As a robustness check, we selected another set of random hyperparameters (labelled as Set 2; see Table 1), and re-ran the experiments. The short-term, mid-term, and long-term prediction results are presented in Tables 2–4. The one-sided paired t-test results for the F-1 scores of each algorithm are shown in Table 5, indicating that three sets of features based on IRL are significantly superior to the benchmarks across all three periods. The consistent results between Set 1 (i.e., the hyperparameters in the submission) and Set 2 (i.e., the alternative hyperparameters) further corroborate the validity of our results.

Table 1. Two Sets of Hyperparameters

Prediction Model	***Set 1 (i.e., original hyperparameter configuration)***	***Set 2***
BNB	*alpha:1.0* *fit_prior=True*	*alpha:1.5* *fit_prior=False*
LR	*penalty='l2'* *C=1.0* *solver='lbfgs'*	*penalty='l1'* *C=0.6* *solver='liblinear'*
SVM	*C=1.0* *kernel='rbf'* *gamma='scale'*	*C=0.7* *kernel='rbf'* *gamma='scale'*
KNN	*n_neighbors=5* *weights='uniform'*	*n_neighbors=10* *weights='distance'*
RF	*n_estimators=100* *max_depth=None* *min_samples_split=2*	*n_estimators=70* *max_depth=9* *min_samples_split=3*
XGBoost	*n_estimators=100* *learning_rate=0.3* *max_depth=6* *objective='reg:squarederror'*	*n_estimators=50* *learning_rate=0.1* *max_depth=8* *objective='reg:squarederror'*
AdaBoost	*n_estimators=50* *learning_rate=1.0* *algorithm='SAMME.R'*	*n_estimators=100* *learning_rate=0.5* *algorithm='SAMME'*
GBDT	*n_estimators=100* *learning_rate=0.1* *max_depth=3* *subsample=1.0*	*n_estimators=200* *learning_rate=0.2* *max_depth=6* *subsample=0.5*
MLP	*hidden_layer_sizes=(100,)* *activation='relu'* *solver='adam'* *max_iter=200*	*hidden_layer_sizes=(100,50)* *activation='relu'* *solver='adam'* *max_iter=150*
LSTM	*units=50* *activation='tanh'* *optimizer='adam'* *batch_size=32*	*units=64* *activation='relu'* *optimizer='adam'* *batch_size=32*
GRU	*units=50* *activation='tanh'* *optimizer='adam'* *batch_size=32*	*units=32* *activation='tanh'* *optimizer='adam'* *batch_size=64*

Table 2. Short-term prediction of the new set of hyperparameters (Set 2)

Algorithm	IRL All-factor				IRL Single-factor			
	Accuracy	Precision	Recall	F-1	Accuracy	Precision	Recall	F-1
BNB	69.29%	63.26%	60.99%	62.10%	70.77%	64.84%	63.68%	64.25%
LR	68.18%	63.28%	54.48%	58.55%	67.90%	66.89%	43.95%	53.04%
SVM	73.73%	68.49%	67.26%	67.87%	71.05%	65.08%	64.35%	64.71%
KNN	83.16%	78.09%	82.29%	80.13%	85.38%	78.57%	88.79%	83.37%
RF	84.37%	84.89%	75.56%	79.95%	87.05%	88.25%	79.18%	83.45%
XGBoost	85.85%	83.37%	82.06%	82.71%	88.07%	83.51%	88.57%	85.96%
AdaBoost	73.91%	72.65%	58.97%	65.10%	68.91%	65.99%	50.90%	57.47%
GBDT	84.74%	80.88%	82.51%	81.69%	85.85%	82.06%	84.08%	83.06%
MLP	69.94%	70.23%	47.09%	56.38%	72.61%	65.96%	69.51%	67.69%
LSTM	81.41%	77.65%	77.13%	77.39%	80.85%	78.93%	73.09%	75.90%
GRU	78.72%	79.35%	65.47%	71.74%	76.87%	70.59%	75.34%	72.89%
Algorithm	IRL Multi-factor				Benchmark			
	Accuracy	Precision	Recall	F-1	Accuracy	Precision	Recall	F-1
BNB	71.79%	64.84%	69.06%	66.88%	68.27%	65.75%	48.21%	55.63%
LR	65.49%	58.80%	54.71%	56.68%	58.46%	49.65%	47.09%	48.33%
SVM	75.76%	69.91%	72.42%	71.15%	59.39%	50.79%	50.45%	50.62%
KNN	84.34%	78.04%	87.67%	82.58%	38.95%	32.63%	45.07%	37.85%
RF	85.48%	86.58%	76.68%	81.33%	70.86%	64.40%	65.70%	65.04%
XGBoost	85.66%	81.29%	84.75%	82.99%	62.53%	54.61%	54.48%	54.55%
AdaBoost	71.60%	70.62%	53.36%	60.79%	63.27%	55.27%	57.62%	56.42%
GBDT	87.23%	83.62%	85.87%	84.73%	59.20%	50.59%	48.21%	49.37%
MLP	74.56%	70.41%	66.14%	68.21%	51.80%	41.30%	39.91%	40.59%
LSTM	83.26%	78.37%	82.06%	80.18%	73.64%	65.28%	77.13%	70.71%
GRU	80.11%	76.07%	75.56%	75.82%	65.95%	57.83%	64.57%	61.02%

Table 3. Mid-term prediction of the new set of hyperparameters (Set 2)

Algorithm	IRL All-factor				IRL Single-factor			
	Accuracy	Precision	Recall	F-1	Accuracy	Precision	Recall	F-1
BNB	66.42%	68.27%	72.08%	70.12%	63.45%	66.61%	66.50%	66.55%
LR	56.71%	59.75%	63.79%	61.70%	56.24%	58.73%	67.17%	62.67%
SVM	56.71%	58.98%	68.36%	63.32%	73.17%	78.34%	70.39%	74.15%
KNN	80.76%	81.24%	84.26%	82.72%	80.85%	80.67%	85.45%	82.99%
RF	79.28%	79.93%	82.91%	81.40%	80.02%	80.39%	83.93%	82.12%
XGBoost	77.89%	79.04%	81.05%	80.03%	78.54%	79.00%	82.74%	80.83%
AdaBoost	69.94%	72.62%	72.25%	72.43%	67.16%	70.21%	69.37%	69.79%
GBDT	79.28%	79.74%	83.25%	81.46%	79.37%	80.36%	82.40%	81.37%
MLP	55.22%	56.16%	82.57%	66.85%	54.58%	55.63%	83.59%	66.80%
LSTM	58.83%	62.17%	63.11%	62.64%	66.88%	67.95%	74.62%	71.13%
GRU	70.77%	71.52%	77.33%	74.31%	73.64%	75.50%	76.65%	76.07%
Algorithm	IRL Multi-factor				Benchmark			
	Accuracy	Precision	Recall	F-1	Accuracy	Precision	Recall	F-1
BNB	67.25%	68.37%	74.62%	71.36%	65.49%	70.72%	62.94%	66.61%
LR	56.71%	59.00%	68.19%	63.27%	50.05%	54.73%	49.92%	52.21%
SVM	73.27%	74.67%	77.33%	75.98%	49.95%	54.63%	49.92%	52.17%
KNN	79.93%	80.36%	83.76%	82.02%	43.20%	47.79%	42.13%	44.78%
RF	78.26%	79.57%	81.05%	80.30%	68.54%	68.87%	77.50%	72.93%
XGBoost	79.19%	80.70%	81.38%	81.04%	48.84%	53.67%	47.04%	50.14%
AdaBoost	71.32%	73.07%	75.30%	74.17%	57.91%	61.81%	60.24%	61.01%
GBDT	80.02%	80.89%	83.08%	81.97%	59.38%	63.63%	60.07%	61.79%
MLP	54.67%	56.43%	74.96%	64.39%	50.05%	54.75%	49.75%	52.13%
LSTM	67.53%	67.91%	76.99%	72.16%	50.23%	55.38%	46.19%	50.37%
GRU	76.04%	76.52%	81.05%	78.72%	57.35%	61.73%	57.87%	59.74%

Table 4. Long-term prediction of the new set of hyperparameters (Set 2)

Algorithm	IRL All-factor				IRL Single-factor			
	Accuracy	Precision	Recall	F-1	Accuracy	Precision	Recall	F-1
BNB	67.62%	68.15%	67.40%	67.77%	64.85%	65.90%	63.00%	64.42%
LR	50.88%	51.11%	63.19%	56.51%	52.08%	52.12%	63.00%	57.05%
SVM	49.49%	50.00%	57.14%	53.33%	49.12%	49.67%	55.68%	52.50%
KNN	80.11%	79.82%	81.14%	80.47%	75.86%	73.87%	80.77%	77.17%
RF	52.08%	52.15%	62.09%	56.69%	64.66%	63.85%	69.23%	66.43%
XGBoost	84.09%	84.89%	83.33%	84.10%	85.56%	86.11%	85.16%	85.64%
AdaBoost	59.39%	58.19%	69.60%	63.39%	54.67%	53.62%	76.01%	62.88%
GBDT	84.09%	84.50%	83.88%	84.19%	85.38%	85.79%	85.16%	85.48%
MLP	53.38%	52.81%	72.16%	60.99%	56.06%	56.22%	58.79%	57.48%
LSTM	72.06%	74.40%	68.13%	71.13%	70.40%	72.69%	66.30%	69.35%
GRU	77.71%	79.73%	74.91%	77.24%	73.54%	72.97%	75.64%	74.28%
Algorithm	IRL Multi-factor				Benchmark			
	Accuracy	Precision	Recall	F-1	Accuracy	Precision	Recall	F-1
BNB	67.99%	68.59%	67.58%	68.08%	62.44%	63.31%	60.99%	62.13%
LR	51.99%	52.02%	63.74%	57.28%	50.42%	50.93%	50.00%	50.46%
SVM	50.51%	50.87%	58.97%	54.62%	52.36%	42.31%	42.60%	42.46%
KNN	79.74%	78.94%	81.68%	80.29%	55.97%	55.85%	61.17%	58.39%
RF	67.25%	66.16%	71.98%	68.95%	43.48%	43.04%	36.81%	39.68%
XGBoost	84.55%	85.69%	83.33%	84.49%	53.65%	54.14%	53.85%	53.99%
AdaBoost	60.22%	59.06%	69.23%	63.74%	54.30%	56.60%	40.84%	47.45%
GBDT	83.90%	85.09%	82.60%	83.83%	52.45%	41.75%	38.57%	40.09%
MLP	52.36%	52.57%	58.06%	55.18%	50.23%	50.74%	50.00%	50.37%
LSTM	71.32%	72.96%	68.68%	70.75%	64.75%	63.29%	71.98%	67.35%
GRU	78.35%	79.66%	76.74%	78.17%	59.94%	59.01%	67.77%	63.09%

Table 5. One-sided paired T-test of F-1

	Total vs. Benchmark		Single vs. Benchmark		Multi vs. Benchmark	
	Set 1	Set 2	Set 1	Set 2	Set 1	Set 2
Short-term	0.002	0.000	0.000	0.001	0.000	0.000
Mid-term	0.003	0.000	0.000	0.000	0.000	0.000
Long-term	0.002	0.001	0.000	0.001	0.000	0.001