Counterfeit Medicines Sales Prediction

Counterfeit medicines are fake medicines which are either contaminated or contain wrong or no active ingredient. They could have the right active ingredient but at the wrong dose. Counterfeit drugs are illegal and are harmful to health. 10% of the world's medicine is counterfeit and the problem is even worse in developing countries. Up to 30% of medicines in developing countries are counterfeit.

Millions of pills, bottles and sachets of counterfeit and illegal medicines are being traded across the world. The World Health Organization (WHO) is working with International Criminal Police Organization (Interpol) to dislodge the criminal networks raking in billions of dollars from this cynical trade.

Despite all these efforts, counterfeit medicine selling rackets don’t seem to stop popping here and there. It has become a challenge to deploy resources to counter these; without spreading them too thin and eventually rendering them ineffective. Government has decided that they should focus on illegal operations of high net worth first instead of trying to control all of them. In order to do that they have collected data which will help them to predict sales figures given an illegal operation's characteristics.

Data Files

Train Dataset = counterfeit_train.csv

Test Dataset = counterfeit_test.csv

Formal Problem Statement

Variable names are self explanatory.

Your task here is to build a predictive model for predicting sales figures given other information related to counterfeit medicine selling operations. You need to build your model on the train dataset. Test dataset does not have a response column; you need to predict those values and submit it in a csv format.

Evaluation Criterion

Part 1:

You will first attempt Part 1 of this project which is a quiz. You can access it through LMS. This quiz needs to be answered based on exploration of the dataset given and some generic questions about algorithms discussed in the course. Consider only the training dataset for data cleaning and exploration to answer the quiz questions. There will be 10 questions of which you need to get at least 7 correct in order to pass the project.

Part 2:

Here you work on creating the machine learning models and choosing the one which gives the best performance. You can refer to the Project Process Guides provided in LMS to understand how to approach and work on a project.

For this project, score will be calculated as:

Score = 1-(MAE/1660)

where MAE is mean absolute error on test file. You need to score more than 0.5 in order to pass the project submission. Don't read too much into score formulation, it is just to scale MAE. You just need to focus on minimizing MAE.

Submission:

Submission CSV should resemble the file:

Sample Submission = 'sample_submission.csv'

Column names, value types should be exactly the same. Also number of rows in the submission csv should be exactly the same as test data. If this is not taken care of, your submission will not be graded.

You can make as many submissions you want if you want. [We might ask you to submit the script which was used to generate the submission at any time].

In order to clear this project, you are required to clear both, Part 1 as well as Part 2 of this assignment.

Wish you all the best!