Analysis of logistic regression results over banking dataset

**Introduction**

In this project we've analysed an open source dataset called Bank Marketing Data Set. The data is obtained from direct contact with clients (often calls) in order to assess if they would or would not open a term deposit. The purpose of this project is to prove that we can build a precise model for predicting the outcome by using logistic regression.

**Data analysis**

The dataset contains 45211 entries with 17 columns that could be split in 6 categorical features, 7 numeric features and 4 binary features, although the least could be seen also as categorical features. To better understand what each column signifies, we present you the following table:

| **Attributes** | **Kind** | **Description** | **Values** |
| --- | --- | --- | --- |
| age | numeric | age of client | values between 18 and 95 |
| job | categorical | type of job | ’management’, ’technician’, ’entrepreneur’, ’blue-collar’, ’unknown’, ’retired’, ’admin.’, ’services’, ’self-employed’, ’unemployed’, ’housemaid’, ’student’ |
| marital | categorical | marital status | ’divorced’, ’married’, ’single’ |
| education | categorical | degree of education | ’primary’, ’secondary’, ’tertiary’, ’unknown’ |
| default | binary | has credit in default? | ’no’, ’yes’ |
| balance | numeric | account balance | values between -8019 and 102127 |
| housing | binary | has housing loan? | ’no’, ’yes’ |
| loan | binary | has personal loan? | ’no’, ’yes’ |
| contact | categorical | contact communication type | ’cellular’, ’telephone, ’unknown’ |
| day | numeric | day in month | Values between 1 and 31 |
| month | categorical | last contact month of year | ’Jan’, ’Feb’, ’Mar’, ’Apr’, ’May’, ’Jun’, ’Jul’, ’Aug’, ’Sep’, ’Oct’, ’Nov’, ’Dec’ |
| duration | numeric | last contact duration, in seconds | values between 0 and 4918 |
| campaign | numeric | number of contacts performed during this campaign and for this client | values between 1 and 63 |
| p-days | numeric | number of days that passed by after the client was last contacted from a previous campaign | values between -1 and 871 |
| previous | numeric | number of contacts performed before this campaign and for this client | values between 0 and 275 |
| p-outcome | categorical | outcome of the previous marketing campaign | ’failure’, ’other’, ’success’, ’unknown’ |
| y | binary | has the client subscribed a term deposit? | ’no’, ’yes’ |

In the next step we will begin to manipulate the data so that we can apply the logistic regression on it.

**Data manipulation**

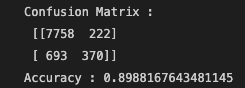
In order for us to be able to analyze the data properly we need to convert all the data into numerical values, so we had to manipulate the data for both binary and categorical values:

* For binary features we simply converted the series values from "yes" to "1" and the "no" to "2"
* For categorical features we used the "get\_dummies()" method from pandas library which creates a series for each categorical value.

Also, after analysing the variance of the mean regarding some categories, we decided to drop the "marital" one because it seemed that it didn't seem a strong predictor for the outcome variable.

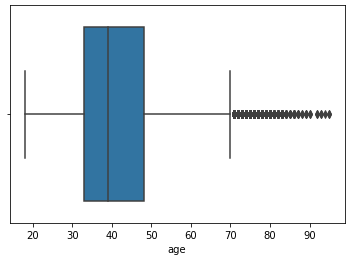
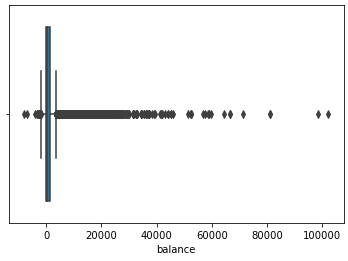


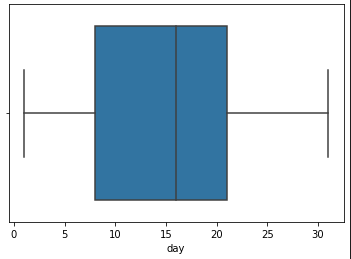
After this process, we now have 70 columns and we move on to the next step, which is the results of the logistic regression without the removal of the irrelevant data. For this, we used some methods from the "sklearn" library. As input we select all the columns except "Target" and as output the "Target" column. We used 20%-80% test to train ratio and so we obtained the following accuracy:



**Optimization**

To better visualise the outliers for some series we used a boxplot, which is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3) and "maximum".





For removing the outliers we calculated the z-score for the series that were most affected by outliers. By using this we observed there are no outliers for the "day" series, while "balance" had a lot. A z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units. The z-score is positive if the value lies above the mean, and negative if it lies below the mean. After calculating this score for all the numerical values we removed all those that resulted in values greater than 3 or less than -3 respectively, thus removing almost 5000 entries.

Moreover, we analyzed the p value for all the series. A p-value, or probability value, is a number describing how likely it is that your data would have occurred by random chance (i.e. that the null hypothesis is true). The null hypothesis states that there is no relationship between the two variables being studied (one variable does not affect the other). It states the results are due to chance and are not significant in terms of supporting the idea being investigated. Thus, the null hypothesis assumes that whatever you are trying to prove did not happen.

The level of statistical significance is often expressed as a *p*-value between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

* A *p*-value less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis.
* A *p*-value higher than 0.05 (> 0.05) is not statistically significant and indicates strong evidence for the null hypothesis. This means we retain the null hypothesis and reject the alternative hypothesis. You should note that you cannot accept the null hypothesis, we can only reject the null or fail to reject it.

Thus we removed all the data that has a p-value higher than 0.05, and after this whole process we arrived at the following results:

