Universitée Paris Nanterre

Aircraft Price Analysis & Prediction Dataset

Auteur : Sika YERIMA 41001058

Professeur : Mélanie Zetlaoui

20 Décembre 2025

Nanterre, France

Contents

Introduction

Le principe de la prédiction des prix par les techniques de **machine learning (ML)** repose sur l'exploitation de données historiques afin de modéliser la relation complexe entre un ensemble de facteurs explicatifs (les features) et la variable cible, ici le prix.

Cette tâche relève généralement de la **régression**, où un algorithme ou un réseau de neurones est entraîné sur des observations passées. L'algorithme apprend les schémas, interactions et dépendances non linéaires présents dans les données pour être capable de généraliser cette connaissance et d'estimer une valeur numérique future avec l'erreur la plus faible possible.

Un enjeu majeur de ce processus consiste à capter simultanément les tendances structurelles et les effets externes tout en évitant le **surapprentissage (overfitting)** sur l'échantillon d'entraînement.

Sur le plan théorique, **l'économiste Sherwin Rosen** formalise en 1979 les **modèles de prix hédoniques**, selon lesquels le prix d'un bien complexe peut s'exprimer comme une fonction de l'ensemble de ses attributs observables.

Autrement dit, qu'il s'agisse d'un logement, d'un véhicule ou d'un appareil aéronautique, la détermination du prix repose sur l'évaluation de ses caractéristiques intrinsèques et de leur contribution marginale à la valeur totale.

Dans ce projet, **l'objectif est précisément de prédire le prix d'un appareil aéronautique à partir de ses caractéristiques techniques**, en mobilisant des méthodes contemporaines d'apprentissage automatique. Les appareils disponibles présentent une grande diversité en termes de puissance moteur, vitesse, autonomie, distances de décollage et d'atterrissage, poids à vide ou encore type de motorisation.

La théorie explique que ces attributs influencent non seulement les performances et l'usage potentiel de l'appareil, mais également son coût de production et donc sa valeur marchande.

Dans un premier temps, avant l'entraînement des modèles, une analyse descriptive approfondie est menée afin d'examiner les distributions, d'identifier les valeurs extrêmes et d'évaluer les corrélations entre variables.

Dans un second temps, nous calibrons plusieurs modèles de régression inspirés du cadre hédonique : une régression linéaire classique, des modèles pénalisés de type Ridge et Lasso, une régression polynomiale pour capter d'éventuelles non-linéarités, ainsi qu'un réseau de neurones multicouches afin d'explorer une approche plus flexible.

Enfin, nous évaluerons la qualité des modèles à l'aide de métriques standard en régression, telles que l'erreur quadratique moyenne (MSE) et le coefficient de détermination (R^2), en mettant en regard performance prédictive et interprétabilité.

Ce projet illustre ainsi l'application concrète des techniques de machine learning à la prédiction de prix dans un contexte hédonique.

1 Données

Le jeu de données utilisé dans ce projet regroupe 517 appareils aéronautiques décrits selon un ensemble de caractéristiques techniques et physiques.

Parmi ces 517 observations, 10 appareils n'ont pas de prix. L'objectif final de ce projet sera de prédire les prix de ces 10 appareils.

Ces variables permettent de représenter les principales performances de l'appareil, ses dimensions structurelles, ainsi que certains paramètres liés à la motorisation.
L'objectif est d'expliquer et prédire le prix des appareils à partir de ces attributs.

Les données comprennent 14 variables numériques et 1 variable catégorielle et 1 variable d'identification des avions.

Le tableau ci-dessous résume l'ensemble des variables, leur signification ainsi que leur unité de mesure.

Table 1: Description des variables du jeu de données

Variable	**Signification**	**Unité**
model_name	Modèle commercial de l'appareil	—
engine_type	Type de motorisation	—
engine_power	Puissance du moteur	Horsepower (hp)
max_speed	Vitesse maximale	Knot (kt)
cruise_speed	Vitesse de croisière	Knot (kt)
stall_speed	Vitesse de décrochage	Knot (kt)
fuel_tank	Capacité du réservoir de carburant	Litres (L)
all_eng_roc	Taux de montée (tous moteurs)	ft/min
out_eng_roc	Taux de montée (moteur en panne)	ft/min
takeoff_distance	Distance de décollage	Mètres (m)
landing_distance	Distance d'atterrissage	Mètres (m)
empty_weight	Masse à vide	Kilogrammes (kg)
length	Longueur de l'appareil	Mètres (m)
wing_span	Envergure	Mètres (m)
range	Autonomie maximale	Kilomètres (km)
price	Prix de l'appareil	Dollars ($)

Les variables retenues couvrent des dimensions essentielles des performances aéronautiques :
les vitesses permettant de caractériser les capacités de vol, les distances opérationnelles liées aux contraintes d'exploitation, les paramètres de motorisation reflétant la puissance et la consommation potentielle, les dimensions qui conditionnent la structure de l'appareil, ainsi que la masse et l'autonomie.

Cet ensemble constitue une base cohérente pour modéliser le prix selon une approche hédonique, où chaque caractéristique contribue à la valeur finale de l'appareil.

2 Analyse descriptive des données

L'analyse descriptive constitue une étape essentielle avant toute modélisation.
Elle permet d'examiner la structure du jeu de données, d'évaluer la distribution des variables, d'identifier d'éventuelles valeurs extrêmes et de comprendre les relations entre les caractéristiques techniques des appareils.

Cette phase a pour objectif de dégager les tendances générales du jeu de données et de préparer le terrain aux choix méthodologiques ultérieurs, notamment en matière de transformations, de normalisation ou d'encodage.

Dans le cadre de ce projet, plusieurs familles de variables sont étudiées :
- la varible cible,
- les caractéristiques de motorisation (type et puissance du moteur),
- les vitesses opérationnelles (vitesse maximale, de croisière, de décrochage),
- les performances opérationnelles (distances de décollage & d'atterrissage, taux montée),
- les variables structurelles (longueur et envergure),
- les capacités et masses (capacité en carburant, masse à vide),
- ainsi que l'autonomie des appareils.

Pour chacune de ces variables, des visualisations ont été réalisées afin d'identifier la forme de la distribution, l'importance des asymétries, la présence d'outliers et la pertinence d'éventuelles transformations telles que le logarithme.

Ces observations guideront ensuite le prétraitement nécessaire pour garantir une modélisation fiable et cohérente.

2.1 les caractéristiques de motorisation

2.1.1 Le type de motorisation

La variable engine_type est une variable catégorielle qui distingue trois types de motorisation présents dans le jeu de données : Piston, Propjet et Jet.

Table 2: Répartition des appareils selon le type de motorisation

Type de motorisation	**Nombre d'appareils**
Piston	378
Jet	71
Propjet	58

Ces catégories correspondent à des technologies de propulsion fondamentalement différentes, associées à des niveaux de performance, de capacité et de coûts distincts.

Les differents type de motorisation équivalent à une hétérogénéité dans les distributions des variables techniques.

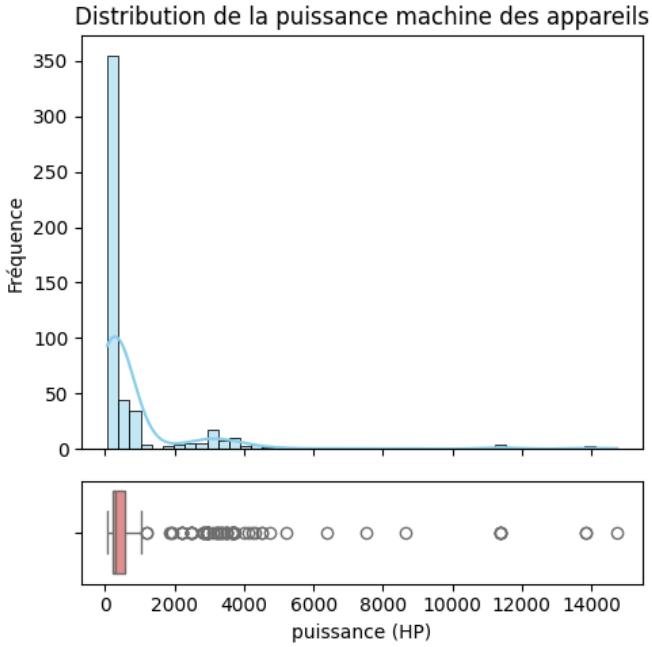
La forte proportion d'appareils à moteur piston concentre les observations sur des valeurs faibles, tandis que les jets et turbopropulseurs contribuent aux queues droites des distributions en raison de leurs performances supérieures.

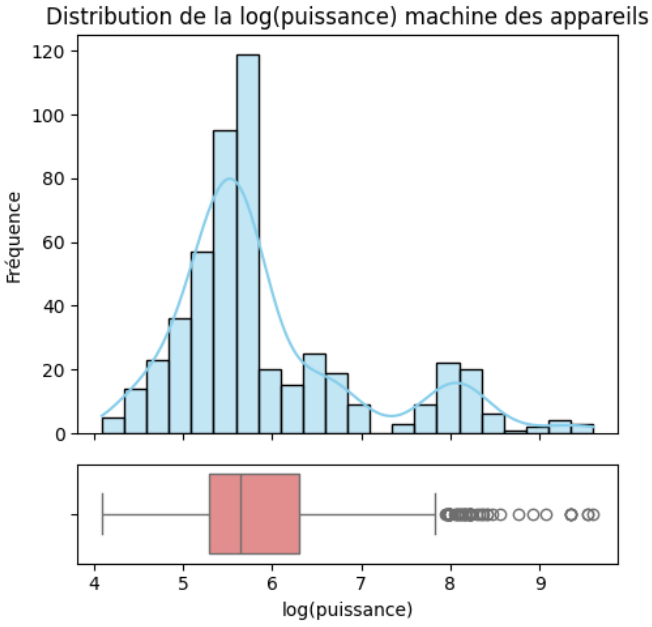
Cette variable joue donc un rôle structurant dans la formation du prix et est donc une varible importante.

Elle sera intégrée aux modèles par encodage catégoriel, afin de permettre aux algorithmes d'apprentissage automatique de capter les différences de prix systématiques liées au type de motorisation.

2.1.2 La puissance du moteur

La variable engine_power décrit la puissance totale du moteur des appareils, exprimée en chevaux (hp).





Elle présente une très forte dispersion, avec des valeurs allant de quelques dizaines de chevaux pour les avions légers à plusieurs milliers de chevaux pour les appareils les plus performants.
Les valeurs élevées correspondent à des appareils de grande capacité et à hautes performances, et ne peuvent être considérées comme aberrantes d'un point de vue technique.

La moyenne est nettement supérieure à la médiane, ce qui traduit une asymétrie marquée de la distribution.
Cette asymétrie s'explique par la coexistence, au sein du jeu de données, d'appareils de catégories très différentes, allant des avions à moteur piston aux jets.

Afin de limiter l'influence disproportionnée des observations extrêmes et de stabiliser la variance, une transformation logarithmique est appliquée à cette variable lors du prétraitement.
Cette transformation permet d'obtenir une distribution plus homogène et facilite l'estimation des modèles de régression utilisés par la suite.

2.2 Le prix de l'appareil

La variable price, qui constitue la variable cible du modèle, présente une forte dispersion. Les prix s'échelonnent d'environ 650 000 dollars à plus de 5 millions de dollars, traduisant la grande hétérogénéité des appareils présents dans le jeu de données.

La distribution est asymétrique à droite, avec une concentration importante d'appareils dans les gammes de prix intermédiaires et quelques modèles nettement plus coûteux.

Cette asymétrie est caractéristique des marchés de biens complexes et différenciés, où une minorité d'appareils haut de gamme exerce une influence significative sur la distribution globale.
Bien que cette variable ne fasse pas l'objet de transformation dans ce projet afin de préserver une interprétation directe des prédictions, son comportement justifie l'utilisation de modèles robustes et régularisés.

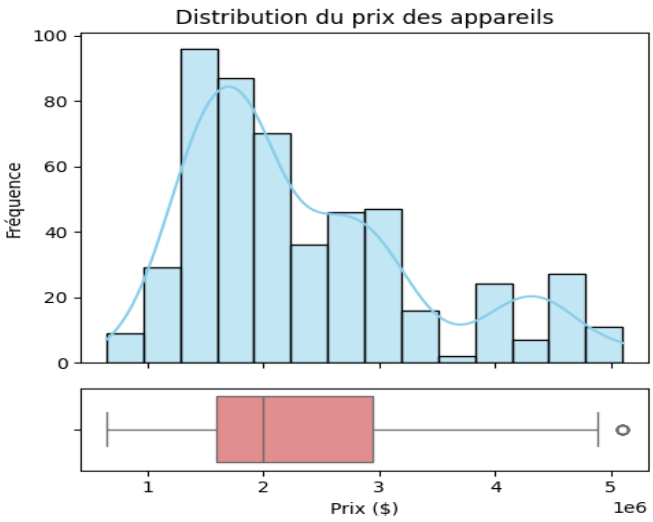


Figure 1: Distribution du prix

2.3 Les capacités et masses

Les variables liées aux capacités et aux masses des appareils décrivent des aspects fondamentaux de leur conception et de leurs performances opérationnelles.

Elles reflètent à la fois la taille de l'appareil, sa capacité d'emport et son autonomie potentielle, et jouent un rôle déterminant dans la formation du prix selon une approche hédonique.

2.3.1 Capacité en carburant

La variable fuel_tank représente la capacité du réservoir de carburant, exprimée en litres.

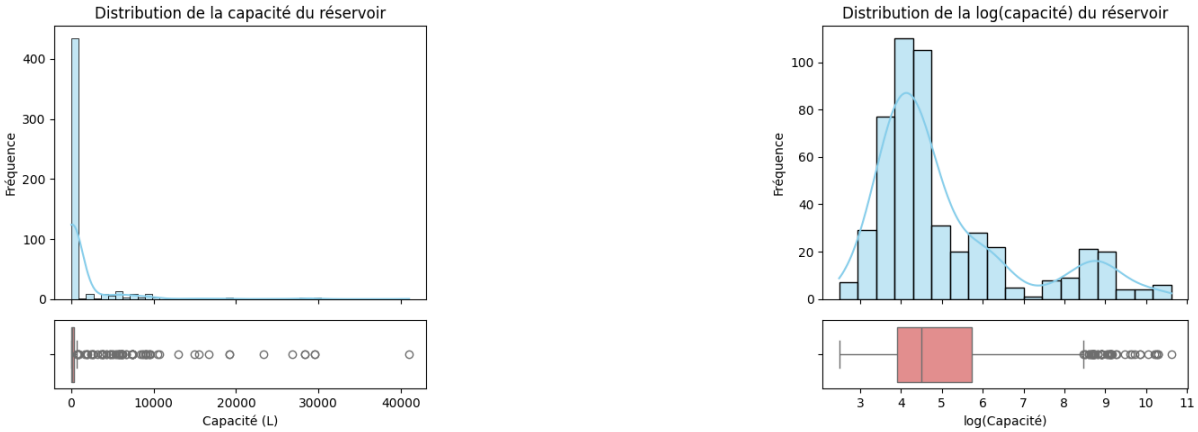


Figure 2: Distribution de la capacité du réservoir

Sa distribution présente une asymétrie très marquée, avec une majorité d'appareils disposant de réservoirs de faible à moyenne capacité, et un nombre restreint d'appareils dotés de capacités nettement plus élevées.

Les valeurs extrêmes observées correspondent donc à des configurations techniques réalistes et ne doivent pas être considérées comme aberrantes.

Toutefois, l'étalement important de la distribution et la présence d'une queue longue justifient l'application d'une transformation logarithmique afin de réduire l'influence des valeurs élevées et de stabiliser la variance lors de la modélisation.

2.3.2 Masse à vide

La variable empty_weight désigne la masse à vide de l'appareil, exprimée en kilogrammes.

Elle présente également une dispersion importante, avec des valeurs allant de quelques dizaines de kilogrammes pour les appareils les plus légers à plusieurs dizaines de tonnes pour les appareils les plus imposants.

La moyenne nettement supérieure à la médiane met en évidence une distribution fortement asymétrique.

Cette asymétrie est structurelle et reflète la diversité des catégories d'appareils présentes dans le jeu de données.

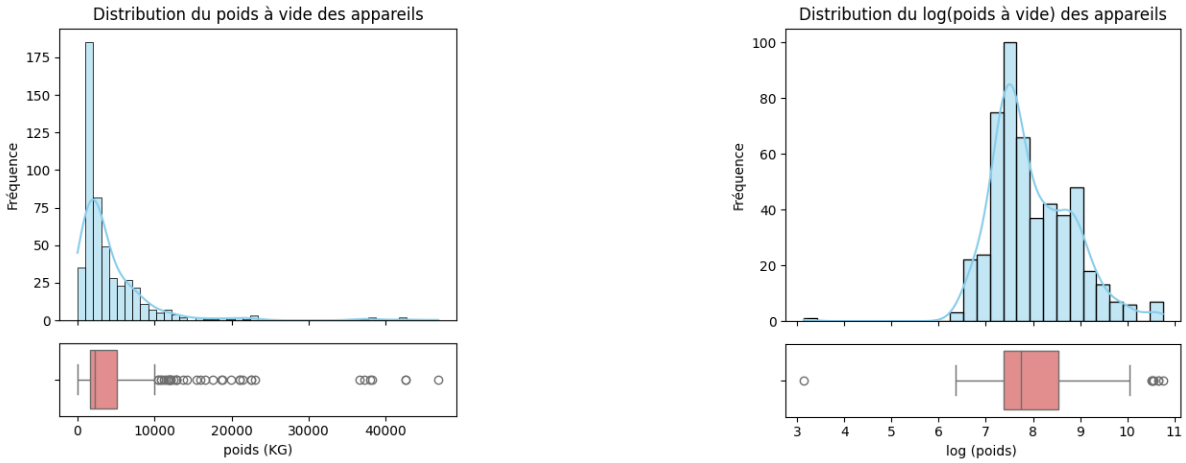


Figure 3: Distribution de la masse à vide

Les valeurs élevées sont cohérentes d'un point de vue technique et traduisent la présence d'appareils de grande capacité.

Afin de limiter l'impact disproportionné de ces observations sur l'estimation des modèles, une transformation logarithmique est appliquée à cette variable lors du prétraitement, permettant d'obtenir une distribution plus homogène et compatible avec les méthodes de régression pénalisée et d'apprentissage automatique.

2.4 Les vitesses opérationnelles

Les vitesses opérationnelless (la vitesse maximale, la vitesse de croisière et la vitesse de décrochage) sont des caractéristiques importantes de l'appareil.

Elles décrivent les performances aérodynamiques fondamentales des appareils et constituent évidemment des déterminants importants du prix.

Une observation générale issue de l'analyse descriptive est la bimodalité de plusieurs distributions, suggérant l'existence de deux grandes catégories d'appareils au sein du jeu de données, correspondant notamment aux avions légers et aux appareils plus performants (turbopropulseurs et jets).

Plus en détail, les trois variables de vitesse présentent des distributions asymétriques ainsi que plusieurs valeurs extrêmes toutefois elles ne constituent donc pas des anomalies et ne doivent pas être supprimées.

Contrairement à des variables telles que la puissance moteur, la masse à vide ou la capacité en carburant, les vitesses, l'application d'une transformation logarithmique n'apporterait ainsi pas de gain significatif

en termes de stabilisation de la variance ou de normalisation de la distribution.
En revanche, une standardisation est nécessaire afin d'assurer une échelle comparable entre les variables.

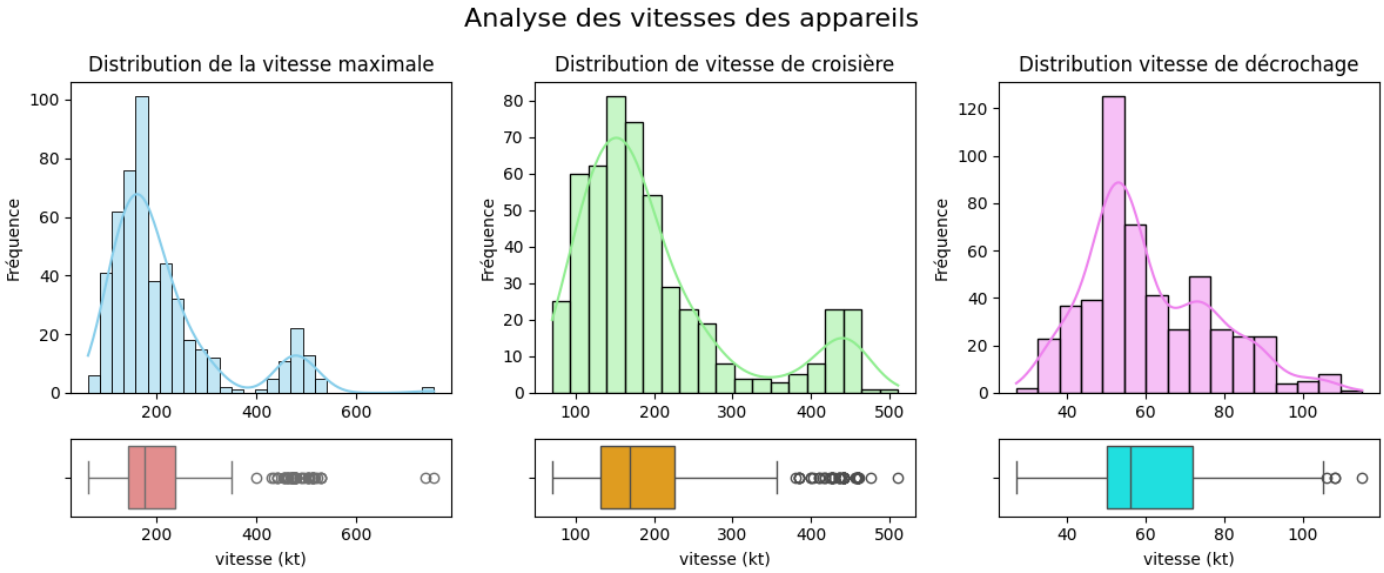


Figure 4: Distribution des vitesses

2.5 Les performances opérationnelles

Les performances opérationnelles décrivent les capacités des appareils à opérer dans des conditions réelles d'exploitation.

Elles traduisent les contraintes liées aux phases critiques du vol, notamment le décollage, l'atterrissage et la montée.

Ces variables influencent directement l'accessibilité des infrastructures, la flexibilité d'usage et la sécurité opérationnelle des appareils et sont donc pertinentes à considerer.

2.5.1 Distances de décollage et d'atterrissage

La distance de décollage (takeoff_distance) présente une distribution asymétrique, avec une majorité d'appareils nécessitant des distances comprises entre des valeurs modérées, et quelques appareils requérant des pistes nettement plus longues.

La distance d'atterrissage (landing_distance) affiche une asymétrie encore plus marquée, avec une queue de distribution particulièrement longue.

Certaines valeurs élevées traduisent des contraintes opérationnelles importantes associées aux appareils les plus lourds ou les plus performants.

Même après transformation logarithmique, cette variable conserve une forte dispersion, ce qui justifie une attention particulière lors du prétraitement afin de limiter l'influence disproportionnée des observations extrêmes.

Dans les deux cas, les valeurs élevées observées correspondent à des configurations techniques réalistes et ne doivent pas être interprétées comme des anomalies.

Une transformation logarithmique est néanmoins appliquée afin de stabiliser la variance et d'améliorer la robustesse des modèles de régression.

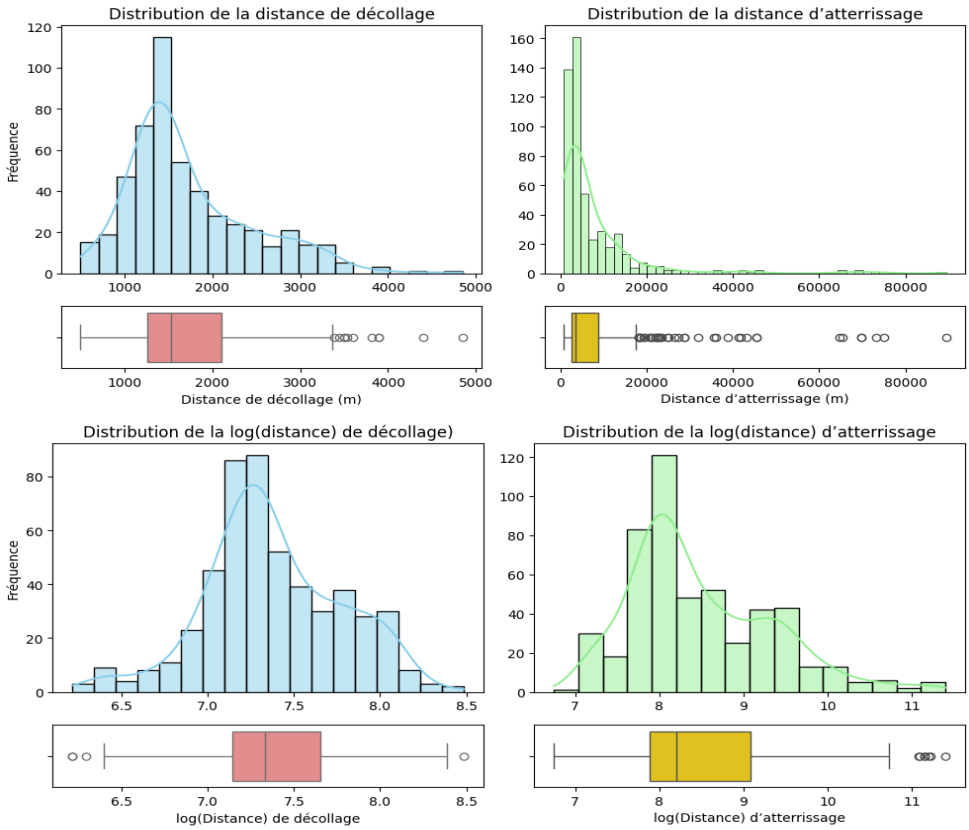


Figure 5: Distribution de la distance

2.5.2 Taux de montée

Les variables all_eng_roc et out_eng_roc décrivent respectivement le taux de montée avec tous les moteurs en fonctionnement et le taux de montée en cas de panne d'un moteur.

Ces indicateurs mesurent la capacité de l'appareil à gagner de l'altitude dans des conditions normales et dégradées, et constituent des critères importants de performance et de sécurité.

Les distributions de ces variables sont caractérisées par une dispersion notable et une asymétrie à droite, traduisant la coexistence d'appareils aux performances très hétérogènes.

Les valeurs les plus élevées correspondent à des appareils dotés d'une puissance importante et de capacités opérationnelles avancées.

Ces observations sont cohérentes d'un point de vue technique et sont conservées dans l'analyse.
Une log standardisation est appliquée afin de permettre leur intégration adéquate dans les modèles de régression pénalisée et d'apprentissage automatique.

2.6 Les variables structurelles

Les variables structurelles décrivent les dimensions physiques des appareils et constituent des indicateurs essentiels de leur conception et de leur capacité.
Elles reflètent la taille globale de l'appareil, influencent les performances aérodynamiques et conditionnent en partie les usages possibles ainsi que les coûts de fabrication.

De maniere generale, elles jouent un rôle important dans la formation du prix, c'est à dire les appareils de plus grande taille étant généralement associés à des niveaux de performance et de prix plus élevés.

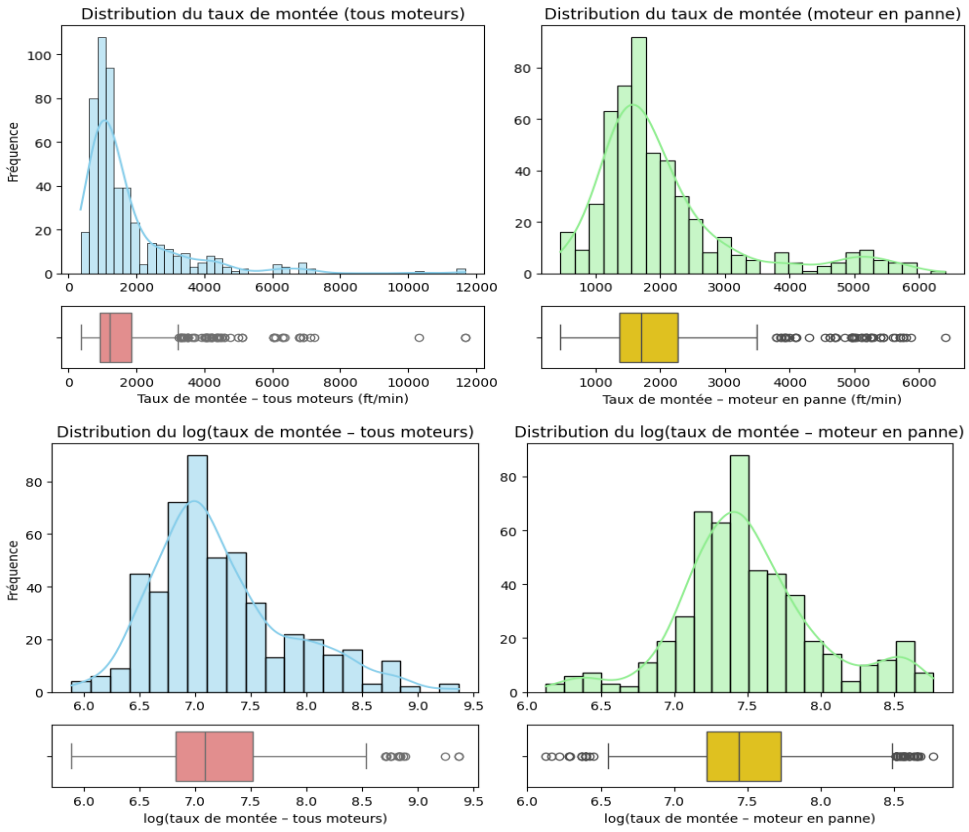


Figure 6: Distribution des taux de montée

Les variables considérées, à savoir la longueur et l'envergure, présentent des distributions asymétriques caractérisées par une forte concentration d'appareils de dimensions modestes et une queue droite correspondant à des appareils de plus grande taille.

Les boxplots mettent en évidence plusieurs valeurs extrêmes qui ne traduisent pas des anomalies, mais reflètent la coexistence d'appareils légers et d'appareils plus imposants au sein du jeu de données.

Cette hétérogénéité est cohérente d'un point de vue technique et structurel.

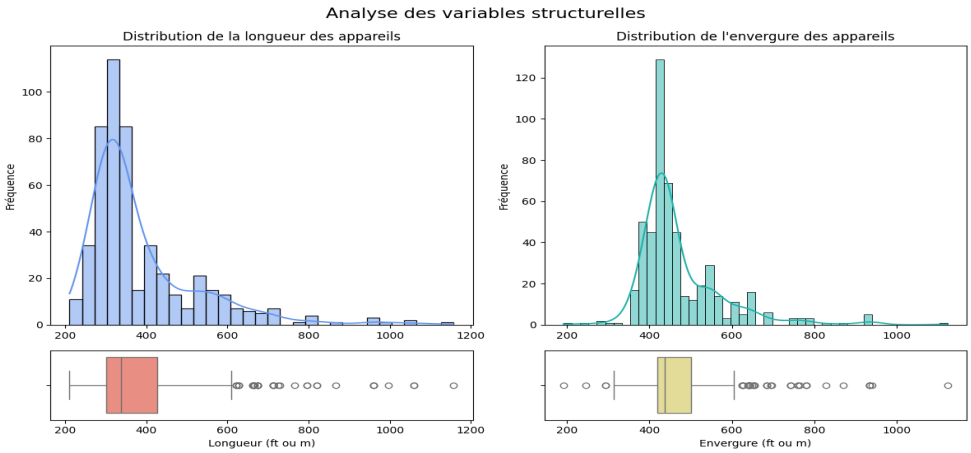


Figure 7: Distribution des dimensions

Afin de réduire l'influence disproportionnée des valeurs élevées et de stabiliser la variance, une transformation logarithmique est appliquée à ces variables lors du prétraitement, permettant d'obtenir des distributions plus homogènes et mieux adaptées aux modèles de régression pénalisée et d'apprentissage automatique.

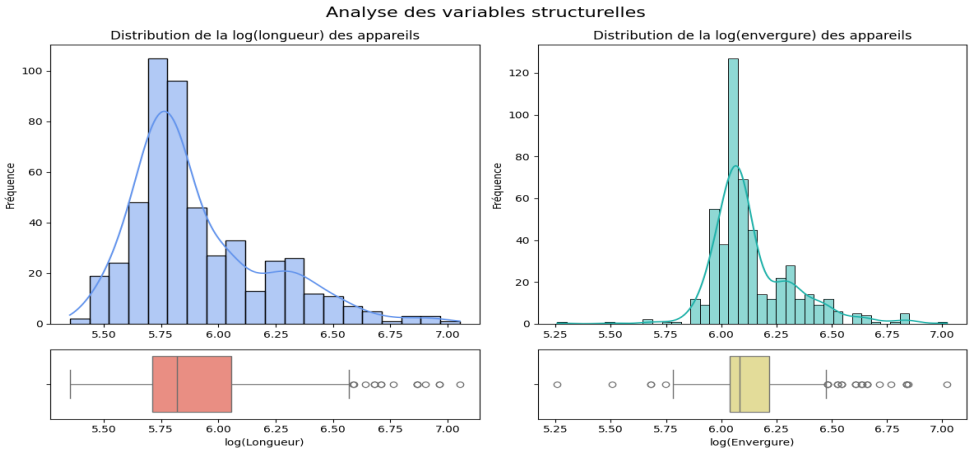


Figure 8: Distribution des dimensions

2.7 L'autonomie des appareils

L'autonomie des appareils, mesurée par la variable *range*, correspond à la distance maximale qu'un appareil peut parcourir sans ravitaillement.

Elle constitue un indicateur central des capacités opérationnelles et conditionne fortement les usages possibles, qu'il s'agisse de vols courts, régionaux ou de missions longue distance.

L'autonomie est généralement associée à des niveaux de prix plus élevés, les appareils capables de parcourir de longues distances intégrant des contraintes techniques et des coûts de conception plus importants.

La distribution de l'autonomie présente une asymétrie marquée, avec une majorité d'appareils offrant une portée limitée à intermédiaire et une queue droite correspondant à des appareils conçus pour des vols longue distance.

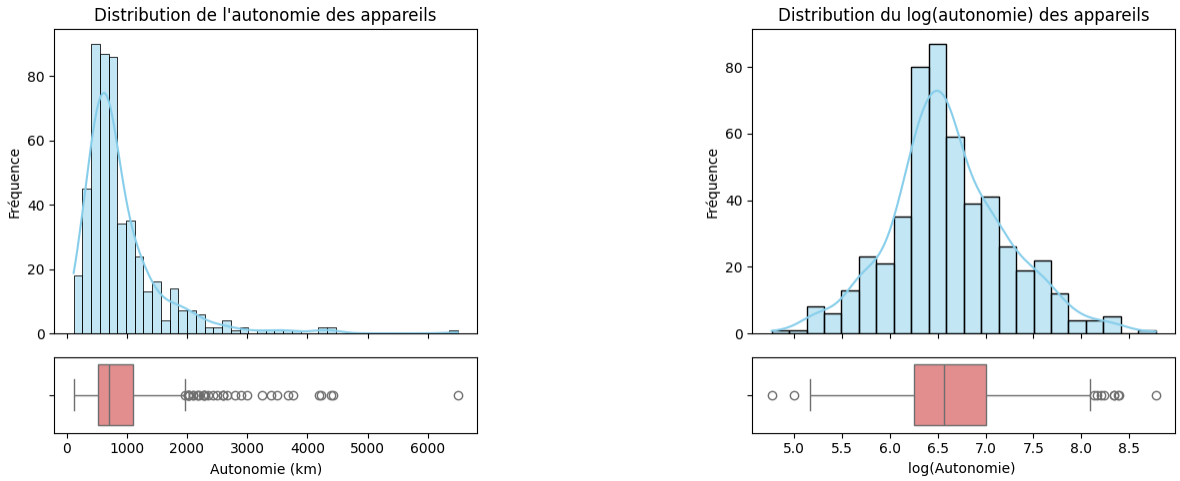


Figure 9: Distribution de l'autonomie

Compte tenu de la dispersion observée et de la présence d'une queue longue, une transformation logarithmique est appliquée à cette variable lors du prétraitement afin de stabiliser la variance et de limiter l'influence disproportionnée des appareils à très grande autonomie sur l'estimation des modèles.

3 Prétraitement

Le prétraitement des données constitue une étape déterminante avant l'entraînement des modèles de machine learning.
Compte tenu de l'hétérogénéité des variables, tant en termes d'unités que de distributions, un ensemble de transformations est nécessaire afin de garantir la stabilité numérique des algorithmes, d'éviter une influence disproportionnée de certaines variables et d'améliorer la capacité de généralisation des modèles.

Dans un premier temps, les observations pour lesquelles la variable cible (*price*) est manquante ont été mises de coté, afin de disposer d'un échantillon exploitable pour l'apprentissage supervisé.

L'analyse descriptive a mis en évidence des distributions fortement asymétriques et des queues longues pour plusieurs variables, notamment la puissance moteur, la masse à vide, la capacité en carburant, les distances opérationnelles ainsi que l'autonomie.
Pour ces variables, une transformation logarithmique de type $\log(x)$ est appliquée.

Cette transformation permet de stabiliser la variance, de réduire l'impact des valeurs extrêmes et de rendre les distributions plus proches d'une forme gaussienne, ce qui est particulièrement bénéfique pour les modèles linéaires pénalisés.

Les variables numériques ne présentant pas de queue lourde, telles que les vitesses ou les taux de montée, sont quant à elles conservées dans leur forme originale mais font l'objet d'une standardisation par centrage-réduction.
Cette étape est indispensable pour les modèles sensibles à l'échelle des variables, comme la régression Ridge, la régression Lasso et les réseaux de neurones.

La variable catégorielle (*engine_ type* est intégrée au moyen d'un encodage catégorielle. Cette méthode permet de représenter les catégories sous forme de variables binaires sans introduire de hiérarchie artificielle entre elles.

L'ensemble des opérations de prétraitement est implémenté au sein d'un pipeline unique, garantissant que les paramètres des transformations (logarithme, standardisation, encodage) sont estimés exclusivement sur l'échantillon d'apprentissage puis appliqués à l'échantillon de test.

Cette approche permet d'éviter toute fuite d'information (*data leakage*) et assure la reproductibilité des résultats 7.1.

4 Méthodologie et calibrage des modèles

L'objectif de ce projet est de prédire le prix des appareils aéronautiques à partir de leurs caractéristiques techniques, en s'inscrivant dans le cadre des modèles de prix hédoniques et en mobilisant des méthodes de machine learning.

La démarche méthodologique adoptée vise à comparer des modèles de complexité croissante, en mettant en regard performance prédictive et interprétabilité.

Dans un premier temps, le jeu de données est divisé en deux:
- un échantillon d'apprentissage, représentant 80 % des observations,
- et un échantillon de test, représentant les 20 % restants.

Cette séparation permet d'évaluer les performances des modèles sur des données non observées lors de l'entraînement et d'apprécier leur capacité de généralisation.

La modélisation débute par l'estimation d'un modèle de **régression linéaire**, utilisé comme modèle de référence. Il permet d'établir un premier lien entre le prix et les caractéristiques des appareils, tout en offrant une interprétabilité directe des coefficients.
Toutefois, l'analyse descriptive a mis en évidence une forte corrélation entre plusieurs variables explicatives, susceptible d'entraîner une instabilité des estimations et de limiter les performances prédictives du modèle linéaire.

Afin de remédier à ces limites, des modèles de **régression pénalisée** sont ensuite implémentés.
La **régression Ridge** introduit une pénalisation de type L_2 sur les coefficients, permettant de réduire la variance des estimations et de gérer efficacement la multicolinéarité.
La **régression Lasso**, reposant sur une pénalisation L_1, permet quant à elle une sélection automatique des variables en annulant les coefficients associés aux variables les moins informatives.
Dans les deux cas, la valeur optimale du paramètre de régularisation est déterminée par validation croisée.

En complément de ces approches linéaires, des modèles plus flexibles sont considérés afin de capturer d'éventuelles relations non linéaires entre les caractéristiques techniques et le prix.
Une **régression polynomiale** est utilisée pour introduire des non-linéarités de manière contrôlée, tandis qu'un **réseau de neurones multicouches** est mis en œuvre afin d'explorer une approche d'apprentissage automatique plus expressive.

Le calibrage des modèles consiste à déterminer les paramètres optimaux permettant d'obtenir les meilleures performances prédictives tout en évitant le surapprentissage.

Dans ce projet, il repose principalement sur la sélection des hyperparamètres associés aux modèles pénalisés et aux modèles non linéaires, à partir des seules données d'apprentissage.

Enfin, les performances des différents modèles sont comparées à l'aide de mesures standard en régression, en particulier l'erreur quadratique moyenne (MSE) et le coefficient de détermination (R^2), calculées sur l'échantillon de test.

Cette comparaison permet d'évaluer les gains apportés par les méthodes pénalisées et non linéaires par rapport au modèle linéaire de référence, et de discuter les compromis entre performance prédictive et interprétabilité.

4.1 Modèle de régression linéaire

Le modèle de régression linéaire est utilisé comme point de départ de l'analyse et constitue le modèle de référence (*baseline*) 7.2.

Dans le cadre d'un modèle de prix hédonique, on suppose que le prix d'un appareil peut être exprimé comme une combinaison linéaire de ses caractéristiques techniques.

Formellement, le modèle s'écrit comme suit :

$$log(price_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \tag{1}$$

où $price_i$ désigne le prix de l'appareil i, X_{ij} représente la j-ième caractéristique technique de cet appareil, β_j le coefficient associé, β_0 l'ordonnée à l'origine, et ε_i un terme d'erreur supposé de moyenne nulle.

Chaque coefficient β_j mesure ainsi l'effet marginal d'une variable explicative sur le prix, toutes choses égales par ailleurs.
Le modèle est estimé sur les données prétraitées, ce qui garantit la comparabilité des coefficients et la stabilité numérique de l'estimation.

Bien que simple et facilement interprétable, la régression linéaire présente certaines limites dans ce contexte.

La forte corrélation entre plusieurs caractéristiques techniques, notamment celles liées à la taille, à la masse et à la puissance, peut entraîner une multicolinéarité importante, rendant les estimations sensibles aux variations de l'échantillon.

Les performances de ce modèle servent néanmoins de point de référence pour évaluer les gains apportés par les modèles pénalisés et les approches plus flexibles présentés par la suite.

4.2 Modèles de régression pénalisée

Les modèles de régression pénalisée étendent la régression linéaire classique en introduisant un terme de pénalisation sur les coefficients.
Ce terme permet de contrôler le compromis entre biais et variance : **une pénalisation trop faible peut conduire à des estimations instables, tandis qu'une pénalisation excessive peut entraîner un sous-ajustement du modèle.**

Afin de sélectionner la valeur optimale du paramètre de régularisation, une procédure de validation croisée est mise en œuvre sur l'échantillon d'apprentissage.

Cette approche consiste à évaluer les performances du modèle sur plusieurs partitions des données et à retenir la configuration offrant le meilleur compromis en termes d'erreur de prédiction.

4.2.1 La régression Ridge

La régression Ridge introduit une pénalisation de type L_2 sur les coefficients du modèle. Elle s'écrit formellement comme la solution du problème d'optimisation suivant :

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(log(price_i) - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \alpha \sum_{j=1}^{p} \beta_j^2, \tag{2}$$

où $\alpha \geq 0$ est le paramètre de régularisation contrôlant l'intensité de la pénalisation.
Lorsque α augmente, les coefficients sont progressivement contraints vers zéro sans jamais être annulés.
Cette pénalisation permet de réduire la variance des estimations et de gérer efficacement la multicolinéarité entre les variables explicatives 7.3.1.

Dans ce projet, la valeur optimale de α est sélectionnée par validation croisée, en minimisant l'erreur quadratique moyenne calculée sur les différentes partitions de l'échantillon d'apprentissage 10.
Sa valeur ici est le α permettant de stabiliser les coefficients sans les annuler : $\underline{\alpha \approx 0{,}305}$

4.2.2 La régression Lasso

La régression Lasso repose sur une pénalisation de type L_1, qui conduit à une formulation différente du problème d'optimisation :

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(log(price_i) - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \alpha \sum_{j=1}^{p} |\beta_j|. \tag{3}$$

Contrairement à la pénalisation Ridge, la pénalisation L_1 a pour effet d'annuler exactement certains coefficients lorsque le paramètre de régularisation est suffisamment élevé.
La régression Lasso permet ainsi non seulement de contrôler la complexité du modèle, mais également d'effectuer une sélection automatique des variables, en ne conservant que les caractéristiques les plus informatives 7.3.2.

Comme pour la régression Ridge, le paramètre de régularisation est déterminé par validation croisée sur l'échantillon d'apprentissage, en cherchant à minimiser l'erreur de prédiction 11.

Le graphique du nombre de coefficients nuls montre une sélection automatique des variables, traduisant une parcimonie croissante du modèle. La valeur optimale de $\alpha \approx 0{,}0002$, déterminée par validation croisée, correspond à un compromis satisfaisant entre réduction de la complexité et maintien des performances prédictives.

4.3 Modèles de régression non linéaires

Les modèles de régression non linéaires sont introduits afin de capturer des relations plus complexes entre les caractéristiques techniques des appareils et leur prix, que les modèles linéaires ou pénalisés ne peuvent pas représenter explicitement.

Ces approches permettent de relâcher l'hypothèse de linéarité stricte et d'explorer des interactions ou des effets de seuil potentiels entre les variables explicatives.

4.3.1 La régression polynomiale

La régression polynomiale constitue une extension du modèle linéaire dans laquelle les variables explicatives sont enrichies par des termes polynomiaux.
Elle permet d'introduire des non-linéarités de manière contrôlée tout en conservant une structure relativement interprétable. Le modèle peut s'écrire comme :

$$price_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \sum_{j=1}^{p} \sum_{k=2}^{d} \beta_{jk} X_{ij}^k + \varepsilon_i, \tag{4}$$

où d représente le degré du polynôme.
Le calibrage du modèle repose sur le choix du degré du polynôme.
Un degré trop faible peut conduire à un sous-ajustement, tandis qu'un degré trop élevé augmente le risque de surapprentissage. Le Tableau ci dessous présente les performances de la régression polynomiale en fonction du degré du polynôme.

Table 3: Performances de la régression polynomiale selon le degré

Degré du polynôme	**MSE**	R^2
1	0.0291	0.8092
2	0.0773	0.4927
3	156.8483	-1028.7958
4	112.0192	-734.4681
5	212.8314	-1396.3560

Le modèle de degré 1, équivalent à la régression linéaire, offre les meilleures performances sur l'échantillon de test. L'augmentation du degré entraîne une dégradation rapide des performances, caractérisée par

une explosion du MSE et des valeurs fortement négatives de R^2, révélant un surajustement sévère. En somme, les modèles polynomiaux de degré élevé ne sont pas adaptés dans ce contexte 7.4.1.

4.3.2 Les réseaux de neurones

Les réseaux de neurones multicouches représentent une approche d'apprentissage automatique plus flexible, capable d'approximer des relations non linéaires complexes entre les variables explicatives et le prix.

Ils reposent sur l'empilement de couches de neurones, chaque neurone appliquant une transformation affine suivie d'une fonction d'activation non linéaire.

Le calibrage du réseau de neurones porte sur plusieurs hyperparamètres, notamment le nombre de couches cachées, le nombre de neurones par couche et le taux de régularisation 7.4.2.

Dans le cas de notre projet, le réseau de neurones est calibré sur l'échantillon d'apprentissage au moyen d'un sous-échantillon de validation.
Le choix de l'architecture, du taux d'apprentissage, et de la régularisation est effectué en comparant les performances sur la perte de validation. Un mécanisme d'*early stopping* est utilisé pour sélectionner automatiquement le nombre d'époques et limiter le surapprentissage.

Comme illustré par la Figure 12, les courbes d'apprentissage et de validation convergent de manière stable et restent proches, indiquant une bonne généralisation du réseau de neurones.
La configuration optimale trouvé est **une architecture (128, 32, 16) avec un taux d'apprentissage de 0,001, sans régularisation supplémentaire**

5 Comparaison et analyse de l'importance des variables

Cette section a pour objectif de comparer les performances des différents modèles estimés et d'analyser les variables qui contribuent le plus à la prédiction du prix des appareils aéronautiques.

La comparaison permet d'évaluer les gains apportés par les méthodes pénalisées et non linéaires par rapport au modèle linéaire de référence, tandis que l'analyse de l'importance des variables apporte des éléments d'interprétation dans le cadre du modèle de prix hédonique.

5.1 Comparaison des performances des modèles

Les performances prédictives des modèles sont évaluées sur l'échantillon de test à l'aide de deux indicateurs standards en régression : l'erreur quadratique moyenne (MSE) et le coefficient de détermination (R^2).
Le MSE mesure l'écart moyen entre les valeurs prédites et les valeurs observées, tandis que le R^2 indique la part de la variance du prix expliquée par le modèle.

Table 4: Comparaison des performances prédictives des différents modèles

Modèle	**MSE**	R^2
Régression linéaire	0.0344	0.7968
Ridge	0.0366	0.7838
Lasso	0.0366	0.7843
Polynomiale	0.0291	0.8092
Réseau de neurones	0.0564	0.6673

Le Tableau 4 présente les performances prédictives des différents modèles estimés sur l'échantillon de test.

Dans l'ensemble, les résultats montrent que les approches linéaires et pénalisées offrent de très bonnes performances, avec des valeurs de R^2 comprises entre 0,78 et 0,81, indiquant qu'une part importante de la variance du prix des appareils est expliquée par leurs caractéristiques techniques.

La régression linéaire constitue un point de référence solide, avec un pouvoir explicatif élevé et une grande interprétabilité. Toutefois, la présence de multicolinéarité entre certaines variables techniques limite sa robustesse.
Les modèles Ridge et Lasso présentent des performances légèrement inférieures en termes de R^2, mais offrent un avantage méthodologique important.

La régularisation Ridge permet de stabiliser les coefficients en réduisant leur variance, tandis que la régularisation Lasso introduit une parcimonie supplémentaire en sélectionnant automatiquement les variables les plus informatives.

La régression polynomiale de degré 1, équivalente à la régression linéaire, apparaît comme le modèle offrant les meilleures performances globales sur l'échantillon de test.
En revanche, l'augmentation du degré du polynôme conduit à une dégradation rapide des performances, traduisant un surajustement sévère.
Ces résultats confirment que l'introduction de non-linéarités explicites n'est pas pertinente dans ce contexte.

Enfin, le réseau de neurones affiche des performances nettement inférieures à celles des modèles linéaires pénalisés. Malgré une convergence stable et un calibrage adéquat, sa flexibilité accrue ne se traduit pas par une amélioration de la prédiction.
Ce constat suggère que, compte tenu de la taille de l'échantillon et de la structure des données, des modèles plus simples et régularisés sont mieux adaptés.

5.2 Analyse de l'importance des variables

L'analyse de l'importance des variables est menée principalement à partir des modèles linéaires et pénalisés, dont les coefficients sont directement interprétables.

Dans le cadre d'un modèle hédonique, les coefficients estimés reflètent la contribution marginale de chaque caractéristique technique à la formation du prix, toutes choses égales par ailleurs.
Dans notre cas, les variables explicatives ayant été préalablement standardisées, les coefficients peuvent être interprétés en termes d'importance relative.

Le Tableau 5 présente les coefficients standardisés estimés par les modèles OLS, Ridge et Lasso, ainsi que l'importance relative des variables mesurée par la valeur absolue des coefficients Ridge.

Les résultats du tableau mettent en évidence le rôle central des variables de performance opérationnelle. Le taux de montée, l'autonomie et la vitesse de croisière apparaissent comme les principaux déterminants du prix, avec des coefficients élevés et stables entre les modèles.
Ces caractéristiques reflètent directement les capacités techniques et les usages potentiels des appareils et ont donc une explication logique à leur importance.

Les variables liées à la taille et aux capacités, telles que la masse à vide et la capacité en carburant, contribuent également de manière significative, bien que leur importance relative soit plus modérée une fois la multicolinéarité prise en compte par la régularisation Ridge.

Table 5: Coefficients standardisés et importance relative des variables

Variable	**OLS**	**Ridge**	**Lasso**	**|Ridge|**
Taux de montée (tous moteurs)	0.718	0.608	0.693	0.608
Autonomie	0.522	0.462	0.517	0.463
Vitesse de croisière	0.487	0.446	0.471	0.446
Distance de décollage	0.382	0.297	0.349	0.297
Taux de montée (un moteur)	-0.320	-0.238	-0.272	0.238
Vitesse maximale	0.220	0.198	0.166	0.198
Masse à vide	0.633	0.195	0.276	0.195
Capacité en carburant	0.262	0.174	0.111	0.174
Vitesse de décrochage	0.049	0.077	0.000	0.077
Envergure	0.211	0.070	0.047	0.070
Puissance moteur	-0.017	-0.022	0.000	0.022
Longueur	-0.023	0.019	0.000	0.019
Distance d'atterrissage	-0.480	0.010	0.000	0.010
Type de motorisation	-0.019	0.003	0.000	0.003

La régression Lasso confirme ces résultats en sélectionnant un sous-ensemble restreint de variables informatives et en annulant les coefficients associés aux caractéristiques redondantes ou marginales.

Dans l'ensemble, la cohérence des signes et des ordres de grandeur entre les modèles renforce la robustesse de l'analyse et confirme la pertinence du cadre de prix hédonique pour expliquer la formation du prix des appareils aéronautiques.

Dans les modèles non linéaires, l'interprétation directe de l'importance des variables est plus limitée. Toutefois, la cohérence globale des résultats avec ceux obtenus à partir des modèles linéaires et pénalisés suggère que les principaux déterminants du prix sont robustes à la méthode de modélisation employée.

6 Application

Cette section illustre l'application du modèle retenu comme le plus performant à des observations pour lesquelles le prix n'est pas renseigné dans le jeu de données initial.
Plus précisément, le modèle est utilisé pour prédire le prix de dix appareils aéronautiques dont la variable cible est manquante.

Conformément à la méthodologie adoptée, les mêmes étapes de prétraitement que celles appliquées aux données d'apprentissage sont reproduites à l'identique sur ces nouvelles observations.

Les variables explicatives sont ainsi encodées et standardisées à l'aide des transformations apprises sur l'échantillon d'apprentissage, afin de garantir la cohérence des prédictions et d'éviter toute fuite d'information.
Le modèle sélectionné est ensuite utilisé pour produire une prédiction du logarithme du prix des appareils 7.5.
Les valeurs prédites sont enfin retranscrites dans l'échelle originale en appliquant la fonction exponentielle, ce qui permet d'obtenir une estimation du prix en niveau.

Cette application met en évidence la capacité du modèle à généraliser à de nouvelles observations et illustre son utilisation concrète dans un cadre de prédiction de prix.

Les résultats obtenus fournissent des ordres de grandeur plausibles des prix des appareils, cohérents avec les caractéristiques techniques observées et avec les relations mises en évidence.

Conclusion

L'objectif de ce projet était de prédire le prix d'appareils aéronautiques à partir de leurs caractéristiques techniques, en s'inscrivant dans le cadre des modèles de prix hédoniques et en mobilisant des méthodes de machine learning.

À partir d'un jeu de données décrivant les performances, les capacités et les dimensions physiques des appareils, plusieurs approches de modélisation ont été mises en œuvre et comparées.
Les résultats montrent que les modèles linéaires constituent une base particulièrement solide pour ce type de données.
La régression linéaire fournit déjà de bonnes performances prédictives et une interprétabilité directe des coefficients.
L'introduction de la régularisation à travers les modèles Ridge et Lasso permet de stabiliser les estimations en présence de multicolinéarité et d'améliorer la robustesse du modèle, tout en conservant un fort pouvoir explicatif.
La régression Lasso se distingue en outre par sa capacité à sélectionner automatiquement les variables les plus informatives.

Les approches plus complexes, telles que la régression polynomiale de degré élevé et le réseau de neurones, n'apportent pas d'amélioration significative des performances dans ce contexte.
Ces résultats suggèrent que, compte tenu de la taille de l'échantillon et de la structure des données, des modèles régularisés et interprétables offrent un meilleur compromis entre performance prédictive et stabilité.

L'analyse de l'importance des variables met en évidence le rôle central des caractéristiques de performance opérationnelle, telles que le taux de montée, l'autonomie et les vitesses, ainsi que des variables liées aux capacités et à la taille des appareils.

Enfin, l'application du modèle retenu à des appareils dont le prix n'était pas renseigné illustre l'intérêt pratique de l'approche proposée.
Les prix prédits présentent des ordres de grandeur plausibles et cohérents avec les caractéristiques techniques observées, confirmant la capacité du modèle à généraliser à de nouvelles observations.

Ce travail pourrait être prolongé par l'intégration de variables supplémentaires, telles que l'âge des appareils ou des informations sur le marché secondaire, ainsi que par l'exploration de méthodes d'interprétabilité avancées.

Néanmoins, les résultats obtenus montrent que les modèles de régression pénalisée constituent un outil pertinent et efficace pour la prédiction du prix des appareils aéronautiques.

7 Annexe

7.1 Prétraitement des données

```
import numpy as np
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split

def pre_traitement(data):
    X = data.drop(columns=["price"])
    y = np.log(data["price"])

    numeric_features =
       X.select_dtypes(include=[np.number]).columns.tolist()

    log_features = [ "engine_power", "fuel_tank",
       "empty_weight","takeoff_distance",
    "landing_distance","range","length", "wing_span", "all_eng_roc",
       "out_eng_roc"   ]

    for col in log_features:
        X[col]=np.log(X[col])

    scaler = MinMaxScaler()
    X_scaled=scaler.fit_transform(X[numeric_features])
    X_scaled=pd.DataFrame(X_scaled,
       columns=X[numeric_features].columns)

    X_scaled['engine_type']=data['engine_type'].map({
    'Piston': 0, 'Propjet': 1, 'Jet':2 }).astype('category')


    return X_scaled, y

X_scaled, y=pre_traitement(aircraft_data)

X_train, X_test, y_train, y_test = train_test_split(
    X_scaled, y,
    test_size=0.2
)

X_train = pd.DataFrame(X_train, columns=X_scaled.columns)
X_test = pd.DataFrame(X_test, columns=X_scaled.columns)
```

7.2 Implémentation des modèles linéaires

```
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.linear_model import LinearRegression

lin_reg = LinearRegression()
lin_reg.fit(X_train, y_train)
```

```
y_pred_lin = lin_reg.predict(X_test)

print("Linear Regression")
print("R_2 :", r2_score(y_test, y_pred_lin))
print("MSE :", mean_squared_error(y_test, y_pred_lin))
```

7.3 Régression pénalisée

7.3.1 Régression Ridge

Listing 1: Calibration et validation du modèle Ridge

```
from sklearn.linear_model import Ridge, RidgeCV
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt
import numpy as np

alphas = np.logspace(-3, 3, 100)
coefs = []

for a in alphas:
    ridge = Ridge(alpha=a)
    ridge.fit(X_train, y_train)
    coefs.append(ridge.coef_)

plt.figure(figsize=(8,5))
plt.plot(alphas, coefs)
plt.xscale("log")
plt.xlabel("Alpha")
plt.ylabel("Coefficients")
plt.title("Profil des coefficients du modèle Ridge")
plt.show()

ridge_cv = RidgeCV(alphas=alphas, cv=10)
ridge_cv.fit(X_train, y_train)

print("Alpha optimal (Ridge) :", ridge_cv.alpha_)

y_pred_ridge = ridge_cv.predict(X_test)
print("Ridge MSE :", mean_squared_error(y_test, y_pred_ridge))
print("Ridge R_2 :", r2_score(y_test, y_pred_ridge))
```

7.3.2 Régression Lasso

Listing 2: Calibration et validation du modèle Lasso

```
from sklearn.linear_model import Lasso, LassoCV
from sklearn.metrics import mean_squared_error, r2_score
import numpy as np
import matplotlib.pyplot as plt

alphas = np.logspace(-3, 3, 100)
coefs = []
nb_nul = []
```

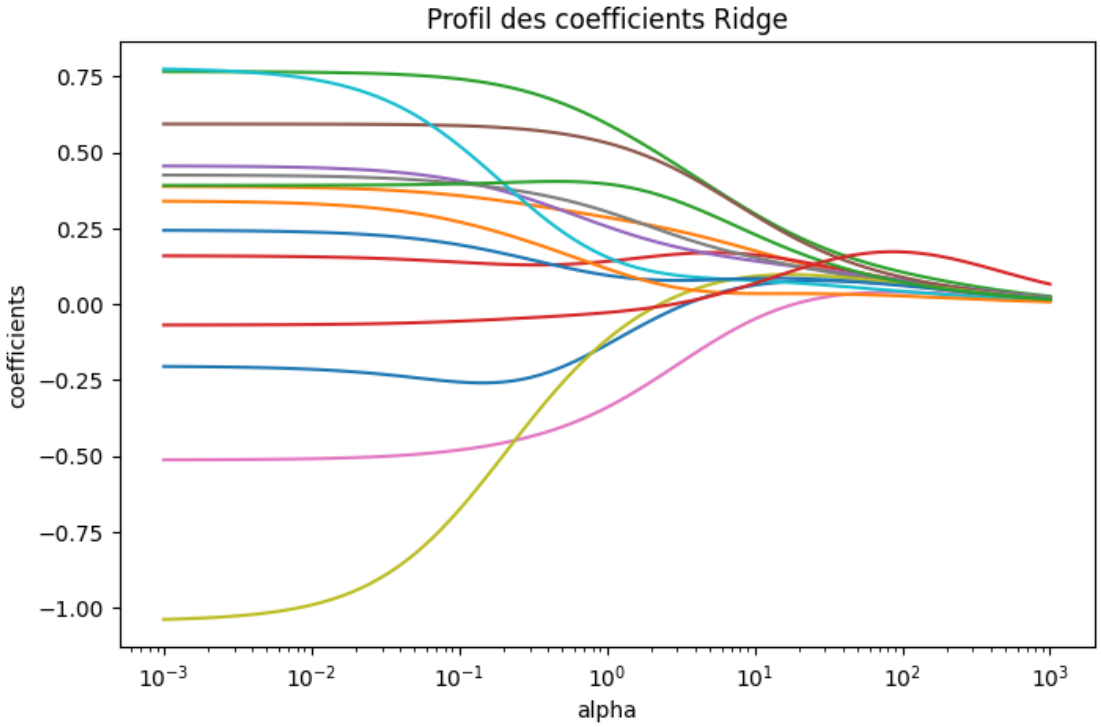


Figure 10

```
for a in alphas:
    lasso = Lasso(alpha=a, max_iter=10_000)
    lasso.fit(X_train, y_train)
    coefs.append(lasso.coef_)
    nb_nul.append(np.sum(lasso.coef_ == 0))

plt.figure(figsize=(12,5))

plt.subplot(1,2,1)
plt.plot(alphas, coefs)
plt.xscale("log")
plt.xlabel("Alpha")
plt.ylabel("Coefficients")
plt.title("Profil des coefficients du modèle Lasso")

plt.subplot(1,2,2)
plt.plot(alphas, nb_nul)
plt.xscale("log")
plt.xlabel("Alpha")
plt.ylabel("Nombre de coefficients nuls")
plt.title("Parcimonie induite par la pénalisation Lasso")

plt.show()

lasso_cv = LassoCV(alphas=alphas, cv=10, max_iter=10_000)
lasso_cv.fit(X_train, y_train)

print("Alpha optimal (Lasso) :", lasso_cv.alpha_)

y_pred_lasso = lasso_cv.predict(X_test)
```

```
print("Lasso MSE :", mean_squared_error(y_test, y_pred_lasso))
print("Lasso R_2 :", r2_score(y_test, y_pred_lasso))
```

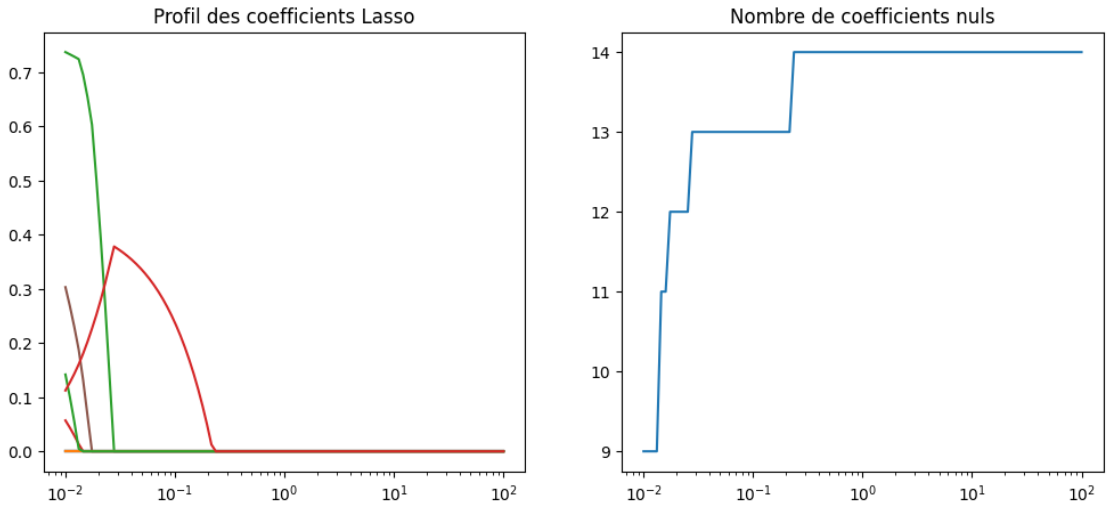


Figure 11

7.4 Modèles de régression non linéaires

7.4.1 Régression polynomiale

```
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import PolynomialFeatures

degrees = [1,2,3,4,5]
results_poly = []
y_pred_poly=-1000000
for d in degrees:
    poly = PolynomialFeatures(degree=d, include_bias=False)
    X_poly = poly.fit_transform(X_scaled)

    X_tr, X_te, y_tr, y_te = train_test_split(
        X_poly, y, test_size=0.2, random_state=42
    )

    model = LinearRegression()
    model.fit(X_tr, y_tr)

    y_pred = model.predict(X_te)

    results_poly.append({
        "degree": d,
        "MSE": mean_squared_error(y_te, y_pred),
        "R2": r2_score(y_te, y_pred)
    })

results_poly=pd.DataFrame(results_poly)
best_degree = results_poly.loc[results_poly["MSE"].idxmin(), "degree"]

print (" Polynomial Regression" )
```

```
print ("␣R_2␣:␣" , results_poly.loc[results_poly["degree"] ==
   best_degree, "R2"].values[0])
print ("␣MSE␣:␣" , results_poly.loc[results_poly["degree"] ==
   best_degree, "MSE"].values[0])
```

7.4.2 Réseau de neurones

```
from tensorflow.keras import Sequential
from tensorflow.keras.layers import Dense, Input, Dropout
from tensorflow.keras.regularizers import l2
from sklearn.model_selection import train_test_split
from tensorflow.keras.callbacks import EarlyStopping
from sklearn.metrics import mean_squared_error, r2_score

X_tr, X_val, y_tr, y_val = train_test_split(
    X_train, y_train, test_size=0.2, random_state=42
)

early = EarlyStopping(
    monitor="val_loss",
    patience=20,
    restore_best_weights=True)

def build_model(n_features, arch=(128,32,16), lr=1e-3, l2_reg=0.0,
   dropout=0.0):
    model = Sequential([Input(shape=(n_features,))])
    for units in arch:
        model.add(Dense(units, activation="relu",
           kernel_regularizer=l2(l2_reg)))
        if dropout > 0:
            model.add(Dropout(dropout))
    model.add(Dense(1))  # sortie linéaire

    model.compile(
        optimizer=tf.keras.optimizers.Adam(learning_rate=lr),
        loss="mse",
        metrics=["mse"]
    )
    return model

search_space = [
    {"arch": (128,32,16), "lr": 1e-3, "l2_reg": 0.0,  "dropout": 0.0},
    {"arch": (128,64,32), "lr": 1e-3, "l2_reg": 1e-4, "dropout": 0.0},
    {"arch": (256,128,64), "lr": 5e-4, "l2_reg": 1e-4, "dropout":
       0.1},]

best = {"params": None, "val_mse": np.inf, "model": None, "history":
   None}

for params in search_space:
    model = build_model(
        n_features=X_tr.shape[1],
        arch=params["arch"],
```

```
            lr=params["lr"],
            l2_reg=params["l2_reg"],
            dropout=params["dropout"]
        )

        history = model.fit(
            X_tr, y_tr,
            validation_data=(X_val, y_val),
            epochs=500,
            batch_size=128,
            callbacks=[early],
            verbose=0
        )

        val_mse = min(history.history["val_mse"])
        if val_mse < best["val_mse"]:
            best.update({"params": params, "val_mse": val_mse, "model":
               model, "history": history})

losses= pd.DataFrame(model.history.history)
losses.plot()
plt.show()

y_pred_rn = best["model"].predict(X_test).flatten()
print("MSE :", mean_squared_error(y_test, y_pred_rn))
print("R_2  :", r2_score(y_test, y_pred_rn))
print("Best params :", best["params"])
```

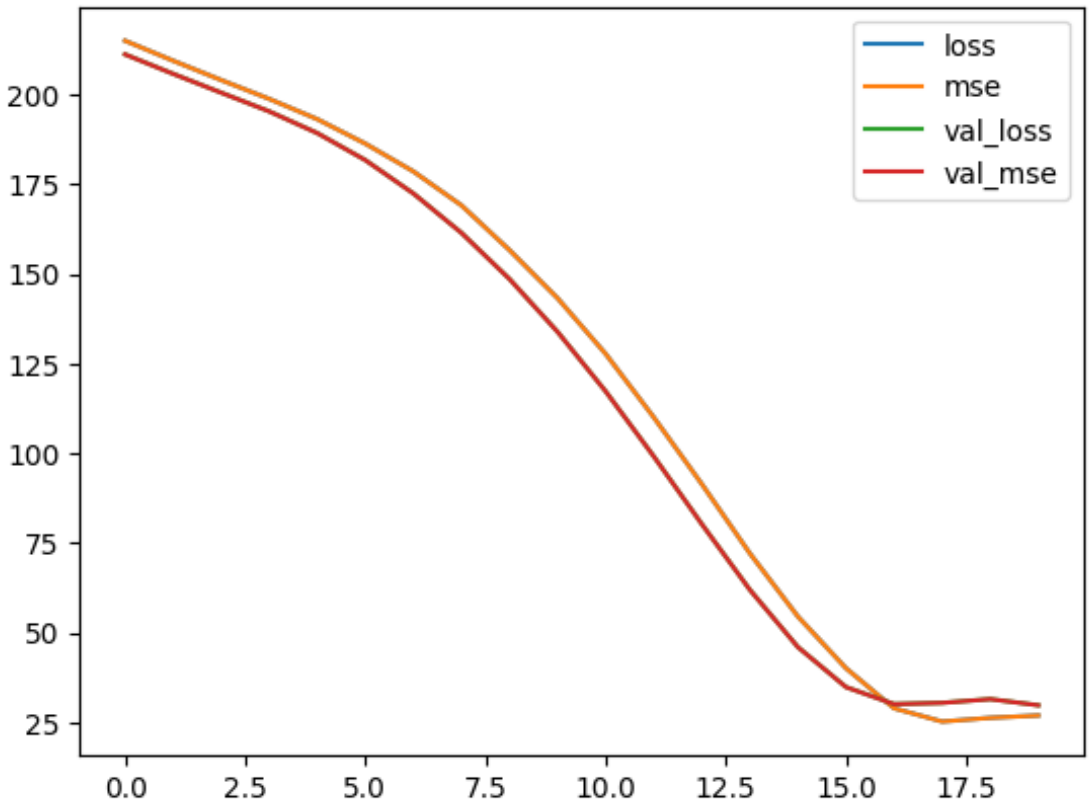


Figure 12: Évolution de la fonction de perte et de l'erreur quadratique moyenne pour le réseau de neurones

7.5 Application du modèle

Listing 3: Code d'application du modèle

```
import numpy as np
import pandas as pd

X_application, _ = pre_traitement(application_data)

y_application = ridge_cv.predict(X_application)

prix = np.round(np.exp(y_application), 2)

application_data['price'] = prix
```