ECOLE SUPÉRIEURE DE LA STATISTIQUE
ET DE L'ANALYSE DE L'INFORMATION

PROJET FIN D'ANNÉE:

MODÉLISATION PRÉDICTIVE POUR LA PROBABILITÉ DE DÉFAUT

PRÉPARÉ PAR :
- BEN ABDESLAM MOHAMED AMINE
- SOUKI NOUR

L'objectif de notre projet est de développer un modèle prédictif capable d'estimer la probabilité de défaut des clients. Pour ce faire, nous avons analysé des données financières, créé des ratios pertinents et ajusté un modèle de Cox. Ce modèle nous permettra d'anticiper les événements de défaut et d'aider les institutions à prendre des décisions éclairées.
Nous sommes fiers de présenter ce travail collaboratif réalisé par moi et mon binôme Souki Nour.

PLAN

- INTRODUCTION
- METHODOLOGIE
- TRAVAIL RÉALISÉ & CONCLUSION

ce plan est structuré de 3 crucial parties :
nous allons commencer par une introduction générale à propos ce projet ,par la suite nous allons parler sur la méthodologie qui nous permet de répartir notre travail étape par étape et enfin nous allons parler sur toutes les étapes et finir par une conclusion.

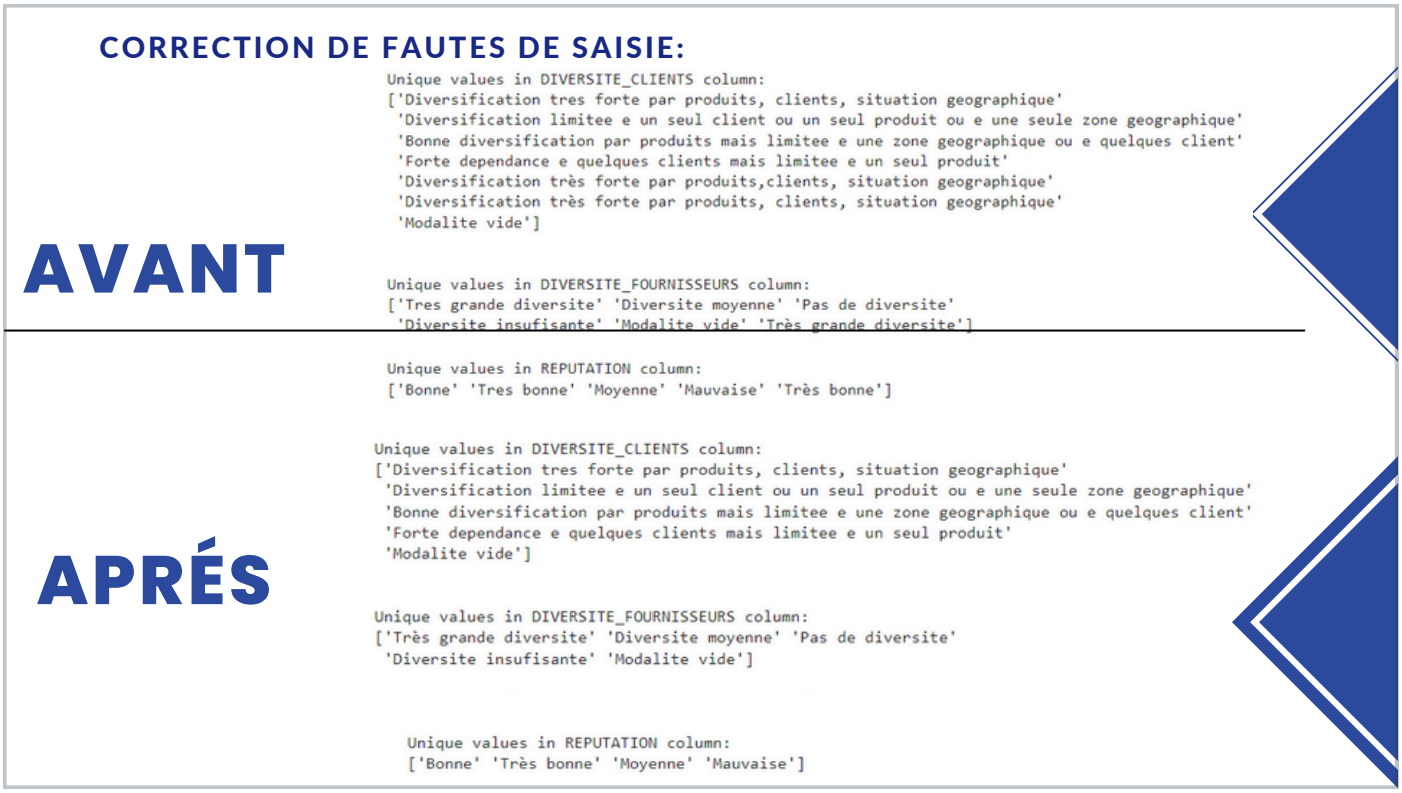
INTRODUCTION

DOMAINE DE LA GESTION DES RISQUES FINANCIERS

Le défaut, défini comme l'incapacité d'une entreprise à honorer ses obligations financières, représente un enjeu crucial pour les investisseurs, les prêteurs et les décideurs économiques. En utilisant des méthodes de modélisation statistique telles que le modèle de Cox, nous cherchons à identifier les facteurs prédictifs du défaut et à évaluer leur impact sur la durée jusqu'au défaut.

MÉTHODOLOGIE

+ PRÉPARATION DE DONNÉES:
  - CORRECTION DE FAUTES DE SAISIE.
  - TRAITEMENT DE VALEURS MANQUANTES.
  - TRAITEMENT DE VALEURS NULLES.
  - CRÉATION DE LA VARIABLE AGE
+ CRÉATION DE RATIOS
  - TRAITEMENT DES VALEURS ABÉRANTES
  - ANALYSE DE RATIOS
  - VISUALISATION DE RATIOS
  - DISCRÉTISATION
+ CALCUL DU WOE & IV
+ MODÉLISATION PRÉDICTIVE
+ VISUALISATION DES RÉSULTATS



nous n'avons pas remarqué cette étape aux autres présentations mais la vérification de données (validation de données) variable par variables est très importante avant de commencer

dans cet diapo on va mettre en évidence les fautes de saisie de notre base de donnée et par la suite les corrections de ces derniers.
donc Avant la Correction : Dans la première partie de la diapositive, vous pouvez remarquez des erreurs.En effet, vous pouvez voir par exemple la colonne Diversité fournisseur il y a deux phrases qui sont identiques.
Après la Correction : La deuxième partie de la diapositive montre les mêmes colonnes après avoir appliqué des corrections.
La correction des erreurs de saisie est une étape cruciale dans le traitement des données. Elle garantit que nos analyses et nos modèles sont basés sur des informations précises et cohérentes.

TRAITEMENT DE VALEURS NULLES:
COMPTER LES ZÉROS POUR CHAQUE VARIABLE

IL EST ACCEPTABLE D'AVOIR DES VALEURS NULLES DANS DES VARIABLES TELLES QUE LES STOCKS, MAIS CE N'EST PAS LE CAS POUR DES VARIABLES COMME LES CAPITAUX PROPRES.

```
numtiers                                  0
Annee                                     0
NUMTIERS_ANNEE                            0
defaut                                 1468
DATE_DE_CREATION_TIERS                    0
DATE_DE_CREATION_ENTREP                   0
CHIFFRE_AFFAIRES                          0
EXCEDENT_BRUT_EXPLOITATION                6
RESULTAT_EXPlOITATION                     4
RESULTAT_NET                             13
FINANCEMENT_PERMANENT                     1
FONDS_DE_ROULEMENT                        0
BESOIN_FONDS_ROULEMENT                    0
CAPITAUX_PROPRES                          4
TRESORIE_NETTE                            1
TOTAL_BILAN                               1
DETTE_FINANCIERE                        563
ACTIF_CIRCULANT                           3
PASSIF_CIRCULANT                          1
TOTAL_ACTIF                               3
TOTAL_PASSIF                              1
DELAI_REGLEMENT_CLIENTS                  32
DELAI_REGLEMENT_FOURNISSEURS              8
AUTO_FINANCEMENT                         18
FRAIS_FINANCIERS                          8
STOCK                                    72
EXPERIENCE_MANAGEMENT_MOYENNE_DIRIGEANT   0
DIVERSITE_CLIENTS                         0
DIVERSITE_FOURNISSEURS                    0
IMPACT_SOCIAUX_ENVIRONNEMENTAL            0
NIVEAU_COMPETITIVITE                      0
QUALITE_INFORMATION_FINANCIERE            0
REPUTATION                                0
STRUCTUREDUMANAGEMENT                     0
SUPPORT                                   0
POSITIONNEMENTMARCHE                      0
Categorie_juridique                       0
Cote en bourse                         1463
Appartenance a un groupe                576
Secteurs                                  0
dtype: int64
```

TRAITEMENT DE VALEURS NULLES:

AVANT

	numtiers	Annee	CAPITAUX_PROPRES
375	5200001299269	2015	0.00
1103	5200001299269	2016	54245661.03
376	5200001299269	2017	40299730.56
1104	5200001299269	2018	57384717.74
1105	5200001299269	2019	39770154.20
399	5200001300441	2015	54175671.89
1117	5200001300441	2016	0.00
1118	5200001300441	2017	60043790.03
400	5200001300441	2018	57058559.74
1119	5200001300441	2019	69106903.78
1292	5200007203033	2015	35875963.98
778	5200007203033	2016	0.00
779	5200007203033	2017	76025022.54
780	5200007203033	2018	0.00
1293	5200007203033	2019	79319783.03

APRÉS

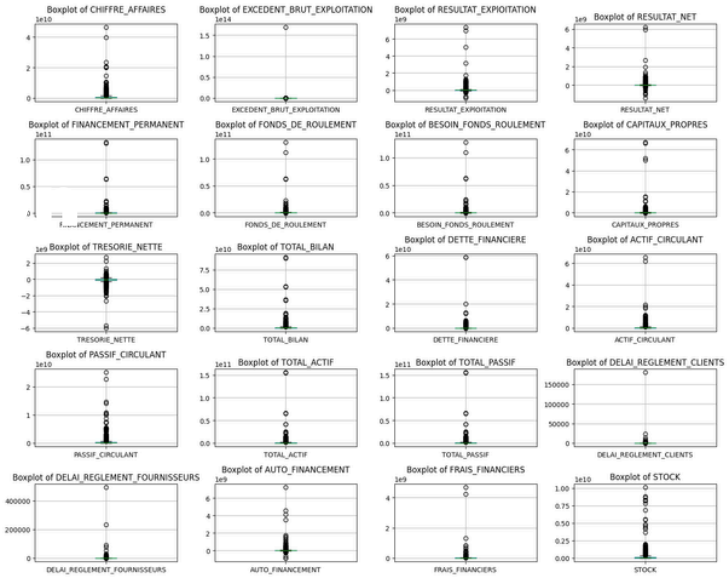
	numtiers	Annee	CAPITAUX_PROPRES
375	5200001299269	2015	54245661.03
1103	5200001299269	2016	54245661.03
376	5200001299269	2017	40299730.56
1104	5200001299269	2018	57384717.74
1105	5200001299269	2019	39770154.20
399	5200001300441	2015	54175671.89
1117	5200001300441	2016	54175671.89
1118	5200001300441	2017	60043790.03
400	5200001300441	2018	57058559.74
1119	5200001300441	2019	69106903.78
1292	5200007203033	2015	35875963.98
778	5200007203033	2016	35875963.98
779	5200007203033	2017	76025022.54
780	5200007203033	2018	76025022.54
1293	5200007203033	2019	79319783.03

CRÉATION D'UNE VARIABLE 'AGE':

```
from datetime import datetime

# Assuming the "creation_date" column is in datetime format
data['DATE_DE_CREATION_ENTREP'] = pd.to_datetime(data['DATE_DE_CREATION_ENTREP'])

# Calculate the age based on the current date
current_date = datetime.now()
data['age'] = (current_date - data['DATE_DE_CREATION_ENTREP']).dt.days/365.25

# Print the DataFrame with the new 'age' column
print(data['age'])
```

```
0       42.272416
1       42.272416
2       42.272416
3       46.242300
4       46.242300
          ...
1516    32.208077
1517    17.563313
1518    22.318960
1519    22.318960
1520    30.365503
Name: age, Length: 1521, dtype: float64
```

VISUALISATION DE LA DISTRIBUTION DES DONNÉES:



LES DONNÉES SONT DE DIFFÉRENTES ÉCHELLES
=> SOLUTION: CRÉATION DE RATIOS

données de différentes échelles
==> ratios

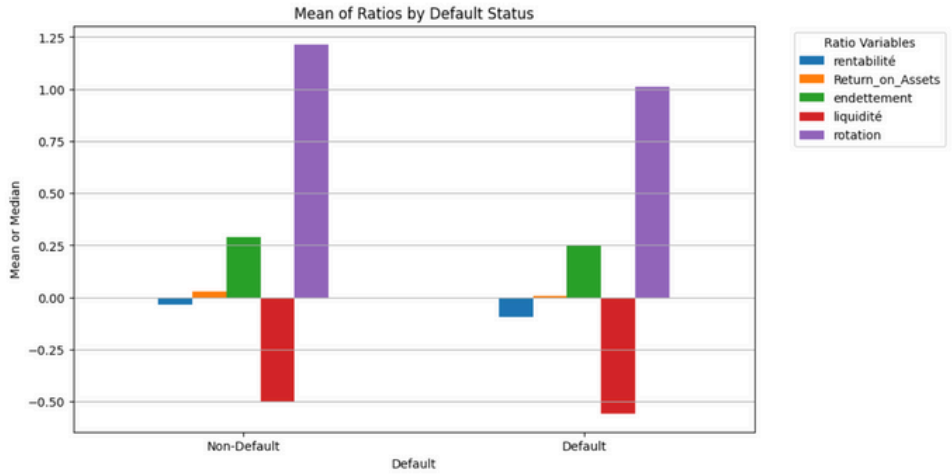
RATIOS:

```
In [35]: data[ratios_columns].describe()
```

Out[35]:

	rentabilité	Return_on_Assets	endettement	liquidité	rotation
count	1521.000000	1521.000000	1521.000000	1521.000000	1521.000000
mean	-0.036972	0.028000	0.289875	-0.503203	1.208791
std	1.619463	0.082679	0.228114	1.762072	1.011653
min	-48.577726	-1.276431	-2.990561	-27.523223	0.000694
25%	0.006488	0.006527	0.153641	-0.675509	0.741968
50%	0.017565	0.019439	0.258256	-0.240616	1.040458
75%	0.047752	0.049928	0.412539	0.009711	1.520506
max	15.941564	0.857895	0.962572	17.906374	26.485113

LES RATIOS FINANCIERS:



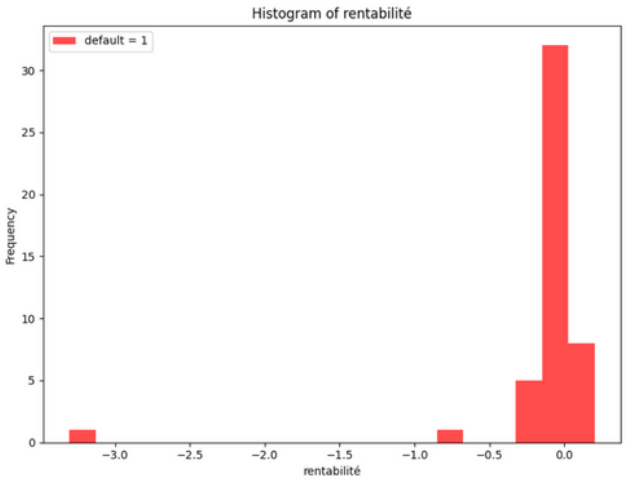
DISCRÉTISATION:

```
from sklearn.preprocessing import KBinsDiscretizer
# Initialize the discretizer with desired parameters
discretizer = KBinsDiscretizer(n_bins=5, encode='ordinal', strategy='quantile')

# Fit and transform the selected columns
discretized_data = discretizer.fit_transform(data[ratios_columns])

# Convert the discretized data back to a DataFrame
discretized_df = pd.DataFrame(discretized_data, columns=ratios_columns)
ratios_df=data
# Merge the discretized data back into the original DataFrame
ratios_df[ratios_columns] = discretized_df
```

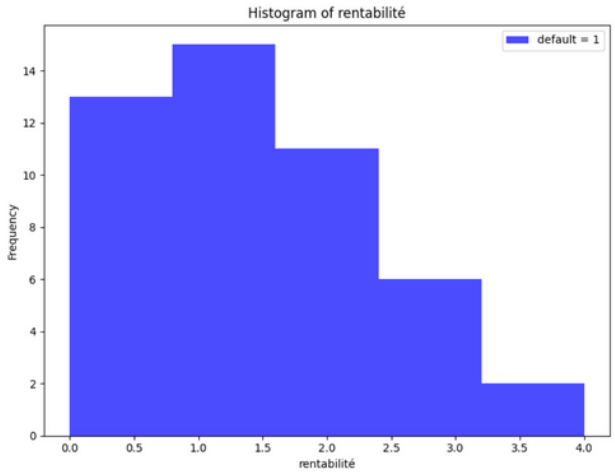
	rentabilité	Return_on_Assets	endettement	liquidité	rotation
0	0.0	0.0	4.0	2.0	2.0
1	1.0	1.0	4.0	2.0	3.0
2	1.0	1.0	4.0	2.0	2.0
3	0.0	0.0	1.0	3.0	0.0
4	1.0	0.0	0.0	2.0	0.0
...	...	...	...	...	...
1516	4.0	4.0	3.0	1.0	2.0
1517	1.0	1.0	1.0	2.0	4.0
1518	2.0	2.0	2.0	1.0	2.0
1519	1.0	1.0	1.0	1.0	3.0
1520	0.0	0.0	0.0	1.0	0.0

1521 rows × 5 columns

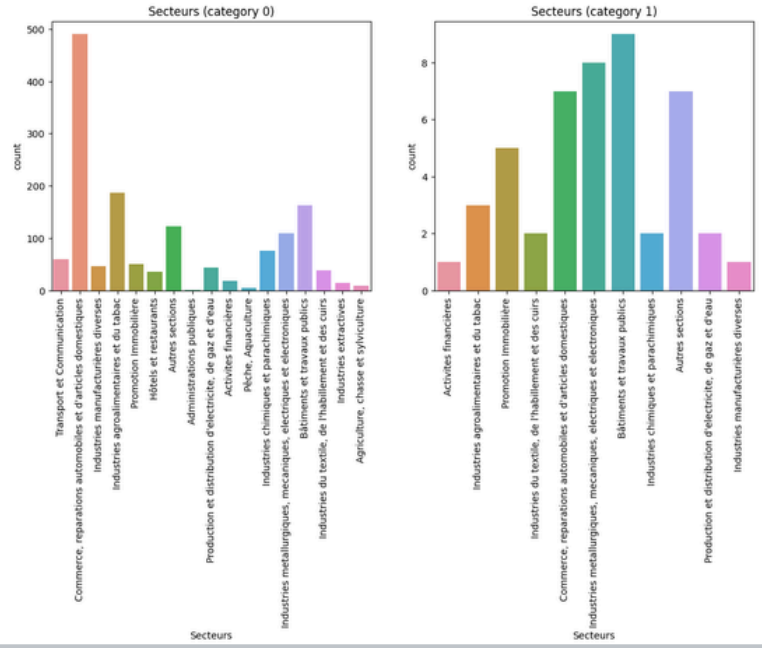
AVANT DISCRÉTISATION:



APRÉS DISCRÉTISATION:



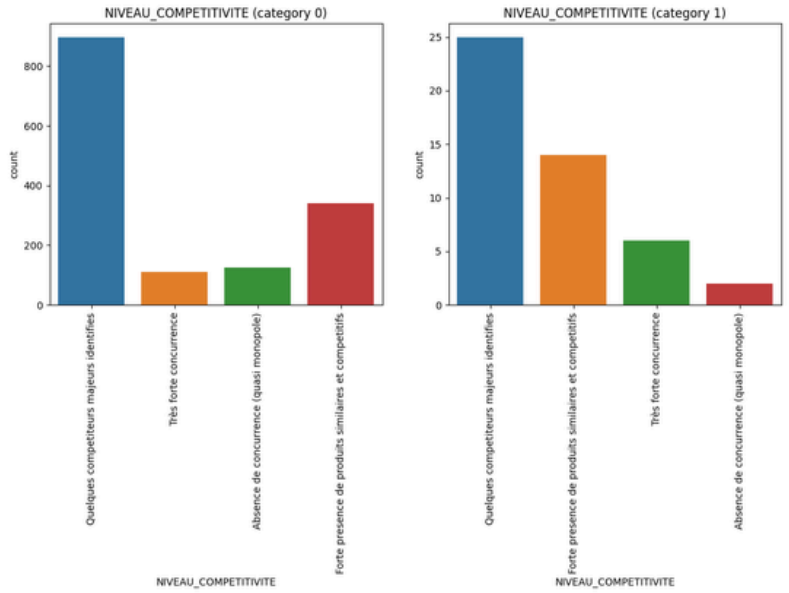
VISUALISATION DE LA DISTRIBUTION DES VARIABLES CATÉGORIELLES :
DÉFAUT =0 VS DÉFAUT= 1



la visualisation de ces deux variables (ce graphe et le graphe suivant) nous donne une idée sur leur IV considérable

VISUALISATION DE LA DISTRIBUTION DES VARIABLES CATÉGORIELLES :

DÉFAUT =0 VS DÉFAUT= 1



CALCUL DE L'INFORMATION VALUE(IV)

Entrée [415]:

```
def calculate_woe_iv(feature_data, target_data):
    total_events = np.sum(target_data)
    total_non_events = len(target_data) - total_events
    woe_iv = 0

    for value in feature_data.unique():
        event_count = np.sum(target_data[feature_data == value])
        non_event_count = np.sum((feature_data == value) & (target_data == 0))

        if event_count == 0 or non_event_count == 0:
            continue

        proportion_of_events = event_count / total_events
        proportion_of_non_events = non_event_count / total_non_events
        woe = np.log(proportion_of_events / proportion_of_non_events)
        iv = (proportion_of_events - proportion_of_non_events) * woe
        woe_iv += iv

    return woe_iv

def calculate_iv(df, target):
    iv_values = {}

    for feature in df.columns:
        if feature != target:
            feature_data = df[feature]
            target_data = df[target]

            if feature_data.dtype == 'object':
                feature_data = feature_data.astype('category').cat.codes

            iv_values[feature] = calculate_woe_iv(feature_data, target_data)

    return iv_values
```

CALCUL DE L'INFORMATION VALUE(IV)

```
IV for Cote en bourse: 0.0005388562524655654
IV for Appartenance a un groupe: 8.274486910661384e-05
IV for Secteurs: 0.454806831469619
IV for rentabilité: 0.383421821182222
IV for Return_on_Assets: 0.3087993667144464
IV for endettement: 0.21132198095272384
IV for liquidité: 0.12237599719896897
IV for rotation: 0.41656917605126165
```

voici le calcul de de l'information value pour chaque variable .vous pouvez remarquer que l'information value pour la majorité de ces derniers est entre 0.1 et 0.5 alors Prédicteur modéré c -a -d La variable a un pouvoir prédictif modéré et peut être considérée comme utile pour prédire la variable cible.

donc ces variables vont être retenues et passées en entrée au modèle de Cox

MODÈLE COX:

In [54]:
```
# Select relevant columns for survival analysis, such as 'defaut' (event), 'age', and other relevant features
selected_columns = [ 'defaut','NIVEAU_COMPETITIVITE', 'QUALITE_INFORMATION_FINANCIERE', 'REPUTATION',
       'Secteurs', 'rentabilité', 'Return_on_Assets', 'endettement',
       'liquidité', 'rotation','age']

data_selected = data[selected_columns]
```

```
# Application de SMOTE
smote = SMOTENC(categorical_features=categorical_features_indices)
```

```
X_NC, y_NC = smote.fit_resample(X, y)

X_train, X_test, y_train, y_test = train_test_split(X_NC, y_NC, test_size=0.2, random_state=42
```

```
# Fit the Cox proportional hazards model
cox_model = CoxPHFitter()
cox_model.fit(X_train, duration_col='age', event_col='defaut')
```

imbalanced data => smote

RÉSULTAT DU MODÈLE DE COX:

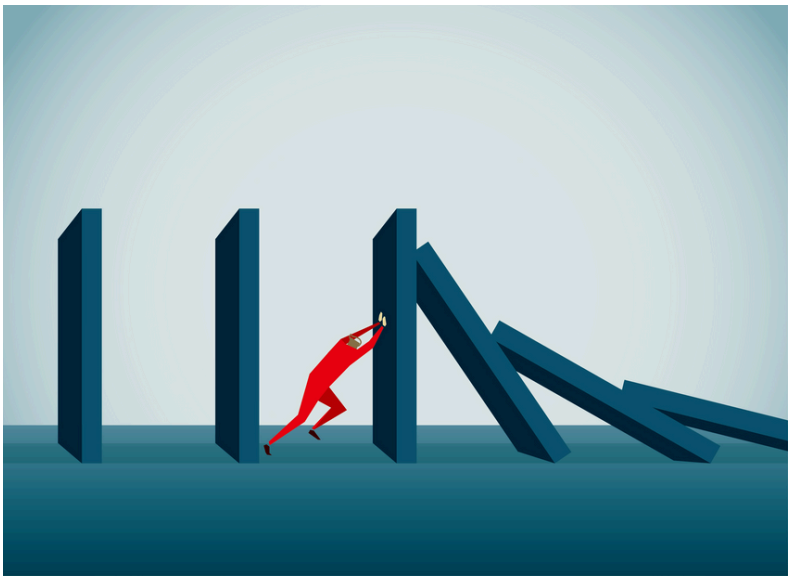
	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
NIVEAU_COMPETITIVITE	-0.24	0.78	0.04	-0.32	-0.16	0.72	0.85	0.00	-5.96	<0.005	28.52
QUALITE_INFORMATION_FINANCIERE	0.20	1.23	0.14	-0.06	0.47	0.94	1.60	0.00	1.49	0.14	2.88
REPUTATION	0.37	1.45	0.03	0.30	0.44	1.35	1.55	0.00	10.92	<0.005	89.83
Secteurs	0.00	1.00	0.01	-0.01	0.02	0.99	1.02	0.00	0.24	0.81	0.30
rentabilité	-0.47	0.62	0.06	-0.60	-0.34	0.55	0.71	0.00	-7.24	<0.005	41.06
Return_on_Assets	0.32	1.38	0.07	0.19	0.46	1.21	1.58	0.00	4.73	<0.005	18.77
endettement	-0.24	0.79	0.03	-0.29	-0.19	0.75	0.83	0.00	-8.87	<0.005	60.31
liquidité	-0.05	0.95	0.03	-0.10	-0.00	0.91	1.00	0.00	-2.00	0.05	4.46
rotation	-0.05	0.95	0.03	-0.11	0.00	0.89	1.00	0.00	-1.83	0.07	3.88

Concordance	0.73

PRÉDICTION DU MODÈLE :

age	defaut	predicted_survival
16.000000	0	23.635481
40.268516	1	23.941081
15.288159	0	24.335760

LIMITATIONS ET DÉFIS:



comme tout modèle statistique, le modèle de cox comporte des limitations et des défis à prendre en compte. Voici quelques-uns :

Proportionnalité des risques : qui signifie que le ratio des risques entre deux groupes est constant dans le temps. Si cette hypothèse n'est pas vérifiée, les résultats du modèle peuvent être biaisés.
Censure et troncature : Le modèle de Cox est adapté aux données de survie censurées, mais il peut être moins efficace lorsque les données sont fortement censurées ou tronquées.
Variables temps-dépendantes : Le modèle de Cox n'est pas conçu pour gérer efficacement les variables qui changent dans le temps pour chaque individu (variables temps-dépendantes). Cela peut limiter sa capacité à modéliser des événements complexes.
Surajustement : Comme avec tout modèle de régression, il existe un risque de surajustement si le modèle est trop complexe par rapport à la taille de l'échantillon ou si trop de variables sont incluses sans justification théorique ou empirique solide.
Dépendance aux données manquantes : La présence de données manquantes peut poser des problèmes dans l'analyse de survie avec le modèle de Cox, nécessitant des techniques appropriées pour gérer ces données manquantes.
Validation du modèle : Valider et évaluer la performance du modèle de Cox peut être complexe, en particulier lorsque l'événement d'intérêt est rare ou lorsque le suivi des individus est de courte durée.
Interprétation des coefficients : Interpréter les coefficients du modèle de Cox peut être délicat, en particulier lorsque plusieurs variables sont incluses et interagissent entre elles.

CONCLUSION :

En conclusion, le modèle de Cox a démontré une bonne capacité prédictive pour évaluer le risque de défaut malgrè les limitations liées à la base de données. Les variables telles que les ratios créés, la réputation et le niveau de compétitivité se sont révélées significatives pour prédire le moment du défaut, offrant ainsi une base rigoureuse pour la prise de décision en matière de gestion des risques.

base de donées


MERCI DE VOTRE
ATTENTION