BANCO DE ESPAÑA
Eurosistema

PUBLIC USE

ARTIFICIAL INTELLIGENCE AND CREDIT RISK: ACCURACY VS. INTERPRETABILITY

José Manuel Carbó (*)

División de Innovación Financiera, Banco de España

Quinta Conferencia Estadística: **"Information to unlock the future"**

Banco de Chile. 26 & 27 September 2023

(*) The opinions and analyses expressed in this paper are the responsibility of the authors and, therefore, do not necessarily match with those of the Banco de España or the Eurosystem.

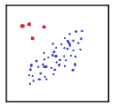
DGA PAYMENTS AND MARKET INFRASTRUCTURE

- The **use of Artificial Intelligence (AI)** is gaining ground in finance.
  - Gains in AI adoption fueled by emerging technologies like **cloud computing and Big Data**.
  - Transitioning from traditional approaches such as linear regressions and dictionary methods to advanced algorithms like deep neural networks and Large Language Models (LLMs).

- Some possible examples at central banks

  - **Natural Language Processing:**
    - Analysis of news for real time economic sentiment tracking
    - Automated extraction and understanding of terms in climate reports, loan guarantees

  - **Image recognition**: Design/production of banknotes, signature verification



  - **Outliers**: Identifying irregularities in both transactions and databases for preventive action

  - **Prediction**: Corporates bankruptcies, credit default

- Zooming In: **Artificial Intelligence and Credit Risk**
  - **Central to Financial Stability**: Accurate credit risk assessment is critical not only for individual financial institutions but also for the overarching stability of the financial system.

- By focusing on AI and credit scoring, we can illustrate the **opportunities and risks of AI in finance**

➢ Consensus on Predictive Ability: Superior **predictive accuracy** in assessing credit risk.

➢ Complexity and Validation Challenges:

  ➢ **Interpretability**: It creates hurdles in understanding the rationale behind predictions

  ➢ **Biases**: May perpetuate biases, raising ethical and regulatory issues.

Opportunities

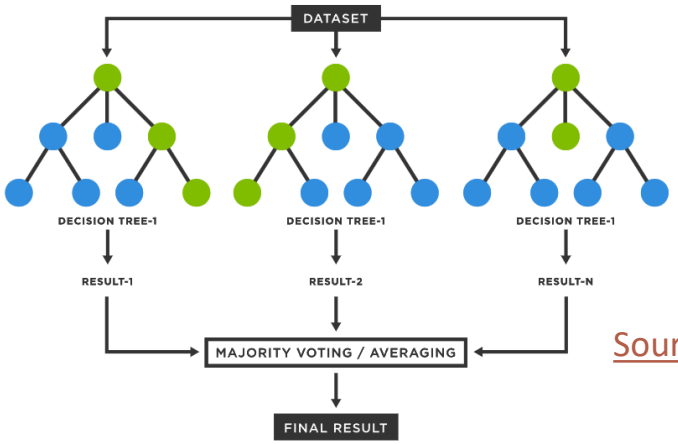
- **Better predictive ability**
  - Gains up to 20% in different statistical metrics
  - Tree base models (like random forest and XGBoost) seem to be best

- **Economic impact**
  - Better prediction of default could decrease loss from default up to 10%
  - Better prediction of default could yield regulatory capital savings up to 20%

- **Increased financial inclusion**
  - Use of alternative data sources

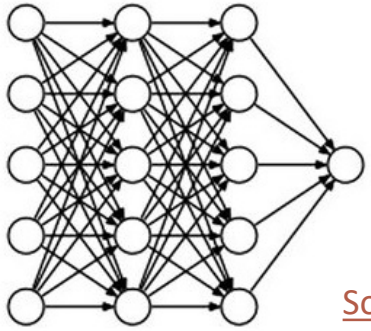
- $ProbDefault_{i,t} = \beta_1 Income_{i,t} + \beta_2 Age_{i,t} + \beta_3 DuePast_{i,t} + \beta_4 DebtRatio_{i,t} + \beta_5 HomeOwner_{i,t} + e_{i,t}$
  - **Logit** only learns linear decision boundaries. Feature interactions must to be manually added

Random forest (RF)



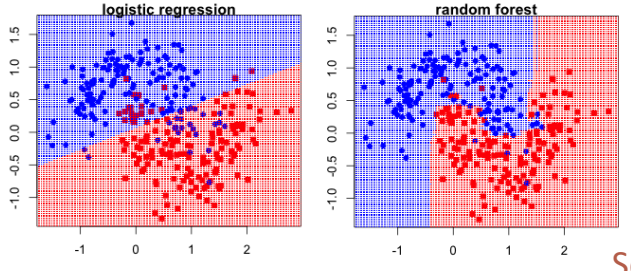
Source

Deep Learning (DL)



Source

- **RF and DL can model more complex relationships, finding interactions between variables**



Source

- Credit default is influenced by a multitude of factors that interact in complex ways

Predictive Ability

Author, year, journal	Data	Sample size	Prediction ML	Prediction Logit
Jones et al (2015)	Corporate loans	5.000 firms	Random forest 93%	Logit 83%.
Petropoulos et al (2019)	Corporate loans	200.000 firms	Gradient boosting 78%	Logit 66%
Sigrist y Hirnschall	Corporate loans	1.000 firms	Gradient boosting 83%	Logit 66%
Kvamme et al (2018)	Mortgages	20.000 mortgages	Conv neural net 91.5%	Logit 86.6%
Sirigniano et al (2019)	Mortgages	120 mill mortgages	Neural net 79%	Logit 59%
Moscatelli (2019)	Consumer loans	300.000 firms	Random forest 75.9%	Logit 73.2
Butaru et al (2015)	Consumer loans	50 million loans	Random forest 66.6%	Logit 59.2%
Albanesi (2019)	**Consumer loans**	**1 million loans**	**Neural net 90%**	**Logit 86%**
Alonso and Carbo (2022)	**Consumer loans**	**80.000 loans**	**XGBoost 85%**	**Logit 78%**

- ✓ Predictive gains up to 20%
- ✓ Huge heterogeneity

Economic impact

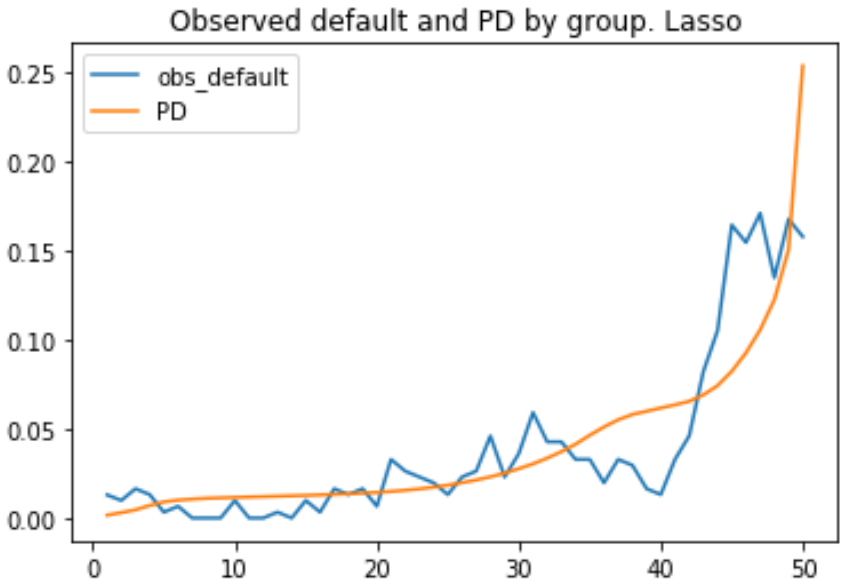
- What does it mean to have a 20% more AUC, or higher true positive rate?

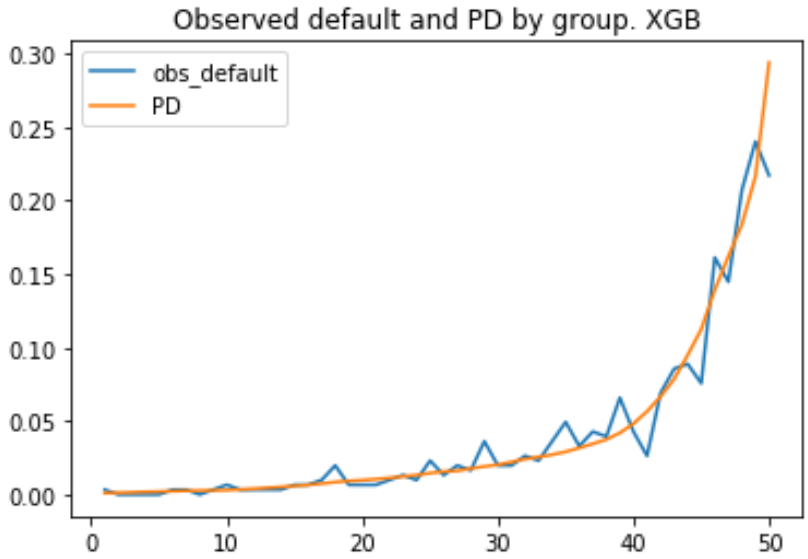
- **Cost Savings in Credit Losses** due to defaults using credit card data **:**
  - Khandani et al. (2010) use random forests, and report cost savings ranging from 6% to 25%
  - Albanesi & Vamossy (2019) use neural networks, savings of up to 9%.

- **Regulatory Capital Savings:**
  - Measurable capital savings can be achieved by using ML over traditional econometric models in a real credit portfolio.
  - Fraisse & Laporte (2022): regulatory capital savings of up to 25% using corporate bonds.
  - Alonso-Robisco & Carbó (2022b) regulatory capital savings of up to 17%
    - We used XGBoost compared to Logit in a Spanish consumer credit database.

Economic impact

- **Regulatory Capital Savings** (Alonso and Carbo 2022b)
  - Concavity in the regulatory capital formula is crucial.
  - Improved PD (Probability of Default) prediction leads to better risk classification.
  - Logit tends to overestimate at low PDs, leading to overestimation of regulatory capital needs.



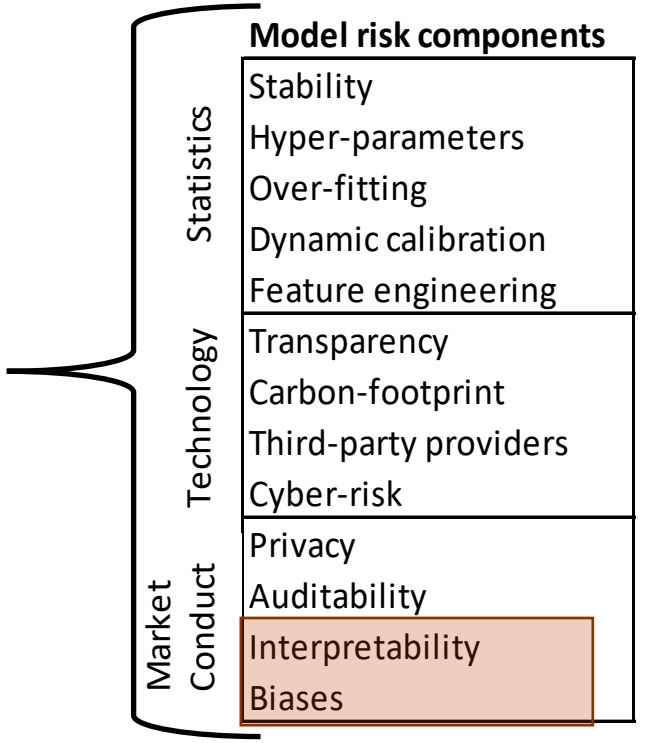


RISKS

- Several reports on the risks of applying ML in credit (Bafin 2022, Dupont et al. 2020).

- In Alonso y Carbó 2022a, we group risks into 3 categories:
  - ✓ Many of these risks apply to traditional econometrics

- Due to its novelty and its importance, in this presentation we are going to talk about two
  - ➢ **Interpretability**
  - ➢ **Biases**



Source: Alonso-Robisco & Carbó (2022a)

Interpretability in Credit Risk Assessment

- **Right to Explanation Debate**: Ongoing discussions about consumers' right to receive explanations for credit decisions, whether it's a denial or unfavorable terms.

- **Legal Landscape**: No legal mandate to "open the black boxes" of AI in credit decision-making.
  - GDPR Article 22: Requires human judgment in automated decision-making and profiling.

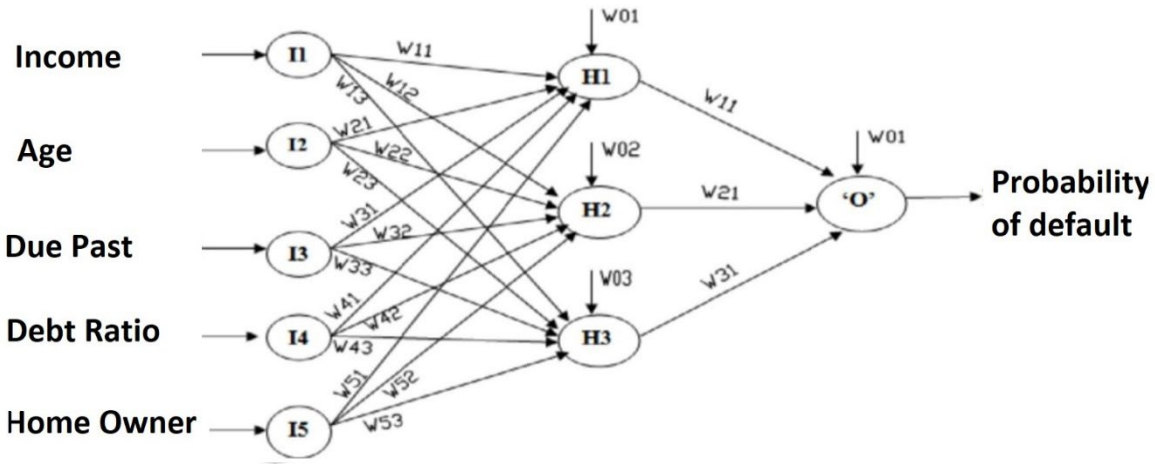
- **High-Stakes Impact, and therefore, regulatory scrutiny**
  - Credit decisions affect individuals' lives, from home ownership to entrepreneurship.
  - EU AI Act labels AI creditworthiness assessment as "high-risk," requiring conformity tests.
  - USA Legislation (ECOA & FHA): Anti-discrimination laws covering disparate treatment and disparate impact. Definition of "Adverse Action Notice"

- **Ethical and Societal Implications:**
  - Decisions shouldn't be based on special categories of personal data

- Let's interpret a logit model

- $ProbDefault_{i,t} = \beta_1 Income_{i,t} + \beta_2 Age_{i,t} + \beta_3 DuePast_{i,t} + \beta_4 DebtRatio_{i,t} + \beta_5 HomeOwner_{i,t} + error_{i,t}$
  - **β_1 captures the effect of income**

- Let's interpret a deep learning model



Source

- In this simple neural network, **there are already 6 weights for income**.
- In a proper deep learning model with three hidden layers and 100 nodes in each one, **there would be 22.220 weights associated with income**

Challenges and Solutions in ML Interpretability for Credit Risk

- **Inherent Interpretability Gap:**
  - Traditional econometric models like Logit are inherently explainable, whereas interpreting machine learning models is complex.

- **Explainable AI (xAI):**
  - Emerging field aimed at enhancing transparency and trust in ML models.
  - Explanations could be local (for a particular loan) or global (for the whole dataset)

- **Post-Hoc Interpretability techniques**
  - Can be applied to any model after it has been trained
  - Shapley Additive Explanations **(SHAP):**
  - Local Interpretable Model-agnostic Explanations **(LIME)**
  - **Permutation Feature Importance**

Permutation Feature Importance

- Basic Idea: Evaluate the model's performance with and without a feature to measure that feature's importance.

1. **Data Shuffling**: For each feature, its values are randomly shuffled (permuted) across all data points, destroying any correlation with the target variable.

2. **Performance Drop**: The model's performance (e.g., accuracy, F1 score, etc.) is measured before and after the permutation.

3. **Importance Metric**: The drop in performance indicates how important the feature is; bigger drops signify greater importance.

4. **Repeat and Average**: The process is often repeated multiple times and averaged to get a more robust measure.

Income	Age	HomeOwner	Default
12.394	35	0	1
93.224	61	1	0
32.210	43	0	0
...	...	...	...
23.400	53	1	0
41.424	61	1	1
43.352	44	1	0

SHAP

- **SHAP** is a technique that measures the contribution of a variable to the predicted outcome, on a given day compared to the average prediction. These contributions are called the Shapley values

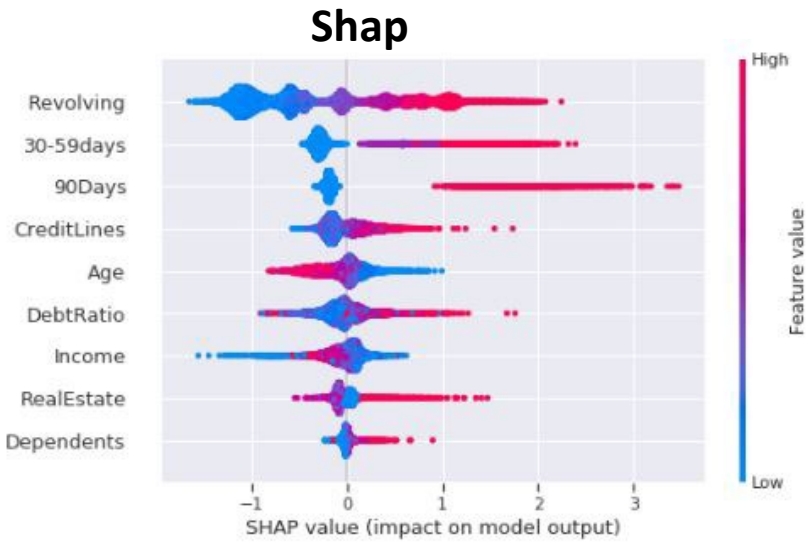
- Suppose we have as variables“*Income*," “*Age*," and “*HomeOwner*”**, and that we want to know the importance of the “Income" in the probability of default of individual i.**
- These are the four possible coalitions of variables without ““*Income*”:
  - No variables
  - *Age*
  - *HomeOwner*
  - *Age* and *HomeOwner*

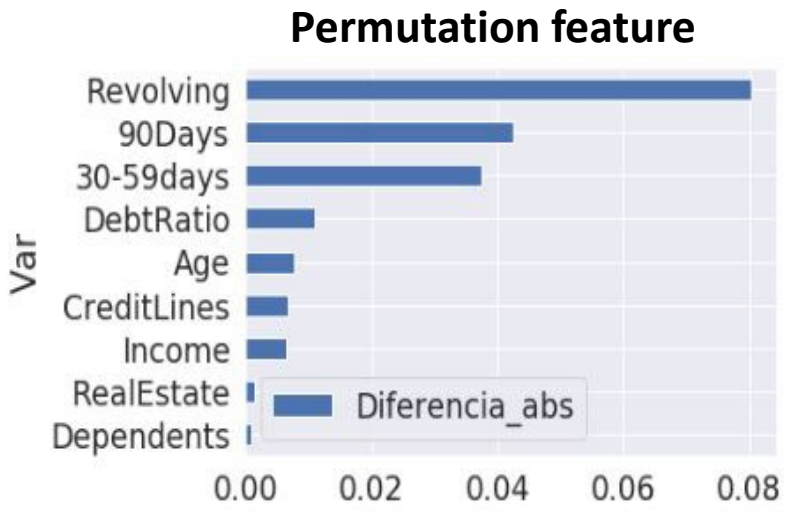
- For the four coalitions, we calculate the price in t with and without “*Income*”.
- The Shapley value of ““*Income*” for the probability of default of individual I is the weighted average of those marginal contributions.
- To obtain the global importance of ““*Income*” in the test sample, we repeat the process for all days and take the mean of the Shapley values.

Formulation

An example with data: Credit database, *Give me some credit* (Kaggle.com)

- 100.000 loans. Binary target: 6% of *default*.
- 11 variables: Income, age, deb ratio, etc.
- ***XGBoost AUC: 0.84. Logit AUC: 0.78***
- ***Let's interpret XGBoost predictions!***



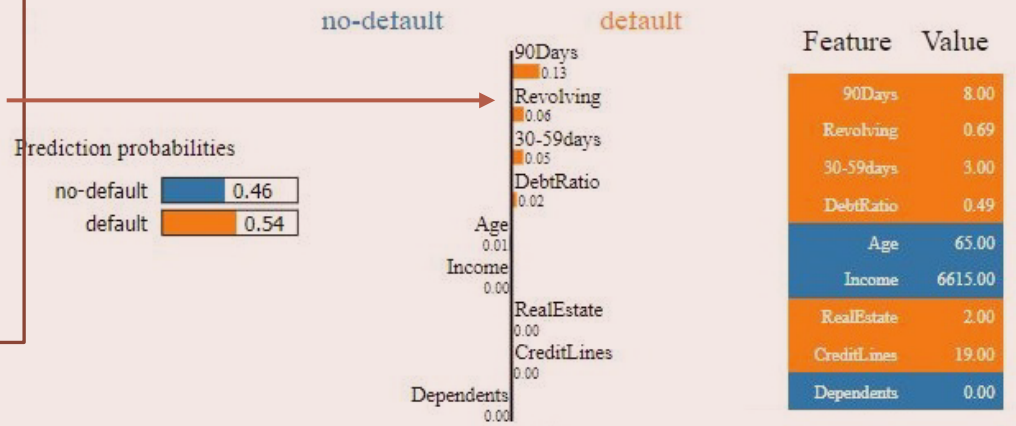


- Not only we have to trust the post-hoc interpretability techniques, but also there exists discrepancies between explanations!

Local discrepancies

Shap XGBoost (local)

Again, there are discrepancies between the explanations (between Shap and LIME), especially about the role of revolving.



LIME XGBoost (local)

- Although there are techniques that interpret machine learning, **many questions arise**:
  - **How to define the discrepancy** between interpretability techniques? (Krishna et al 2022)
  - **How worrisome is the discrepancy?**
  - **There exist also discrepancies within models!**
  - ¿Is it enough as an explanation? (Miller, 2019)
  - **Literature still incipient, few articles on economics.**

- In Alonso-Robisco & Carbó (2022c) we propose the generation of synthetic data to create a *stress test* for post-hoc interpretability techniques.
  - Our synthetic data represents credit-like scenarios.
  - We assign the importance of the variables for the *default*.
  - We compare the importance given to features by interpretability models with the importance assigned by us.

CONCLUSIONS

- **ML brings great benefits** in credit risk models:
  - **Better prediction**
  - **Better financial inclusion:** Yes, but…
  - **It has unintended effects:** Biases and interpretability.

- **xIA provides techniques to mitigate them**, but it does not solve all problems.
  - The problem of discrepancy between explanations(Krishna et al. 2022).

- **Other mitigating factors:**
  - **Human-in-the-Loop:** human-machine collaboration can contribute to better decision making
  - Alignment between **ethical principles** of finance and ML (Rizinski et al. 2022).
  - **Collaboration between ML and econometrics**. P.ej.: Kaji et al. (2022).
  - **Formation**: *ML methods that economists should know about* (Athey & Imbens 2019).

Appendix

Biases

- Innovations in statistical technology have caused an increased risk of redistributive impacts on protected social classes such as religion, gender, or race.
  - ➢ The impacts can come from the greater flexibility to discover structural relationships or from triangulation of other previously excluded characteristics..

- Fintech lenders process mortgage applications (USA) 20% faster than traditional lenders, and their default rates are 25% lower, controlling for loan characteristics (Fuster et al. 2022):
  - ✓ **Are Fintech more lax or do they have better credit scoring systems with ML?**

Return

Biases

- **Fuster et al. (2022)**: finds that black and Hispanic lenders are disproportionately disadvantaged by the introduction of ML.
- **Bartlett et al. (2022)** finds that although Fintech can reduce discrimination, they don't eliminate it, and observe that black and Hispanic lenders pay a positive interest differential on mortgage loans
- **Dobbie et al. (2021)** finds that ML models guided by long-term objectives can increase the benefit of entities and reduce biases, but those guided by short-term objectives penalize minorities such as the elderly or immigrants.

➢ **Can we extrapolate the results?**

- ✓ How does this discrimination happen? More or less competition.
- ✓ Who asks for a loan in Fintech? It's probably not random.

Return

Financial Inclusion

- The use of financial technology, particularly machine learning, **is democratizing access to financial services.**
  - **Philippon (2019):** Big Data can reduce negative biases in credit allocation but may compromise the protection of specific minority groups.
  - **Barruetabeña (2019):** The new generation of financial services accessible via mobile phones and the Internet is fostering progress and encouraging the entry of Bigtech companies into the financial sector.
  - **Huang et al. (2020):** Big Data and ML provides significant advantages in predicting small and medium enterprise (SME) defaults. Benefits arise from the use of alternative new data (Information Advantage) and new predictive methods (Model Advantage).

Return

SHAP (Formulas)

- The Shapley value or contribution Ø of a given feature *i* in a prediction *p* is :

$$\emptyset_i = \sum_{S \in N/i} \frac{|S|!\,(n - |S| - 1)}{n!} (p(S \cup i) - p(S))$$

  - S represents a coalition of features
  - N is the total number of features,
  - $S \in N/i$ represents all possible coalitions of features excluding feature *i*, considering all possible orders,
  - $p(S \cup i) - p(S)$ represents the difference in the predicted outcome *p* when we consider a particular coalition of features and feature *i* minus the predicted outcome when we consider the coalition of features without feature *i*.

- The term $\frac{|S|!(n-|S|-1)}{n!}$ assigns different weights to the differences, depending on the features that are in the set $|S|!$, the features that have to be added $(n - |S| - 1)$, and all normalized by the features that we have in total.

Return

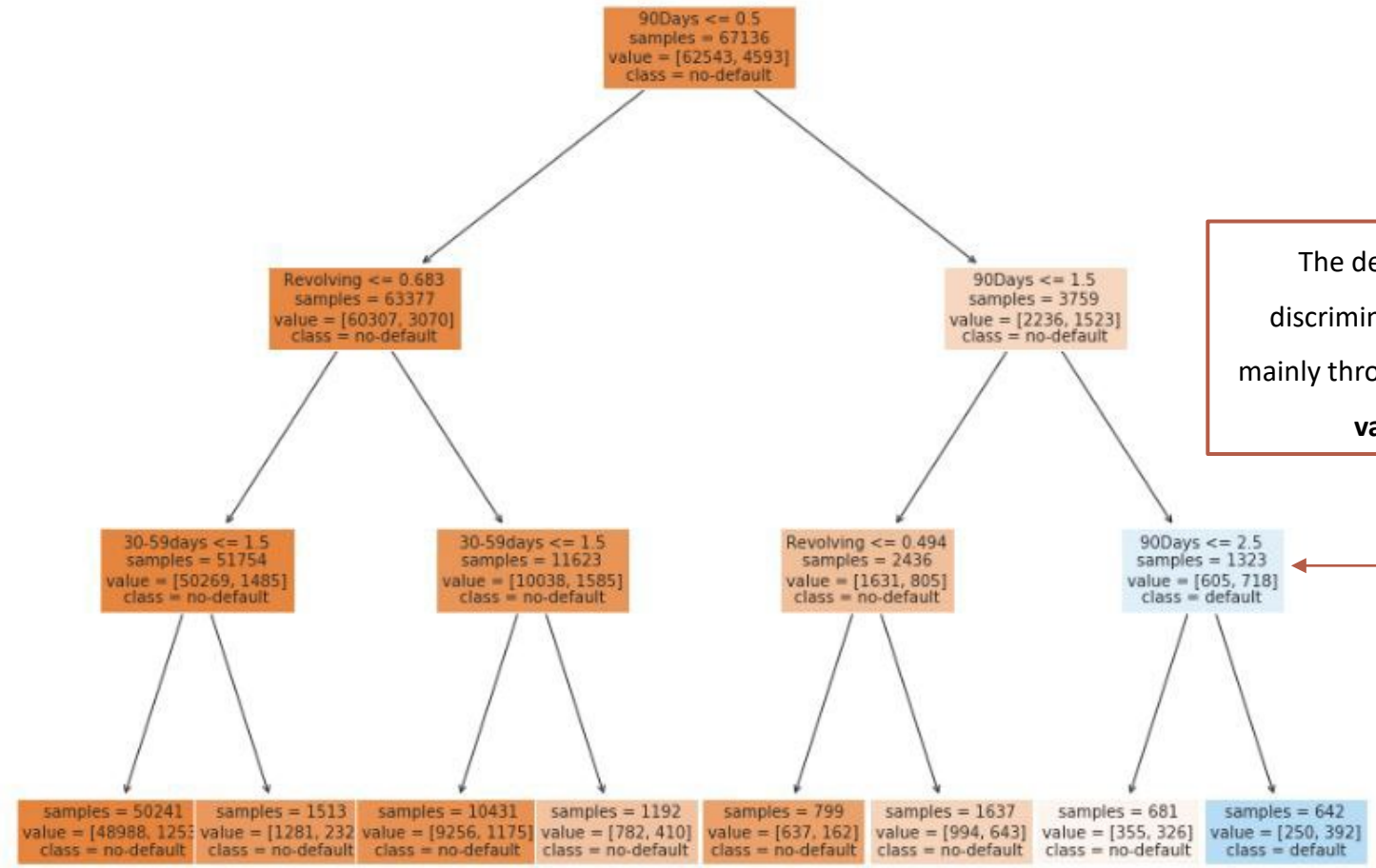
- We measure predictive capacity with different statistical metrics

	Logit	Logit cuadrado	Tree	XGBoost	Deep learning
AUC-ROC	0.64	0.78	0.81	0.84	0.82
TPR / Recall	0.61	0.61	0.71	0.74	0.75
F1	0.17	0.31	0.31	0.32	0.20

- **Logit is underperforming**
  - Therefore we add variables to the square (it will have consequences for interpretability)
- Results in line with the literature: **Complex ML models have profits.**
  - **Although the most complex ones (Deep learning) do not necessarily predict better.**

Return

Another interpretable model



The decision tree discriminates defaults mainly through the **90Days variable**

Return

- **In Alonso-Robisco & Carbó (2022c)** we propose the generation of synthetic data to create a stress test in a controlled environment.

  1. We decide the **number of observations**.
     - Integer between 75.000 and 225.000
  2. We select **the percentage of 0s and 1s** in our Target variable..
     - *Random float* between 3% and 5%
  3. We select the number of explanatory variables and their distribution.
     - Statistical distributions: Normal, Beta, Gamma, Cauchy.
     - Between 75 and 120 *features*.
     - We determine the importance of each feature on the target.
     - We use 4 parameters: **Overlap, noise, sparsity** y **corruption.**
       - ✓ **¿Overlap?**
         - ❖ Parameter between 0 y 1.
         - ❖ The lower the overlap, the greater the separation 0s y 1s.

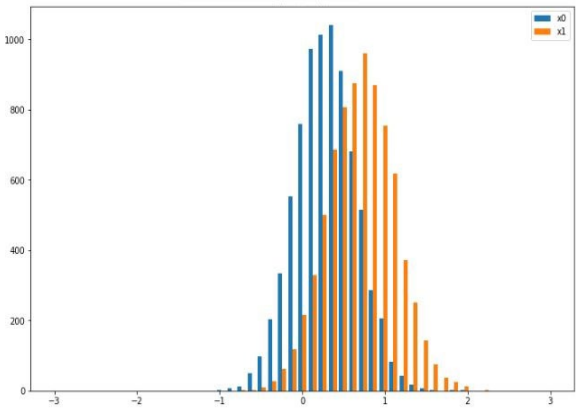
Return

Alonso-Robisco & Carbó (2022c)

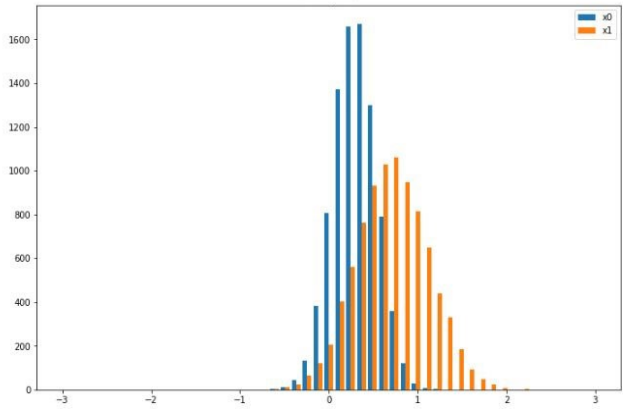
¿How does it work Overlap?

- $\mu_0, \mu_1 \in [0, 1]$. For instance, μ_0=0.3 and μ_1=0.7
- $\sigma_1 = |\mu_1 - \mu_0| = 0.4$
- $\sigma_0 = \sigma_1$*Overlap

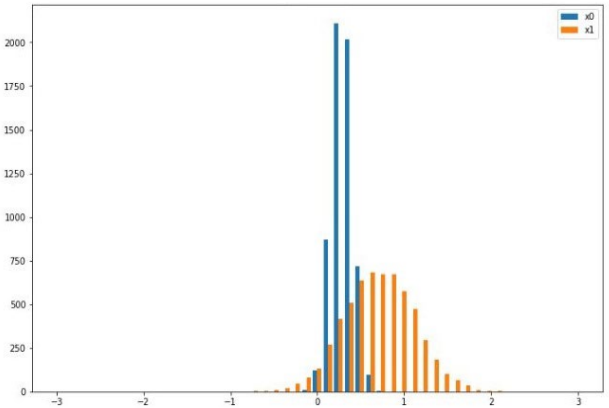
Overlap = 0.9



Overlap = 0.6

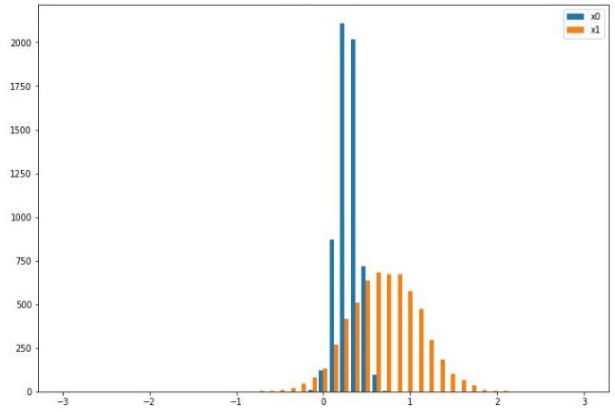


Overlap = 0.3



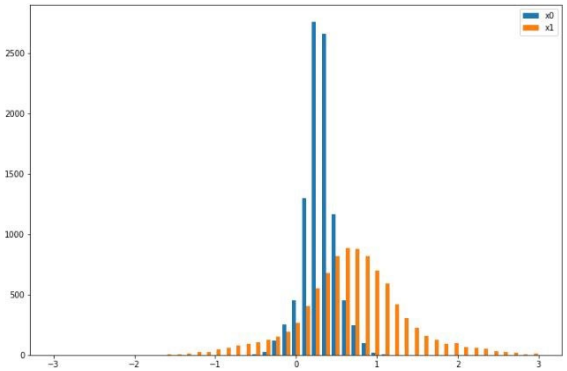
Return

Overlap = 0.3



1

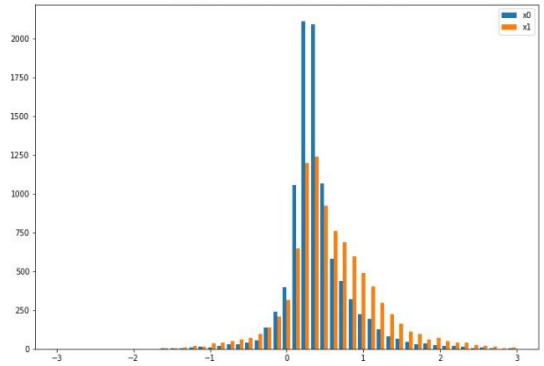
We add Noise



2

Noise:
% of observations to which we add random noise.

We add Corruption



3

Corruption:
% of observations to which we change the position of the 0s and 1s without taking into account their Target values.

Return

Alonso-Robisco & Carbó (2022c)

Table 1. Synthetic datasets vs real credit datasets

	Number of features	Number of rows	Target	25 percentile correlation	50 percentile correlation	75 percentile correlation	Top correlation	(%) Categorical variables	(%) Binary variables
Mean datasets Kaggle	63.1	140000	0.078	0,013	0,042	0,116	0,95	20%	13%
Mean 500 Simulations Synthetic dataset	50.2	120000	0.067	0,022	0,057	0,116	0,7	20%	10%

- Give me some credit.
- Lending Club.
- Geekbrains AI/Big Data Loan Default Prediction Competition
- Home Credit Default Risk
- Default of Credit Card Clients Dataset:
- Loan Default Prediction - Imperial College London:
- Development of Credit Risk Model & Scorecard:
- XYZCorp_LendingData:

Return

- We simulate many datasets.
- In each dataset, we create a ranking of the features by adding these 4 parameters
  - ✓ **Less overlap, noise, corruption and sparsity → The more important the *feature*.**

Name	Class	Parameters	Sum
Var A	Normal	Overlap = 0.5 Noise = 0.5 Corruption = 0.5 Sparsity = 0.5	2
Var B	Beta	Overlap = 0.8 Noise = 0.7 Corruption = 0.9 Sparsity = 0.7	3.1
Var C	Gamma	Overlap = 0.3 Noise = 0.3 Corruption = 0.2 Sparsity = 0.3	1.1

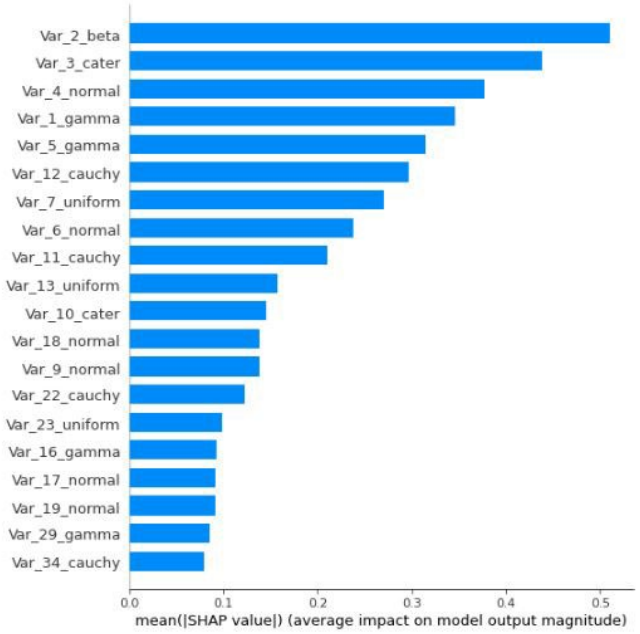
→

Ranking	Name
1	Var C
2	Var A
3	Var B

Return

- We look at a dataset with 40 variables, and 100,000 observations.
- **We train an XGBoost** and make 20,000 predictions (test sample).
- We explain these predictions with **SHAP**, which gives us a ranking.

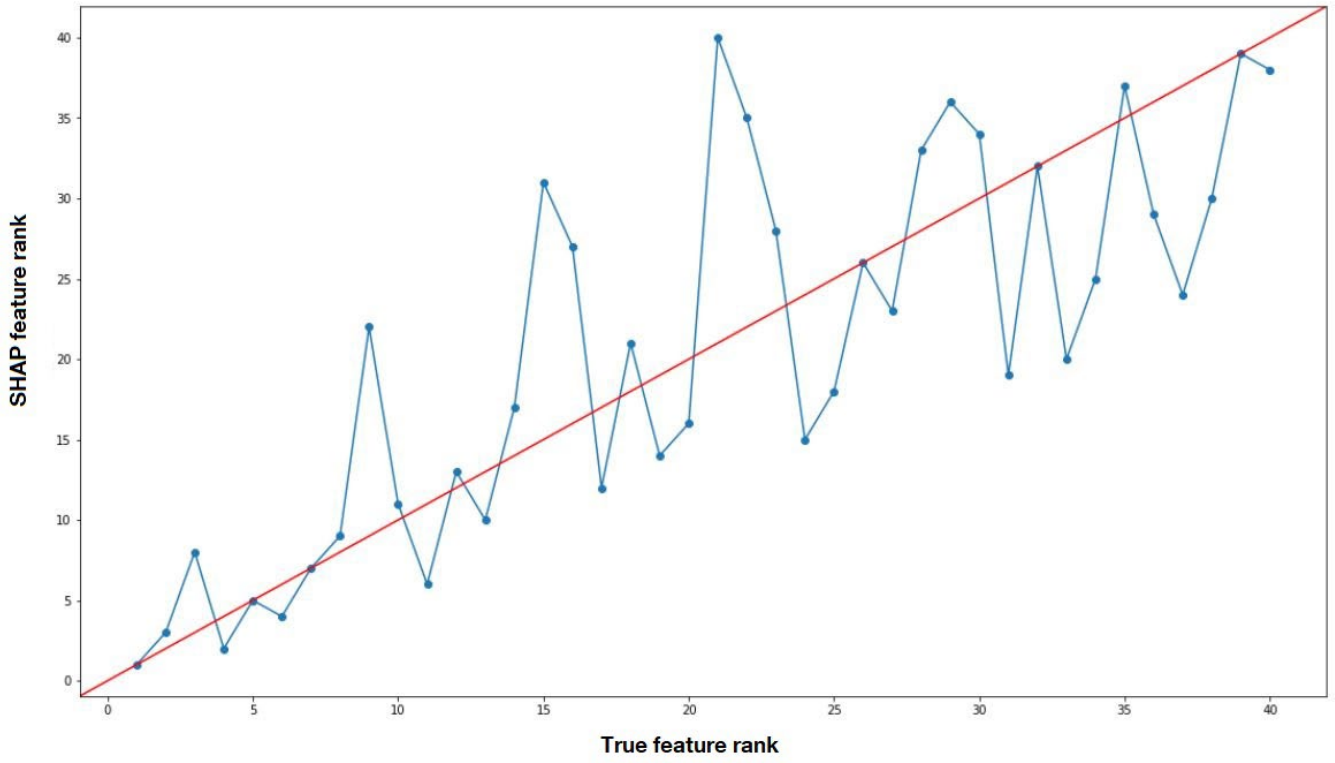
SHAP ranking



Real (generated) ranking

Var	Overlap	Noise	Sparsity	Corruption	suma	Ranking
Var_2_beta	0.806351	0.052004	0.030041	0.043461	0.931857	1
Var_4_normal	0.805112	0.051229	0.028378	0.048670	0.933389	2
Var_6_normal	0.806415	0.050326	0.037270	0.045383	0.939394	3
Var_3_cater	0.805745	0.050847	0.046444	0.051950	0.954985	4
Var_5_gamma	0.822488	0.055972	0.039620	0.040541	0.958620	5
Var_1_gamma	0.845670	0.055180	0.042595	0.050137	0.993582	6
Var_7_uniform	0.836853	0.052936	0.072182	0.048675	1.010646	7
Var_11_cauchy	0.863123	0.141602	0.113560	0.141610	1.259895	8
Var_8_beta	0.869465	0.147632	0.156775	0.104585	1.278458	9
Var_10_cater	0.857541	0.165070	0.162743	0.110153	1.295507	10

Return



- We use **Rank Biased Overlap (RBO)** to compare the result between models.

Return

References

- *Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767-2787.*
- *Albanesi, S., & Vamossy, D. F. (2019). Predicting consumer default: A deep learning approach (No. w26165). National Bureau of Economic Research.*
- *Alonso-Robisco, A., & Carbó, J. M. (2022b). Can machine learning models save capital for banks? Evidence from a Spanish credit portfolio. International Review of Financial Analysis, 102372.*
- *Fraisse, H., & Laporte, M. (2022). Return on investment on artificial intelligence: The case of bank capital requirement. Journal of Banking & Finance, 138, 106401.*

Mainly economic journals.

References Biases

- *Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. The Journal of Finance, 77(1), 5-47.*
- *Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 143(1), 30-56.*
- *Dobbie, W., Liberman, A., Paravisini, D., & Pathania, V. (2021). Measuring bias in consumer lending. The Review of Economic Studies, 88(6), 2799-2832.*

Publicaciones TOP ranked …

Return

References images

- Ikasari, I. H., Ayumi, V., Fanany, M. I., & Mulyono, S. (2016, October). Multiple regularizations deep learning for paddy growth stages classification from LANDSAT-8. In *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (pp. 512-517). IEEE. Return

- https://www.tibco.com/reference-center/what-is-a-random-forest Return

- https://www.oreilly.com/library/view/data-analysis-with/9781788393720/7d7c538e-2f7f-4e7e-9661-cd34f37c4711.xhtml Return

- https://deepai.org/machine-learning-glossary-and-terms/weight-artificial-neural-network Return