Supervised learning, M2 DS2E

Marion, Lyna, Jeanne

2023-09-20

Presentation

Source of Data : the Survey on Business Strategies

- 2000 manufacturing companies
- 121 992 observations
- 16 variables period : 1990-2012

GOAL : Predict which company is going to be a HGF in the last of year of the sample

Steps

NA Treatment

- For the yearest, use the minimum of the id
- Replace sales and va missing by mean of values before and after
- Use the median for variables with remaining missing values

Outliers Treatment

- Delete outliers in “gom”, extremely high comparing to the rest

Creating new variables

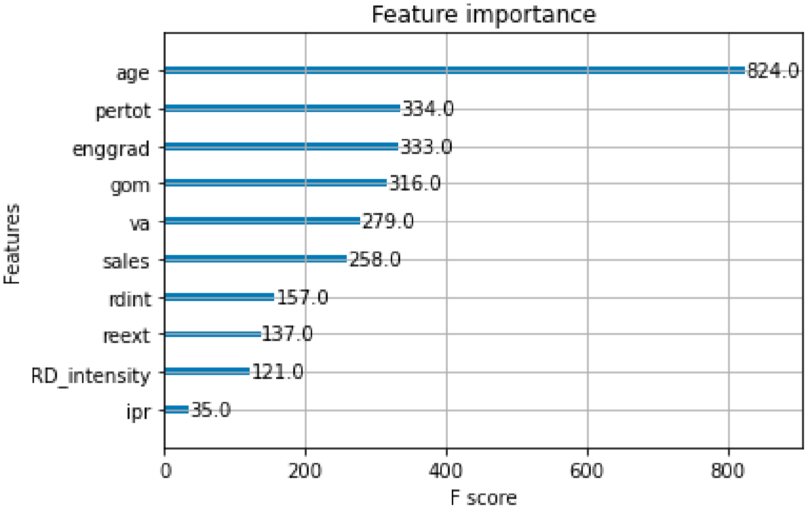
- Age, HGF and R&D

Model

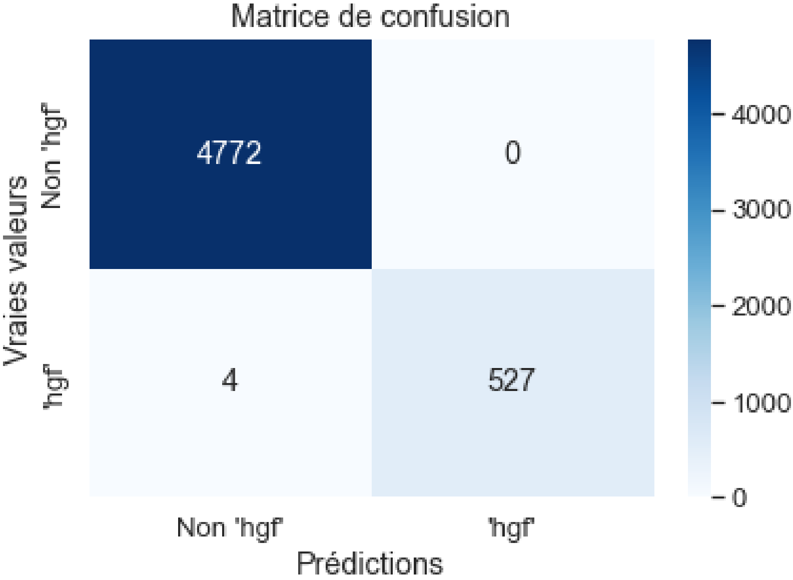
Table 1: *Accuracy of the models*

	Logistic regression	SVM	KNN	Decision tree	Random forest	XG Boost
accuracy test	0.117	0.994	0.997	0.997549	0.999057	0.999246

Feature importance



Confusion matrix


Matrice de confusion
Vraies valeurs
Non 'hgf'
'hgf'
Non 'hgf'
'hgf'
Prédictions