LISIUM13: 30- Group Project

Week 8: Project Deliverables

Team member's details :

Group Name : Banana Peels

Names: Devika Chandnani, Dylan Hugey, Camillo, Farha Jabin Oyshee

Email: devika.chandnani@gmail.com, dylan94539@gmail.com, fj.oyshee@gmail.com

Country: United States,Bangladesh

College/Company: Fordham University,Independent University, Bangladesh, University of Illinois at Urbana Champaign

Specialization: Data Science

Link to the data: UCI

Data understanding

We have been provided a large data set provided by UCI and collected from a Portuguese banking institution. There are a number of different data points which were collected from a number of clients which will help determine if they purchased the bank's new term deposit product. There are a few different versions of the dataset where some are the full version with all inputs and others are more compact with fewer lines of data.

What type of data do we have for the analysis

The data is provided to us in the form of a csv. The columns which we are provided include both categorical and numerical data. For example, the columns which describe each user's job, marital status, and education are categorical and the columns describing their current balance and number of times contacted are numerical.

There are also 3 different categories of data within the large dataset labeled bank client data, social and economic context attributes, and other attributes.

What are the problems in the data (number of NA values, outliers , skewed etc)

The data thankfully has no NA values and does not need to be cleaned in that regard. Within the data description it is said that there were multiple calls to each user and this may be reflected within the data. However, we will assume that no two rows correspond to the same client and that the data already handles duplicate clients.

What approaches are you trying to apply on your data set to overcome problems like NA value, outlier etc and why?

We have not yet explored the quantitative data to determine if there are any outliers; however, in the case that there are outliers we would most likely conduct two experiments or create two models. One with the outliers and one without the outliers to determine if they have a significant impact on the final results on the model.