

PAP_OSo2

Machine Learning Model for Risk Qualification of R Libraries

Maripaz Montero
Felipe Jimenez
Luis Diego Leal

Motivation

Open-Source adoption & challenges

- Open Source brings **opportunities** for innovation and efficiency.
- But also brings **challenges** in terms of infrastructure and **software validation**.

R Validation Hub

- Provides a framework, guidelines and tools to **implement R Libraries Validation system**.

The Opportunity

- R Validation Hub **doesn't provide guidelines** on how to measure the libraries risk.
- Most methodologies implement **traditional scoring systems**.

Our solution

- Leverage simple yet powerful **Machine Learning models** as a robust alternative to operationalize **R libraries Validation Risk Assessment**.

Features

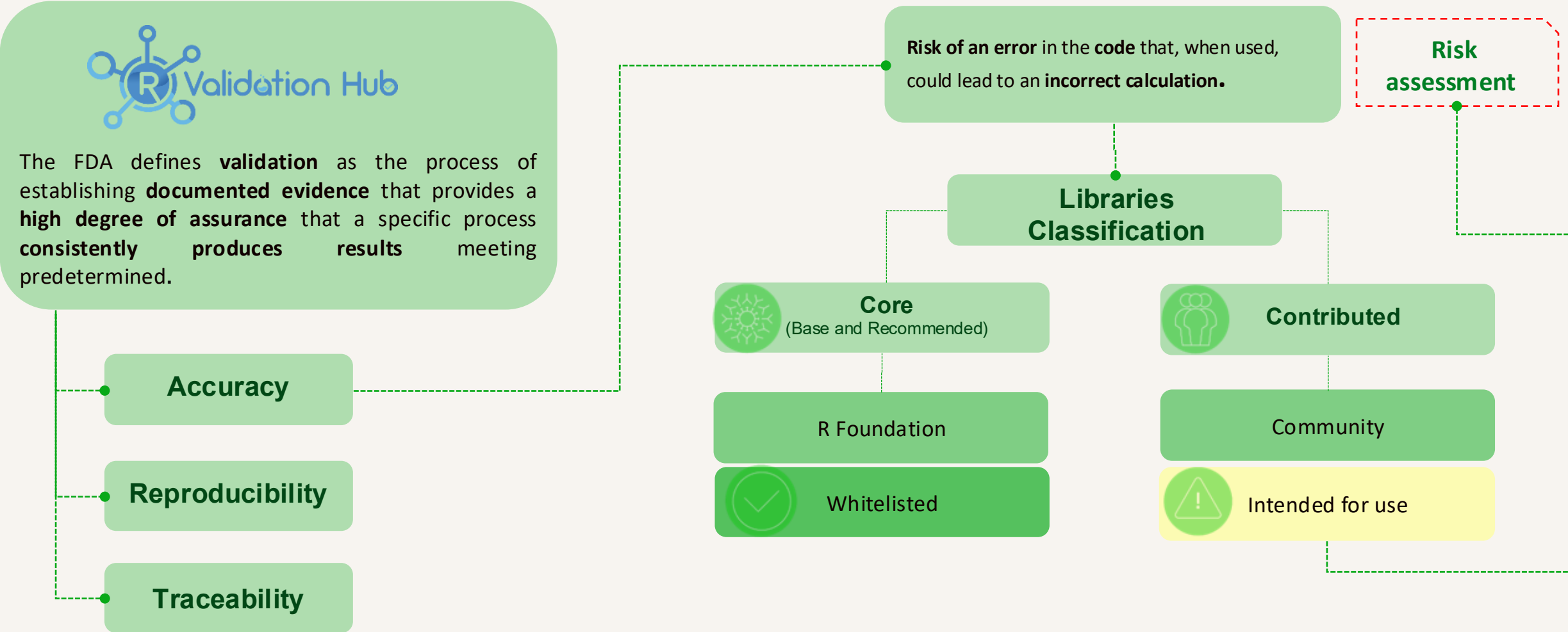
- Identifies automatically most predictive variables and assign weights.
- Dynamic, data-driven process that generates a reproducible risk score.
- Capable of identifying novel risk signals through continuous learning.

Comparison of Traditional vs. Machine Learning-Based Risk Scoring

Traditional Scoring	Aspect	Machine Learning
Expert-defined variable weights based on experience	Methodology	Learns variable importance from historical data
Results may vary across reviewers or organizations	Reproducibility	Consistent scoring when model and data are version-controlled
Limited ability to detect new or evolving risk patterns	Emerging Risk Detection	Capable of identifying novel risk signals through continuous learning
Simple to implement; minimal technical infrastructure required	Implementation Complexity	Requires data infrastructure, model training, and validation expertise

Methodology and data

R Validation Hub Revision



R Validation Hub Revision



The FDA defines **validation** as the process of establishing **documented evidence** that provides a **high degree of assurance** that a specific process **consistently produces results** meeting predetermined.

Accuracy

Risk of an error in the code that, when used, could lead to an **incorrect calculation**.

Risk assessment

Categories for Risk Assessment

Purpose

Two classifications are proposed:
Statistical or Non-Statistical.

Good Practice in Maintenance

Presence of vignettes, websites, bug tracking, source control, release frequency, and codebase size.

Testing

Presence and quality of testing practices, which support functional reliability and maintainability.

Community usage

Considers indicators such as time since first release and download volume, reflecting real-world exposure.

Operationalization with Riskmetric



The model leverages **riskmetric R library** to extract **metadata** from **contributed R libraries**.

These metrics **evaluate categories for risk assessment** (purpose, good practice in maintenance, testing & community usage).

Accuracy

Risk of an error in the code that, when used, could lead to an **incorrect calculation**.

Risk assessment

Metrics

Good Practice in Maintenance

Active maintainer

public code base

Number of dependencies

URL to report Bugs

News updated

CMD check

has news file

Proportion of Bugs solving status

Testing

Code has use examples

Code coverage

Code has vignettes

Community usage

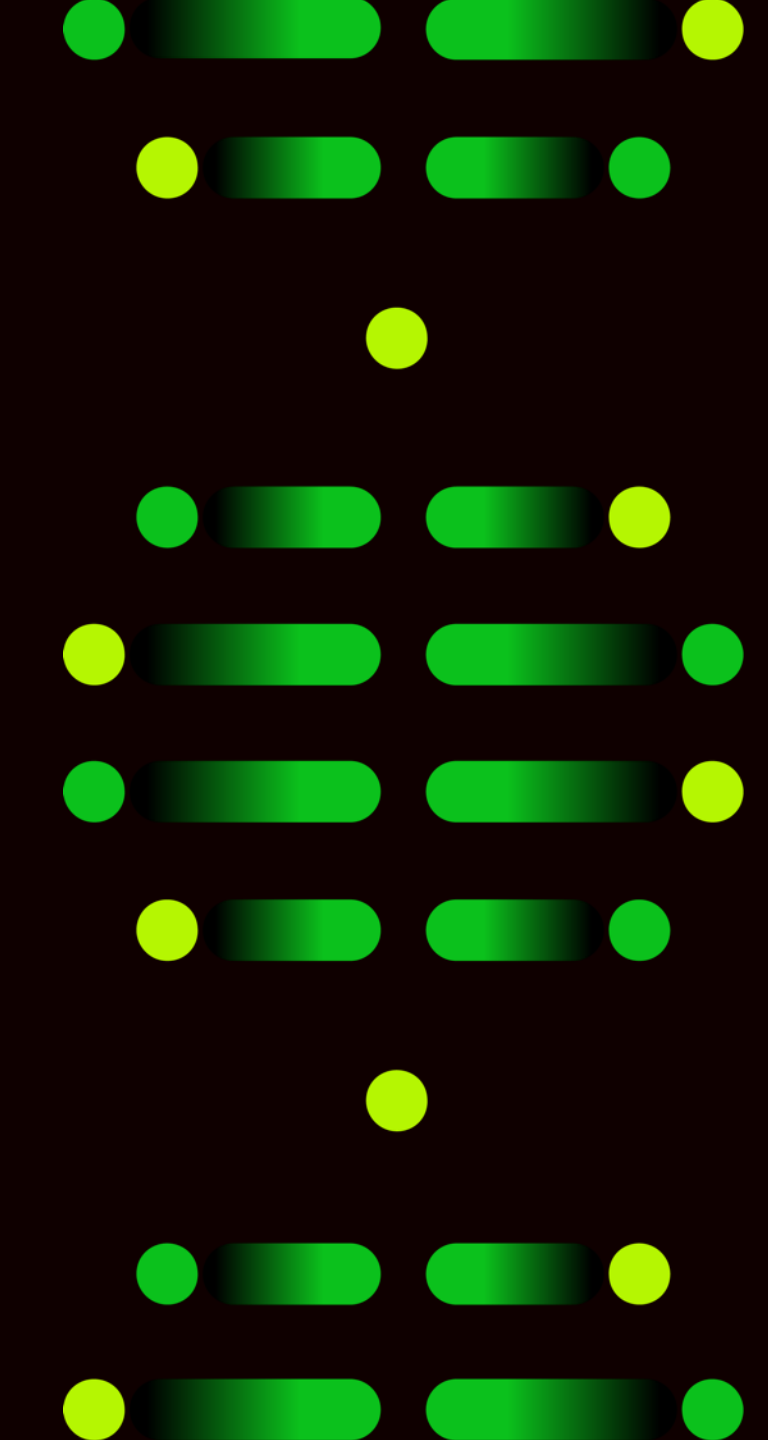
N downloads

Package has a website

Availability of documentation for exported values

continuous

binary



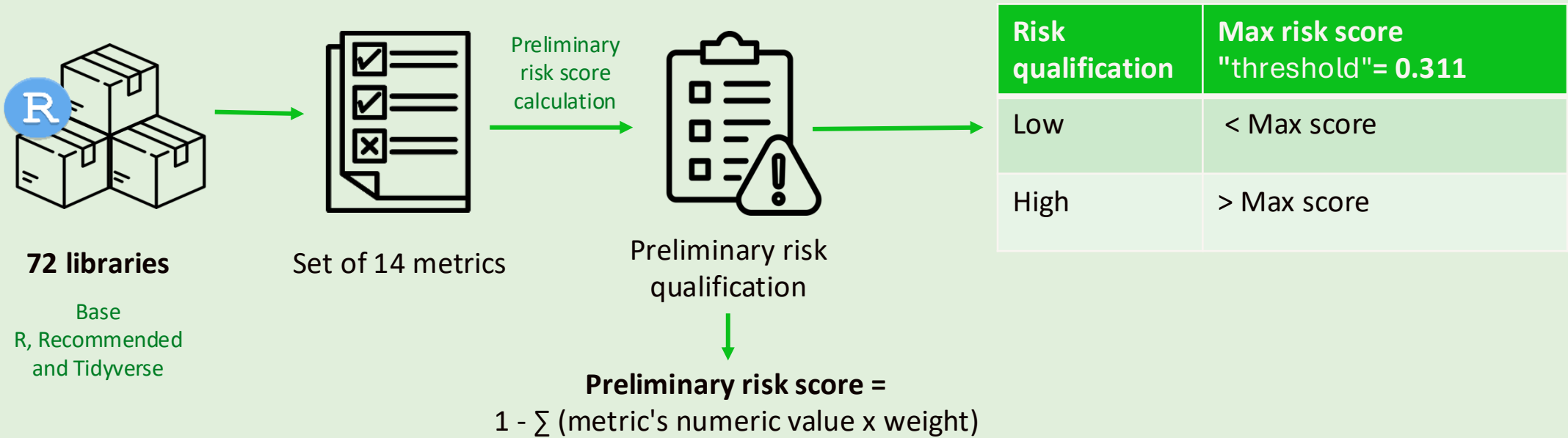
Proposed ML Framework for R Package Risk Qualification

Labcorp's Clinical Development and Commercialization Services business is now Fortrea.
[Learn more at fortrea.com](https://www.fortrea.com). For Internal Use Only. Not For Distribution.
©2023 Fortrea Inc. All rights reserved.



Model Data labeling

Stage 1. Preliminary Risk level qualification



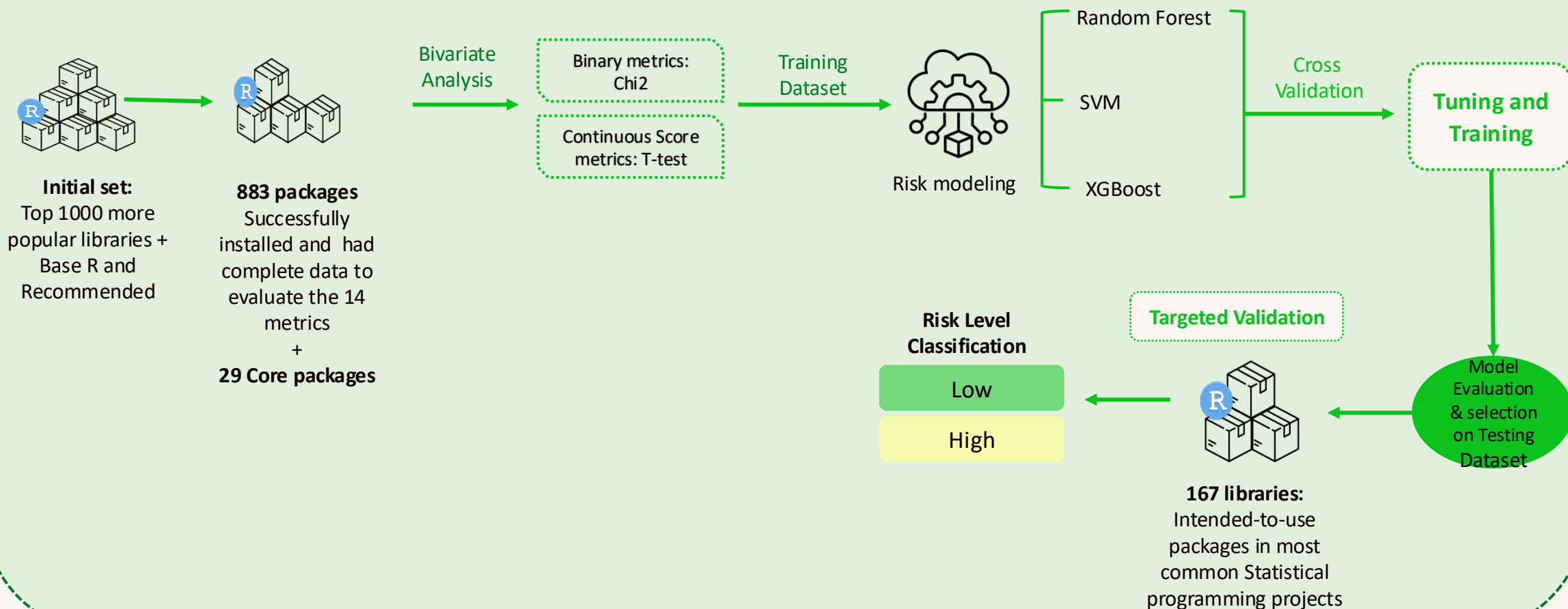
Model Data labeling

Stage 1. Preliminary Risk level qualification



Model Data labeling

Stage 2. Training and Validation datasets

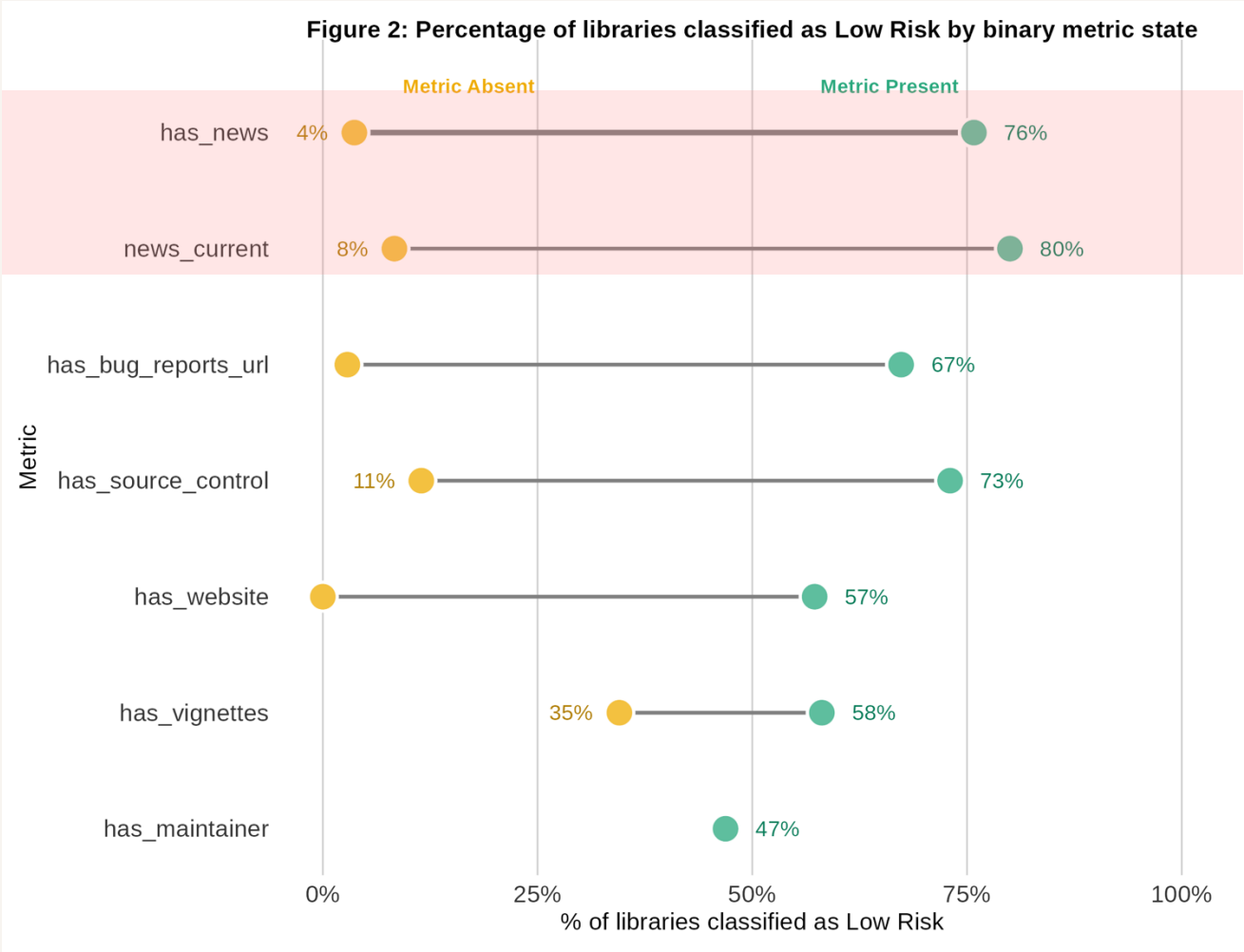
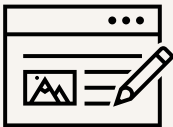


Results

Labcorp's Clinical Development and Commercialization Services business is now Fortrea.
[Learn more at fortrea.com](https://www.fortrea.com). For Internal Use Only. Not For Distribution.
©2023 Fortrea Inc. All rights reserved.



Bivariate Analysis: Binary metrics

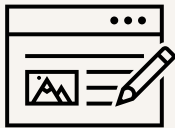


Relation between the preliminary risk score and the binary metrics:

Evaluation of the discriminative power towards the score, and therefore are likely to receive a greater weight in the machine learning models.

Most metrics demonstrated strong discriminative power. For example, 80% of libraries with an **up-to-date NEWS file** were classified as **low risk**, compared to only 8% without one.

Bivariate Analysis: Continuous metrics



Metrics such as **code coverage** and **number of downloads** showed **significant** separation between groups. (p-value < 0.05)

Model Results

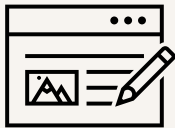


Table 3: Model Comparison metrics

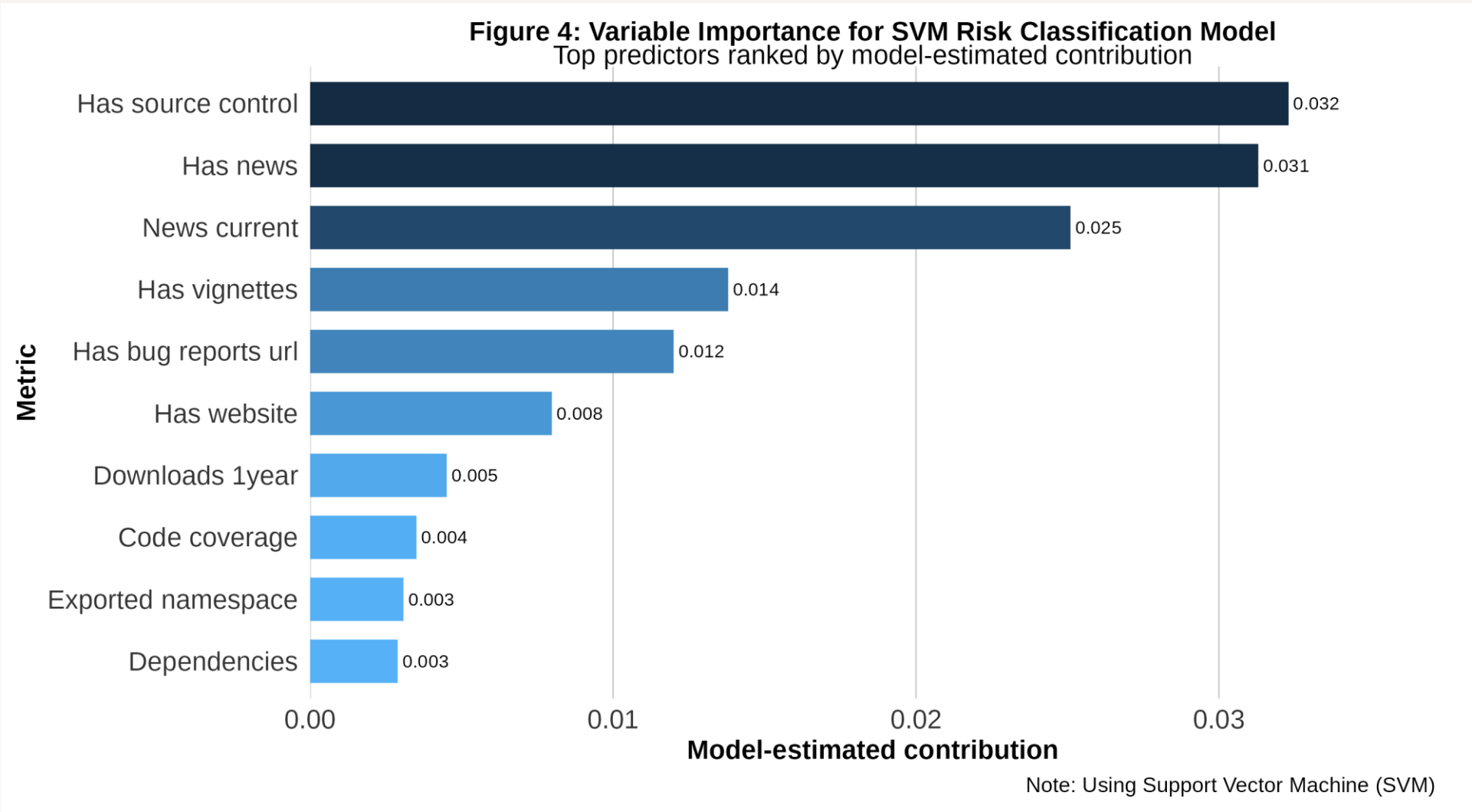
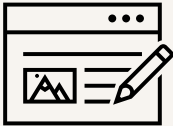
Model	Sensitivity	Specificity	Precision	Recall	F1
Random Forest	0.99	0.941	0.936	0.99	0.963
Support Vector Machine	0.99	0.949	0.945	0.99	0.967
XGBoost	0.99	0.941	0.956	0.99	0.963

Table 4: SVM Risk Libraries Model Confusion Matrix

Predicted	Actual	
	Low risk	High risk
Low risk	103	6
High risk	1	112

- 46.89% were pre-classified as **low risk** and 53.11% as **high risk**.
- All models were trained using 75/25 train-test split and evaluated via 10-fold cross-validation.
- All models demonstrated high performance, with F1 scores above **0.96**.

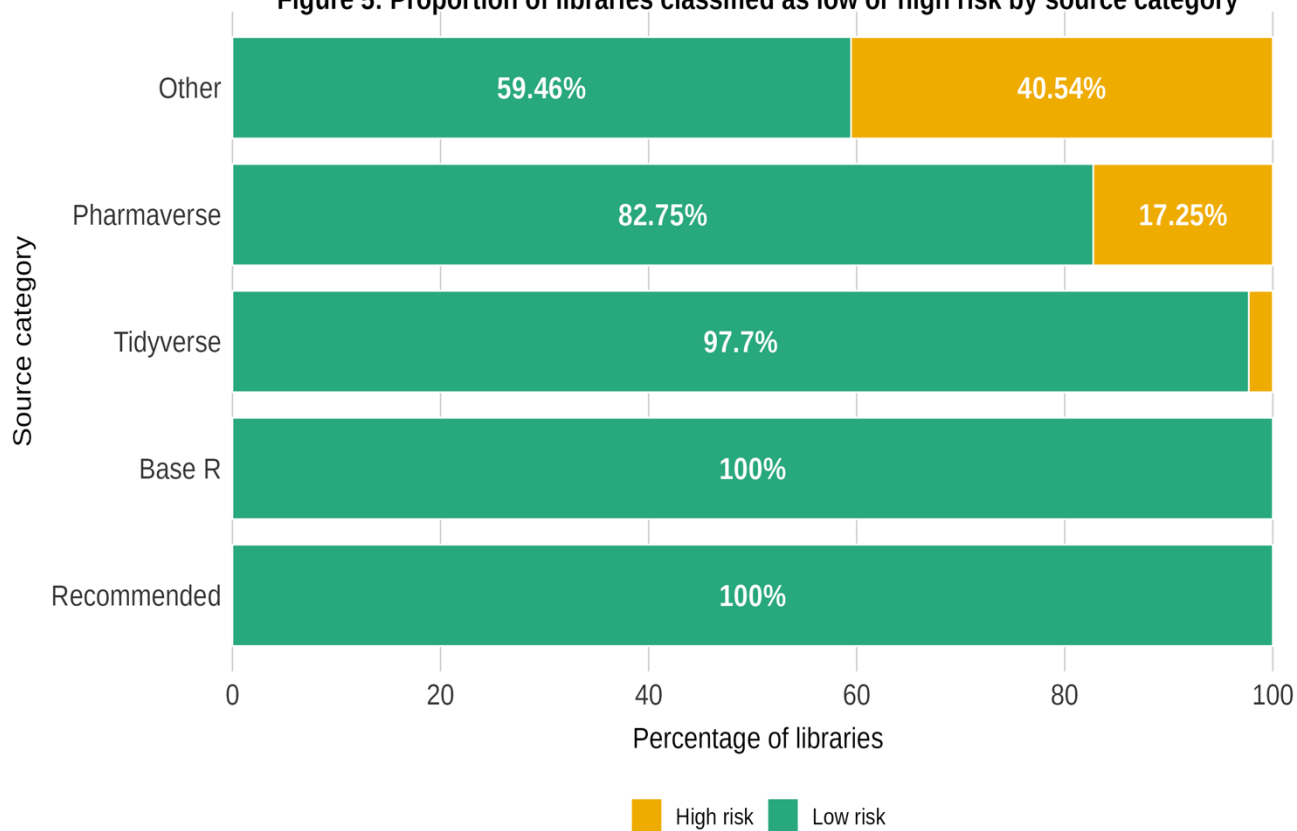
Model Results



Model Validation and implementation

External Data Validation

Figure 5: Proportion of libraries classified as low or high risk by source category



Manual certification

- For libraries **near the risk threshold**. Common causes for borderline classifications included:
 - incomplete metadata or
 - information located in sources not captured by Riskmetric.
- To address this, a **temporary “quarantine”** category was introduced to flag such libraries for further investigation before assigning a final risk classification.

Implementation and deployment



API



Shiny



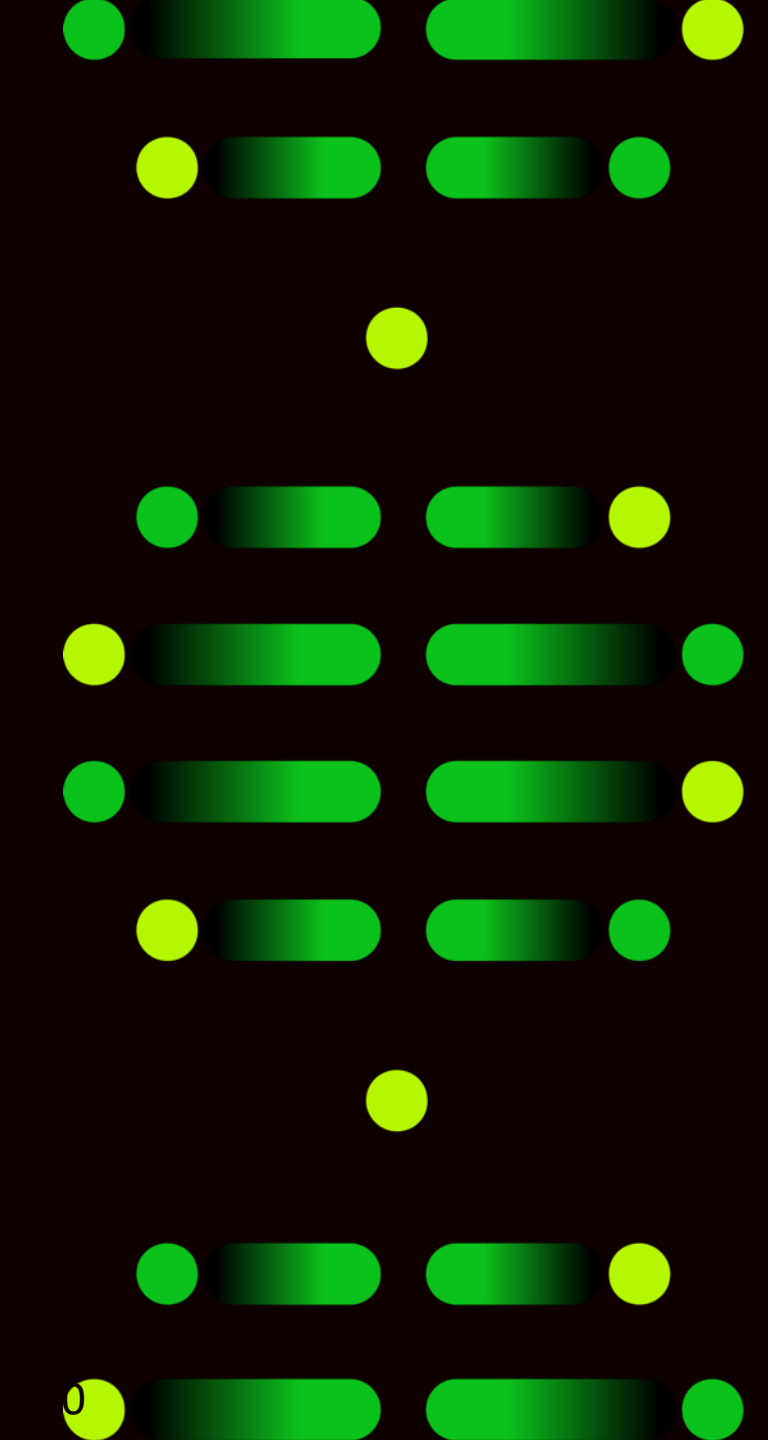
Pins

Conclusions

Labcorp's Clinical Development and Commercialization Services business is now Fortrea.
[Learn more at fortrea.com](https://www.fortrea.com). For Internal Use Only. Not For Distribution.
©2023 Fortrea Inc. All rights reserved.



Final development	Machine-learning framework to assess risk in R open-source libraries for clinical trial workflows that provides a reproducible, data-driven alternative to subjective expert-based evaluations.
High classification performance	Effectively differentiates high-risk from low-risk packages (F1 = 0.97)
Industry Regulation Standards	Integrated 14 quality metrics aligned with the R Validation Hub methodology. Ensures compatibility with existing regulatory-focused validation approaches.
Operationalizing open-source risk evaluation.	Reduces subjectivity and supports continuous adaptation as new data becomes available.
Limitations	Reliance on expert oversight, lack of public labeled datasets, and metadata gaps.
Future work	Extend to other ecosystems (e.g., Python), incorporate additional metrics, and validate with broader datasets.



Thank you

felipe.jimenez@fortrea.com
maripaz.montero@fortrea.com

Labcorp's Clinical Development and Commercialization Services business is now Fortrea.
[Learn more at fortrea.com](https://www.fortrea.com). For Internal Use Only. Not For Distribution.
©2023 Fortrea Inc. All rights reserved.

