Case Study Description

A company launches several products yearly in different markets. They would like to have a tool that helps them identify which of their products will be successful or not. To develop such solution, they are providing 1716 hypothetical products that have been launched already. The dataset includes both categorical and numerical predictors along with the response variable (Market share). The response variable represents the market share of the product one year after launch. A threshold of 0.7% is used to classify if a product is successful or not. Products with market share above the threshold are considered as successes and the ones below the threshold as failures. We ask you to create a model that predicts the success of a product and associated presentation that is going to be pitched in-front of the business owner.

Data Sources

Use the following files for the analysis:

- score_data.xlsx – data you are going to use to score your model
- use_case_data.xlsx – training data

Objectives

1. Import the data into a Python-based environment.
2. Preprocess data to select the most informative predictors and apply any necessary transformations.
3. Train a model that predicts the market share after one year in the market. Feel free to use any modelling technique suitable for the data provided.
4. Evaluate the model performance using the area under and the confidence interval to measure how confident we are that the true success probability will fall within the range of the interval. Feel free to use other metrics suitable for this use case and makes your model as robust as possible. You can experiment with different values of threshold that potentially increase the accuracy of your model.
5. Score the new products in the file score_data.xlsx using the model you created.
6. Create a presentation (15 minutes + 5 minutes for Q&A) with your results for a managerial audience with some technical knowledge using necessary visualizations. Your presentation must include which predictors were selected and how your model performed. Add any recommendations and next steps if you had to continue the analysis and what additional data sources you could possibly use during the next sprint.
7. In the technical part of discussion, you'll have to justify and explain:
   a. Any assumptions you made during the preprocessing phase,
   b. How you split datasets into training and testing,
   c. Techniques that you used for modelling,
   d. The metric that you used for quantifying the performance of your model.
8. Provide us with your presentation and well commented source code in Python (preferably via GitHub).