FRISS Data Scientist Case

For this case you will try to find fraudulent cases in the Database with Historical Claims of our customer “FRISS Insurances”. FRISS Insurances is using only one Line of Business, which is Motor. This means that every claim in the Database will contain an object, which is a car.

There are three CSV files needed for the case, called “fraud_cases.csv”, “FRISS_ClaimHistory_training.csv” and “FRISS_ClaimHistory_test.csv”. Note that the data is made “tidy” already, meaning that each row in the data represents a claim and each column in the data represents a variable belonging to the claim.

You are asked to solve the problems stated below. You are free to report the results in a way that you think is suitable for the problem, but just remember to keep the scripts that you coded in.

1) Find out which of the 80.000 cases in the training data are fraudulent cases, by mapping the file “fraud_cases.csv” to the training data. How many fraud cases can you find?
2) The customer wants to know when the number of days between the occurrence of the claim and the report date of the claim is high. Based on this information, an indicator at FRISS can be made. Explain what you think is a high number, based on the training data.
3) Clean the data by removing irrelevant features, cleaning up features and by engineering any fun new feature that you might think of.
4) Train a model based on the training data and the responding fraud cases. Choose any model you like (and Google anything you like!). Keep in mind that the data is very imbalanced and keep in mind that you should be able to explain your model to anyone!
5) Classify the test set (20.000 cases), obviously by not using the column “sys_fraud” in the test set.
6) Report the results! (The confusion matrix might be a good way to start).
7) Create a REST API with a ”/score” endpoint where the payload is a single claim from the dataset, and the response is the model’s prediction.
8) Put the REST API in a Docker container that can make predictions on a local computer.

Final Note

This is a chance for you to shine and showcase your technical skills, do the best you can, pay attention to details which you find the most important, and make sure you use the latest and greatest of the technologies required. This is a timeboxed exercise, we understand you cannot finish everything on time, so please focus on what you believe is important. Good luck! 😊