Reports on some conclusions made after analyzing the data

Looking at our database we can see that although we have a lot of columns with a lot of information, the number of customers is not very large, we have 10000 data and 18 columns, we can verify that we have categorical variables and continuous variables, something that facilitates the our work is that we do not have null values, which does not require treatment.

Speaking about the analysis of our data, we can see that we have an absolute correlation between the Complain variables and our Target variable, with that we exclude the Complain variable so that we can run our models, looking at the exploratory analysis we can verify that a good part of the our data are well distributed, we can see some peculiarities such as the majority of customers are from France, usually between 1 and 2 products, a good part of our customers do not have money in their account, and we can see that the age of our customers meets a distribution normal, a very important thing we saw is that our Target variable is unbalanced.

We have some small Outliers but it is nothing that influences or harms our data, when we look at our Bivariate analysis a variable that caught my attention was the Age variable, when we look at it we can see that older people are more prone to have churn.

Entering the Machine Learning part, we removed some variables that do not make sense for our models, we transformed our Categorical variables into Continuous variables using the OneHot Label Encoder (I had a better result using OneHot than the Label Encoder), we separated our data into training with 70% of our data and testing with 30%, as mentioned earlier our Target variable is unbalanced, so after running the models I was able to confirm this, without balancing the data our results were good but the model learned only the model with negative churn and not the result we seek.

After balancing our Target class and running the Machine Learning models we got some good results and some not so much, most of the models had a satisfactory result learning both the negative result and the positive result, and others learned only the negative result and not the our objective result, in terms of accuracy, the best model was the Random Forest with 83.63%, but the model that best managed to predict our Target variable was the Ada Boost model, where we had 78% accuracy but a greater accuracy when we speak in predicting Churn.

When we look at the most important variables, we have Age, NumOfProducts and Balance, with Age being the main one, which confirms what we saw in our exploratory analysis.