SAFE Artificial Intelligence in Finance

Paolo Giudici

Professor of Statistics and Machine Learning
University of Pavia

AI: opportunities and risks

- AI can improve operational efficiency, consumer experience and inclusion.
- AI can also bring risks, such as cyber risk, model risk, governance risks, discrimination risks.
- For this reason, codes of practice and regulations are being developed for AI deployers and providers. Among them, The EU AI Act is a risk based law, which distinguishes:
  - Prohibited AI practices (e.g. social scoring)
  - High risk AI practices: allowed, subject to a risk management system (e.g. credit scoring, life insurance pricing)
  - Limited risk practices: allowed, subject to transparency requirements (e.g. chatbots)

Proposal

In the line with the EU AI Act, We propose a S.A.F.E. measurement model for the governance of Artificial Intelligence, based on four main principles:

- Security: measures the robustness of AI systems
- Accuracy: measures the truthfulness of AI systems
- Fairness: measures the inclusiveness of AI systems
- Explainability: measures the controllability of AI systems

For the measurement, we propose four consistent metrics, that extend the Area Under the ROC Curve (AUC) to all types of response variables, leveraging the properties of the Gini index.

The Gini index

Measure of inequality in a population variable Y:

$$G(Y) = \frac{E|Y_i - Y_j|}{2|E(Y)|}$$

Also a measure of variability, alternative to the coefficient of variation:

$$CV(Y) = \frac{\sqrt{E(Y_i - Y_j)^2}}{|E(Y)|}$$

It measures classification accuracy in machine leraning:

$$G(\hat{Y}) = 1 - 2 * AUC(\hat{Y}),$$

Gini index and Concordance

Let Y be $i = 1, \ldots, n$ response values to be predicted. Recall:

- Lorenz curve (L_Y): $(i/n, \sum_{j=1}^{i} y_{r_j}/(n\bar{y}))$, where r_j indicates the non-decreasing ranks of Y and $\bar{y}$ the mean of Y.
- dual Lorenz curve: $(L_Y^{'})$: $(i/n, \sum_{j=1}^{i} y_{r_{n+1-j}}/(n\bar{y}))$, where r_{n+1-j} indicates the non-increasing ranks of Y.
- Lorenz Zonoid: area between the L_Y and $L_Y^{'}$, corresponding to the Gini coefficient.

Definition

A concordance curve (C) can be defined by: $(i/n, \sum_{j=1}^{i} y_{\hat{r}_j}/(n\bar{y}))$, where $\hat{r}_j$ indicates the non-decreasing ranks of $\hat{Y}$.

Gini index and Concordance

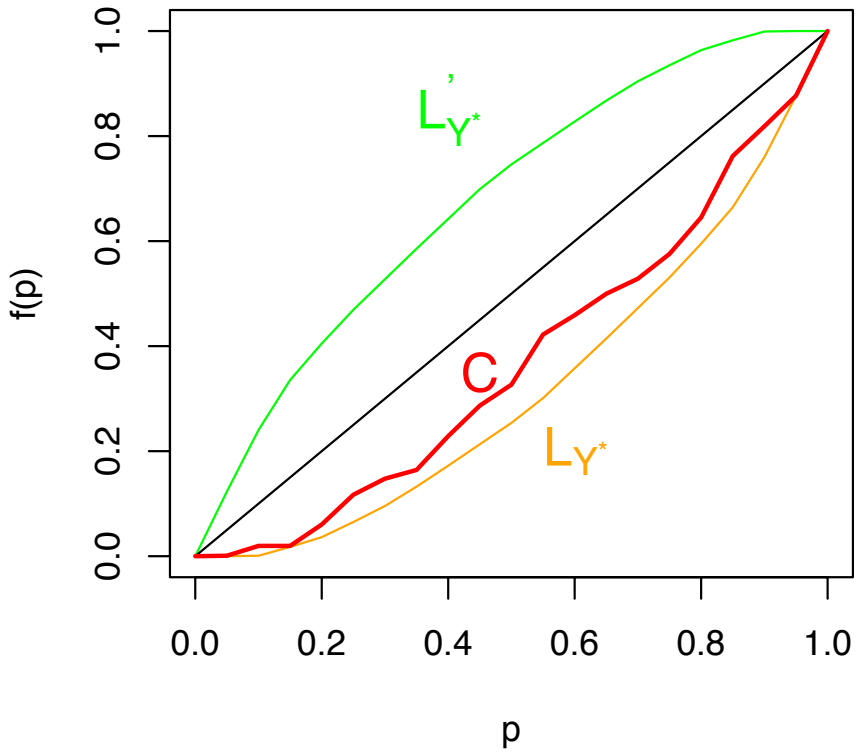

Figure: The L_Y and $L_Y^{'}$ Lorenz curves and the C concordance curve, where p and $f(p)$ are the cumulative values of the x and y coordinates of the L_Y, $L_Y^{'}$ and C curves.

The Rank Graduation Accuracy metric (RGA)

- Dividing the area between the concordance curve and the dual by its maximum value (the Lorenz Zonoid) we obtain:

$$RGA = \frac{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{y}} \left(\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{\hat{r}_j} \right) \right\}}{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{y}} \left(\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{r_j} \right) \right\}}.$$

- It can be shown that: $0 \leq \text{RGA} \leq 1$, with RGA$= 1$ for a perfectly concordant model; RGA$= 0$ for a perfectly discordant model; RGA$= 0.5$ for random predictions;
- When the response Y is binary, RGA=AUC;
- RGA can however be calculated for all types of response variables: ordinal, continuous and multivariate.

Application: accuracy in personal lending

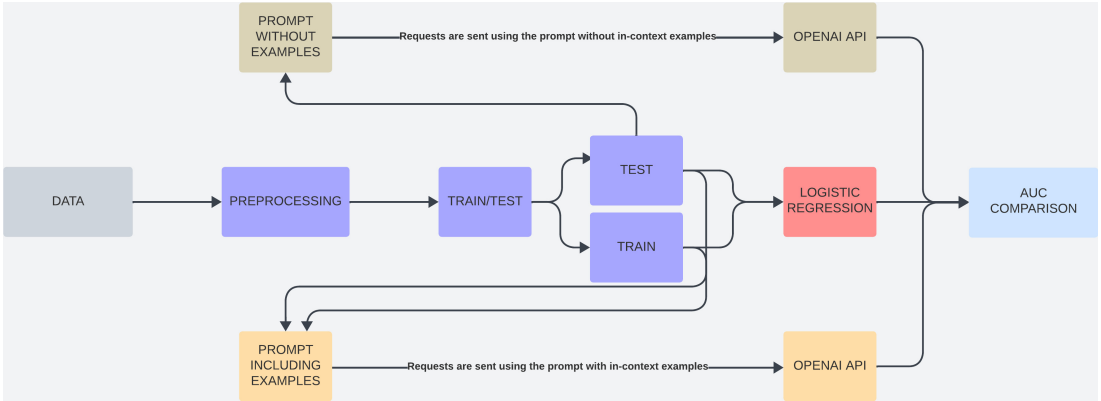

Figure: Generative AI

Method	Min RGA	Max RGA	Mean RGA	Std RGA
LR (30,000 ex.)	0.7018467852	0.7950752394	0.75318741451	0.0314667993
GPT(0 ex.)	0.5895348837	0.6445964432	0.61264021887	0.0212562962
GPT (100 ex.)	0.6260601915	0.6963064295	0.66655266764	0.0261630545

RGA as a Cramer distance

Definition

Given any two cumulative distribution functions $F, G : \mathbb{R} \to [0,1]$, the Cramer - Von Mises divergence between F and G is

$$\mathrm{CvM}(F,G) = \int_{-\infty}^{\infty} \left|F(u) - G(u)\right| dF(u). \tag{1}$$

Theorem

Let F_Y and $F_{\hat{Y}}$ be the cumulative distributions of Y and $\hat{Y}$. Then:

$$\mathrm{RGA}(Y,\hat{Y}) = 1 - \frac{\mathrm{CvM}(F_Y, F_{\hat{Y}})}{G(Y)}. \tag{2}$$

RGA and R^2

Definition

For any type of response, we can measure accuracy with:

$$\mathrm{RGA} = 1 - \frac{\mathrm{CvM}(F_Y, F_{\hat{Y}})}{G(Y)}.$$

Definition

If the response is continuous we can, in addition, measure accuracy with the predictive R^2:

$$\mathrm{R}^2 = 1 - \frac{MSE(\hat{Y}, Y)}{Var(Y)}.$$

RGX: a Rank Graduation Box

Definition

The mathematical derivation of RGA can be extended to all AI principles. Define:

$$\mathrm{RGX}(Y,Z) = 1 - \frac{\mathrm{CvM}(Y,Z)}{G(Y)}.$$

Different pairs (Y,Z) lead to different metrics:

- $(\hat{Y}, \hat{Y^p})$ lead to RGR, which measures robustness against adversarial perturbations p;
- $(\hat{Y}, \hat{Y^{-k}})$ lead to RGE, which measures the explainability gain attributed to variable(s) k;
- $(\hat{Y^A}, \hat{Y^B})$ lead to RGF, which measures fairness between groups A and B.

Application: robustness of Natural Language Processing

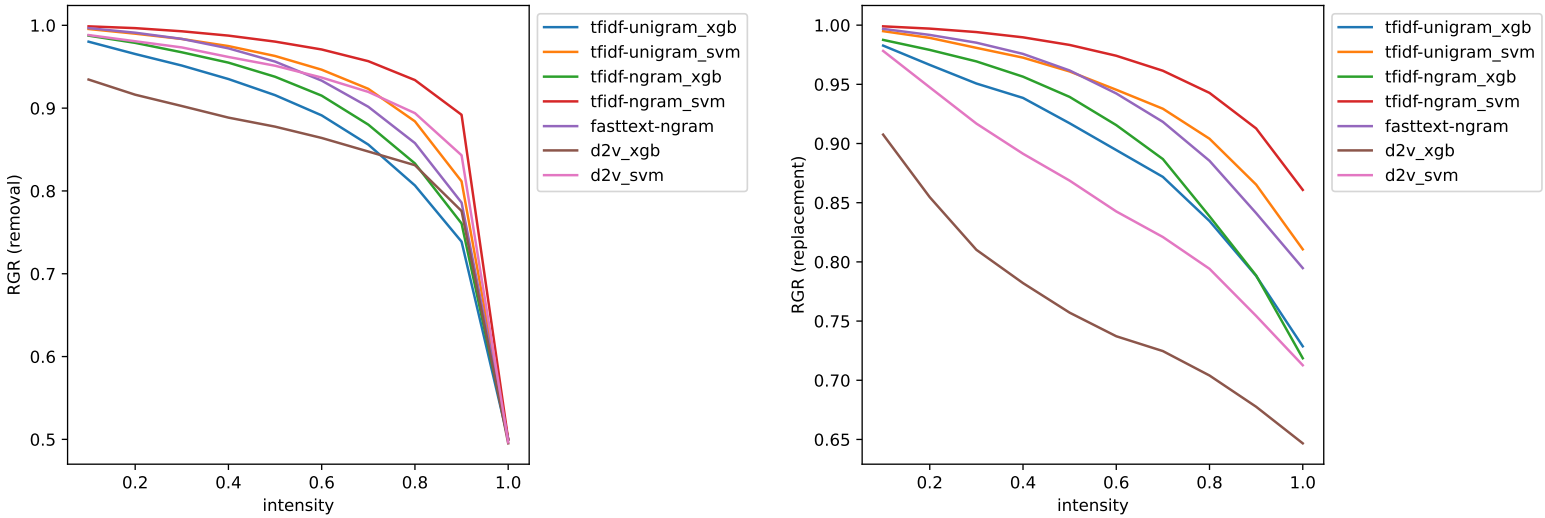

Figure: Robustness of about 68,000 decisions about consumer claims changes using word removal (left) and word replacement (right), with a growing intensity, for different AI models, based on different Natural Language Processing methods.

Application: explainability of financial predictions

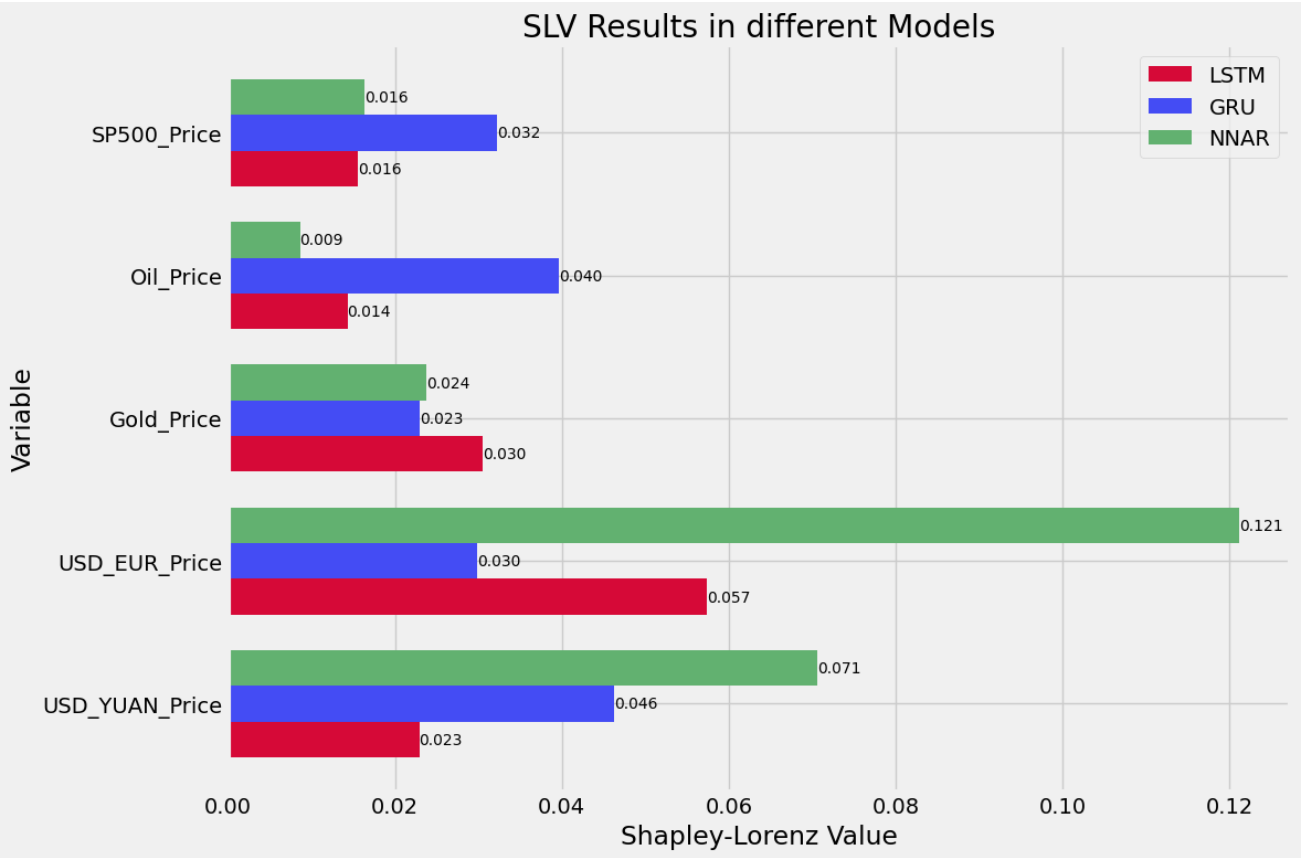

Figure: Model explanations for three neural networks: (NNAR, LSTM, GRU). The Shapley Lorenz value are percentages of explained accuracy.

Application: phishing detection

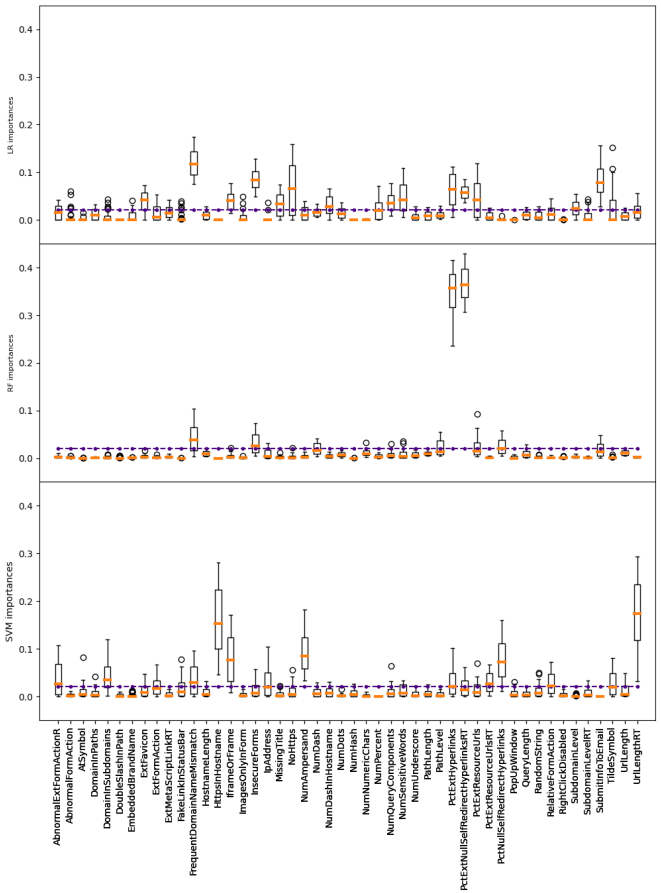

Figure: Boxplots representing the distribution of the importance of each feature, with different classifiers

Application: fairness in corporate lending

Chen et al. (2024) consider:

- a credit rating panel data of about 100,000 small and medium-sized enterprises (SMEs);
- 4 European countries: Germany (DEU), France (FRA), Italy (ITA), and Spain (ESP); 11 industry sectors from 2015 to 2020;
- 6 financial accounting variables (predictors), including operating revenue (Turnover), operating profit/loss (EBIT), profit/loss after tax (P/L), Leverage, return on equity (ROE), and total assets (TA);
- the 10 credit ratings (response variable) aggregated into three groups: Strong (AAA; AA; A), Normal (BBB; BB; B) and Vulnerable (CCC; CC; C; D);
- the application of a random forest model.

Application: fairness in corporate lending

Percent share of Shapely Lorenz for countries

		Leverage	P/L	EBIT	ROE	TA	Turnover
ITA		0.3001	0.2139	0.1922	0.1455	0.0958	0.0525
	2015	0.3121	0.2126	0.1834	0.1471	0.0895	0.0554
	2020	0.3069	0.2108	0.1925	0.1555	0.0905	0.0437
FRA		0.3131	0.1916	0.1948	0.1466	0.1034	0.0506
	2015	0.3147	0.1957	0.2000	0.1474	0.0928	0.0495
	2020	0.3221	0.1924	0.1909	0.1509	0.1012	0.0424
ESP		0.3168	0.1988	0.1815	0.1478	0.1029	0.0522
	2015	0.3288	0.2005	0.1756	0.1506	0.0906	0.0540
	2020	0.3113	0.2014	0.1875	0.1642	0.0934	0.0422
DEU		0.3205	0.1439	0.1301	0.1874	0.1671	0.0508
	2015	0.3304	0.1468	0.1176	0.1952	0.1674	0.0426
	2020	0.3364	0.1540	0.1453	0.1808	0.1517	0.0317
Benchmark		0.3013	0.2018	0.1871	0.1515	0.1065	0.0518

Gini-Fairness coefficient and Kolmogorov-Smirnov test for variable Country (Shapley Lorenz) - Year 2020

	Leverage	P/L	EBIT	ROE	Turnover	TA
$\mathcal{F}_k$	0.0362	0.0591	0.0506	0.0378	0.0566	0.1095
p-value	> 0.10	> 0.10	> 0.10	> 0.10	> 0.10	> 0.10

Integrating the metrics for AI assessment

Model	Sust.	Acc.	Exp.	Prob	RMSE
MLP	0.9661	0.4518	0.5114	0.7768	0.1046
RBF	0.9538	0.4519	0.5443	0.75454	0.0982
NNAR	0.7157	0.3718	0.2405	0.9361	0.1358
LSTM	0.9607	0.8186	0.1122	0.9118	0.0561
GRU	0.9244	0.8865	0.1778	0.8543	0.0439

Table: AI metrics integration: comparison of MLP, RBF, NNAR, LSTM and GRU models to predict bitcoin prices, in terms of three S.A.F.E. AI metrics, in comparison with the "classic" Root MSE metric.

Integrating the metrics in Agentic AI

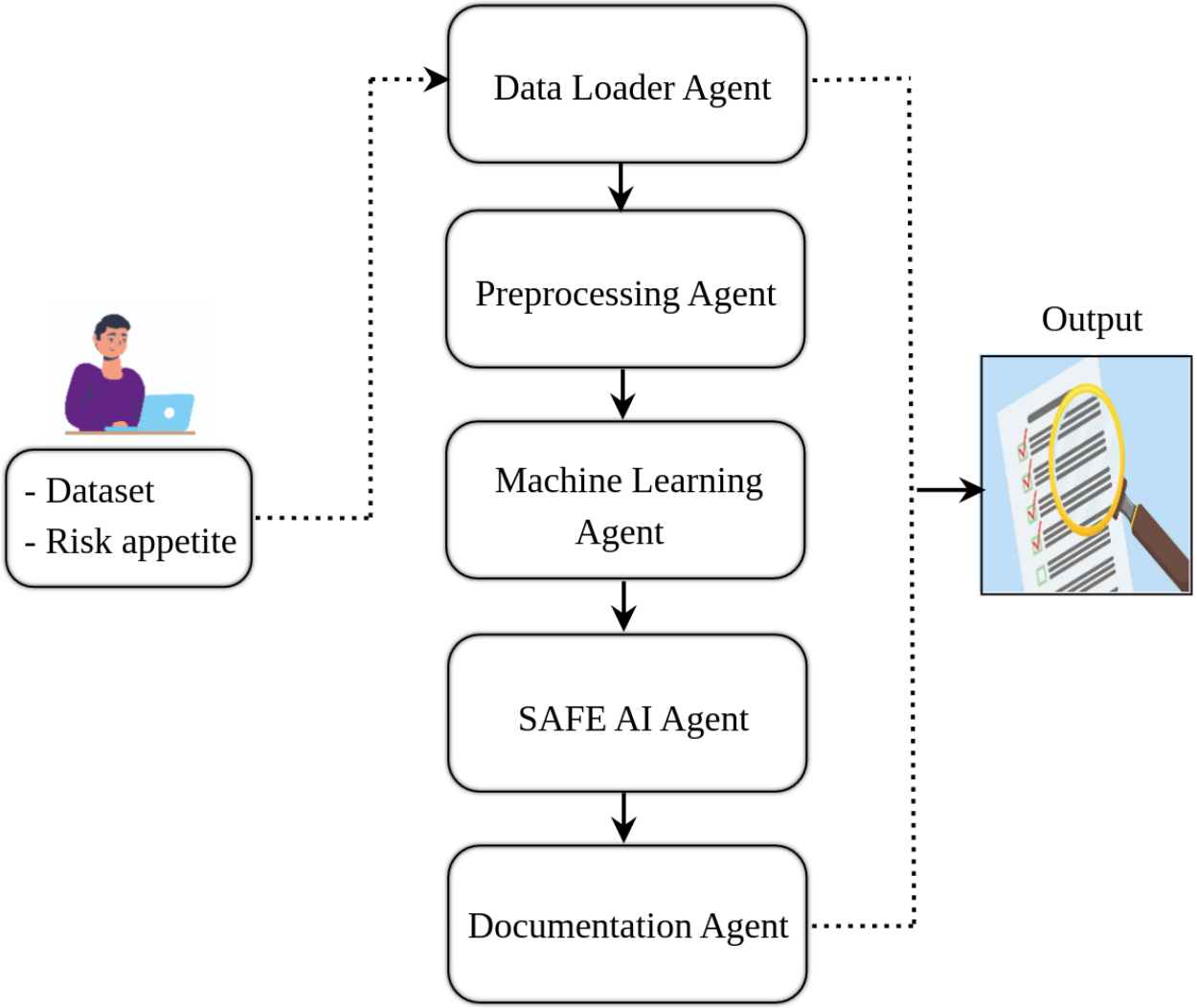

Figure: SAFE AI by Agentic Design

Agents' report

- *The computed RGA score is* 0.6875*. This significant score is above the acceptable threshold of* 0.6*, indicating that the model exhibits a strong alignment with true labels...*
- *...The computed RGR score is* 0.91*. This indicates the model's strong performance in making predictions, regardless of variations in the data...*
- *...The Highest Explainability is for Gender (0.1673) and Loan Purpose (0.1106). The Lowest Explainability is for Race (0.0001), indicating a lack of understanding of the influence of this feature on model predictions....*
- *...In summary, the model displays a high level of accuracy and robustness across all features, making it a reliable choice for classification tasks. However, there is an evident disparity in explainability across the features, which suggests a need for further investigation into how different features contribute to model predictions to enhance transparency.*

References


Giudici P, Raffinetti E. (2024). RGA: a unified measure of predictive accuracy. *Advances in Data analysis and classification.*


Giudici P, Raffinetti E. (2025).A rank graduation Box for SAFE AI. *Expert Systems With Applications*


`https://github.com/GolnooshBabaei/safeaipackage`.


Babaei, G., Giudici, P. (2024). GPT classifications, with application to credit lending. Machine learning with applications, 16.


Giudici, Piergallini, Raffinetti, Recchioni (2025). Explainable AI methods for financial time series. Physica A: statistical mechanics with applications, 2024, 655, 130176.


Chen, Y., Giudici P, Liu, K. and Raffinetti E. (2024). Measuring fairness in credit ratings. *Expert Systems With Applications* 258, 125184.