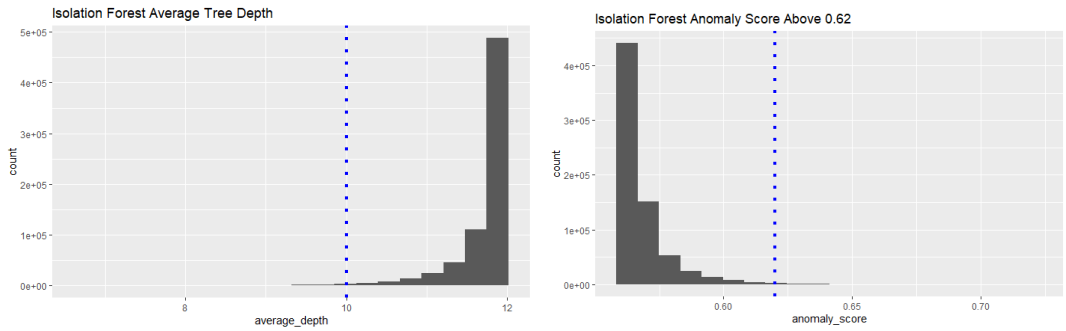
Using Isolation Forest as a surrogate labeling strategy

Emma Wang (06648801), wangj422@wfu.edu, 12/12/2022

Choose threshold

Average depth = 10, score = 0.62. Over 10% of frauds are detected.



Recall	Synthetic Target vs Event Label	Synthetic Target2 vs Event Label
Training	0.1219	0.1147
Testing	0.1308	0.1195

Build Models to Predict Synthetic Target and Event Label

Predict synthetic target

Model	Accuracy	AUC	Precision	Recall	Log loss
Train	0.99994	1.0000	0.99980	0.99185	0.00270
Test	0.99706	0.9986	0.90279	0.67331	0.00747

Predict synthetic target2

Model	Accuracy	AUC	Precision	Recall	Log loss
Train	0.99996	0.99999	1.00000	0.86762	0.00237
Test	0.99732	0.99907	0.93305	0.60625	0.00657

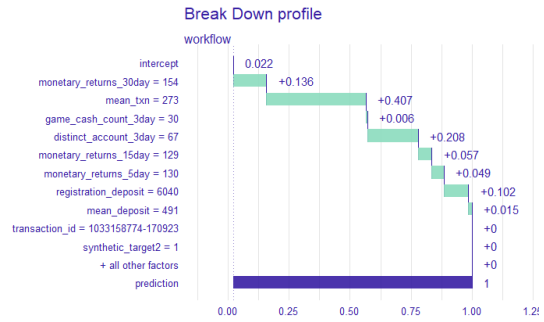
Predict event label

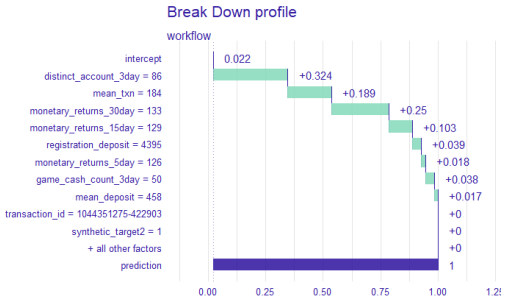
Model	Accuracy	AUC	Precision	Recall	Log loss
Train	0.99565	0.99999	0.99863	0.70645	0.00970
Test	0.99281	0.83293	0.98929	0.52004	0.10217

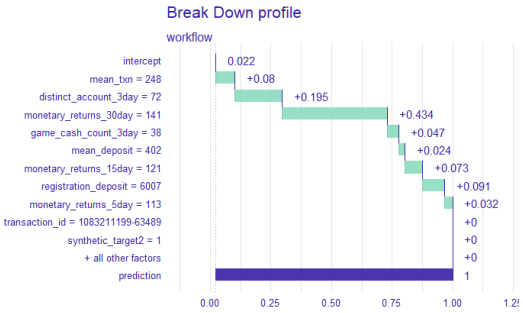
When predicting synthetic target, the recall is 67.33%, compared to 52.00% when predicting event label.

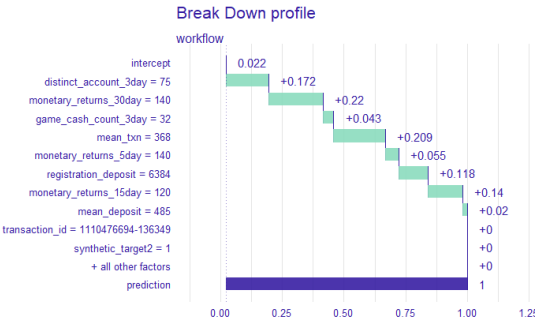
Top TP Records

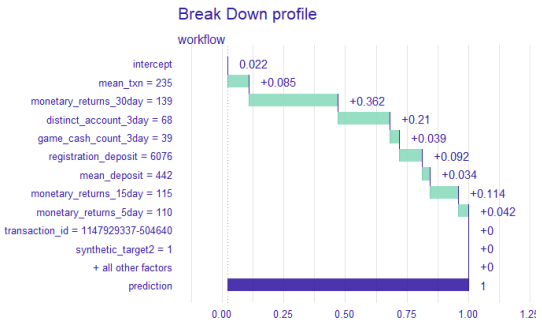
Top 10 TP when predicting Event Label

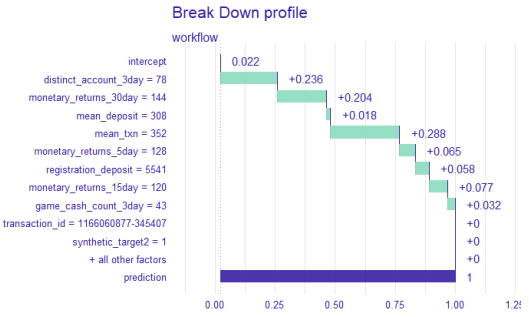


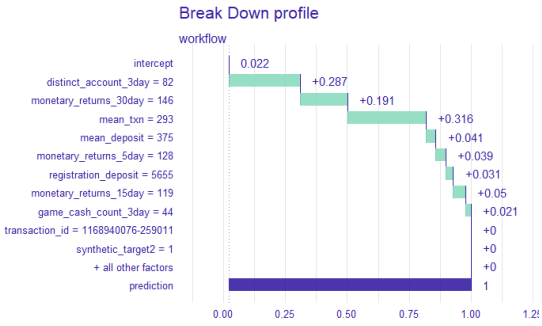


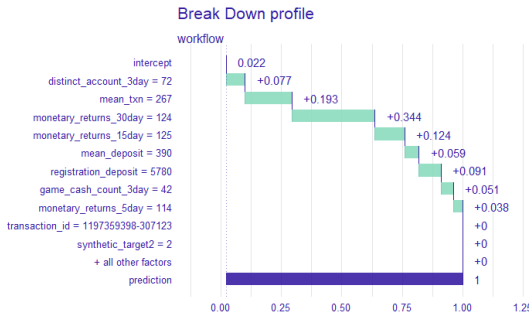


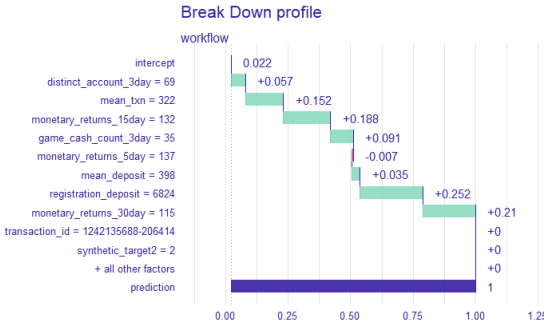


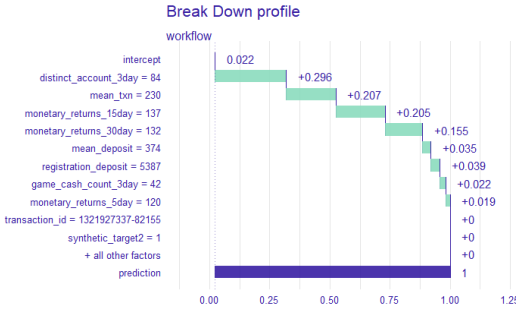




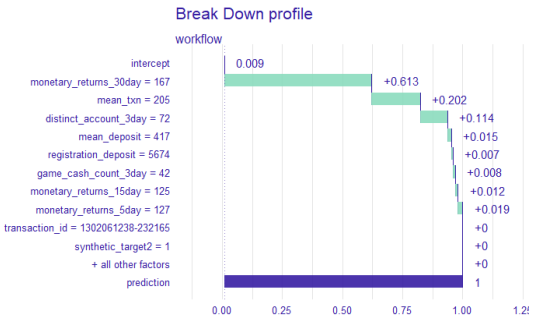


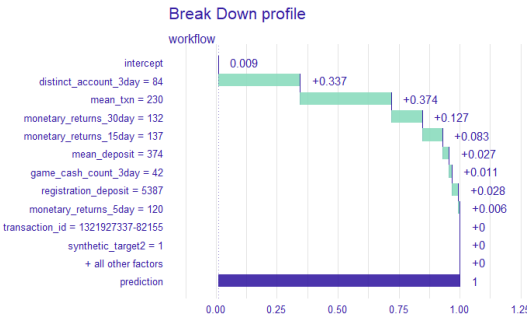


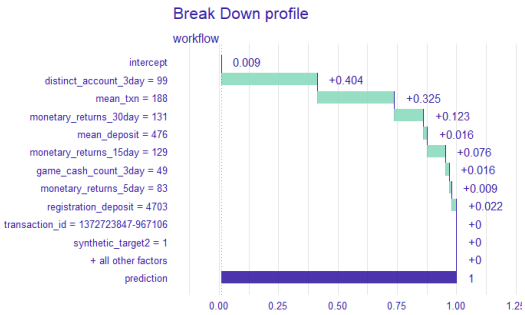


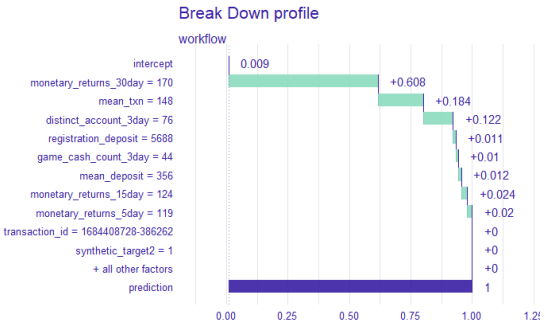


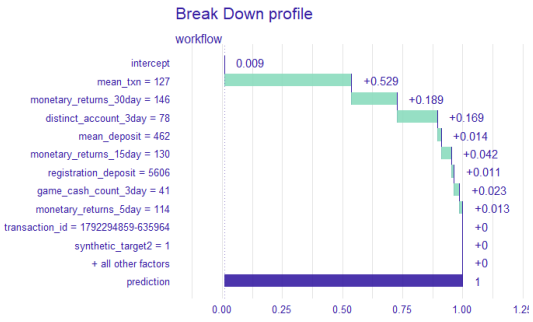
Top 10 TP when predicting Synthetic Target

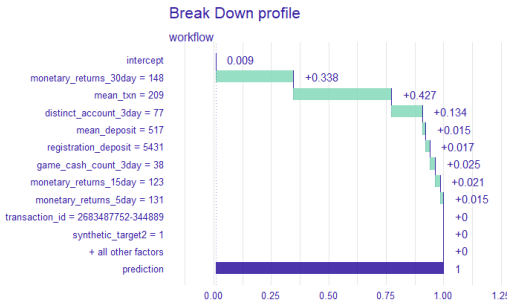


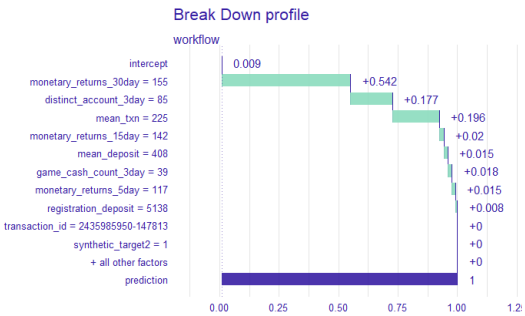


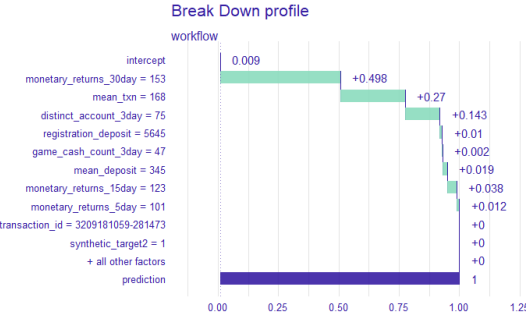


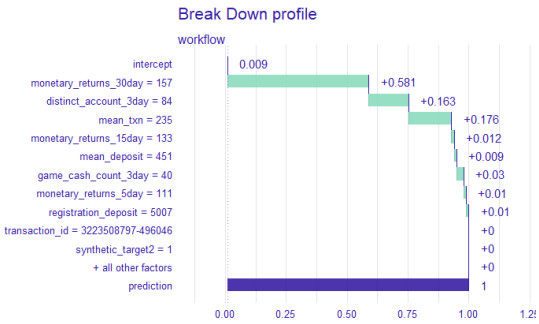


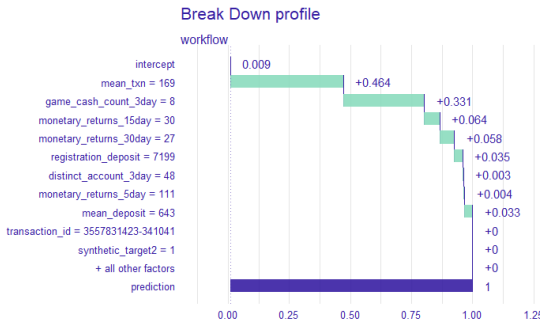












Discussion

The idea of this approach is to equal anomalies to fraud, which is partially true because fraud cases will share some similarities which are different from the legit cases. However, this method can only detect part of the true fraud. This method does save us some time when trying to predict fraud, but this costs our models' accuracy. What's more, it is unrealistic to detect over half of the total frauds since it will need a bigger threshold of average depth when assigning anomaly rules, which will make the anomalies pretty normal (because it will be a very big group).