



Analyse prédictive de churn dans le domaine des télécommunications

Encadré Par :

Madame Ferihane Kboubi

Réalisé Par :

Chaima Ammeri

PLAN



1

Introduction

2

Solution proposée

3

Mise en place

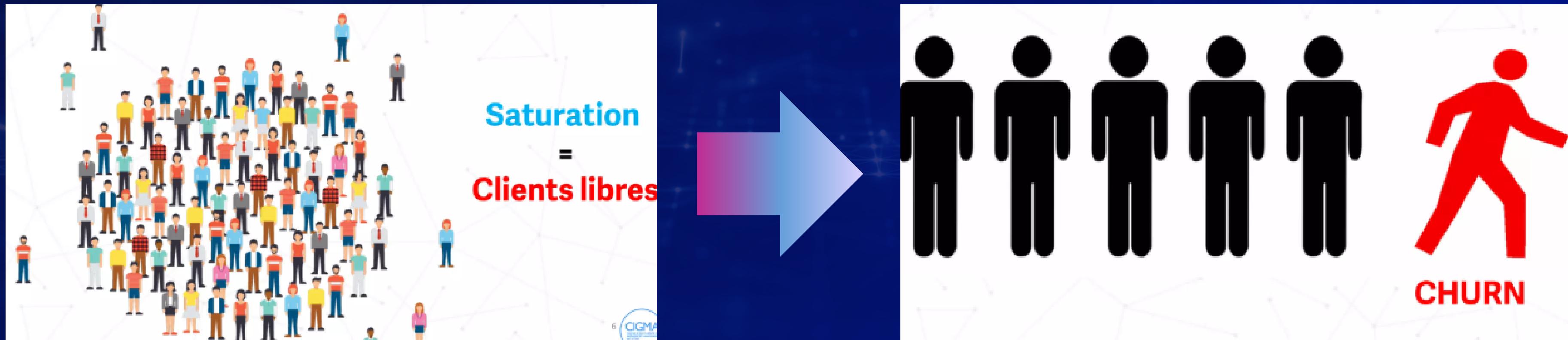
4

Conclusion et Perspectives

Introduction

Le secteur des télécommunications poursuit sa transformation.

+92% de taux de penetration du marché des Télécommunications



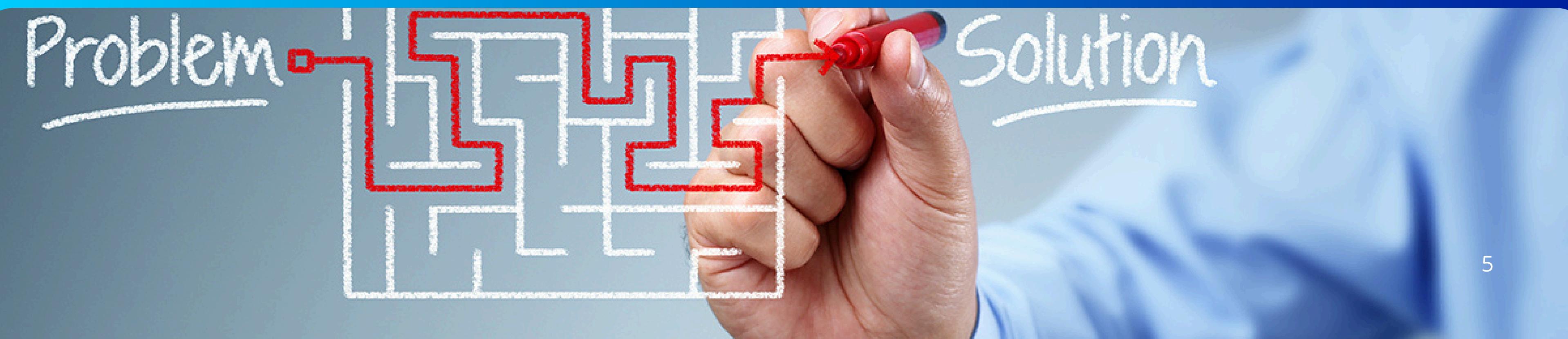
Définition :

Churn ou [Attrition] : Exprime le taux de perdition de clients pour une entreprise
Les couts d'acquisition d'un nouveau client présente **5X** Les couts de maintien d'un client actuel

→ L'entreprise doit se concentrer sur ces clients avant qu'ils ne partent



Solution proposée



Solution proposée

Objectifs de la solution

L'objectif est de prédire les clients susceptibles de résilier leur abonnement (churn) en utilisant des techniques de data mining et machine learning. Cette solution vise à identifier les facteurs influençant la fidélité des clients et à anticiper leur départ pour optimiser les actions de rétention.



Modèle CRISP-DM

Le projet suit la méthodologie CRISP-DM en six étapes :

- 1. Compréhension des objectifs et des données :** Identifier les clients à risque de churn et explorer les variables du dataset (démographie, services, modes de paiement).
- 2. Préparation des données :** Nettoyage des données avec Pandas, encodage et normalisation via Scikit-learn, et visualisation des relations avec Matplotlib et Seaborn.
- 3. Modélisation :** Dans cette phase, des modèles prédictifs ou descriptifs sont construits à l'aide de techniques d'analyse de données adaptées aux objectifs du projet. Application Random Forest et Gradient Boosting
- 4. Évaluation des modèles :** Comparaison des modèles à l'aide des métriques de performance (AUC-ROC, précision, rappel, F1-score).
- 5. Déploiement et Suivi :** Mise en production du modèle via Flask ou Streamlit et suivi continu des performances.
- 6. Analyse et recommandations :** Identification des actions de rétention basées sur les prédictions de churn.

Bénéfices attendus

- **Prédiction fiable du churn** : Identifiant les clients à risque avant qu'ils ne quittent, ce qui permet d'agir proactivement.
- **Optimisation des coûts de rétention** : En concentrant les efforts sur les clients à risque, l'entreprise maximise l'impact des actions de fidélisation.
- **Amélioration de la satisfaction client** : En proposant des offres personnalisées aux clients à risque, l'entreprise peut améliorer la fidélité et la satisfaction client.



Mise en place



Mise en place

Compréhension du problème métier

La perte de la clientèle ou d'abonnés est toujours un problème grave pour l'industrie des télécommunications, car les clients n'hésitent pas à désabonner ou de changer l'opérateur, s'ils ne trouvent pas ce qu'ils recherchent. Les clients veulent certainement des prix compétitifs, et de la valeur ajoutée pour l'argent qu'ils payent et surtout, un service de haute qualité.

Le churn est directement lié à la satisfaction du client. C'est un fait connu que le coût de l'acquisition d'un client est beaucoup plus élevé que le coût de la fidélisation d'un client, ce qui fait de la rétention des clients un prototype d'entreprise crucial. Il n'existe pas de modèle standard permettant de résoudre avec précision les problèmes de clientèle des fournisseurs de services telecoms mondiaux.

L'analyse Big Data avec le Machine Learning et Data mining s'est révélée être un moyen efficace d'identifier et prédire le désabonnement des clients. Afin d'anticiper la rupture des clients avant qu'elle se produise.

Mise en place

Compréhension des données

Description du dataset

Le dataset provient de Kaggle et contient des informations sur les abonnés d'une entreprise de télécommunications. Il comprend 21 colonnes couvrant plusieurs aspects des clients :

- **Démographie des clients**
- **Détails de l'abonnement**
- **Services utilisés**
- **Modes de paiement**
- **Variable cible (Churn)**

Feature vectors	Types
Customer id	alpha numeric
gender	categorical
Senior citizen	numeric
Partner	categorical
Dependents	categorical
tenure	numeric
Phone service	categorical
Multiple lines	categorical
Internet service	categorical
Online security	categorical
Online backup	categorical
Device protection	categorical
Tech support	categorical
Streaming Tv	categorical
Streaming movies	categorical
Contract	categorical
Paperless billing	categorical
Payment method	categorical
Monthly charges	numeric
Total charges	numeric
Churn	categorical



Objectifs du dataset :

- Analyser les facteurs influençant la fidélité des clients.
- Construire des modèles prédictifs pour anticiper le churn.
- Améliorer les stratégies de rétention en ciblant les clients à risque.

Mise en place

Compréhension des données

Analyse Exploratoire des données (EDA) :

Une analyse exploratoire des données (EDA) a été menée pour comprendre la structure et les particularités du dataset.

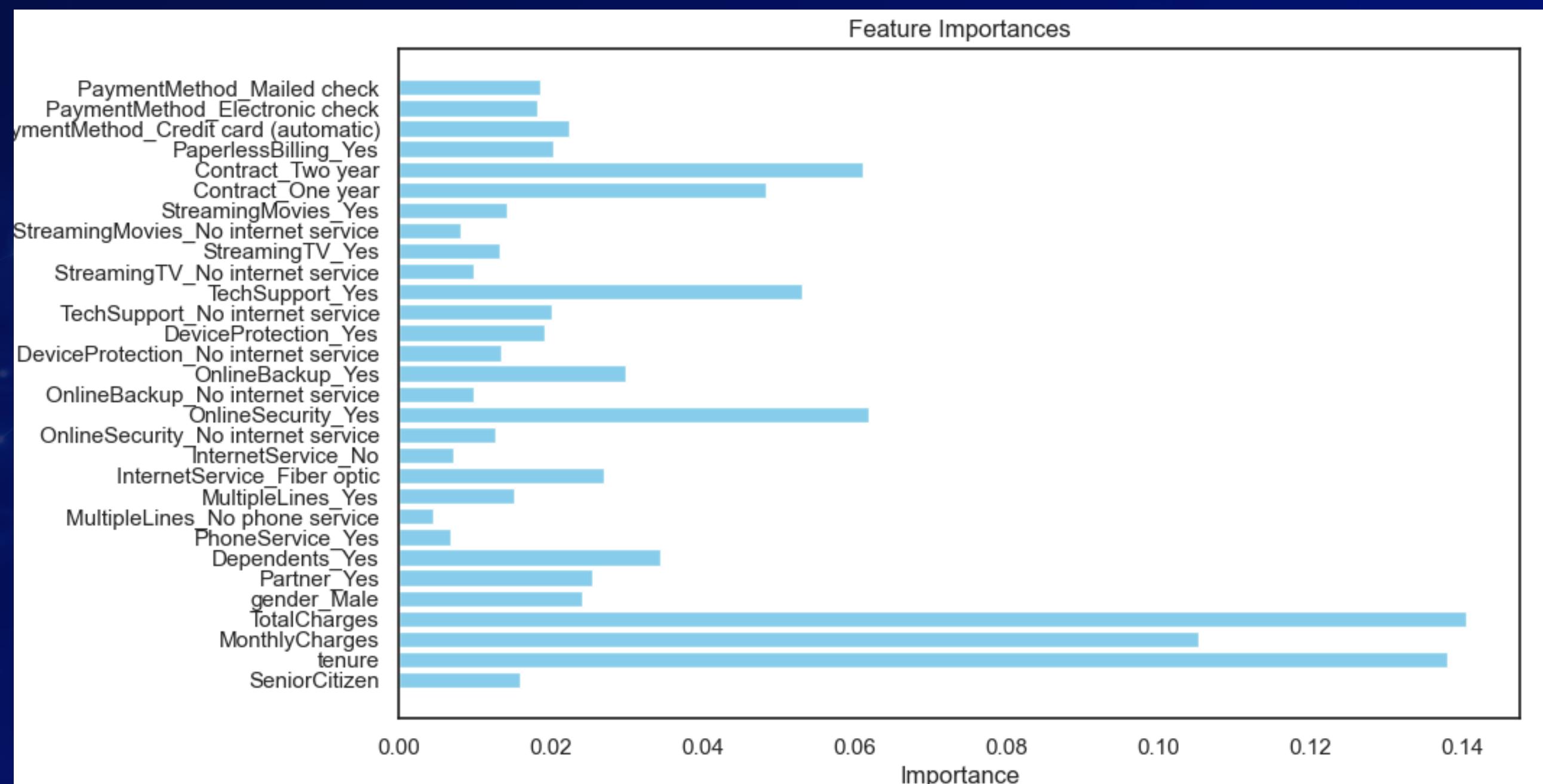
- Un aperçu sur l'ensemble des données:

	customerID	gender	SeniorCitizen	...	MonthlyCharges	TotalCharges	Churn
0	7590-VHVEG	Female	0	...	29.85	29.85	No
1	5575-GNVDE	Male	0	...	56.95	1889.5	No
2	3668-QPYBK	Male	0	...	53.85	108.15	Yes
3	7795-CFOCW	Male	0	...	42.30	1840.75	No
4	9237-HQITU	Female	0	...	70.70	151.65	Yes

Mise en place

Compréhension des données

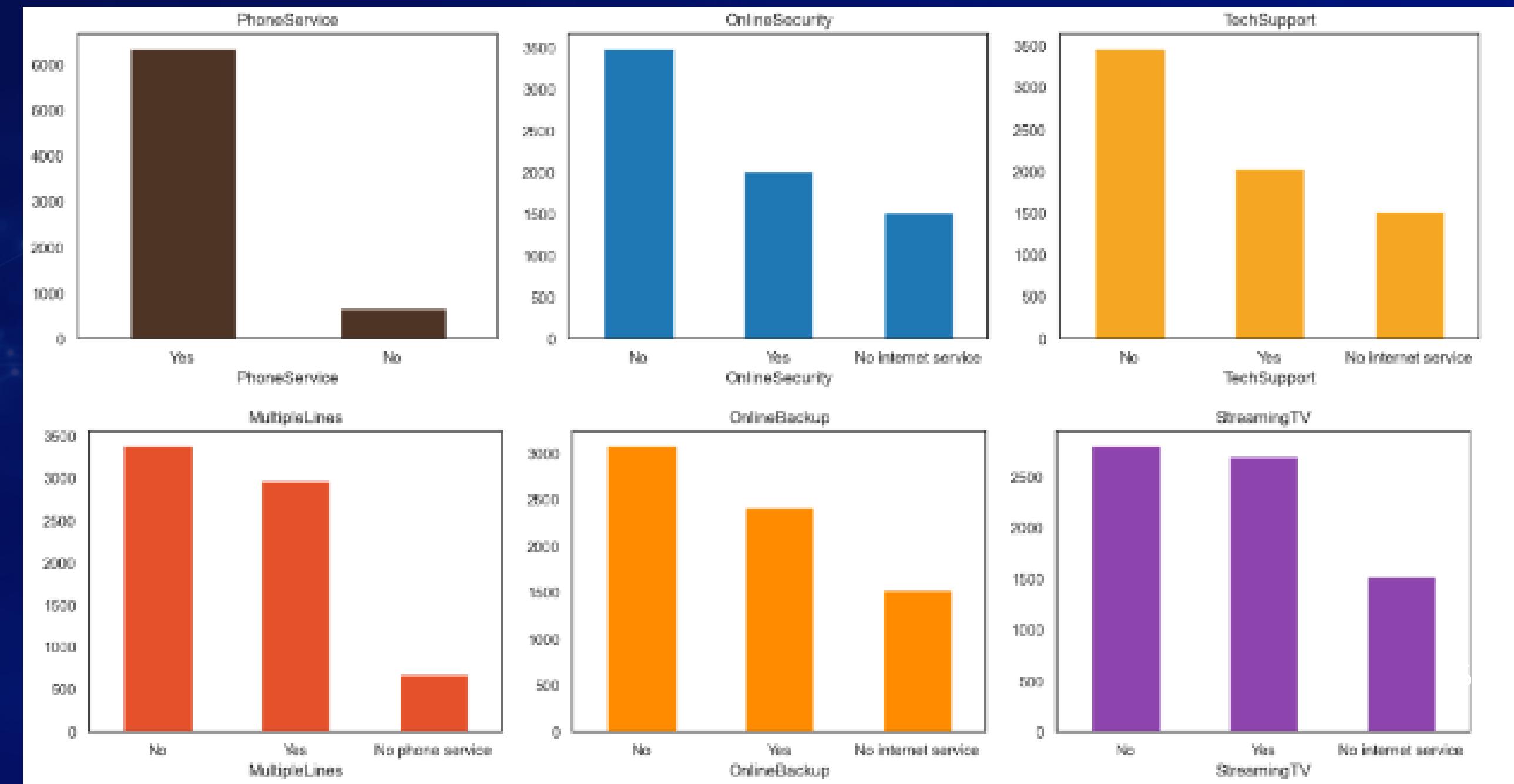
- Importances des variables les plus prédictives:



Mise en place

Compréhension des données

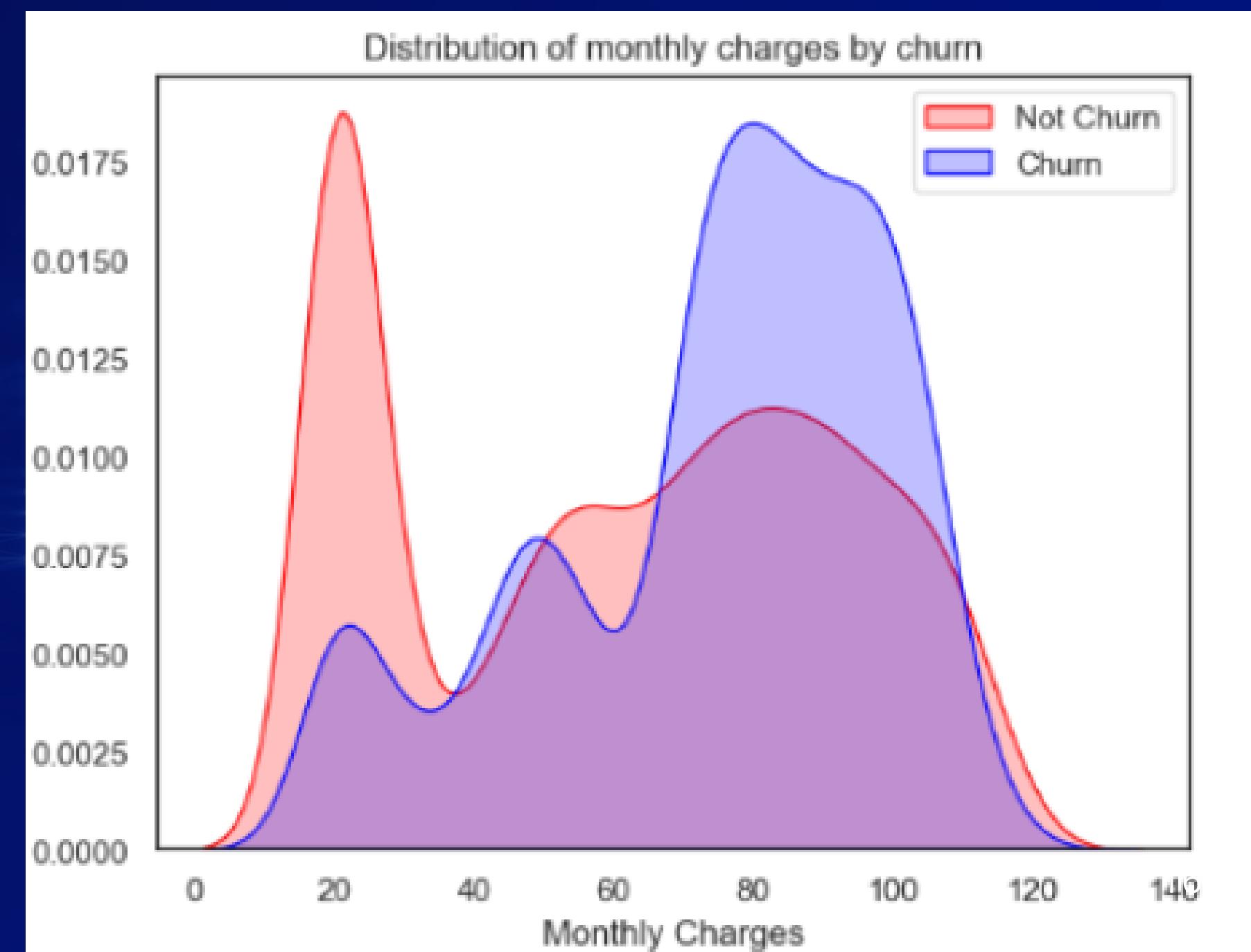
- La distribution des caractéristiques :



Mise en place

Compréhension des données

- La distribution des caractéristiques :



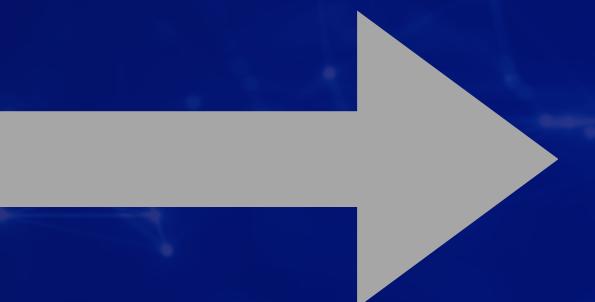
Mise en place

Préparation des données

- **Gestion des valeurs manquantes :**

Notre dataset contient 11 valeurs manquantes on les a supprimé afin d'assurer l'intégrité et la qualité des données avant l'analyse prédictive

customerID	0
gender	0
SeniorCitizen	0
Partner	0
Dependents	0
tenure	0
PhoneService	0
MultipleLines	0
InternetService	0
OnlineSecurity	0
OnlineBackup	0
DeviceProtection	0
TechSupport	0
StreamingTV	0
StreamingMovies	0
Contract	0
PaperlessBilling	0
PaymentMethod	0
MonthlyCharges	0
TotalCharges	11
Churn	0



customerID	0
gender	0
SeniorCitizen	0
Partner	0
Dependents	0
tenure	0
PhoneService	0
MultipleLines	0
InternetService	0
OnlineSecurity	0
OnlineBackup	0
DeviceProtection	0
TechSupport	0
StreamingTV	0
StreamingMovies	0
Contract	0
PaperlessBilling	0
PaymentMethod	0
MonthlyCharges	0
TotalCharges	0
Churn	0
InternetUsage	0

Mise en place

Préparation des données

- **Encodage des variables :**

De nombreux algorithmes d'apprentissage automatique ne peuvent pas fonctionner directement sur les données catégorielles. Ils exigent que toutes les variables d'entrée et les variables de sortie soient numériques.

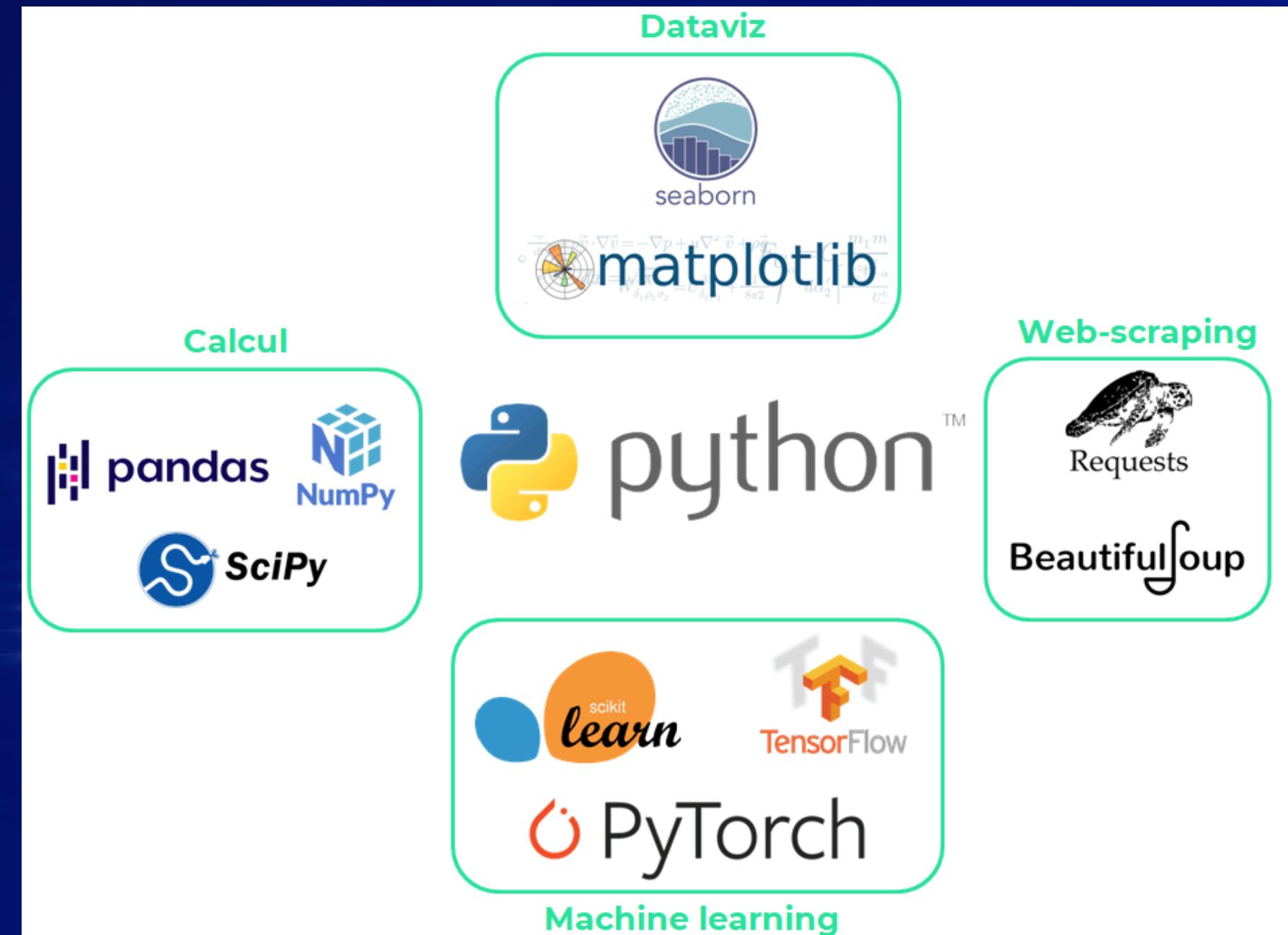
SeniorCitizen	tenure	MonthlyCharges	TotalCharges	Churn	gender_Male	Partner_Yes	Dependents_Yes	PhoneService_Yes	MultipleLines_No phone service
0	0	1	29	29	0	0	1	0	0
1	0	34	56	1889	0	1	0	0	1
2	0	2	53	108	1	1	0	0	1
3	0	43	42	1840	0	1	0	0	0
4	0	2	70	151	1	0	0	0	1
StreamingTV_No Internet service	StreamingTV_Yes	StreamingMovies_No Internet service	StreamingMovies_Yes	Contract_One year	Contract_Two year	PaperlessBilling_Yes	PaymentMethod_Credit card (automatic)		
0	0	0	0	0	0	1	0	0	
0	0	0	0	0	1	0	0	0	
0	0	0	0	0	0	1	0	0	
0	0	0	0	0	1	0	0	0	
0	0	0	0	0	0	1	0	0	

Mise en place

Modélisation

Bibliothèques utilisées

- **Pandas** : Utilisé pour la gestion des données (chargement, nettoyage, transformation).
- **NumPy** : Pour les opérations numériques et le traitement des matrices.
- **Scikit-learn** : Contient des outils pour la normalisation des données, l'encodage des variables catégorielles (ex. LabelEncoder), et la gestion des valeurs manquantes (ex. SimpleImputer).
- **Matplotlib et Seaborn** : Pour la visualisation des données, permettant de comprendre les distributions et les relations entre variables.



Mise en place

Modélisation

On a utilisé 2 modèles de machine learning :

Random Forest : Random Forest est un algorithme d'apprentissage automatique qui utilise un ensemble d'arbres de décision pour effectuer des tâches de classification et de régression. Il fonctionne en construisant plusieurs arbres à partir de sous-ensembles aléatoires de données et d'attributs, ce qui permet de réduire le risque de surapprentissage.

Gradient Boosting Est un algorithme d'apprentissage supervisé basé sur l'optimisation de l'erreur en ajoutant successivement des modèles faibles (généralement des arbres de décision) de manière itérative. Chaque modèle corrige les erreurs du précédent en se concentrant sur les instances mal classées.

Implémentation du modèle : Random Forest

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix, precision_score, recall_score, f1_score, roc_auc_score
from sklearn.model_selection import RandomizedSearchCV
import matplotlib.pyplot as plt

# Entrainer un modèle Random Forest
rf = RandomForestClassifier(random_state=42)
rf.fit(X_train_sm, y_train_sm)

# Prédictions sur les données de test
y_pred = rf.predict(X_test_scaled)
```

Implémentation du modèle : Gradient Boosting

```
[69]: from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix, precision_score, recall_score, f1_score, roc_auc_score
import matplotlib.pyplot as plt

# Initialiser et entraîner le modèle Gradient Boosting
gb = GradientBoostingClassifier(random_state=42)
gb.fit(X_train_sm, y_train_sm)

# Prédictions sur les données de test
y_pred = gb.predict(X_test_scaled)
```

Évaluation des Modèles



Évaluation des Modèles

- Les modèles sont évalués en fonction de leur Accuracy F1-score, précision et rappel. Ces métriques permettent de mesurer la performance des modèles pour prédire correctement les clients à risque de churn.
- Les modèles mis en œuvre ont déjà été décrits et analysés séparément, mais l'objectif principal de cette recherche était d'identifier quel modèle peut être considéré comme le meilleur pour résoudre ce problème.

Évaluation des Modèles

```
Classification Report:  
precision    recall   f1-score   support  
  
          0       0.88      0.70      0.78      1033  
          1       0.47      0.73      0.57      374  
  
accuracy                           0.71      1407  
macro avg       0.67      0.71      0.67      1407  
weighted avg     0.77      0.71      0.72      1407
```

```
Confusion Matrix:  
[[720 313]  
 [101 279]]
```

```
Métriques supplémentaires :  
Accuracy: 0.7058  
Précision: 0.4659  
Recall: 0.7299  
F1-Score: 0.5687  
ROC-AUC: 0.7735
```

• Random Forest

• XGBOOST

```
Classification Report:  
precision    recall   f1-score   support  
  
          0       0.82      0.88      0.84      1033  
          1       0.57      0.46      0.51      374  
  
accuracy                           0.76      1407  
macro avg       0.69      0.67      0.68      1407  
weighted avg     0.75      0.76      0.76      1407
```

```
Confusion Matrix:  
[[904 129]  
 [203 171]]
```

```
Métriques supplémentaires :  
Accuracy: 0.7640  
Précision: 0.5700  
Recall: 0.4572  
F1-Score: 0.5074  
ROC-AUC: 0.7666
```

Évaluation des Modèles

Métrique	Random Forest	Gradient Boosting
Précision	0.71	0.76
Rappel (Classe 0)	0.88	0.88
Rappel (Classe 1)	0.46	0.46
F1-Score (Global)	0.57	0.51
ROC-AUC	0.77	0.77

Selon le tableau précédent, **Gradient Boosting** semble légèrement plus performant que **Random Forest** en termes de précision globale (0.76 contre 0.71), mais les deux modèles affichent des performances similaires dans la capacité à séparer les classes, comme en témoigne leur score AUC presque identique (0.77).

Conclusion

Ce projet permet de prédire le churn des clients d'une entreprise de télécommunications en utilisant des techniques de data mining et machine learning. En analysant les données démographiques, des services et modes de paiement, nous avons développé une solution permettant d'anticiper les départs et d'optimiser les stratégies de rétention. Les modèles appliqués (Random Forest et XGBOOST) ont fourni des résultats fiables pour une prise de décision en temps réel.

Perspectives

1. **Amélioration continue** : Suivi et mise à jour du modèle avec de nouvelles données pour garantir sa précision.
 2. **Exploration de nouveaux modèles** : Tests d'approches avancées comme les réseaux de neurones profonds.
 3. **Extension à d'autres secteurs** : Adaptation du modèle pour prédire le churn dans d'autres industries (bancaire, e-commerce).
 4. **Personnalisation de la fidélisation** : Développement de stratégies ciblées pour chaque segment de clients à risque.
- Ce projet ouvre la voie à une gestion proactive du churn et à une meilleure fidélisation des clients.

-



Merci !