THESIS DEFENSE FOR ENGINEERING DEGREE

Distillation Learning

RISQ/MRM

Omar Elghaffouli **Imen Fourati**

Zineb Baroudi & Nada Amini

Société Générale : 17, Cours Valmy, 92000, Puteaux, France

C'EST VOUS L'AVENIR SOCIETE GENERALE

Centrale CASABLANCA

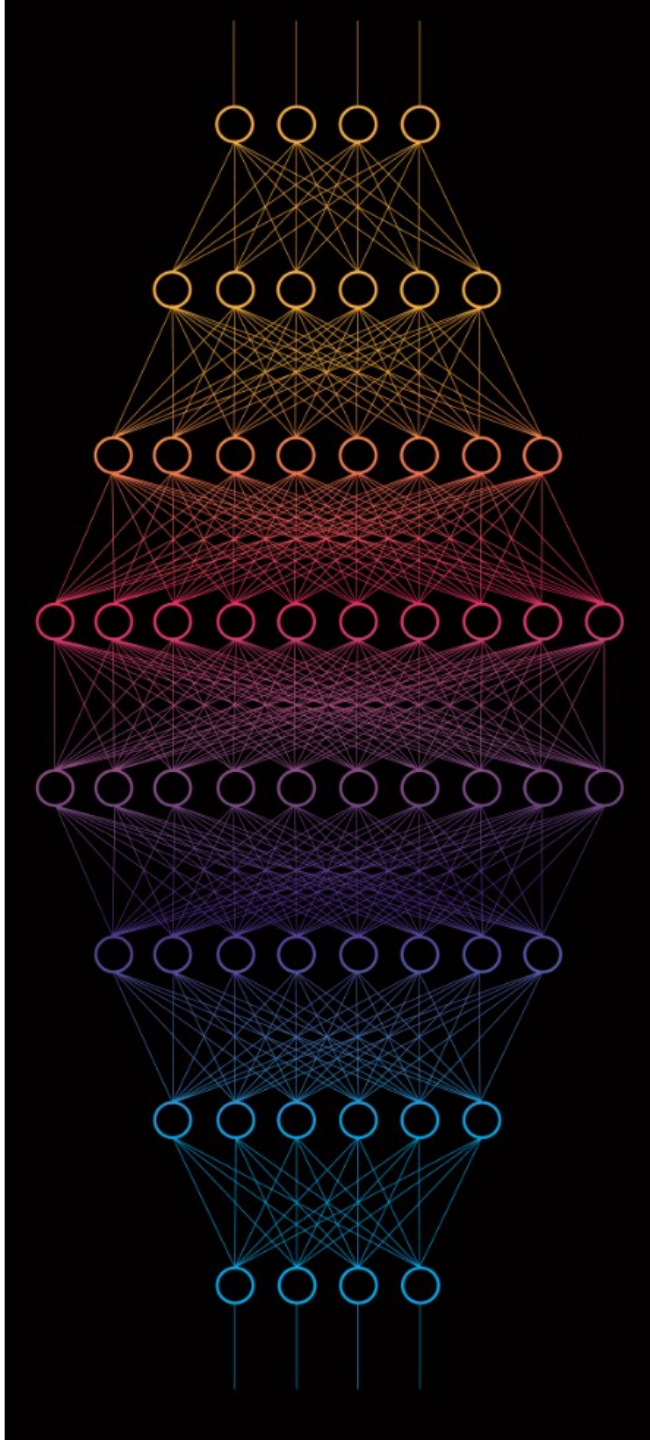
OUTLINES

1. INTRODUCTION

A. Context and Motivation
B. Knowledge Distillation Basic Definition
C. Missions

SOCIETE GENERALE

Centrale CASABLANCA



A. CONTEXT AND MOTIVATION (1/2)

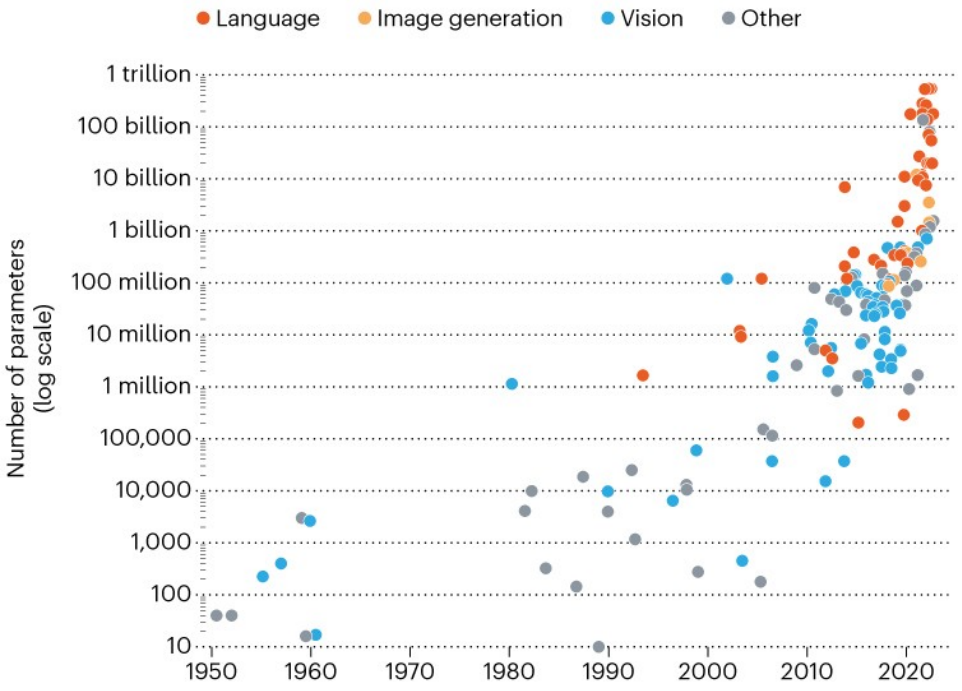


Figure 1. The drive to bigger AI models. The scale of AI neural networks models is growing exponentially, as measured by the models' trainable parameters, source: Nature

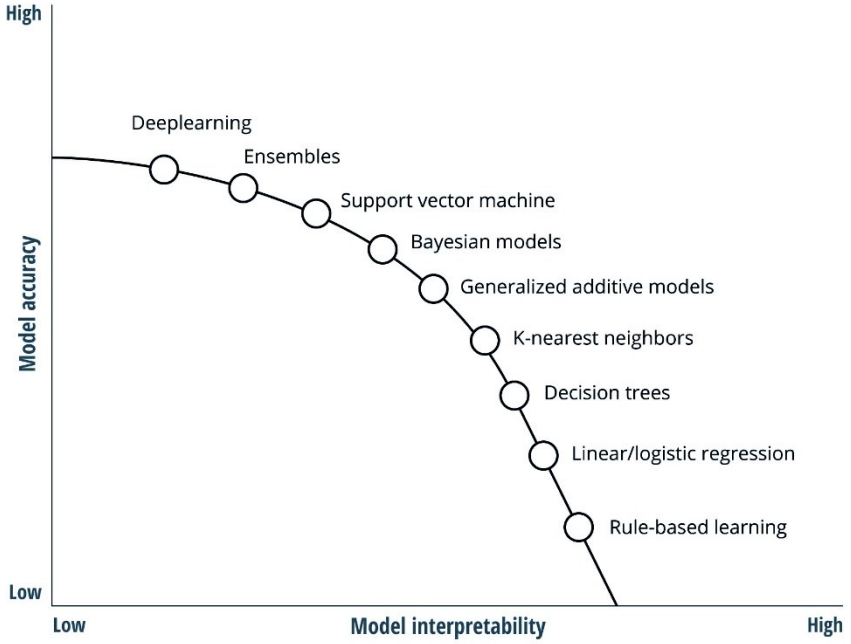


Figure 2. Inherent trade-offs between model accuracy, complexity and model interpretability. Source: Deloitte Insights

The final report of European Banking Authority (2020) urges banking institutions to **"understand the models used, and their methodology, input data, assumptions, limitations and outputs".**

A. CONTEXT AND MOTIVATION (2/2)

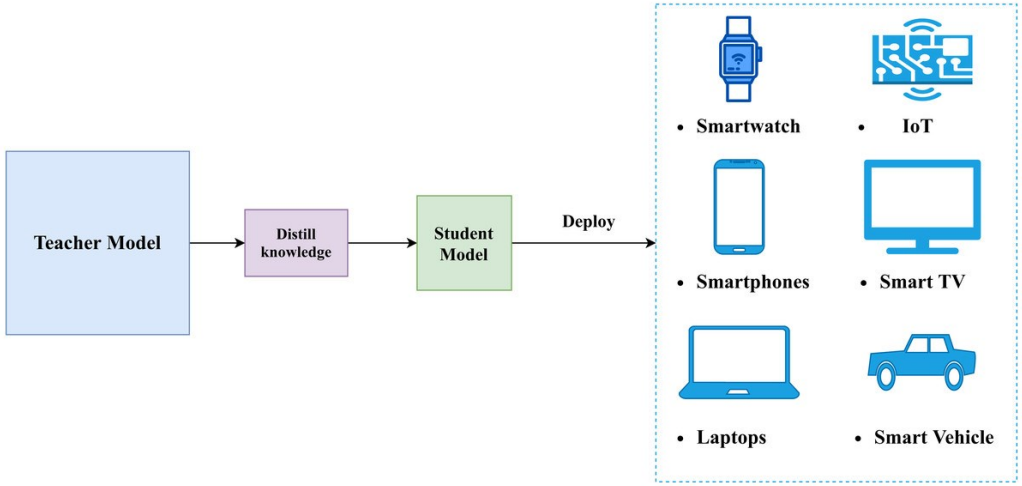


Figure 3. Knowledge Distillation for Edge Computing

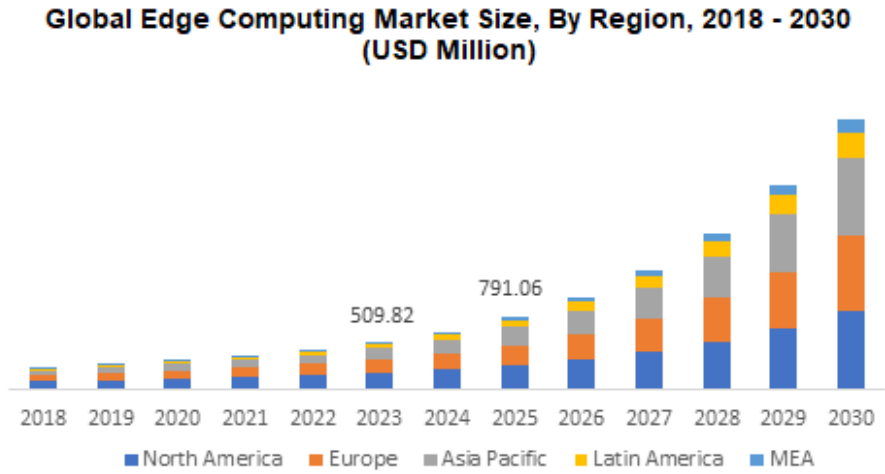


Figure 4. Edge Computing Market is gaining in size over years. Development of small and less cumbersome models is then mandatory to fit in edge device computing capacity. Source

B. KNOWLEDGE DISTILLATION FUNDAMENTAL DEFINITION

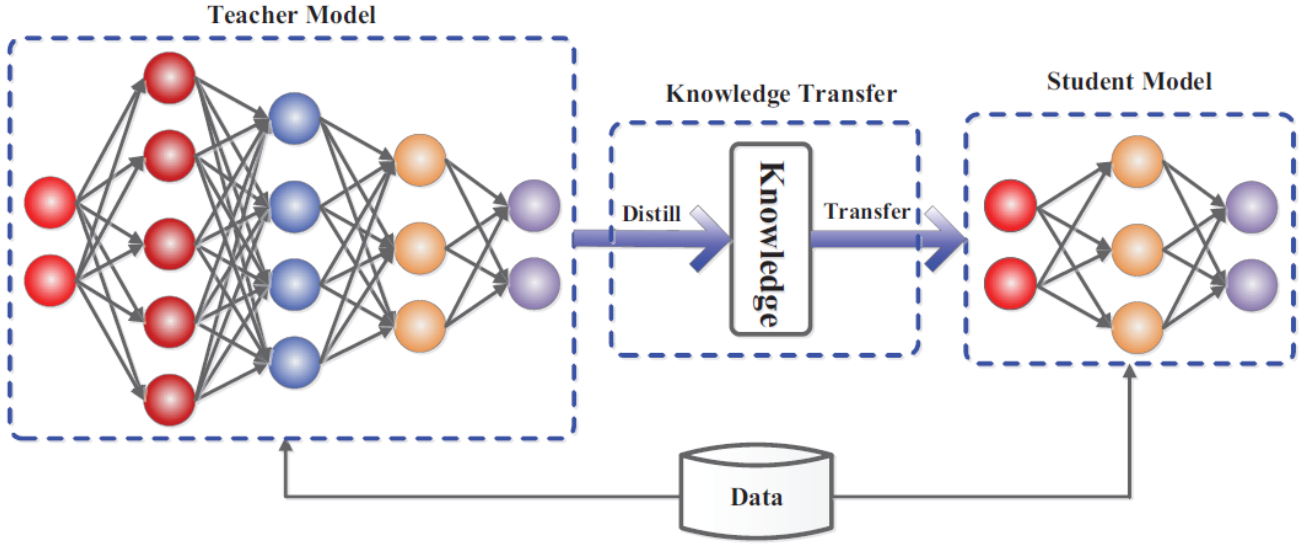


Figure 5. The generic teacher-student framework for knowledge distillation.
Source: Arxiv

2 Main Questions in the Context of RISK/MRM at Société Générale

- ❑ What are the main uses of knowledge distillation in the context of model risk management ?
- ❑ How can distillation learning be used in improving models' interpretability and performance ?

C. MISSIONS

			April				May				June				July				August				September			
Missions	**START**	**END**	W1	W2	W3	W4	W1	W2	W2	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4
Literature Review	4/3/23	5/24/23																								
Workshop 1	5/20/23	25/5/23																								
Test on PD Estimation models	6/1/23	24/8/23																								
Test on Lending Club Dataset	8/1/23	9/15/23																								
Workshop 2	9/6/23	9/20/23																								

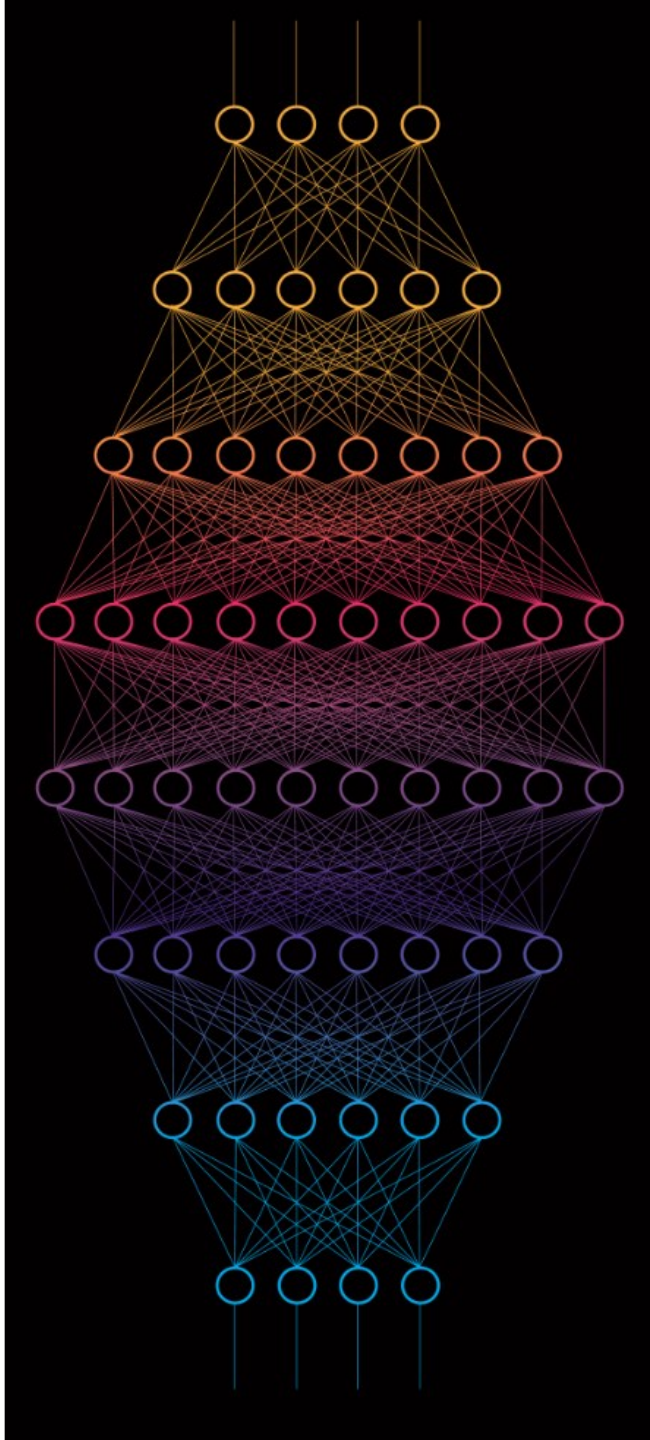
Figure 6. Gant chart of internship's missions

2. LITERATURE REVIEW OVERVIEW

A. Methodology
B. Distillation learning fundamentals
C. Relevant frameworks
D. Potential Uses in RISK/MRM

SOCIETE GENERALE

Centrale CASABLANCA



A. METHODOLOGY

Figure 6. Literature Review Milestones, 29 Scientific paper on knowledge distillation were reviewed.

B. DISTILLATION LEARNING FUNDAMENTALS (1/3)

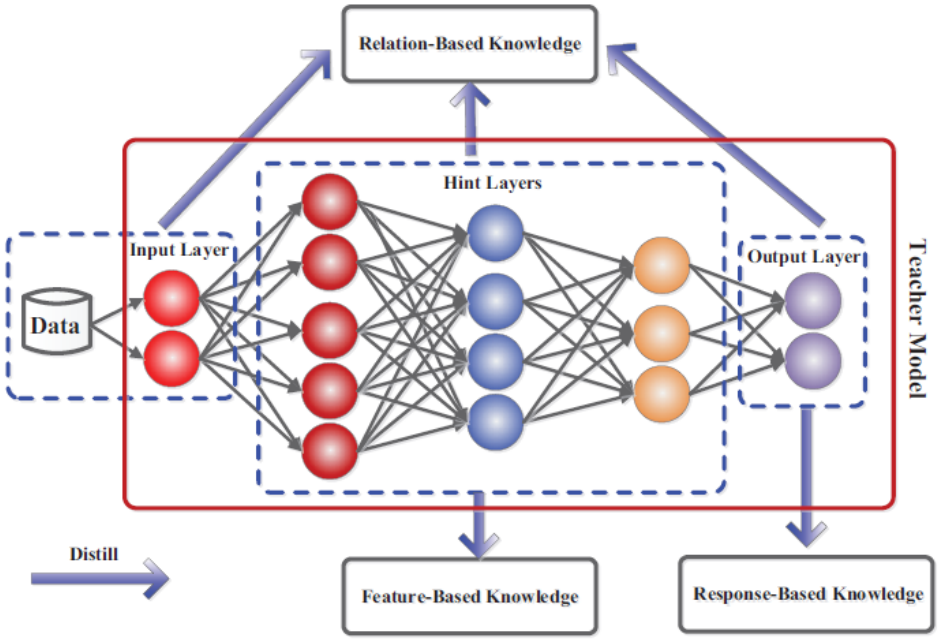


Figure 7. The schematic illustrations of sources of knowledge in a deep teacher network

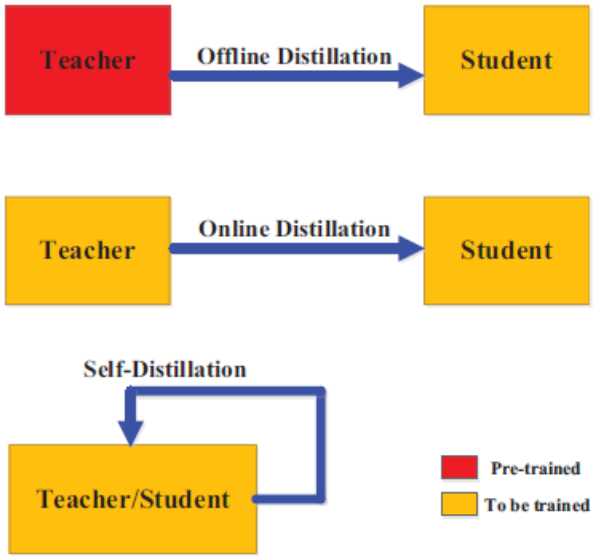


Figure 8. Different Distillation Training Modes

A. DISTILLATION LEARNING FUNDAMENTALS (2/3)

RESPONSE-BASED KNOWLEDGE

The classical framework for knowledge distillation. The student tries to mimic as good as possible the output predictions of the teacher model in a response-based manner. Practically, we use logits (Neurons outputs before SoftMax) because they contain dark knowledge which is the deep knowledge learnt by the teacher.

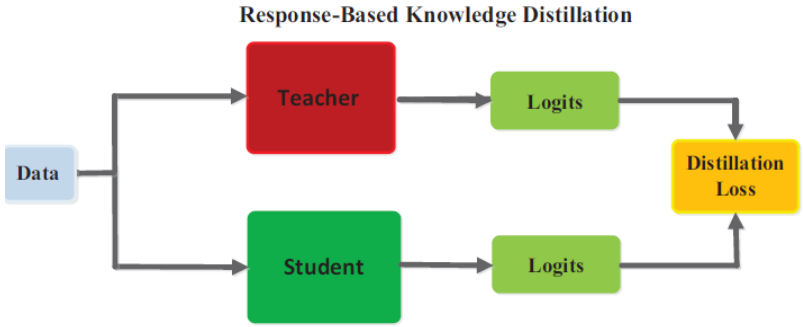


Figure 9. The specific architecture of the benchmark knowledge distillation. The student model can learn to mimic teacher's predictions and ground truth labels.

Pros	Knowledge	Limits
Easy-to-use, straight-forward Fast, efficient	Predictions of the teacher model Dark knowledge embedded in soft targets or in logits (Hinton and al, 2015, Caruana and al, 2014).	Limited to supervised learning Relies on the final output fails to address intermediate-level supervision

Table 1. Response-based distillation investigation

$$\mathcal{L}_{KD} = \sum_{(x_t, y_t) \in (X_t, Y_t)} [\alpha \mathcal{L}_{CE}(f_S, x_t, y_t) + \beta \mathcal{L}_{KL}(f_S, f_T, x_t)]$$

Formula 1. Hinton Loss for Response-Based KD, Source, Hinton and al, 2015

$$p(z_i, T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

Formula 2. Hinton Soft-Targets for Response-Based KD, Source, Hinton and al, 2015; very high T values correspond approximately to matching logits.

A. DISTILLATION LEARNING FUNDAMENTALS (3/3)

OTHER TYPES OF KNOWLEDGE DISTILLATION

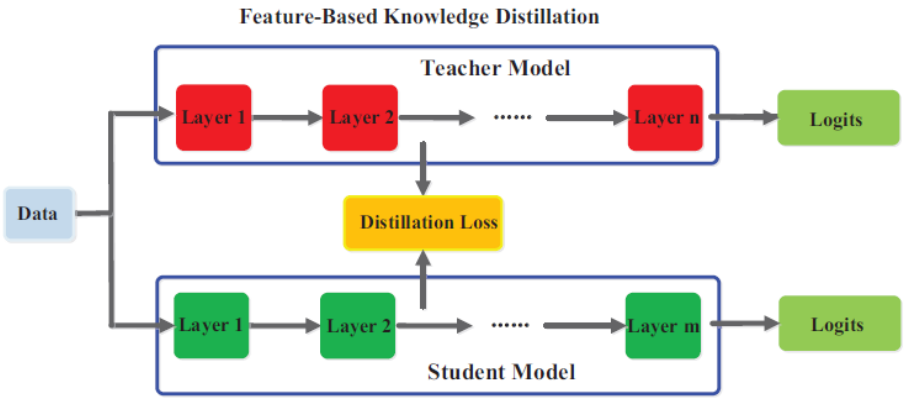


Figure 10. The generic feature-based knowledge distillation.

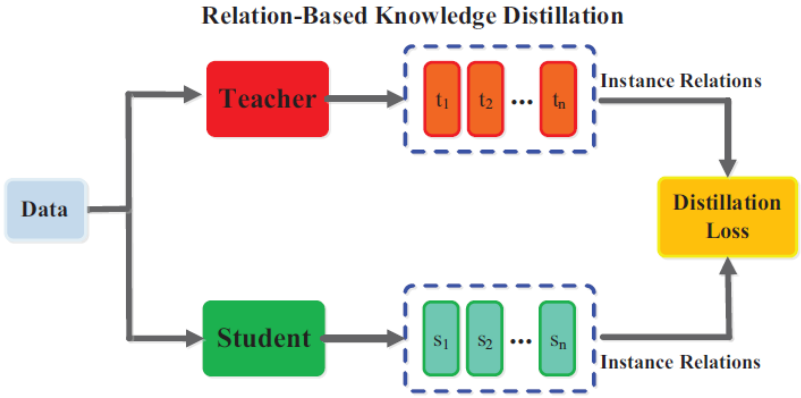


Figure 11. The generic relation-based knowledge distillation.

Pros	Knowledge	Limits
Learn multiple levels of feature representation.	1) Feature representation, hint layers (Romero et al., 2015) 2)Parameter distribution, multi-layer group (Liu et al., 2019c) 3)Feature Maps, hint layers (Chen et al., 2021)	Effectively choose the hint layers from the teacher model and the guided layers from the student model with optimum training complexity is questionable.

Table 2. Feature-based distillation investigation

Pros	Knowledge	Limits
Explores the relationship between layers or data samples.	1) FSP matrix, End of multi-layer group (Yim et al., 2017) 2) Logits graph, hint layers (Zhang and Peng, 2018) 3) Similarity Matrix, hint layers (Tung and Mori, 2019)	Relation modeling difficulties

Table 3. Relation-based distillation investigation

B. RELEVANT FRAMEWORKS (1/3)

ADVERSARIAL KNOWLEDGE DISTILLATION

An effective framework to enhance the power of student learning via the teacher knowledge distillation using GAN. This framework tackles two main problems;1) Difficulty for the teacher to learn the true data distribution (lack of data, unrepresentative data, small model, etc.); 2) Small capacity of the student and difficulties to mimic accurately the teacher (Capacity gap, Unreliable teachers)

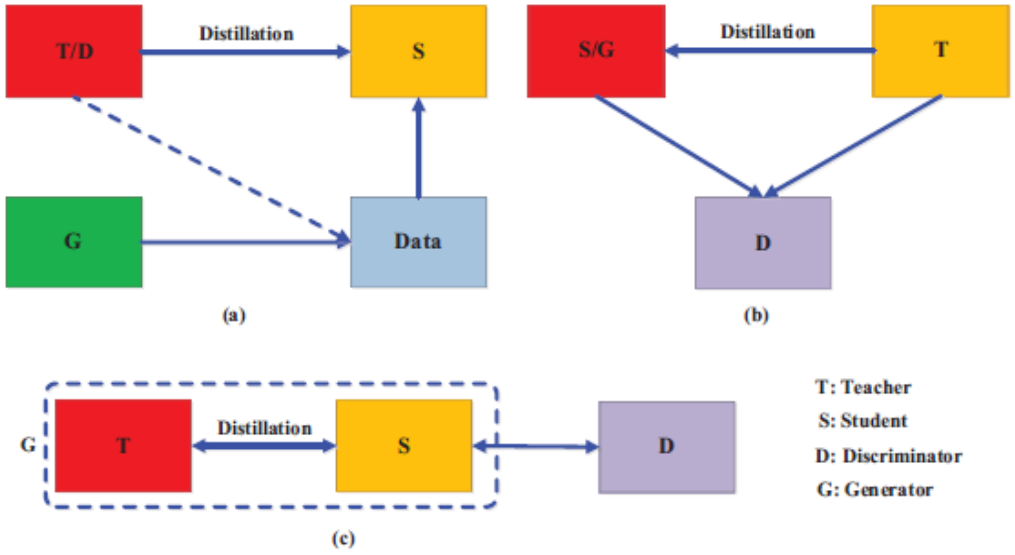


Figure 12. The different categories of the main adversarial distillation methods.

Scheme	Explanation
(a)	A generator is trained on true data distribution. Generated Data go then through teacher discrimination based on its proper data distribution. Student learns then teacher knowledge from 2 sources; 1) classical distillation process, 2) through generated data embedding teacher's internal feature representation.
(b)	A discriminator is trained on teacher's feature distribution. In addition to traditional distillation process, the student will generate new data based on its internal feature distribution corrected each time by the discriminator. The generated data is not used for training.
(c)	A discriminator is trained on true data distribution and corrects feature distribution of generators which are the student and teacher in an online setting.

Table 4. Adversarial Knowledge Distillation Framework's detailed explanation.

B. RELEVANT FRAMEWORKS (2/3)

INTERPRETABILITY DISTILLATION

Teacher explanation are important features driving a specific prediction. However, traditional distillation doesn't distill explanation and thus, student predictions are not driven by the same features due to explanation inconsistency between the teacher and the student.

Alharbi and al., 2021 have proposed a novel framework to distill explanation in addition to dark knowledge called XDistillation (XD). The framework has outperformed all traditional distillation methods.

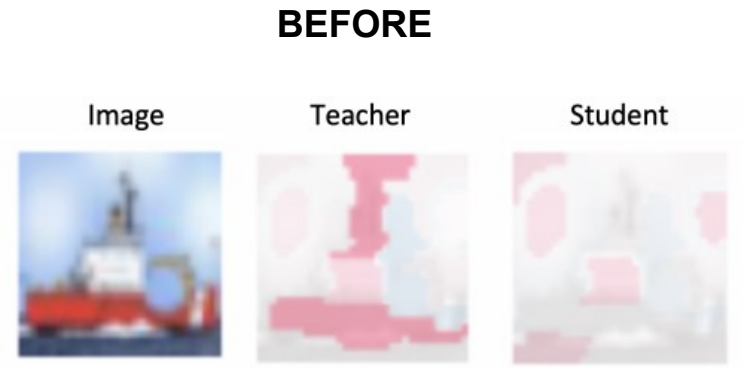


Figure 12. Inconsistency between teacher and student explanation

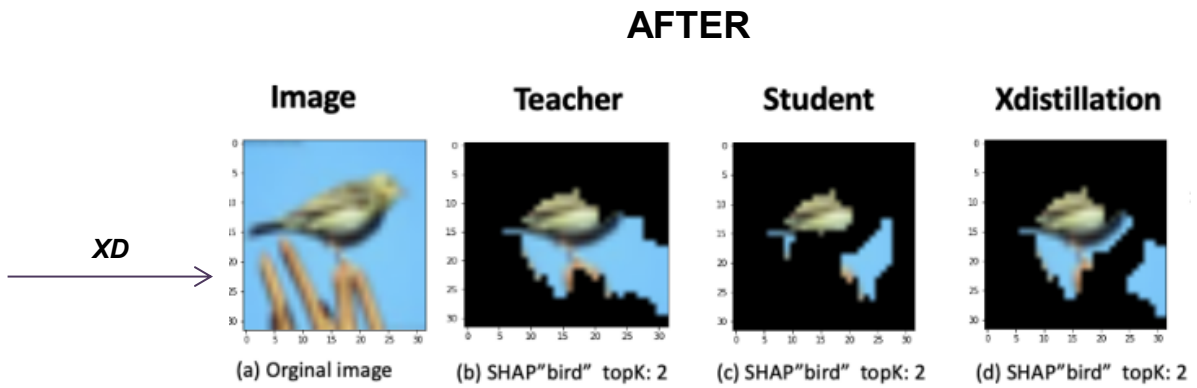


Figure 13. The overlapping explanation area of teacher, KD and XD.

Model	Accuracy %	#Parameters
Teacher	93.78	14,728,266
Baseline Student	89.2	2,781,386
Response-Based Distillation	90.2	2,781,386
X-Distillation	**90.9**	**3,521,276**

Table 5. X-Distillation performance Comparing with response-based distillation

MAIN USE OF KNOWLEDGE DISTILLATION IN RISK/MRM

3 Main Potential Uses of Knowledge Distillation in RISK/MRM

1 XAI

- If we have an inexplicable teacher such as a deep neural network or a random forest, we can use distillation of the teacher to train an interpretable and transparent model such as a decision tree along with being close to the teacher performance.
- In this case, the trade off performance/interpretability must be balanced depending on the situation.
- Usually, we use the teacher for inference alongside with student's interpretability insights.

2 Enhancing Performance

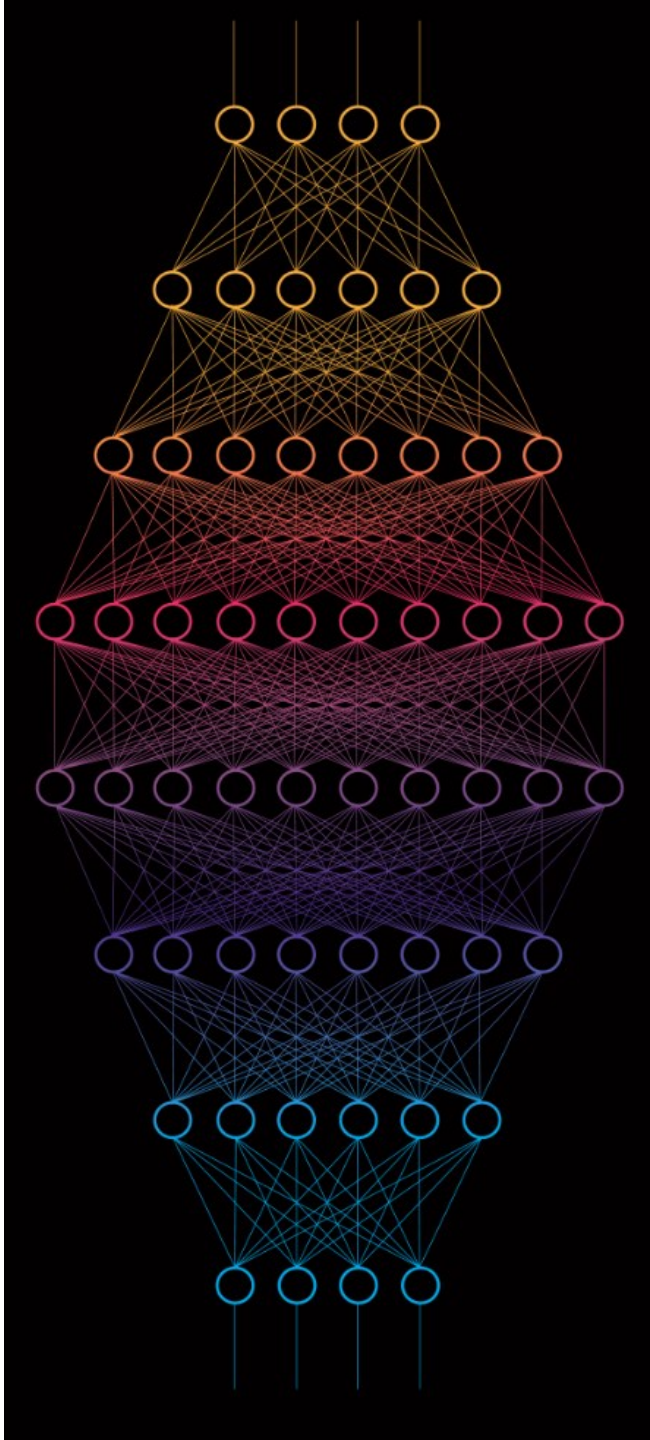
- A simple model is such as logistic regression, random forest, decision tree, linear regression or a simple neural network.
- Training a simple model through distillation of a more complex model usually outperforms training directly the same simple model.
- In MRM context , performance of PD Estimation Models can be enhanced by training a complex model such as deep neural network and then distilling it into a student model.
- We can also enhance performance of more sophisticated models by using self-distillation and transfer learning frameworks

3 Reducing Models Complexity

- In the context of compression, distillation can be used to reduce models' complexity.
- However, the student cannot outperform the teacher in general due capacity gap.
- In result, several frameworks were developed by scholars to reduce the performance gap between the teacher and the student.
- Some of the most relevant frameworks are adversarial knowledge distillation, interpretability distillation and transfer learning.

2. DISTILLATION OF PD ESTIMATION MODELS

A. Teacher Training
B. Response-Based Distillation
C. Adversarial Knowledge Distillation Framework
D. X-Distillation Framework



SOCIETE GENERALE

Centrale CASABLANCA

INTRODUCTION TO PD ESTIMATION MODELS FRAMEWORK

PD ESTIMATION MODELS ARCHITECTURE FOR FRANCE'S SME PORTFOLIO

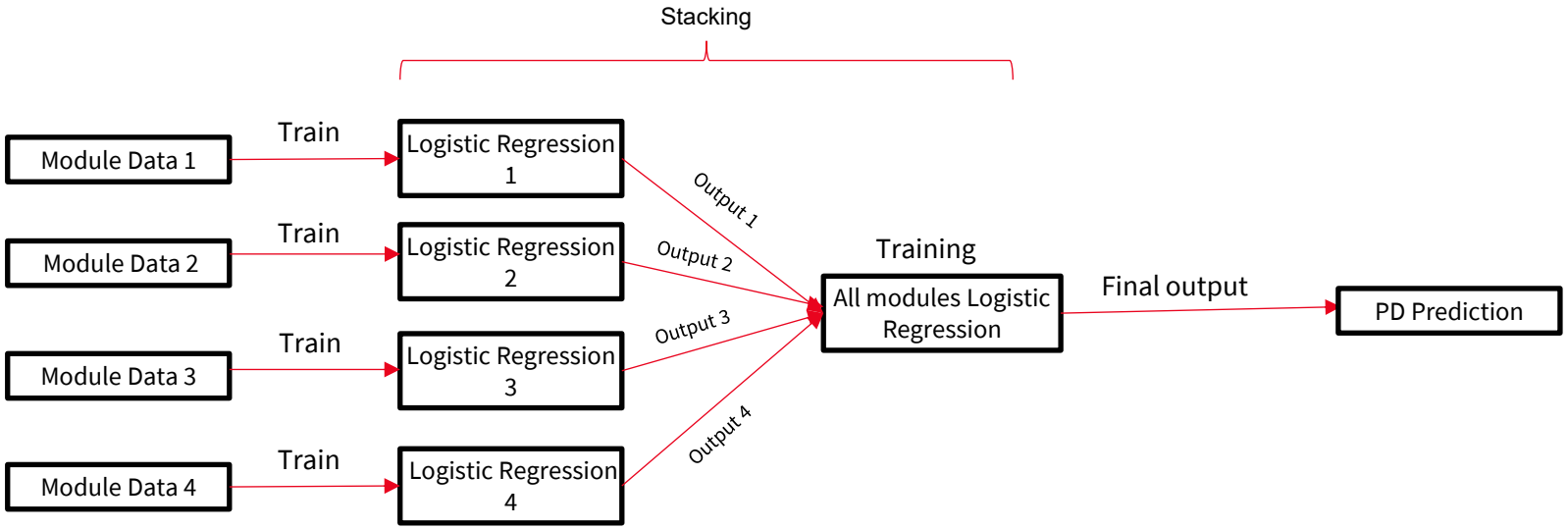


Figure 15. PD Estimation Models Training Workflow

Modules	Training AR*	WT Test AR	OOT Test AR
Module 1	55,3%	53,2%	60,2%
Module 2	53,91%	52,98%	57,43%
Module 3	64,7%	64,4%	66,76%
Module 4	28,86%	28,88%	32,78%
All Modules	65,4%	66,2%	66,4%

Table 7. Performance of PD Estimation Models

2 TYPES OF TEST DATA ARE USED:

- Within Time TEST SET : Each obligor (borrower) has **data records in different date times (panel data)**
- OUT-OF-TIME SET (OOT): each obligor has **one data record at a date fixed in 2018**.

Our goal is to enhance the performance of the overall model on the test and out-of-time sets.

*Accuracy Ratio = 2 * ROC AUC – 1

PD ESTIMATION MODELS' DISTILLATION FRAMEWORK

The objective of the distillation framework is to **enhance the overall PD estimation models performance**.

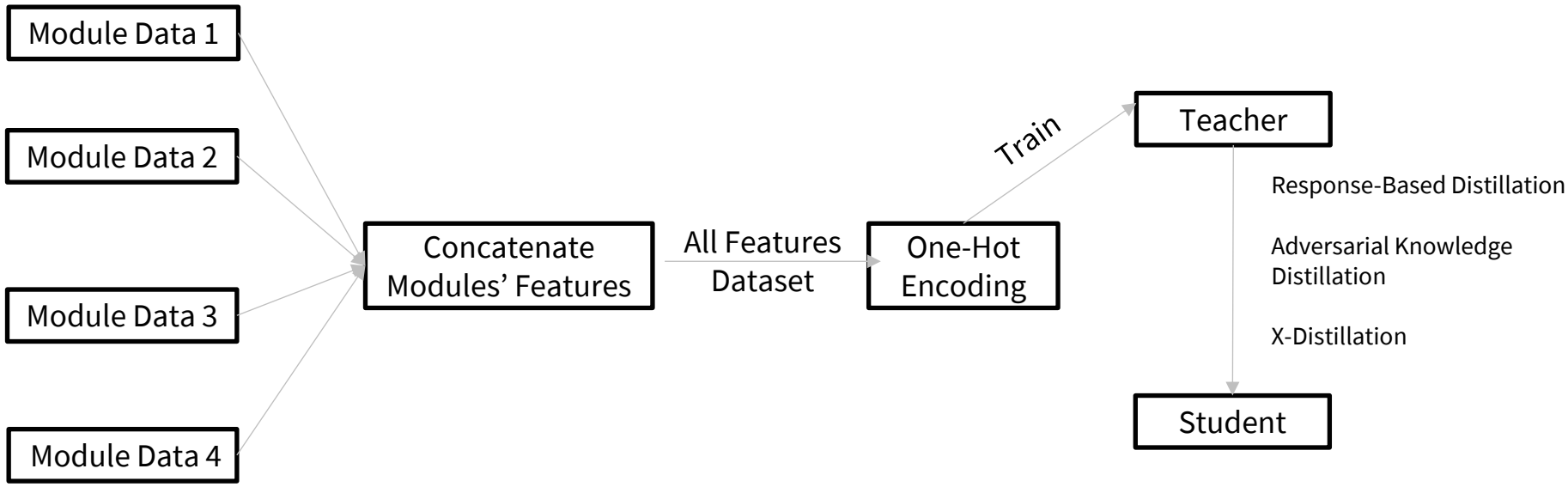


Figure 16. PD Estimation Models Distillation Framework

TEACHER TRAINING

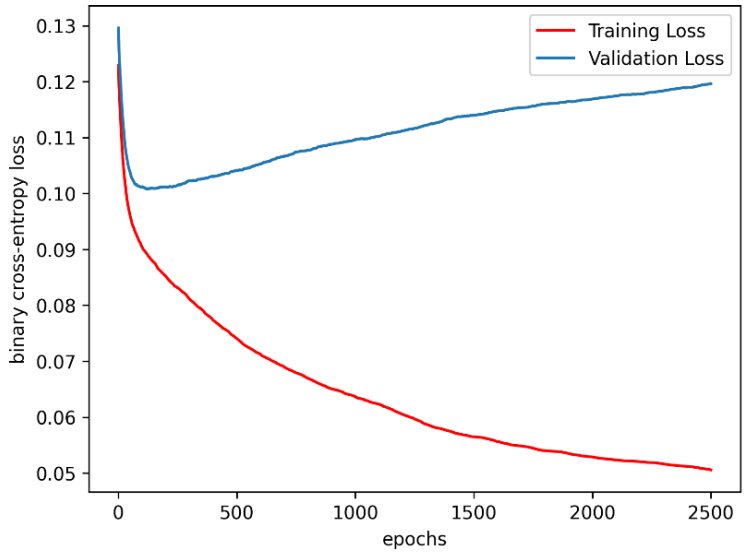


Figure 17. LightGBM teacher learning curves without regularization

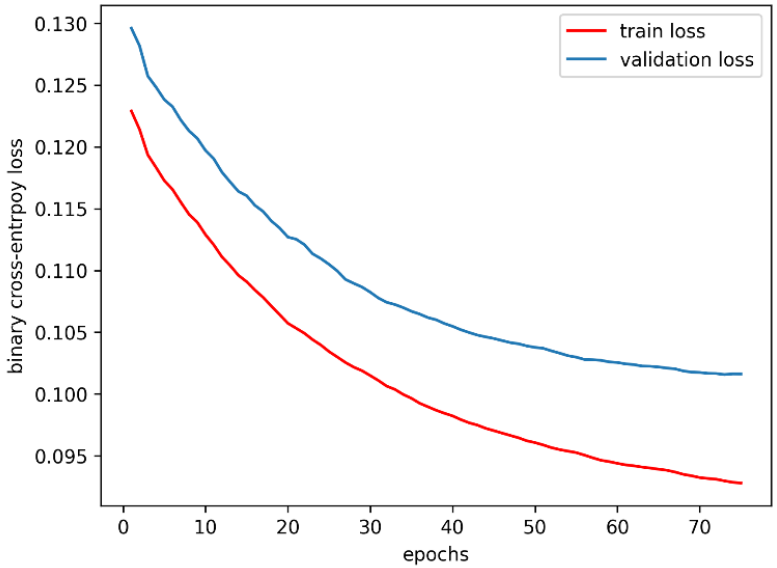


Figure 18. LightGBM teacher learning curves with regularization (L1 = 0.01 and Early Stopping = 7)

Models	Role	Training AR (%)	WT Test AR (%)	OOT Test AR (%)
PD estimation models	Baseline	65,40	66,20	66,4
Feed-Forward Neural Networks (FFNN)	Teacher	72.28	63.25	71.28
LightGBM with regularization	Teacher	70.87	67.41	71.44
LightGBM without regularization	Teacher	89.16	58.55	58.84

Selected Teacher

Table 8. Performance of trained teachers using the framework in Figure 16

RESPONSE-BASED DISTILLATION (1/3)

Response-Based Online Distillation

For comparison matters with PD Estimation Models, we train a logistic regression using response-Based distillation. Practically, we train a linear regression on teacher soft prediction. However, to constraint values to lie between 0 and 1, we apply the following transformation on teacher soft predictions before training:

$$y_{transformed,i} = log\left(\frac{pi}{1-pi}\right) \qquad pi = \frac{exp(z_i)}{\sum_j ex\,p(z_j)}$$

Where pi is teacher's soft prediction for the class i, z_i are teacher's logits and $y_{transformed,i}$ is the new target ranging on the negative real numbers line.

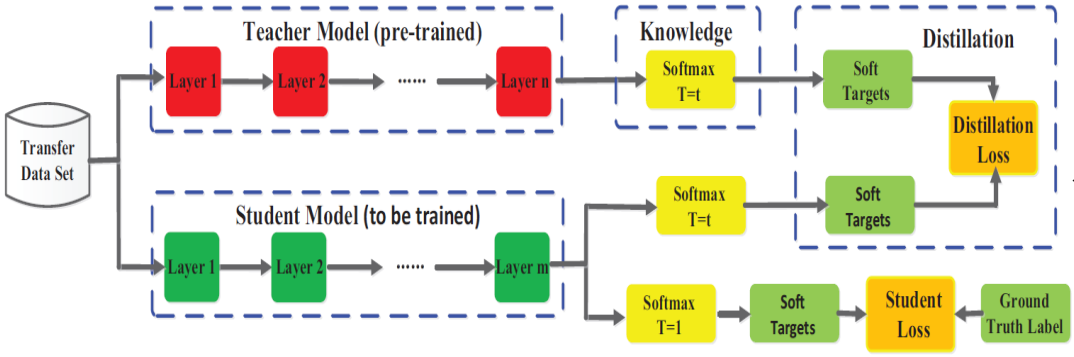


Figure 19. Response-Based Distillation Framework

$$\mathcal{L}_{kD} = \sum_{(x_t,y_t)}[\alpha\mathcal{L}_{CE}(f_s,x_t,y_t) + (1-\alpha)\,\mathcal{L}_{kL}(f_s,f_T,x_t)]$$

Formula 1. Response-Based Knowledge Distillation Loss, $\boldsymbol{\alpha}$ is the contribution of the loss comparing to ground truth labels.

RESPONSE-BASED DISTILLATION (2/3)

Response-Based Online Distillation

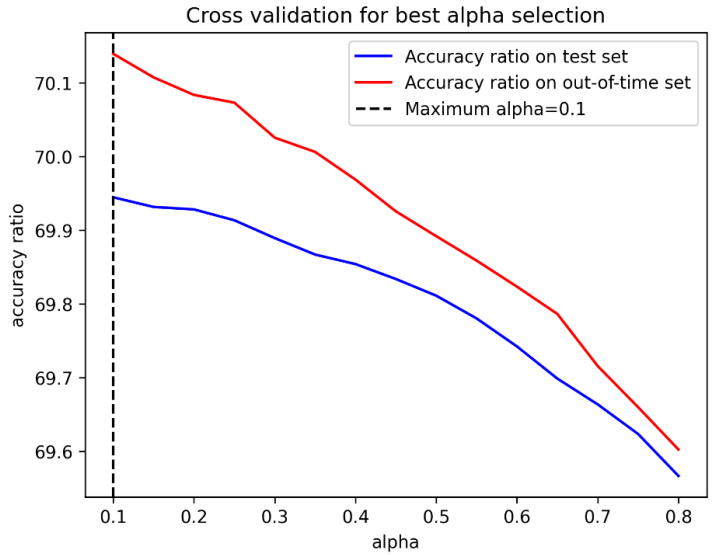


Figure 20. Cross-validation for best α selection

Models	Role	Training AR (%)	WT Test AR (%)	OOT Test AR (%)
PD estimation models	Baseline	65,40	66,20	66,40
LightGBM with regularization	Teacher	70.87	67.41	71.44
PD estimation models Distilled without Temperature	Student	68.03	63.94	70.13

Remarkable Improvement on out-of-time set

Table 9. Response-Based Students' Performance

RESPONSE-BASED DISTILLATION (3/3)

Response-Based Distillation with Temperature T

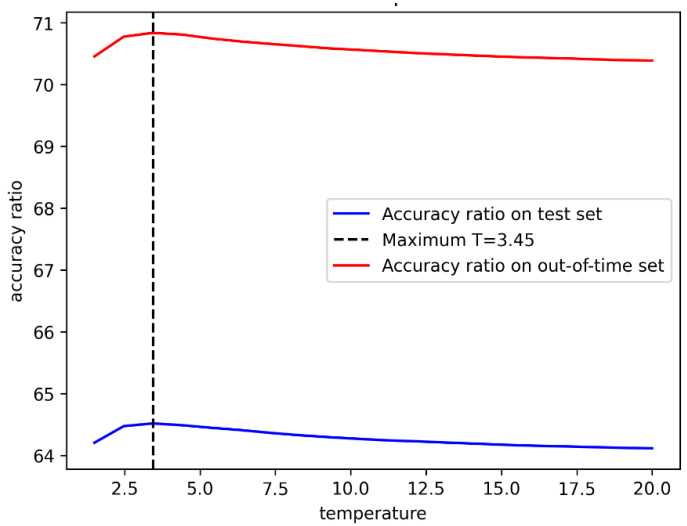


Figure 21. Cross-validation for best T selection, T = 3.45 corresponds to the best accuracy ratio on the test set.

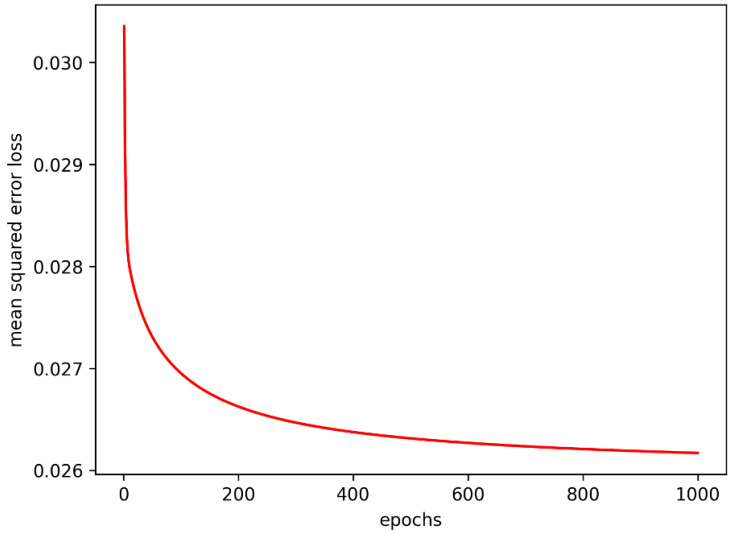


Figure 22. learning curve's training convergence with T = 3.45

Models	Role	Training AR (%)	WT Test AR (%)	OOT Test AR (%)
PD estimation models	Baseline	65,40	66,20	66,40
LightGBM with regularization	Teacher	70.87	67.41	71.44
PD estimation models Distilled without Temperature	Student	68.03	63.94	70.13
PD estimation models Distilled with Temperature	Student	68.63	64.51	70.83

Table 5. LightGBM Students' Performance. Softening teacher's outputs enhance student performance.

$$\mathrm{pi} = \frac{exp\left(\frac{z_i}{T}\right)}{\sum_j ex\,p\left(\frac{z_j}{T}\right)}$$

Formula 2. Teacher's softened predictions with temperature T

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (1/8)

Exploring Generative Knowledge Distillation Technique to Improve the Student AR TEST

This method will help the student to generalize well. In our case, student's parameters are updated only when the generated instance is labeled as fake which help update parameters only when the model is mistaken and adjust its data distribution.

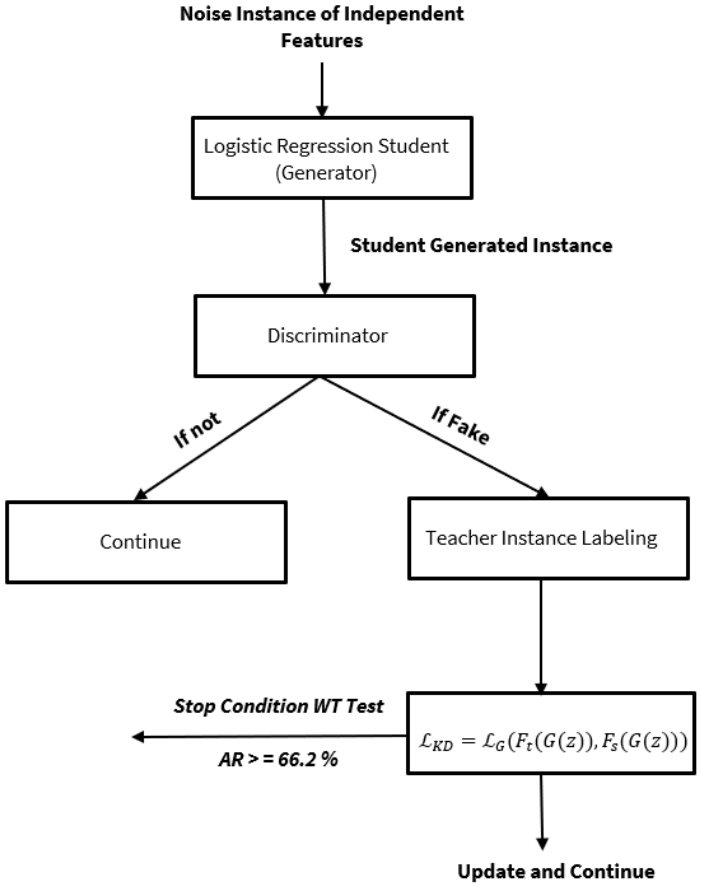


Figure 23. Adversarial Knowledge Distillation Training Framework Performed on PD Estimation Models. In this case, the generator is also the student which is a logistic regression. The teacher is the LightGBM used in the response-based framework.

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (2/8)

Discriminator Training Framework

The discriminator is trained to **predict either a generated instance from the generator (the student in our case) is real or fake**. In other words, it's used to predict if a generated instance follows the real data distribution or not.

To train it, we generate fake instances including the dependent variable then, we label it as fake (1). In the other hand, we label real training instances as real (0). Then we concatenate and shuffle fake and real instances and finally train the discriminator model.

Fake data should be a random noise, containing examples that are both far away from real data distribution and near to real data distribution to challenge the discriminator model training.

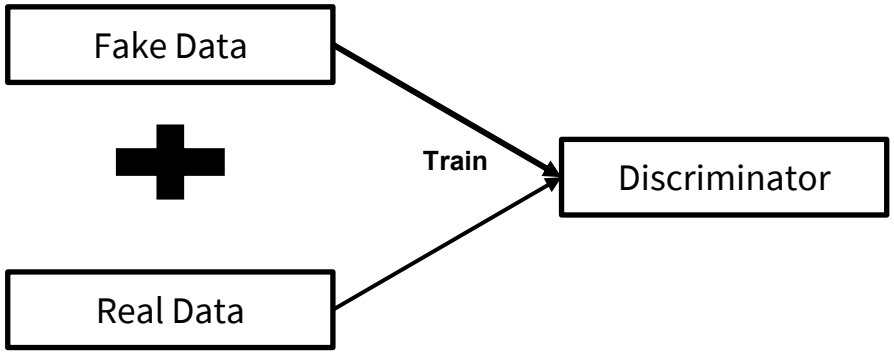


Figure 24. Discriminator Training Flow

How to Construct Fake Data ?

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (3/8)

Discriminator Training Framework - Fake Data Construction

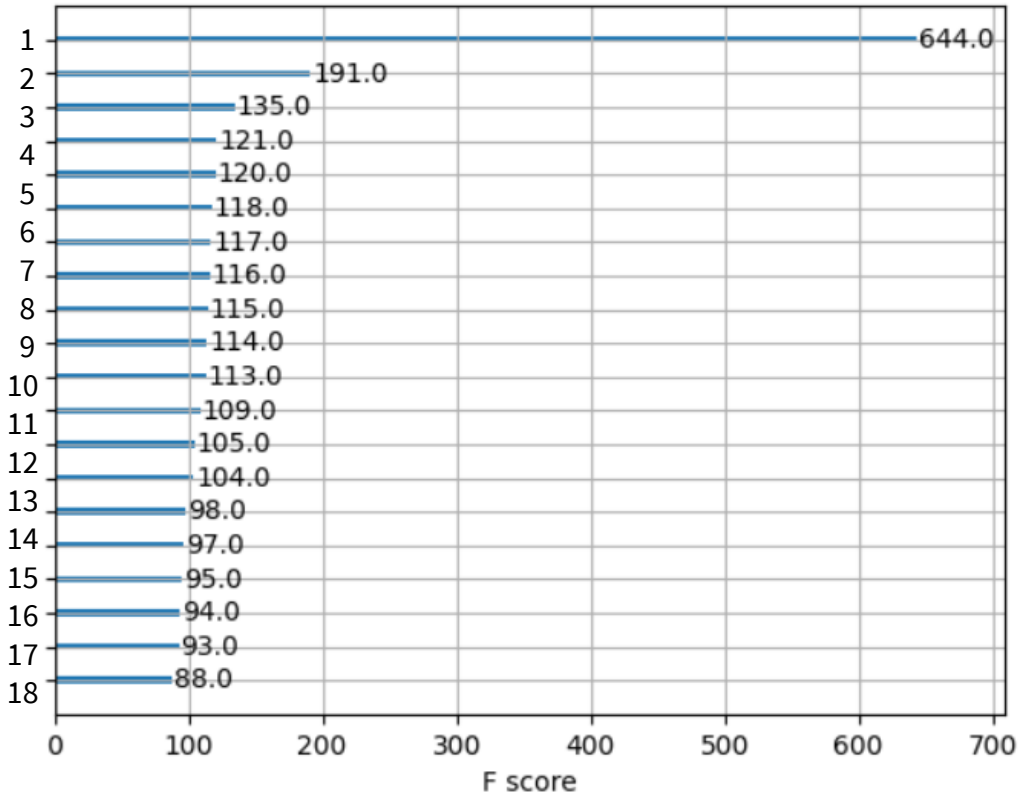


Figure 25. Preliminary discriminator (LightGBM) training. The continuous variable has almost 6 times more importance comparing to other features.

The discriminator relies the most on the continuous feature to make its prediction. How to construct fake examples of this variable ?

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (4/8)

Constructing Fake Instances of the Continuous Feature – Continuous Variable Distribution Approximation

Constructing an estimator of the probability distribution of the continuous variable with non-parametric estimation using kernel density

The idea is to sample the noise from a close estimation of the probability distribution of the continuous variable to avoid triviality and challenge the discriminator since the continuous variable is the most important feature of the model. The model relies mainly on it to make its prediction.

Kernel density, 1962, Parzen-Rosenblatt :

$$\widehat{p_n^h}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

where :

$\widehat{p_n^h}(x)$: is the estimated probability density at point x with bandwidth h.

K : is the kernel function, which depends on the choice of kernel (e.g., Gaussian, Epanechnikov).

X_i is the random variable which we want to estimate its distribution, in this case, the continuous variable and $x_1, x_2, \ldots\ldots, x_i, \ldots\ldots x_n$ the realisation of the random variable X_i.

In our case, we will choose the normal gaussian kernel defined as:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (5/8)

Constructing Fake Instances of the Continuous Feature - Cross Validation to determine the optimal bandwidth h

$$\mathrm{MISE(h)} = \mathrm{E}\left[\int \left(\mathrm{p(x)} - \widehat{p_n^h}(x)\right)^2 \mathrm{dx}\right] = \mathrm{E}\left[\int \left(\widehat{p_n^h}(x)\right)^2 - 2\,\widehat{p_n^h}(x)\mathrm{p(x)} + \left(\mathrm{p(x)}\right)^2 \mathrm{dx}\right]$$

$$= \mathrm{E}\left[\int \left(\widehat{p_n^h}(x)\right)^2 \mathrm{dx}\right] - 2 \times \mathrm{E}\left[\int \widehat{p_n^h}(x)\mathrm{p(x)}\,\mathrm{dx}\right] + \mathrm{E}\left[\int \left(\mathrm{p(x)}\right)^2 \mathrm{dx}\right]$$

Don't depend on h

The goal is constructing an unbiased estimator of MISE then we minimize it according to h :

$$CV(h) = \int \left(\widehat{p_n^h}(x)\right)^2 dx - 2\hat{G} \;\; ; h_{CV} \in arg\min_{h>0} CV(h) \;\; ; \hat{G} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{k \neq i}^{n} \frac{1}{h} K\left(\frac{X_k - X_i}{h}\right)$$

Computing is done using Trapezoidal Method and data

Computing is done from data

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (6/8)

Constructing Fake Instances of the Continuous Feature - Cross Validation to determine the optimal bandwidth h

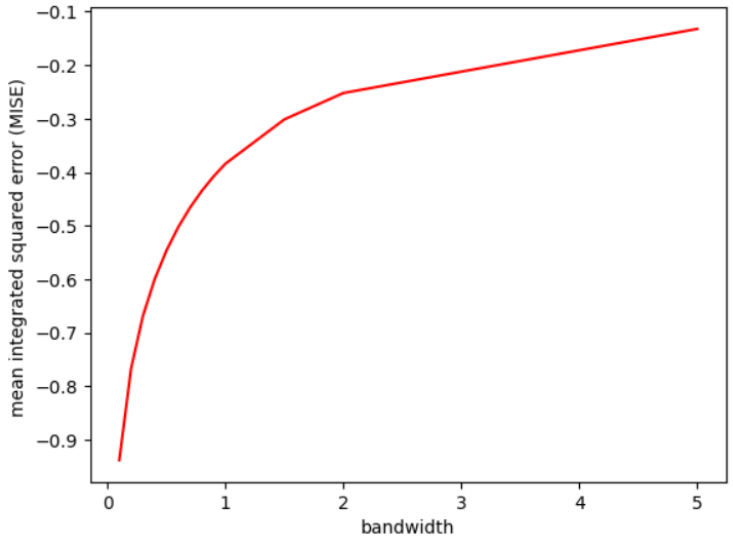


Figure 26. Cross validation to select the best bandwidth h

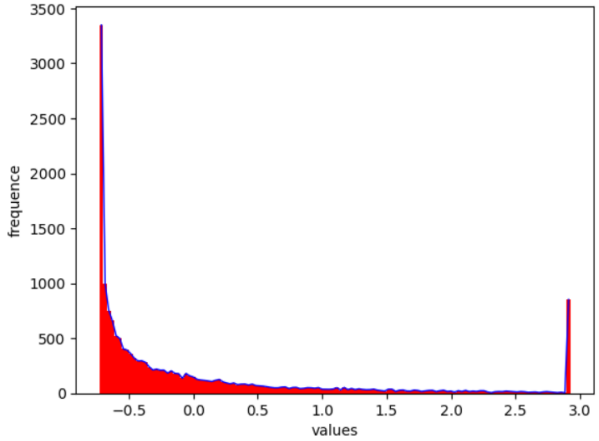


Figure 27. Continuous variable distribution in the training data

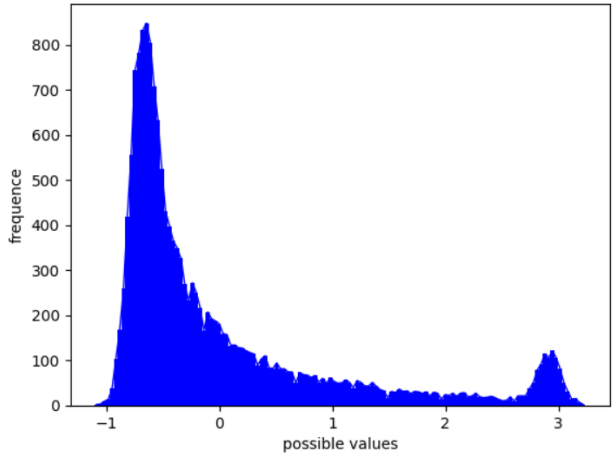


Figure 28. Continuous variable's distribution using Kernel Density Approximation with h = 0.1

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (7/8)

Discriminator Training

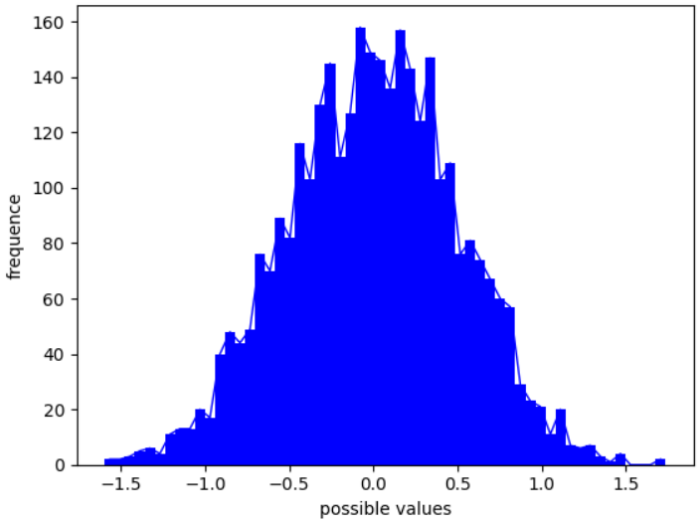


Figure 29. Noise Distribution of Continuous Variable During Test

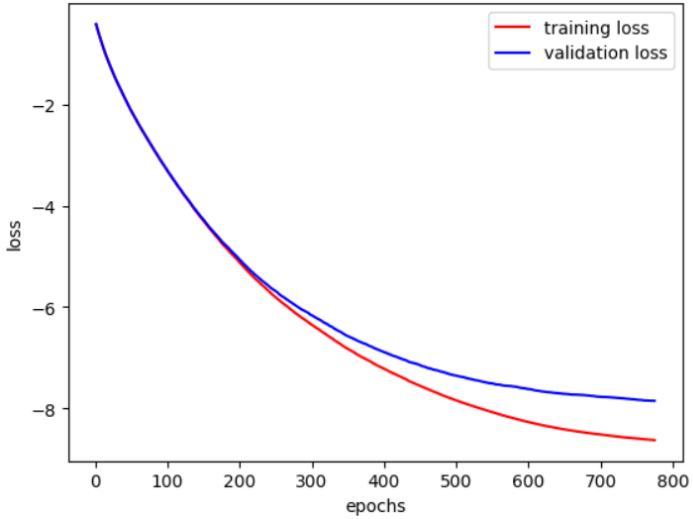


Figure 30. Discriminator (LightGBM) Learning Curve

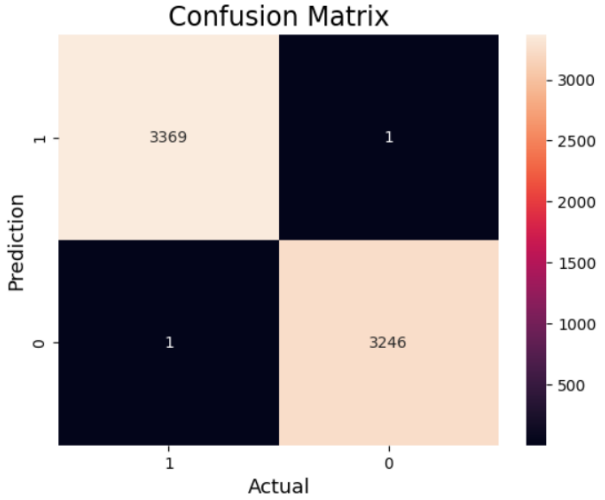


Figure 31. Confusion Matrix of the LightGBM Discriminator

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (8/8)

Framework Results

Models	Role	Training AR (%)	WT Test AR (%)	OOT Test AR (%)
PD estimation models	Baseline	65,40	66,20	66,40
LightGBM with regularization	Teacher	70.87	67.41	71.44
PD estimation models distilled with response-based distillation with temperature	Student	68.63	64.51	70.83
PD estimation models distilled with response-based with temperature + Adversarial Knowledge Distillation	Student	68.99	66.20	72.14

Table 6. Students' Performance using Adversarial Distillation Framework

X-DISTILLATION FRAMEWORK (1/8)

Exploring Interpretability Distillation (XD) Technique to Improve the Student AR TEST

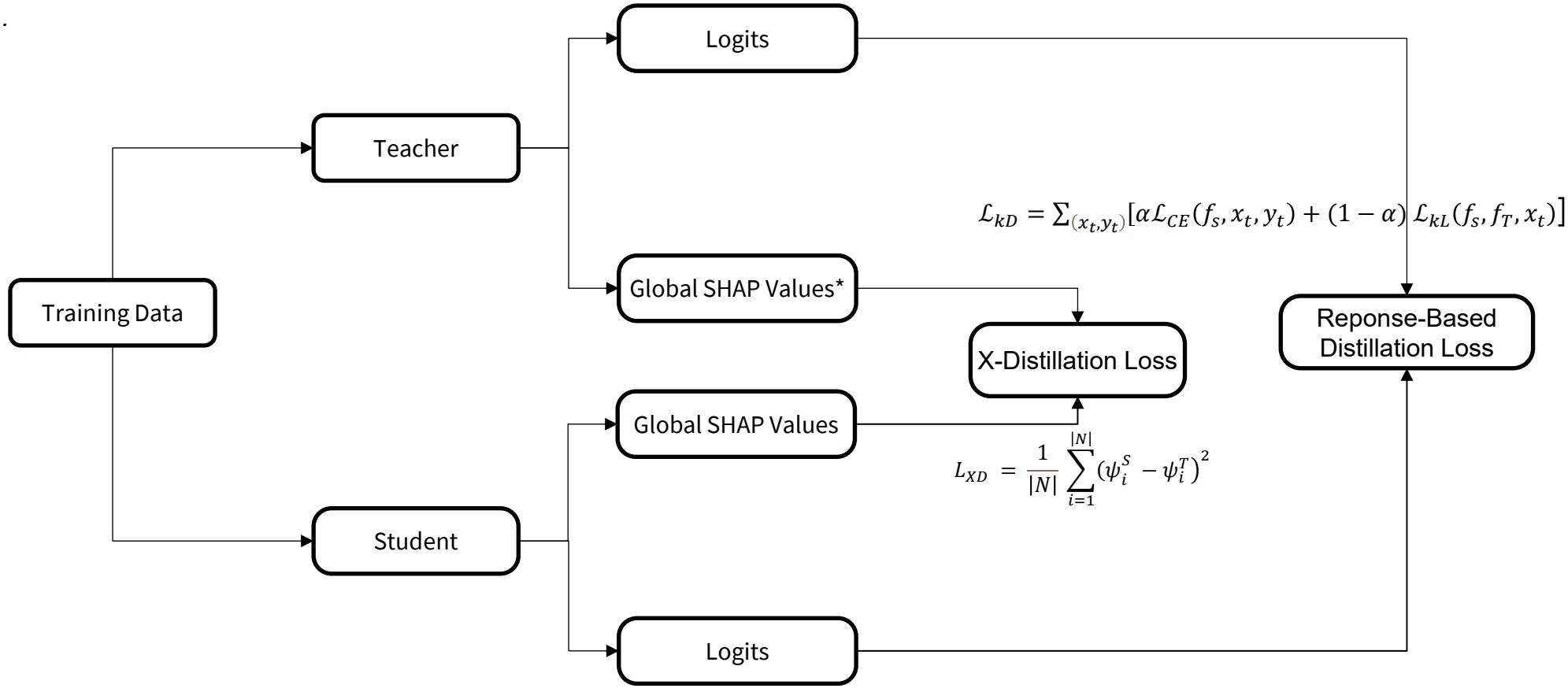


Figure 32. X-Distillation framework used to enhance student's performance

** Corresponds to the absolute mean of local SHAP values across training instances*

X-DISTILLATION FRAMEWORK (2/8)

Student's Global SHAP Values Expression as a Function of Logistic Regression Parameters – Approximation

In our case, the expression of X-distillation loss (or the optimization objective) can be expressed as the following:

$$L_{XD} = \frac{1}{|N|}\sum_{i=1}^{|N|}(\psi_i^S - \psi_i^T)^2$$

With :

ψ_i^S : Global SHAP Value of feature i using the student model

ψ_i^T: Global SHAP Value of feature i using the teacher model

N : The set of features aka independent variables

*To update our student's (Logistic Regression) parameters β_i. **ψ_i^S must be expressed as a function of β_i** in order to perform gradient descent. However, calculating the theorical expression of ψ_i^S as a function of logistic regression parameters is not straightforward.*

X-DISTILLATION FRAMEWORK (3/8)

Student's Global SHAP Values Expression as a Function of Logistic Regression Parameters – Approximation

The SHAP values kernel explainer can be expressed as:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

With :

f : the model we want to explain, in our case, the logistic regression student; $\phi_i(f)$: Local shapley value for the instance i; S : the set of a possible coalition within all features except I; N : The set of all features AKA independent variables; $f(S \cup \{i\})$: instance i prediction using the model trained only on the set $S \cup \{i\}$; $f(S)$: Instance i prediction using the model trained only on the set S.

In our case, f is a linear regression model. Let x_j be an instance vector, and v a specific combination within $|S|$ elements defined as : $v(X, S) = \{X \,/\, X \subseteq S \text{ and } |X| = |S|\}$. $\forall\, S \subseteq N \setminus \{i\}$ we have :

$$f(S \cup \{i\})(x_j) = \beta_0^i + \beta_{v(1)}^i x_{v(1),j} + \beta_{v(2)}^i x_{v(2),j} + \ldots\ldots\ldots\ldots\ldots + \beta_i^i x_{i,j} + \ldots\ldots\ldots + \beta_{v(|S|)}^i x_{v(|S|),j}$$

$$f(S)(x_j) = \beta_0^{-i} + \beta_{v(1)}^{-i} x_{v(1),j} + \beta_{v(2)}^{-i} x_{v(2),j} + \ldots\ldots\ldots\ldots\ldots + \beta_i^{-i} x_{i,j} + \ldots\ldots\ldots + \beta_{v(|S|)}^{-i} x_{v(|S|),j}$$

With:

$\beta_{v(k)}^i$ are the coefficient obtained by training a linear regression model on $S \cup \{i\}$;

$\beta_{(k)}^{-i}$ are the coefficient obtained by training a linear regression model on S;

X-DISTILLATION FRAMEWORK (4/8)

Student's Global SHAP Values Expression as a Function of Logistic Regression Parameters – Approximation

So far: $f(S \cup \{i\})(x_j) = \sum_{k \neq i}^{|S|} \beta_{v(k)}^{i} \, x_{v(k),j} \quad + \quad \beta_i^i \, x_{i,j} \quad and \quad f(S)(x_j) = \sum_{k \neq i}^{|S|} \beta_{v(k)}^{-i} \, x_{v(k),j}$

So, the Shapley value of a feature i is simplified as :

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[\sum_{k \neq i}^{|S|} \left(\beta_{v(k)}^{i} - \beta_{v(k)}^{-i} \right) x_{v(k),j} \quad + \quad \beta_i^i \, x_{i,j} \right]$$

Let's denote:

$$C_1^S = \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \qquad ; \; C_{2,i}^S = \sum_{k \neq i}^{|S|} \left(\beta_{v(k)}^{i} - \beta_{v(k)}^{-i} \right) x_{v(k),j}$$

In this case, we have:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} C_1^S \left[C_{2,i}^S \; + \; \beta_i^i \, x_{i,j} \right] .$$

Approximation : Let's assume that $C_{2,i}^S = 0$, which means that $\beta_{v(k)}^{i} = \beta_{v(k)}^{-i}$. Training f for each coalition S is cumbersome. Instead, we attribute the value zero to features excluded S.

X-DISTILLATION FRAMEWORK (5/8)

Student's Global SHAP Values Expression as a Function of Logistic Regression Parameters – Approximation

- To compute global Shapley Values for each feature i in the dataset, we take the mean absolute value of $\phi_i(f)\ \forall\ i\ \in\ N$.
- **The motivation behind taking the absolute mean is that we are interested in Shapley values magnitude, and we do not want any compensation effect between positive and negative values due to the mean summation.**
- Let's denote ψ_i^S the global Shapley value of feature i using the student model. we have :

$$\psi_i^S\ =\ E\,|\phi_i(f)|\ =\ \sum_{S\subseteq N\setminus\{i\}} C_1^S\ \times\ E_j[\,|\,\beta_i^i\ x_{i,j}\,|\,]$$

$$\boldsymbol{\psi_i^S\ =\ |\beta_i|\ \times\ \sum_{S\subseteq N\setminus\{i\}} C_1^S\ \times\ E_j|\,x_{i,j}\,|}$$

because $\sum_{S\subseteq N\setminus\{i\}} C_1^S\ > 0\ \ and\ \ \beta_i\ is\ not\ a\ random\ variable\ depending\ on\ instances.$

Now that we have the expression of the Shapley values of our linear model f expressed using the coefficients β_i . we can calculate the gradient of XDistillation loss between student's Shapley values and teacher Shapley values

X-DISTILLATION FRAMEWORK (6/8)

Student's Global SHAP Values Expression as a Function of Logistic Regression Parameters – Approximation

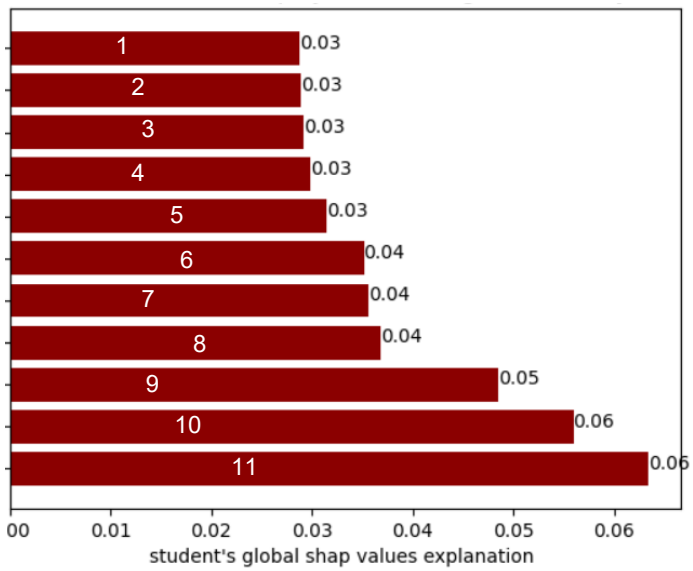


Figure 33. Student's global Shapley values using SHAP library in Python

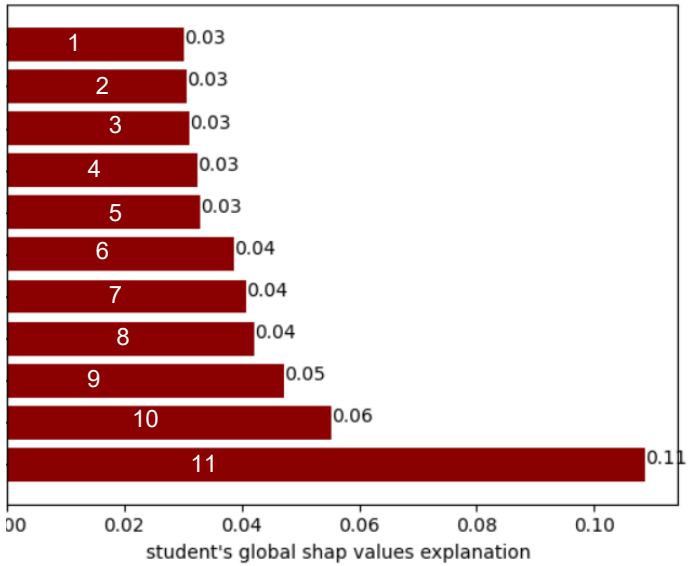


Figure 34. Student's global Shapley values using kernel explainer approximation

XDISTILLATION FRAMEWORK (7/8)

Student's Global SHAP Values Expression as a Function of Logistic Regression Parameters – Approximation

The XDistillation loss is expressed as :

$$L_{XD} = \frac{1}{|N|} \sum_{i=1}^{|N|} (\psi_i^S - \psi_i^T)^2$$

Let's denote : $Cte_s^i = \sum_{S \subseteq N \setminus \{i\}} C_1^S \times E_j \left| x_{i,j} \right|$ thus, $L_{XD} = \frac{1}{|N|} \sum_{i=1}^{|N|} (Cte_s^i \times |\beta_i| - \psi_i^T)^2$

The gradient of the absolute value function is not well-defined at zero because it's a non-smooth function at that point. However, we can compute sub-gradients of the absolute value when $\beta_i > 0$ *and* $\beta_i < 0$.

$$\frac{\partial L_{XD}}{\partial \beta_i} = \frac{2}{|N|} \times (-1)^{\mathbb{1}(\beta_i < 0)} \times Cte_s^i \times (|\beta_i| \times Cte_s^i - \psi_i^T)$$

The gradient descent optimization will be then :

$$\beta_{i,m+1} = \beta_{i,m} - \alpha \times \frac{\partial L_{XD}}{\partial \beta_i}$$

X-DISTILLATION FRAMEWORK (8/8)

X-Distillation Framework Results

Models	Role	Training AR (%)	WT Test AR (%)	OOT Test AR (%)
PD estimation models	Baseline	65,40	66,20	66,40
LightGBM with regularization	Teacher	70.87	67.41	71.44
PD estimation models distilled with response-based distillation only without temperature	Student	68.03	63.94	70.13
PD estimation models trained with response-based distillation only with temperature	Student	68.63	64.51	70.83
PD estimation models distilled with response-based distillation with temperature + adversarial knowledge distillation	Student	68.99	66.20	72.14
PD estimation models trained with Response-Based Distillation with temperature + X-Distillation	Student	70.09	66.01	72.78

Table 7. X-Distillation framework performance comparing with other frameworks

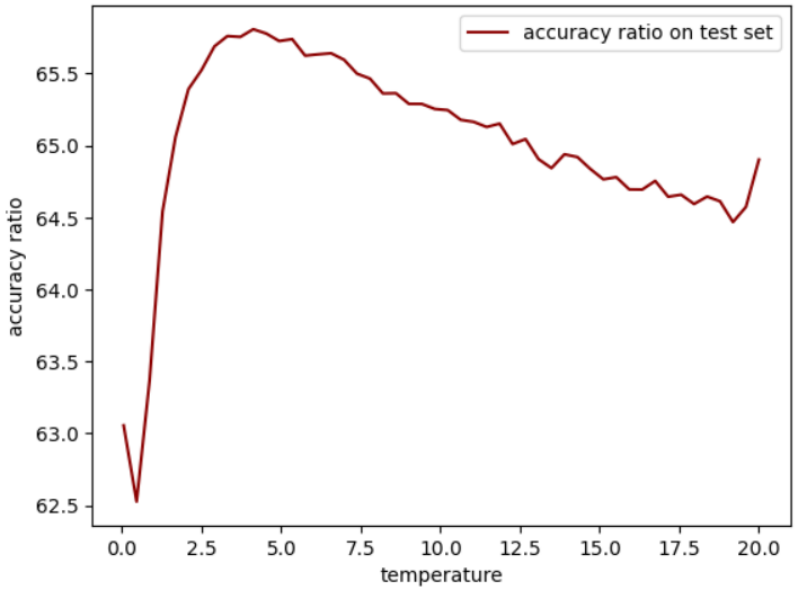


Figure 35 . Cross validation for best temperature T selection in X-Distillation. Here T = 4.12

Takeaways :

- ❑ **+ 6.38 gain in performance** on the OOT Test AR.
- ❑ Same performance on WT Test AR due to dataset variability.

THANK YOU FOR YOUR ATTENTION

C'EST VOUS L'AVENIR

SOCIETE GENERALE

Centrale CASABLANCA

REFERENCES

1. Craven and al., 1995: Extracting Tree-Structured Representations of Trained Networks
2. Caruana and al.,2006: model compression
3. Hinton and al.,2015: Distilling the Knowledge in a Neural Network
4. Han and al., 2015: Learning both Weights and Connections for Efficient
5. Hoffman and al., 2015: Cross Modal Distillation for Supervision Transfer
6. Zagoruyko and al., 2017: Attention Transfer
7. Huang and al., 2017: Knowledge Distill via Neuron Selectivity Transfer
8. Caruana and al., 2017: Interpretable & Explorable Approximations of Black Box Models
9. Yim and al, 2017: A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning
10. Burda, Edwards and al., 2018: Exploration by Random Network Distillation
11. Caruana and al., 2018: Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation
12. Liu and al., 2018: Improving the Interpretability of Deep Neural Networks with Knowledge Distillation
13. Asadulaev and al., 2019: Interpretable Few-Shot Learning via Linear Distillation
14. Bastani and al., 2019: Interpreting Blackbox Models via Model Extraction
15. Zhang and al, 2021: Adversarial co-distillation learning for image recognition

APPENDIX- DISTILLATION TESTS ON LENDING CLUB DATASET (1/7)

Teacher Training – Feed-Forward Neural Networks

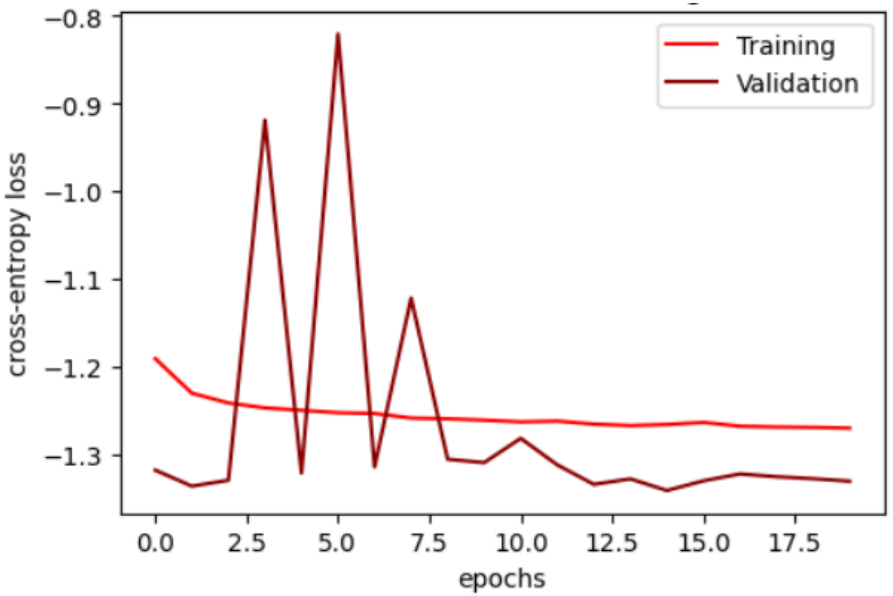


Figure 36. Neural Network Teacher's Learning Curve

Metric	Train	Test
ROC AUC	0.909	**0.906**
F1-Score	0.6281	**0.6249**

Table 8. Teacher Performance

APPENDIX- DISTILLATION TESTS ON LENDING CLUB DATASET (2/7)

Teacher Training – XGBOOST

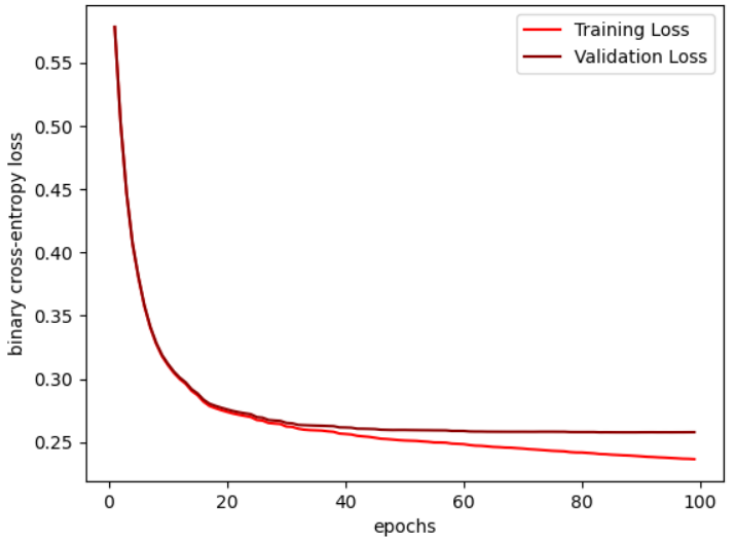


Figure 37. XGBOOST Teacher's Learning Curve

Metric	Train	Test
ROC AUC	0.909	**0.906**
F1-Score	0.651	**0.627**

Table 9. XGBOOST Teacher Performance

APPENDIX- DISTILLATION TESTS ON LENDING CLUB DATASET (3/7)

Student Training – Feed-Forward Neural Network

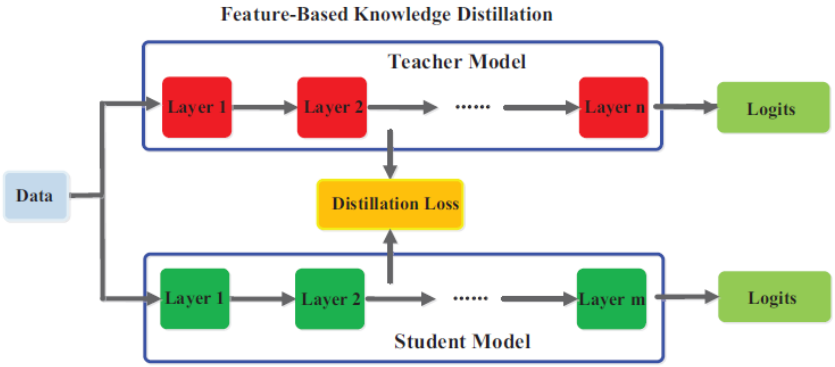


Figure 38. The generic feature-based knowledge distillation

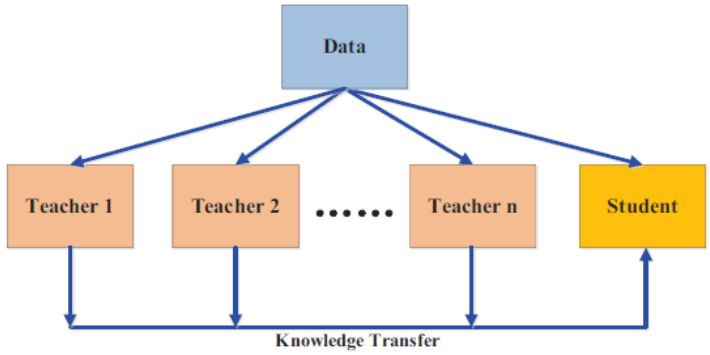


Figure 39. Multi-teacher Distillation Framework's Illustration

Models	ROC AUC TRAIN	ROC AUC TEST
FFNN Teacher	90.96%	90.65%
XGBOOST Teacher	92,86%	90.73%
Student with **Hinton-Based Distillation**	71.45%	71.48%
Student with **Multi-Teacher Distillation**	88.95%	**88.97%**
Student with **Feature-Based Distillation**	71.44%	71.47%

Selected Teacher

Table 10. Student's Performance using different Distillation Frameworks

Compression Ratio = 51.41

Model	#parameters
Teacher	58,357
Student	1,135

Table 11. Student's performance

APPENDIX (4/7)

Cross Validation to determine the optimal bandwidth h

$$\mathrm{MISE(h)} = \mathrm{E}\left[\int \left(\mathrm{p(x)} - \widehat{p_n^h}(x)\right)^2 \mathrm{dx}\right] = \mathrm{E}\left[\int \left(\widehat{p_n^h}(x)\right)^2 - 2\,\widehat{p_n^h}(x)\mathrm{p(x)} + \left(\mathrm{p(x)}\right)^2 \mathrm{dx}\right]$$

Don't depend on h

$$= \mathrm{E}\left[\int \left(\widehat{p_n^h}(x)\right)^2 \mathrm{dx}\right] - 2 \times \mathrm{E}\left[\int \widehat{p_n^h}(x)\mathrm{p(x)}\,\mathrm{dx}\right] + \mathrm{E}\left[\int \left(\mathrm{p(x)}\right)^2 \mathrm{dx}\right]$$

The goal is construct an unbiased estimator of MISE then we minimize it according to h. Let's denote :

$$\tau(h) = \mathrm{E}\left[\int \left(\widehat{p_n^h}(x)\right)^2 \mathrm{dx}\right] - 2 \times \mathrm{E}\left[\int \widehat{p_n^h}(x)\mathrm{p(x)}\,\mathrm{dx}\right]$$

- $\int \left(\widehat{p_n^h}(x)\right)^2 \mathrm{dx}$ is an unbiased estimator of $\mathrm{E}\left[\int \left(\widehat{p_n^h}(x)\right)^2 \mathrm{dx}\right]$
- Let's construct an unbiased estimator for the second term:

$$E_p\left[\int \widehat{p_n^h}(x)p(x)\,dx\right] = E_p\left[\int \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{X_i - x}{h}\right)\right)p(x)\;dx\right]$$

APPENDIX (5/7)

Cross Validation to determine the optimal bandwidth h

$$E_p\left[\int \widehat{p_n^h}(x)p(x)\,dx\right] = \int E_p\left[\frac{1}{h}K\left(\frac{X_i - x}{h}\right)\right]p(x)\;dx = \int\int \frac{1}{h}K\left(\frac{y-x}{h}\right)p(y)p(x)\;dy\,dx$$

Using the leave one-out estimator leave-one out of p, let's denote : $\hat{G} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{k\neq i}^{n}\frac{1}{h}K\left(\frac{X_k - X_i}{h}\right)$

Which give this passing the expectancy $E_p(\hat{G}) = \frac{1}{n(n-1)}\sum_{k\neq i}\frac{1}{h}E_p\left[K\left(\frac{X_k - X_i}{h}\right)\right]$

The joint distribution of (X_k, X_i) is $p(y)p(x)$ (Independent Variables), Thus,

$$E_p(\hat{G}) = \frac{1}{h}\int\int K\left(\frac{y-x}{h}\right)p(y)p(x)\;dy\,dx$$

So $\hat{G}$ is an unbiased estimator of $\mathrm{E}\left[\int \widehat{p_n^h}(x)\mathrm{p}(\mathrm{x})\,\mathrm{dx}\right]$. Thus, the cross-validation objective is defined as:

$$CV(h) = \int\left(\widehat{p_n^h}(x)\right)^2 dx - 2\hat{G}\;;\; h_{CV} \in arg\min_{h>0} C\,V(h)$$

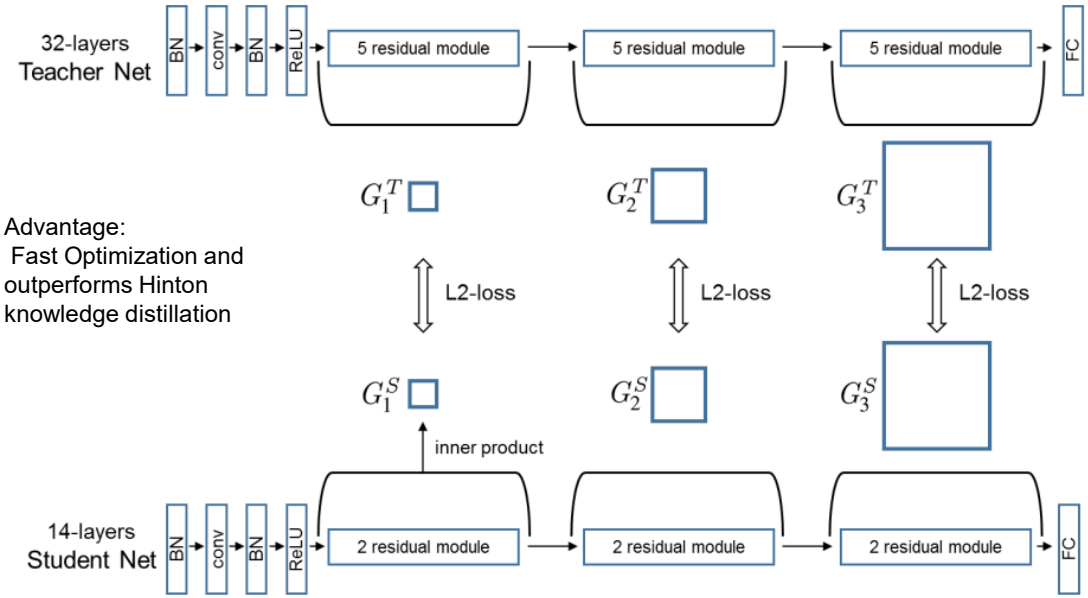
Trapezoidal method

From data

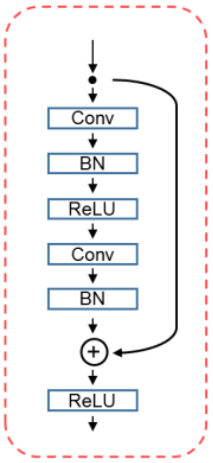
APPENDIX (6/7)

EXAMPLE OF RELATION-BASED DISTILLATION FRAMEWORK

A novel example is distilling the flow of solution process (FSP) matrix Yim and al, 2017. The FSP matrix is generated by crossing feature representation of two selected layers across data using inner product, gram representation, or other capturing-information product depending on the problem.



Advantage:
Fast Optimization and outperforms Hinton knowledge distillation



Residual module

Algorithm 1 Transfer the distilled knowledge

Stage 1: Learning the FSP matrix

Weights of the student and teacher networks: W_s, W_t

1: $W_s = \arg\min_{W_s} L_{FSP}(W_t, W_s)$

Stage 2: Training for the original task

1: $W_s = \arg\min_{W_s} L_{ori}(W_s)$

$$G_{i,j}(x;W) = \sum_{s=1}^{h}\sum_{t=1}^{w} \frac{F^1_{s,t,i}(x;W) \times F^2_{s,t,j}(x;W)}{h \times w}$$

$$L_{FSP}(W_t, W_s) = \frac{1}{N}\sum_{x}\sum_{i=1}^{n} \lambda_i \times \|(G_i^T(x;W_t) - G_i^S(x;W_s)\|_2^2.$$

Figure 40. Complete architecture of FSP KD. There are two stages. In stage 1, the student network is trained to minimize the distance between the FSP matrices of the student and teacher networks. Then, the pretrained weights of the student DNN are used for the initial weight in stage 2. Stage 2 represents the normal training procedure

Example of FSP construction as in Yim and al, 2017: 1) Take the feature map matrix of layer i and layer j. 2) Construct the Gramian matrix; 3) sum up over all terms; 4) repeat the process across selected layers. NB: 1) the number of selected layers must be the same in the teacher and the student; 2) this is only a construction example.

APPENDIX (7/7)

EXAMPLE OF SELF-DISTILLATION FRAMEWORK

Self-distillation is a variant of online distillation where the teacher and student share the same architecture. It is often used to enhance neural networks performance comparing to the traditional training mode. As a framework's example, adversarial co-distillation (ACN) by Zhang and al, 2021 is a novel technique to enhance the performance of a CNN in the image recognition task by generating divergent examples where models do not totally agree. The goal is to have them make the same prediction accurately based on a majority vote mindset.

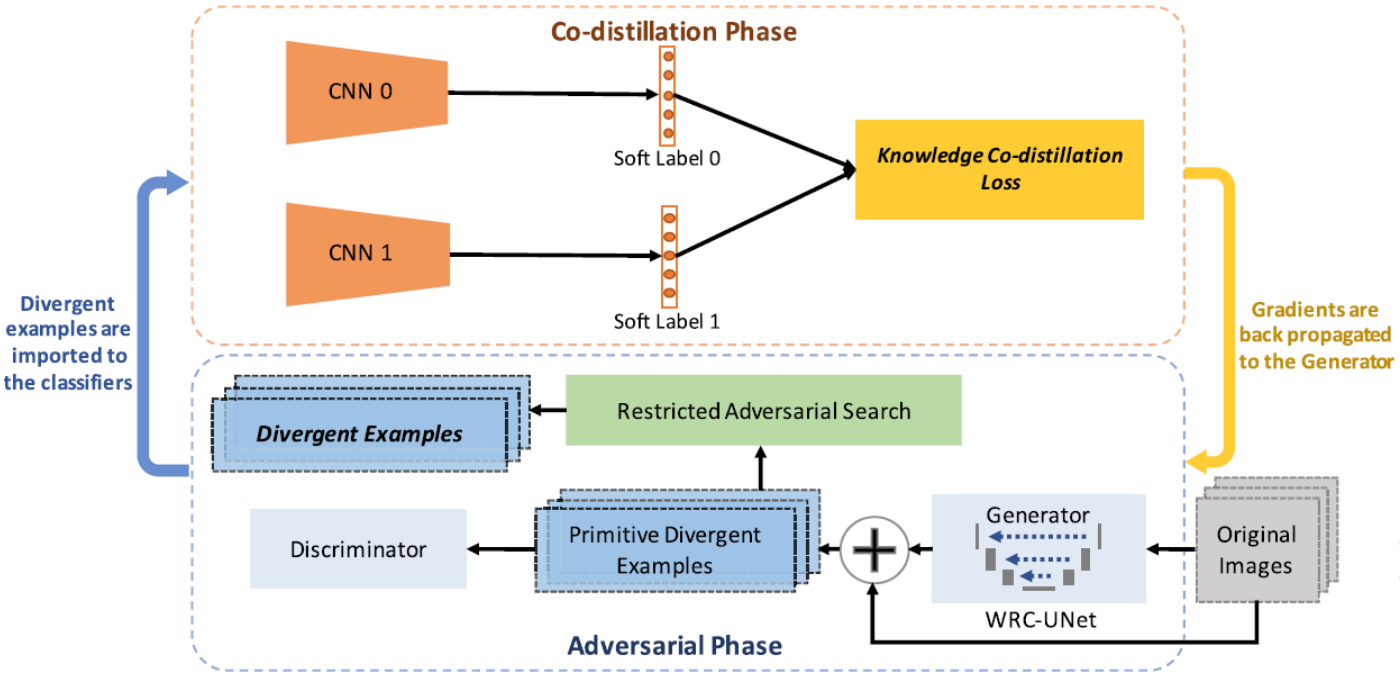


Figure. 41. The framework illustration of ACNs. ACNs consist of an Adversarial Phase and a Co-distillation Phase. The Adversarial Phase generates the divergent examples, and the Co-distillation Phase learn the divergent examples. The Adversarial Phase is designed according to the GANs framework.

Model	Original trained	ACN
Resnet-20	68.22%	70.67%
VGG11	67.38%	70.11%
AlexNet	39.45%	46.27%

Table 12. Distillation learning can be used to enhance complex models' performance without compression

B. RELEVANT FRAMEWORKS (3/3)

MULTI-TEACHER DISTILLATION

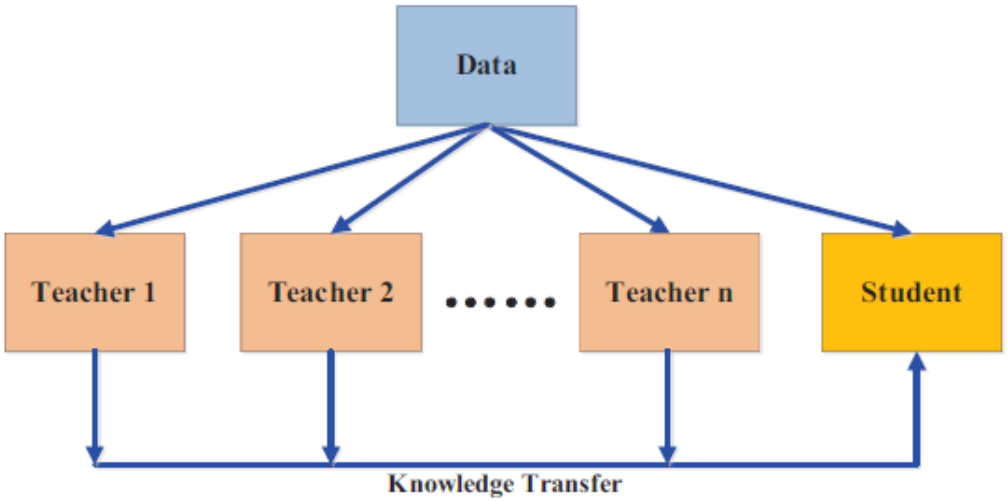


Figure 14. The generic multi-teacher knowledge distillation framework

Problem	Usage example	Pros
1) Bias coming from one the teacher 2) Lack of knowledge using one teacher	2 teachers, one transfers response-based knowledge and the other transfers feature-based knowledge (Chen et al. 2019b).	Provide richer knowledge to the student Straightforward

Table 6. Multi-teacher Distillation Framework's detailed explanation.