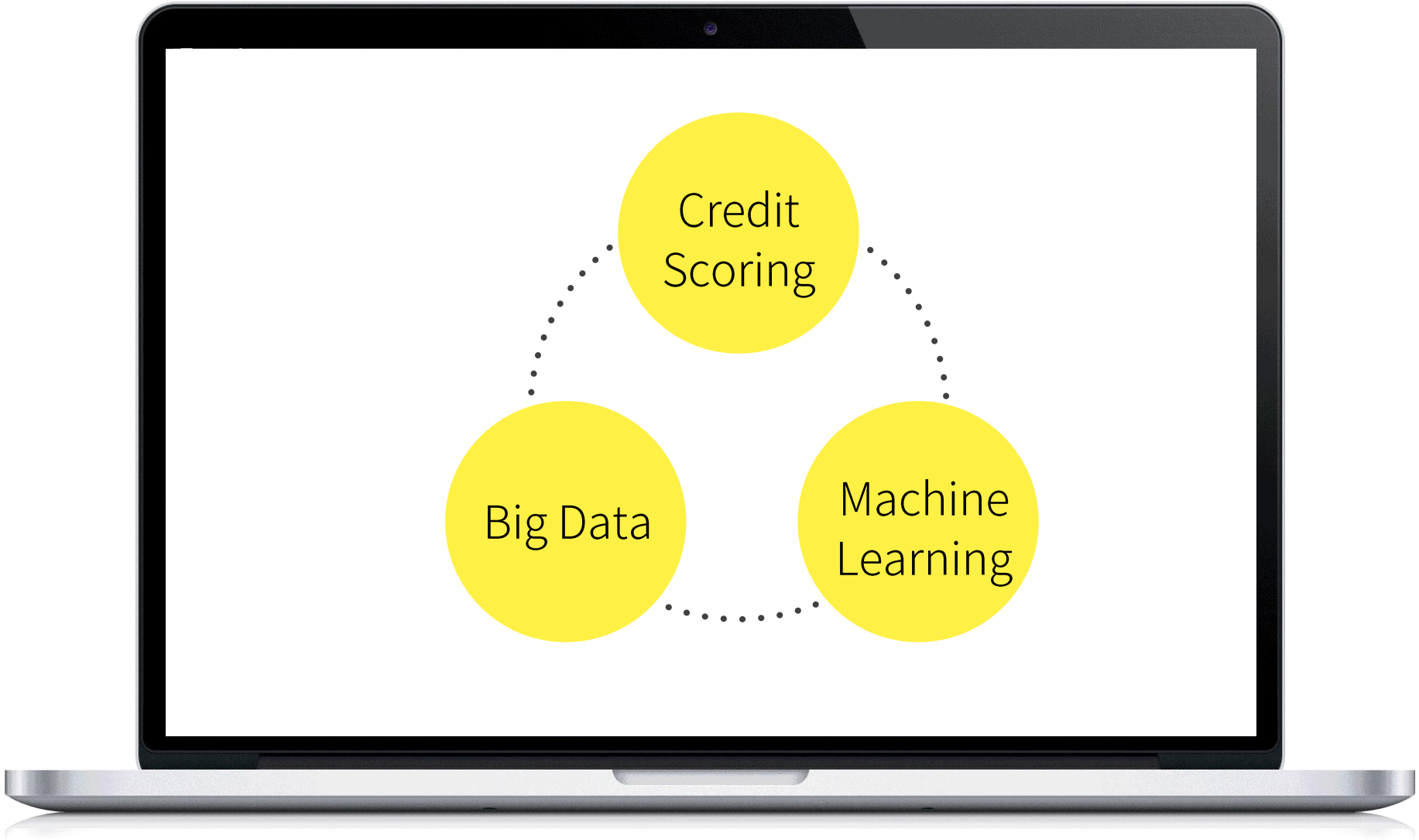
# GUIDE TO SPARK MACHINE LEARNING FOR CREDIT SCORING

ÁLVARO ORGAZ EXPÓSITO

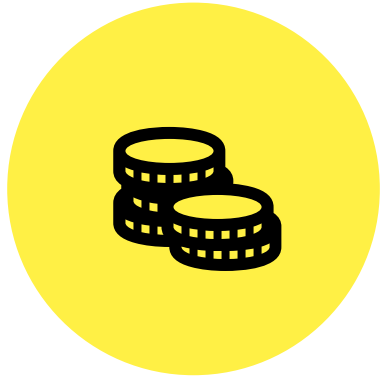
# THESIS GLOBAL PICTURE



# AIMS

Introduce the Credit Sector & Machine Learning algorithms

Real case application coded in Spark R



New point of view for the banking regulator

Create a predictive analytics guide for developing a credit score with Spark

# HYPHOTESIS

## 1 Algorithms theory

Black-box models such as neural networks or tree-based use algorithmic theories that are not as much harder to interpret than simpler models.

## 2 Real case approach

If the model complexity increases, the predictive power or accuracy is higher but the interpretability decreases. However, in small datasets, complex algorithms do not have enough data for learning and beating simpler models.

## 3 Big Data engine Spark

Develop the entire thesis with the Big Data engine Spark for the first time will be time-consuming.

# CREDIT SECTOR: Journey Map

1. Customer application declaring information
2. Pre-acceptation decision
3. Verification of declarative information (e.g., with a call center)
4. Financial decision
5. Companies track the risk of financed customers

# CREDIT SECTOR: Key points

Marketing cost balanced with financing rates

Digitalization and automatization

Risk metrics balanced with interest rates

# CREDIT SECTOR: Scoring

## How to select well the customers ?

### CREDIT SCORING

Model or method for punctuating customers assessing their potential risk. The data is mainly obtained from loan application and from credit bureaus (e.g., *Equifax*, *Experian*).

### HISTORY

It began with simple statistical models in 1950 with Cox, Fair and Isaac, evolving over the years toward complex machine learning algorithms.

# BIG DATA

The 3 V of Big Data:

1. **Volume.** The volume of data is massive and it is growing exponentially
2. **Velocity.** Fast rate at which data is received and acted on
3. **Variety.** Structured in relational database or unstructured (text, audio, video)

Typical tends for Big Data processing:

- More powerful hardware, for example, NVIDIA GPUs with Cuda Python package.
- Distribute data in clusters working in a remote server with Big Data engines such as Spark, Hadoop, Cloudera, Hive, and Amazon Web Service.

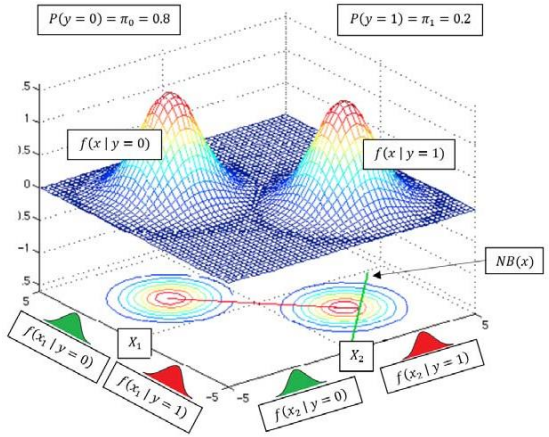
**This thesis has been coded in** APACHE Spark™

# MACHINE LEARNING: Binary classification

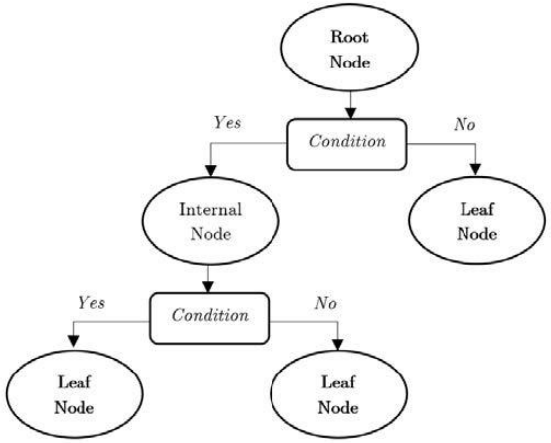
- **Supervised**
  The aim is to predict a target variable
  - Discriminative classification
    - Naive Bayes
  - Tree-based
    - Decision tree
    - Random forest
    - Gradient boosted trees
  - Neutral networks
    - Multilayer perceptron classifier
  - Generalized linear models
    - Logistic regression
- **Unsupervised**
  The aim is to explore data without a target variable
  - Principal components analysis (PCA)

# MACHINE LEARNING: Binary classification

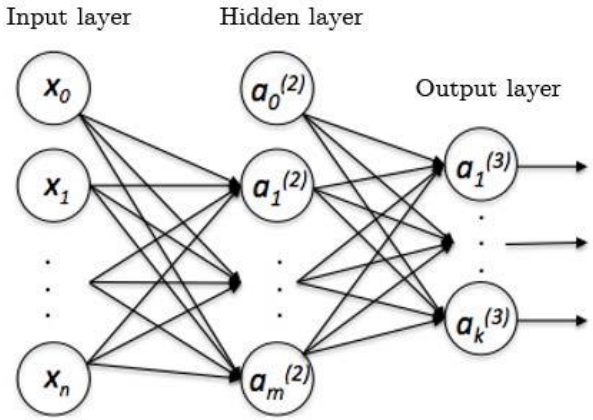
**Discriminative classification**
Naive Bayes



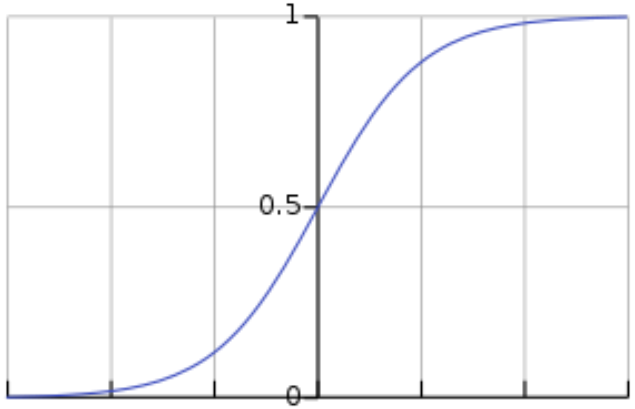

**Tree-based models**
Decision tree, Random forest, Gradient boosted trees




**Neural networks**
Multilayer perceptron classifier




**Generalized linear models**
Logistic regression



# REAL CASE

## DATABASE

3468 financed customers of a Fintech

24 features related to:
- Loan application
- Profile
- Professional situation
- Housing situation
- Expenses
- Revenues

Binary target variable:
- Defaulted or not (20-80%)

## MODEL APPLICATIONS

Credit score

Reject applications higher than a threshold

Risk categories for different pricings

Marketing campaign

Detect clusters of customers by characteristics

# REAL CASE: Protocol of model validation

## PHASE 1

Create training and test sets as well as the training folds for CV

## PHASE 2

Find the best parametrization of every model with the training set

## PHASE 3

Optimal models parametrizations with training set

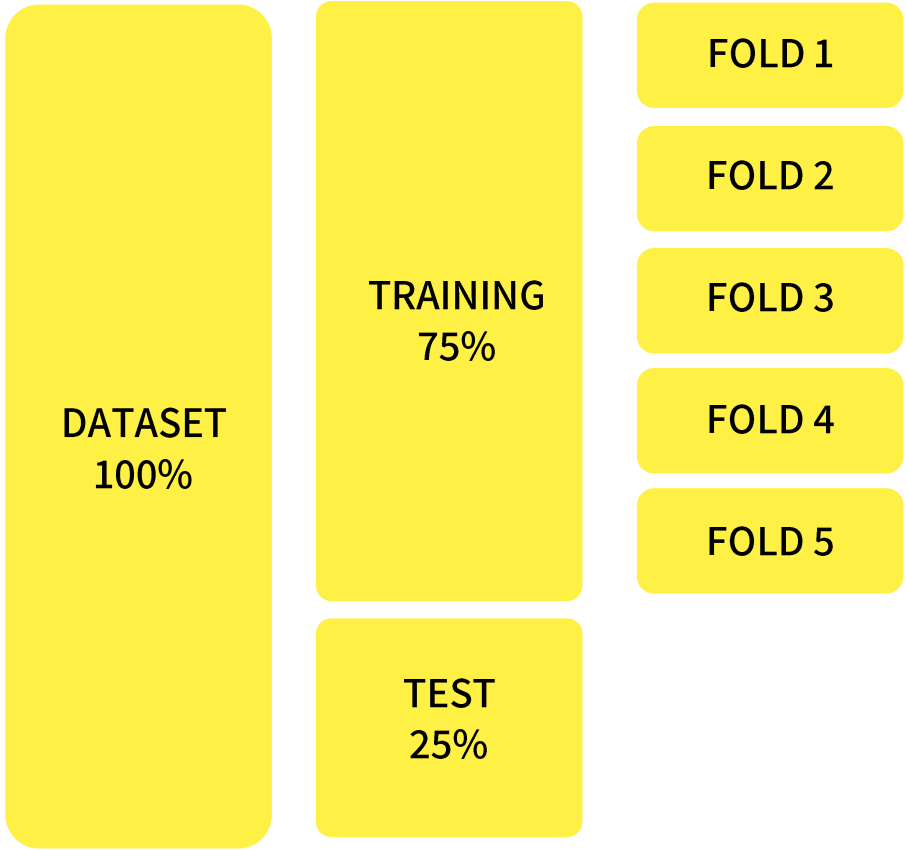
Performance measures by sets

## PHASE 4

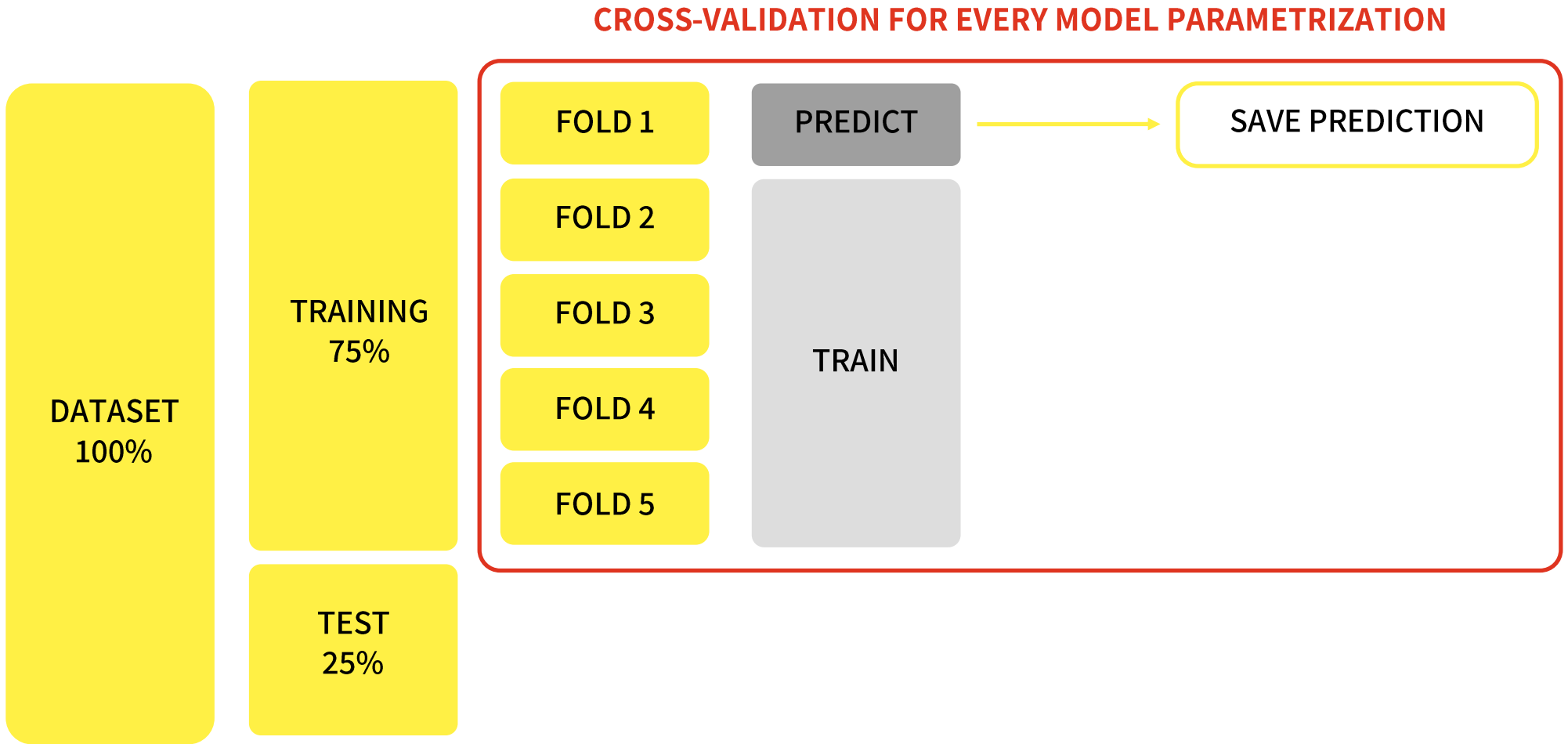
Compare models with measures

Analysis of features with PCA and feature importance

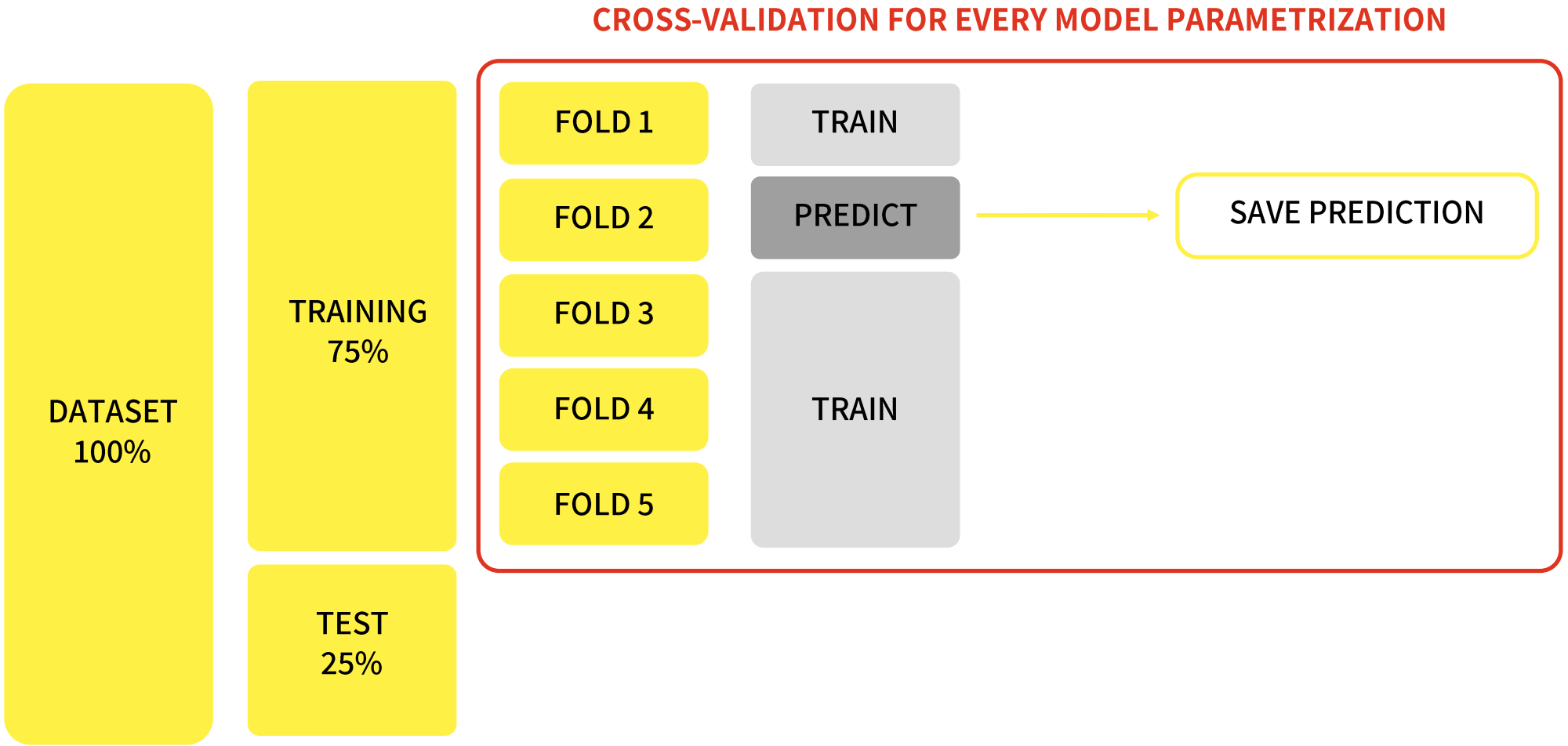
# REAL CASE: Protocol of model validation (Phase 1)



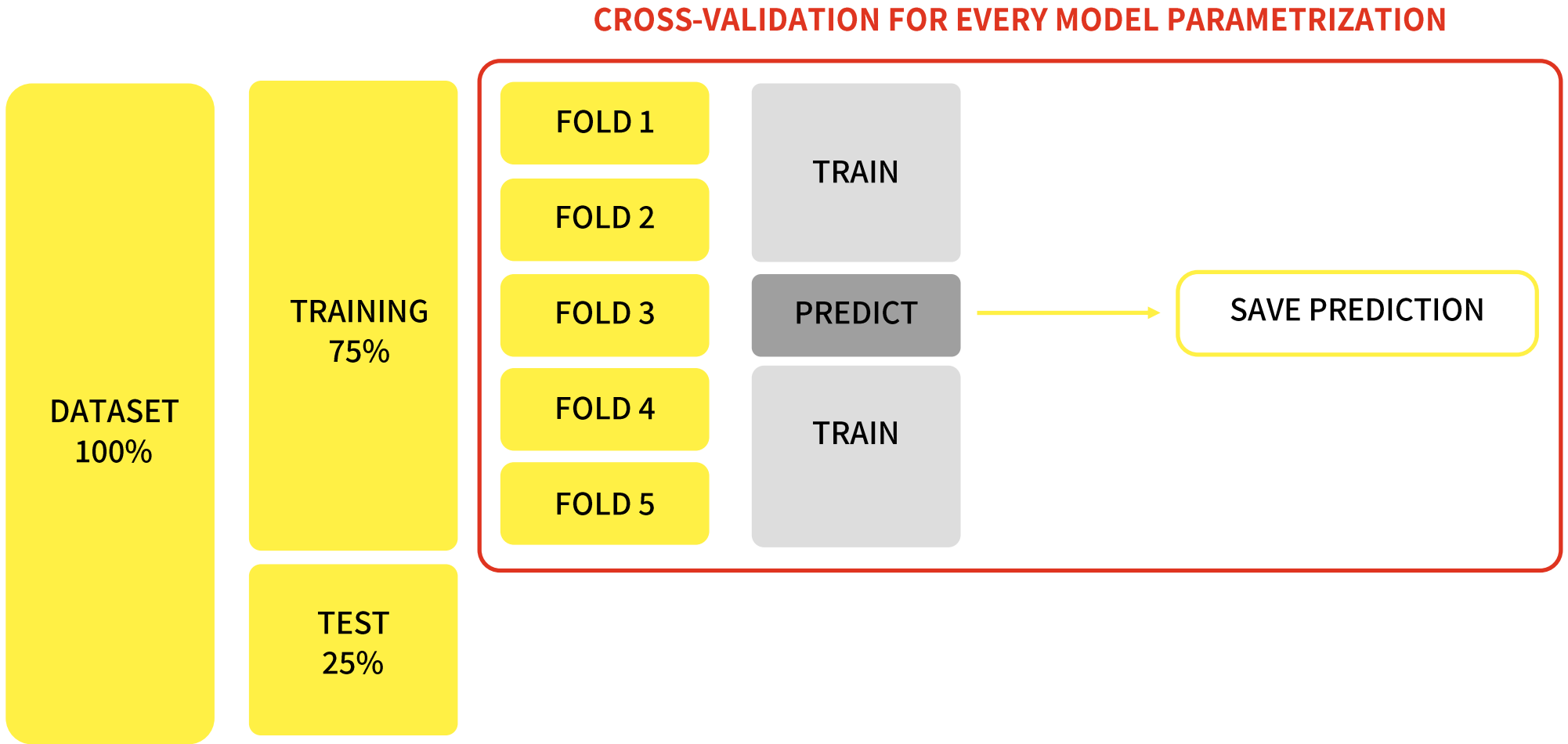
# REAL CASE: Protocol of model validation (Phase 2)



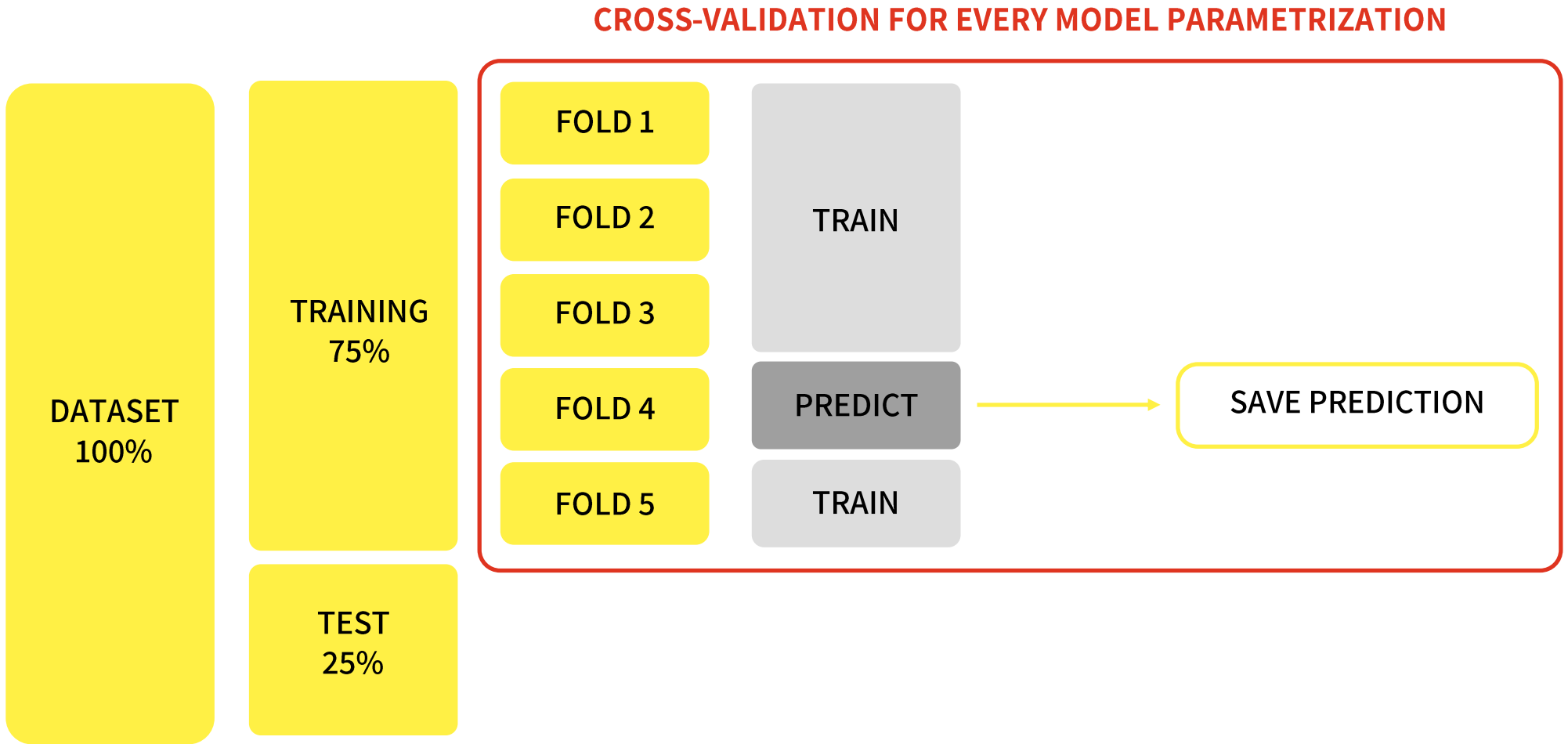
# REAL CASE: Protocol of model validation (Phase 2)



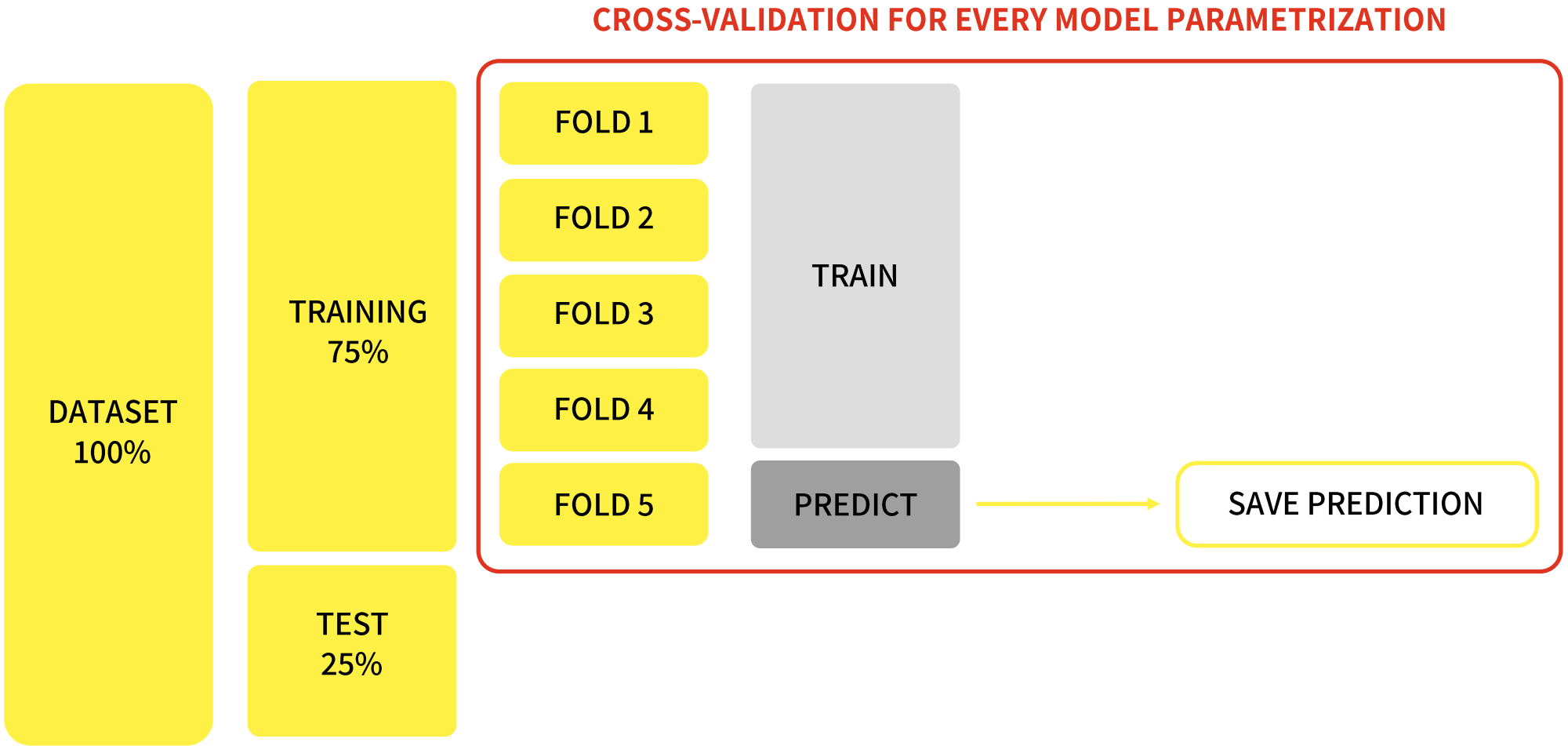
# REAL CASE: Protocol of model validation (Phase 2)



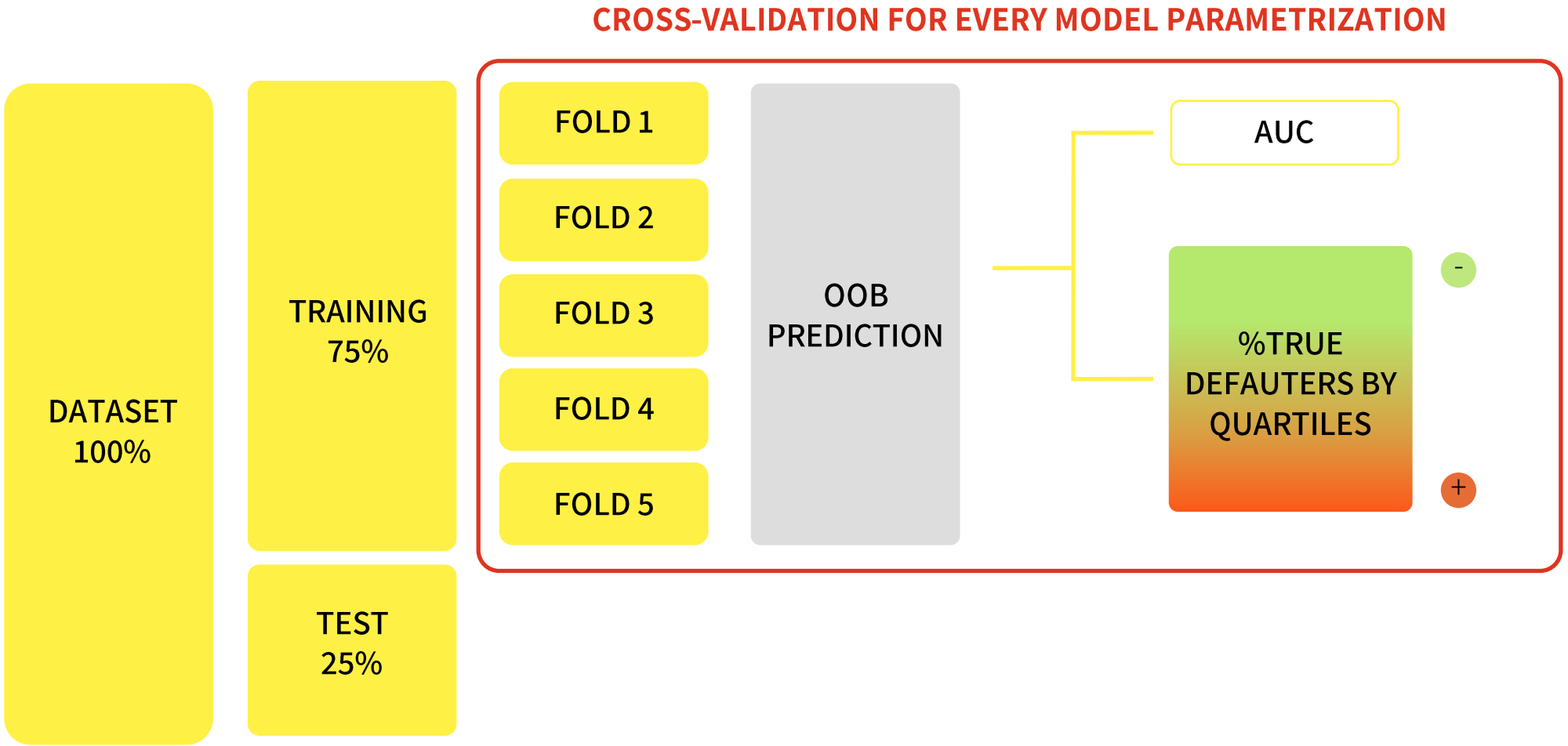
# REAL CASE: Protocol of model validation (Phase 2)



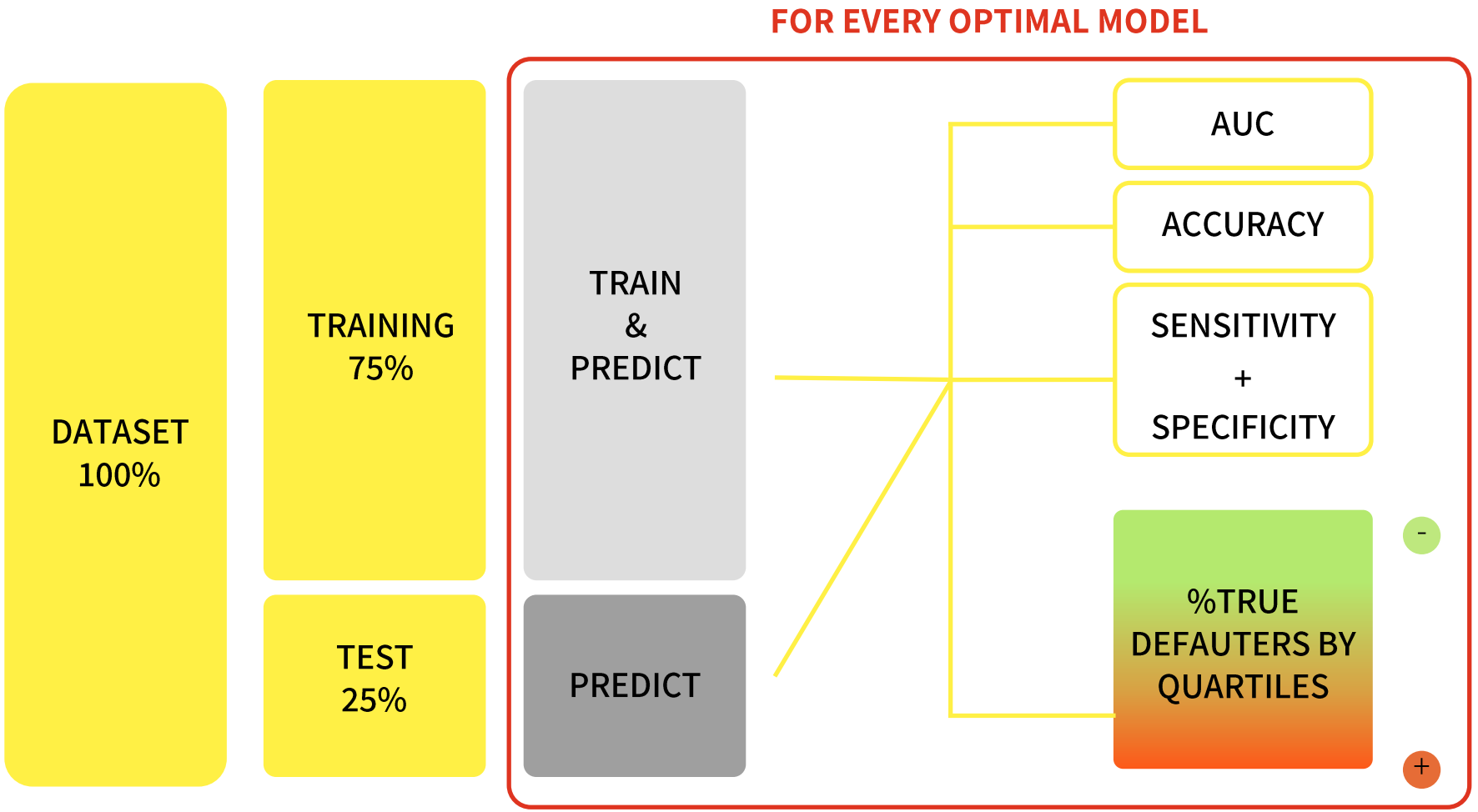
# REAL CASE: Protocol of model validation (Phase 2)



# REAL CASE: Protocol of model validation (Phase 2)



# REAL CASE: Protocol of model validation (Phase 3)



# REAL CASE: Results of phase 2

| MODEL | OPTIMAL PARAMETRIZATION | AUC BY CV | % OF TRUE POSITIVE BY QUARTILES | | | |
|---|---|---|---|---|---|---|
| | | | Q1 | Q2 | Q3 | Q4 |
| Random forest | max_bins = 30; max_depth = 5; num_tres = 30; min_instances_node=9 | 0.5904 | 0.1415 | 0.1760 | 0.2134 | 0.2706 |
| Logistic regression | all features | 0.5733 | 0.1446 | 0.1900 | 0.2181 | 0.2488 |
| Gradient boosted trees | max_depth = 10; max_iter = 15; step_size = 0.1 | 0.5714 | 0.1462 | 0.1822 | 0.215 | 0.2582 |
| Naive Bayes | categorical features | 0.5621 | 0.1711 | 0.1838 | 0.1931 | 0.2535 |
| Decision tree | max_bin = 30; max_depth = 5; min_instance = 5 | 0.5601 | 0.1562 | 0.1831 | 0.2332 | 0.2321 |
| Neural network MLPC | hidden_layer_1 = 12; hidden_layer_2 = 4 | 0.5359 | 0.1757 | 0.1905 | 0.1982 | 0.2379 |

# REAL CASE: Results of phase 3

## TRAINING SET

Clear winner: GBT. Best performance measures with 0% defaulters in Q1 and 80% in Q4.

NN is the worst model with a low AUC and illogical distribution of defaulters by quartiles.

Other models perform quite good and have similar performance.

## TEST SET

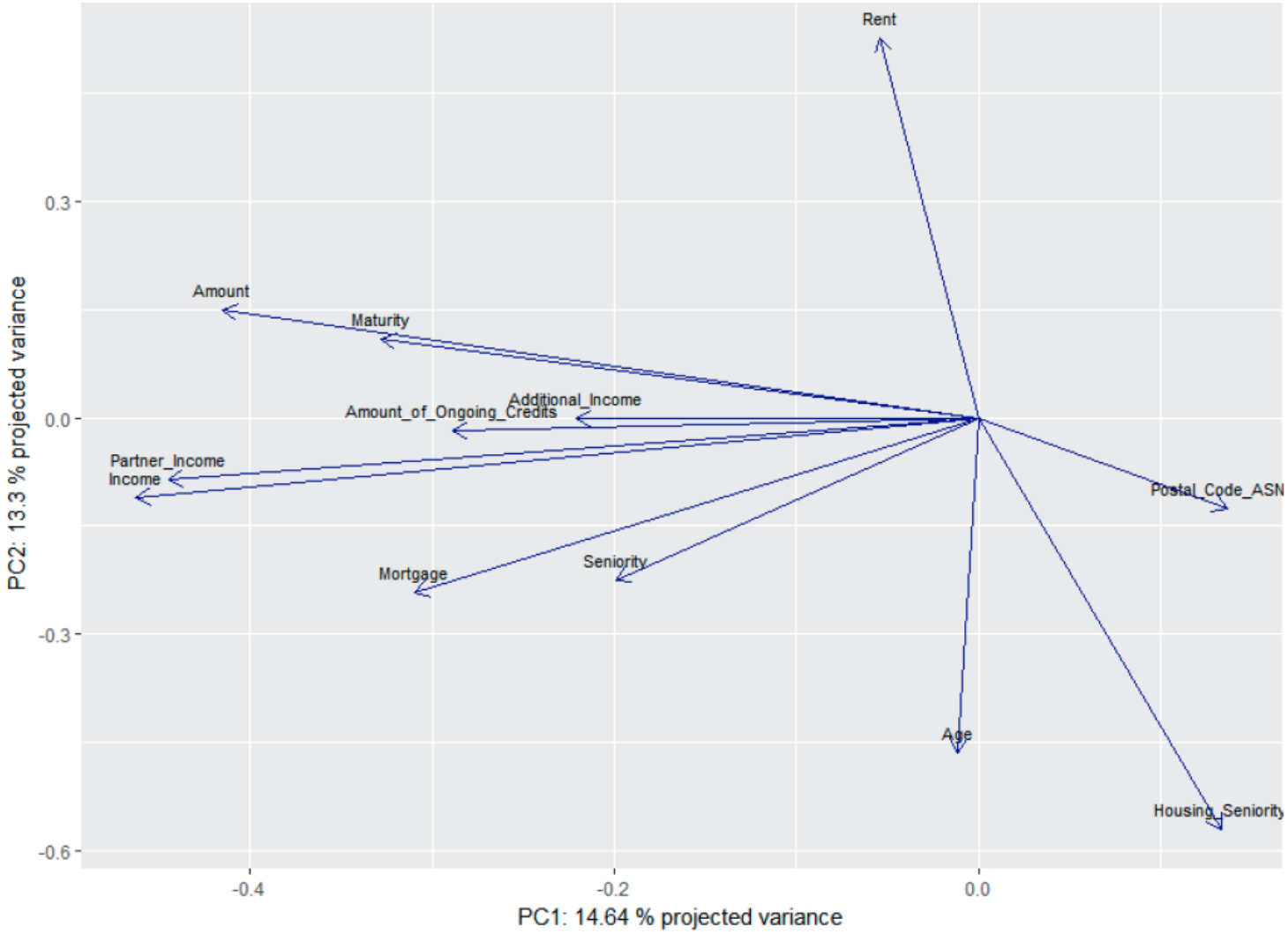
Not a clear winner, various models have similar performance.

GBT is not the best model, it seems to have been over-fitted, and NN continues being the worst model.

LR and RF have the best AUC and defaulters distribution by quartiles.

Optimal measures cut-off differ between training and test in DT & GBT.

# REAL CASE: Results of  phase 4




Analysis of features correlations with the unsupervised ML technique PCA

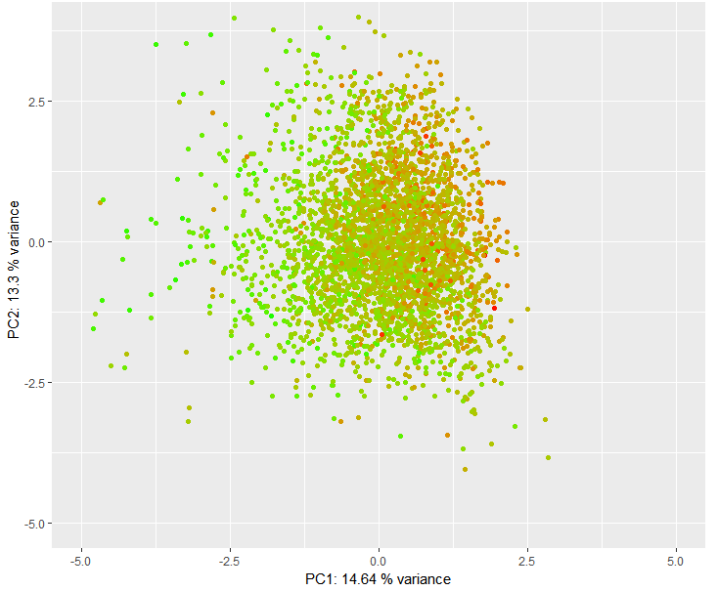
# REAL CASE:
# Results of phase 4

All models except the neural network have same patterns.
In general, the profile of a defaulter is a customer that:

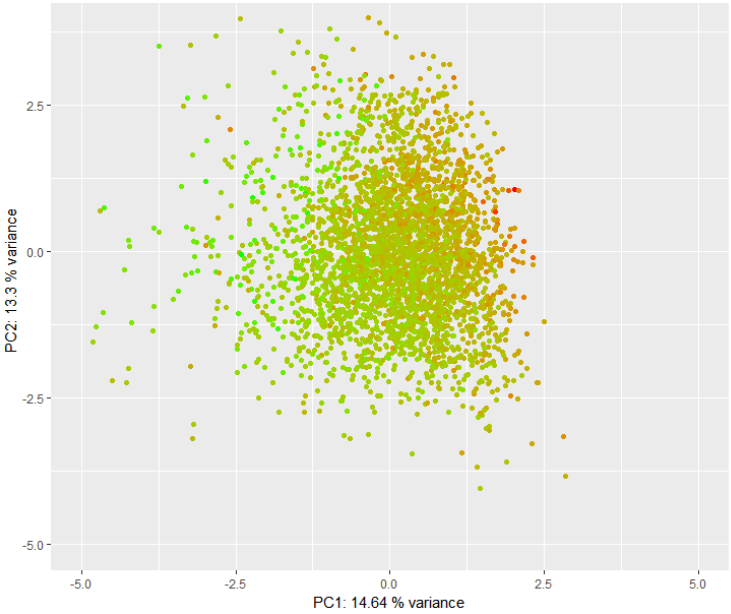
- Has a residence in a postal code with a high probability of being in ASNEF.
- Has low income, partner income, mortgage, seniority and amount of ongoing credits.
- Does not have a concrete age, monthly rent cost and housing seniority.

Nevertheless, these principal components have projected only a 30% of the global variance

PCA Projections: LOGISTIC REGRESSION




PCA Projections: RANDOM FOREST



# REAL CASE: Results of phase 4

| LOGISTIC REGRESSION | |
|---|---|
| Variable | Coefficient* |
| Postal_Code_ASNEF | 2.0552 |
| Profession_Code_MIDDLEGRADEMANAGER | 0.9489 |
| Profession_Code_ADMINISTRATIVE | 0.8904 |
| Profession_Code_OPERATOR | 0.8643 |
| Profession_Code_OTHERS | 0.7602 |
| Profession_Code_TECHNICIAN | 0.7396 |
| … | |
| Additional_Income | 0.0000 |
| Partner_Income | 0.0000 |
| Income | 0.0000 |

| RANDOM FOREST | |
|---|---|
| Variable | Importance |
| Amount | 0.1064 |
| Amount_of_Ongoing_Credits | 0.0973 |
| Age | 0.0864 |
| Housing_Seniority | 0.0570 |
| Seniority | 0.0560 |
| Income | 0.0534 |
| … | |
| Application_Week_Day_6 | 0.0024 |
| Profession_Code_MIDDLEGRADEMANAGER | 0.0023 |
| Num_Ongoing_Credits_2 | 0.0009 |

# CONCLUSIONS about hypothesis

## 1 Algorithms theory

The logic behind the building process of these models is not so complicated but a large number of iterations when optimizing the models makes them a black-box in terms of interpretability.

## 2 Real case approach

The simplest model, logistic regression, has performed as the best in the test set. But it is possible that the small dataset dimension has influenced the results because complex algorithms are the best in a competition such as Kaggle.

## 3 Big Data engine Spark

Use the coding language Spark for first time has been so time-consuming, but the Big Data engine Spark has a really user-friendly integration with statistical language R.

# **CONCLUSIONS** about aims

**5** This thesis provides a hybrid guide for machine learning with succinct theory and practice aimed to different audiences.

**6** According to the bank regulator laws, the wrong way of interpreting the results is that logistic regression performs well enough and for this reason, it would be not necessary that banking regulators attempt to be more flexible.

# **CONCLUSIONS** about thesis extension

**7** Analyse the business impact of the models, in other words, researching how applying these algorithms will impact a company in economic terms.

# THANK YOU

ÁLVARO ORGAZ EXPÓSITO

www.github.com/alvarorgaz

www.linkedin.com/in/alvaro-orgaz-exposito