Prédire le turn-over de salariés

Problème de régression logistique / classification

Présentation des variables

- **attrition_train.csv** : historique des employés avec **1100 lignes**
- **attrition_test.csv** : employés actuellement dans l'entreprise avec **370 lignes**
- Il y a **0 valeurs manquantes**.
- Il y a 3 colonnes qui affichent 1 seule valeur : **'Over18', 'EmployeeCount','StandardHours'**
- La colonne **'EmployeeNumber'** donne un numéro d'employé qui n'a pas de lien apparent avec la variable cible.
- Il y a une colonne **index1** dans attrition_test sans lien avec la variable cible.

Présentation des variables

23 variables numériques :

'Age', 'DailyRate', 'DistanceFromHome', 'Education', 'EnvironmentSatisfaction', 'HourlyRate', 'JobInvolvement', 'JobLevel', 'JobSatisfaction', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'PercentSalaryHike', 'PerformanceRating', 'RelationshipSatisfaction', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager'

Présentation des variables

7 variables catégorielles :

'BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'OverTime'

Traitements réalisés sur les variables

Suppression des colonnes : 'Over18', 'EmployeeCount', 'StandardHours', 'EmployeeNumber' et ‘index1’

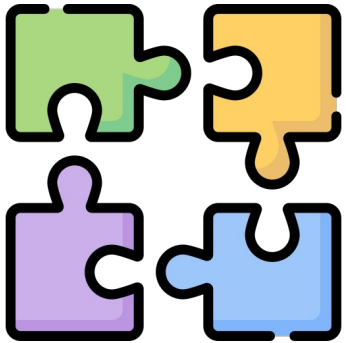
Preprocessor

Transformation des variables numériques en variables centrées réduites :

```
StandardScaler()
```

Encodage des variables catégorielles en variables numériques binaires :

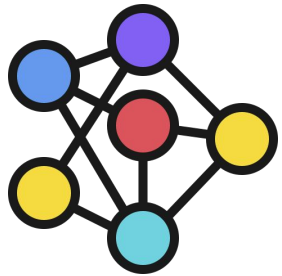
```
OneHotEncoder()
```



Modèle de classification

Classifier : **LogisticRegression()**

class sklearn.linear_model.LogisticRegression(*penalty='l2'*, ***, *dual=False*, *tol=0.0001*, *C=1.0*, *fit_intercept=True*, *intercept_scaling=1*, *class_weight=None*, *random_state=None*, *solver='lbfgs'*, *max_iter=100*, *multi_class='auto'*, *verbose=0*, *warm_start=False*, *n_jobs=None*, *l1_ratio=None*)

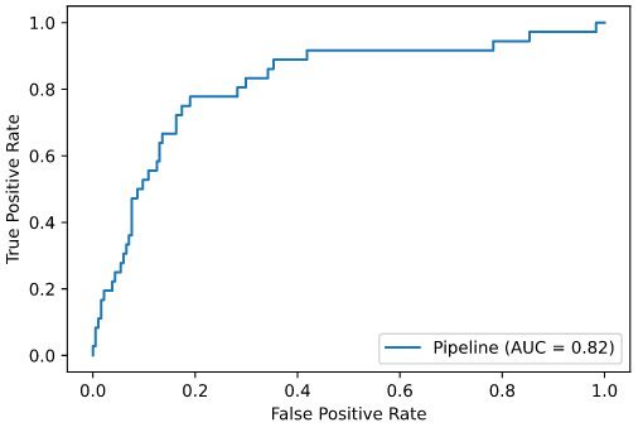


Évaluation - f1-score - roc_auc

Recherche des meilleurs paramètres par une validation croisée GridSearchCV, avec comme métriques le F1-score (qui combine la sensibilité et la spécificité) et le ROC AUC (Area Under Curve).

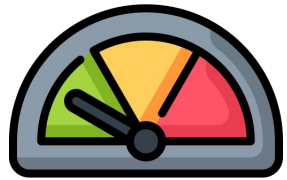
Une courbe ROC trace les valeurs TVP et TFP pour différents seuils de classification. Diminuer la valeur du seuil de classification permet de classer plus d'éléments comme positifs, ce qui augmente le nombre de faux positifs et de vrais positifs.

```
param_grid = {
    'classifier__C': [1, 3, 10, 30, 100],
    'classifier__dual': [True, False],
    'classifier__penalty' : ['l1', 'l2', 'elasticnet', 'none'],
    'classifier__solver' : ['newton-cg', 'lbfgs', 'liblinear', 'sag'],
    'classifier__l1_ratio' : [0.3, 0.5, 0.7],
    'classifier__class_weight' : ['balanced', None]
}

scorers = {'f1': make_scorer(f1_score),
           'roc_auc': make_scorer(roc_auc_score)}

grid_search = GridSearchCV(clf, param_grid, scoring=scorers, refit='roc_auc', cv=5)

grid_search.fit(X_train, y_train)
```



Évaluation - f1-score - roc_auc

```
Pipeline(steps=[('preprocessor',
                 ColumnTransformer(transformers=[('num',
                                                  Pipeline(steps=[('scaler',
                                                                   StandardScaler())]),
                                                  ['Age', 'DailyRate',
                                                   'DistanceFromHome',
                                                   'Education',
                                                   'EnvironmentSatisfaction',
                                                   'HourlyRate',
                                                   'Job...
                                                   'YearsInCurrentRole',
                                                   'YearsSinceLastPromotion',
                                                   'YearsWithCurrManager']),
                                                 ('cat',
                                                  OneHotEncoder(handle_unknown='ignore'),
                                                  ['BusinessTravel',
                                                   'Department',
                                                   'EducationField', 'Gender',
                                                   'JobRole', 'MaritalStatus',
                                                   'OverTime'])])),
                ('classifier',
                 LogisticRegression(C=1, class_weight='balanced', l1_ratio=0.3,
                                    max_iter=1500, penalty='l1', random_state=0,
                                    solver='liblinear'))])
```



Modèle de classification

LogisticRegression()

Coefficients des variables :

	features	coef
45	JobRole_Sales Representative	1.164649
29	EducationField_Human Resources	1.136431
24	BusinessTravel_Travel_Frequently	0.989427
34	EducationField_Technical Degree	0.808924
19	YearsAtCompany	0.782316
21	YearsSinceLastPromotion	0.657373
48	MaritalStatus_Single	0.503620
50	OverTime_Yes	0.395513
11	NumCompaniesWorked	0.331009
2	DistanceFromHome	0.286214
5	HourlyRate	0.117706
28	Department_Sales	0.101258
13	PerformanceRating	0.080190
39	JobRole_Laboratory Technician	0.046842
3	Education	0.018884
10	MonthlyRate	0.017792
40	JobRole_Manager	0.000000
27	Department_Research & Development	0.000000
38	JobRole_Human Resources	0.000000
37	JobRole_Healthcare Representative	0.000000
36	Gender_Male	0.000000
42	JobRole_Research Director	0.000000
32	EducationField_Medical	0.000000
44	JobRole_Sales Executive	0.000000
31	EducationField_Marketing	0.000000
47	MaritalStatus_Married	0.000000
0	Age	0.000000
25	BusinessTravel_Travel_Rarely	0.000000
18	WorkLifeBalance	-0.091566
1	DailyRate	-0.172419
7	JobLevel	-0.227144
8	JobSatisfaction	-0.248735
30	EducationField_Life Sciences	-0.256777
41	JobRole_Manufacturing Director	-0.261924
46	MaritalStatus_Divorced	-0.296422
17	TrainingTimesLastYear	-0.297115
12	PercentSalaryHike	-0.299704
33	EducationField_Other	-0.363513
26	Department_Human Resources	-0.364334
20	YearsInCurrentRole	-0.390196
6	JobInvolvement	-0.399982
14	RelationshipSatisfaction	-0.414038
15	StockOptionLevel	-0.429031
35	Gender_Female	-0.442874
16	TotalWorkingYears	-0.471693
9	MonthlyIncome	-0.515209
4	EnvironmentSatisfaction	-0.550361
22	YearsWithCurrManager	-0.682622
43	JobRole_Research Scientist	-0.736262
23	BusinessTravel_Non-Travel	-0.927701
49	OverTime_No	-1.474297

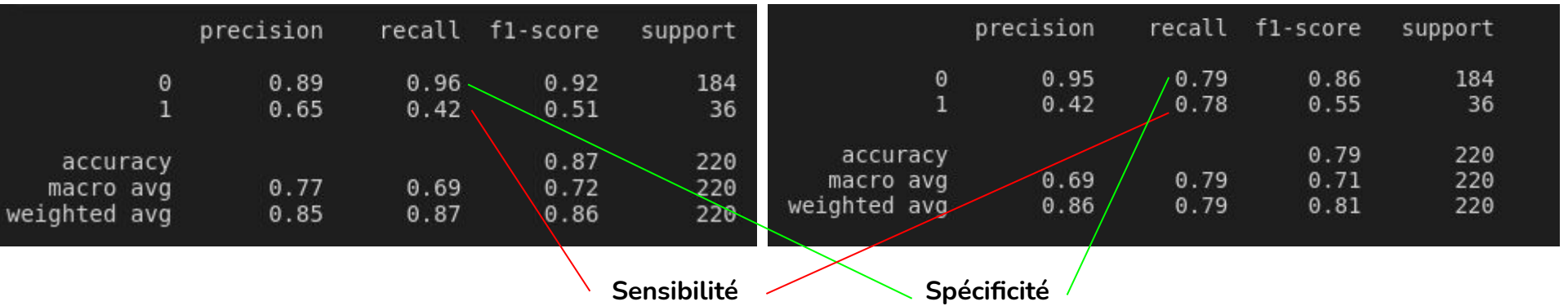
Évaluation

Avant recherche des paramètres

	precision	recall	f1-score	support
0	0.89	0.96	0.92	184
1	0.65	0.42	0.51	36
accuracy			0.87	220
macro avg	0.77	0.69	0.72	220
weighted avg	0.85	0.87	0.86	220

Après recherche des paramètres

	precision	recall	f1-score	support
0	0.95	0.79	0.86	184
1	0.42	0.78	0.55	36
accuracy			0.79	220
macro avg	0.69	0.79	0.71	220
weighted avg	0.86	0.79	0.81	220

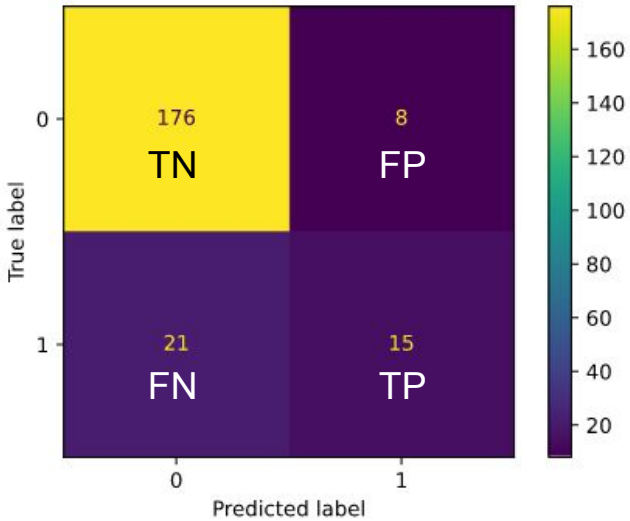


Notez que dans la classification binaire, le rappel de la classe positive est également appelé «sensibilité»;

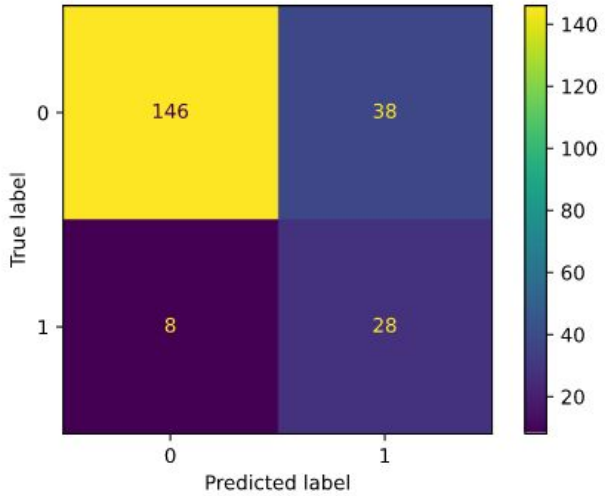
le rappel de la classe négative est la «spécificité».

Évaluation

Avant recherche des paramètres



Après recherche des paramètres



Le nombre de faux négatifs passe de 21 à 8. (Meilleure sensibilité)

Le nombre de faux positifs passe de 8 à 39. (Moins bonne spécificité)

Il faudra déterminer le coût du programme d'accompagnement comparativement

au risque de voir partir les employés dont on a pas détecté l'envie de quitter l'entreprise.

Prédiction

Nous prédisons un score AttritionScore% et une prédiction PredictWillLeave (0 ou 1) dans le fichier predictions.csv

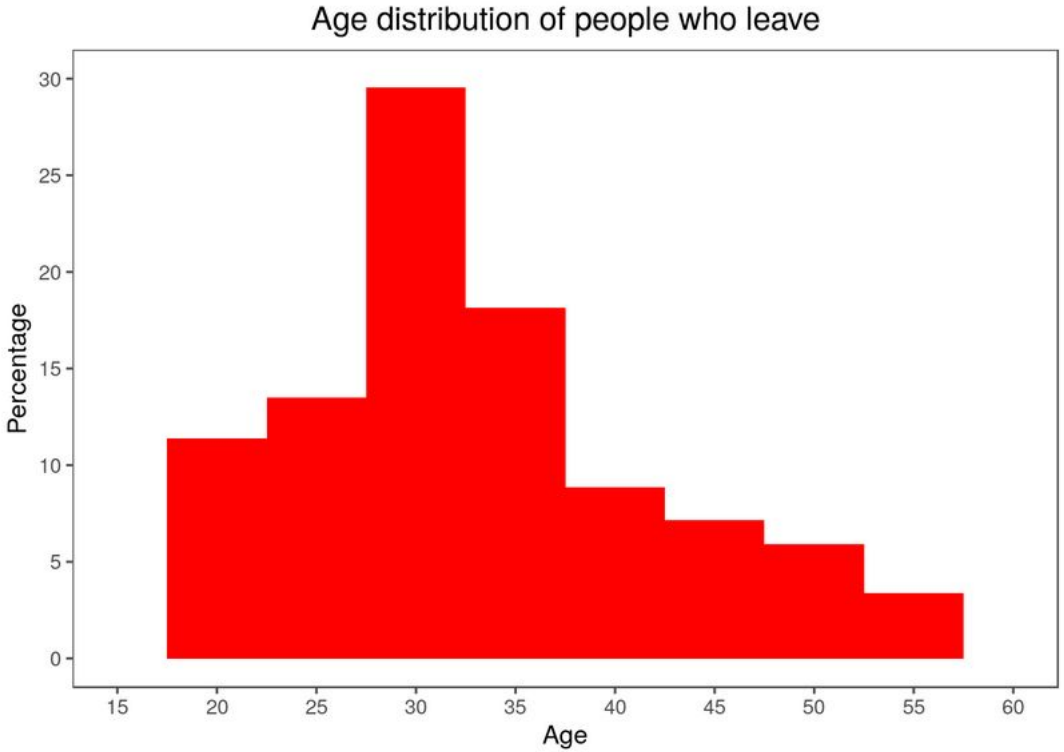
Il y a sur les 370 employés, 113 employés dont on prédit le départ selon notre modèle, avec un seuil à 50% de probabilité.

Si on augmente le seuil :

- à 60 % : on prédit le départ de 79 employés
- à 70 % : on prédit le départ de 54 employés
- à 80 % : on prédit le départ de 30 employés
- à 90 % : on prédit le départ de 10 employés

Pour définir la taille optimale du programme d'accompagnement, il faudra déterminer le seuil choisi en fonction du coût de ce programme, et du coût que représente le départ d'un employé. On pourra aussi décider de ne pas sélectionner dans le programme les employés qui sont proches de la retraite.

Prédiction


Age distribution of people who leave
Percentage
30
25
20
15
10
5
0
15
20
25
30
35
40
45
50
55
60
Age