REGRESSION LOGISTIQUE

# Introduction

# Qu’est-ce que la régression logistique ?

**La régression logistique** est une technique d’analyse de données qui utilise les mathématiques pour trouver les relations entre deux facteurs de données. Elle utilise ensuite cette relation pour prédire la valeur de l’un de ces facteurs en fonction de l’autre.

Technique utilisée pour prédire une variable binaire ou trouver une relation mathématique entre une variable binaire et plusieurs autres variables.

La prédiction a généralement un nombre fini de résultats, comme Oui ou Non.

**Exemple** :

Par exemple, supposons que vous souhaitez savoir si le visiteur de votre site web cliquera ou non sur le bouton de paiement dans son panier d’achat. La régression logistique examine le comportement des visiteurs précédents, tels que le temps passé sur le site web et le nombre d’articles dans le panier.

En s’appuyant sur des données du passé, elle détermine que si les visiteurs ont passé plus de 5 minutes sur le site et ajouté plus de 3 articles au panier, ils ont cliqué sur le bouton paiement.

A l’aide de ces informations, la fonction de régression logistique peut alors prédire le comportement d’un nouveau visiteur du site web.

# Pourquoi la régression logistique est-elle importante ?

C’est une technique importante dans le domaine de l’IA et du Machine Learning.

**Les modèles de Machine Learning** sont des logiciels que vous pouvez entrainer pour effectuer des tâches complexes de traitement de données sans intervention humaine.

 Les modèles de Machine Learning créés à l'aide de la régression logistique aident les entreprises à obtenir des informations exploitables à partir de leurs données commerciales.

À l'aide de ces informations, elles peuvent réaliser une analyse prédictive afin de réduire les coûts opérationnels, d'accroître l'efficacité et d'accélérer la mise à l'échelle.

Par exemple, les entreprises peuvent découvrir des modèles qui améliorent la rétention des employés ou conduisent à une conception de produits plus rentable.

# Les Avantages et Inconvénients de la Régression Logistique par rapport à d’autres technique de Machine Learning ?

## Les avantages

* **Simplicité**
  + Les modèles de régression logistique sont mathématiquement simples, avec des bases algébriques claires.
  + Leur implémentation ne nécessite pas de capacités de calcul importantes, ce qui les rend accessibles même sans expertise approfondie en Machine Learning.
* **Rapidité**
  + En raison de leur faible complexité algorithmique, les modèles s'entraînent rapidement, surtout avec des ensembles de données de petite ou moyenne taille.
* **Flexibilité**
  + La régression logistique peut être adaptée à une variété de cas pratiques : classification binaire (oui/non), classification multiclasse (avec extensions comme Softmax), etc.
  + Elle accepte des données de types différents, bien qu'un prétraitement soit nécessaire.
* **Visibilité**
  + Les coefficients de régression sont facilement interprétables et permettent de comprendre l'impact de chaque variable explicative.
  + C'est un modèle transparent, contrairement à des techniques comme les réseaux neuronaux, souvent qualifiés de "boîtes noires".
* **Stabilité**
  + Fonctionne bien avec des données où les relations entre variables explicatives et la variable cible sont linéaires.
  + Moins sensible au bruit dans les données par rapport aux techniques non paramétriques.
* **Performance**
  + Idéal pour les problèmes où les variables explicatives sont linéairement séparables.
  + Efficace sur des ensembles de données simples, équilibrés, et bien préparés.

## Les inconvénients

* **Hypothèse de Linéarité**
  + La régression logistique suppose que les variables explicatives ont une relation linéaire avec la variable cible, ce qui peut limiter son efficacité dans des scénarios plus complexes.
* **Sensible aux Outliers**
  + Les valeurs aberrantes peuvent affecter de manière significative les résultats, surtout si elles ne sont pas traitées au préalable.
* **Non adaptée à des problèmes non linéaires**
  + Pour des données où les relations sont hautement non linéaires ou trop complexes, des modèles comme les forêts aléatoires ou les réseaux neuronaux sont bien plus efficaces.
* **Manque de puissance prédictive pour les données multiclasses**
  + Bien que des extensions comme Softmax existent, d'autres algorithmes comme les SVM ou les arbres de décision sont souvent plus performants dans les problèmes multiclasses complexes.
* **Nécessité de Prétraitement**
  + La régression logistique nécessite que les données soient normalisées ou transformées si elles incluent des valeurs de magnitudes très différentes ou des variables catégoriques.
* **Moins performante avec des ensembles de données très grands ou déséquilibrés**
  + Avec des ensembles de données massifs ou des classes déséquilibrées, des algorithmes comme XGBoost ou LightGBM offrent souvent de meilleures performances.

# Quelles sont les applications de la régression logistique ?

La régression logistique a plusieurs applications concrètes dans de nombreux secteurs d'activité différents.

* **Fabrication** :

Les entreprises manufacturières utilisent une analyse de régression logistique pour estimer la probabilité de défaillance des pièces des machines. Elles planifient ensuite les calendriers de maintenance en fonction de cette estimation afin de minimiser les défaillances futures.

* **Soins médicaux** :

Les chercheurs en médecine planifient les soins et les traitements préventifs en prédisant la probabilité de maladie chez les patients. Ils utilisent des modèles de régression logistique pour comparer l'impact des antécédents familiaux ou des gènes sur les maladies.

* **Finance** :

Les sociétés financières doivent analyser les transactions financières à la recherche de fraudes et évaluer les risques liés aux demandes de prêt et d'assurance. Ces problèmes sont adaptés à un modèle de régression logistique, car ils ont des résultats distincts. Par exemple, une transaction financière est frauduleuse ou non, et une demande de prêt présente un risque faible ou élevé.

* **Marketing** :

Les outils de publicité en ligne utilisent le modèle de régression logistique pour prédire si les utilisateurs cliqueront sur une publicité. En conséquence, les spécialistes du marketing peuvent analyser les réponses des utilisateurs à différents mots et images et créer des publicités très performantes avec lesquelles les clients interagiront.

# Comment fonctionne l’analyse de régression ?

Pour comprendre la régression logistique, nous devons d’abord comprendre l’analyse de régression de base.

Ci-dessous, nous montrons comment fonctionne l’analyse de régression au moyen d’un exemple d’analyse de régression linéaire.

## Identifier la question

Toute analyse de données commence par une question commerciale. Pour la régression logistique, vous devez formuler la question pour obtenir des résultats spécifiques :

* Les jours de pluies ont-ils un impact sur nos ventes mensuelles ? (Oui ou Non)
* Quel type d’activité de carte de crédit le client effectue-t-il ? (Activité autorisée, Frauduleuse ou Potentiellement Frauduleuse)

## Collecter des données historiques

Après avoir identifié la question, vous devez identifier **les facteurs** de données impliqués. Vous collecterez ensuite les données passées pour tous les facteurs.

Par exemple, pour répondre à la première question ci-dessus, vous pouvez collecter le **nombre de jours de pluie** et vos données de **ventes mensuelles** pour **les trois dernières années**.

## Entraîner le modèle d'analyse de régression

Vous allez traiter les données historiques à l'aide d'un logiciel de régression. Le logiciel traitera les différents points de données et les reliera mathématiquement à l'aide d'équations.

Par exemple, si le nombre de jours de pluie pendant trois mois est de 3, 5 et 8 et que le nombre de ventes pendant ces mois est de 8, 12 et 18, l'algorithme de régression reliera les facteurs à l'équation :

**Nombre de ventes = 2\*(nombre de jours de pluie) + 2**

## Faire des prévisions pour des valeurs inconnues

Pour les valeurs inconnues, le logiciel utilise l'équation pour effectuer une prédiction.

Si vous savez qu'il pleuvra pendant six jours en juillet, le logiciel estimera la valeur de vente de juillet à 14.

# Comment fonctionne le modèle de régression logistique ?

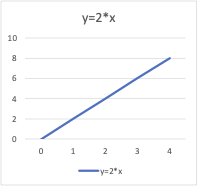
Pour comprendre le modèle de régression logistique, commençons par comprendre les équations et les variables.

## Équations

En mathématiques, les équations donnent la relation entre deux variables : x et y.

Vous pouvez utiliser ces équations, ou fonctions, pour tracer un graphique le long de l'axe des x et de l'axe des y en saisissant différentes valeurs de x et y.

Par exemple, si vous tracez le graphique pour la fonction y = 2\*x, vous obtiendrez une ligne droite comme indiqué ci-dessous. Par conséquent, cette fonction est également appelée fonction linéaire.



## Variables

En statistique, les variables sont les facteurs ou attributs de données dont les valeurs varient. Pour toute analyse, certaines variables sont des variables indépendantes ou explicatives. Ces attributs sont à l'origine d'un résultat.

Les autres variables sont des variables dépendantes ou des variables de réponse, c'est-à-dire que leurs valeurs dépendent des variables indépendantes.

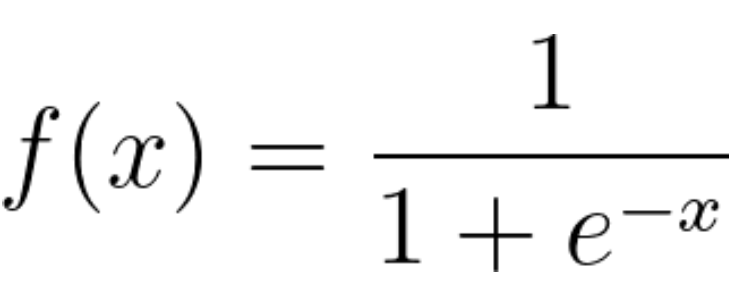
En général, la régression logistique explore la façon dont les variables indépendantes affectent une variable dépendante en examinant les valeurs de données historiques des deux variables.

Dans notre exemple ci-dessus, *x* est appelé variable indépendante, variable de prédiction ou variable explicative, car sa valeur est connue. *Y* est appelée variable dépendante, variable de résultat ou variable de réponse, car sa valeur est inconnue.

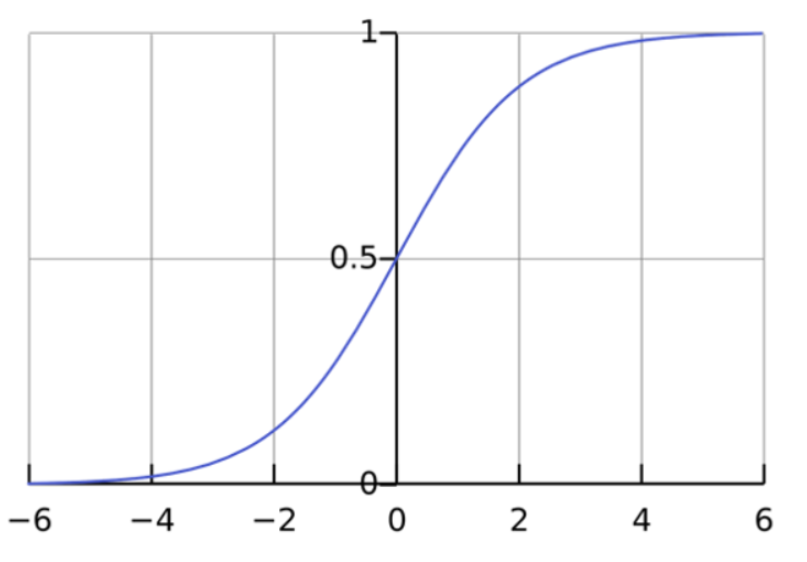
## Fonction de régression logistique

La régression logistique est un modèle statistique qui utilise la **fonction logistique**, ou **fonction logit**, en mathématiques comme l'équation entre *x* et *y*.

La fonction logit mappe *y* en tant que fonction sigmoïde de *x*.



Si vous tracez cette équation de régression logistique, vous obtenez une courbe en S comme illustré ci-dessous.



Comme vous pouvez le constater, la fonction logit renvoie uniquement des valeurs comprises entre 0 et 1 pour la variable dépendante, quelles que soient les valeurs de la variable indépendante.

C'est ainsi que la régression logistique estime la valeur de la variable dépendante. Les méthodes de régression logistique modélisent également des équations entre plusieurs variables indépendantes et une variable dépendante.

## Analyse de régression logistique avec plusieurs variables indépendantes

Dans de nombreux cas, plusieurs variables explicatives affectent la valeur de la variable dépendante. Pour modéliser de tels jeux de données d'entrée, les formules de régression logistique supposent une relation linéaire entre les différentes variables indépendantes.

Vous pouvez modifier la fonction sigmoïde et calculer la variable de sortie finale comme suit :

***y* = *f*(β0 + β1*x*1 + β2*x*2+… βn*x*n)**

Le symbole β représente le coefficient de régression. Le modèle logit peut effectuer un calcul inverse de ces valeurs de coefficient lorsque vous lui donnez un jeu de données expérimental suffisamment grand avec des valeurs connues de variables dépendantes et indépendantes.

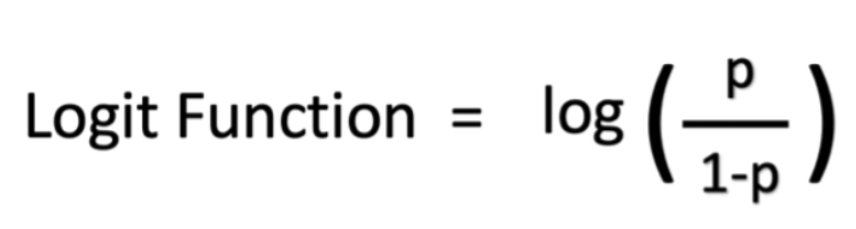
## Cotes logarithmiques

Le modèle logit peut également déterminer **le ratio succès/échec**, également appelé **cotes logarithmiques**.

Par exemple, si vous jouez au poker avec vos amis et que vous gagnez quatre parties sur 10, vos chances de gagner sont de quatre sixièmes, ou quatre sur six, ce qui correspond à votre ratio réussite/échec. La probabilité de gagner, en revanche, est de quatre sur 10.

Mathématiquement, vos cotes en termes de probabilité sont ***p*/(1 - *p*)**, et vos cotes logarithmiques sont **log (*p*/(1 - *p*))**.

Vous pouvez représenter la fonction logistique sous forme de cotes logarithmiques comme indiqué ci-dessous :



# Quels sont les types d'analyse de régression logistique ?

Il existe trois approches d'analyse de régression logistique basées sur les résultats de la variable dépendante.

## Régression logistique binaire

La régression logistique binaire fonctionne bien pour les problèmes de classification binaire qui n'ont que deux résultats possibles. La variable dépendante ne peut avoir que deux valeurs, par exemple oui ou non ou 0 ou 1.

Même si la fonction logistique calcule une plage de valeurs comprise entre 0 et 1, le modèle de régression binaire arrondit la réponse aux valeurs les plus proches.

En général, les réponses inférieures à 0,5 sont arrondies à 0, et les réponses supérieures à 0,5 sont arrondies à 1, de sorte que la fonction logistique renvoie un résultat binaire.

## Régression logistique multinomiale

La régression multinomiale peut analyser des problèmes qui ont plusieurs issues possibles tant que le nombre de résultats est limité. Par exemple, elle peut prédire si les prix des maisons augmenteront de 25 %, 50 %, 75 % ou 100 % sur la base des données démographiques, mais elle ne peut pas prédire la valeur exacte d'une maison.

La régression logistique multinomiale fonctionne en mappant les valeurs de résultat à différentes valeurs comprises entre 0 et 1. Comme la fonction logistique peut renvoyer une plage de données continues, telles que 0,1, 0,11, 0,12, etc., la régression multinomiale regroupe également la sortie aux valeurs les plus proches possibles.

## Régression logistique ordinale

La régression logistique ordinale, ou modèle logit ordonné, est un type spécial de régression multinomiale pour les problèmes dans lesquels les nombres représentent des rangs plutôt que des valeurs réelles. Par exemple, vous utiliseriez la régression ordinale pour prédire la réponse à une question de sondage qui demande aux clients de classer votre service comme mauvais, passable, bon ou excellent sur la base d'une valeur numérique, telle que le nombre d'articles qu'ils vous achètent au cours de l'année.

# Comment la régression logistique se compare-t-elle aux autres techniques de machine learning ?

Les deux techniques d'analyse de données les plus courantes sont l'analyse de régression linéaire et le deep learning.

## Analyse de régression linéaire

Comme expliqué ci-dessus, la régression linéaire modélise la relation entre les variables dépendantes et indépendantes en utilisant une combinaison linéaire. L'équation de régression linéaire est

***y*= β0*X*0 + β1*X*1 + β2*X*2+… βn*X*n+ ε, où β1 à βn et ε sont des coefficients de régression.**

## Comparaison entre régression logistique et régression linéaire

**La régression linéaire** prédit une variable dépendante continue en utilisant un ensemble donné de variables indépendantes. Une variable continue peut avoir une plage de valeurs, telles que le prix ou l'âge. La régression linéaire peut donc prédire les valeurs réelles de la variable dépendante. Elle peut répondre à des questions telles que « Quel sera le prix du riz après 10 ans ? ».

Contrairement à la régression linéaire, la régression logistique est un algorithme de classification. Elle ne peut pas prédire de valeurs réelles pour des données continues. Elle peut répondre à des questions telles que « Le prix du riz augmentera-t-il de 50 % en 10 ans ? ».

## Deep learning

Le [deep learning](https://aws.amazon.com/what-is/deep-learning/) utilise des réseaux neuronaux ou des composants logiciels qui simulent le cerveau humain pour analyser l’information. Les calculs de deep learning sont basés sur le concept mathématique des vecteurs.

Comparaison entre régression logistique et deep learning

La régression logistique est moins complexe et nécessite moins de calculs que le deep learning. Plus important encore, les calculs de deep learning ne peuvent pas être étudiés ni modifiés par les développeurs, car ils sont complexes et effectués par des machines. D'autre part, les calculs de régression logistique sont transparents et plus faciles à résoudre.

# Comparaison :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Critère | Régression Logistique | Forêts Aléatoires (Random Forest) | Réseaux de Neurones (ANN) | Support Vector Machines (SVM) |
| Coût Financier | Faible, nécessite peu de ressources. | Modéré, coût de calcul plus élevé. | Élevé, nécessite beaucoup de ressources matérielles. | Modéré à élevé, selon le noyau et les données. |
| Coût Matériel | Faible, pas de matériel spécialisé nécessaire. | Plus élevé, nécessite plus de puissance de calcul, notamment pour les grands jeux de données. | Très élevé, nécessite des GPU pour les calculs intensifs. | Moyen à élevé, dépend de la taille des données et du noyau. |
| Coût Humain | Faible, facile à comprendre et à mettre en œuvre. | Modéré, nécessite un réglage des hyperparamètres. | Très élevé, nécessite une expertise approfondie en Deep Learning. | Moyen, demande une expertise dans le choix du noyau. |
| Temps d'Entraînement | Rapide, surtout pour les petits jeux de données. | Modéré, plus long pour entraîner plusieurs arbres. | Très long, surtout pour des réseaux profonds avec de grandes données. | Modéré, dépend de la taille des données et du noyau. |
| Domaine d'Application | Classification binaire, détection de fraude, marketing. | Prédiction des zones à risques, classification complexe. | Traitement d'images, NLP, reconnaissance vocale. | Problèmes complexes, classification non linéaire. |
| Précision | Modérée sur des données linéaires, bonne pour des tâches simples. | Haute précision sur des données complexes, robustesse face au surapprentissage. | Très haute précision, surtout dans les tâches complexes comme la vision par ordinateur et le NLP. | Haute précision sur des données non linéaires et complexes. |
| Exemples Concrets | Prédiction du comportement des utilisateurs en ligne (ex : clic sur bouton de paiement). | Prédiction des zones sensibles aux glissements de terrain en Iran (précision de 98,4%). | Prédiction des risques de maladies cardiaques à partir de données complexes. | Prédiction des zones à risques pour les glissements de terrain, précision de 98,4% sur des routes montagneuses en Iran. |

# Comment pouvez-vous exécuter une analyse de régression logistique sur AWS ?

Vous pouvez exécuter une régression logistique sur AWS en utilisant [Amazon SageMaker](https://aws.amazon.com/sagemaker/). SageMaker est un service de machine learning entièrement géré avec des algorithmes intégrés pour la régression linéaire et la régression logistique, parmi plusieurs packages de logiciels statistiques.

* Chaque spécialiste des données peut utiliser SageMaker pour préparer, créer, former et déployer rapidement des modèles de régression logistique.
* SageMaker facilite chaque étape du processus de régression logistique afin de rendre plus aisé le développement de modèles de haute qualité.
* SageMaker fournit tous les composants dont vous avez besoin pour la régression logistique dans un seul jeu d'outils. Vous pouvez ainsi mettre les modèles en production plus rapidement, plus facilement et à moindre coût.