Campus Tech Data Science

Coding Assignment

**Notes:**

* You can use any analysis software which you are comfortable with for e.g. Python, R, etc.
* Share both the analysis results and source code when you are done.
* Your analysis results can be a list of observations, tables, supporting visualizations etc.
* The source code should be such that it can be executed to reproduce the analysis results.
* It is important that you share all steps not just the results to illustrate your thought process.

**DataSet2**

The file “DataSet2.csv” consists of data obtained from an experiment. The output of the experiment is indicated in the column “y”. The output is either 0 or 1. The other columns indicated with a leading “x” are the input variables. Imagine you are tasked with building a predictive model which takes the x’s as input and produces y as the output.

Share your code and any supporting write-up (in-line comments are also ok) for this predictive model. Your code and the write-up should be enough to reproduce the steps which were taken. Do make sure to include every step you took (for e.g. reading the data, any pre-processing, any visualization, any exploratory analysis, key observations, model training etc.) and mention any assumptions you made.

Imagine that this dataset is from a project and after your modeling the code and comments your share should be enough to convey:

1. To your team: why it is the best model? Give an example of a data point that model doesn’t do well and explain the characteristics of that observation. How- do you see the limitations of your model? How do you plan to improve it?
2. Assume your model performance is acceptable.
3. A few days after you finish the initial model, you learn new information that the value of the X2 feature makes sense only if it is positive (i.e. X2 > 0 should be in the model instead of original X2). Show how you would address this issue.
4. Later, you additionally learn that X4 should’ve been used as an indicator variable (e.g. X4 that is greater than a certain threshold should be treated differently than X4 values below that threshold). However, you don’t know the value of threshold. Adapt your code to support a systematic way of finding threshold which leads to improved model performance.
5. Now that you realize feature engineering is a frequent need for improving our models, you want to create a general framework to tackle such situations (e.g. conversion to indicators, defining certain thresholds, etc.). What function/module do you develop to enable user-defined column conversions for given thresholds? Suppose thresholds are given to you as arguments to your function.