prometeia

A Novel AI-based Model Validation Framework and Use Cases in Banking Sector

Data Science Turkey
May 2022

smartcon

- EXECUTIVE SUMMARY
- AI MODEL VALIDATION FRAMEWORK: PROMETEIA VALUE PROPOSITION
- METHODOLOGICAL INSIGHTS
- USE CASES

prometeia

EXECUTİVE SUMMARY

prometeia

Scope

- A specific insight on "**AI Model Validation Standards**" framework related to **validating machine learning** and even **deep learning** models on the distinctive elements of validation (e.g., performance, interpretability, data management, process, governance…) as emphasized in the recently published 'consultation paper' by European Banking Authority.

- A high-level **overview of the validation areas** with a particular **insight** details on **methodology validation**.

- The **Prometeia value proposition** which combines **experience** on **real-life AI models** in banking as well as **advanced technical deep-dive** know-how on machine learning and deep learning best practices.

- **Use cases** and outputs on previously completed AI validation projects.

prometeia

Our AI approach

Set of techniques that emulate cognitive human processes such as learning, reasoning, understanding and acting, by exploiting typical computers' strengths: speed, replicability and scalability.

Rule based system

Intelligence is developed on a knowledge base expressed by hand-written rules.

- No hypothesis on the data
- Requires: Cascade of rules, Domain experts, No data
- Allows for human interpretability
- Heavy workload for maintenance
- Allows for quick small fixes

Statistical Learning

Intelligence is developed on a knowledge base expressed by statistical models and specialised by data.

- Hypotheses on the data
- Requires: Statistical model, Statistics experts, Data-driven fitting
- Allows for statistical interpretability
- Limited maintainability (if the hypotheses are still met)
- No small fixes in the model are allowed

Machine Learning

Intelligence is learnt directly from data without any need of human input.

- No hypotheses on the data
- Requires: Learning algorithms, Machine learning experts, Data-driven modelling
- Interpretability tied with the specific algorithm
- Limited workload for maintenance (re-training, continuous learning)
- No small fixes in the model are allowed

AI for risk initiatives

1 CREDIT RISK

→ ***Model Enhancement***
Enhancement of PD models

→ ***Augmented Credit Analysis*** Dashboard to support credit analysis based on unstructured and qualitative data (supplementary notes, news, ...)

→ ***Augmented Early Warning System***
Enrichment of monitoring systems through the use of AI techniques (using internal, external, qualitative, sector and relational data)

2 FINANCIAL RISK

→ ***Behavioral Models***
Evolution of prepayment behavioral models using internal data (CRM) and external information (online aggregators)

→ ***Financial Data Anomaly Detection***
Automatic identification of anomalies on input and output data produced by Financial Risk Management systems in Asset Management

3 OPERATIONAL RISK

→ ***Data driven Operational Risk***
Integration of Operational Risk models with AI solutions able to: identify new emerging phenomena (New Phenomena recognition), early identification of loss events (Early Warning).

EBA’s Take on Machine Learning & AI

Discussion Paper on Machine Learning for IRB Models

EBA DISCUSSION PAPER ON MACHINE LEARNING FOR IRB MODELS

EBA/DP/2021/04

11 NOVEMBER 2021

DISCUSSION PAPER ON MACHINE LEARNING FOR IRB MODELS

EBA EUROPEAN BANKING AUTHORITY

For complex ML models with limited explainability or for frequently updated models, a reliable **validation** is particularly important and might require increased depth and/or frequency. The institutions are recommended to take care of:

i. Overfitting issues: ML models are very prone to suffer from overfitting, i.e. performance optimisation of the development sample, which leads to very high performance of the development sample that may not be confirmed on the current and foreseeable application portfolio. Therefore, they should put particular attention on the comparison of the model performances measured within the development sample with those obtained using out-of-sample and out-of-time samples.

ii. Challenging the model design: the hyperparameters used to describe the structure of the model and to customise the learning algorithm are often based on human judgement. The validation unit should therefore place particular attention on verifying the rationale behind the choice of these hyperparameters. This check may prove to be particularly challenging for complex models considering that a deep knowledge of the methodology is required to understand all implications of hyperparameters. If, on the contrary, hyperparameters are selected by minimising the error of the model, it should be ensured that this process does not introduce an undesired bias.

iii. Representativeness and data quality issues: If ML techniques used for risk differentiation and risk quantification purposes are fed with a large amount of data, sufficient data quality needs to be ensured. Where these data are external data, institutions are recommended to place particular care on the assessment of the representativeness of the external data with respect to the application portfolio. In particular, institutions are recommended to verify whether a diminished representativeness leads to a reduction in the performance of the model measured strictly on the internal customers. Institutions should also be particularly careful when using unstructured data in ensuring accuracy, completeness and appropriateness of the data.

iv. Analysis of the stability of the estimates, also in light of the institution’s rating philosophy. It is useful to analyse both the stability:

- In the assignment process of each debtor/exposure to grades or pools. Indeed, ML algorithms may introduce point-in-time (PiT) elements in the models that may hamper the stability of the rating assignment process compared to more through-the-cycle (TtC) models leading to potential rapid changes in capital requirements;
- Of the relationship between the output variable and the drivers in subsequent releases of the model based on ML techniques especially in light of the model change policy, to provide an assessment of whether changes between inputs and outputs require regulatory approval, *ex ante* or *ex post* notification.

EBA took the first step for the future use of artificial intelligence-based models for IRB by publishing a "consultation paper" concerning the **usability of Machine Learning models within the scope of IRB**. (document link)

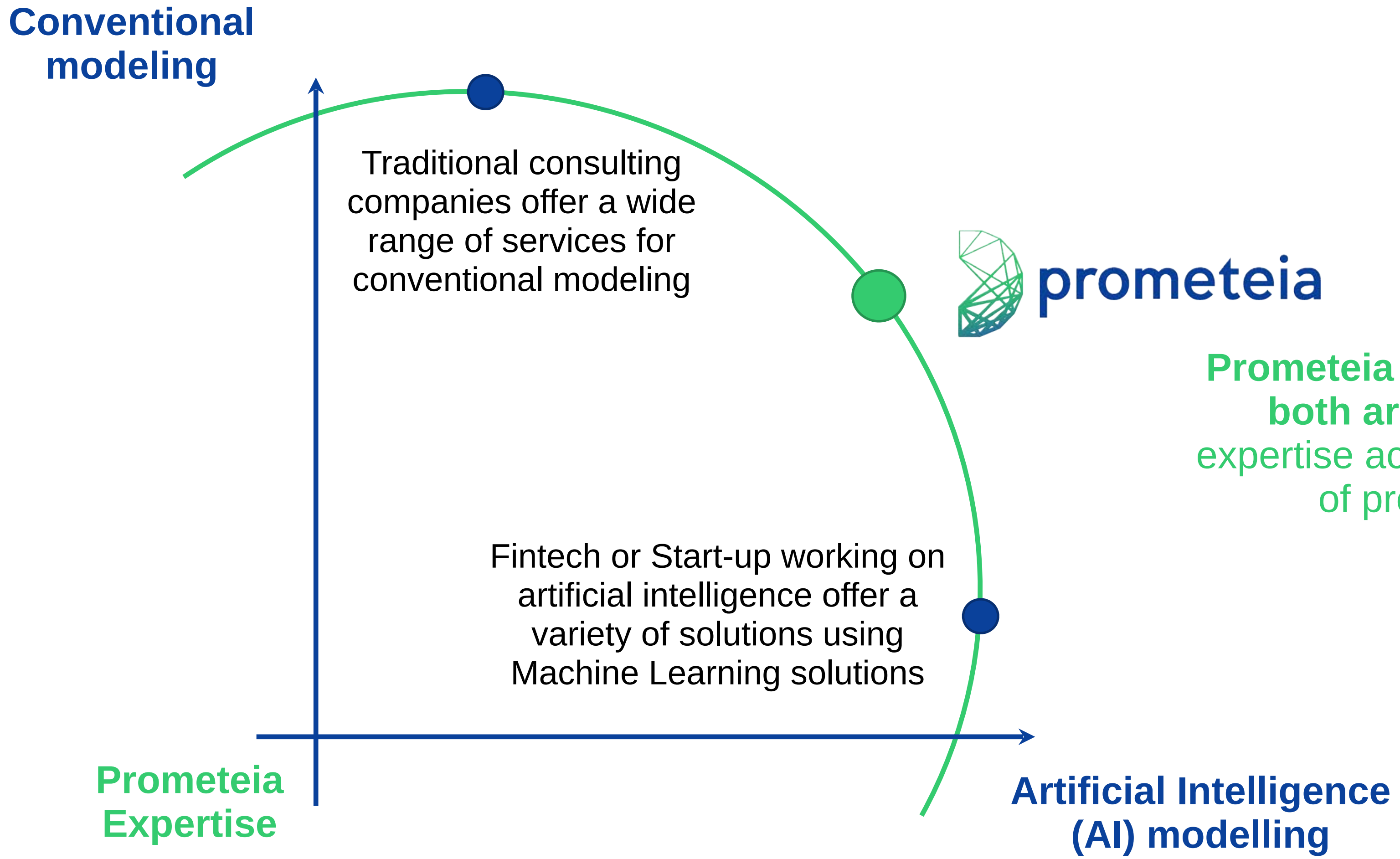
Principle-based recommendations regarding the integration of Machine Learning models into IRB models and the **explainability of these models**, their openness to expert intervention, **validation and precautionary practices** are also envisaged in the post-consultation process.

Prometeia, accompanies clients in this journey by;

- Creating a **four-dimensional** (data, methodology, process, governance) **validation framework** for the validation of the IRB models generated with Machine Learning,
- Organizing **comprehensive training sessions** on Data Science and Artificial Intelligence.

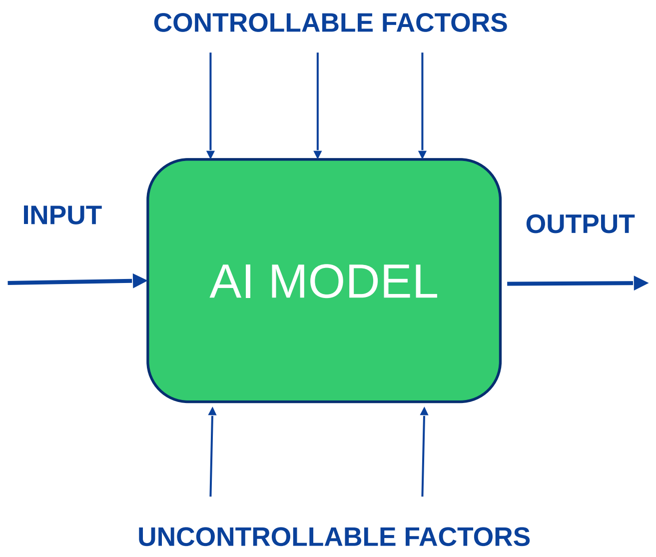
AI Model Validation Framework: Prometeia Value Proposition

Prometeia Value proposition



Prometeia is able to fully cover both areas combining the expertise acquired in a wide range of projects delivered

Machine Learning validation approach



- **Controllable Factors**:
  - Learning algorithm used
  - Hyperparameters of the algorithm (the number of hidden units for a multilayer perceptron, k for k-nearest neighbor, C for support vector machines)
- **Uncontrollable Factors**:
  - Noise in the data
  - Randomness in the optimization process (initial state of gradient descent with multilayer perceptrons)

VALIDATE CONTROLLABLE FACTORS

DIAGNOSE UNCONTROLLABLE FACTORS

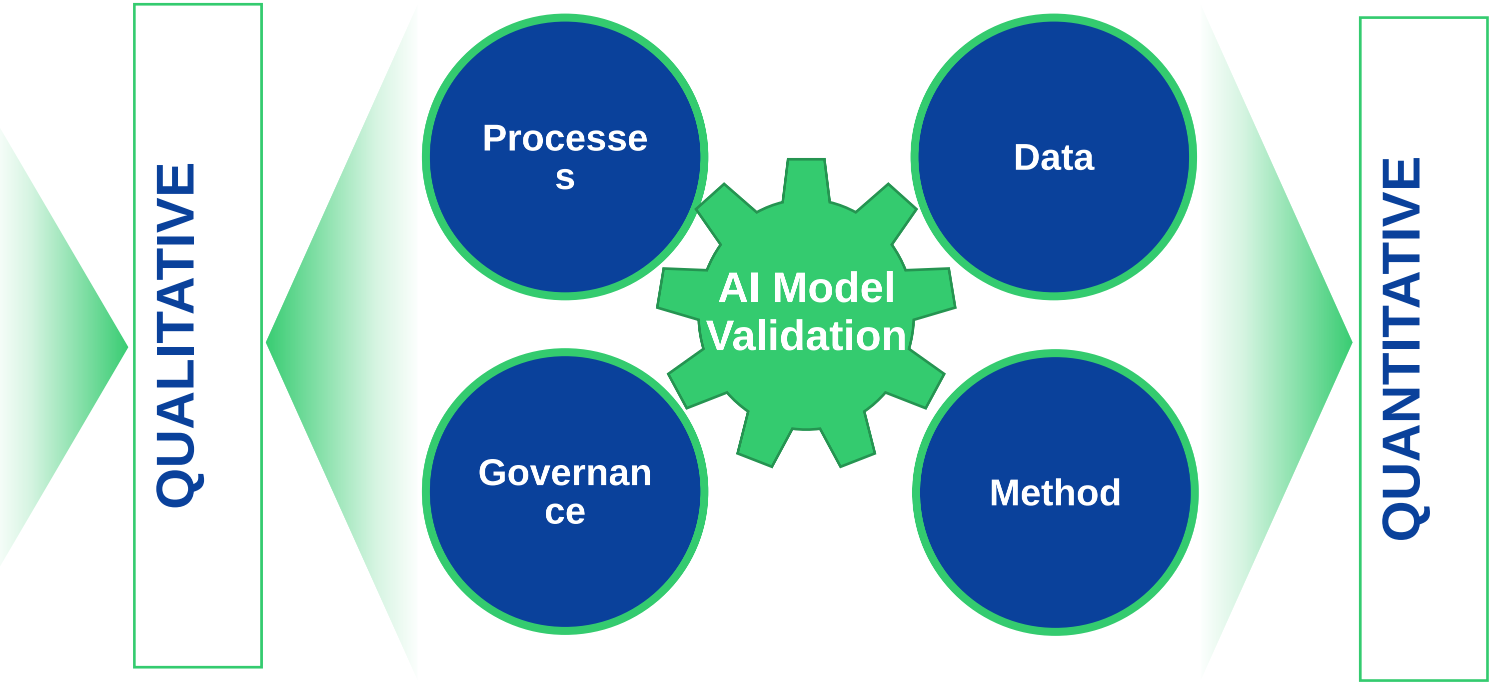
Validation landscape

PROMETEIA AI VALIDATION FW

MODEL DEVELOPMENT
ML Engineering specific KPIs for model development & implementation

MODEL VALIDATION
Validating black-box or white-box (inhouse developed) ML models

MODEL END-USER
Business-wise KPIs followed on dash-boards
- Coverage
- Edit/Acceptance percentages



- **End-to-end validation framework**: Our framework covers all 4 pillars of AI model validation.
- **Two main spectrums**: Corresponding pillars serve to 2 different spectrum of AI models: quality (discriminative power) and quantitative (calibration power).

AI Validation Framework: DATA

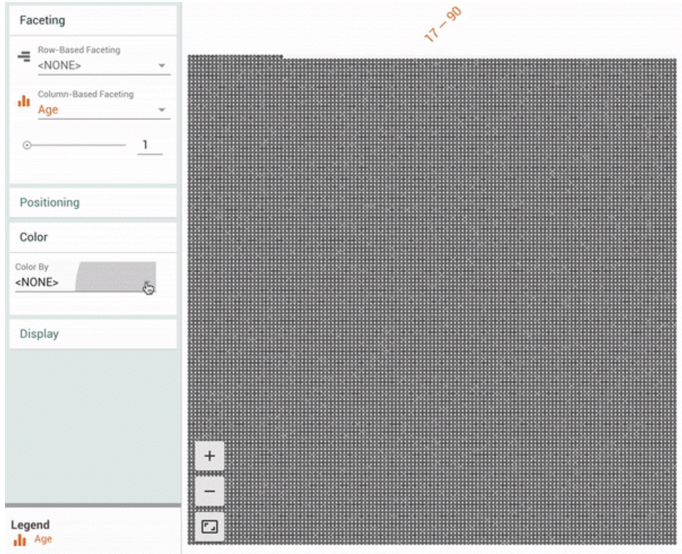
Data

Raw Data Quality Checks	Raw Data to Model Input	Target Value in Raw Data	Model Data Hold-out

• Sparsity of data:
  • Text data surely includes sparsity
• Are there categorical info?
• Is normalization required?
  • How multi-source joins performed?
• …
	• What-if raw data design changes?
	  • Is there an alert mechanism for model owners?
	• How data category changes or new additions handled?
	  • One-hot encoding
	  • Label-encoding
		• Are target values balanced or imbalanced?
		• If target value does not exist, how manual labeling performed?
		  • How many annotators?
		  • Annotator sanity chekcs?
		  • …
			• SOT train/test/validation splits:
			  • %70-%30 splits, randomization?
			• What-if small sample size?
			  • Data Augmentation?
			  • Cross-validation
			• Automated data slicing?

Data Visualization Tools & Technologies

FACETS - KNOW YOUR DATA



No Free-Lunch Theorem(*)

- THERE İS NO SUCH THİNG AS THE "BEST" LEARNİNG ALGORİTHM.
- FOR ANY LEARNİNG ALGORİTHM, THERE İS A **DATASET WHERE İT İS VERY ACCURATE** AND **ANOTHER DATASET** WHERE İT İS **VERY POOR.**

AI Powered Data Slicing

PROMETEIA VP

- TREE-BASED SLİCİNG
- LATTICE SEARCHING

(*)Wolpert

Metho d

AI Validation Framework: METHODOLOGY

PERFORMANCE VALIDATION

		Metrics & Scoring	Feature Selection & Extraction	Feature Interpretability	Hyperparameter Tuning	Overfitting (Bias & Variance)
SUPERVISED	REGRESSION	MALLOW'S CP R SQUARED MEAN SQUARED ERROR	LOW VARIANCE ELIMINATION CHI-SQUARED STATS F-TEST		EXHAUSTIVE GRID SEARCH RANDOMIZED SEARCH	K-FOLD CROSS VALIDATION REGULARIZATION NOISE ADDITION
SUPERVISED	CLASSIFICATION Binary Multi-Class Multi-Label	BALANCED: ACCURACY, PRECISON & RECALL, CONFUSION MATRIX IM-BALANCED: F1 SCORE, ROC & AUC	TIME-SERIES: TS-FRESH, AUTOML, LDA IMAGE: ISOMAP	LOCAL: LIME, SHAP, PAIR WHAT-IF-TOOL GLOBAL: SHAPLEY VALUES, ANCHORS	TREE-BASED ML: MAX DEPTH, MIN CHILD WEIGHT, SAMPLING RATE DEEP LEARNING: XAVIER INIT, TRANSFER LEARNING	TREE-BASED ML: BAGGING, BOOSTING, SAMPLING RATE
UN-SUPERVISED	CLUSTERING	FOWLKES MALLOWS HOMOGENITY SCORE RECONSTRUCTION ERROR	PCA FACTOR ANALYSIS MDS		LEARNING RATE DROP-OUT EARLY STOPPING	ENSEMBLING DROP-OUT EARLY STOPPING

METHODOLOGİCAL INSİGHTS

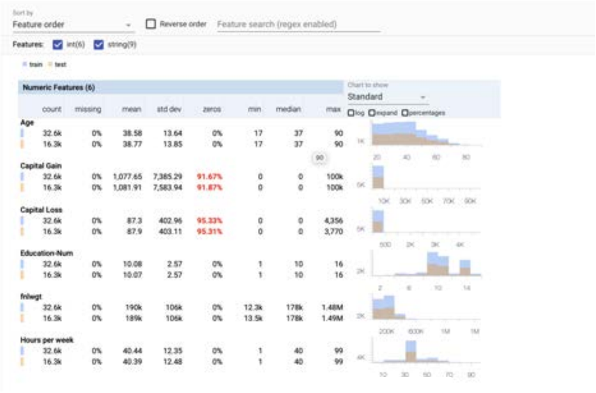
prometeia

Methodology: Feature Interpretability

Method

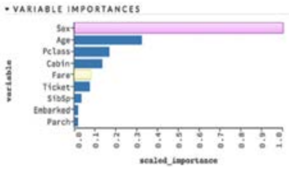
FAIRNESS | **ROBUSTNESS** | **TRUST**

UNDERSTAND FEATURE DISTRIBUTIONS



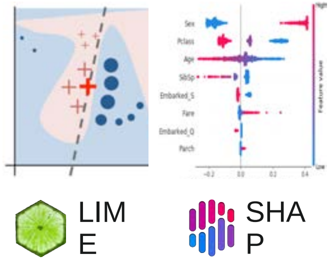
FACETS – KNOW YOUR DATA

INTERPRET SINGLE PREDICTIONS(LOCAL)



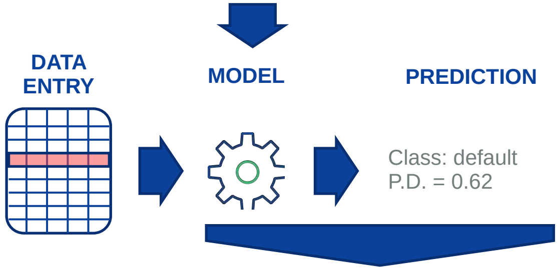
The **global feature importance** indicates the impact of a feature on the model for the entire training data set.

It is obtained by **analyzing the model's parameters** (e.g. weights for regression, gain/coverage/ frequency of features in tree-based models)



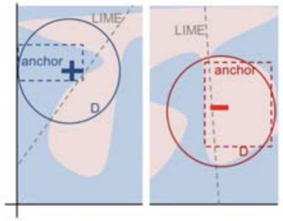
Local interpretation: approximate local feature importance values ranked to create reason codes for every prediction.

Explaining how the **model behaves in the vicinity of** the **instance** being predicted.



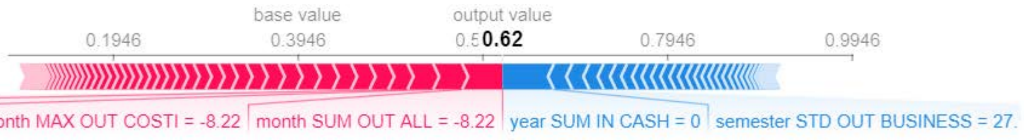
SHAPLEY VALUES

INTERPRET PREDICTIONS (GLOBAL)



Anchors: explains individual predictions of any black-box classification model by finding a decision rule that "anchors" the prediction sufficiently. **A rule anchors a prediction** if changes in other feature values **do not affect the prediction (*)**.

PAIR – WHAT-IF ANALYSIS



REASONS IN FAVOUR OF CLASS DEFAULT | **REASON AGAINST CLASS DEFAULT**

(*) Ribeiro, Singh, and Guestrin 2018

Methodology: Performance Metrics & Scoring

Metho
d

What type of technical problem?

Metrics Validation → Regression, Clustering, Classification

Regression → R-Squared, MAPE, …

Clustering → Homogenity Score, Reconstruction Error, …

Data type differentiates validation methods: Each problem may contain different types of data like text, image, tabular data. Classification problems on text data or image data may require different questions to be asked.

Classification

A sample deep-dive validation view for classification models: Each step should be evaluated as a check-list (questions to be asked to model developers) in order to find the optimal metrics & scoring.

What type of classification problem? Binary | Multi-label | Multi-class **Are all classes equally important?**

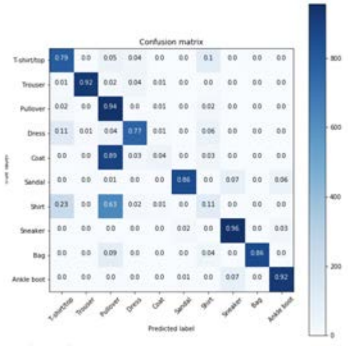
Is there enough training data for deep learning? Deep Learning | Machine Learning | Balanced Class Types | Imbalanced Class Types **How imbalanced data is managed (sub-sampling vs. over-sampling vs as-is)?**

Precision & Recall

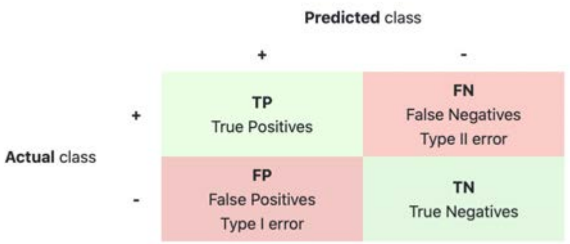
Accuracy

F1 Score

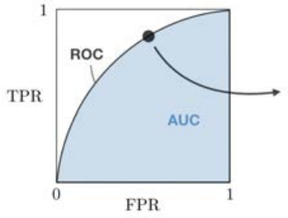
Multi-class confusion matrix

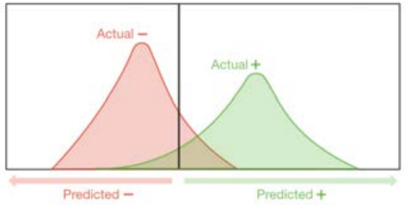


Binary confusion matrix(*)



ROC Curve Area Under Curve(*)





Accuracy dilemma?

(*)Stanford CS-229 ML

Methodology: Overfitting & Underfitting (Bias vs Variance)

Metho
d

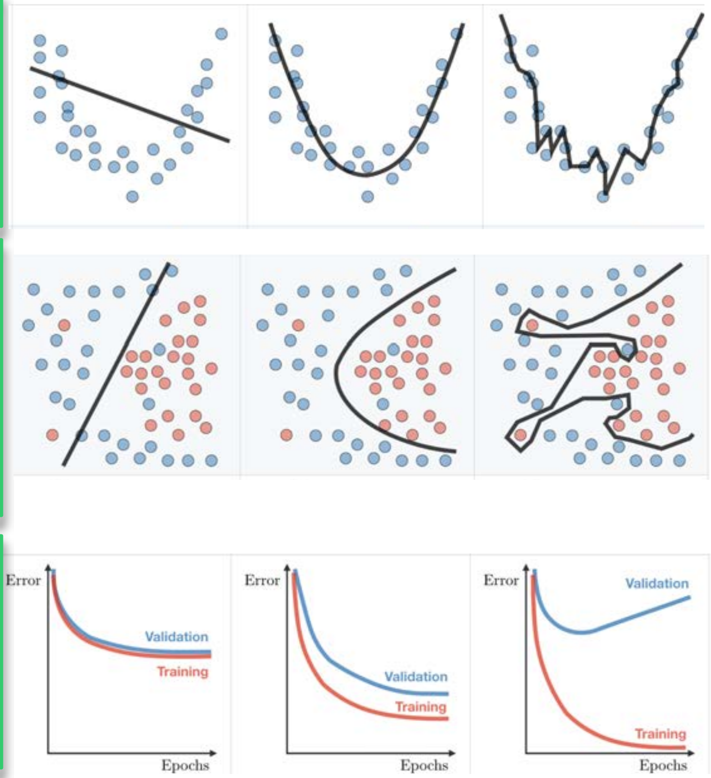
DETECT OVERFIT & UNDERFIT

UNDERFITTING | GENERALIZATION | OVERFITTING

REGRESSION

CLASSIFICATION

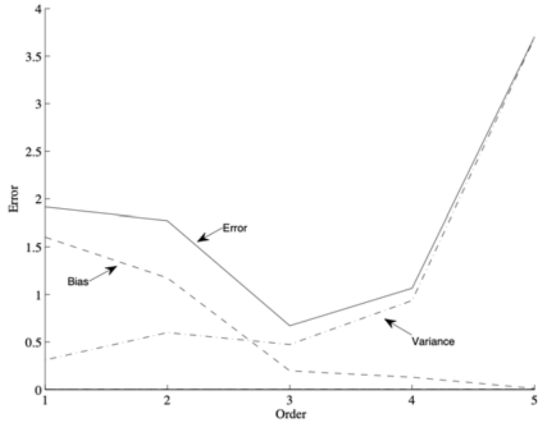
DEEP LEARNING

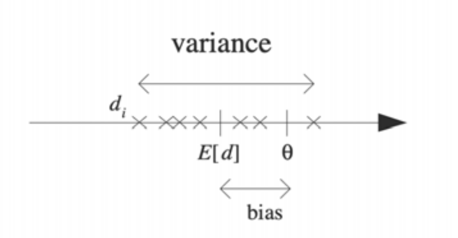


HIGH BIAS

HIGH VARIANCE

BIAS vs VARIANCE DILEMMA

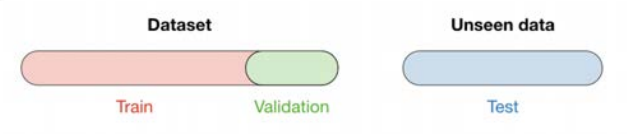




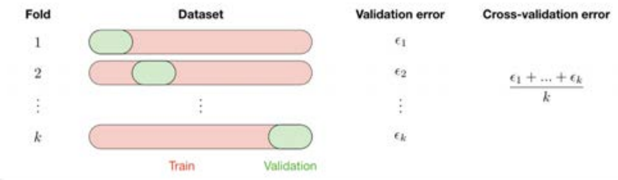
If there is **bias**, this indicates that our model class **does not contain the solution**; this is **underfitting**. If there is **variance**, the model class is **too general** and also learns the noise; this is **overfitting.**

SOLVE OVERFIT PROBLEM

DATA HOLD-OUT (ENOUGH DATA SIZE)



CROSS VALIDATION (SMALL DATA SIZE)



AI Validation Framework: PROCESSES

Process

Performance (error rate) is **the only one metric** that affect our decision on validation of the AI model. We should take care of **process specific additional factors**:

Identify **train timeliness** of the model by analyzing **training time** and **space complexity** of the model

Identify **production timeliness** of the model by analyzing **testing time** and **space complexity** of the model

Identify **interpretability** of the model by **end-users** of the application: method should allow **knowledge extraction** which can be checked & validated by business experts

Identify **easy programmability** feature of the model whether **easy to train, easy to test** and **make it** deployed to production

Identify **documented information** of the model about **layout of the algorithm**, **architecture visualization** of the deep neural network

Identify **processing power and memory requirements** of the model to analyze its deployment environment

prometeia

AI Validation Framework: GOVERNANCE

Gover
n

Governance of an AI model requires nearly same -even sometimes more- effort than model development. AI models may become **obsolete**, may have **different live performance** than test.

Identify whether AI model is **obsolete** and can be prevented by structuring metric **threshold control dashboards** and supported with **continuous learning** mechanisms

Support governance of AI models with **continuous-learning** structure

Identify possible novel state-of-the art **benchmark algorithms** outperform the deployed model

Identify **the gap** between **test dataset** and the **production deployed model in terms of metrics & scores.** Automatic or even in some-cases manual **live metric calculation** structures required

prometeia

USE CASES

prometeia

Use Cases

Bank #1

Bank #2

Bank #3

1

- **5-day ML training** to prepare the staff for technical and theoretical aspects of AI model validation.
- **AI Validation Framework mapped** to client's needs
- Validation of sample **XGBoost** model (EWS scoring) for different customer segments
- User-friendly functional libraries that contains code blocks on validation delivered.

2

- 5-day hands-on **ML training**
- **AI Validation Framework** prepared based on four pillars: Data, Methodology, Process, and Governance
- Validation of sample **XGBoost** model (EWS scoring) for different customer segments
- User-friendly functional libraries that contains code blocks on validation delivered.

3

- Validation of sample **XGBoost** model (Collection Analytics)
- 5 different customer segments simultaneously
- Thanks to the previous ML training and validation framework, a **smooth and time-efficient validation process**.

VALIDATION FINDINGS

- ML models are **more complex than traditional techniques**, hence the documentation is more complex (performance metrics, visualizations)
- Difficulties in **interpreting** model results (Shapley Analysis)
- Some important aspects:
  - Quality of data
  - Preprocessing (Missing values, encoding techniques)
  - Information leakage
  - Hyperparameter optimization
  - Overfitting/Representability
  - Using the right performance metrics
  - Model retraining

prometeia

CONFİDENTİALİTY AND CONTACTS

prometeia

Confidentiality

Contacts

Bologna
Piazza Trento e Trieste, 3
+39 051 6480911
italy@prometeia.com

Milan
Via Brera, 18
Viale Monza, 265
+39 02 80505845
italy@prometeia.com

Rome
Viale Regina Margherita, 279
italy@prometeia.com

London
Dashwood House 69 Old Broad Street
EC2M 1QS
+44 (0) 207 786 3525
uk@prometeia.com

Istanbul
River Plaza, Kat 19
Büyükdere Caddesi Bahar Sokak
No. 13, 34394
| Levent | Istanbul | Turkey
+ 90 212 709 02 80 – 81 – 82
turkey@prometeia.com

Prometeia
@PrometeiaGroup
Prometeiagroup
Prometeia

Cairo
Smart Village - Concordia Building, B2111
Km 28 Cairo Alex Desert Road
6 of October City, Giza
egypt@prometeia.com

Moscow
ul. Ilyinka, 4
Capital Business Center Office 308
+7 (916) 215 0692
russia@prometeia.com

www.prometeia.com

prometeia