Week 9 deliverables

Group Detail

Name: Jose Vicente Solorzano

Email: solorzano.vco@gmail.com

Country: Argentina

College: Montpellier Business School

Specialization: Data Science

What are the problems found in data?

- There are instances with an "unknown" string in some features. This is going to be treated as null values. (Features with "unknown": job 0.8%, marital 0.2%, education 4.2%, default 20.9%, housing 2.4% and loan 2.4%).
- Target "y" is imbalanced. No = 89% and yes = 11%.
- There is a lot of outliers in the features.
- "Duration" is left skewed. Equally, it is a feature to use with careful because after the end of the call y is obviously known. It could not be useful, so for now, the skewed is not a problem.

Problems Approach

For nulls, we will analyze if it is more convenient imputation or deletion. We have to be careful not to cause bias while imputing. We will experiment with both approaches in order to choose the best for the model.

Outliers going to be treated with IQR and WOE.

Regarding the unbalance of the target "y", we will consider two approaches. In one hand, a more passive approach, we will not generate nothing new. We will just consider the right metric for the model, in order to avoid the impact of the unbalance (recall, F1 score, ROC, etc).

In the other hand, we will apply SMOTE (Synthetic minority over-sampling technique), a common solution for this kind of problems in machine learning classification models.

Gitub Repo Link

https://github.com/jvsolorzano96/bank_marketing_campaign/tree/main/Week%209%20-%20Data%20Cleaning%20and%20transformation