DISTILLATION LEARNING TESTS

WORKSHOP 2

RISQ/MRM

C'EST VOUS L'AVENIR

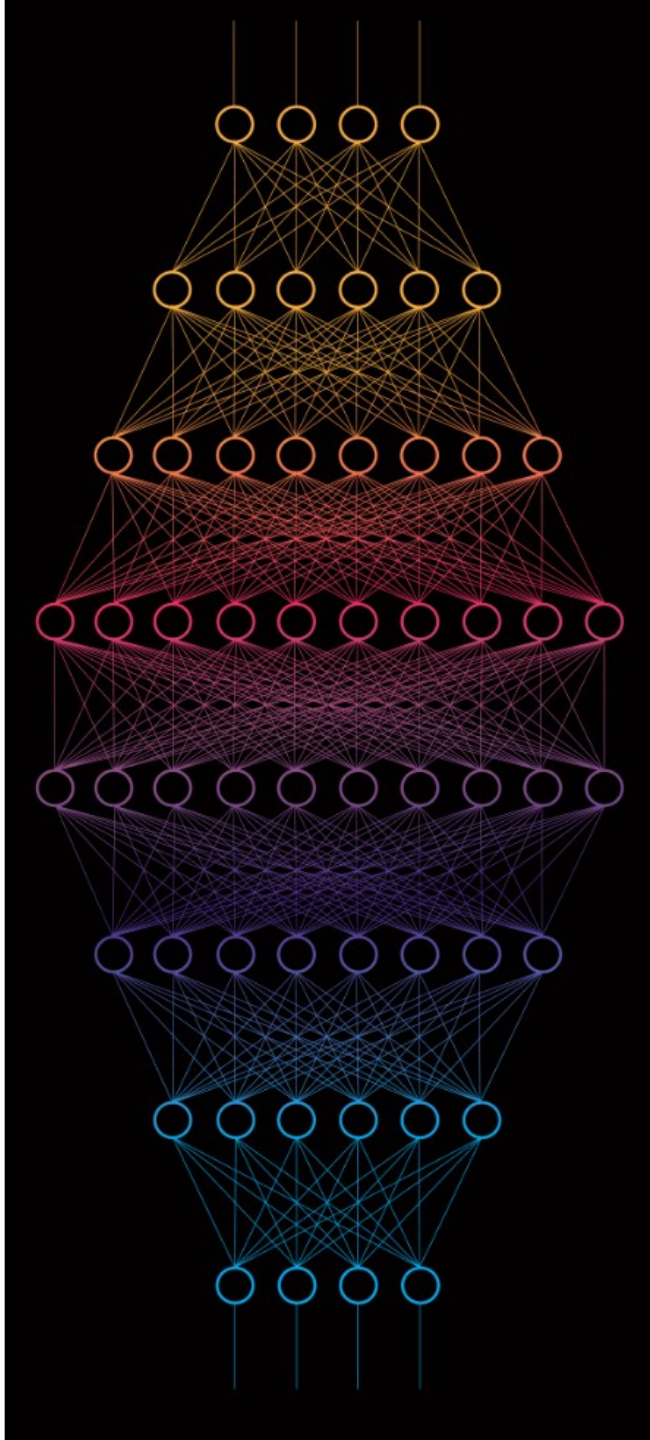
SOCIETE GENERALE

OUTLINES

1. INTRODUCTION

A. Reminder of Knowledge Distillation Fundamentals
B. Main Use of Knowledge Distillation
C. Relevant Applications in Scientific Work

SOCIETE GENERALE



A. REMINDER OF KNOWLEDGE DISTILLATION FUNDAMENTALS

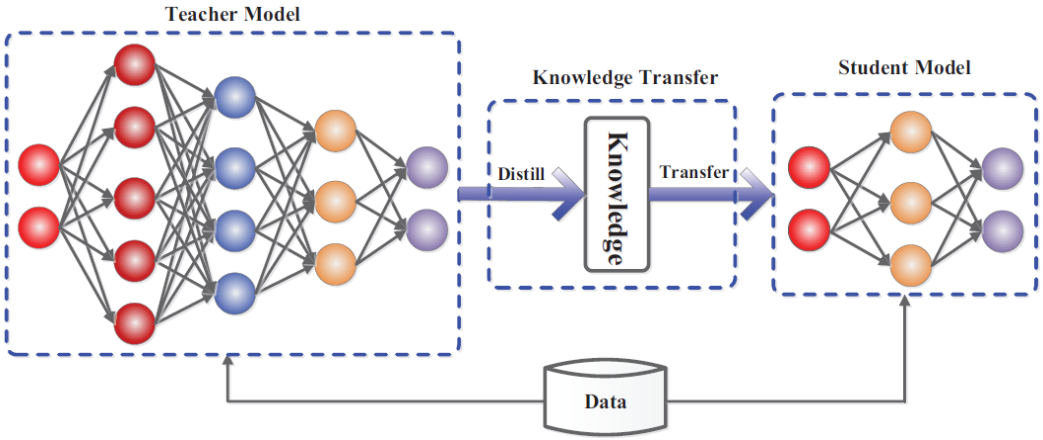


Fig. 1 The generic teacher-student framework for knowledge distillation

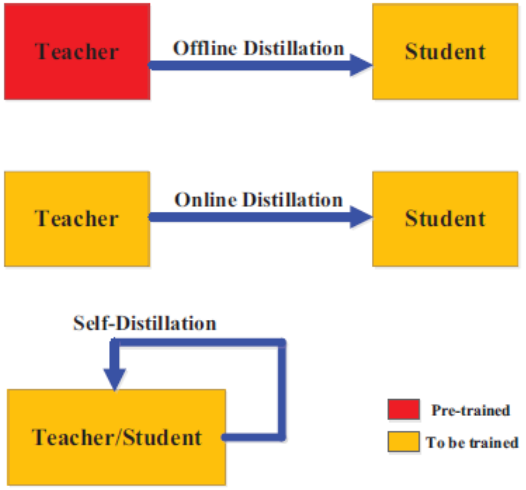


Fig 3. Different Distillation Training Modes

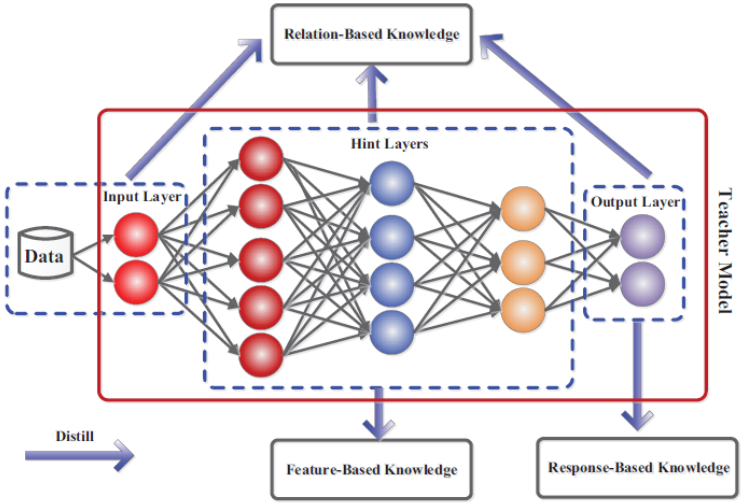


Fig 2. The schematic illustrations of sources of knowledge in a deep teacher network

B. MAIN USE OF KNOWLEDGE DISTILLATION

3 Main Potential Uses of Knowledge Distillation

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.

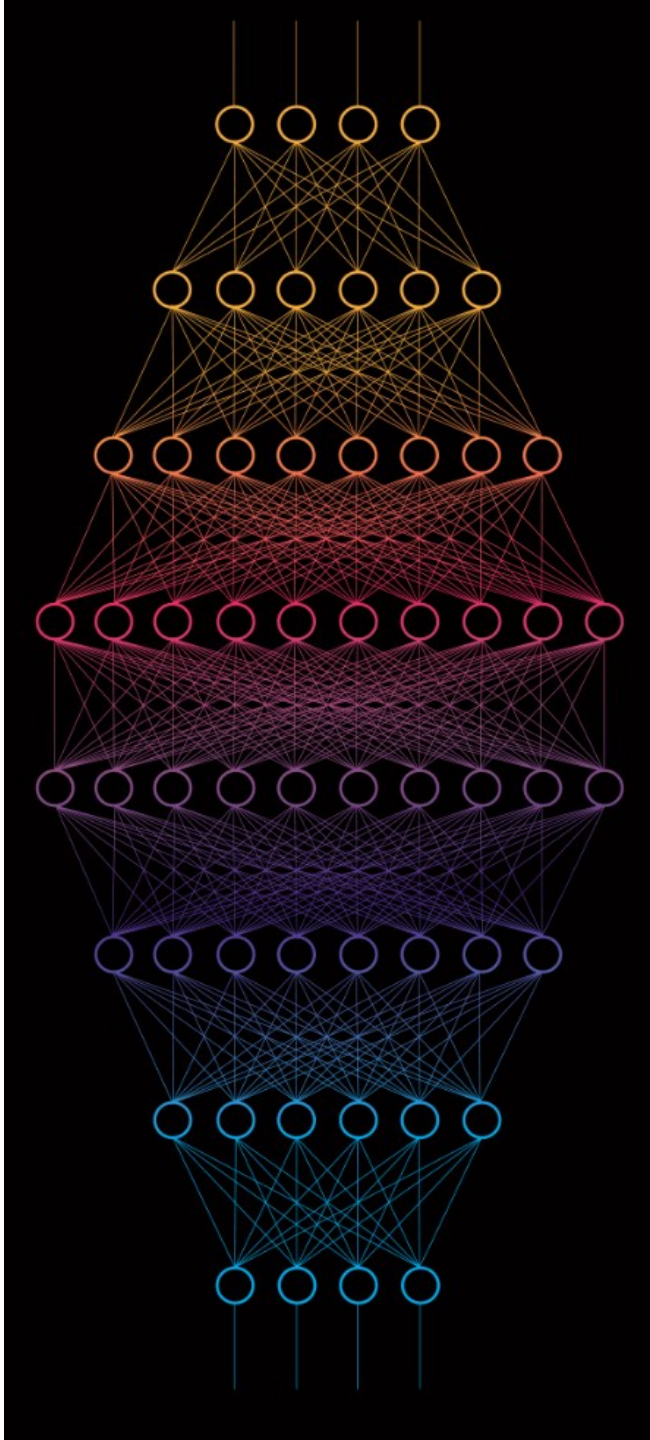
C. RELEVANT APPLICATIONS IN SCIENTIFIC WORK

Article	Description	Performances	Commentary
Improving the Interpretability of Deep Neural Networks with Knowledge Distillation (Liu, et al., 2018).	Distill **Deep Neural Networks (CNN)** into **decision trees** for the MNIST Task. In this work, Application in **XAI and model performance enhancing.**	**Baseline student accuracy: 84%** Teacher accuracy: 99.25% **Distilled student accuracy: 86.6 %**	The choice of an interpretable student always creates a **performance gap** between the student and the teacher. However, **outperforming the baseline is always possible in interpretability applications**.
Natural Language Generation for Effective Knowledge Distillation (Tang, et al., 2019)	**Distill a BERT** (Devlin, et al., 2018) teacher into a **smallest BiLSTM** using a transfer set constructed by **GPT-2** (Radford, et al., 2019) and **TXL** (Dai, et al., 2019), **application in model complexity reduction**.	**BiLSTM baseline accuracy: 87.6%** BERT Teacher accuracy: 94.9% **Distilled BiLSTM student** on GPT-2 transfer dataset accuracy: 92.7 %	BERT models are very heavy and requires important computational resources for training. Distillation learning **reduces complexity** while staying faithful to the original performance.

Table 1. Relevant Distillation Learning Applications in Literature

2. DISTILLATION OF PD ESTIMATION MODELS

A. Teacher Training
B. Hinton-Based Distillation
C. Adversarial Knowledge Distillation Framework
D. Xdistillation Framework



SOCIETE GENERALE

INTRODUCTION TO PD ESTIMATION MODELS FRAMEWORK

PD ESTIMATION MODELS

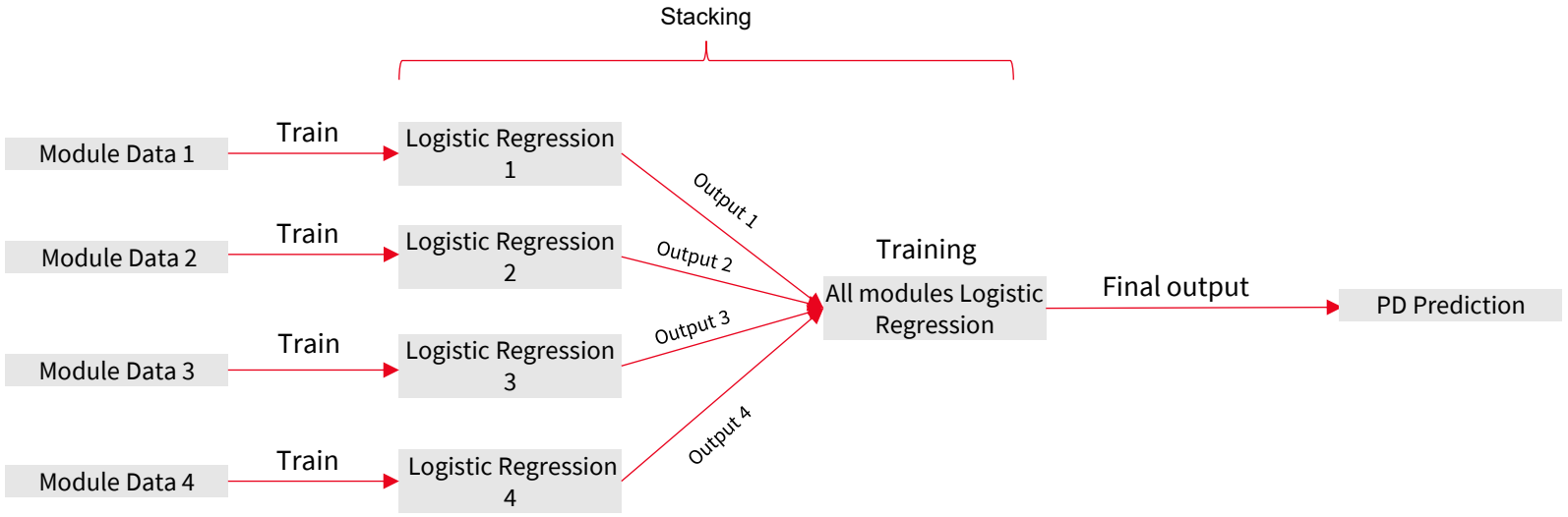


Figure 3. PD ESTIMATION MODELS Training Workflow

Modules	AR* Train	AR Test	AR OOT
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 2. Performance of PD ESTIMATION MODELS

2 TYPES OF TEST DATA ARE USED:

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

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

*Accuracy Ratio = 2 * ROC AUC – 1

PD ESTIMATION MDOELS' DISTILLATION FRAMEWORK

The objective of the distillation framework is to **enhance the overall *PD ESTIMATION MODELS* performance**.

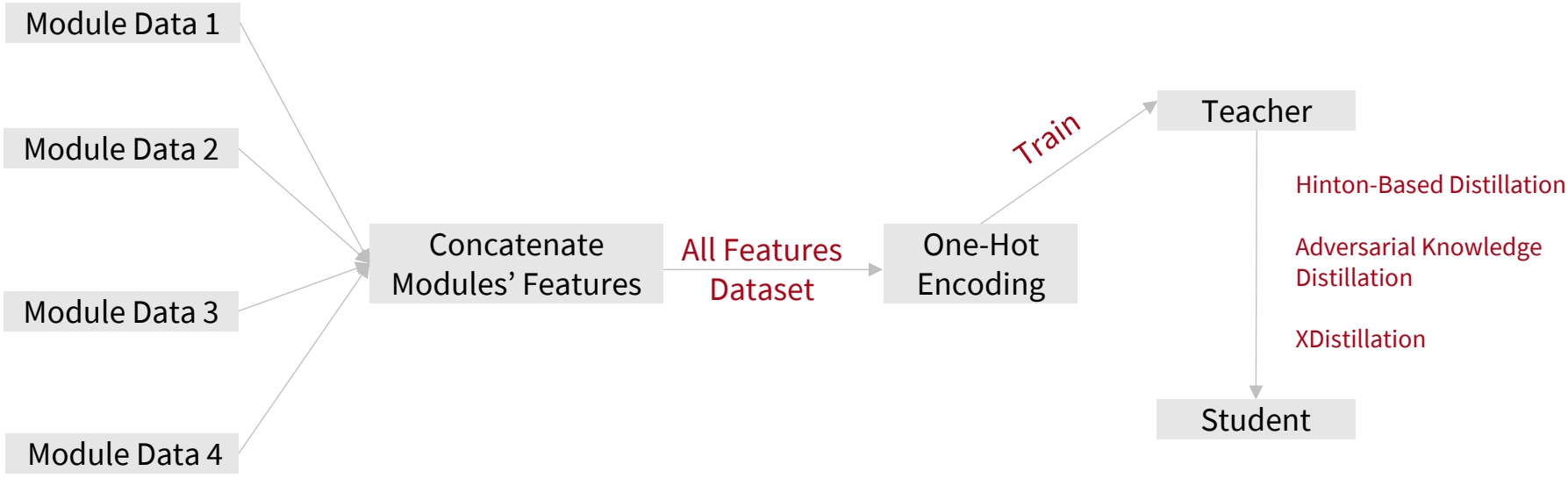


Figure 4. PD ESTIMATION MODELS Distillation Framework

TEACHER TRAINING

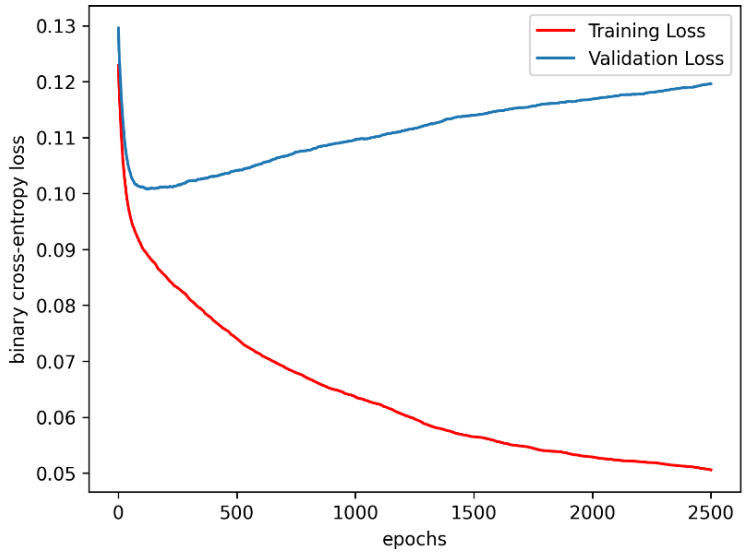


Figure 5. LightGBM teacher learning curves without regularization

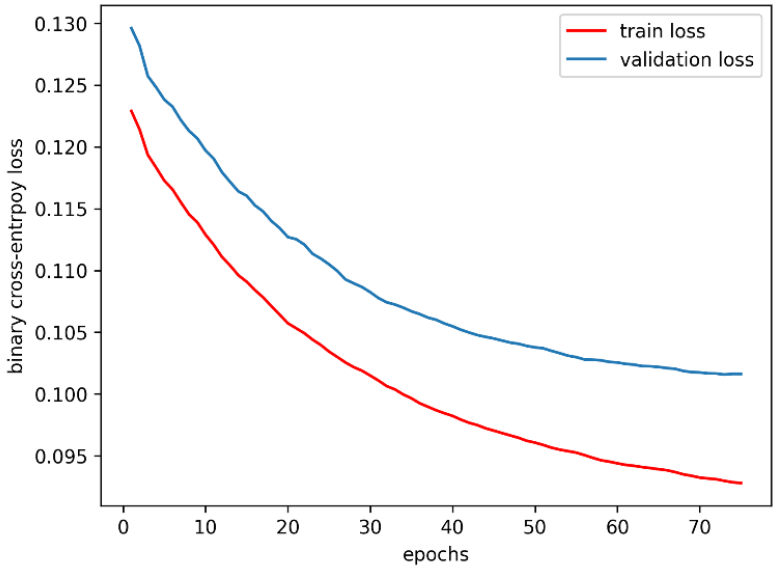


Figure 6. LightGBM teacher learning curves with regularization (L1 = 0.01 and Early Stopping = 7)

Models	AR TRAIN	AR TEST	AR OOT
PD estimation models	65,4%	66,2%	66,4%
Feed-Forward Neural Networks (FFNN)	72.28 %	63.25 %	71.28 %
LightGBM with regularization	**70.87 %**	**67.41 %**	**71.44 %**
LightGBM without regularization	89.16 %	58.55 %	58.84 %

Selected Teacher

Table 3. Performance of offline trained teachers using all modules' features

HINTON-BASED DISTILLATION (1/3)

Response-Based Online Distillation

As a baseline, we choose a decision tree student model as a reference. For comparison matters with *PD ESTIMATION MODELS*, we train a logistic regression using Hinton-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.

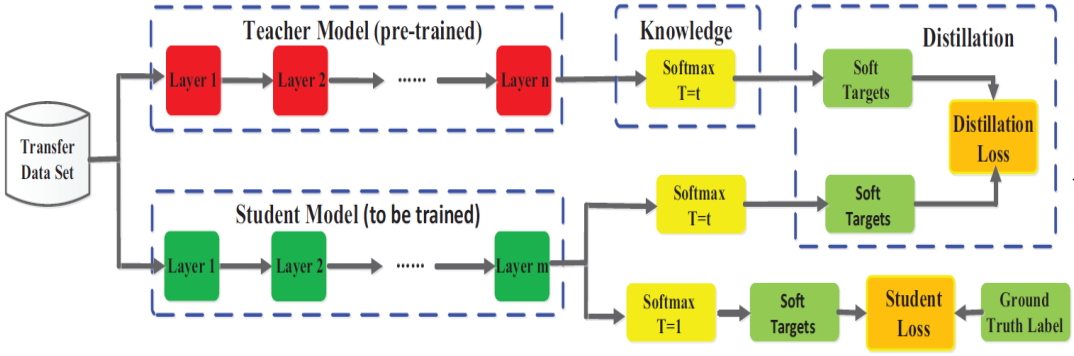


Figure 7. Hinton-Based Distillation Framework

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

Formula 1. Response-Based Knowledge Distillation Loss, α is the contribution of the loss comparing to ground truth labels.

HINTON-BASED DISTILLATION (2/3)

Response-Based Online Distillation

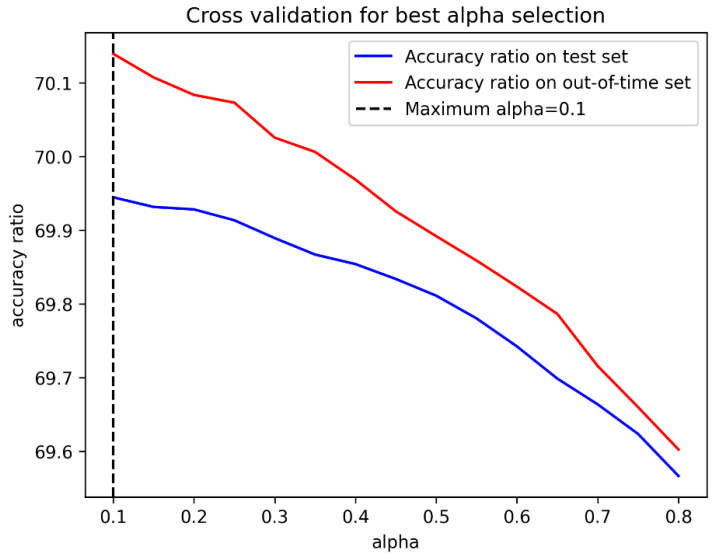


Figure 8. Cross-validation for best α selection

Models	Role	AR TRAIN	AR TEST	AR OOT
PD estimation models models	Baseline	65,4%	66,2%	**66,4%**
LightGBM with regularization	Teacher	70.87 %	67.41 %	71.44 %
Decision Tree Distilled	Student	68.5 %	64.48 %	70.16 %
Logistic Regression Distilled	Student	68.03 %	63.94 %	70.13 %

Remarkable Improvement on out-of-time set

Table 4. Hinton-Based Students' Performance

HINTON-BASED DISTILLATION (3/3)

Response-Based Online Distillation with Temperature T

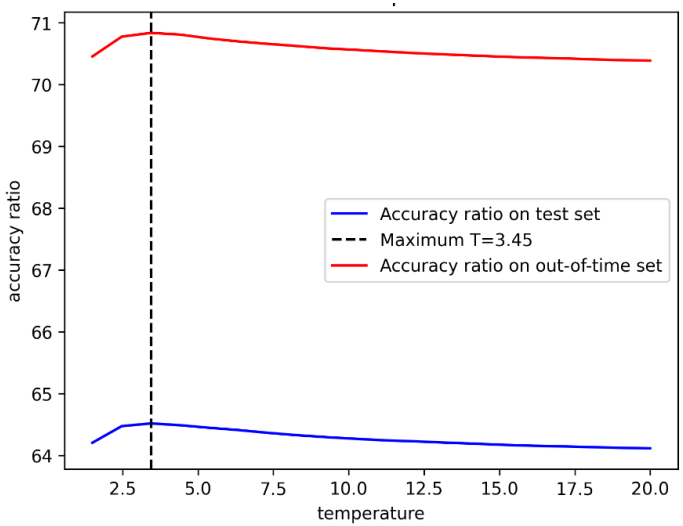


Figure 9. Cross-validation for best T' selection, T = 3.45 corresponds to the best accuracy ratio on the test set.

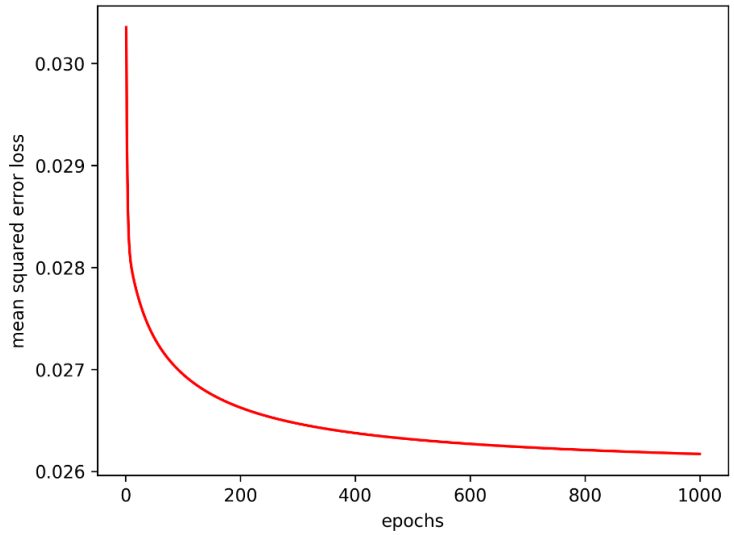


Figure 10. learning curve's training convergence with T = 3.45

Models	Role	AR TRAIN	AR TEST	AR OOT
PD ESTIMATION MODELS	Baseline	65,4%	66,2%	**66,4%**
LightGBM with regularization	Teacher	70.87 %	67.41 %	71.44 %
Logistic Regression Distilled	Student	68.03 %	**63.94 %**	70.13 %
Logistic Regression Distilled with Temperature	Student	68.63 %	**64.51 %**	70.83 %

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

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

Formula 2. Teacher's softened predictions with temperature T

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (1/10)

Exploring Generative Knowledge Distillation Technique to Improve the Student AR TEST

So far, we have achieved great result on the OOT set (AR OOT). However, the student accuracy ratio on the test set (AR TEST) still need to be improved. To achieve that, we will be performing an advanced distillation framework called adversarial knowledge distillation.

An effective framework to enhance the power of student learning via the teacher knowledge distillation using GANs. This framework tackles the problem arising from the small capacity of the student and difficulties to mimic accurately the teacher which lead to the performance gap.

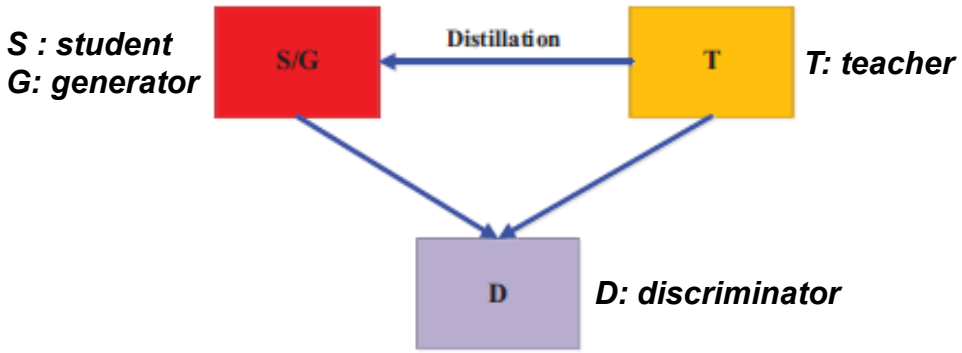


Figure 11. In addition to Hinton-based distillation process, the student will generate new data based on its internal feature distribution corrected each time by the discriminator* which *is trained to discriminate real and fake feature distribution.

** More details about adversarial knowledge distillation framework is provided in Appendix*

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (2/10)

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.

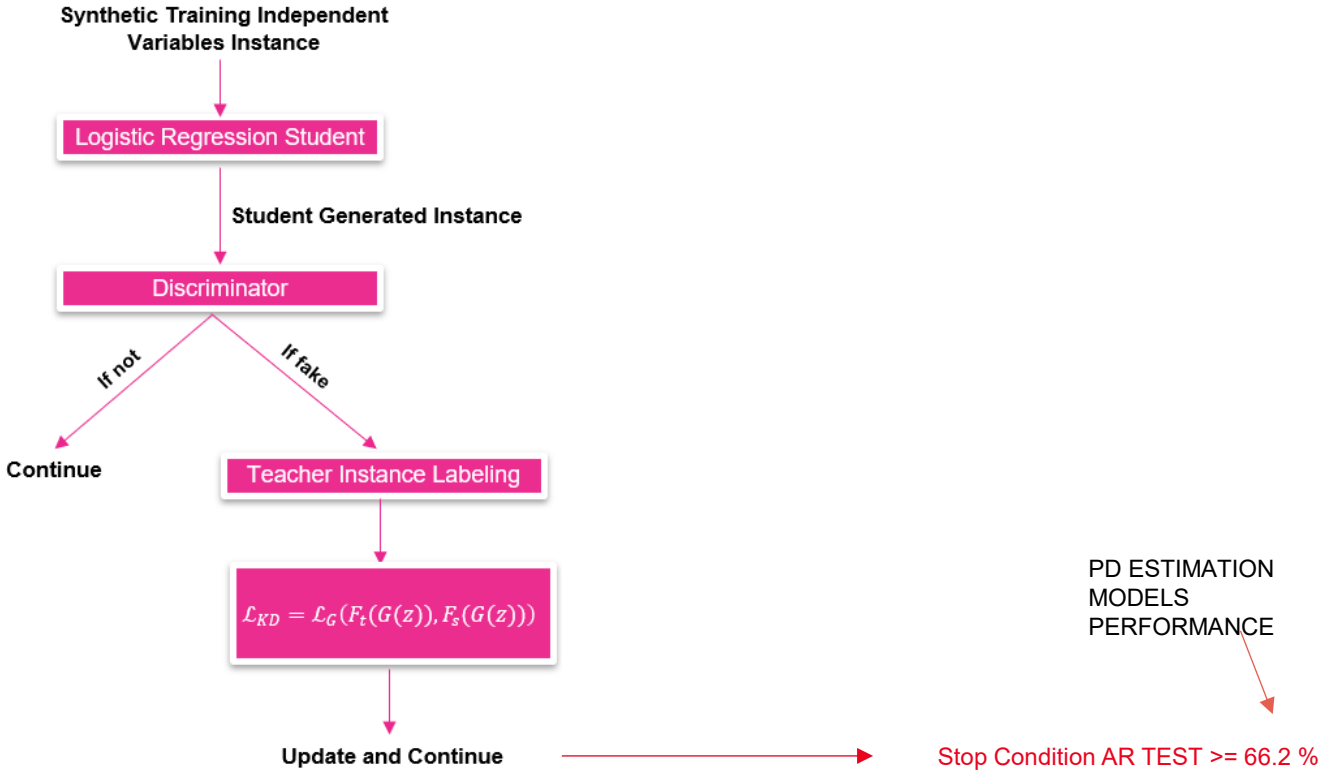


Figure 12. Adversarial Knowledge Distillation Training Framework Performed on *PD ESTIMATION MODELS*. In this case, the generator is also the student which is a logistic regression. The teacher is the LightGBM used in the Hinton-based Framework.

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (3/10)

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.

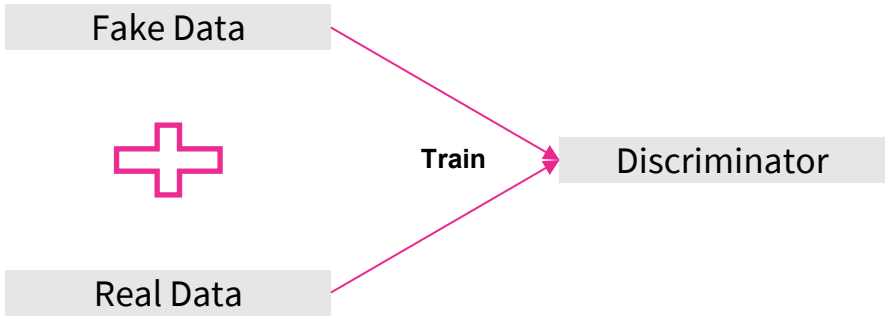


Figure 13. Discriminator Training Flow

How to Construct Fake Data ?

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (4/10)

Discriminator Training Framework - Fake Data Construction

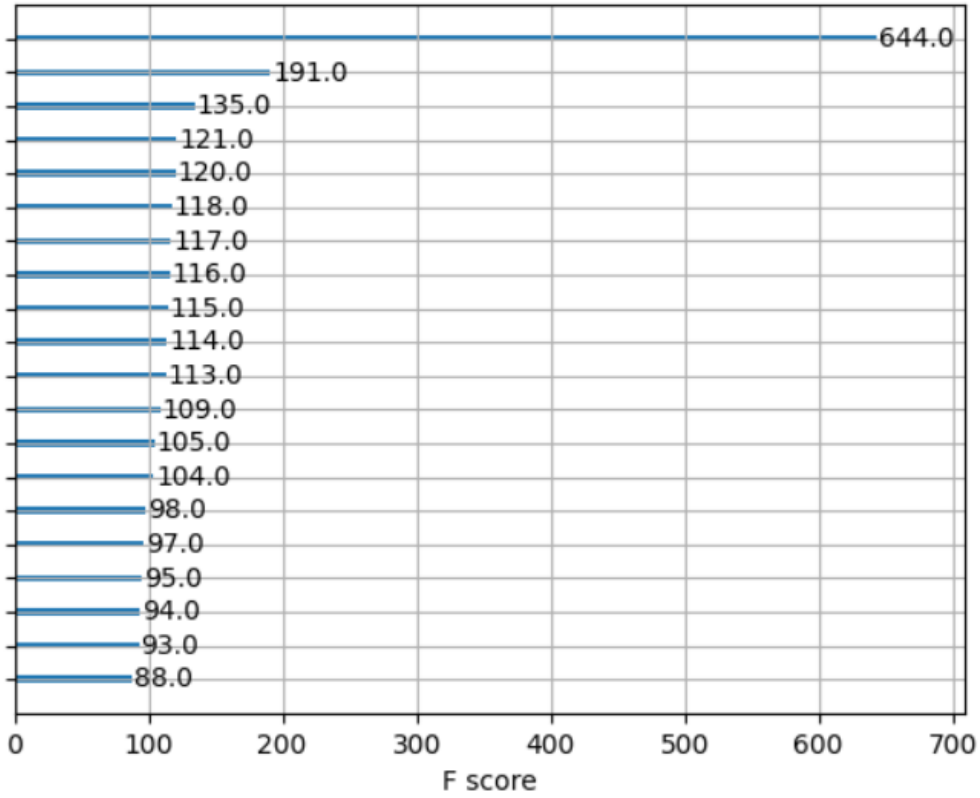


Figure 14. Preliminary discriminator (LightGBM) training. The continuous variable has almost 6 times more importance comparing to other features.

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 (5/10)

Discriminator Training Framework - Constructing Fake Instances of the Continuous Feature

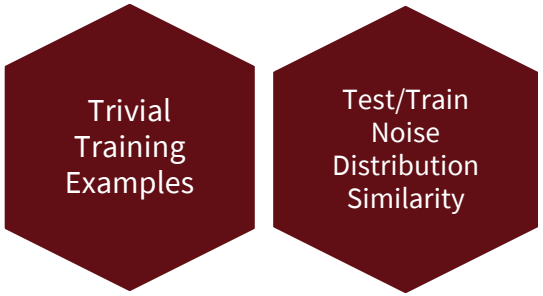


Figure 15. Discriminator Training Main Challenges

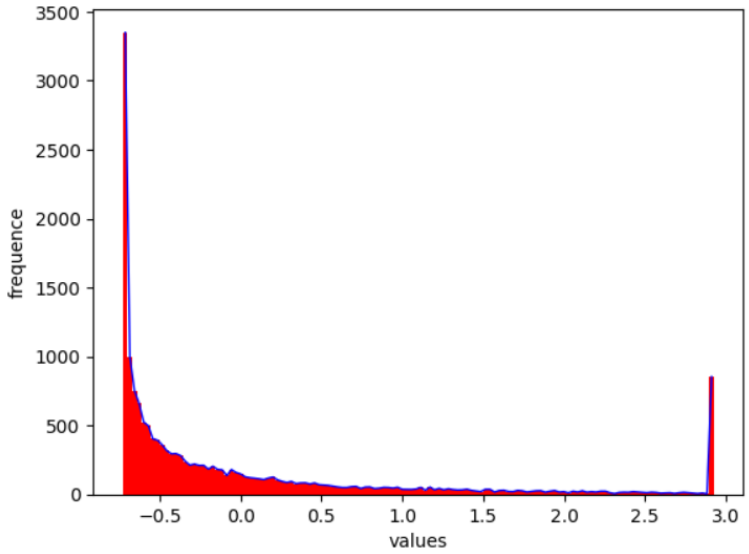


Figure 16. Continuous feature distribution in the training data

- **Trivial training examples**: Since the continuous variable has a very consequent weight on discriminator's decision, constructing fake example of it following regular distributions can be a trivial task for the discriminator during test. It is easy for the model to discriminate fake and real data during test since he overfit on fake data distribution. This led to biased performance metrics in test set.
- **Test/Train Noise Distribution Similarity**: Noise distribution used during training must be different (not strictly) from noise distribution during test especially for the continuous variable to avoid biased metrics.

Solution →

- **Construction of fake continuous variable from multiple distributions** (either estimate the real distribution of data via kernel density method to make it hard for the discriminator or use a mix of multiple random distribution)
- During test phase, we will **not use the same fake continuous variable distribution**.

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (6/10)

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 (7/10)

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 construct 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

1. Full demonstration is provided in Appendix

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (8/10)

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

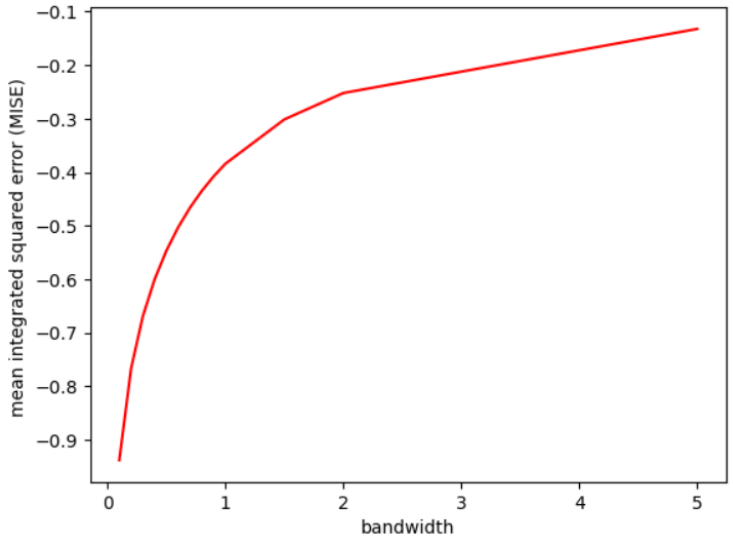


Figure 17. Cross validation to select the best bandwidth h

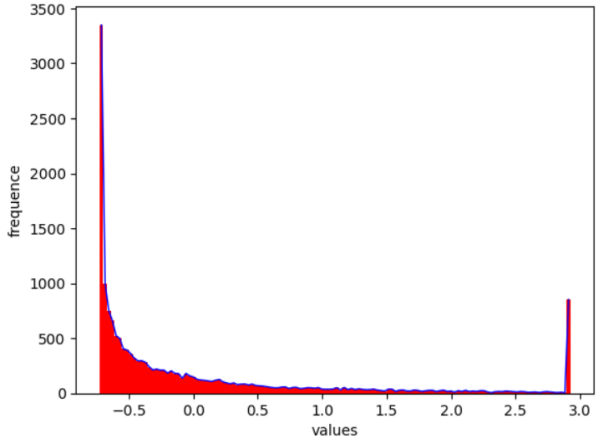


Figure 19. Continuous variable distribution in the training data

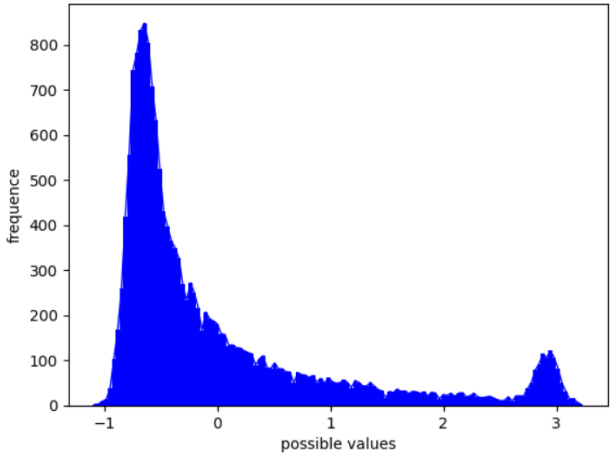


Figure 18. Continuous variable's distribution using Kernel Density Approximation with h = 0.1

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (9/10)

Discriminator Training

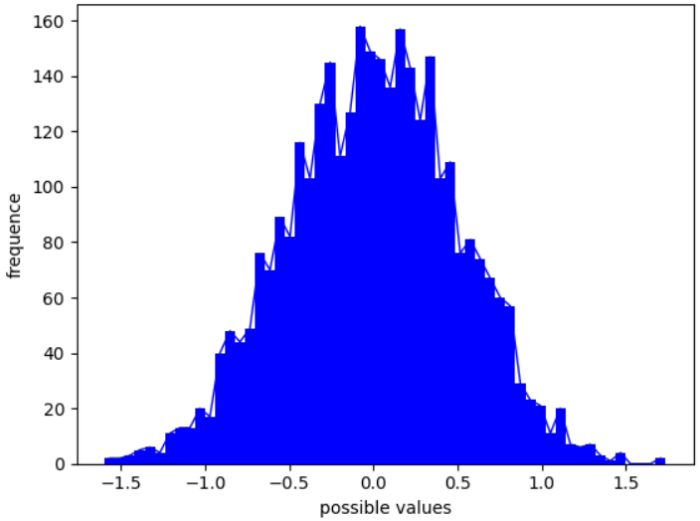


Figure 20. Noise Distribution of Continuous Variable During Test

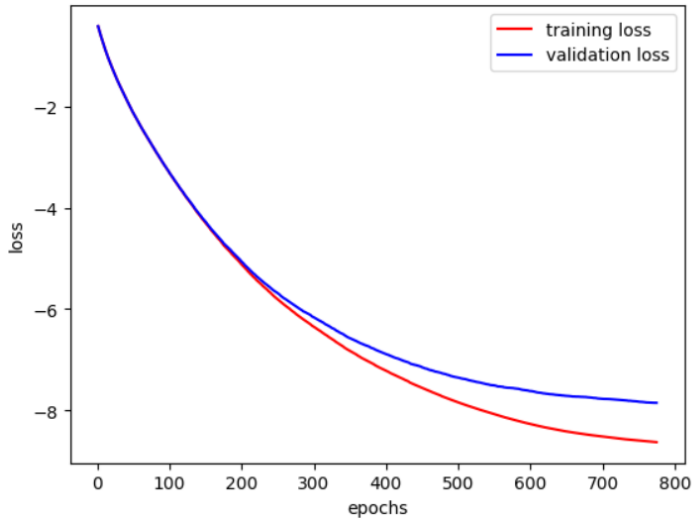


Figure 21. Discriminator (LightGBM) Learning Curve

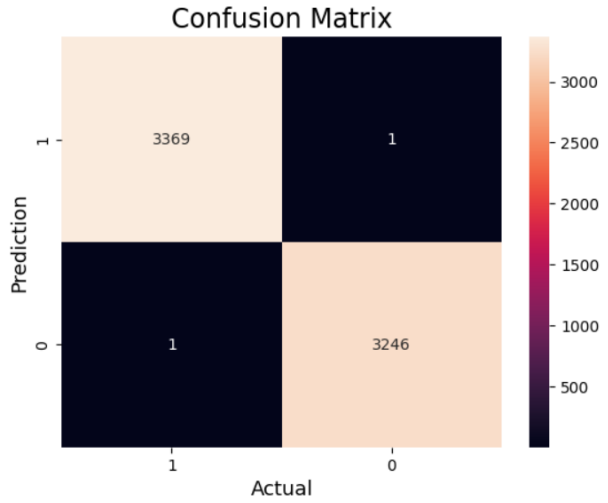


Figure 22. Confusion Matrix of the LightGBM Discriminator

ADVERSARIAL KNOWLEDGE DISTILLATION FRAMEWORK (10/10)

Framework Results

Models	Role	AR TRAIN	AR TEST	AR OOT
PD ESTIMATION MODELS	Baseline	65,4%	66,2%	66,4%
LightGBM with regularization	Teacher	70.87 %	67.41 %	71.44 %
Logistic Regression trained with **Hinton-Based distillation**	Student	68.63 %	64.51 %	70.83 %
Logistic Regression trained with **Hinton-Based + Adversarial Knowledge Distillation**	Student	**68.99%**	**66.2%**	**72.14%**

Table 6. Students' Performance using Adversarial Distillation Framework

XDISTILLATION FRAMEWORK (1/7)

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

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.

Fig 23. Inconsistency between teacher and student explanation

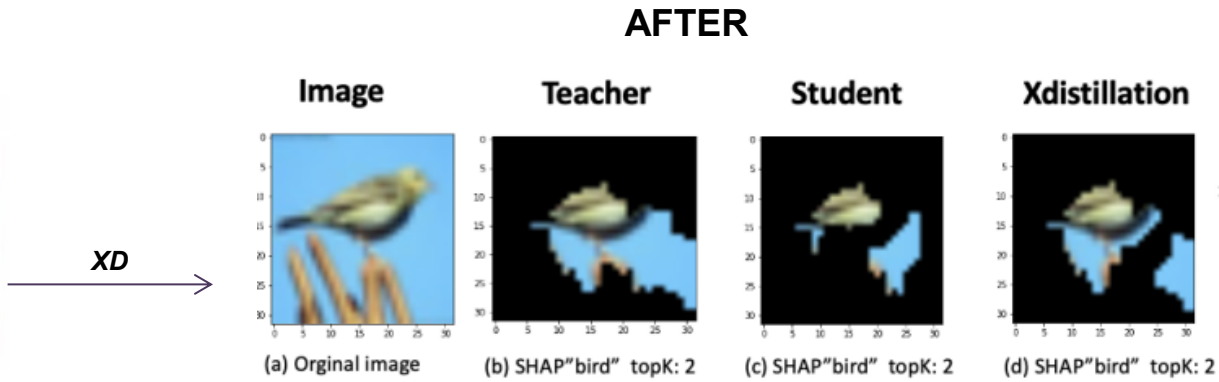


Fig 24. The overlapping explanation area of teacher, KD and XD.

** More details about XDistillation framework is provided in Appendix*

XDISTILLATION FRAMEWORK (2/7)

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

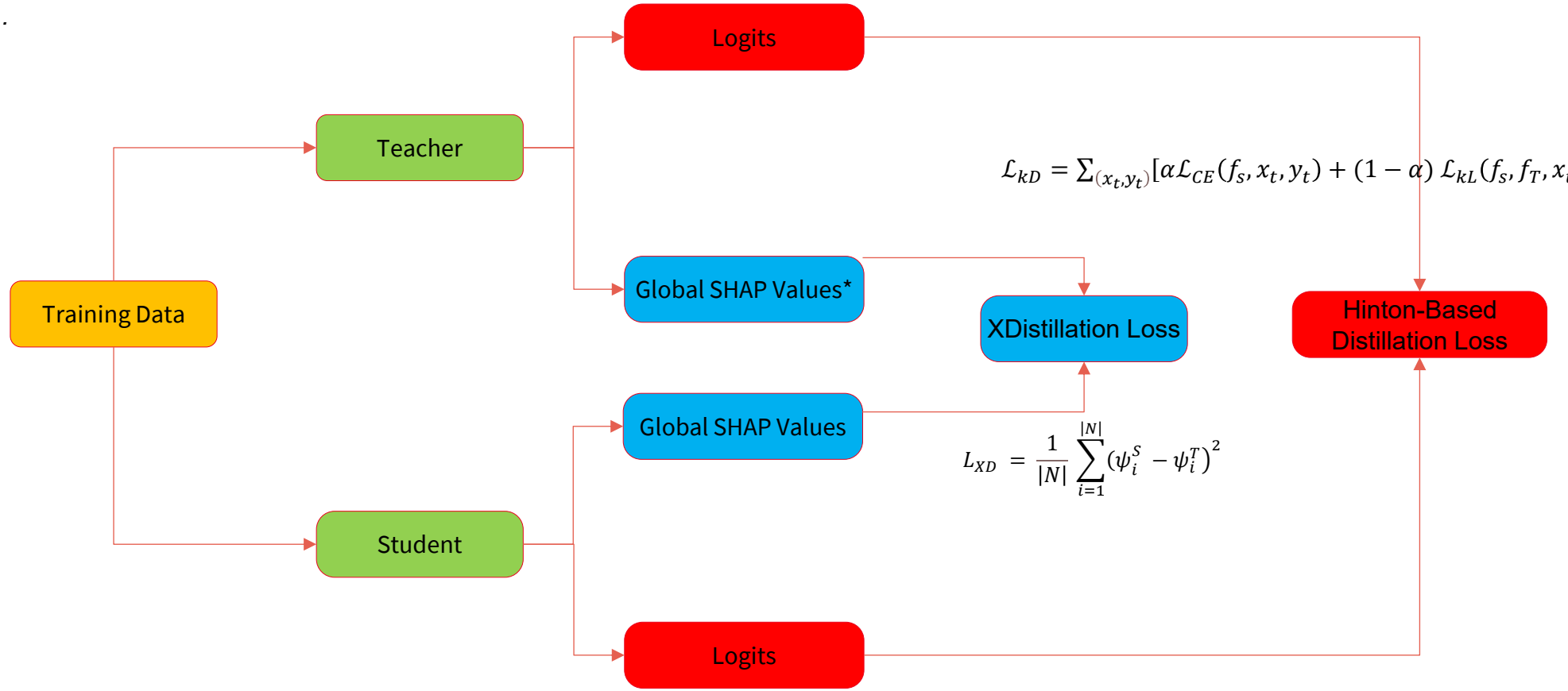


Fig 25. XDistillation Framework Used to Enhance performance

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

XDISTILLATION FRAMEWORK (3/7)

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

In our case, the expression of Xdistillation 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.*

XDISTILLATION FRAMEWORK (4/7)

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;

XDISTILLATION FRAMEWORK (4/7)

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 :

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

Let's denote:

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

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.

XDISTILLATION FRAMEWORK (4/7)

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

XDISTILLATION FRAMEWORK (5/7)

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

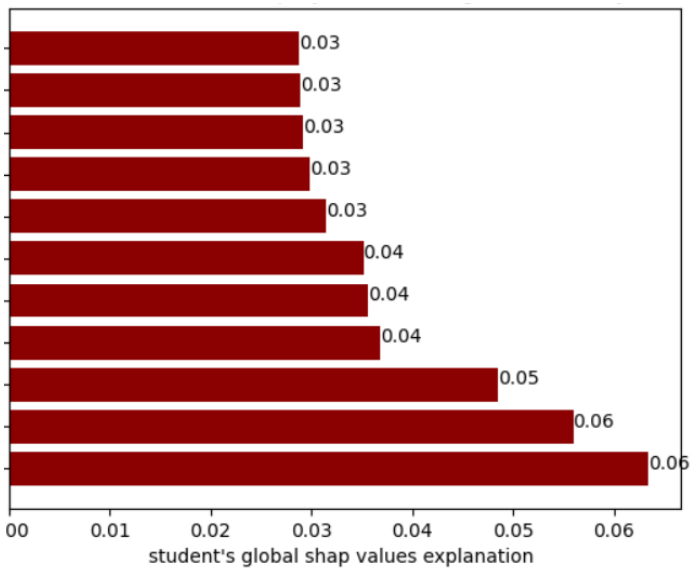


Figure 26. Student's global Shapley values using SHAP library in Python

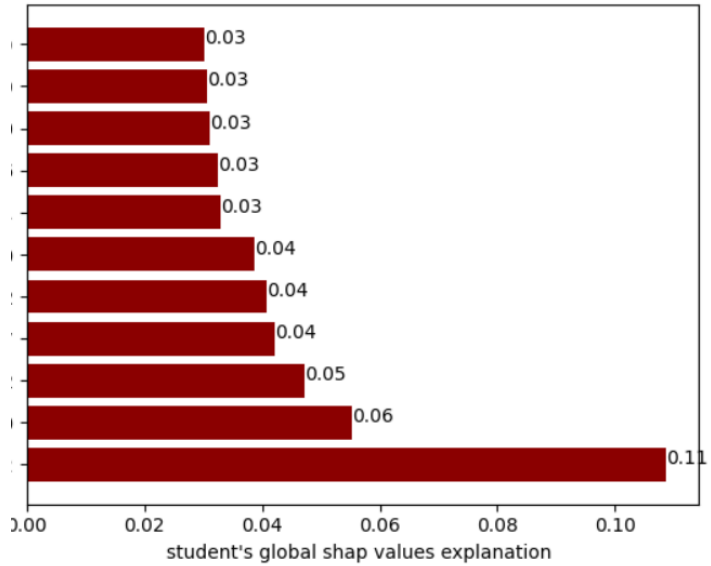


Figure 27. Student's global Shapley values using kernel explainer approximation

XDISTILLATION FRAMEWORK (6/7)

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 :

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

XDISTILLATION FRAMEWORK (7/7)

Xdistillation Framework Results

Models	Role	AR TRAIN	AR TEST	AR OOT
PD ESTIMATION MODELS	Baseline	65,4%	66,2%	66,4%
LightGBM	Teacher	70.87 %	67.41 %	71.44 %
Logistic Regression trained with **Hinton-Based distillation** with temperature	Student	68.63 %	64.51 %	70.83 %
Logistic Regression trained with **Hinton-Based** with temperature **+ Xdistillation**	Student	**70.09%**	**66.01%**	**72.78%**

Table 7. XDistillation framework performance comparing with Hinton-Based distillation with temperature

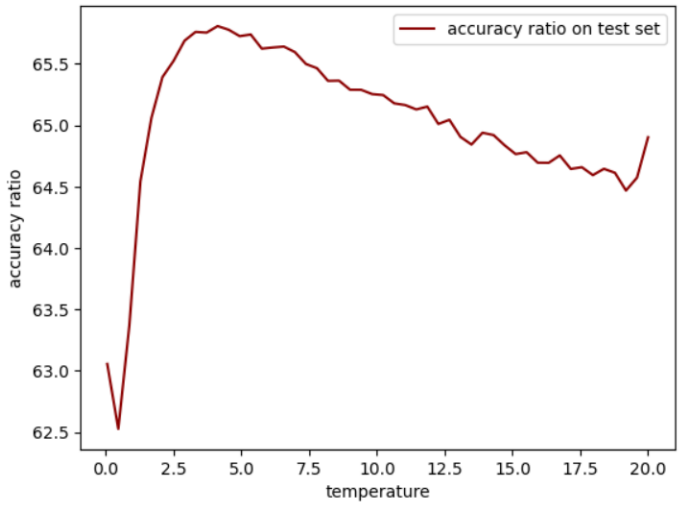


Figure 28 . Cross validation for best temperature T selection in Xdistillation. Here T = 4.12

PD ESTIMATION MODELS' DISTILLATION TAKEAWAY

WRAP-UP

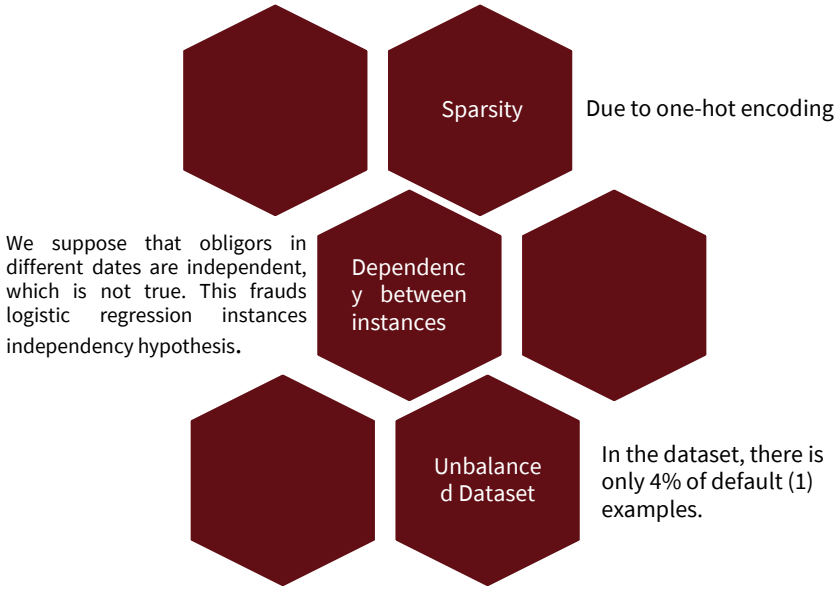
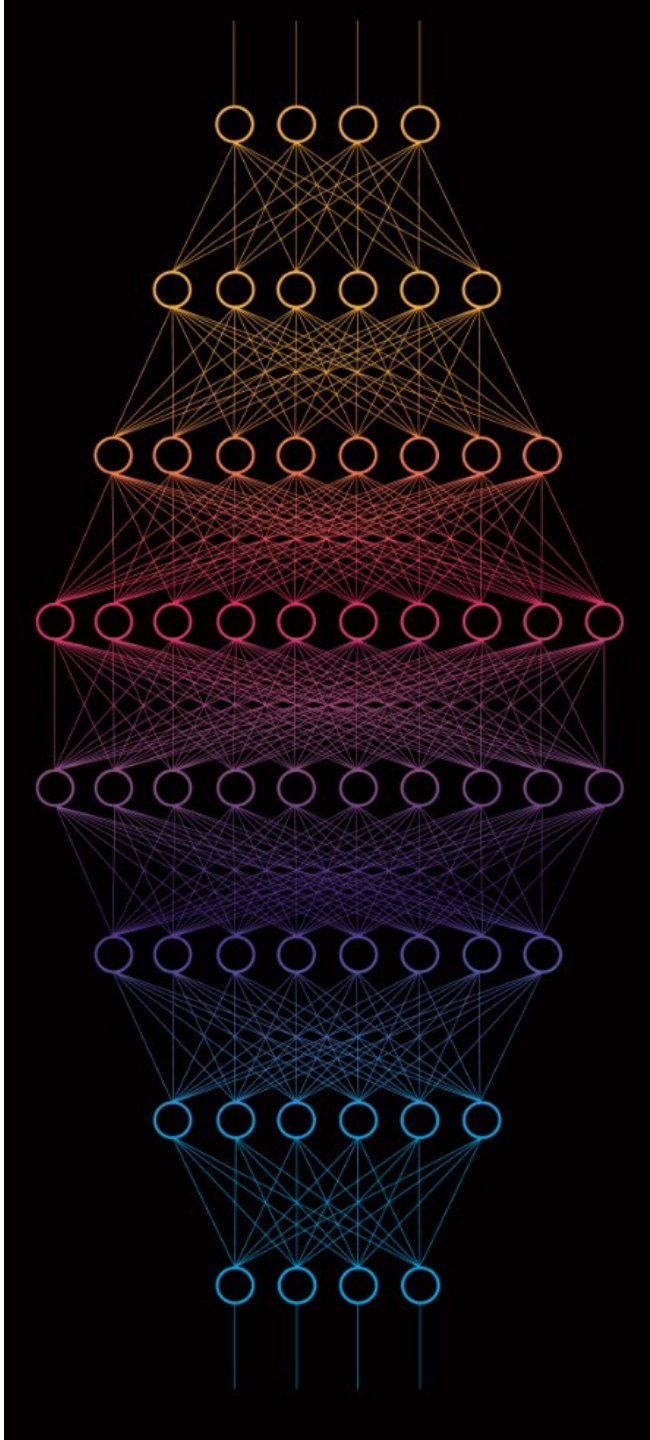


Figure 29. Main problems in training data. This explains also why achieving better performance on the test set was challenging.

3. DISTILLATION ON LENDING CLUB DATASET

A. Teacher Training – Feed-Forward Neural Networks
B. Teacher Training – XGBOOST
C. Student Distillation Training – Feed-Forward Neural Network

SOCIETE GENERALE



DISTILLATION TESTS ON LENDING CLUB DATASET (1/3)

Teacher Training – Feed-Forward Neural Networks

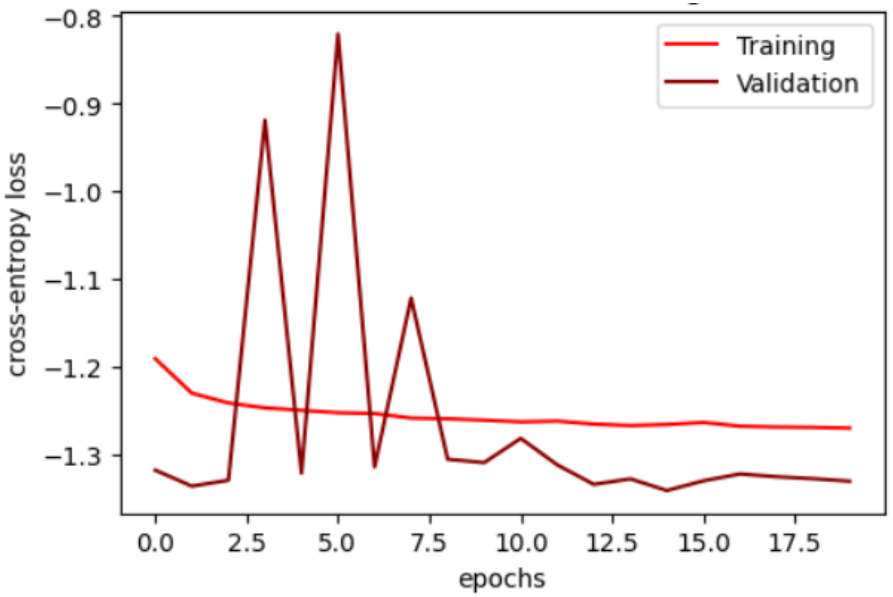


Figure 31. Neural Network Teacher's Learning Curve

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

Table 8. Teacher Performance

DISTILLATION TESTS ON LENDING CLUB DATASET (2/3)

Teacher Training – XGBOOST

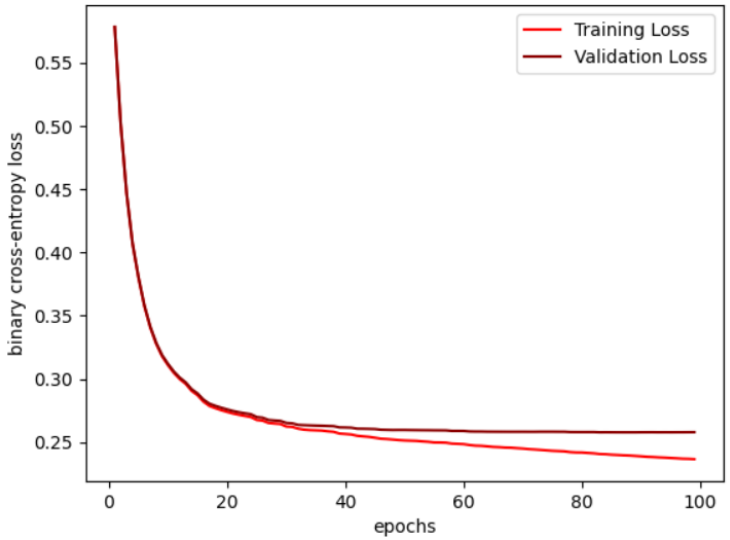


Figure 32. XGBOOST Teacher's Learning Curve

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

Table 9. XGBOOST Teacher Performance

DISTILLATION TESTS ON LENDING CLUB DATASET (3/3)

Student Training – Feed-Forward Neural Network

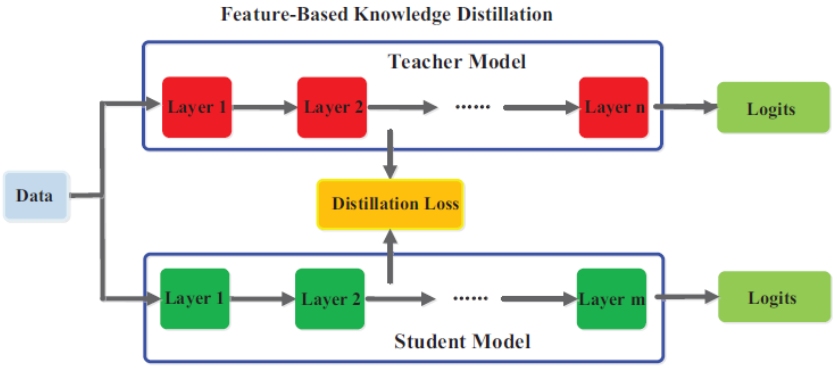


Figure 33. The generic feature-based knowledge distillation

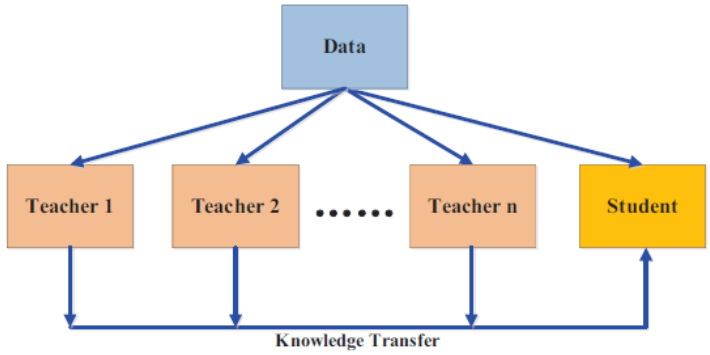


Figure 34. Multi-teacher Distillation Framework's Illustration

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 11. Student's Performance using different Distillation Frameworks

Compression Ratio = 51.41

Model	#parameters
Teacher	58,357
Student	1,135

Table 10. Student's performance

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 (1/6)

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.

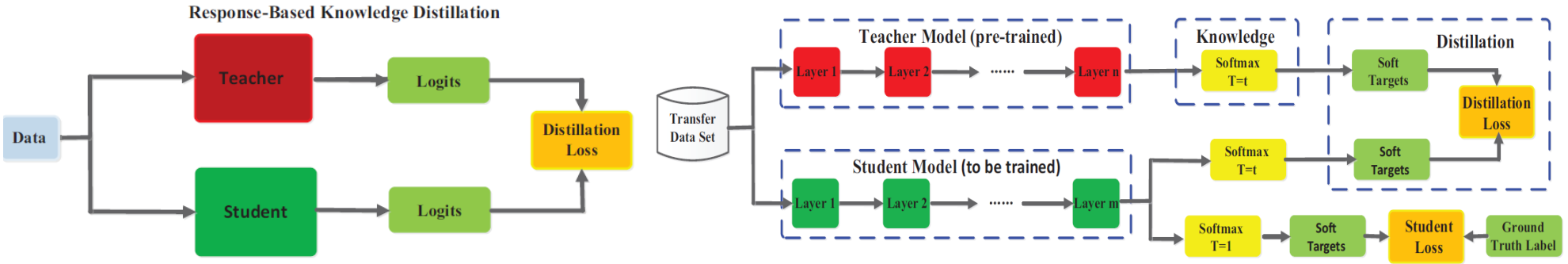


Fig 4. The specific architecture of the benchmark knowledge distillation. The student model can learn to mimic teacher's predictions and also ground truth labels.

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 2. 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.

APPENDIX (2/6)

Feature-Based Knowledge Distillation

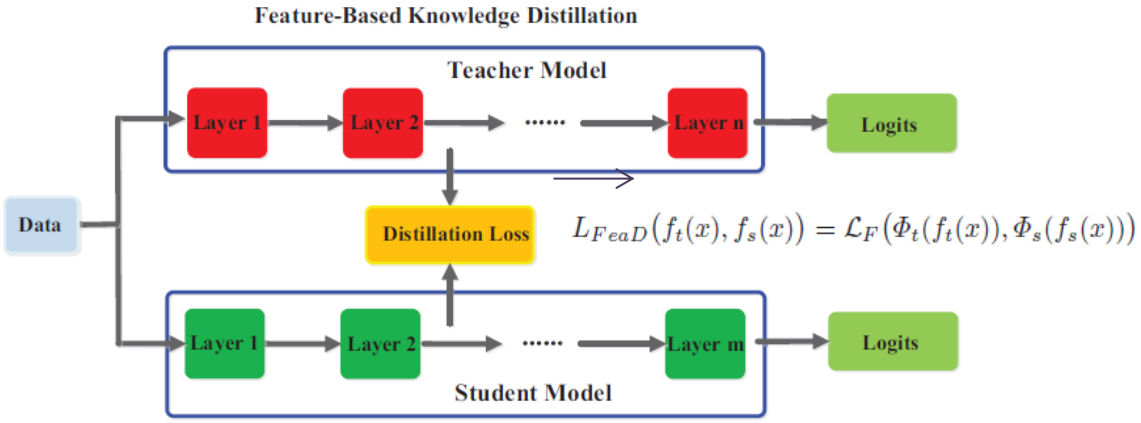


Fig 5. The generic feature-based knowledge distillation.

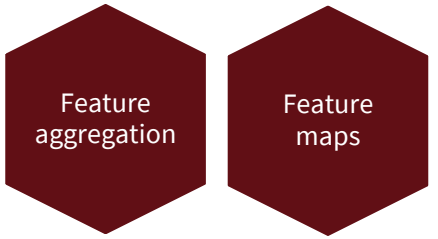


Fig 6. Some types of feature-based knowledge

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.

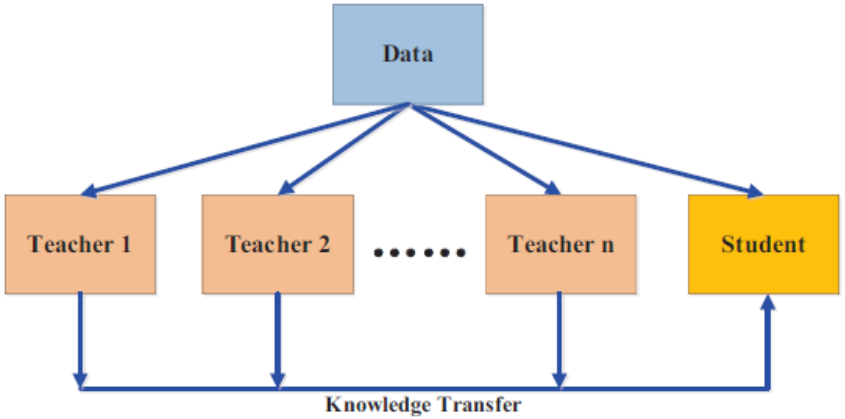
Table 3. Feature-based distillation investigation

.

APPENDIX (3/6)

SOME DISTILLATION FRAMEWORKS

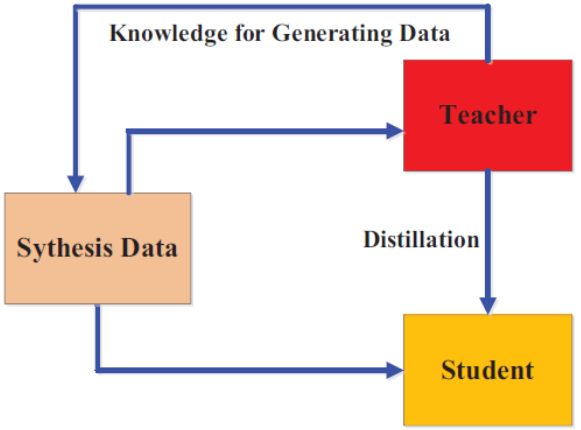
1. Multi-teacher Distillation



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.

2. Data-Free Distillation



Problem	Usage example	Pros
1) Unavailable data arising from privacy, legality, security and confidentiality	Data is generated from the feature representations from the pre-trained teacher model and used to train the student model in addition to traditional distillation.	Powerful, Uses GAN Straightforward

Table 7. Data-Free Distillation Framework's detailed explanation.

APPENDIX (4/6)

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)

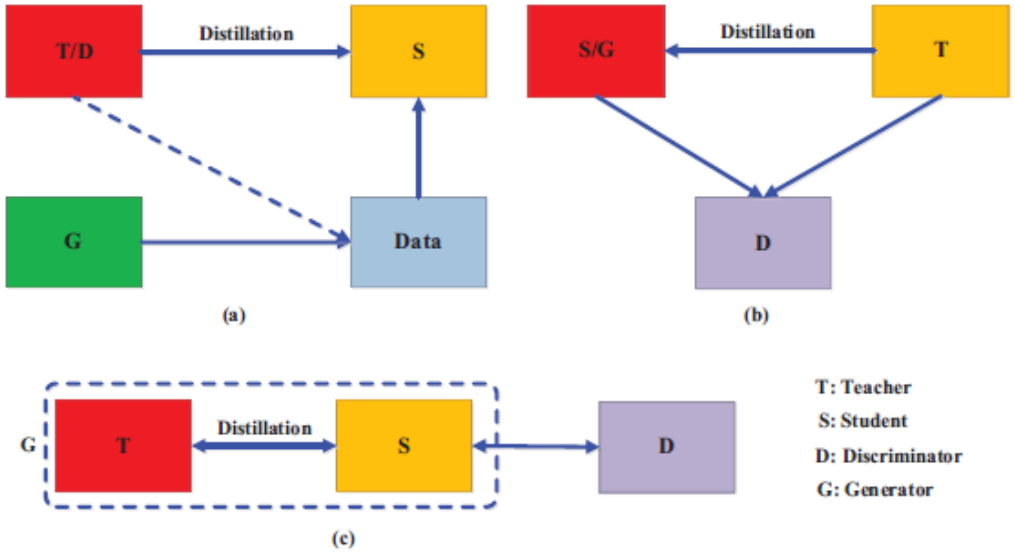


Fig 10. The different categories of the main adversarial distillation methods.

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 8. Adversarial Knowledge Distillation Framework's detailed explanation.

APPENDIX (5/6)

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 (6/6)

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

C'EST VOUS
L'AVENIR

SOCIETE
GENERALE