TECHNICAL TEST

Narrative:
In this assignment you will come across concepts and datasets similar to those on which MTS Data Scientists/ML Engineer works on.

You will read a research paper and use related dataset to develop a classification model.

The goal of the assignment is not to develop as good model as possible, but rather to demonstrate your ability to understand the topic and how you approach to such tasks.

Go to Kaggle and search for Dataset "Beat The Bookie: Odds Series Football Dataset" https://www.kaggle.com/datasets/austro/beat-the-bookie-worldwide-football-dataset
There's a link to the research paper "Beating the bookies with their own numbers - and how the online sports betting market is rigged", by Lisandro Kaunitz, Shenjun Zhong and Javier Kreiner https://arxiv.org/pdf/1710.02824.pdf

Objectives:

- read the paper, the content will be discussed on next interview
- download the Dataset (relevant is only closing_odds.csv.gz)
- prepare some descriptive analysis to get familiar with the closing odds dataset (e.g. number of matches per league, distribution of odds overall and per one team from English Premier League, count of top bookie per bookmaker, ...)
- develop a model for "probability that a match will be picked by the betting strategy (eq. 7) before closing odds are known".

    - i.e. we want to assess probability that a match will be picked by the betting strategy described in the article, at the time when the match is scheduled (it is known who will play against who and when) but no odds for that match are offered on the market yet
    - this means that info about odds can't be used for a current match, but it can be used for previous matches (odds are known for past matches but not for a recently scheduled match)
    - the model should have AUC at least 0.51, but the objective is not to develop as good model as possible, but rather to demonstrate:
        - understanding the concepts around development of a model
        - forming sensible features
        - evaluating quality of the model
        - capability to present all these in a notebook

Instructions:

Data analysis and model development should be done in a notebook (python). Compressed notebook should be uploaded to personal GitHub and link sent to s.gabor@sportradar.com (or the compressed file directly) in 1 week time.