الكلية متعددة التخصصات -ورزازات
ⵜⴰⵖⵉⵍⵜ ⵜⴰⵣⵜⵉⵔⵜ – ⵡⵔⵣⵣⴰⵣⴰⵜ
FACULTE POLYDISCIPLINAIRE DE OUARZAZATE

Master Mathématiques Appliquées pour la Science de Données

Progiciel de gestion Intégré
Rapport

Analyse du Taux de désabonnement des clients et prédiction

Élèves :
Youssef Moknia

Enseignant :
M. Charaf Hamidi

20 janvier 2024

Table des matières

Table des figures

1 Introduction

Dans le monde concurrentiel d'aujourd'hui, la rétention des clients est un enjeu majeur pour de nombreuses entreprises. Le taux de désabonnement des clients, ou "churn", est un indicateur clé de la satisfaction des clients et de la performance de l'entreprise. Dans ce projet, nous avons cherché à comprendre les facteurs qui contribuent au désabonnement des clients et à développer des stratégies pour réduire ce taux.

Le contexte de ce projet est une entreprise de télécommunications où nous avons accès à un ensemble de données contenant des informations sur les clients, leurs utilisations des services et leur statut d'abonnement. La motivation de ce projet est d'aider l'entreprise à mieux comprendre ses clients, à améliorer la satisfaction des clients et à augmenter la rétention des clients.

Les objectifs de ce projet étaient de réaliser une analyse exploratoire des données pour identifier les caractéristiques importantes, de développer un modèle prédictif pour prédire le désabonnement des clients, et d'interpréter les résultats pour formuler des recommandations stratégiques. Nous avons testé l'hypothèse que certaines caractéristiques des clients, comme la durée de l'abonnement, le type de contrat et la méthode de paiement, ont un impact significatif sur le taux de désabonnement.

2 Méthodologie

Dans le cadre de ce projet, nous avons suivi une série d'étapes méthodologiques pour analyser et modéliser les données. Voici une description détaillée de ces étapes :

2.1 Définition du Problème

L'objectif principal de ce projet est de prédire si un client individuel va se désabonner ou non. Pour ce faire, nous avons opté pour l'utilisation de modèles d'apprentissage automatique. Ces modèles ont été entraînés sur 80% des données disponibles, tandis que les 20% restants ont été utilisés pour évaluer la performance des modèles.

Une question secondaire que nous avons abordée était d'identifier les caractéristiques spécifiques des clients qui contribuent le plus au taux de désabonnement. Ces informations peuvent être utilisées pour élaborer des stratégies de rétention ciblées.

2.2 Collecte des Données

L'ensemble de données utilisé dans ce projet provient d'une entreprise de télécommunications et contient 7043 entrées. Chaque entrée représente un client unique, identifié par `customerID`. Les données comprennent 21 caractéristiques, notamment le type de contrat, la durée de l'abonnement, le mode de paiement, etc. Nous avons exploré ces données pour mieux comprendre la nature des informations disponibles.

```
In [4]: df
```

Out[4]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechS
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	
...	...	...	...	...	...	...	...	...	...	...	...	...	
7038	6840-RESVB	Male	0	Yes	Yes	24	Yes	Yes	DSL	Yes	...	Yes	
7039	2234-XADUH	Female	0	Yes	Yes	72	Yes	Yes	Fiber optic	No	...	Yes	
7040	4801-JZAZL	Female	0	Yes	Yes	11	No	No phone service	DSL	Yes	...	No	
7041	8361-LTMKD	Male	1	Yes	No	4	Yes	Yes	Fiber optic	No	...	No	
7042	3186-AJIEK	Male	0	No	No	66	Yes	No	Fiber optic	Yes	...	Yes	

7043 rows × 21 columns

FIGURE 1 – Partie de l'ensemble de donnée

2.3 Analyse Exploratoire des Données (EDA)

L'EDA a été réalisée pour explorer la structure des données, effectuer un prétraitement initial et identifier des modèles ou des incohérences. Voici quelques points saillants de cette étape :

2.3.1 Nettoyage de Données

(a) **Conversion de 'TotalCharges' en Numérique :** Nous avons converti la variable 'TotalCharges' en format numérique pour une manipulation plus aisée.

(b) **Gestion des Valeurs Manquantes :** Nous avons identifié 11 valeurs manquantes dans la colonne 'TotalCharges'. Ces valeurs ont été remplacées par la moyenne.

(c) **Suppression des Doublons :** Nous avons vérifié et supprimé les éventuels doublons dans l'ensemble de données.

2.3.2 Analyse des Données :

(a) **Distribution de 'Churn'**

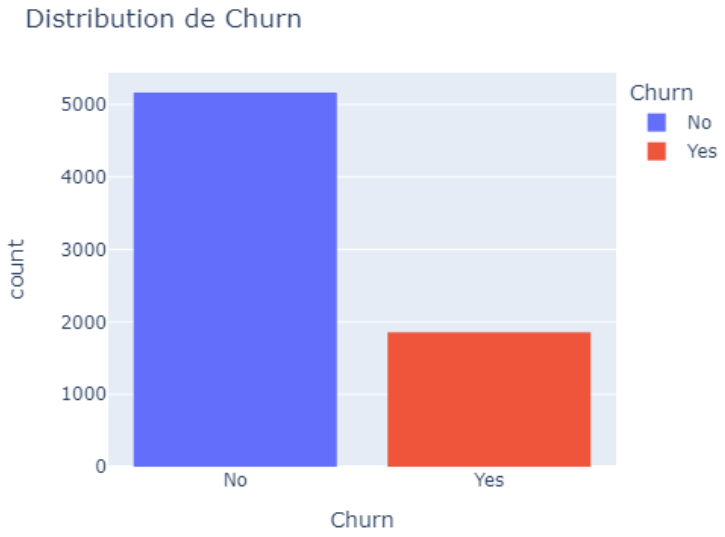

FIGURE 2 – Distribution de 'Churn'

L'ensemble de données est déséquilibré avec une proportion plus élevée de clients qui ne se sont pas désabonnés.

(b) **Distribution de ‘Churn’ par rapport à ‘Gender’**

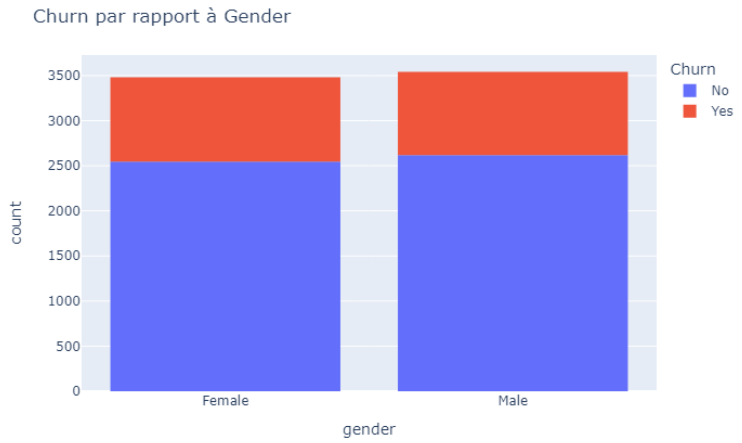

FIGURE 3 – ‘Churn’ par rapport à ‘Gender’

Il semble y avoir une légère différence dans le taux de désabonnement entre les hommes et les femmes, avec un peu plus de femmes qui se sont désabonnées que d’hommes.

(c) **la distribution de la ’Tenure’ par rapport au ’Churn’**

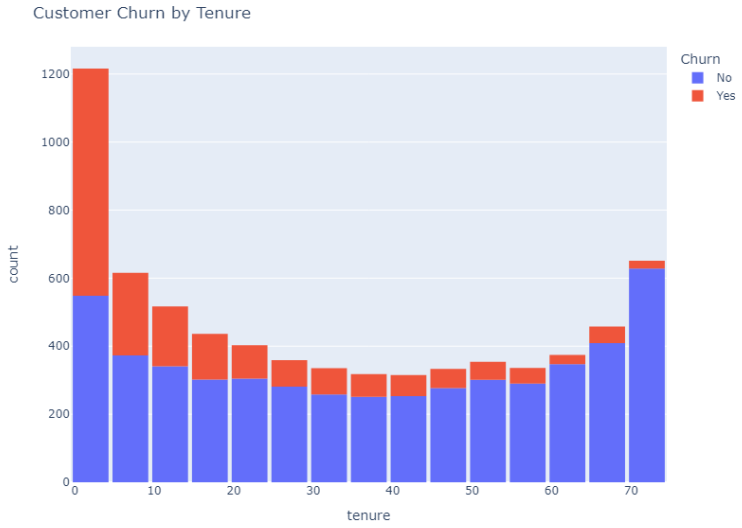

FIGURE 4 – ’Tenure’ par rapport au ’Churn’

Il semble y avoir un nombre significatif de clients qui se désabonnent à court terme (tenure proche de 0). Cela pourrait indiquer que de nombreux clients ne sont pas satisfaits au début de leur abonnement.

À mesure que la durée de l’abonnement augmente (tenure), le nombre de clients qui se désabonnent diminue. Cela pourrait indiquer que les clients qui restent abonnés plus longtemps sont généralement plus satisfaits du service.

(d) **La distribution de 'Churn' pour chaque type de contrat**

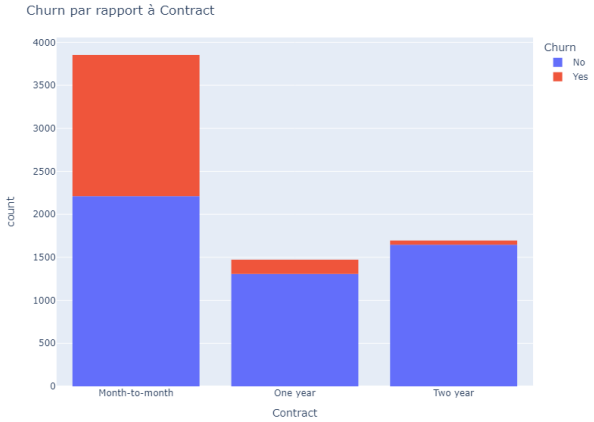

FIGURE 5 – 'Churn' par rapport au 'Contrat'

— **Contrats mensuels :** Les clients avec des contrats mensuels semblent avoir le taux de désabonnement le plus élevé en raison de la flexibilité de ces contrats, permettant aux clients de se désabonner plus facilement.

— **Contrats annuels :** Les clients avec des contrats d'un an et de deux ans ont un taux de désabonnement beaucoup plus faible. Cela pourrait indiquer une plus grande satisfaction du service ou un engagement plus élevé en raison de la durée plus longue de leur contrat.

— **Stratégies de rétention :** Pour réduire le taux de désabonnement, il pourrait être bénéfique d'encourager les clients à opter pour des contrats plus longs. Cela pourrait être réalisé en offrant des incitatifs ou des réductions pour les contrats d'un an ou de deux ans.

(e) **La distribution de 'Churn' pour chaque type de service d'internet**

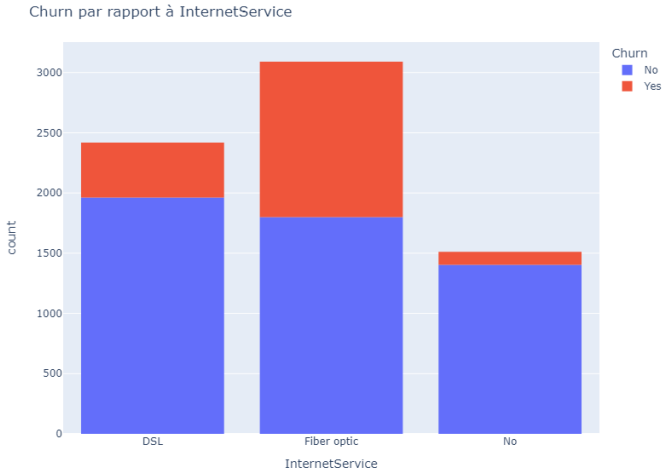

FIGURE 6 – 'Churn' par rapport au type de service d'internet

— **DSL :** La majorité des clients avec un service Internet DSL ne se sont pas désabonnés. Cela pourrait indiquer que les clients sont généralement satisfaits de ce service.

— **Fibre optique :** Il y a presque autant de clients qui se sont désabonnés que de clients qui sont restés abonnés. Cela pourrait indiquer un problème avec le service de fibre optique, comme un coût élevé ou une qualité de service insatisfaisante.

— **Pas de service Internet :** La majorité des clients sans service Internet ne se sont pas désabonnés. Cela pourrait indiquer que ces clients sont satisfaits des autres services qu'ils reçoivent.

(f) **La distribution de 'Churn' pour chaque méthode de paiement**

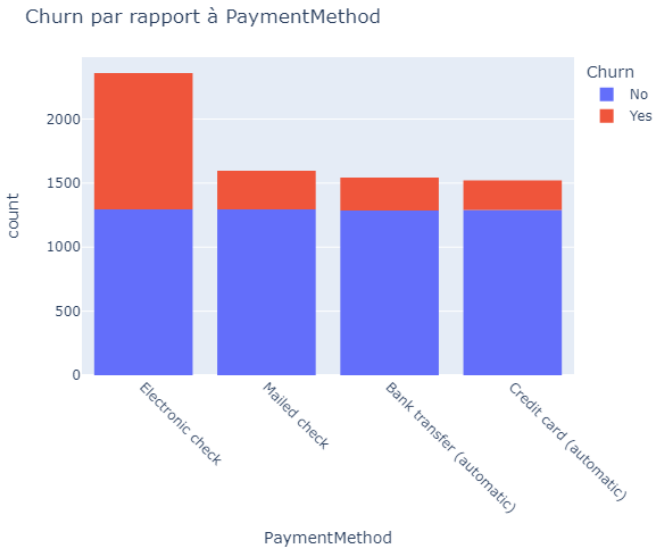

FIGURE 7 – 'Churn' par rapport au méthode de paiement

— **Chèque électronique :** Les clients qui utilisent le chèque électronique comme méthode de paiement semblent avoir le taux de désabonnement le plus élevé. Cela pourrait indiquer que ces clients sont moins satisfaits du service ou qu'ils trouvent cette méthode de paiement moins pratique.

— **Virement bancaire et carte de crédit :** Les clients qui utilisent le virement bancaire automatique ou la carte de crédit automatique comme méthode de paiement ont un taux de désabonnement plus faible. Cela pourrait indiquer que ces méthodes de paiement sont plus pratiques pour les clients ou qu'ils sont généralement plus satisfaits du service.

— **Stratégies de rétention :** Ces observations suggèrent que l'une des stratégies pour réduire le taux de désabonnement pourrait être d'encourager les clients à utiliser des méthodes de paiement automatiques, comme le virement bancaire ou la carte de crédit. Cela pourrait se faire en offrant des incitatifs ou des réductions pour ces méthodes de paiement.

(g) **Heatmap de corrélation**

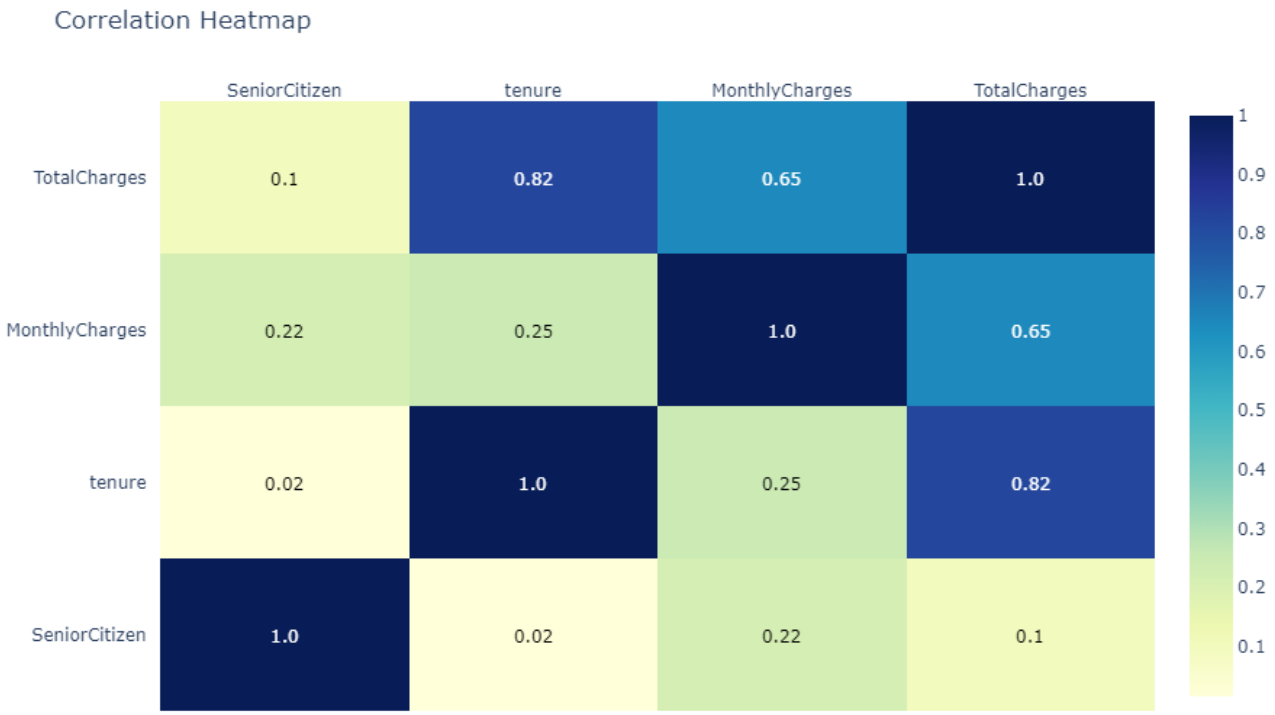

FIGURE 8 – Heatmap de corrélation

— **Corrélation entre tenure et MonthlyCharges :** La corrélation entre la durée de l'abonnement (tenure) et les frais mensuels (MonthlyCharges) est de 0,25. Cela indique une corrélation positive faible, ce qui signifie que lorsque la durée de l'abonnement augmente, les frais mensuels ont tendance à augmenter légèrement.

— **Corrélation entre tenure et TotalCharges :** Il y a une forte corrélation positive (0,82) entre la durée de l'abonnement (tenure) et les frais totaux (TotalCharges). Cela est logique car plus un client reste longtemps, plus il est susceptible d'avoir payé des frais totaux plus élevés.

— **Corrélation entre MonthlyCharges et TotalCharges :** Il y a également une corrélation positive modérée (0,65) entre les frais mensuels (MonthlyCharges) et les frais totaux (TotalCharges). Cela suggère que les clients qui paient des frais mensuels plus élevés ont tendance à avoir des frais totaux plus élevés.

— **Les autres corrélations sont faibles,**

2.4 Prétraitement des Données

Le prétraitement des données est une étape cruciale pour préparer les données à être utilisées dans les modèles d'apprentissage automatique. Les étapes spécifiques du prétraitement que nous avons suivies sont les suivantes :

(a) **Transformation des Valeurs 'Yes' et 'No' :**
Certaines colonnes, comme 'Churn', utilisaient des valeurs 'Yes' et 'No'. Pour permettre l'utilisation de ces variables dans les modèles, nous avons effectué une transformation en remplaçant 'Yes' par 1 et 'No' par 0. Cela assure que les modèles puissent interpréter ces variables comme des variables binaires.

(b) **Encodage des Variables Catégorielles :**
Les modèles d'apprentissage automatique requièrent souvent que les variables catégorielles soient converties en un format numérique. Nous avons utilisé la technique d'encodage one-hot **(One-Hot Encoding)**, qui crée des variables binaires distinctes pour chaque catégorie d'une variable catégorielle. Par exemple, la variable 'Contract' avec les catégories 'Month-to-month', 'One year', et 'Two year' a été transformée en trois variables binaires distinctes. Cela garantit que toutes les informations de la variable catégorielle sont prises en compte sans introduire d'ordre artificiel.

(c) **Normalisation des Données :**
La normalisation est une étape importante pour mettre toutes les variables à la même échelle, ce qui facilite la convergence des modèles d'apprentissage automatique. Nous avons appliqué la normalisation des variables numériques, en particulier 'tenure', 'MonthlyCharges', et 'TotalCharges'. La normalisation standardise les données, les ramenant à une échelle commune et facilitant ainsi l'entraînement des modèles.

(d) **Gestion du Déséquilibre de la Classe 'Churn' :**
Le jeu de données présentait un déséquilibre important entre les classes 'Churn' et 'Not Churn', avec une majorité de clients ne se désabonnant pas. Pour remédier à cela, nous avons utilisé des techniques de suréchantillonnage (over-sampling) de la classe minoritaire ('Churn') en utilisant la méthode **SMOTE (Synthetic Minority Over-sampling Technique)**. Cela crée des exemples synthétiques de la classe minoritaire, équilibrant ainsi les classes et améliorant les performances des modèles pour détecter les cas de désabonnement.

Ces techniques de prétraitement garantissent que les données sont prêtes à être utilisées efficacement par les modèles d'apprentissage automatique, en prenant en compte les spécificités des différentes variables du jeu de données.

2.5 Entraînement et Évaluation des Modèles

La dernière étape de la méthodologie consiste à développer et à optimiser un modèle de réseau de neurones pour prédire le désabonnement des clients. Nous avons utilisé un réseau de neurones artificiels en raison de sa capacité à modéliser des relations complexes entre les caractéristiques des clients. Voici une explication détaillée de cette étape :

1. **Division des Données :** Avant d'entraîner le modèle, j'ai divisé le jeu de données en ensembles d'entraînement et de test pour évaluer la performance du modèle.

2. **Optimisation par GridSearchCV :** Pour maximiser les performances du modèle, j'ai utilisé la technique de GridSearchCV (validation croisée avec recherche sur grille) pour explorer différents hyperparamètres. Les hyperparamètres tels que le nombre de neurones, l'optimiseur et le nombre d'époques ont été ajustés pour trouver la combinaison optimale. Cette approche garantit que le modèle est ajusté de manière optimale aux données d'entraînement tout en généralisant bien aux données de test.

3. **Création du Modèle :** Le modèle de réseau de neurones a été construit en utilisant la bibliothèque TensorFlow en conjonction avec Keras. J'ai choisi une architecture de réseau de neurones simple, composée de couches d'entrée, de couches cachées et d'une couche de sortie. La fonction d'activation ReLU a été utilisée pour les couches cachées, et la fonction d'activation sigmoïde pour la couche de sortie, adaptée à un problème de classification binaire.

4. **Évaluation des Performances :** Une fois le modèle entraîné, j'ai évalué sur l'ensemble de test en utilisant des métriques de performance telles que la précision, le rappel, la F1-score et l'aire sous la courbe ROC (AUC-ROC). Ces métriques me permettent de mesurer l'efficacité du modèle à prédire avec précision les cas de désabonnement.

3 Résultats

Après avoir optimisé mon modèle de réseau de neurones, j'ai obtenu les meilleurs paramètres suivants :

— Taille du lot : 40
— Époques : 100
— Neurones : 5
— Optimiseur : 'adam'

Avec ces paramètres, j'ai atteint une précision de 0.80 lors de l'évaluation du modèle sur l'ensemble de test.

1. Le rapport de classification est le suivant :

	precision	recall	f1-score	support
0	0.85	0.74	0.79	1034
1	0.77	0.87	0.81	1032
accuracy			0.80	2066
macro avg	0.81	0.80	0.80	2066
weighted avg	0.81	0.80	0.80	2066

FIGURE 9 – Le rapport de classification

Le rapport de classification fournit plusieurs mesures statistiques pour évaluer la performance d'un modèle de classification. Voici ce que signifient ces mesures et comment les calculer :

— **Précision :** La précision est le rapport entre les vrais positifs (VP) et la somme des vrais positifs et des faux positifs (FP). Elle est calculée comme suit :

$$Prcision = \frac{VP}{VP + FP} \tag{1}$$

Une précision élevée indique que le modèle a correctement prédit la majorité des points positifs.

— **Rappel :** Le rappel est le rapport entre les vrais positifs et la somme des vrais positifs et des faux négatifs (FN). Il est calculé comme suit :

$$Rappel = \frac{VP}{VP + FN} \tag{2}$$

Un rappel élevé indique que le modèle a correctement identifié la majorité des points positifs réels.

— **Score F1 :** Le score F1 est la moyenne harmonique de la précision et du rappel. Il est calculé comme suit :

$$F1 = 2 \times \frac{Prcision \times Rappel}{Prcision + Rappel} \tag{3}$$

Un score F1 élevé indique que le modèle a une précision et un rappel élevés.

— **Support :** Le support est le nombre total d'occurrences de la classe dans l'ensemble de données. Il donne une idée de la distribution des classes.

Dans mon cas, pour la classe 0, la précision est de 0.85, le rappel est de 0.74, le score F1 est de 0.79 et le support est de 1034. Pour la classe 1, la précision est de 0.77, le rappel est de 0.87, le score F1 est de 0.81 et le support est de 1032. Cela signifie que mon modèle a une bonne performance globale, avec une légère tendance à mieux prédire la classe 1.

2. La matrice de confusion est la suivante :

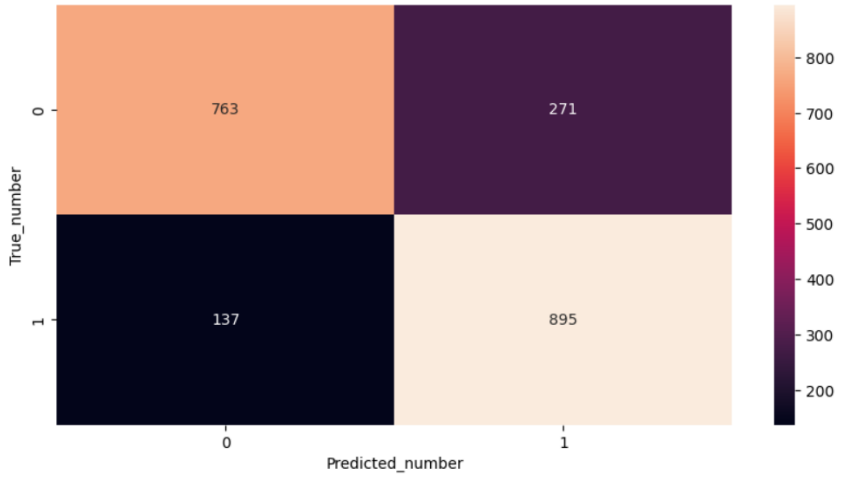

FIGURE 10 – La matrice de confusion

— **Vrais positifs (VP) :** Ce sont les cas où le modèle a correctement prédit la classe positive. Dans mon cas, il y a 895 vrais positifs, ce qui signifie que le modèle a correctement prédit 895 cas de désabonnement.

— **Vrais négatifs (VN) :** Ce sont les cas où le modèle a correctement prédit la classe négative. Dans mon cas, il y a 763 vrais négatifs, ce qui signifie que le modèle a correctement prédit 763 cas d'abonnement.

— **Faux positifs (FP) :** Ce sont les cas où le modèle a incorrectement prédit la classe positive. Dans mon cas, il y a 271 faux positifs, ce qui signifie que le modèle a prédit à tort 271 cas d'abonnement comme des désabonnements.

— **Faux négatifs (FN) :** Ce sont les cas où le modèle a incorrectement prédit la classe négative. Dans mon cas, il y a 137 faux négatifs, ce qui signifie que le modèle a prédit à tort 137 cas de désabonnement comme des abonnements.

3. **La courbe ROC**

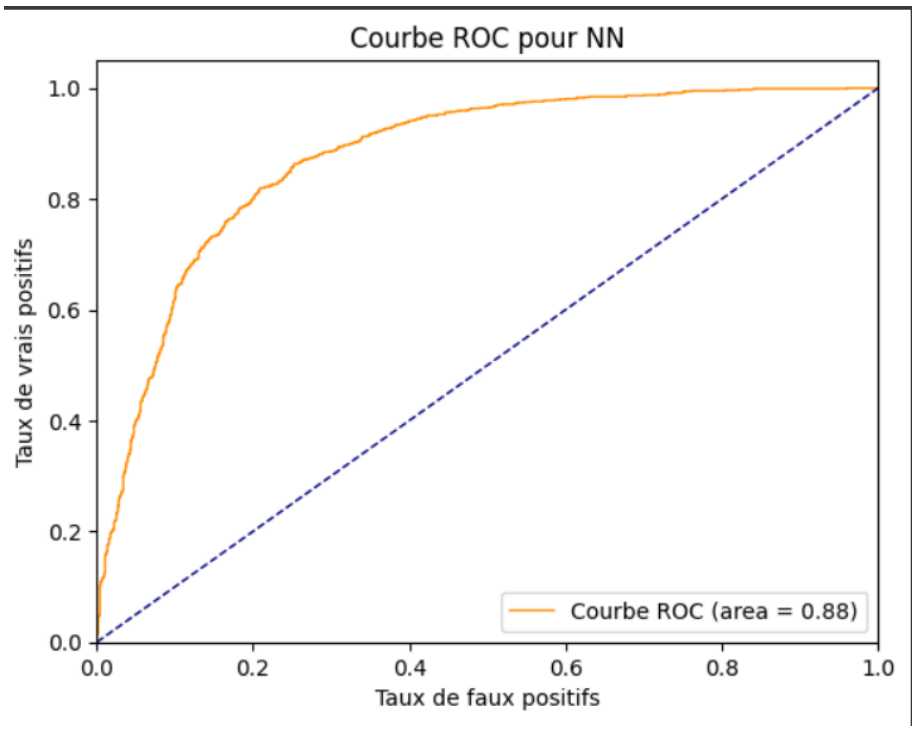

FIGURE 11 – La courbe ROC

La courbe ROC de mon modèle a une AUC de 0.88. Cela signifie que mon modèle a une bonne capacité à distinguer entre les abonnements et les désabonnements. Plus l'AUC est proche de 1, meilleur est le modèle. Donc, un AUC de 0.88 indique que mon modèle est assez performant.

.

4 Discussion

4.1 Points forts et limites

Mon approche a plusieurs points forts. Tout d'abord, j'ai utilisé un modèle de réseau de neurones, qui est capable de capturer des relations complexes dans les données. De plus, j'ai optimisé les paramètres du modèle pour obtenir une précision de 0.80, ce qui est assez élevé.

Cependant, mon approche a aussi ses limites. Par exemple, le modèle pourrait être amélioré en utilisant plus de données ou en intégrant plus de caractéristiques pertinentes. De plus, bien que la précision soit élevée, le modèle a tendance à mieux prédire la classe 1, ce qui pourrait être un problème si la classe 0 est également importante.

4.2 Implications et conclusions

Les résultats de mon modèle ont plusieurs implications. Ils montrent que le modèle peut être utilisé pour prédire efficacement le désabonnement des clients, ce qui pourrait aider les entreprises à prendre des mesures proactives pour retenir leurs clients.

En conclusion, mon modèle de réseau de neurones a montré une bonne performance dans la prédiction du désabonnement des clients. Cependant, il y a encore place à l'amélioration.

4.3 Directions de recherche future et améliorations possibles

Pour les recherches futures, je suggère d'explorer d'autres types de modèles de machine learning, comme les forêts aléatoires(RF) ou les machines à vecteurs de support(SVM). De plus, il serait intéressant d'explorer l'impact de différentes caractéristiques sur le désabonnement des clients.

En termes d'améliorations possibles, je pourrais essayer d'obtenir plus de données ou d'intégrer plus de caractéristiques pertinentes dans le modèle. De plus, je pourrais essayer d'améliorer la performance du modèle sur la classe 0.

5 Conclusion

En conclusion, je suis satisfait des résultats obtenus dans ce projet de prédiction du désabonnement des clients en utilisant un modèle de réseau de neurones. En optant pour cette approche, j'ai pu capturer des motifs complexes dans les données, atteignant une précision de 0.80 lors de l'évaluation du modèle sur l'ensemble de test.

Mon modèle a démontré une capacité significative à prédire avec précision les cas de désabonnement, comme le montre le rapport de classification et la matrice de confusion. Cependant, il existe toujours des opportunités d'amélioration, notamment en explorant d'autres modèles de machine learning et en étudiant davantage l'impact des différentes caractéristiques.

Les implications de ce travail sont importantes, offrant aux entreprises la possibilité de prendre des mesures proactives pour retenir leurs clients. Pour l'avenir, je suggère d'explorer d'autres types de modèles et d'enrichir le jeu de données pour une meilleure généralisation.

Ce projet m'a permis de développer mes compétences en machine learning et de mieux comprendre l'application pratique des réseaux de neurones dans le domaine de la prédiction de comportement client. C'est une étape prometteuse, mais il reste toujours des défis à relever pour perfectionner davantage les modèles.

6 Les references

1. **Customer Churn Prediction Using Deep Learning**
   Lien :
   https://www.researchgate.net/publication/350906568_Customer_Churn_Prediction_Using_Deep_Learning

2. **Customer Churn Prediction Using Artificial Neural Network**
   Lien :
   https://www.analyticsvidhya.com/blog/2021/10/customer-churn-prediction-using-arti

3. **Muhammad Muzzamil : Customer Churn Prediction Using Artificial Neural Networks (ANN)**
   Lien :
   https://medium.com/@muzammilmuhammad196/customer-churn-prediction-using-artificia
   EF%B8%8F-%EF%B8%8F-74ee0ea33b59

4. **Bank Customer Churn Prediction Using Machine Learning | Machine Learning Projects**
   Lien :
   https://www.youtube.com/watch?v=VpMGXfhDQXc

5. **Deep Learning Customer Churn Predication for Beginners**
   Lien :
   https://www.youtube.com/watch?v=iRWI5b0aYRI&t=29s

6. **Neural Network + GridSearchCV Explanations**
   Lien :
   https://www.kaggle.com/code/aaryandhore/neural-network-gridsearchcv-explanations

7. **Mémoire de Fin d'Études : Prévision du désabonnement de clients dans le secteur de télécommunication**
   Lien :
   https://dspace.univ-bba.dz/bitstream/handle/123456789/2217/memoir-m2.pdf?sequence=1&isAllowed=y

8. **Étude de cas d'apprentissage automatique : prédiction du taux de désabonnement des clients des opérateurs de télécommunications**
   Lien :
   https://ichi.pro/fr/etude-de-cas-d-apprentissage-automatique-prediction-du-taux-