Model and feature manipulation to predict specific anomalies

Oftentimes, a data science model is used to replace an existing model so that human resources can be better spent doing other tasks and to increase the speed at which that process is done.
This requires creating a model that emulates the decision making that the humans were doing before.
This task is to build a modelling process to replace a system that identifies risk events in some spending data.

The data is time-series for many individuals who each have a different number of events associated with them, making this longitudinal or panel data[1]. Each event has several features, so is also multivariate.
We are looking for an unsupervised model but have provided targets (risk_flag) for testing. We believe anomaly detection models are the way to approach this problem.
This is distinct from a regular anomaly detection problem in that we are not actually interested in detecting anomalies, we are interested in detecting risks, which usually do appear as anomalies. Generally all risks will be anomalous, but not all anomalies will be threats **so you need to come up with ways to alter the features and model so that it outputs anomalies that are risks.**
We have provided targets to simulate the human knowledge that we are wanting you to transfer into the models. Usually, this takes the form of collaborating with the existing operations team to produce a model that makes the same decisions that they would make, we have simplified this and just provided the labels.
To reiterate, the target series (risk_flag) is for checking progress of you unsupervised model at detecting threats and not for training.

Below are some tips, we have provided a list of difficulties that you could face due to the data and some possible approaches you could take.

Issues of the data:
- If you were to run anomaly detection on the data as it is now, you would see that many events identified as anomalous are not deemed threats;
- There is great variation in the number of events associated with each individual;
- There are extreme outliers;
- There are some low frequency categorical features (eg events on weekends are much rarer).

Recommendations:
- Look at ways to alter features so that they are more likely to lead to anomaly detection identifying anomalies that are threats instead of just identifying any anomaly;
- Construct new features;
- Look at some methods for dealing with the high variation in the number of events per person (eg hypernetworks[2]).
- Consider removing the time-series element and instead using statistics that encompass the idea of time (eg moving averages);
- Ensemble models to treat different (groups of) features differently.

A good metric to use to check your work is f1, but we are interested in the methodologies employed, not achieving 100% by any metric.

Citations:

(1) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4669300/

(2) https://ora.ox.ac.uk/objects/uuid:59a94527-0a79-4a14-8d37-94c5835e3971

Research papers:

https://www.mdpi.com/1424-8220/23/5/2844

https://www.nature.com/articles/s41598-022-12792-3