more complex techniques.

1.1. Univariate analysis

The analysis starts with a brief description and summary of all the variables. For numerical variables, mean, standard deviation and following percentiles are provided. The summary includes variables modified and results from data cleaning. In the original dataset there were 1,238 observations, of which 40 were removed since they were containing missing values. The procedures applied and the created variables can be found in the technical appendix.

Data summary

Name	dataset2
Number of rows	1198
Number of columns	23
_____	
Column type frequency:	
character	5
numeric	18
_____	
Group variables	None

Variable type: character

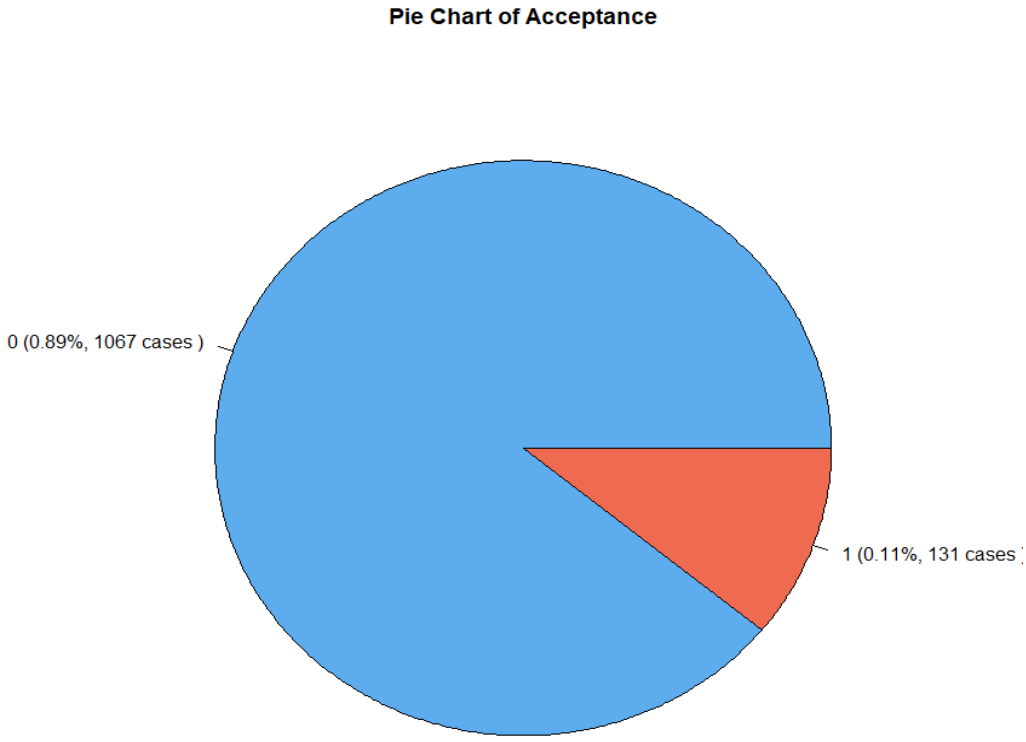
skim_variable	min	max	empty	n_unique	whitespace
Type_Income	7	20	0	4	0
EDUCATION	15	29	0	5	0
Marital_status	5	20	0	5	0
Housing_type	12	19	0	6	0
Type_Occupation	0	21	373	19	0

Variable type: numeric

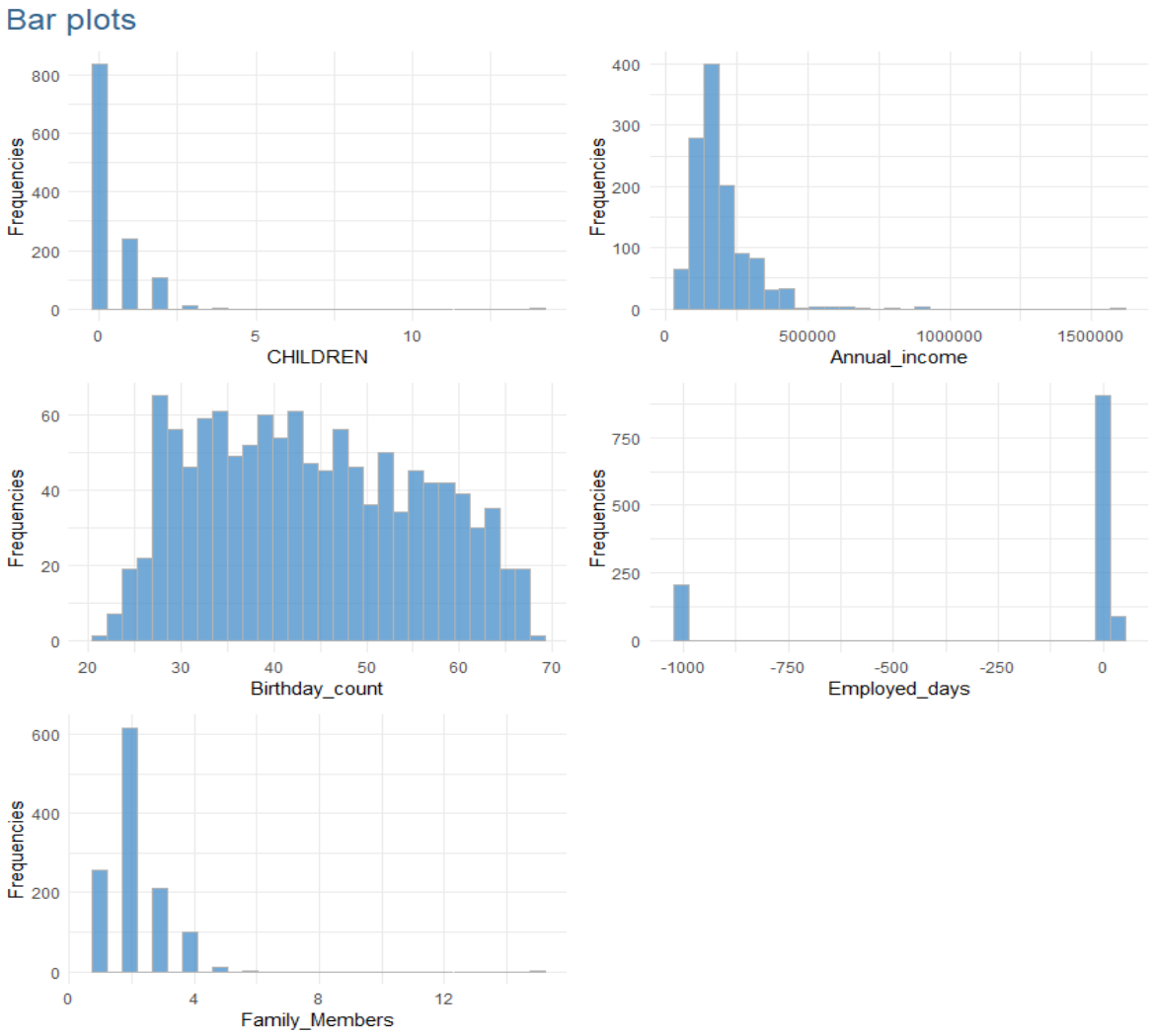
skim_variable	mean	sd	p0	p25	p50	p75	p100	hist
GENDER	0.64	0.48	0.00	0.00	1.00	1.00	1.000e+00	
Car_Owner	0.40	0.49	0.00	0.00	0.00	1.00	1.000e+00	
Propert_Owner	0.65	0.48	0.00	0.00	1.00	1.00	1.000e+00	
CHILDREN	0.42	0.80	0.00	0.00	0.00	1.00	1.400e+01	
Annual_income	190063.46	111989.58	33750.00	117000.00	157500.00	225000.00	1.575e+06	
Birthday_count	43.83	11.61	21.11	33.84	42.68	53.34	6.835e+01	
Employed_days	-163.46	378.36	-1000.67	1.16	4.30	8.77	4.079e+01	
Mobile_phone	1.00	0.00	1.00	1.00	1.00	1.00	1.000e+00	
Work_Phone	0.20	0.40	0.00	0.00	0.00	0.00	1.000e+00	
Phone	0.29	0.46	0.00	0.00	0.00	1.00	1.000e+00	
EMAIL_ID	0.10	0.29	0.00	0.00	0.00	0.00	1.000e+00	
Family_Members	2.17	0.96	1.00	2.00	2.00	3.00	1.500e+01	
acceptance	0.11	0.31	0.00	0.00	0.00	0.00	1.000e+00	

skim_variable	mean	sd	p0	p25	p50	p75	p100	hist
Type_Income_mean	0.11	0.03	0.05	0.10	0.10	0.13	1.500e-01	
EDUCATION_mean	0.11	0.02	0.00	0.10	0.10	0.11	2.800e-01	
Type_Occupation_mean	0.11	0.05	0.00	0.09	0.11	0.12	1.000e+00	
Marital_status_mean	0.11	0.03	0.05	0.10	0.10	0.10	1.500e-01	
Housing_type_mean	0.11	0.04	0.05	0.10	0.10	0.10	4.000e-01	

This is the pie chart that represents the proportion of units characterized by acceptance = 1 with respect to the other ones: this proportion is very low and this leads to a strongly unbalanced data set.

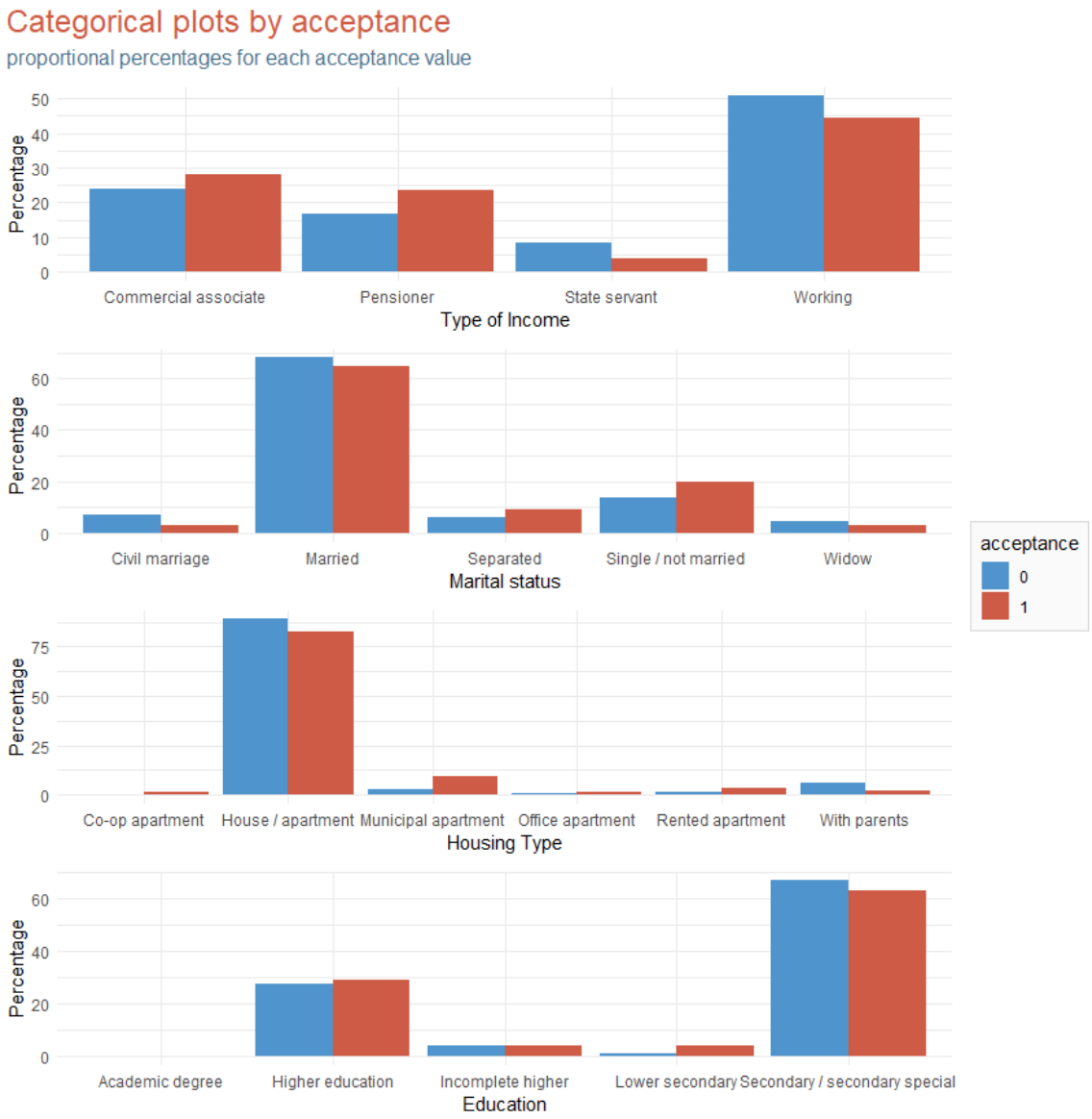


Now the graphical visualization of both numerical and categorical variables will be displayed: for both of them, the most appropriate graphic is the barplot, due to the fact that there is interest in looking at the conditional relative frequencies distributions of such variables concerning the values assumed by "acceptance".



It can be noticed how the distribution of Employed_days is characterized by a strong presence of extreme values. This might suggest potential outliers, but it is an intrinsic characteristic of the dataset. All observations in the dataset with an Employed_days value of -1000 belong to people with a Type of Income equal to pensioner, suggesting that the Employed_days entries are actually default values that appear when the applicant is a pensioner. Other features do not show such peculiar values.

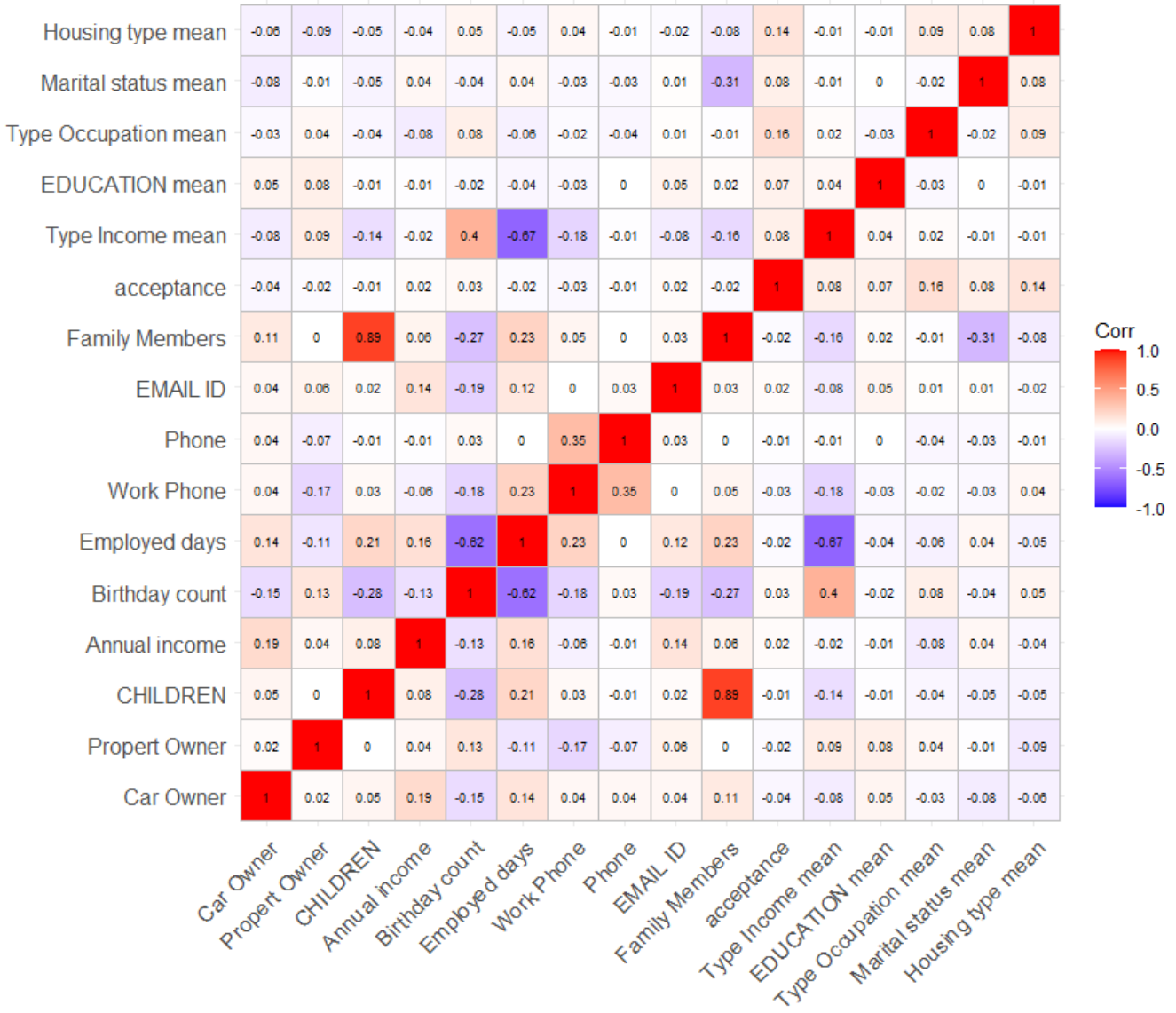
1.2. Multivariate analysis

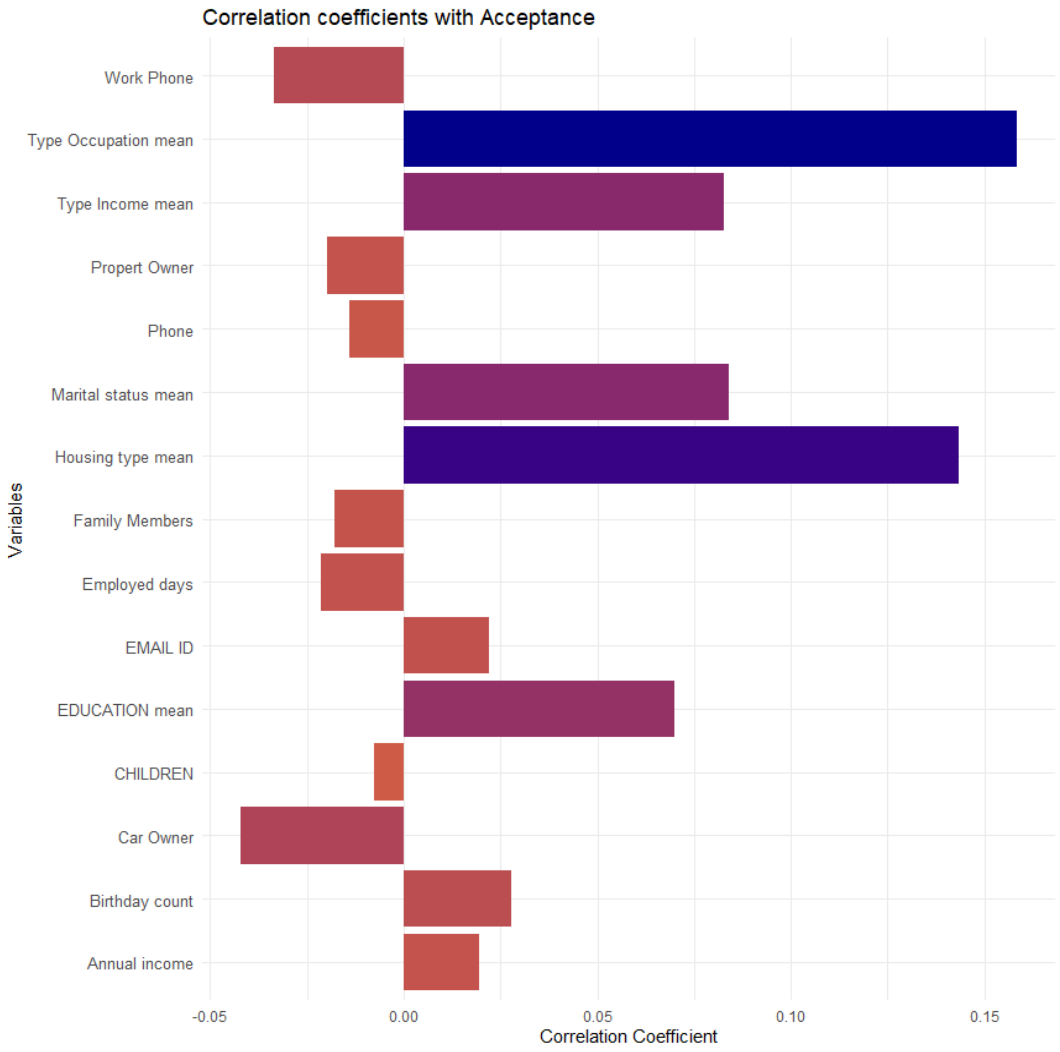


In the sequence of plots in the figure, each category is expressed as a percentage of its frequency in the dataset, grouped by acceptance. The values that are more frequent in the dataset naturally have a higher percentage of frequency. This allows us to observe that there are no significant imbalances between the two acceptance modes for the category values. However, the differences are more pronounced in the case of "Working" within the "Type of Income" variable, where the percentage difference between observations with acceptance = 1 that have "Working" as the Type of Income and those same observations with acceptance = 0 is over 7 percentage points.

The correlation between the variables analyzed and shown in the graph as a matrix of pairwise correlations, denotes a mild correlation amongst the variables, with the exception of Employed Days and Birthday Count with a negative correlation of -0.67, and Family Members

and Children, with a strong correlation of 0.89.





From the graph showing the correlation between the response variable and all the other variables, it is possible to see that the most positive-correlated variables - with respect to "acceptance" - are the type of occupation and the housing price; on the other hand, the ownership of a car is the most negative-correlated variable with respect to the acceptance of a credit card. Finally, other variables have a less significant relationship with the considered target variable.

2. Classification of the Dataset

2.1. Classifiers implementation and model selection

The classification models are now implemented. The models that have been used include the most popular ones for classification problems:

1. Linear Regression
2. Linear Discriminant Analysis (LDA)
3. Quadratic discriminant analysis (QDA)
4. Naive Bayes
5. Decision Tree
6. Random Forest
7. Boosting

The training set has been divided into a validation set (80% of the observations) and a test set (20%), using stratified sampling in order to maintain the original proportions of acceptance classes in both the two sets. All of these models have been trained on the validation set and then applied to the test set: this procedure leads to the creation of seven confusion matrices (one for each fitted classification model). Finally, a set of six indices (accuracy, precision, misclassification error, F1 score, sensitivity and specificity) have been calculated in order to evaluate the overall performances of the models and make comparisons between them.

The results in the validation set have been grouped in a table:

```
##                          accuracy  precision      error         f1 sensitivity
## GLM                     0.8902954 0.25000000 0.10970464 0.07142857  0.04166667
## GLM with SMOTE          0.7004219 0.16901408 0.29957806 0.25263158  0.50000000
## LDA                     0.8818565 0.16666667 0.11814346 0.06666667  0.04166667
## LDA with SMOTE          0.6877637 0.14285714 0.31223629 0.21276596  0.41666667
## QDA                     0.8649789 0.30000000 0.13502110 0.27272727  0.25000000
## QDA with SMOTE          0.8354430 0.22222222 0.16455696 0.23529412  0.25000000
## Naive Bayes             0.8619247 0.30000000 0.13807531 0.26666667  0.24000000
## Naive Bayes with SMOTE  0.8493724 0.26086957 0.15062762 0.25000000  0.24000000
## Random Forest           0.9163180 0.80000000 0.08368201 0.44444444  0.30769231
## Boosting                0.8912134 0.00000000 0.10878661         NA  0.00000000
## Boosting with SMOTE     0.7071130 0.04081633 0.29288703 0.05405405  0.08000000
##                          specificity
## GLM                        0.9859155
## GLM with SMOTE             0.7230047
## LDA                        0.9765258
## LDA with SMOTE             0.7183099
## QDA                        0.9342723
## QDA with SMOTE             0.9014085
## Naive Bayes                0.9345794
## Naive Bayes with SMOTE     0.9205607
## Random Forest              0.9906103
```

```
## Boosting                 0.9953271
## Boosting with SMOTE      0.7803738
```

Now here's an overall evaluation of the quality of the implemented models according to different metrics.

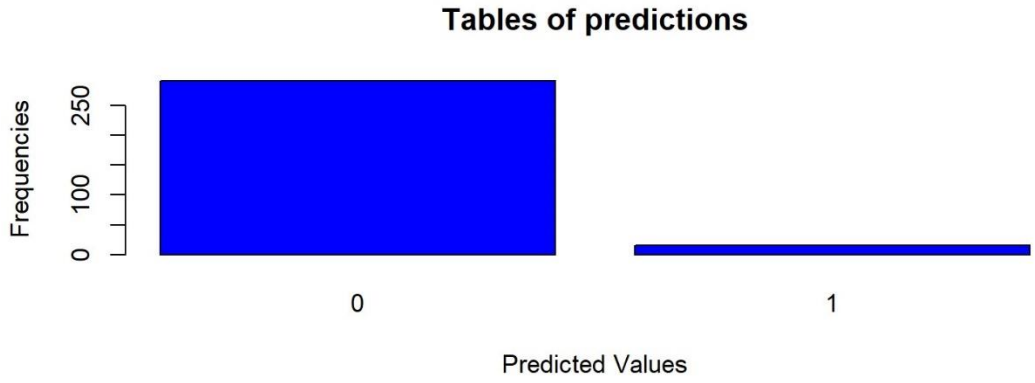
- In terms of **accuracy**, random forest (91.6%) and logistic regression (89%) have performed the best on the imbalanced data set.
- In terms of **precision**, random forest records again had the best performance on the imbalanced data set (83.3% of accurate positive predictions).
- Looking at the **misclassification error**, the random forest again records the lowest proportion of incorrect classified units (8% of incorrect classifications).
- About the **F1 score**, the random forest resulted as the best model with 44%, even though the QDA and Naive Bayes classifier on the imbalanced data set performed nearly good results, with 27% and 26.6% respectively.
- In terms of **sensitivity**, logistic regression and LDA in a balanced data set have provided the highest results (both 51.5% of correctly predicted true instances).
- In terms of **specificity**: boosting and random forest recorded the highest - and similar - results on the imbalanced data set, with proportions of true negatives identified of 99.5% for the first model and 99.1% for the second one.

As a consequence of different index comparisons, the **random forest** results in being the most appropriate model for the classification of this data set since it resulted in the most accurate performances for almost all the provided measures.

2.2. Model assessment on the test set and conclusions

Since the random forest has been selected as the most appropriate model in the training data, it will be now applied to the test data set, verifying its prediction capabilities.

```
##
##   0   1
## 291  16
```



The barplot shows that, according to the prediction on the test set performed by the random forest model, around the 95.08% of individuals's request for a credit card will likely accepted against the remaining 4.92% who instead are likely to get it rejected.

It is worth noting that 2 observations of the test set have not been predicted for the lack of the values for GENDER in the set. Given the impossibility to address this issue, the observations have been removed.

In conclusion, the classification model predicts an overall major probability for an individual to get a credit card.

Technical Appendix

1. Data cleaning

A preliminary process of data cleaning has been performed on the data set: the aim is to deal with missing data and to make some transformations on the categorical variables in order to make them able to be implemented in a regression model (which requires only numerical variables).

A) Firstly, a set of packages has been downloaded in order to simplify some procedural steps related to the circulation of the R file and the graphical visualization of the results.

B) The next step is to put into practice the division of the dataset into two subsets: the training data and the test data, with the second one lacking the target variable "acceptance".

C) Some variable transformations have been performed: C1) "birthday_count" and "employed_days" are expressed in years by dividing them by 365 and multiplying times -1 as well because it is easier to conceive these variables as characterized by positive supports rather than negative ones. C2) Two dichotomic variables, "car_owner" and "propert_owner", have been transformed into binary ones due to the fact that their possible modalities, y and n, can be easily associated with 1 and 0.

D) The problem of transformation of the categorical variables into numerical ones is solved through the target encoding procedure, whose basic idea is to calculate the average of the target variable (in this case, "acceptance") for every possible category of the predictors, and to substitute the obtained average values to the original categories as a consequence.

E) The presence of missing values has been identified with respect to two variables: "annual_income" and "birthday_count". There are many approaches to deal with missing values, but the one which will be applied in this case is, for the annual income, the substitution of such values with the average income per category of occupation (expressed by the categorical variable "type_occupation"). Consequently, the missing values for "birthday_count" are substituted with the average values of the age on the basis of the type of annual income.

F) It can be noticed that the data set is strongly unbalanced: this can be evaluated by measuring the proportion of units which assumes value = 0 for the variable "acceptance" (1,098 units, 89% of the units) against the ones which assume value = 1 for the same output variable (140 units, 11% of the observations). The problem will be faced through the usage of the oversampling technique, which solves the issue of incorrect proportion of units attributed to the classes of "acceptance" by artificially creating new observations for the under-represented class (which is the one for which

acceptance = 1) which are similar to the ones that already exist. This is done through the implementation of SMOTE (Synthetic Minority Over-sampling Technique).[1]

2. Interpretation of common classification metrics

In the field of classification problems, it is crucial to understand and accurately interpret various performance metrics. This chapter of the appendix provides detailed explanations of several key metrics: accuracy, precision, misclassification error, F1 score, sensitivity (recall), and specificity, which all have been used in this report.

Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

Accuracy = (TP + TN) / N

It is a general measure of how well a model performs across all classes. While it provides a quick overview of model performance, it can be misleading in the case of imbalanced datasets where one class dominates. High accuracy in such cases might not necessarily reflect good performance in the minority class.

Precision, also known as Positive Predictive Value, is the proportion of true positive results among all positive results predicted by the model.

Precision = TP / TP + FP

This index is crucial when the cost of false positives is high. It indicates the accuracy of the positive predictions.

Misclassification error, or error rate, is the proportion of incorrect predictions (both false positives and false negatives) out of the total number of cases examined.

Misclassification Error = (FP + FN) / N

[1] This technique aims to reduce the inner imbalance in the distribution of the observations in the two classes by essentially generating new observations for the minority class, characterized by a lower amount of units. Through the steps of neighbour selection (in which, considering the observations in the minority class, the distance between that observation and its nearest neighbours is calculated), synthetic example generation (in which, for each observation in the minority class, new observations are created on the basis of some of the nearest neighbours randomly selected) and the addition of the new examples (last step in which the considered new observations are added to the training set, increasing the number of the units contained in the minority class), a new balanced data (or training) set is created, hopefully improving the quality of the statistical learning techniques that will be applied (Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P (2002). SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, 16, 321–357).

This metric provides the rate at which the model makes incorrect predictions. It is simply the complement of accuracy (i.e., 1 - Accuracy). A lower misclassification error indicates better model performance.

The **F1 score** is the harmonic mean of precision and recall (sensitivity). It provides a balance between the precision and the recall.

F1 Score= 2 x [(Precision x Recall) / (Precision + Recall)]

This index is particularly useful when dealing with imbalanced datasets. It gives a single metric that considers both false positives and false negatives, making it a good measure for evaluating the performance of a model in such contexts.

Sensitivity, also known as Recall or True Positive Rate, is the proportion of true positive results out of all actual positive cases.

Sensitivity = TP / TP + FN

It measures the model's ability to correctly identify all positive cases.

Specificity, also known as True Negative Rate, is the proportion of true negative results out of all actual negative cases.

Specificity = TN / TN + FP

It measures the model's ability to correctly identify all negative cases. It is important in contexts where the cost of a false positive is high.

Each of these metrics offers unique insights into the performance of a classification model. Accuracy provides a general overview but can be misleading with imbalanced datasets. Precision and recall (sensitivity) offer deeper insights into the performance of positive predictions, which can be balanced using the F1 score. Misclassification error provides a straightforward measure of prediction errors, and specificity helps understand performance on negative predictions. When evaluating a model, it is essential to consider these metrics together to gain a comprehensive understanding of its performance.