FINTECH RISK MANAGEMENT
www.fintech-ho2020.eu

REGTECH WORKSHOP I

Credit scoring model development:

comparison of different models for predictive classification problems

Sabrina Stella, Fintech analyst

modefinance

European Commission

CREDIT SCORING MODEL

The goal is to build a model that **foresees company status**, active or defaulted, taking as input financial data.

Learning from data: testing algorithms for predictive modeling in **supervised** form.

- Logistic regression → parametric model
- Tree model
- Random forest

→ non-parametric models

parametric models

Train a target function *(f)* that best maps input variables *(X)* to output variables *(Y)*.

$$Y = f(X)$$

1) Select a form of the function (linear, quadratic, etc.) $y = a1 + a2*x$ → **model selection**

2) Find the best value for the function's parameters → **fitting** procedure

Training set

- How good our model is in describing the data (distance measure)
- Evaluate **residuals**

3) Use the obtained model to perform **prediction**

Test set

- What is correctly captured by the model → testing **accuracy**

linear regression

Linear regression is one of the most popular and widely used modeling approach in social sciences.

- Very robust;
- Provides a basis for more advanced empirical methods;
- Transparent and relatively easy to understand;
- Useful for both descriptive and structural analysis.

The contents here presented are strictly confidential and intended for educational purposes only. It is strictly forbidden to share any part of this presentation with any third party, without a written consent of the sender.

linear regression

The simple linear regression model (the predictor is one dimensional $(X = X_1)$).

$y(X)$ is assumed to be linear, hence

$$y_i = \alpha + \beta X_i + \varepsilon_i$$

where
α = intercept parameter
β = slope parameter
y_i = dependent variable
X_i = independent variable
ε_i = error term

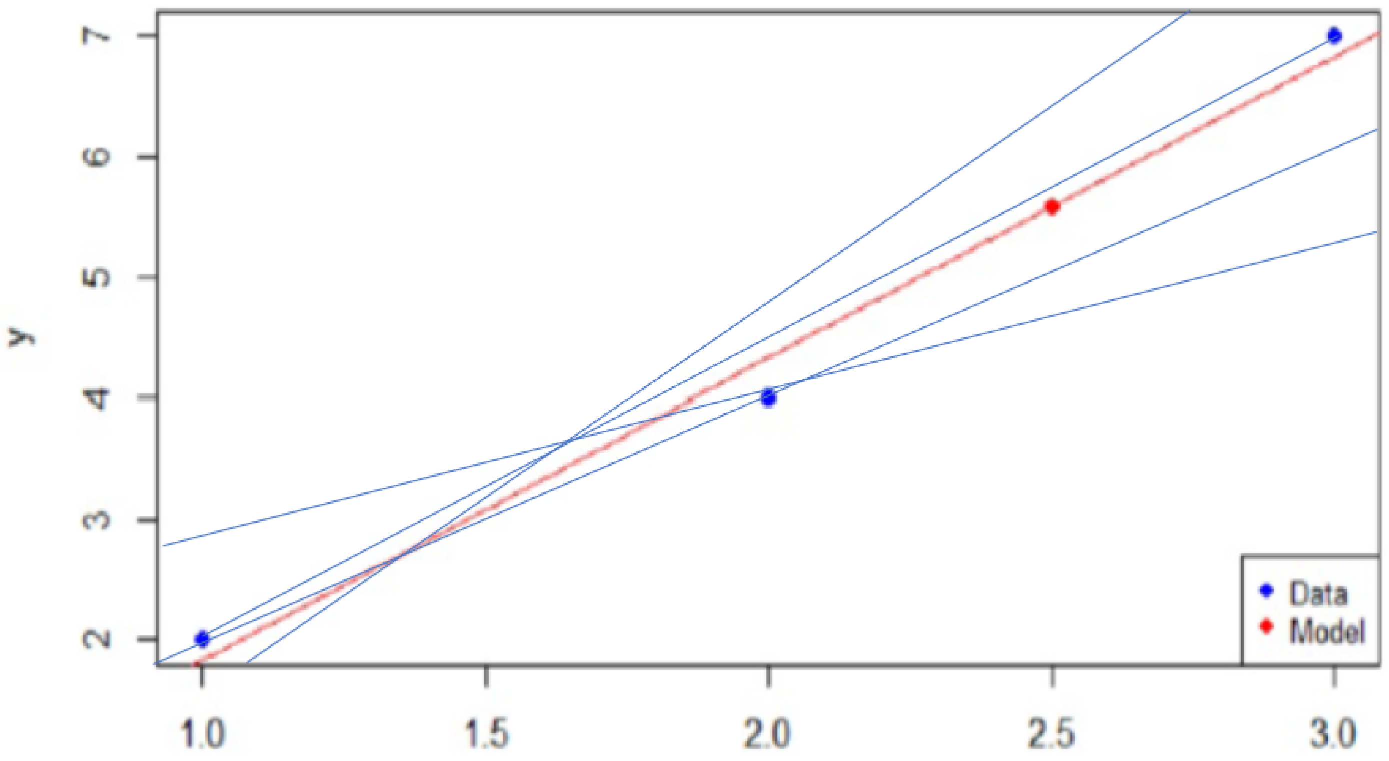
Estimated Regression Line:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} X_i + \varepsilon_i$$

graphical representation

Given the list of data: (1,2), (2,4), (3,7)

Predict a value for *x=2.5*

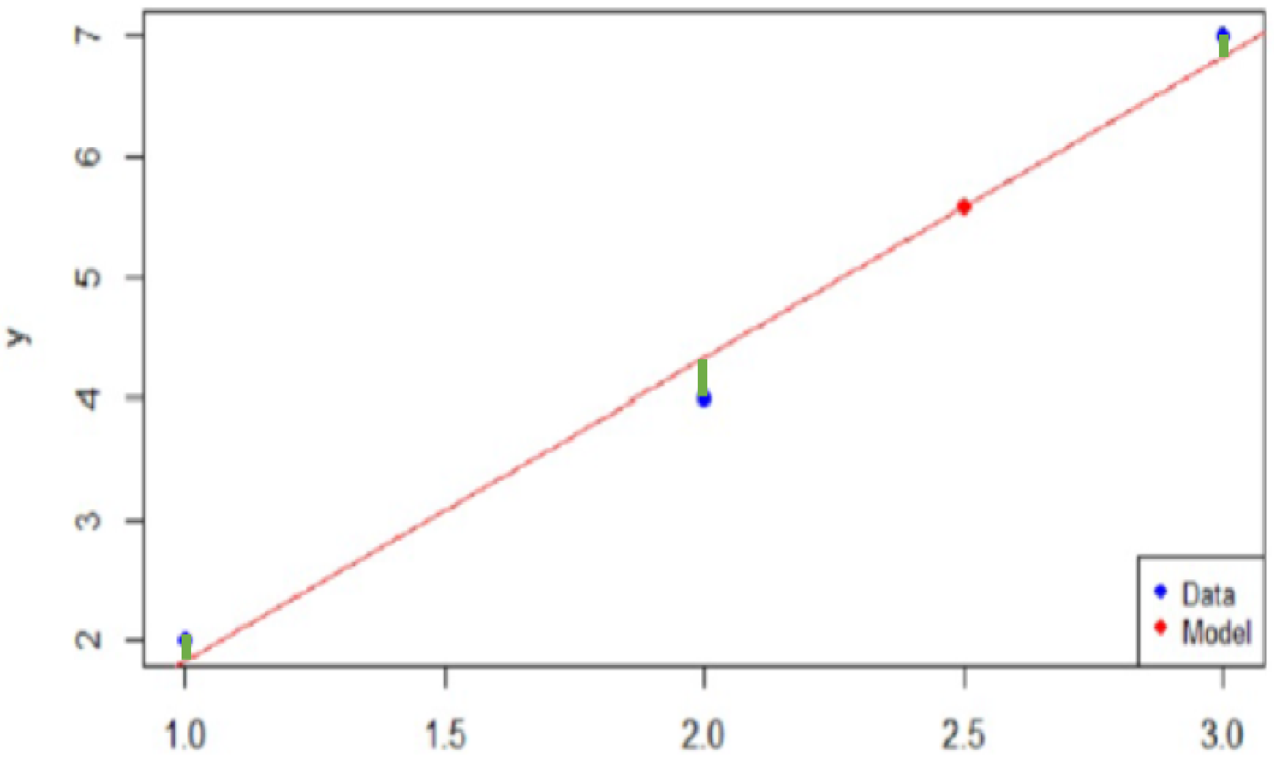


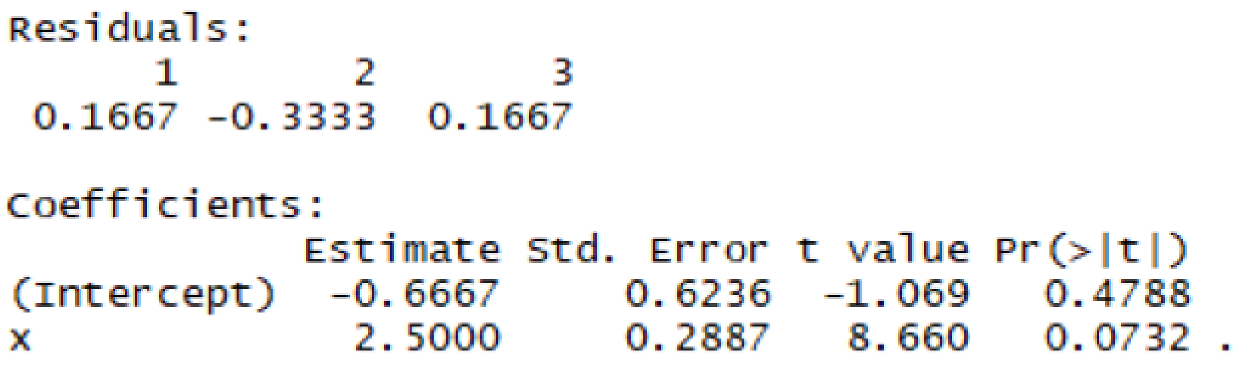
Residual: the difference between the real data and the predicted value.

$$e_i = y_i - \hat{y}_i .$$

It's important to find the best fit.

Linear regression result





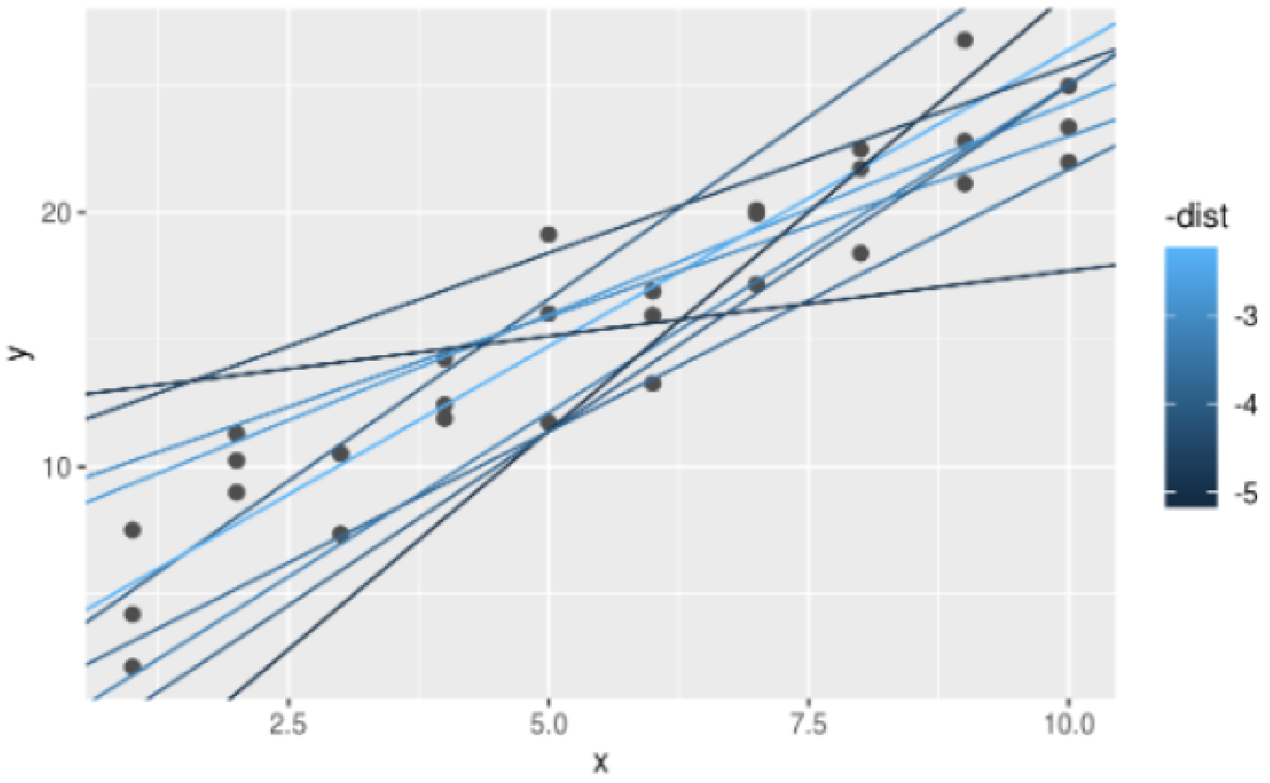
```
Residuals:
      1       2       3
 0.1667 -0.3333  0.1667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6667     0.6236  -1.069   0.4788
x             2.5000     0.2887   8.660   0.0732 .
```

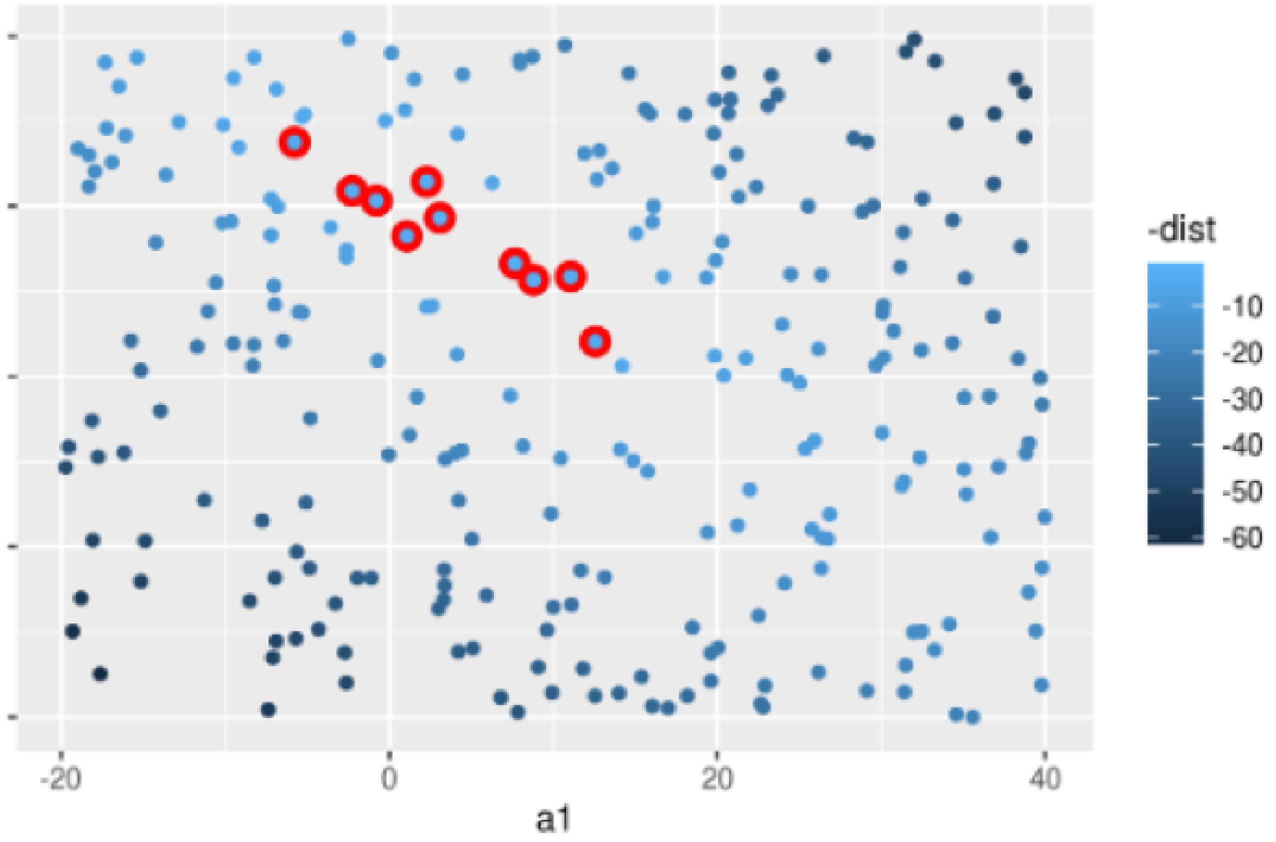
The contents here presented are strictly confidential and intended for educational purposes only. It is strictly forbidden to share any part of this presentation with any third party, without a written consent of the sender.

least squares method - best fit

The method of least squares is often used to find the best curve to fit the data. It chooses the first degree polynomial that **minimizes** the sum of the squares of the errors of the fit (SSE), defined by:

$$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$





The contents here presented are strictly confidential and intended for educational purposes only. It is strictly forbidden to share any part of this presentation with any third party, without a written consent of the sender.

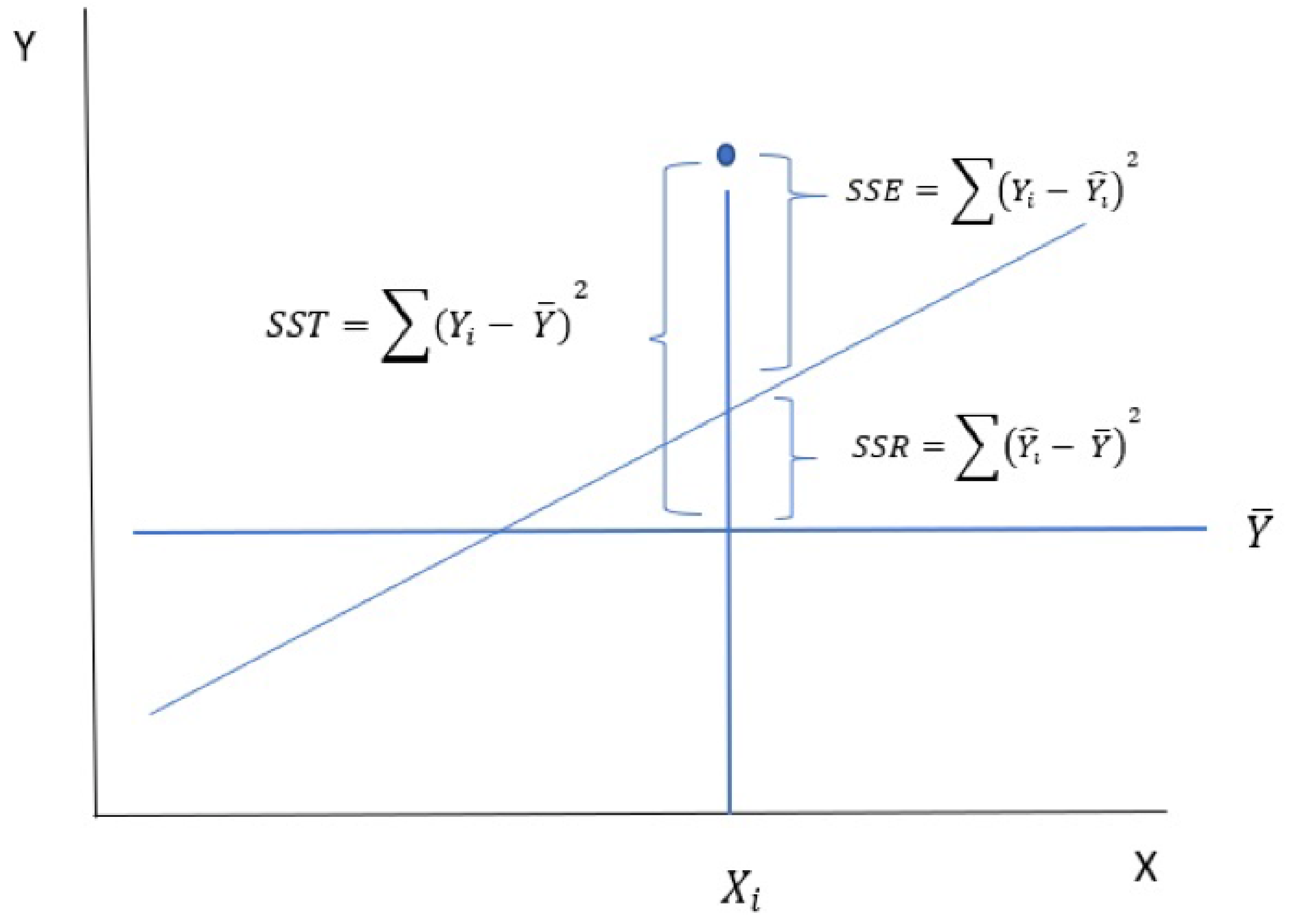
model fit evaluation

To compute the goodness-of-fit statistic R square is a commonly accepted metric, based on residuals analysis. The closest R is to 1, the better the fit.

R-squared:

$$R^2 = 1 - \frac{SSE}{SST}$$

Sum of Squares Regression (**SSR**) is the sum of the squared differences between the predictor for each observation and the population mean.


Y
$SSE = \sum(Y_i - \hat{Y}_i)^2$
$SST = \sum(Y_i - \bar{Y})^2$
$SSR = \sum(\hat{Y}_i - \bar{Y})^2$
$\bar{Y}$
X_i
X

model fit evaluation

The goodness of fit can also be investigated through the *F* test.

$$F = \frac{R^2}{(1-R^2)/n-2} = t^2$$

In a practical sense, higher R^2 will correspond to high F statistic providing evidence that the estimated coefficient is different from zero.

multiple linear regression

Multiple linear regression means that we have a model with *n* independent predictor variables **x_1, x_2,..., x_k** (input) and one response variable **y** (output).

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$$

If we have *k* observations of *n* predictors:

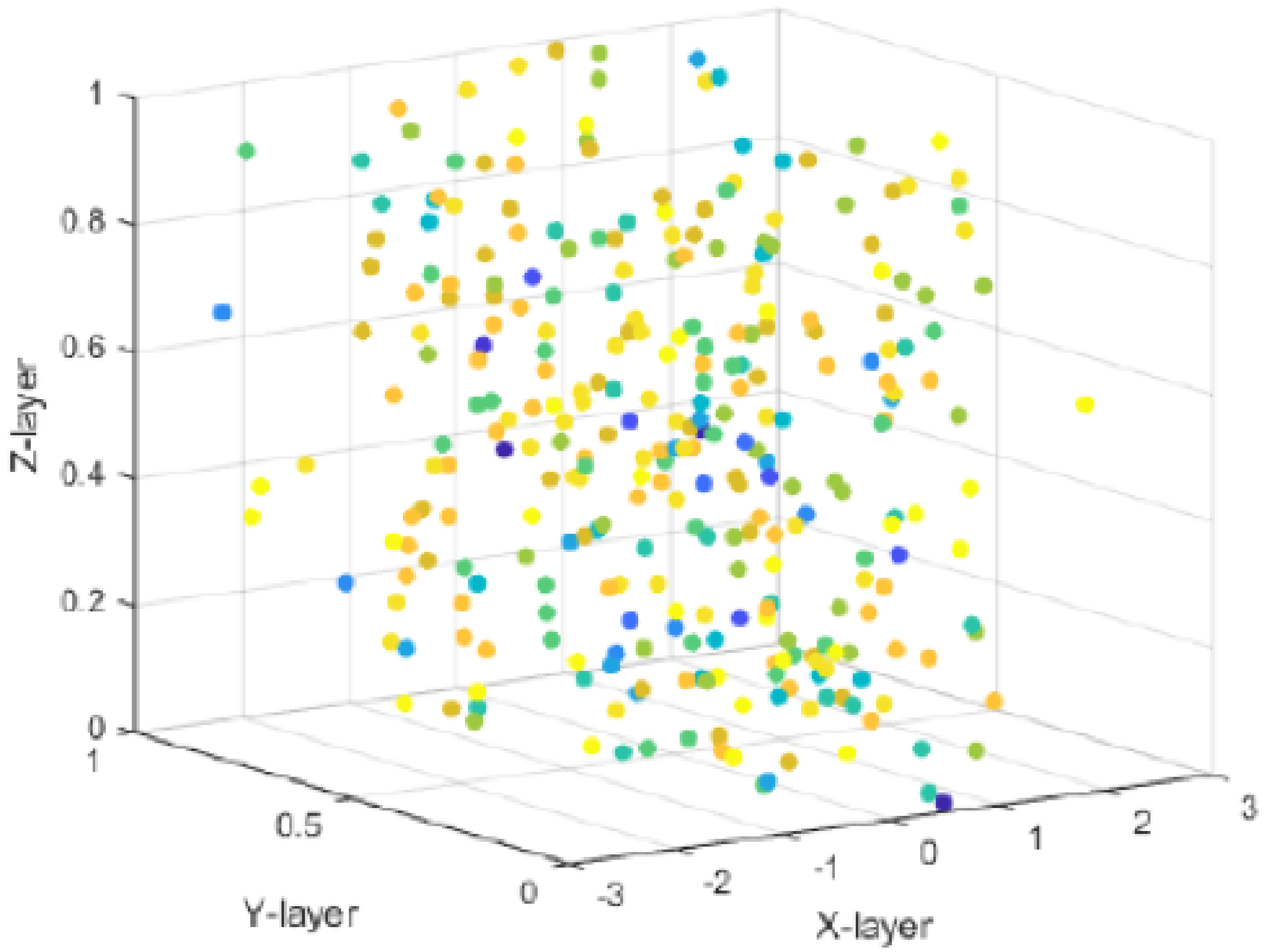
$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} \qquad i = 1 \ldots . n$$

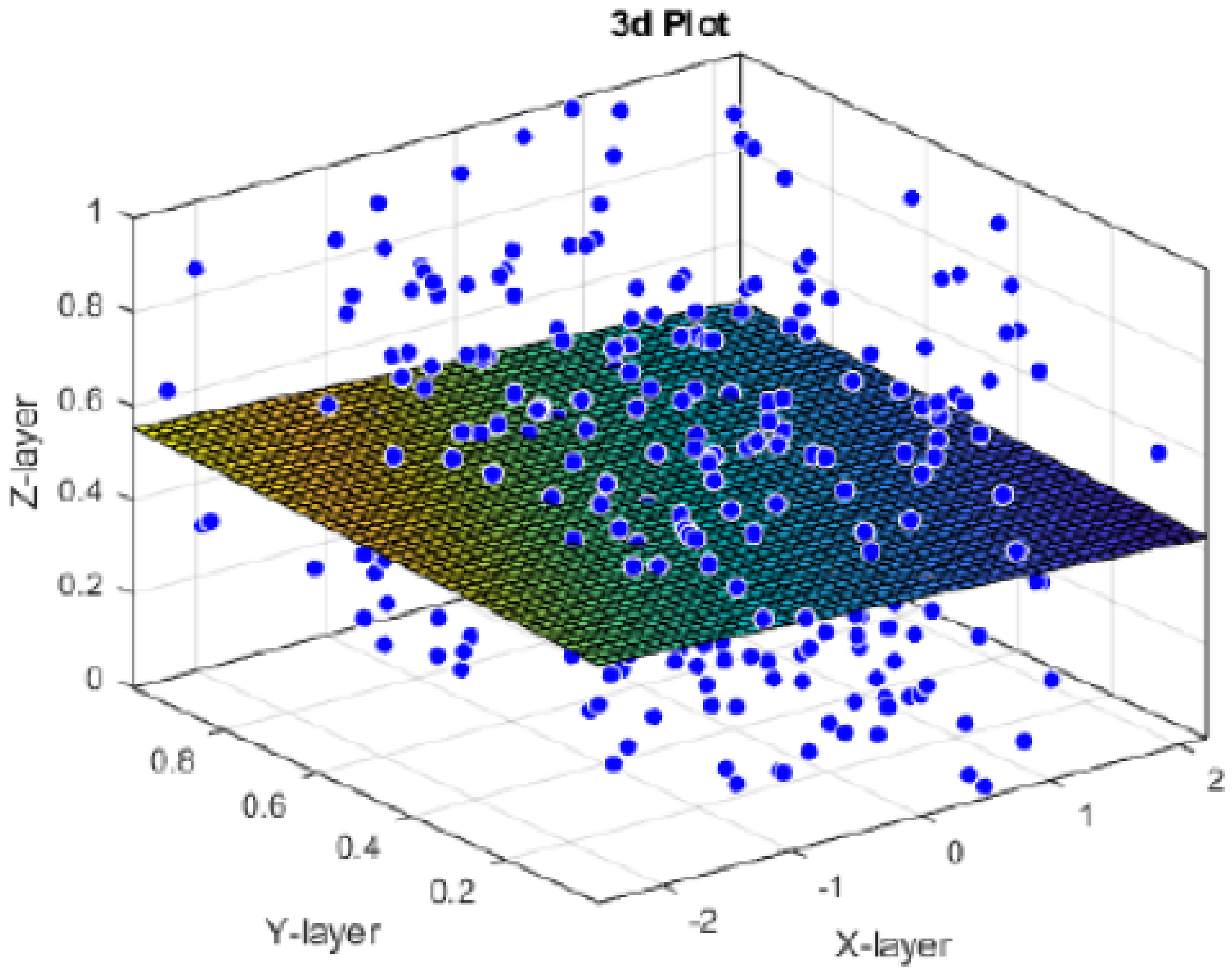
Least-squares regression, this time, turns into fit the hyper-plane in to k+1 dimensional space that minimize the sum of squared residual:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \alpha - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

The contents here presented are strictly confidential and intended for educational purposes only. It is strictly forbidden to share any part of this presentation with any third party, without a written consent of the sender.

graphical representation





The contents here presented are strictly confidential and intended for educational purposes only. It is strictly forbidden to share any part of this presentation with any third party, without a written consent of the sender.

MODEL SELECTION: LOGISTIC REGRESSION

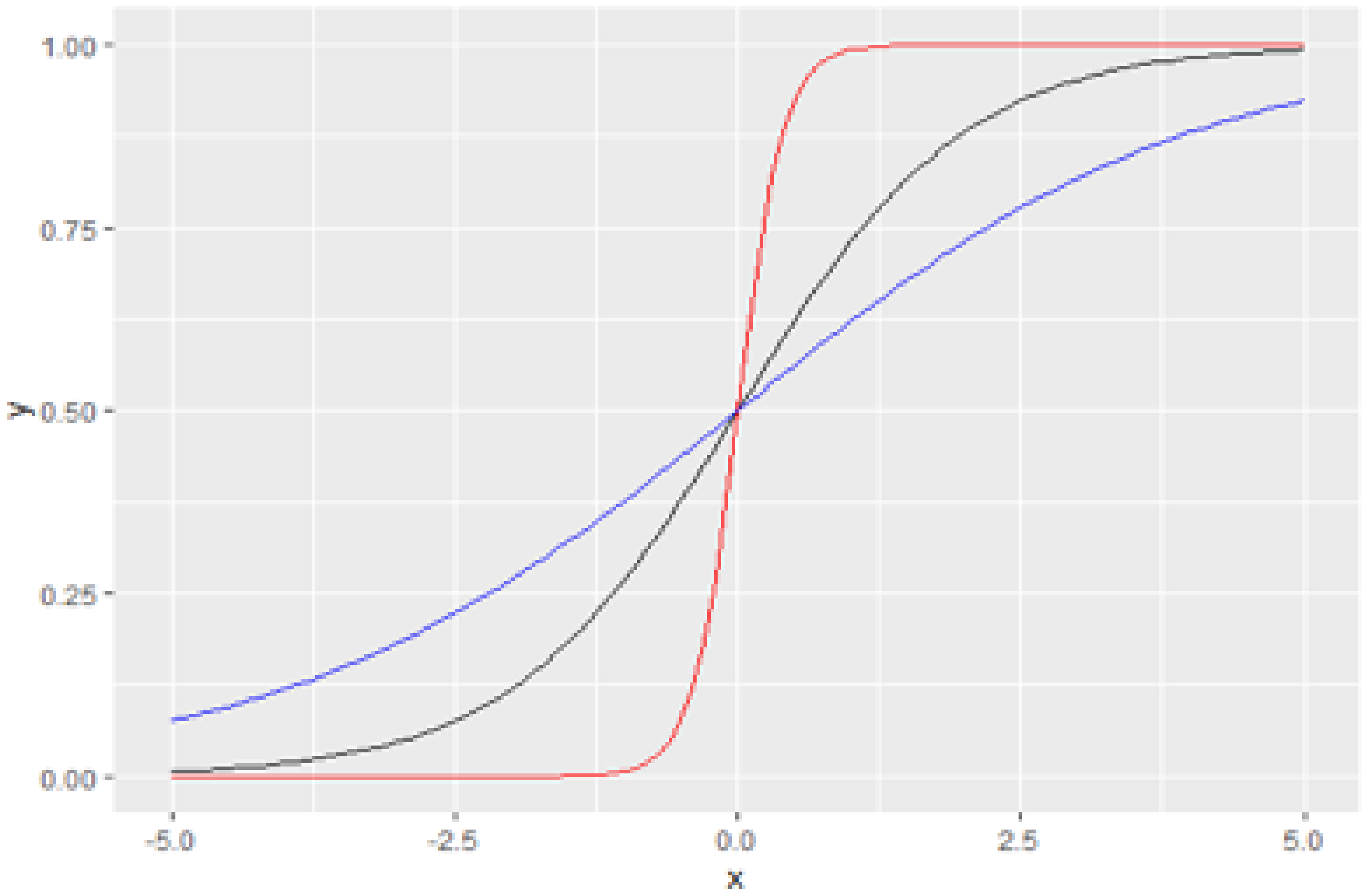
When the response variable y is binary, i.e. it gives a result *"0"* or *"1"*. It can be modeled using a logistic regression, the **logit function** is a linear combination of the independent variables.

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$$

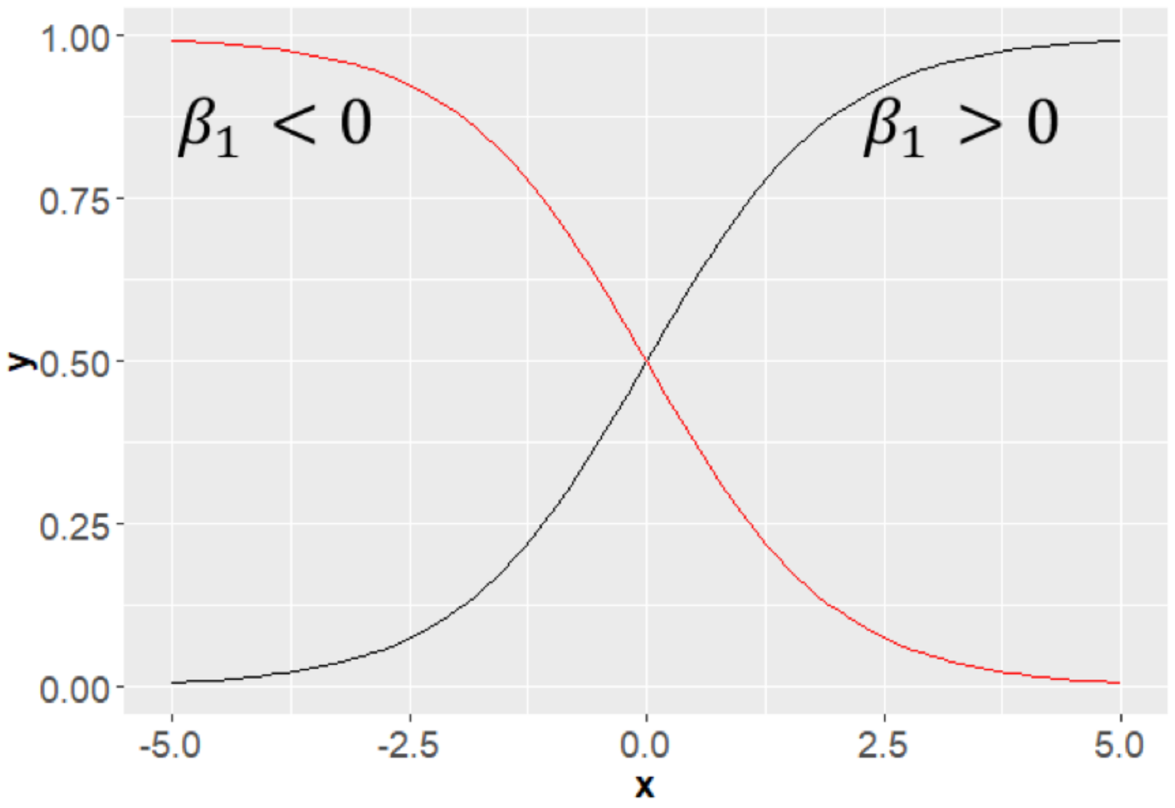
- π = P(y = 1) is the probability of "success" for a given event(in our case, default).
- $Odds_1 is\ defined\ as = \frac{\pi}{1-\pi}$

The probability *π* is a non-linear combination of the explanatory variable, according to a sigmoid function – consider the single explanatory variable case, for simplicity:

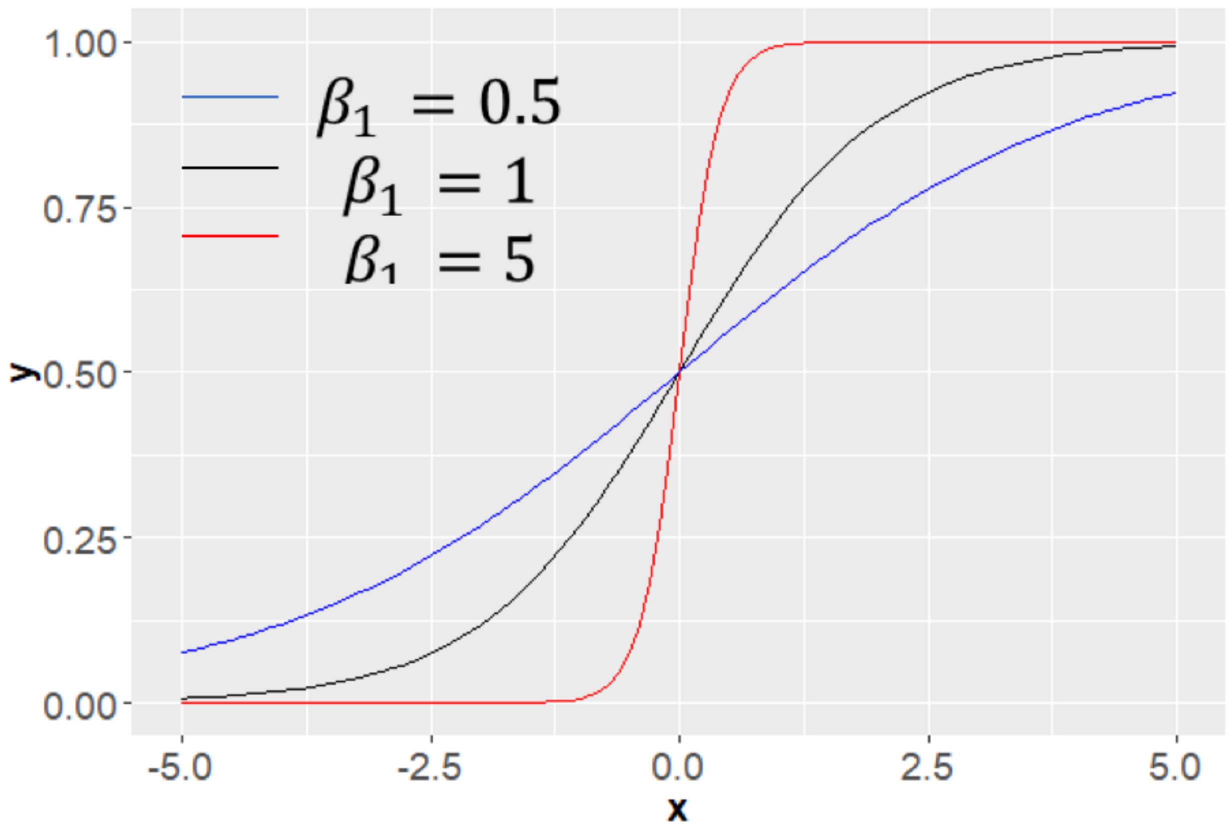
$$\pi_i = \frac{1}{1 - e^{-(a+\beta_1 x_1)}} = \frac{e^{a+\beta_1 x_{i1}}}{e^{(a+\beta_1 x_{i1})+1}}$$



logistic regression



- $\beta_1 > 0$, π(x) increase as x increase
- $\beta_1 < 0$, π(x) increase as x increase



- β determines the rate of growth or increase of the curve.
- The magnitude of β determines the rate of that increase or decrease

Value of β_1 can be used to find the significative variable

β_1 = 0 → the variable does not affect the result

$$\left(\frac{\pi}{1-\pi}\right) = e^{\alpha} + \left(e^{b}\right)^{x} e^{\beta}$$

best fit criteria

To compute the goodness-of-fit statistic for logistic regression following techniques are the most commonly used:

- **Deviance statistic**, a measure of discrepancy between observed and fitted values:

$$D = 2\sum\{y_i \log(\frac{y_i}{\hat{\mu}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{\mu}_i})\}$$

- **Pearson's chi-squared statistic**:

$$\chi_P^2 = \sum_i \frac{n_i(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(n_i - \hat{\mu}_i)}$$

accuracy: measure of predictive power

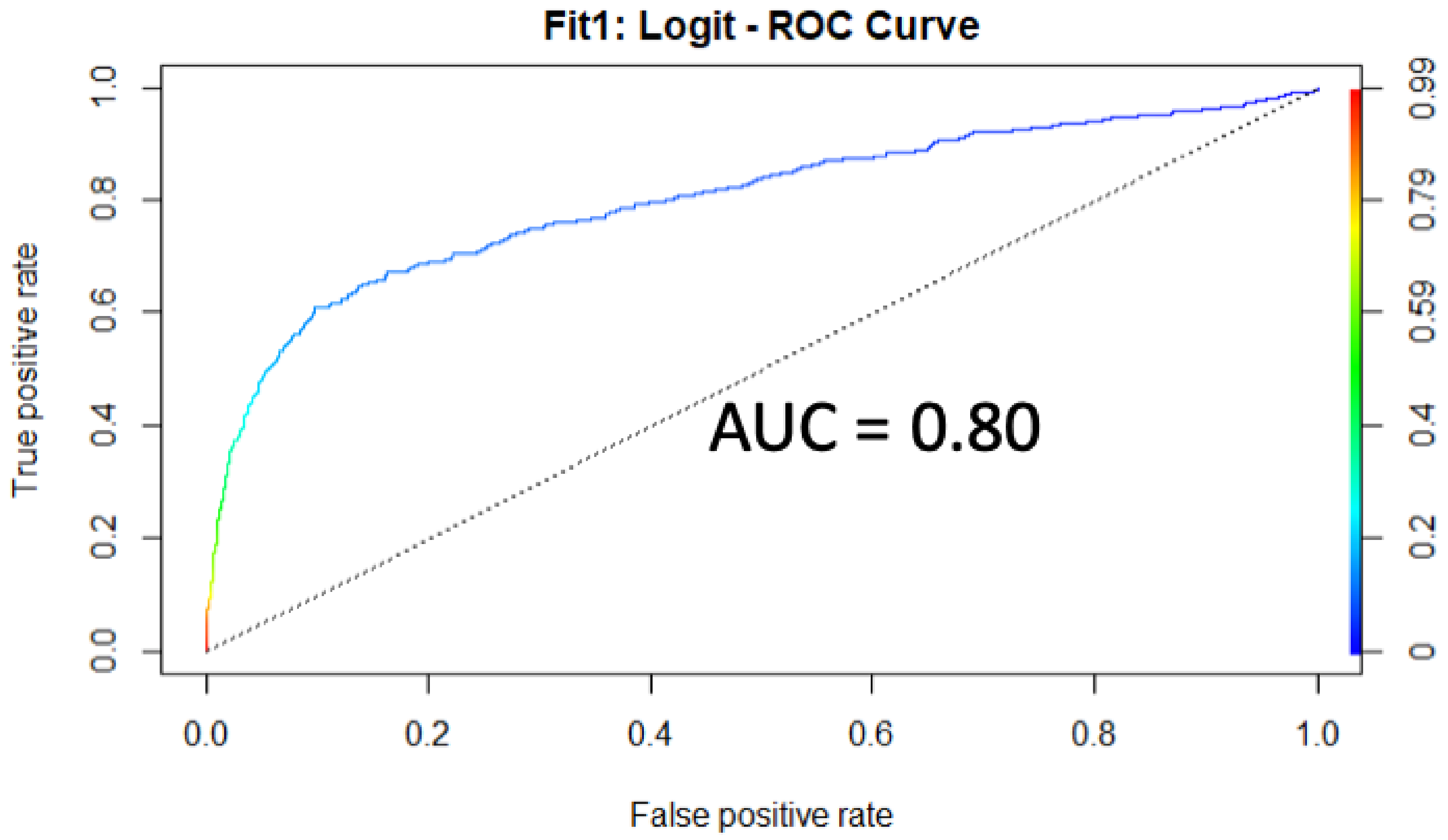
It is possible to test the validity of the model by measuring how well you can predict the dependent variable based on the independent variables, i.e.

the ability of the model to differentiate between default or non-default event based on its input value.

Receiver Operating Characteristics (**ROC**) curve is the plot that displays the full picture of trade-off between the sensitivity and (1-specificity) across a series of cutoff points.

Sensitivity (or True Positive Rate) is the percentage of 1's (default) correctly predicted by the model, while, **specificity** is the percentage of 0's (non-default) correctly predicted. Specificity can also be calculated as 1 – False Positive Rate.

Area Under the ROC curve (AUROC or AUC) is a statistic that summarizes the predictive accuracy. AUC value range is between 0 and 1.



The contents here presented are strictly confidential and intended for educational purposes only. It is strictly forbidden to share any part of this presentation with any third party, without a written consent of the sender.

other accuracy methods

Other measures of model discrimination, frequently used in credit risk models, are:

- **Gini coefficient** or **accuracy ratio** (**AR**): it is related to the AUC metrics by the following formula: G = (2*AUC)-1

- **Kolmogorov-Smirnov** (**KS**) metrics: it is related to the cumulative distribution of true positive (event captured) and false positive (event non-captured).
It is a measure of the maximum separation between the two curves.

Both Gini and KS metrics values range from 0 to 1.
They can be used to perform direct comparisons across the models.

LET'S SEE THIS !

TREE MODEL

Tree models are an important type of non-parametric models: any assumption of the functional form is set *a priori*.

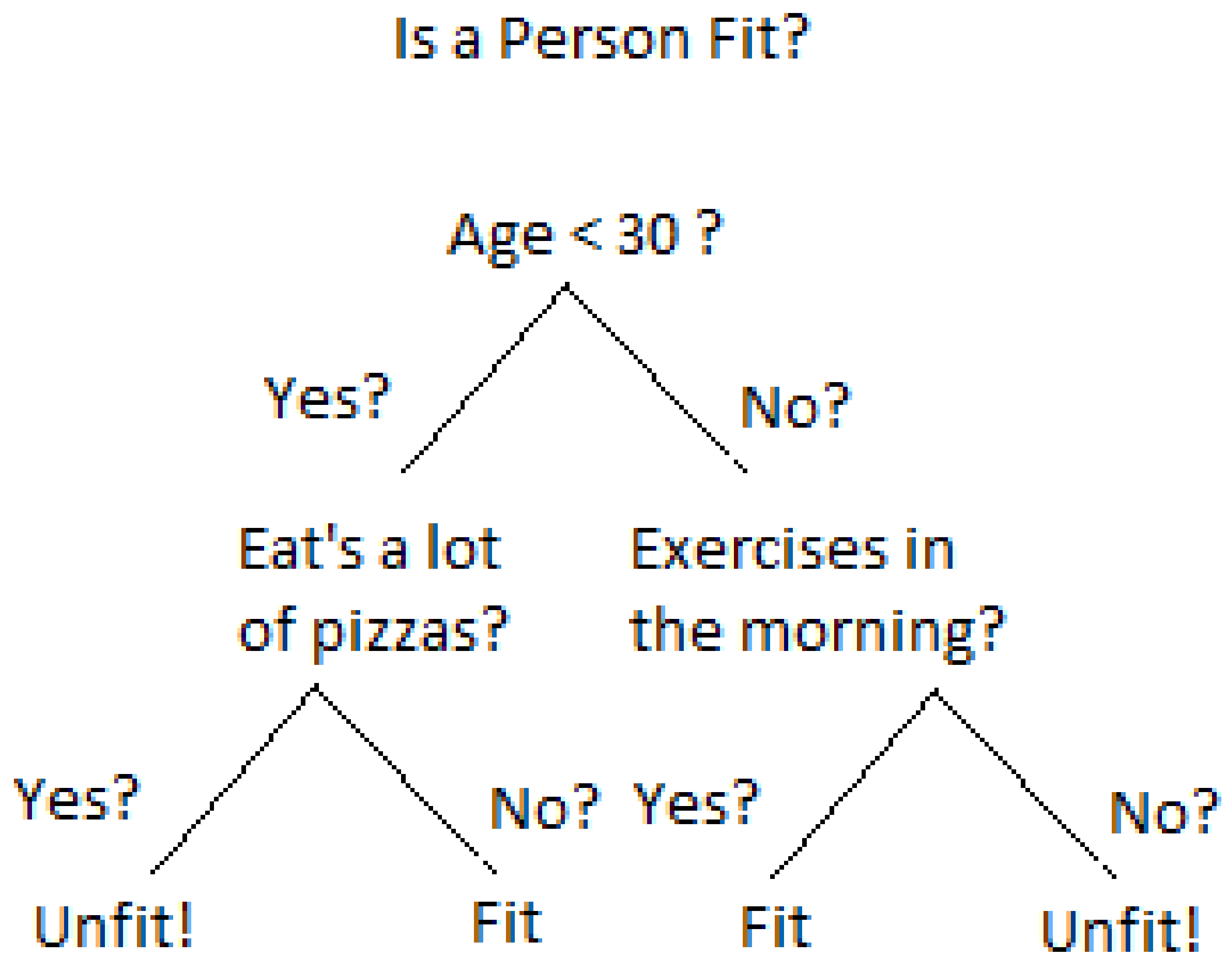
While linear and logistic regression methods produce first a score and then may produce a classification, according to a discriminant rule, tree models first produce a classification of groups observations and then a constant score for each of them

Tree models are fast to learn and do not require any special preparation for your data. We can distinguish between:

- **Regression tree** → **continuous response variable**
- **Classification tree** → **discrete response variable**

how do they work?

n statistical units are progressively divided in subgroups according to a splitting rule that maximizes purity (homogeneity) of the *Y* values within each obtained group.



how do they work?

Suppose that a final partition is achieved, with *g* < *n* total number of groups:

Regression trees produce a **fitted value** $\boldsymbol{\hat{y}_i}$ for group, which is equal to the mean response value of the group to which the observation *i* belongs (*m* is the group index):

$$\hat{y}_i = \frac{\sum_{l=1}^{n_m} y_{lm}}{n_m}$$

Classification trees produce a fitted affiliation probability to a single group:

$$\pi_i = \frac{\sum_{l=1}^{n_m} y_{lm}}{n_m}$$

If only two classes are possible (binary classification) the value of *y* can be only 1 or 0, and therefore the fitting probability corresponds to the observer proportion of success (for example default-event in the group).

the splitting criteria to be maximized

For a tree algorithm the splitting criteria to be maximized is the criterion function:

$$\Phi(s,t) = I(t) - \sum_{r=1}^{s} I(t_r)p_r$$

t is the new node;
p_r is the proportion of the observation that are allocated to each child node $(\sum p_r = 1)$;
I is the impurity function, whose definition depends on the tree model type (regression or classification).

$$I_v(m) = \frac{\sum_{l=1}^{n_m}(y_{lm} - \hat{y}_m)^2}{n_m}$$

m is the group index

Impurity for regression tree is linked to the variance of data.

The contents here presented are strictly confidential and intended for educational purposes only. It is strictly forbidden to share any part of this presentation with any third party, without a written consent of the sender.

misclassification & Gini

Misclassification:

$$I_M(m) = \frac{\sum_{l=1}^{n_m} 1(y_{lm} y_k)}{n_m}$$

Impurity for classification tree.

Gini:

$$I_G(m) = 1 - \sum_{i=1}^{k(m)} \pi_i^2$$

π_i fitted probabilities of the level at node *m*

The recursive process goes on until a stopping criteria is introduced, i.e.:

- Set a maximum number of groups.
- Set a maximum number of steps.
- Introducing a minimum value for impurity decrease.

CART TREE MODEL

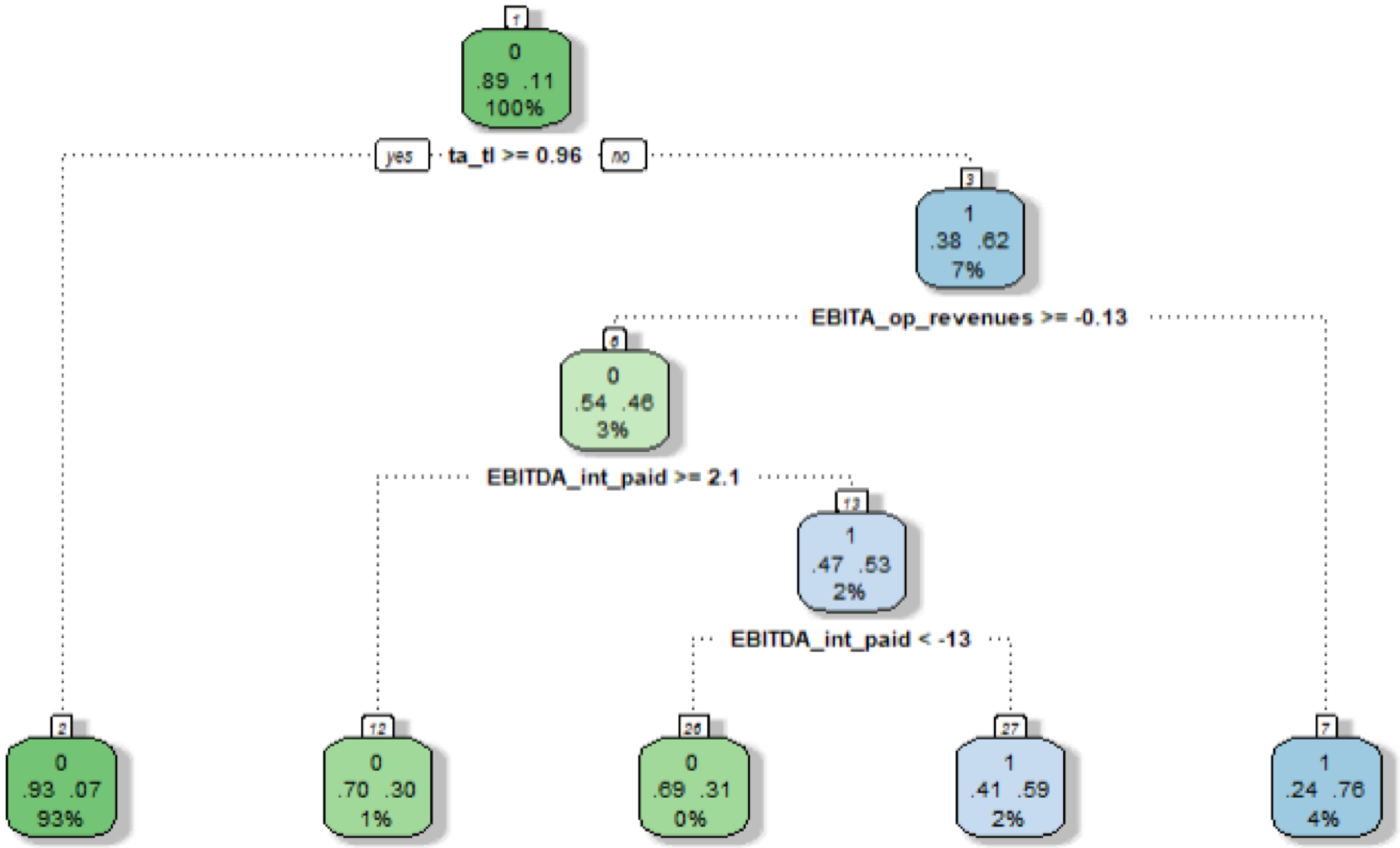
A CART tree algorithm has a different stopping criteria: **pruning strategy**.

- First of all, the tree is grown at its maximum level.
- Then, the pruning proceeds backward, eliminating at each step the node if this results in a drop within the overall cost function on the entire test set.
Stop removing nodes when no further improvement will it be made.

Complexity pruning is a more sophisticated method where a learning parameter *alpha* is used to weigh whether nodes can be removed, based on the size of the sub-tree.

CART tree output

Each root node represents a single variable input *X* and a single split point.
The leaf node of the tree contains an output variable used to make predictions.

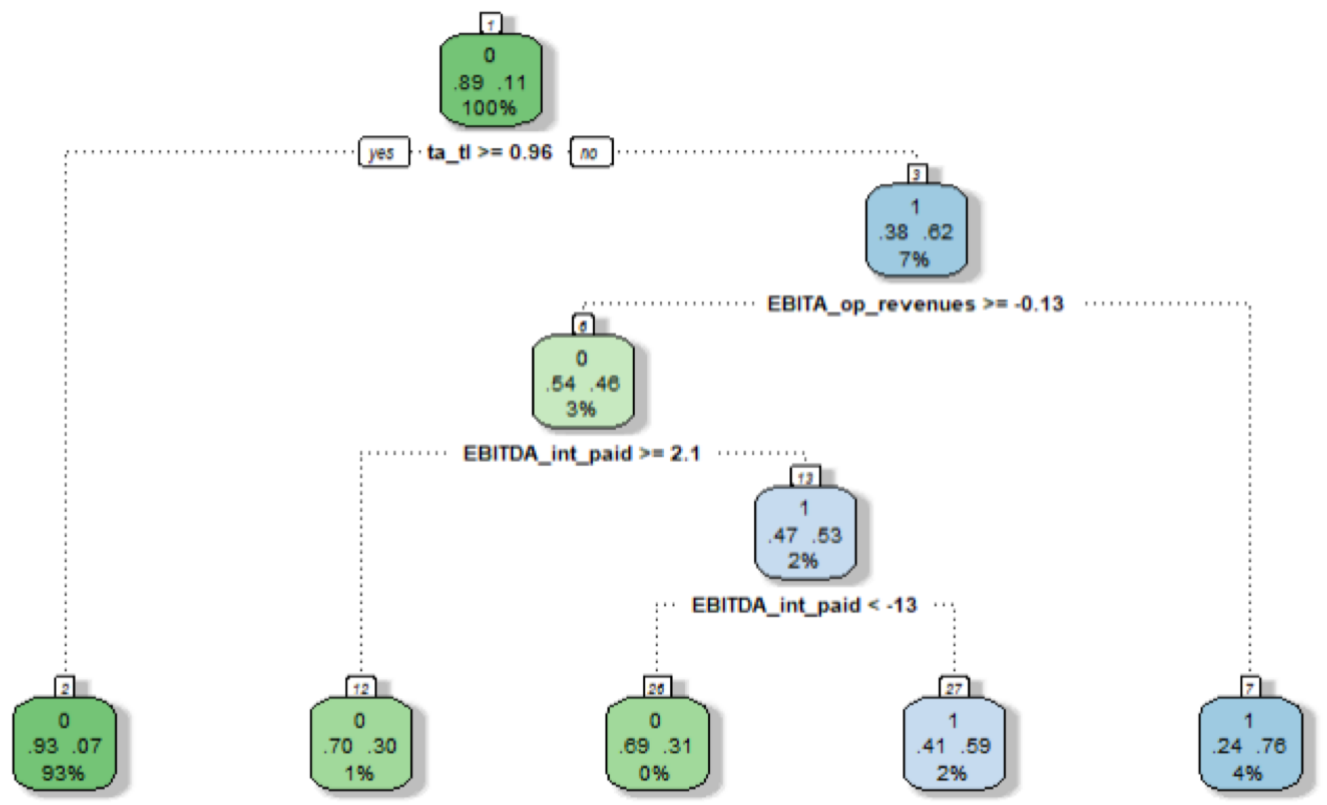


The contents here presented are strictly confidential and intended for educational purposes only. It is strictly forbidden to share any part of this presentation with any third party, without a written consent of the sender.

CART model output

The output of a CART model is a set of rules on input variables:

```
1)root 10533 1143 0 (0.8914839 0.1085161)
2) ta_tl>=0.955 9784 677 0 (0.9308054 0.0691946) *
3) ta_tl< 0.955 749 283 1 (0.3778371 0.6221629)
6) EBITA_op_revenues>=-0.135 341 156 0 (0.5425220 0.4574780)
12) EBITDA_int_paid>=2.05 108 32 0 (0.7037037 0.2962963) *
13) EBITDA_int_paid< 2.05 233 109 1 (0.4678112 0.5321888)
26) EBITDA_int_paid< -12.865 51 16 0 (0.6862745 0.3137255) *
27) EBITDA_int_paid>=-12.865 182 74 1 (0.4065934 0.5934066) *
7) EBITA_op_revenues< -0.135 408 98 1 (0.2401961 0.7598039) *
```



Default probability

0.0691946 when ta_tl>=0.955 *(2)*

0.7598039 when ta_tl< 0.955 & EBITA_op_revenues< -0.135 *(7)*

…

Decision rules for the previous binary tree.

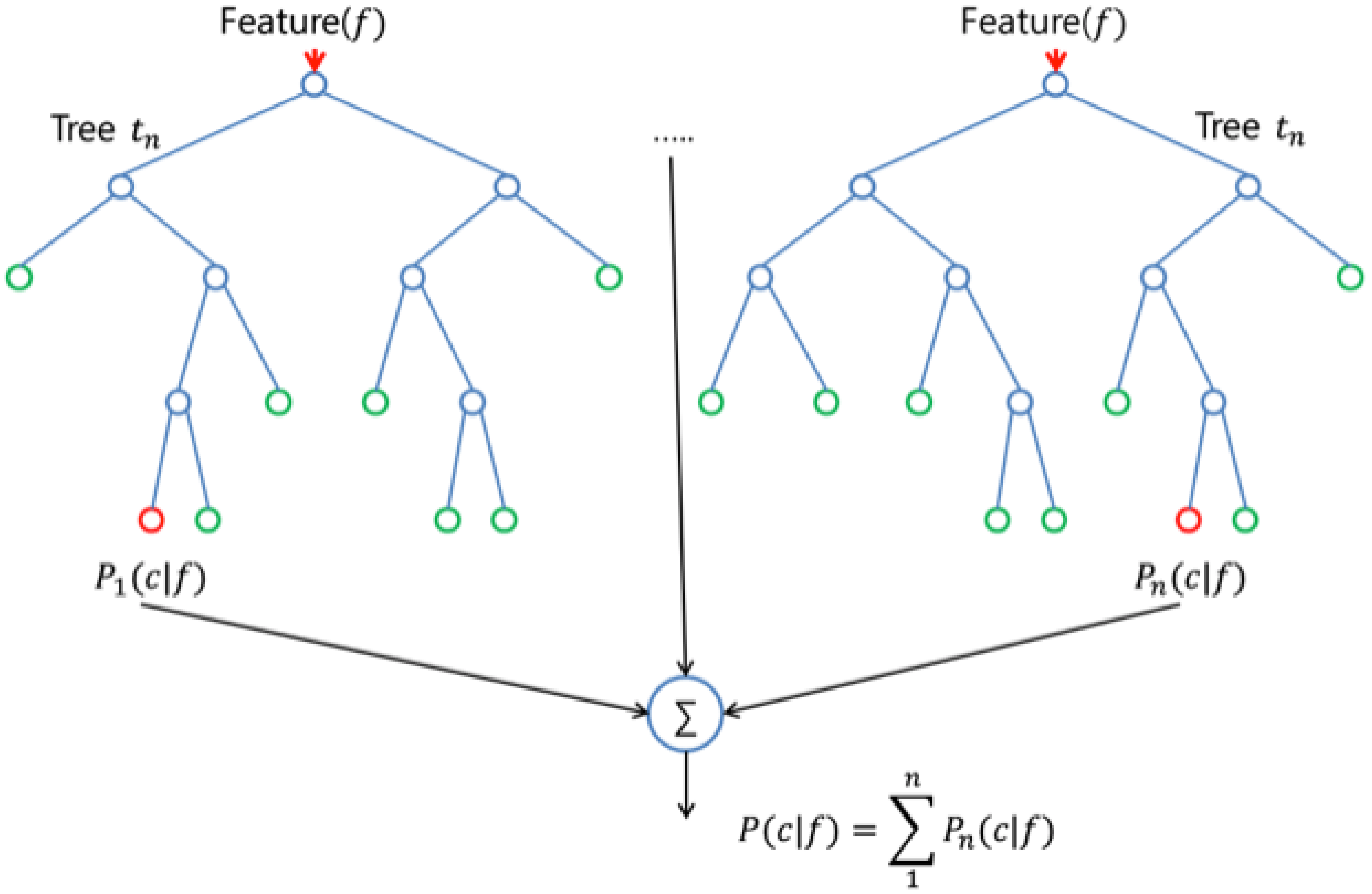
LET'S SEE THIS !

RANDOM FOREST

Random Forest is an ensemble method that uses multiple decision tree models.

It can be used for both classification and regression.

Accuracy and variable importance information is provided with the results.



how does it work?

Decision trees tend to **overfit** on training data therefore reducing the predictive power of the model.

- **Bagging tree**: the use of multiple decision trees reduces overfitting, BUT:

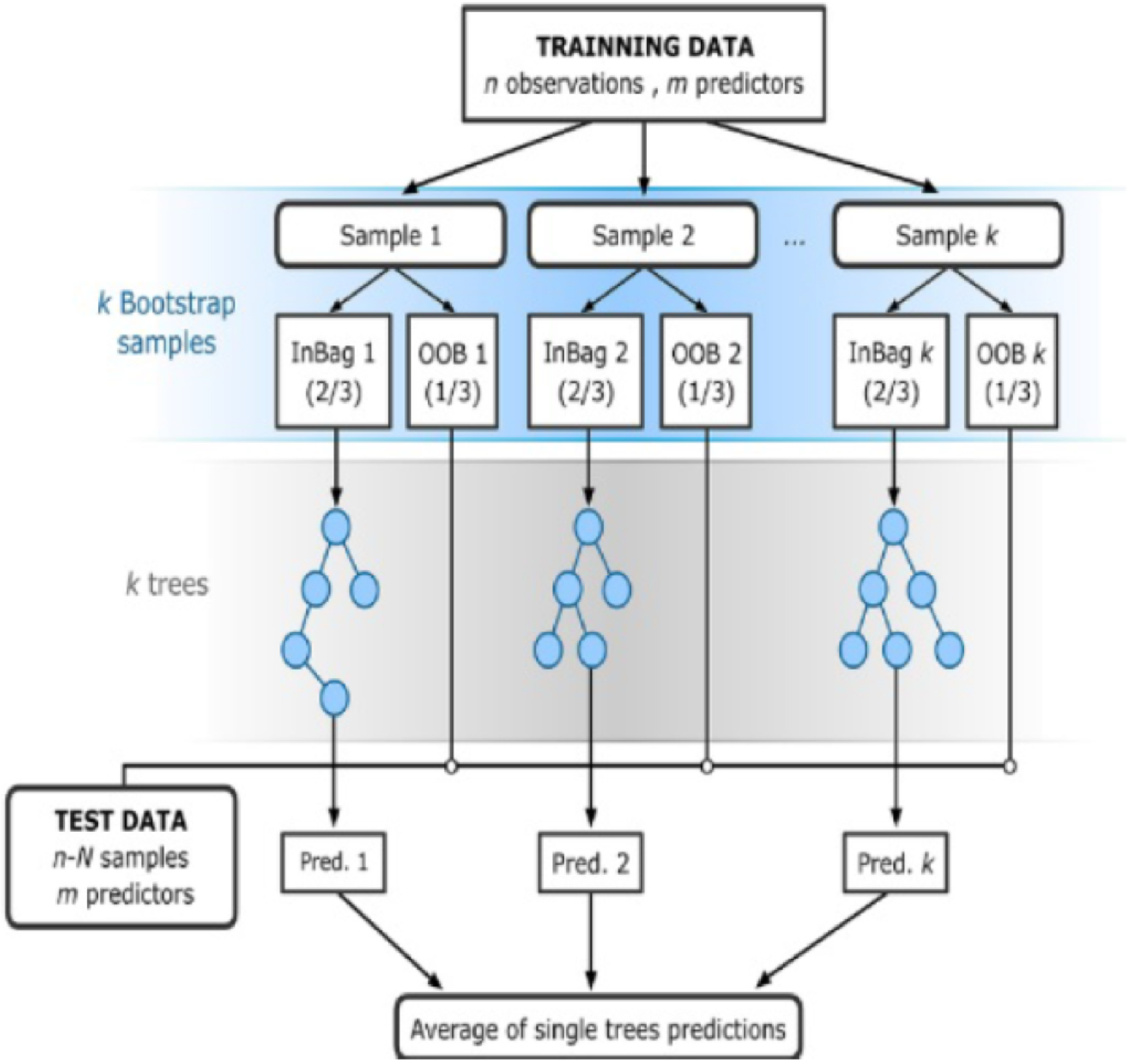
A bootstrap aggregating algorithm is used to create several subsets of data, from training samples chosen randomly with replacement (use of same values multiple times is accepted).
A decision tree is trained from each subset, ending up with an ensemble of different models.
Average predictions from different trees are used, obtaining a more robust result than a single decision tree.

Why Random Forest:
- Tree models in Bagging tree can have high correlations;
- Random forest reduces model correlations, performing a random selection of input variable to grow trees.

algorithm flow chart

Random forest crosses validation method:
Data in sub-sample **split in training and test set** (Out-Of-Bag samples).
OOB is used to evaluate the model's performance.

Hyperparameter of the model affects both predictive power and speed of the model, we will use:
- **ntree,** number of tree models;
- **mtry,** number of variables randomly sampled as candidates at each split.



LET'S SEE THIS !

THANK YOU

see you in Frankfurt

FINTECH RISK MANAGEMENT
www.fintech-ho2020.eu

Sabrina Stella, Fintech analyst

modefinance

European Commission

The contents here presented are strictly confidential and intended for educational purposes only. It is strictly forbidden to share any part of this presentation with any third party, without a written consent of the sender.