**Data Scientist Challenge**

**Guillermo Fremd – 24th July 2020**

**Part 1: Explore the Data**

Before analysing individual variables, I decided to visualise the kurtosis and the skewness of all variables in order to observe whether a few (or many) of them presented anomalous distributions. While skewness measures the symmetry of a distribution, kurtosis indicates if the data is heavy-tailed or light-tailed. A distribution is considered to be non-normal when its skewness or kurtosis are above 1 or below -1.

A screenshot of a cell phone

Description automatically generated

The scatterplot above shows that there is a large number of variables presenting non-normal kurtosis and skewness. Interestingly, while we can observe variables with kurtosis both above 1 and below -1, the skewness of the variables in the dataset appear to be either close or above 0, but we do not see any variables presenting significantly negative skewness. A positive skewness, such as the one that we can observe in many of the variables above, indicates that the long tail in the distribution is on the positive side of the median.

It should be noted, however, that kurtosis and skewness are very sensible to the presence of outliers, and therefore it is useful to observe both measures once we neutralise their effect in the dataset. I have defined as outliers those values that are 1.5 interquartile range above the third quartile or below the first quartile, and for those methods than allow the user to choose whether to include or exclude the outliers (*without\_outliers=True*), they are substituted by the mean of the corresponding variable.

A screenshot of a cell phone

Description automatically generated

Please note that when a user chooses to see the name of the variables that present abnormal shapes (using *identify\_abnormal=True*), apart from showing them in the scatterplot, the kurtosis and skewness of all variables with non-normal distributions are printed, facilitating their identification.

The plot above shows that, after removing the outliers, variables 653 presents the largest skewness value. This is coherent with the histogram and the boxplot of the variable, presented below, which show the presence of a long tail to the right :

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

Describe\_Variable provide us with the following summary statistics of this variable:

* Number of observations: 272
* Range: (-1.547458, 1.254303)
* Mean: -0.32896379044117646
* Variance: 0.21541593104913304
* Standard deviation: 0.464129218051539
* Skewness: 1.2288821239162357
* Kurtosis: 0.9503020550412944

**Part 2: Principal Components Analysis**

The Principal Components Analysis plot is shown below:

**A screenshot of a cell phone

Description automatically generated**

**Part 3 (a) - Calculate a statistic for every variable that describes its relationship with the class column**

Considering that Class is a dichotomous variable and that the rest of the variables are continuous, I decided to evaluate the Point Biserial Correlation to asses which of the variables presents the strongest relationship with Class. Variable\_1497 presents a point Biserial Correlation of 0.36 with the Class variable (p-value< 0.01).

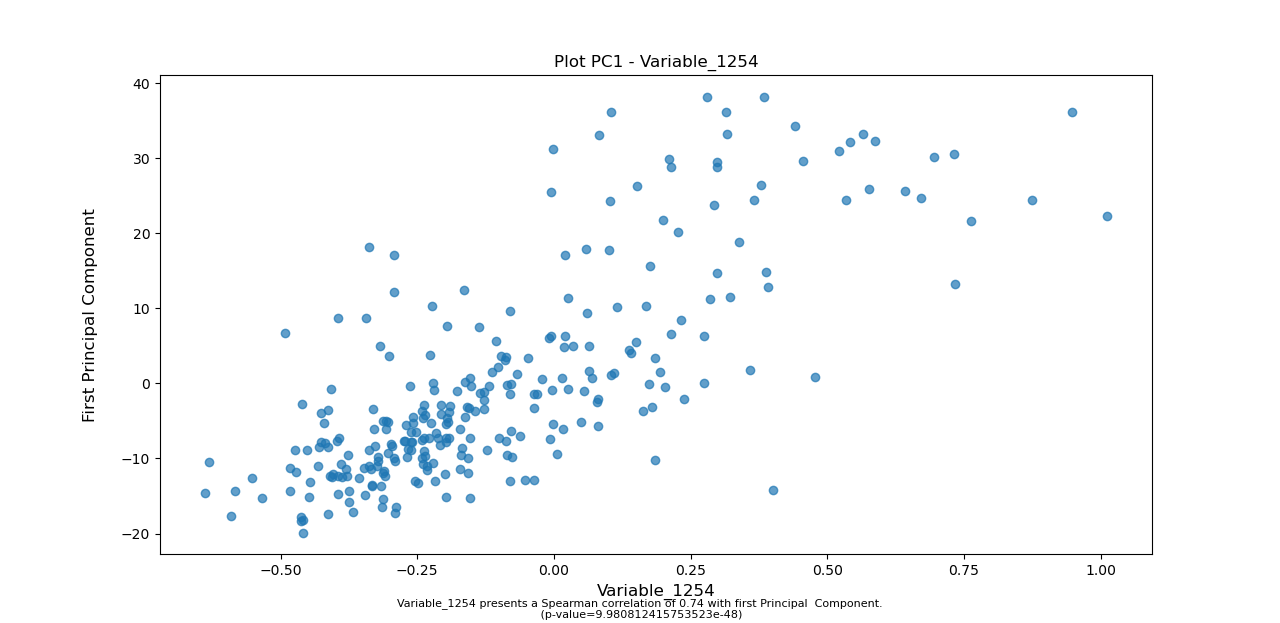
Besides the correlation coefficient, the relationship that exists between variable 1497 and the Class is observed also in the boxplot below, which indicates that, while not perfect, there is a significant relation between both variables, with those observations that belong to Class 1 presenting a mean significantly higher of variable 1497 than those observations In Class 0.

A picture containing clock

Description automatically generated

**Part 3 (b) - Calculate a statistic for every variable that describes its relationship with** **PC1**

In this case, and given that both elements are continuous variables, I calculated the Spearman correlation between every variable and the PC1. Variable 1254 is the one that presents the strongest relationship with PC1, with a Spearman correlation of 0.74 (P value < 0.01). The positive direction of the relationship between this variable and PC1 is clearly observed in the plot below:



**Part 4 -** **Create a classifier model**

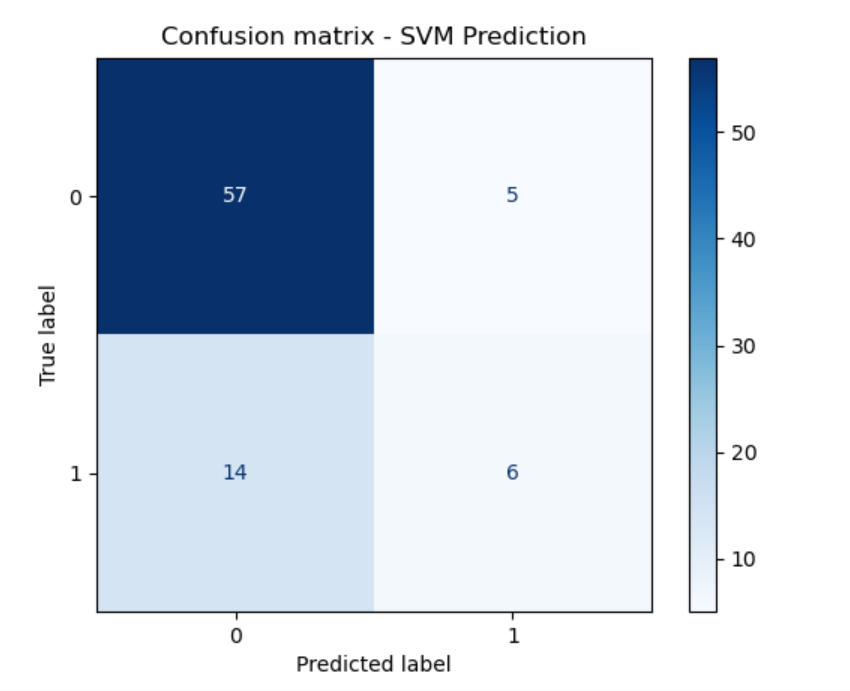
I created an SVM classifier using all the variables, and Grid Search to cross validate and tune its parameters:

* Cost: 1, 10, 100 or 1000
* Kernel: lineal of radial
* Gamma (for the radial kernel only): 0.001 or 0.01

The grid search returned the following results:

* Best Cost: 100
* Best Kernel: rbf
* Best Gamma: 0.001

Subsequently, I divided the data in two subsets: one for training the SVM and one for testing it. After fitting the SVM with the training data, the fitted SVM was used to predict the value of the class in the testing dataset. The classifier presented an accuracy of 76.9%, and an error rate of 23.1%. However, the confusion matrix shown below indicates a precision of 0.54 (6/11) and a recall of 0.3 (6/20), which are not very strong.



Should I have more time, I would attempt to try other classifiers, such as Logistic Regression or Random Forests. It is possible, also, that using only a limited number of features having a strong relationship with the Class (as measured in part 3) may yield a better accuracy in the test dataset.

It would be relevant to understand, also, if the end-users are more interested in reducing the number of false positives or false negatives, as it would allow us to tune the threshold of our classifiers to address such aim. Finally, it would also be insightful to have information about what the variables are about; the analysis and the use of machine learning achieve their full potential when they are coupled with knowledge on the matters they try to address.