Method	News-2 ϵ_{ATE}	News-4 ϵ_{mATE}	News-8 ϵ_{mATE}	News-16 ϵ_{mATE}
kNN	7.83 ± 2.55	19.40 ± 3.12	15.11 ± 2.34	17.27 ± 2.10
PSM	4.89 ± 2.39	30.19 ± 2.47	22.09 ± 1.98	18.81 ± 1.74
RF	5.50 ± 1.20	18.03 ± 3.18	12.40 ± 2.29	15.91 ± 2.00
CF	4.02 ± 1.33	13.54 ± 2.48	9.70 ± 1.91	8.37 ± 1.76
BART	5.40 ± 1.53	17.14 ± 3.51	14.80 ± 2.56	17.50 ± 2.49
GANITE	4.65 ± 2.12	13.84 ± 2.69	11.20 ± 2.84	13.20 ± 3.28
PD	4.69 ± 3.17	8.47 ± 4.51	7.29 ± 2.97	10.65 ± 2.22
TARNET	4.58 ± 1.29	13.63 ± 2.18	9.38 ± 1.92	8.30 ± 1.66
CFRNET	4.54 ± 1.48	12.96 ± 1.69	8.79 ± 1.68	8.05 ± 1.40
SITE	4.53 ± 1.32	12.75 ± 1.88	9.01 ± 1.86	8.63 ± 1.41
PM	$\mathbf{3.99 \pm 1.01}$	10.04 ± 2.71	6.51 ± 1.66	5.76 ± 1.33
TDCFD	4.25 ± 0.98	$\mathbf{8.77 \pm 2.49}$	$\mathbf{5.93 \pm 1.25}$	$\mathbf{5.04 \pm 1.19}$

Table 1: Performance on News data sets. We present the mean $\pm$ standard deviation for ϵ_{ATE} and ϵ_{mATE} on the test sets. We list the available results reported by the original authors (**?**).

Method	Synthetic data			Real corporate risk data		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
LR	0.92	0.64	0.54	0.83	0.21	0.16
SVM	0.94	0.68	0.65	0.87	0.40	0.27
KNN	0.91	0.55	0.60	0.91	0.62	0.47
RF	0.95	0.72	0.78	0.90	0.60	0.43
XGBoost	0.94	0.67	0.83	0.91	0.61	0.63
DNN	0.95	0.73	0.80	0.93	0.70	0.66
Transformer	0.96	0.77	0.85	0.93	0.71	0.71
TDCFD	**0.97**	**0.82**	**0.90**	**0.96**	**0.86**	**0.80**

Table 2: Performance on synthetic risk prediction task and real corporate risk prediction task.

Method	No hidden variable			Hidden variable		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
DNN	0.98	0.89	0.92	0.94	0.72 $\downarrow$	0.65 $\downarrow$
XGBoost	**0.99**	**0.91**	**0.96**	0.95	0.78 $\downarrow$	0.70 $\downarrow$
TDCFD	0.98	0.90	0.95	**0.97**	**0.82**	**0.89**

Table 3: Performance on synthetic risk prediction data with and without hidden variables.

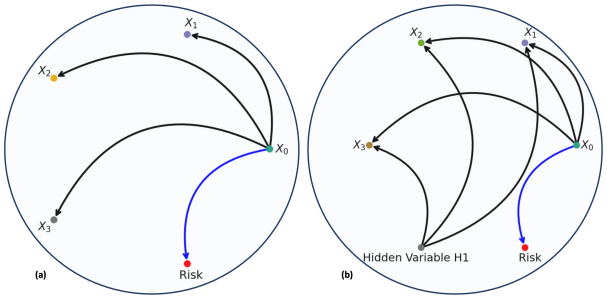

Figure 3: The DAGs of risk data generation with and without hidden variables.

$\epsilon_{\text{mATE}} = \frac{1}{\binom{k}{2}} \sum_{i=0}^{k-1} \sum_{j=0}^{i-1} \epsilon_{\text{ATE},i,j}$, where k is the number of intervention options.

Table 1 shows the performance of our method and baseline methods on the News datasets with $2, 4, 8,$ and 16 intervention options. TDCFD achieves the best performance with respect to ϵ_{ATE} on News datasets with $4, 8,$ and 16 intervention options. The results of these benchmarks for causal effects estimation can demonstrate that our method is capable of precisely estimating causal effects.

Experiments of Risk Prediction on Synthetic

Simulation Procedure. Because, in the real observational data, the true data generation procedure is unknown, we cannot effectively evaluate the explainability and the true feature contributions.

We generate a synthetic dataset that can not only reflect the complexity of real data but also help to explore the reason

why our model can outperform the general machine learning models for the risk prediction task. As shown in Fig. 2, our synthetic data includes 20 features and a binary risk label. In order to incorporate the underlying causal relationships among the features and between the features and the risk outcome, we randomly generate a directed acyclic graph (DAG) to represent the conditional dependency relationships and then utilize the Bayesian networks (**?**) to simulate the data. Bayesian networks are a type of probabilistic graphical model that uses Bayesian inference for probability computations. It aims to model conditional dependence, and therefore causation, by representing conditional dependence by edges in a directed graph. Data is simulated from a Bayes net by first sampling from each of the root nodes, then followed by the children conditional on their parents until data for all nodes have been drawn. To realistically simulate the risk data, we generate $1,000$ samples with the positive label and $9,000$ with the negative label.

Baseline Methods. We apply some classical classification models to this risk prediction task, such as Logistic Regression (LR), Support Vector Machine (SVM) (**?**), K-Nearest Neighbours (KNN) (**?**), Random Forest (RF) (**?**), DNN (**?**), Transformer (**?**), and XGBoost (**?**).

Evaluation Metrics. To evaluate the effectiveness of a model, we adopt Precision ($\frac{TP}{TP+FP}$), Recall ($\frac{TP}{TP+FN}$), and Accuracy ($\frac{TP+TN}{Total}$). Both precision and recall are defined in terms of the positive class. Precision measures the quality of model predictions for positive class and recall, on the other hand, measures how well the model did for the actual observations of the positive class. Compared to accuracy, precision and recall are more important in the risk prediction task.

Results. Table 2 shows that TDCFD achieves the best performance with respect to precision and recall in the synthetic data experiment. To further explore the reason why there exist large differences in precision and recall between our model and the baseline models, we performed ablation studies on two more datasets by predicting the risk outcome based on four observed variables (X_0, X_1, X_2, and X_3). The first one (Fig.3 (a)) contains 4 feature variables and a risk outcome variable, where only X_0 is the cause of the outcome also related to X_1, X_2, and X_3. In the second data (Fig.3 (b)), except for the four observed feature variables, there exists another hidden variable $H1$. X_1, X_2, and X_3 all depend on this hidden variable $H1$. According to Table 3, we can find that based on the spurious correlations in the first data, the positive samples can still be accurately captured, but in the second data, the precision and recall decrease dramatically due to ignorance of the real cause.

Experiments of Risk Prediction on Real Data

Real Data. To evaluate the model performance for risk prediction tasks, we adopt a real dataset collected from Alipay, the top Fintech company that offers billions of customers equal access to sustainable financial services and capital. This corporate risk data includes $16,409$ observations with $1,867$ positive samples and $14,542$ negative samples. It contains 114 feature variables, such as corporate financial statement data, public opinion data, corporate event data, and so on. The baseline methods and evaluation metrics are identical to those in synthetic data experiments.

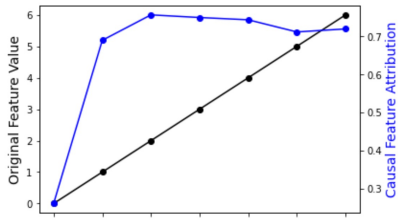

Figure 4: Original values and causal feature attributions.

Results. Table 2 shows the performance of our method and baseline methods on the real corporate risk prediction task. TDCFD achieves the best performance with respect to precision and recall. To figure out the reasons for the model's performance, we did two studies on the original feature data (original feature values) and causal feature distilled data, where original feature values are replaced by causal feature attributions. We exhibit a typical categorical feature with original values and causal feature attributions in Fig. 4. We can find the original values are uniformly increased from 0 to 6, but there is a huge gap between the first value and other 5 values in this feature's causal feature attribution range. The causal feature attributions for the original values from 1 to 6 are very close and have similar risk probabilities. However, the original data cannot reflect such information. In addition, we did the t-test for continuous variables and the chi-square test for categorical variables for both original feature data and causal feature distilled data. In the original feature data, 64 variables significantly differ between positive and negative classes. However, in the causal feature distilled data, there are only 52 variables that are significantly different. Furthermore, the 52 variables do not all come from the 64 variables of the original feature data. Therefore, the TDCFD filters out a part of spurious correlations and discovers some new causal relationships that did not appear in the original data.

Conclusion

We propose a novel Task-Driven Causal Feature Distillation model (TDCFD) for trustworthy risk predictions, which incorporates the POF to distill causal feature contributions and make predictions based on them. We conduct comprehensive experiments on both synthetic and real datasets to illustrate our model can perform well in risk prediction tasks with significantly improved precision and recall and generate causal-based interpretability.

Dolorum inventore temporibus perspiciatis quaerat nihil libero, voluptatibus natus enim, cupiditate voluptatibus numquam soluta dicta tempore aperiam, vero ducimus voluptate dolorem doloribus alias sequi tempora beatae blanditiis atque id, maiores culpa dolore ea consequatur aperiam eaque minima at delectus labore.Magni hic soluta eum repellat voluptates ipsam ipsa sit, pariatur suscipit nobis voluptas et tempore perferendis possimus, vitae tempora velit beatae delectus ullam modi consectetur, dolor voluptates deleniti iusto dolores ducimus, aliquam excepturi nostrum veritatis.Distinctio explicabo obcaecati deserunt impedit error nesciunt dolorem soluta, quasi eum maxime, tenetur