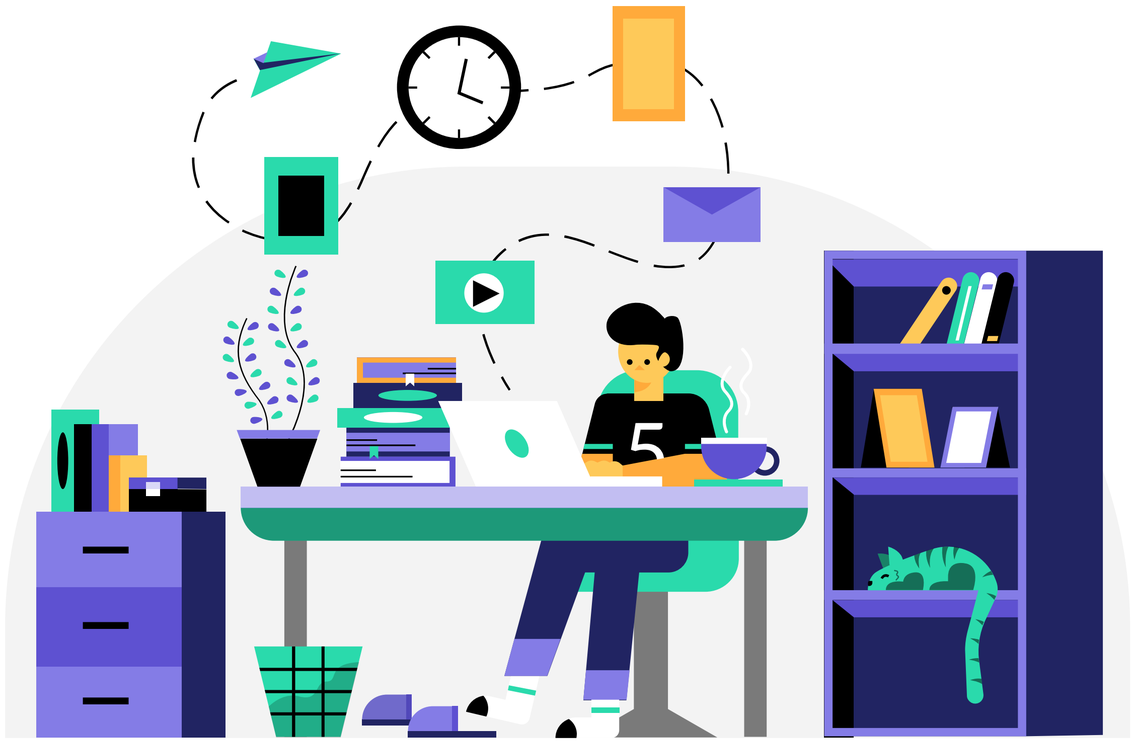


Projet

–

Prédire le turn over des employés d’une sociétés

INTRODUCTION

1. Contexte du projet
2. Présentation du modèle logistique
3. Variables du jeu de données
4. Traitement et nettoyage
5. Evaluation du modèle
6. Réponses aux problématiques

Contexte du projet

La société X est confronté à un problème de turn-over de ses employés. Trop de personnes quittent l'entreprise pour aller à la concurrence .

Ceci représentant un coût important pour l’entreprise, le service des ressources humaines souhaitent mettre en place un programme d’accompagnement ciblant les salariés dont les chances de départ sont les plus élevées.

Notre objectif est d’identifier ces salariés à partir d’un ensemble de données fournit par l’entreprise.

Présentation du modèle logistique

Pour réaliser notre prédiction, nous avons choisi le modèle de machine learning SVC, qui nous a fournit les meilleurs performances apres un ensemble de test.

Les variables du modèle

Pour réaliser nos prédictions de l'attrition, nous avons utilisés au total **27 variables sur les 35 variables** présentent.

Nous avons notamment exclut les variables qui :

qui ne contenaient qu'une **valeur unique** et qui ne permettaient donc pas d'apporter des informations sur le départ ou non d'un salarié (ex : Over18, Standard Hours, ...)

qui possédaient une **corrélation très forte avec une autre variable** (ex : niveau hiérarchique et revenus mensuels)

qui fournissaient une **même information de différentes manières** (ex : revenus mensuels et taux horaire)

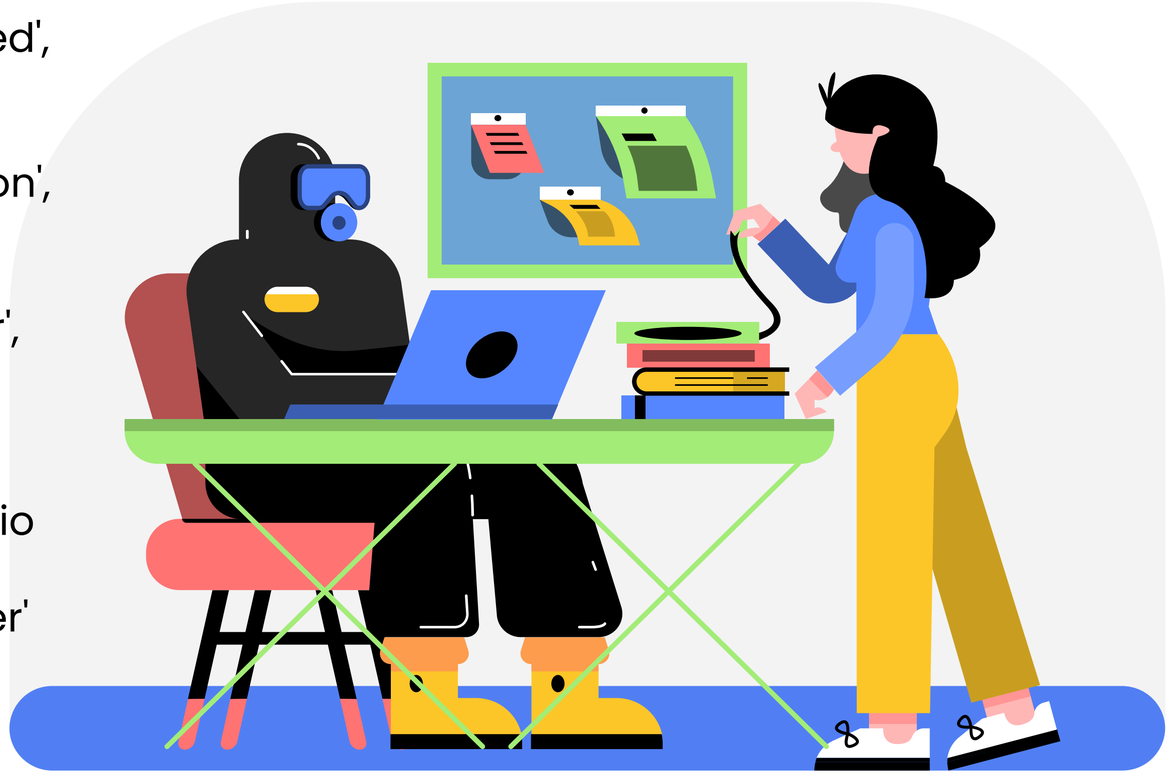
qui r**éduisaient la performance du modèle** lorsqu'elles étaient ajoutées (ex : évaluation de la performance, niveau d'éducation, ...)

enfin nous avons utilisé un algorithme (RFE) qui nous a selectioner ,sur les valeur restantes, les variables les plus importantes a utiliser sur notre modèle

Les variables du modèle

Les variables utilisés dans le modèle sont :

- 'Age',
- 'Attrition',
- 'BusinessTravel',
- 'Department',
- 'DistanceFromHome',
- 'EducationField',
- 'EmployeeNumber',
- 'EnvironmentSatisfaction',
- 'Gender',
- 'HourlyRate',
- 'JobInvolvement',
- 'JobRole',
- 'JobSatisfaction',
- 'MaritalStatus',
- 'MonthlyIncome',
- 'MonthlyRate',
- 'NumCompaniesWorked',
- 'OverTime',
- 'PercentSalaryHike',
- 'RelationshipSatisfaction',
- 'StockOptionLevel',
- 'TotalWorkingYears',
- 'TrainingTimesLastYear',
- 'WorkLifeBalance',
- 'YearsAtCompany',
- 'YearsInCurrentRole',
- 'YearsSinceLastPromotion',
- 'YearsWithCurrManager'

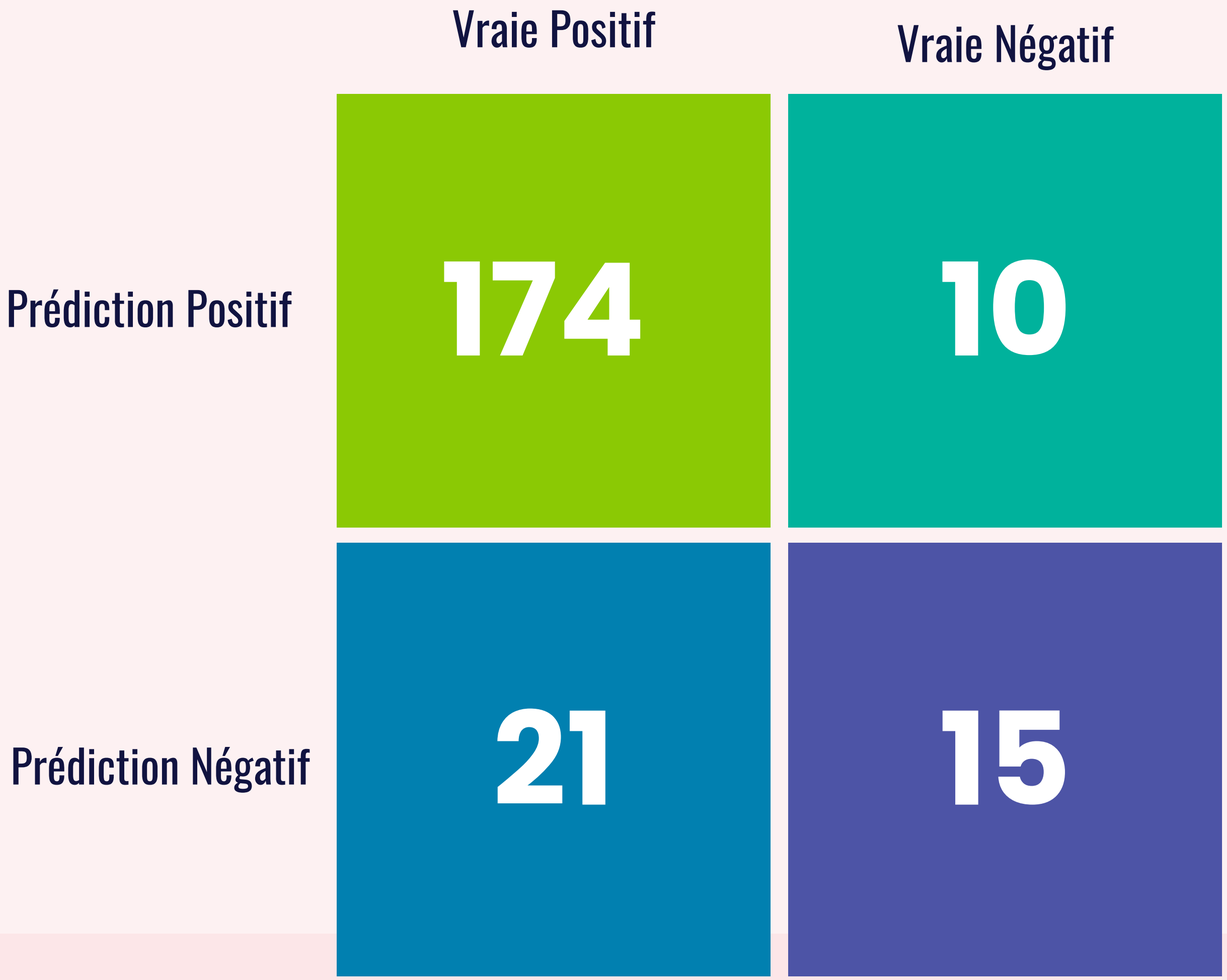


Le traitement des variables

	Variables numériques	Variables catégoriques
Traitement des valeurs nulles	Correction par la médiane	Correction par la valeur la plus fréquente
Encodage	Non Necessaire	Création de nouvelles colonnes exprimant les valeurs sous forme numérique
Imputation	Suppression des lignes contenant des valeurs nulles	Suppression des lignes contenant des valeurs nulles

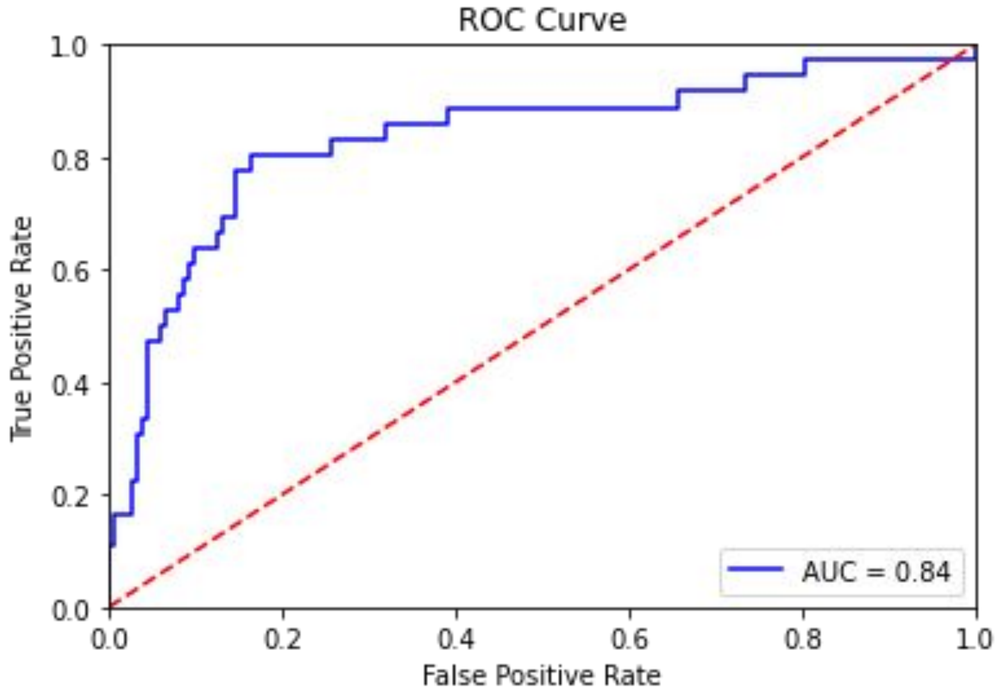
Evaluation du modèle

Présentation de la matrice de confusion :



Le modèle utilisé obtient un score de **95%** pour la prédiction des personnes qui sont restés dans l'entreprise, contre **42%** pour la prédiction des personnes ayant quitté l'entreprise

Evaluation du modèle



Une courbe **ROC** trace le taux de vrais positifs en fonction du taux de faux positifs

AUC signifie "aire sous la courbe ROC". Cette valeur mesure l'intégralité de l'aire située sous l'ensemble de la courbe ROC, plus le chiffre est élevé , moins il y a de faux positif

La réponse aux problématiques

Prédiction de la liste d'admission au programme d'accompagnement

Cahier des charges :

- 10 Employés Minimum
- 100 Employés Maximum
- Liste Index des employés

Grâce à notre modèle, nous avons pu déterminer les employés qui sont susceptibles de partir de l’entreprise. Sur les 370 salariés, 36 peuvent potentiellement quitter l’entreprise (soit 9.7% des salariés).

Voici la liste des index de ces personnes :

[0, 2, 5, 10, 11, 23, 25, 31, 49, 54, 72, 74, 76, 80, 89, 101, 145, 149, 157, 177, 179, 187, 206, 228, 250, 260, 265, 273, 276, 306, 329, 347, 351, 353, 356, 367]